Leximancer User Guide

Release 4.5

Leximancer Pty Ltd

Mar 10, 2021

CONTENTS

Contents:

CHAPTER

ONE

INTRODUCTION TO LEXIMANCER

- The learning materials in this section are designed to give the new Leximancer user an introduction to the workings of the program;
- This section also identifies some common applications of the software;
- Theoretical background on content analysis is then provided.

1.1 What is Leximancer?

Leximancer is a text analytics tool that can be used to analyse the content of collections of textual documents and to display the extracted information visually. The information is displayed by means of a conceptual map that provides a bird's eye view of the material, representing the main concepts contained within the text as well as information about how they are related (Fig. 1).

Essentially, this map allows the user to view the conceptual structure of a body of text, as well as perform a directed search of the documents. The interactive nature of the map permits the user to explore examples of concepts, their connections to each other, as well as links to the original text (Fig. 2).

In this way, Leximancer provides a means of **quantifying** and **displaying** the conceptual structure of text, and a means of using this information to explore interesting conceptual features.

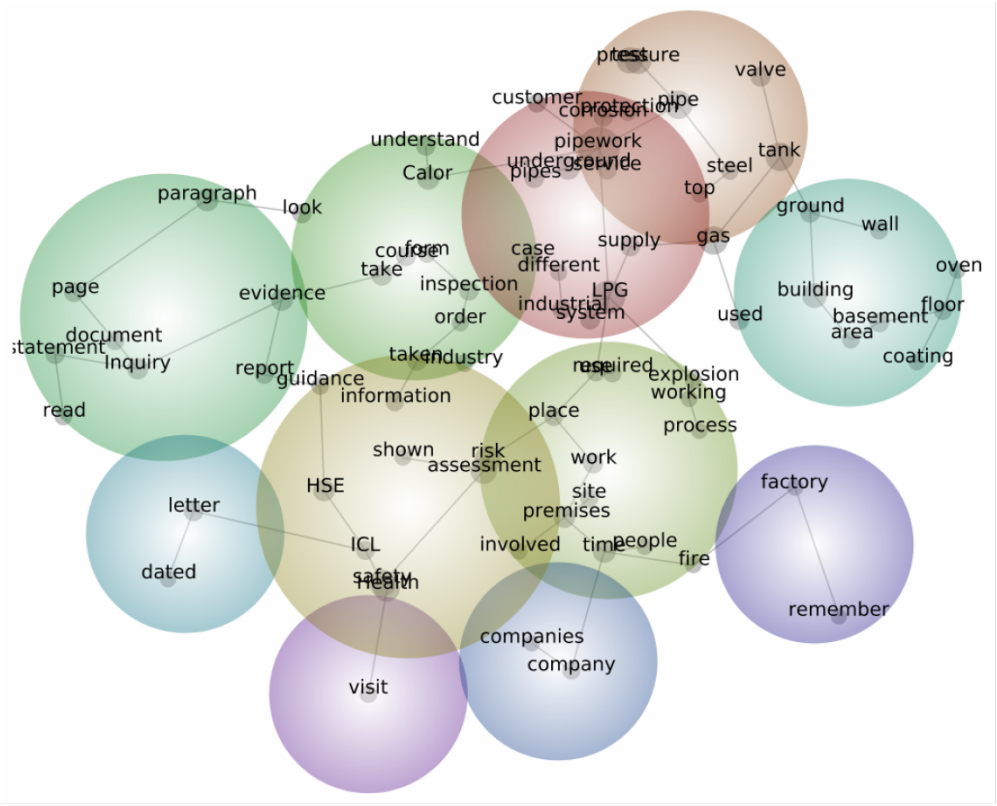


Fig. 1: Leximancer Concept Map

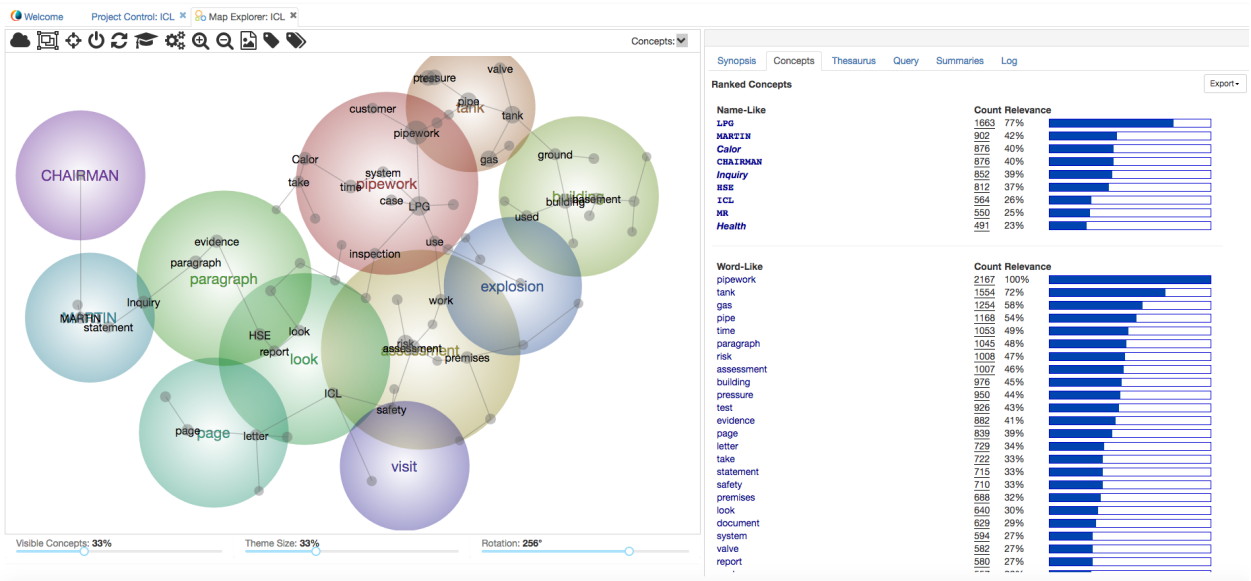


Fig. 2: Leximancer Concept Map and Ranked Concept List

1.2 Leximancer Applications

Table 1: Applications of Leximancer

Application	Type of Text	Output Options	Possible Projects
Basic Text Analysis	Any non-protected text: Word Docs, PDF, online content (html), .txt, .xml, etc.	Visual, via the Concept Map; Report, via the Insight Dashboard; Statistical, via Leximancer data exports.	Communication research; Analysis of speeches over time; Blog analysis.
Coding Open-ended Surveys	Customer feedback data; Call Centre data; Qualitative research spreadsheet data.	Statistical, via Leximancer data exports; Link open-ended questions to metadata.	Employee satisfaction survey; Net Promoter Score analysis.
Site and Archive Concept Navigation	Electronic content; Litigation evidence (electronic form)	Profile concepts being investigated; Concept co-occurrence data.	Legal e-discovery; Alternative to manually maintained site maps.
Media Analysis	Electronic media articles	Profile of company or issue	Competitor analysis; Online opinion analysis.
Customer Relationship Management (CRM)	Communication from customers	Current issues and concerns of customers	Policy and campaign development
Academic Research	Any	Concept Map; Statistical output.	History, Literature, Media Studies, Sociology, Politics

If you would like to see Leximancer in action, example maps can be found under on the Science page of the Leximancer website:

https://info.leximancer.com/

1.3 Theory: Content Analysis

- This section of the manual is for those wishing to understand more about the theoretical underpinnings of Leximancer;
- More practical, instructional chapters are to follow.

1.3.1 What is Content Analysis?

Content analysis is a research tool used for determining the presence of words or concepts in collections of textual documents. It is used for breaking down the material into manageable categories and relationships in order to quantify and analyze text.

Once extracted, these measurements can be used to make valid inferences about the ideas contained within the text (such as the presence of propaganda), properties of the writer or speaker (such as his or her psychological state), the audience to which the material is presented, or properties of the culture of the time in which the material was written.

Content analysis is an important research methodology as it can be used to analyse any form of verbal communication from written to spoken forms. As text documents tend to exist over long periods of time, the technique can be used to extract valuable historical and cultural insights.

As content analysis can be performed on numerous forms of data ranging from political speeches and open-ended interviews to newspaper articles and historical documents, it is invaluable to many researchers. Such uses include:

- historical analysis of political speeches
- detecting the existence and level of propaganda
- coding surveys that ask open-ended questions
- determining the psychological state of the writers
- assessing textual content against measures (e.g. censoring)
- assessing cultural differences in populations

1.3.2 Types of Content Analysis

In general, approaches to content analysis fall into two major categories: conceptual analysis and relational analysis.

In **conceptual analysis**, documents are measured for the presence and frequency of concepts. Such concepts can be words or phrases, or more complex definitions, such as collections of words representing each concept. One of Leximancer's main features is that it can automatically extract its own dictionary of terms for each document set using this information. That is, it is capable of inferring the concept classes that are contained within the text, explicitly extracting a thesaurus of terms for each concept. This approach also relieves the user of the task of formulating their own coding scheme.

Relational analysis, by contrast, measures how such identified concepts are related to each other within the documents. Leximancer measures the co-occurrence of concepts found within the text, automatically extracts this information, and represents the information visually for comparison. By doing so it displays the main relationships between concepts.

One of the strengths of the Leximancer system is that it conducts both forms of analysis, measuring the presence of defined concepts in the text as well as how they are interrelated. The following sections describe Leximancer's method for extraction of these concepts and their interrelationships.

1.3.3 Interested in Learning More about Content Analysis?

If you are interested in learning more about the issues relating to content analysis and the various techniques that are used, we recommend the following book: Weber, R.P. (1990) Basic Content Analysis. Newbury Park, Calif.: Sage Publications, 2nd ed.

CHAPTER

TWO

THE CONCEPT MAP

2.1 Theory: Concepts and Conceptual Mapping in Leximancer

Concepts in Leximancer are collections of words that generally travel together throughout the text. For example, a concept building may contain the keywords mill, warrant, tower, collapsed, etc.

These terms are weighted according to how frequently they occur in sentences containing the concept, compared to how frequently they occur elsewhere. Sentences are tagged as containing a concept if enough accumulated evidence is found.

Terms are weighted so the presence of each word in a sentence provides an appropriate contribution to the accumulated evidence for the presence of a concept. That is, a sentence (or group of sentences) is only tagged as containing a concept if the accumulated evidence (the sum of the weights of the keywords found) is above a set threshold (Fig. 3).

Aside from detecting the overall presence of a concept in the text, the concept definitions are also used to determine the frequency of co-occurrence between concepts. This co-occurrence measure is what is used to generate the concept map.

2.1.1 Concept Seed Words

In Leximancer, the definition of each concept (i.e. the set of weighted terms) is automatically learned from the text itself. Concept seed words represent the starting point for the definition of such concepts, with each concept definition containing one or more seeds.

They are called seeds as they represent the starting point of the concept, with more terms being added to the definition through learning. Occasionally, more appropriate central terms may be discovered, pushing the seeds away from the centre of the concept definition.

Leximancer automatically identifies concept seeds by looking for words that most frequently appear in the text. Alternatively, the user can manually provide seed words.

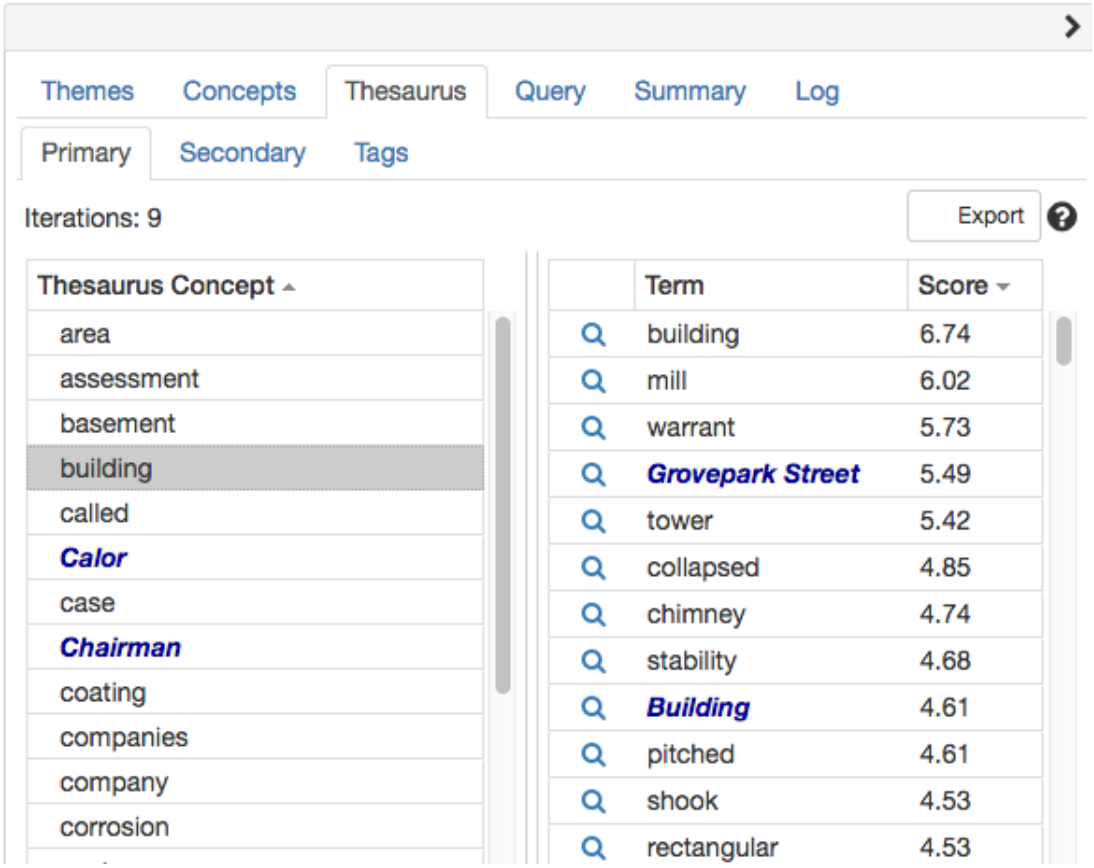


Fig. 3: Leximancer concept ‘building’ and associated thesaurus terms

2.1.2 Concept Learning

Leximancer begins with a set of seed words, as defined above. During the learning process, words highly relevant to the seed are continuously updated, and eventually form a thesaurus of terms for each concept.

Apart from adding highly relevant words to a concept, Leximancer may also add words that are negatively correlated with the concept (i.e. words that rarely appear in sentence blocks containing the concept and frequently appear elsewhere).

The aim of concept learning is to discover clusters of words which, when taken together as a concept, maximise the relevancy of all the other words in the document.

2.1.3 The Concept Map

Once Leximancer has run the learning process and developed a list of concepts contained in the text, and their relationship to each other, the information is presented via the Concept Map (Fig. 4).

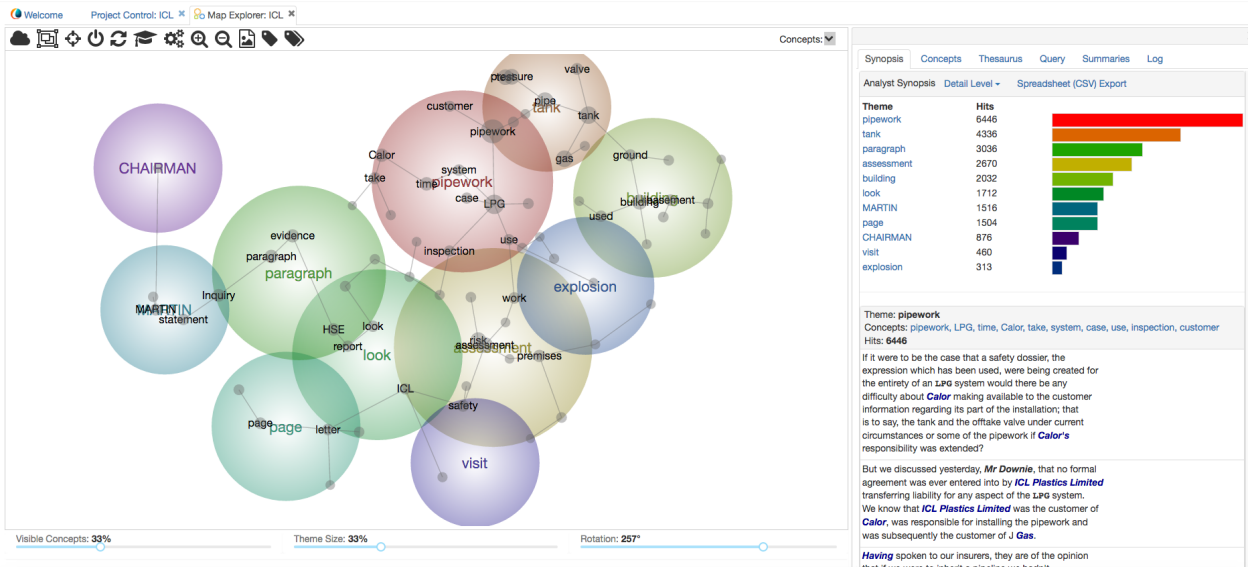


Fig. 4: Leximancer Concept Map

2.1.4 Understanding the Concept Map

The Concept Map is divided into two sections: a visual display of concepts and their relationships to each other on the left; and report tabs on the right for interacting with the concept map.

2.2 The Initial Display

When the map first opens, the top 50% of concepts are visible on the map. These are the concepts that appear most frequently in the text, and those that are most-connected to other concepts on the map.

Use the % Visible Concepts slider (beneath the map) to change the number of concepts visible on the map. Moving the slider all the way to the left hides all the concepts, and moving it all the way to the right reveals all the concepts.

2.3 Themes

The concepts are clustered into higher-level 'themes' when the map is generated. Concepts that appear together often in the same pieces of text attract one another strongly, and so tend to settle near one another in the map space. The themes aid interpretation by grouping the clusters of concepts, and are shown as coloured circles on the map (Fig. 5):

Here, a cluster of conceptually related concepts is grouped by the theme 'pipework'.

The themes are heat-mapped to indicate importance. This means that the 'hottest' or most important theme appears in red, and the next hottest in orange, and so on according to the colour wheel.

When the map first opens, the Theme Size is set to 33%, but you can move the Theme Size slider beneath the map to adjust the grouping of concepts on the map. Move the slider to the right to make fewer, broader themes, and move it to the left to make more, tighter themes (Fig. 6):

When the map first opens, the tab on the right presents a Summary of the Themes. A bar chart ranks the most important themes relative to one another, and beneath that the concepts visible within each theme are listed. A list of representative text excerpts is included for each theme, so that you can read some examples quickly to understand how and why the concepts in that theme appear together in the text. Click on the list of concepts for each theme to see all the text segments which contribute to the theme. At the top of this *full* list of text segments you can also see the query syntax that is used to discover the thematic texts.

In the screen shot below, the Pipework theme (shown as a red circle on the map) contains concepts such as *pipework, LPG* and *inspection.* Excerpts linking these concepts are shown in the Analyst Synopsis Tab on the right (Fig. 7):

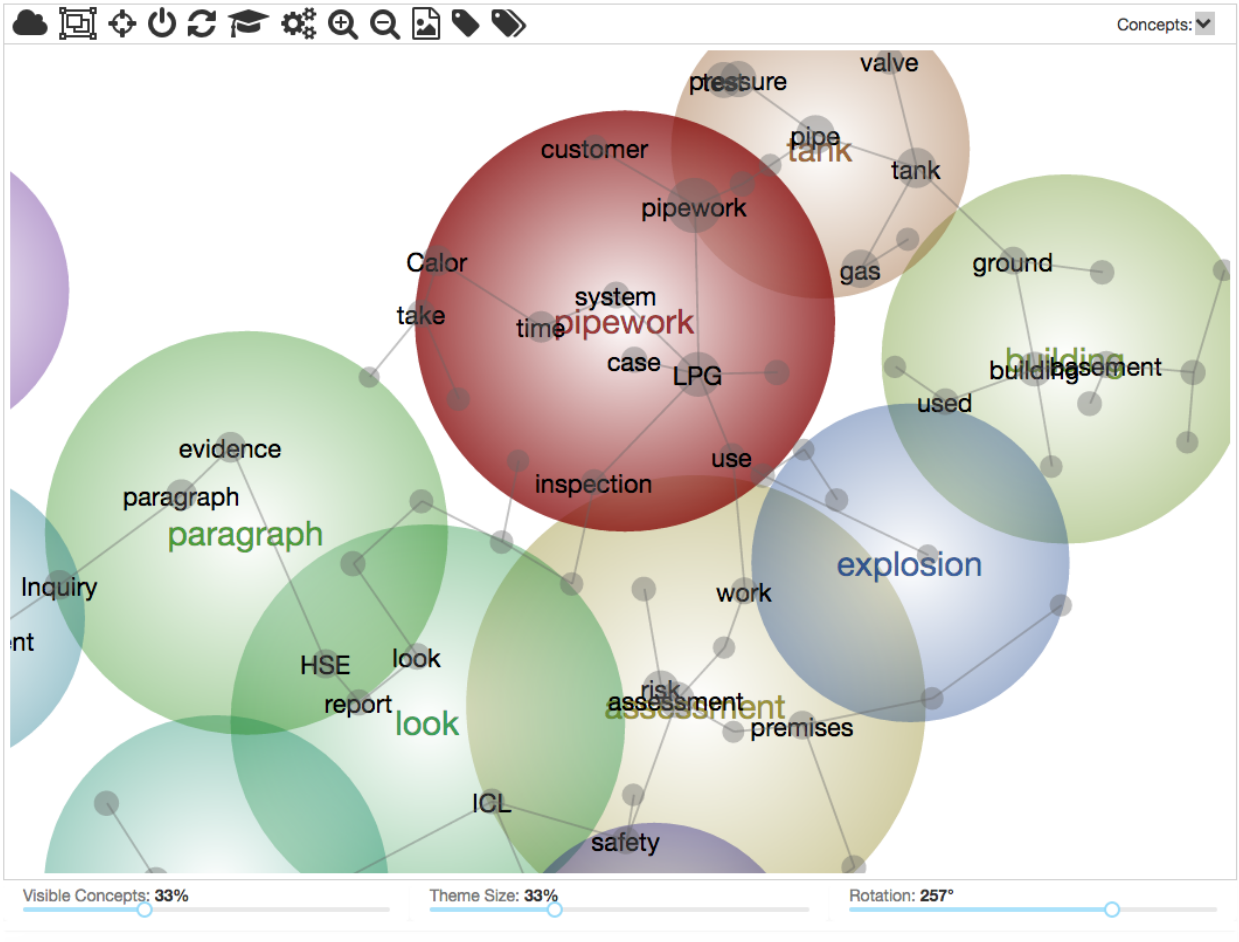


Fig. 5: Pipework Theme

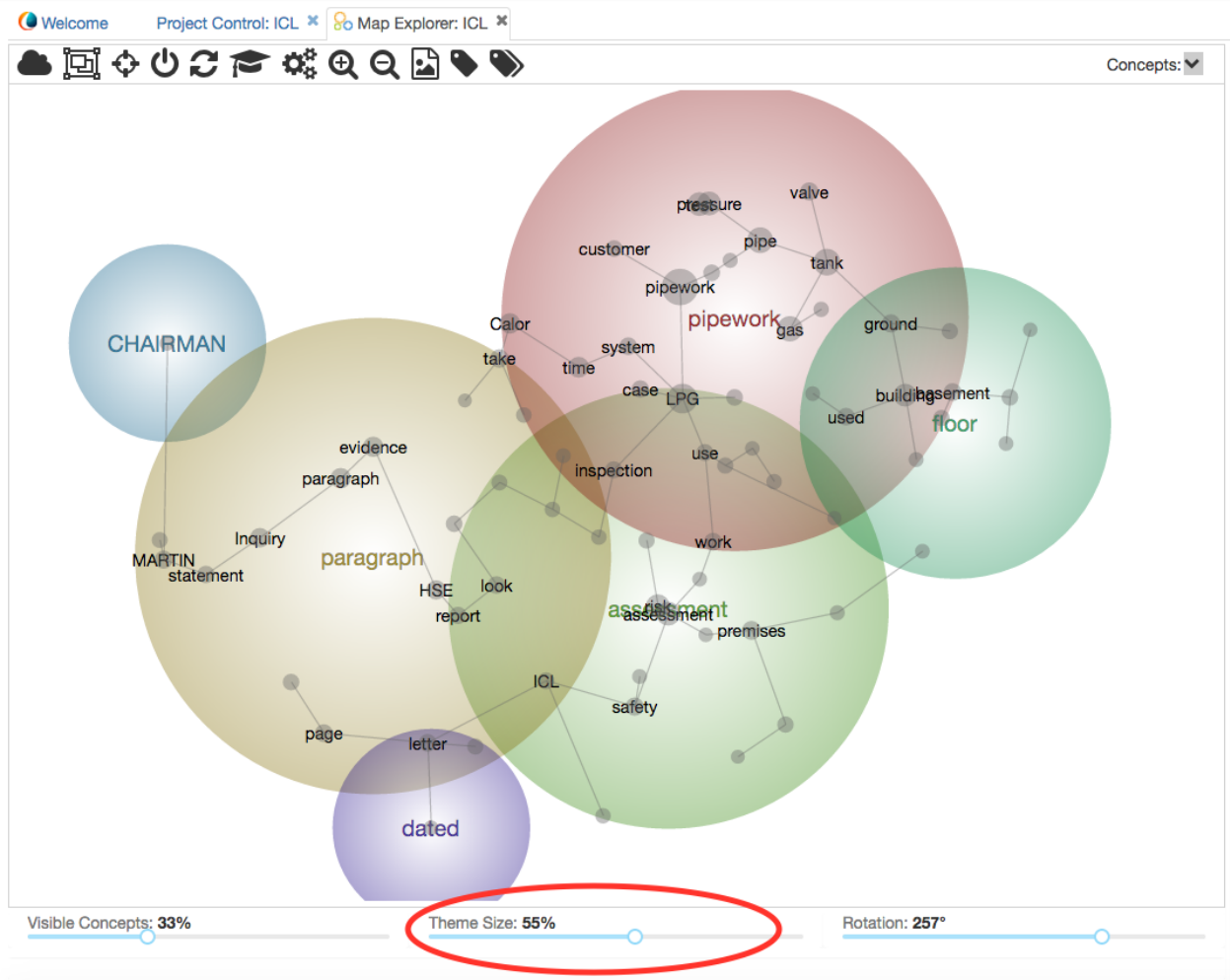


Fig. 6: Map Theme Slider Adjustment

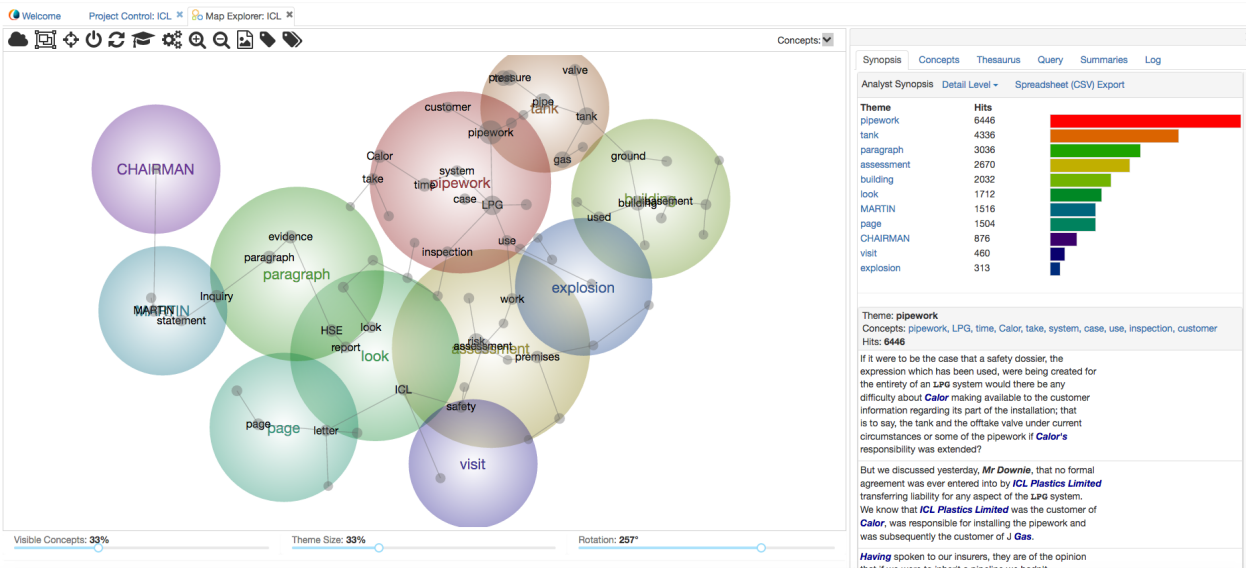


Fig. 7: Leximancer Concept Map and Analyst Synopsis Panel

If you adjust the size of the theme circles using the slider beneath the map, the Themes Summary updates to represent the new groups you have created on the map.

You can make all the themes disappear from the map by moving the Theme Size slider all the way to the left (0%).

If you hover your mouse over a theme circle on the map, the name of that theme will appear, if it isn't always visible (this behaviour is configurable).

Initially, each theme takes its name from the most connected concept within that circle.

You can change the names of the themes if you right click on the map near the theme name. A list of nearby concepts and themes will appear. Hover your mouse over the concept or theme of interest to get an option to Rename it (Fig. 8):

You can make all the theme names visible permanently on the map by clicking the Map Settings (gears) button in the header above the map and tick Theme Names Always Visible.

2.3.1 Concepts

The Concept Map contains the names of the main concepts that occur within the text. These are shown as grey labels on the map (Fig. 9).

Concepts written with an upper case first letter represent name-like (proper noun) concepts. These often include the names of people or locations, and appear with a capital first letter on the map. Examples below include *LPG* and *Calor*. All other word-like concepts, such as *tank* and *pressure*, appear in lower case on the map, and refer to objects, actions and so on.

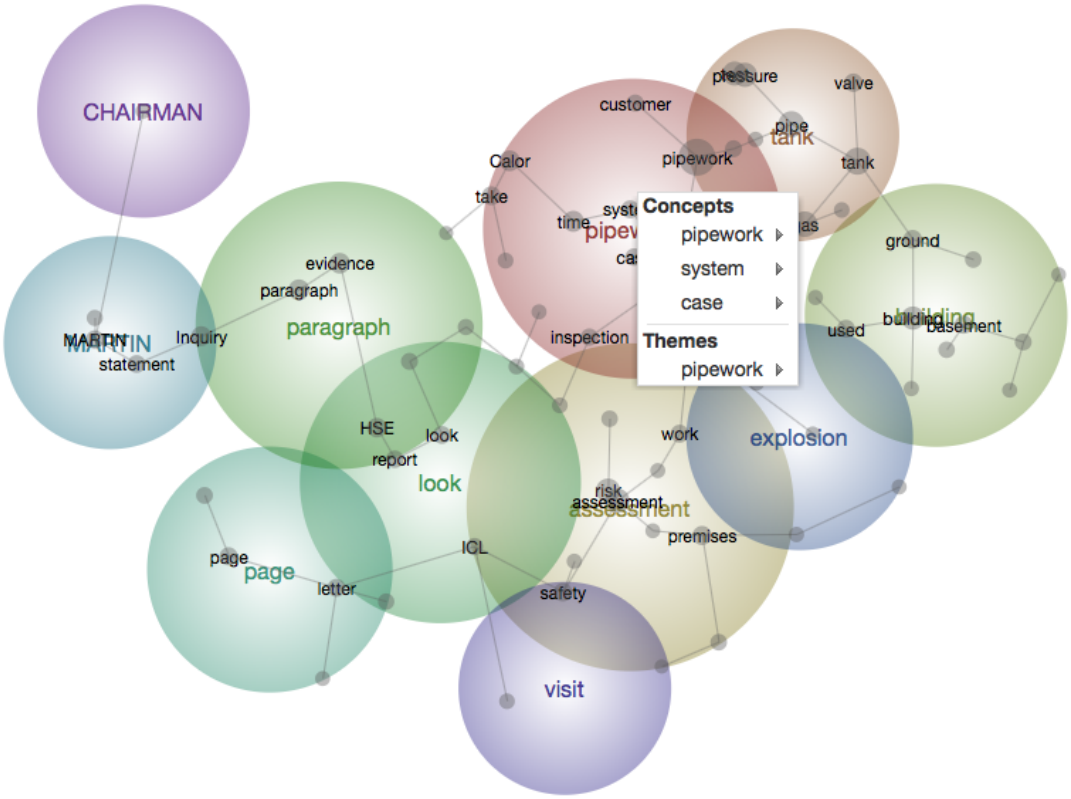


Fig. 8: Renaming a Theme

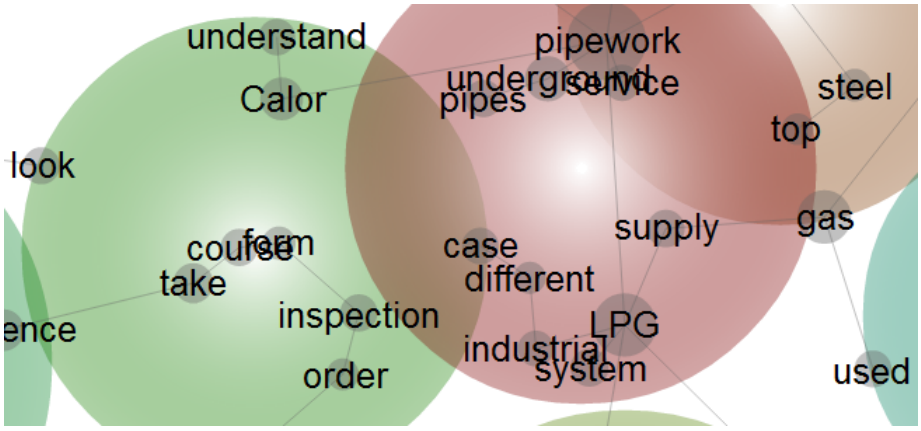


Fig. 9: Concept Map with Name-like and Word-like Concepts

2.3.2 Concept Display

The frequencies with which the name- and word-like concepts appear in the text are also listed separately in the Concepts tab on the right of the map (Fig. 10):

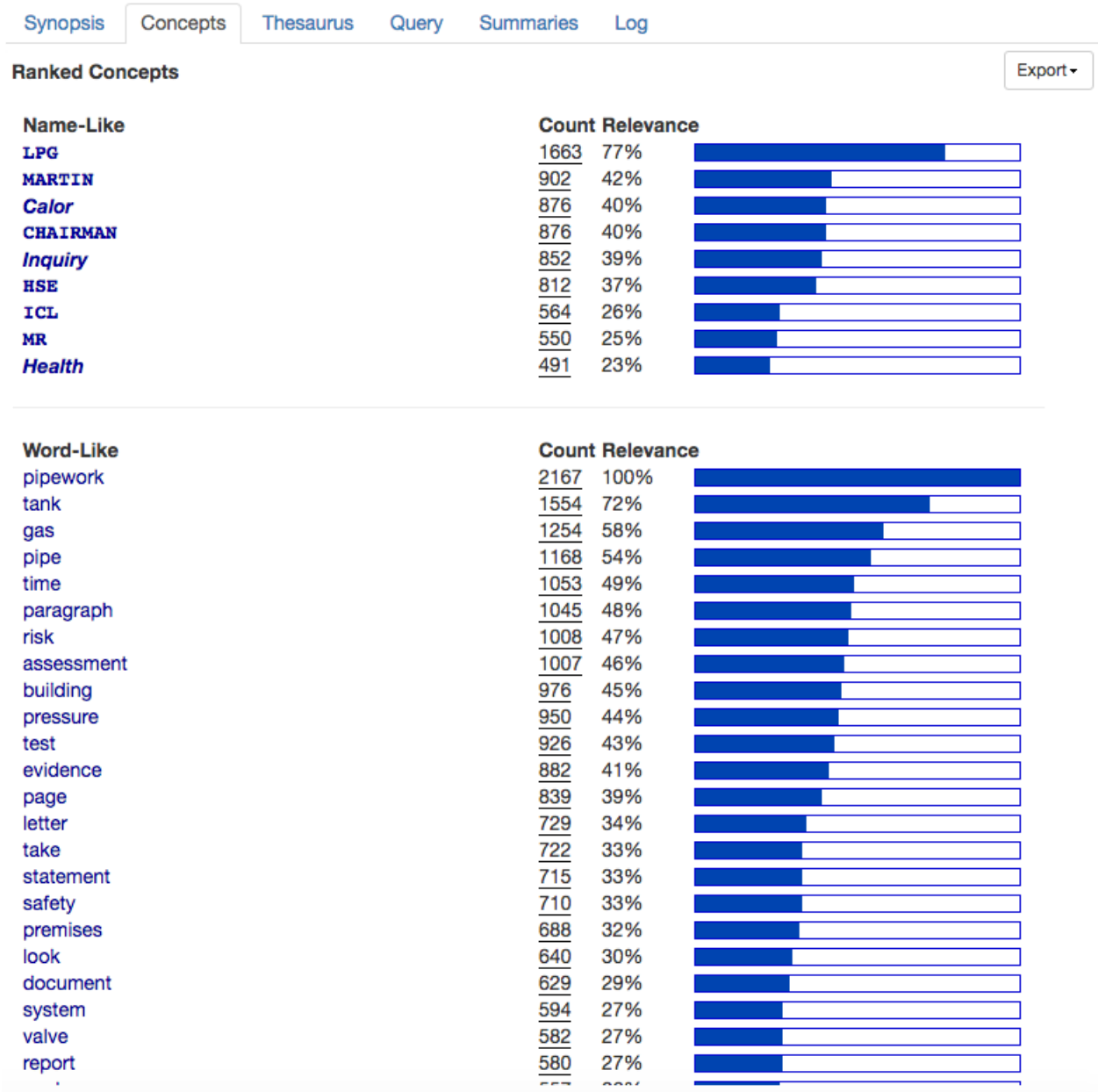


Fig. 10: Ranked List of Name-Like and Word-Like Concepts

The size of a concept's dot reflects its connectivity in the concept map. In other words, the larger the concept dot, the more often the concept is coded in the text along with the other concepts in the map. Connectivity in this sense is the sum of all the text co-occurrence counts of the concept with every other concept on the map.

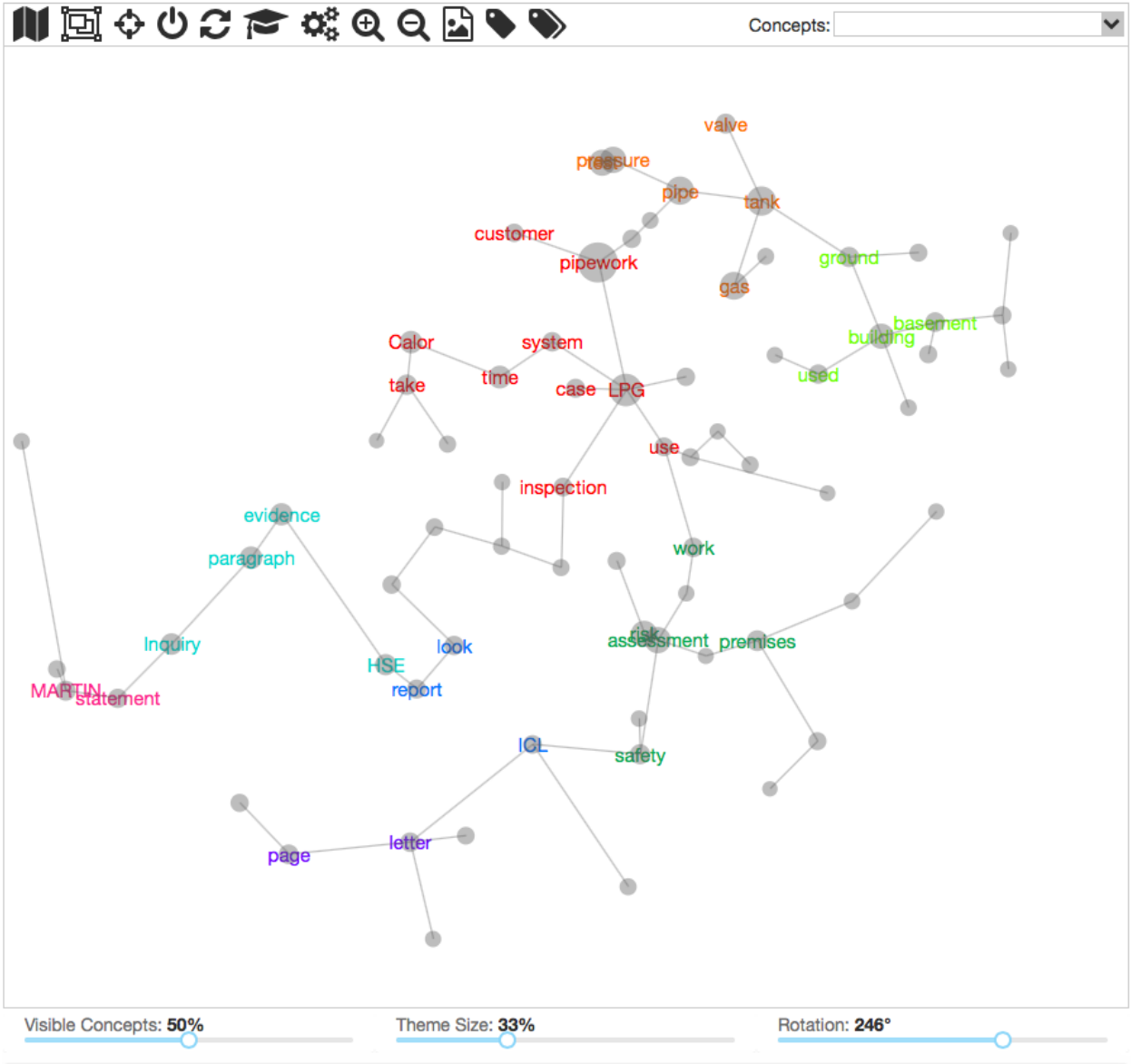


Fig. 11: Map Slider Controls

With concept visibility at 50%, some concepts do not appear on the map. The grey nodes illustrate where they would have appeared.

You can reveal hidden concepts by moving the % Visible Concepts slider (underneath the map) to the right. To reveal the most important concepts in order, move the slider far left then slowly drag the pointer to the right (Fig. 11).

2.3.3 Concept Map Toolbar

Fig. 12: Concept Map Toolbar

Hover your mouse over a button above the map to see its name.

Starting from the left, they include:

The *Concept Cloud* button reveals a graphical alternative to the concept map. The Concept Cloud will be explained in a separate section to follow.

The *Switch to Social Network* (Gaussian) button causes the concept to be reclustered using a Gaussian algorithm. Clicking the same button again returns you to the original view, created using a topical (Linear) clustering algorithm.

The *Center Map* button centers the map image in the screen space.

The *Reset Map* button returns the map to the way it looked when it first opened.

The *Recluster Map* button scatters the concepts randomly in the map space initially, then uses a clustering algorithm to allow the concepts to attract one another once more so as to lay the map out on screen.

The *Cluster Map* button allows the concepts more iterations of attracting one another to settle in stable locations on the map (without randomising them first).

The *Map Settings* button allows you to change various visual aspects of the concept map. Clicking the gear wheels button above the map opens this interface (Fig. 13):

In the Map Settings dialogue box you can increase the Font Size of labels on the map, and change the background colour from white to black.

It also allows you to make the theme names always visible on the map by ticking the *Themes Names Always Visible* checkbox.

You can choose whether to show the spanning tree on the map. The spanning tree appears as a grey network of connections between concepts (like a spider web) beneath the concept network. It shows the most-likely connections between concepts (like a road map of highways), but there are other (less-strong) connections between concepts (like backstreets).

Untick the Name, Word or Tag boxes to hide certain types of items on the map.

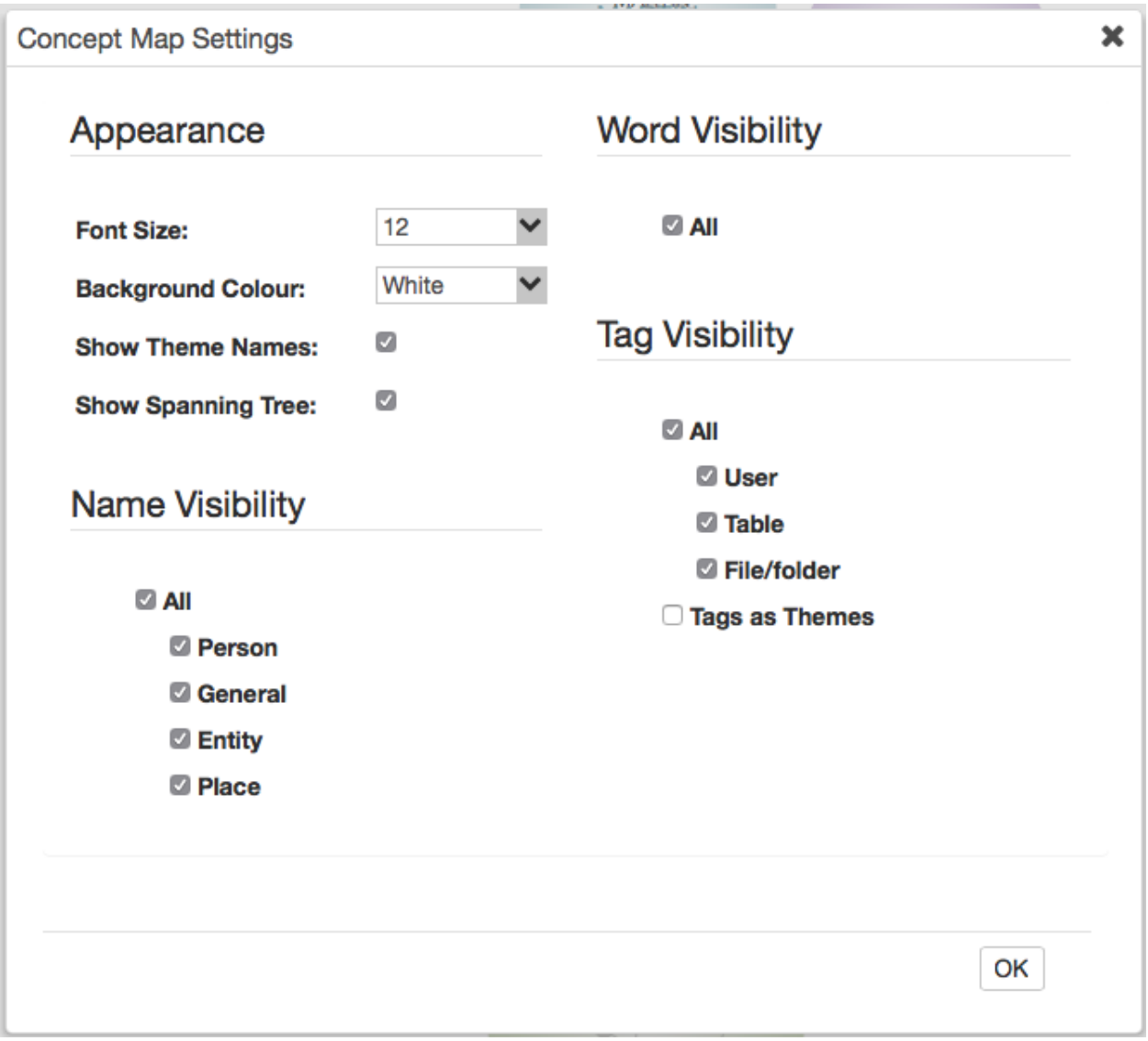


Fig. 13: Concept Map Settings

If your map contains tags, ticking Tags as themes allows the tags to give their names to themes if appropriate. Normally tags are not allowed to give their names to themes.

This map (Fig. 14) was created using the map settings shown above.

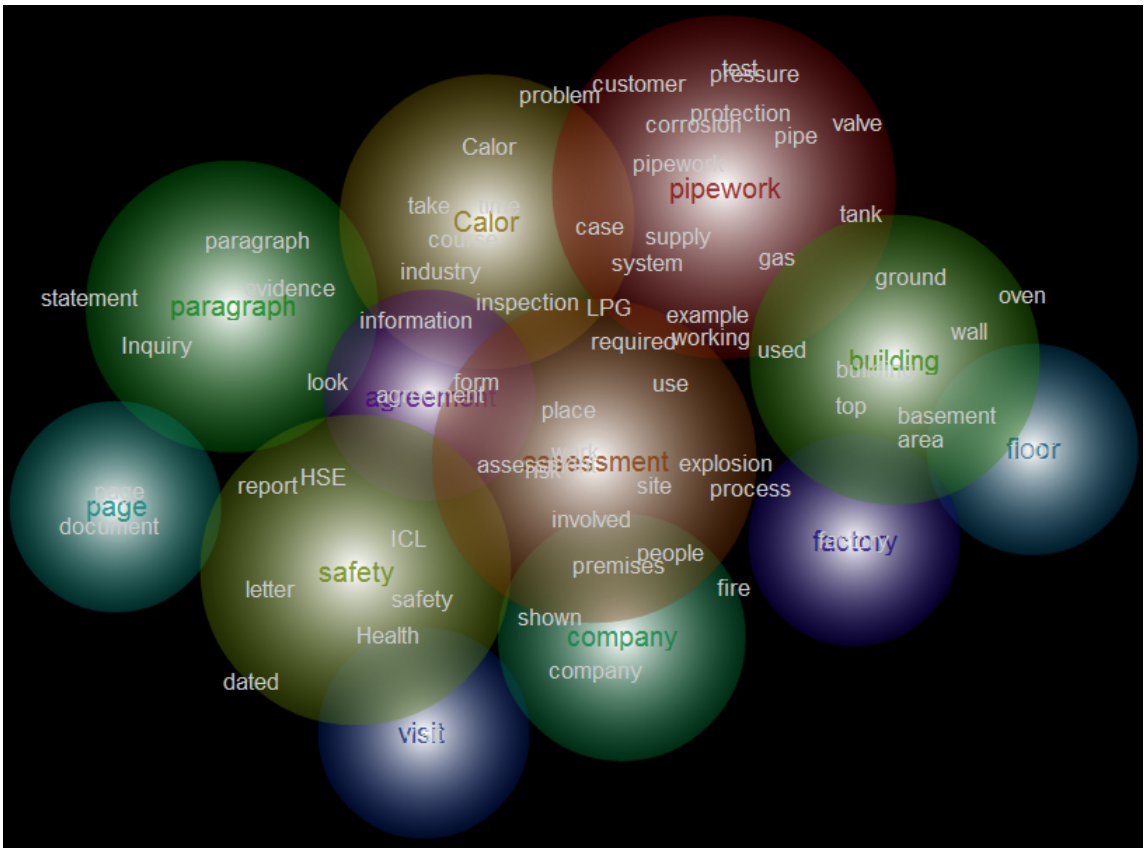


Fig. 14: Concept Map With Black Background

The *Increase and Decrease Map Zoom* buttons allow you to zoom your map view in or out.

The *Save Image* Map button allows you to export an image of the concept map in a number of format including png, jpg, and svg (Fig. 15).

You can choose the file type and resolution of the image.

Note that pop-ups must be enabled in the browser so that the map image can appear in a new window. The image can then be saved to local disk.

The *Bookmark Map* button lets you save the current map configuration (in case you have changed theme names etc.) so that you can return to it later. Simply enter a name for the map image, and click OK.

The *Load Bookmarked Map* button lets you reload a map configuration saved previously using the Bookmark Map button. Simply select the name of the map image you require and click Load.

The *Concepts* Dropdown List on the right above the map is an alphabetised concept finder. It allows you to locate (and select) a concept of interest in the map space. This is useful where you

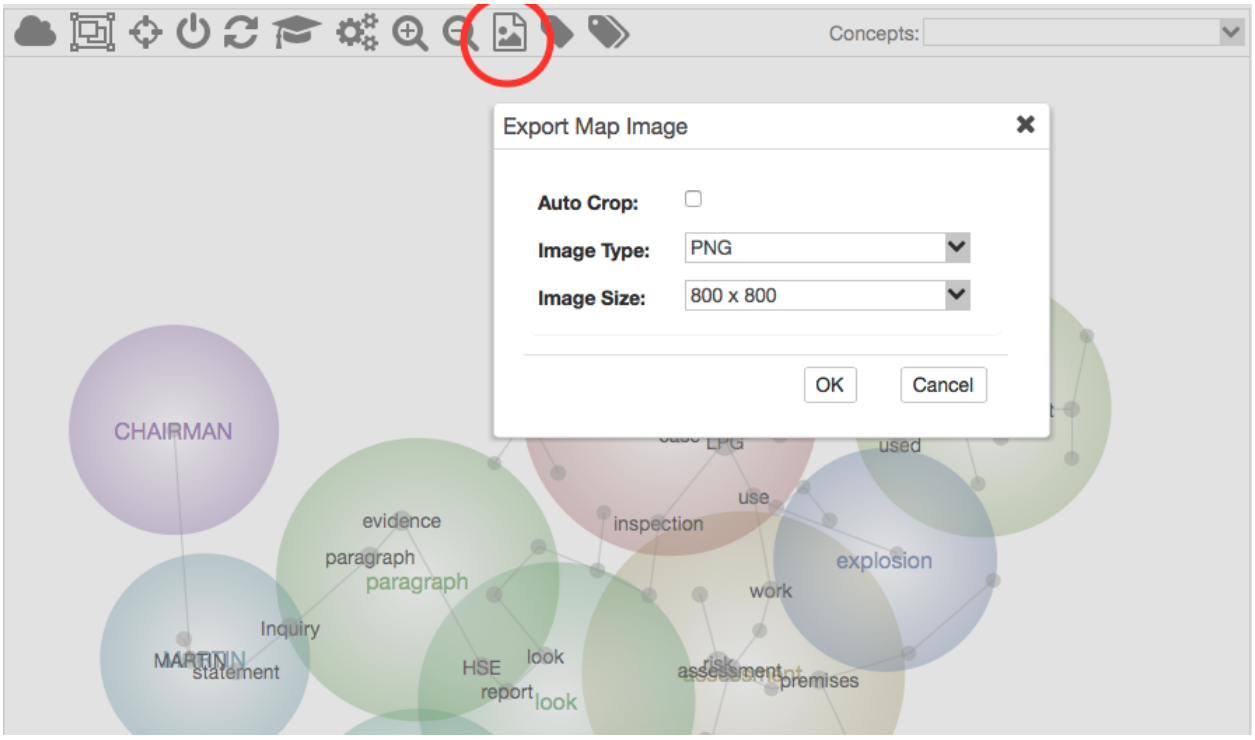


Fig. 15: Concept Map Image Export

know the name of a concept of interest, but cannot see it on the map easily.

2.4 The Concept Cloud

The project results can also be presented via the Concept Cloud. This is an alternative visual display tool, familiar to anyone who has seen a Tag Cloud on the Internet.

To display the cloud, press the button at the top-left of the Concept Map (Fig. 16):

The Concept Cloud, like the concept map, is heat-mapped, in that hot colours (red, orange) denote the most relevant concepts, and cool colours (blue, green), denote the least relevant. The font size of each concept's label denotes its frequency in the text.

The Concept Cloud is fully interactive. It behaves like the concept map, in that you can click on a concept (or tag) to select it and see the list of related concepts in the right-hand tab (Fig. 17):

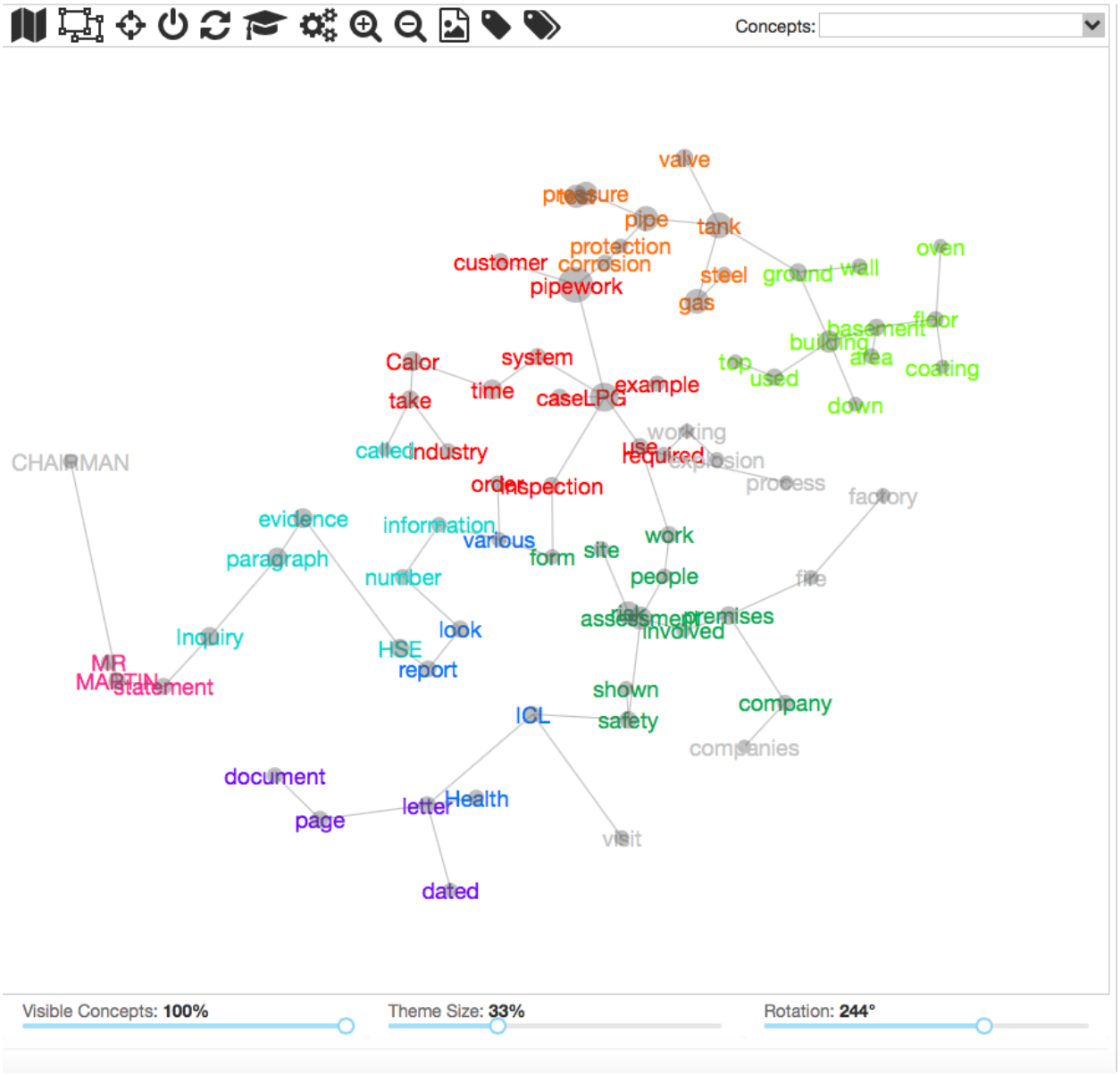


Fig. 16: Concept Cloud View

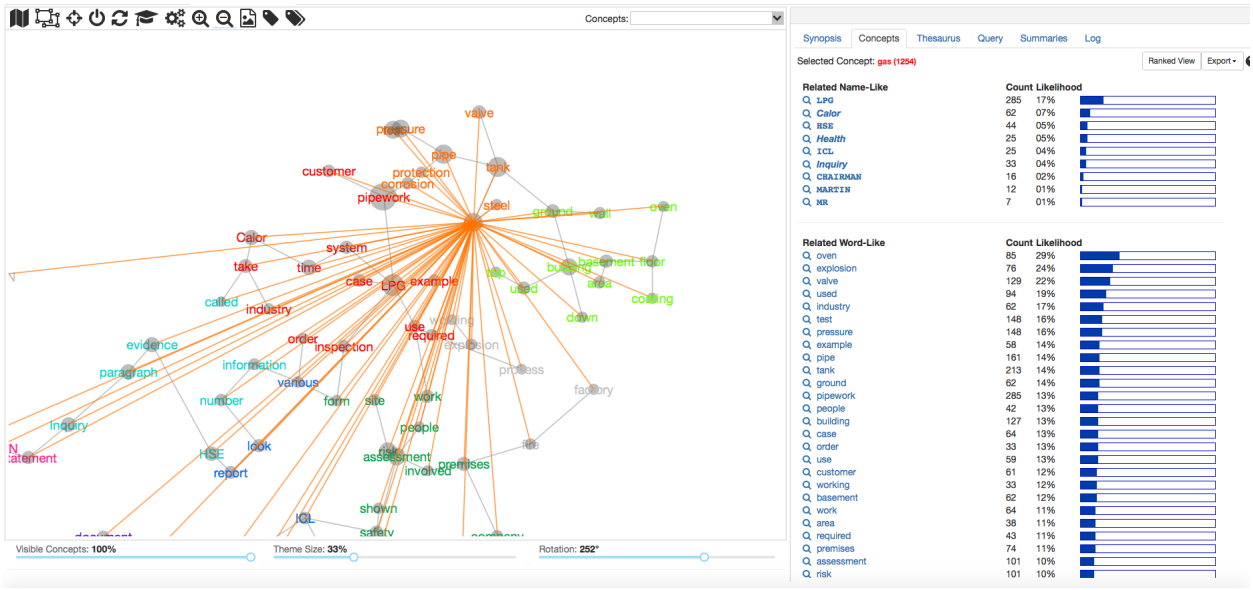


Fig. 17: Clicking on Concept Map to see Related Concept Details Panel

2.5 Report Tabs

The right-hand window contains several report tabs that allow for further interaction with the Concept Map. Each tab represents a different way to interact with the map and explore the results.

2.5.1 Analyst Synopsis

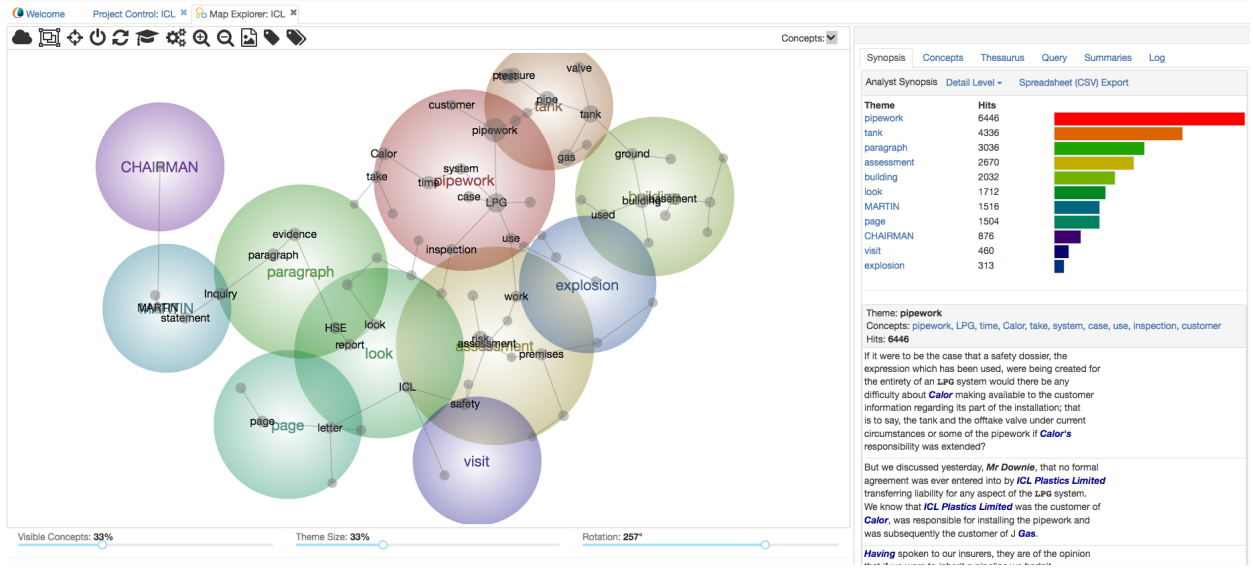


Fig. 18: Analyst Synopsis

As mentioned previously, themes are the coloured circles that group clusters of concepts. The themes are heat-mapped, meaning that hot colours (red, orange) denote the most important themes, and cool colours (blue, green), denote those less important.

The Analyst Synopsis tab (Fig. 18) is divided into two sections. The top section shows the themes ranked by their relative importance. The *Hits* column denotes the number of text blocks in the project associated with the Theme.

The bottom section shows each Theme, the concepts that are part of the theme, and the top 5 text hits for the theme. The list of concepts for a theme is a live query link. If you click on the theme's concept list, you will be taken to the Query Tab and can examine all of the text blocks that match this theme.

If you click on a theme label in the top section, the Analyst Synopsis Panel will scroll down to the details for that theme.

The *Detail Level* dropdown menu will adjust the visible concepts displayed in the Analyst Synopsis and the Concept Map on the left.

The *Spreadsheet (CSV) Export* will export all of the information contained in the Analyst Synopsis for further processing and analysis in other programs such as Excel.

2.5.2 Concepts

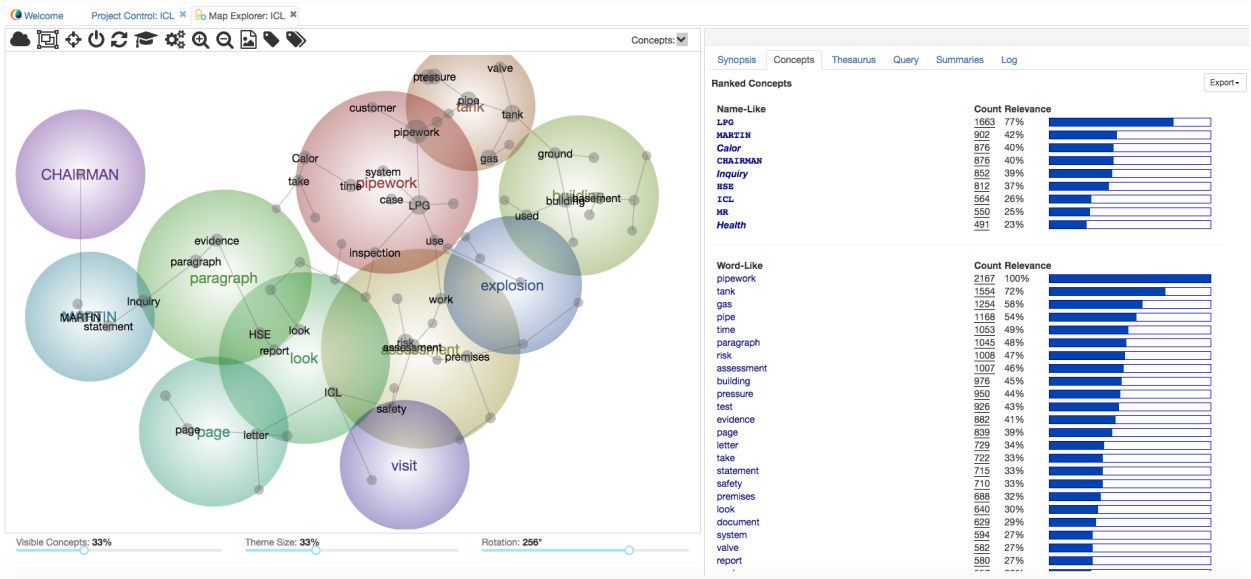


Fig. 19: Concept Frequency

Clicking the Concepts tab (Fig. 19) displays a list of name-like and word-like concepts, ranked by their frequency of occurrence in the text. Clicking on a concept in this list reveals its connections with other concepts (Fig. 20):

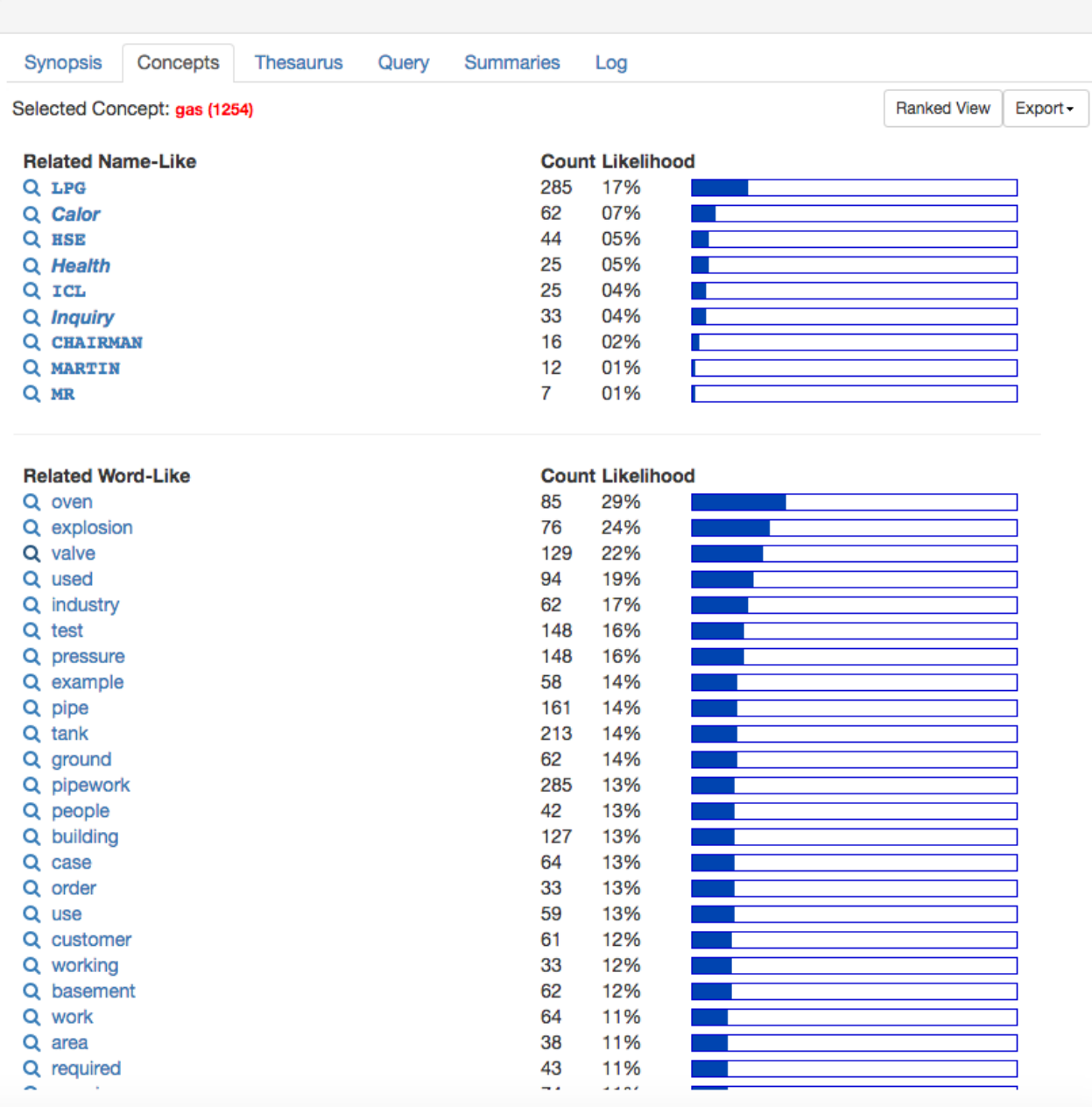


Fig. 20: Selected Concept Details

This information may also be accessed and displayed visually by clicking on a concept on the map itself (Fig. 21). The brightness of the ray indicates the strength of relationship (co-occurrence) between the concepts. A ranked list of the related name- and word-like concepts is displayed automatically in the Concepts tab on the right:

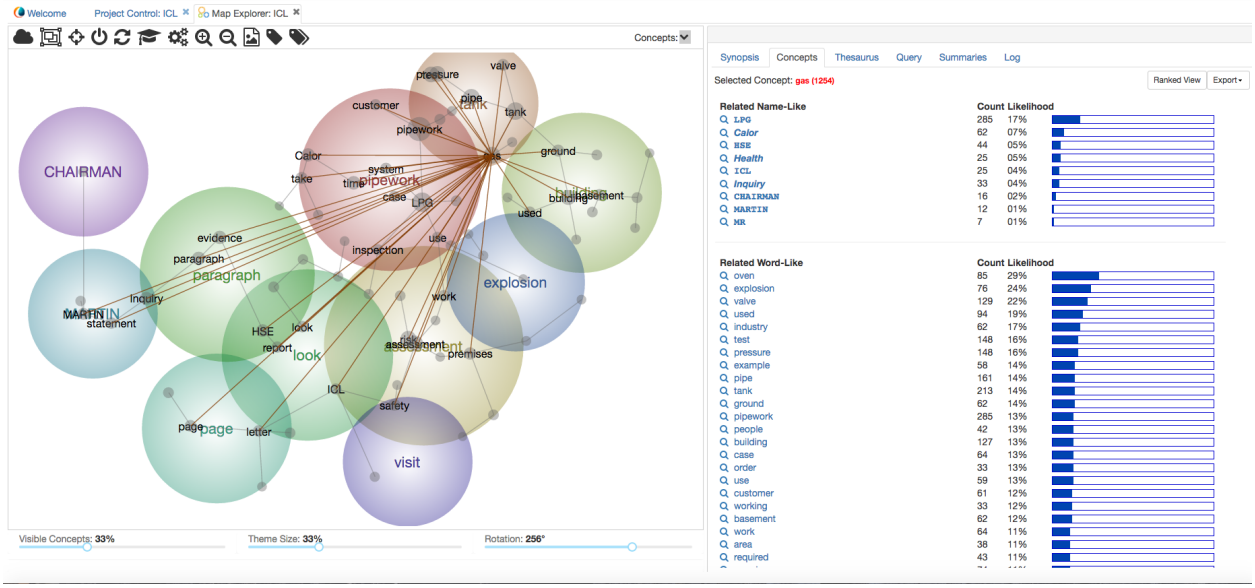


Fig. 21: Selecting a Concept on the Map

From the related concepts lists, you can browse the locations in the document where concepts co-occur by clicking in the Browse button (the magnifying glass icon) (Fig. 22).

Browsing extracts automatically takes you to the Query tab in the right-hand panel, and displays instances where the concepts of interest co-occur in the text (Fig. 23).

Click on the text block title in any of the query result instances to read the text extract in context (Fig. 24):

From the Query Results, click on Add to Log to add an extract of interest to the LogBook for export or reporting. When an extract is logged, the Add to Log button changes to a View Log option. Click View Log to review the list of extracts in the LogBook, then select Edit to add your own notes about an excerpt (Fig. 25):

2.5.3 Thesaurus

The Thesaurus tab (Fig. 26) displays a list of your concepts, the number of iterations performed by the Thesaurus Learning system while generalising from the seed words to the concepts, and a ranked list of the thesaurus words that define and describe each concept.

Click on a concept in the alphabetical list on the left to reveal the list of words Leximancer found to be associated with that concept on the right. The thesaurus list also shows the relevancy weightings associated with each indicative word. The iterations count (top left) tells you the number of times

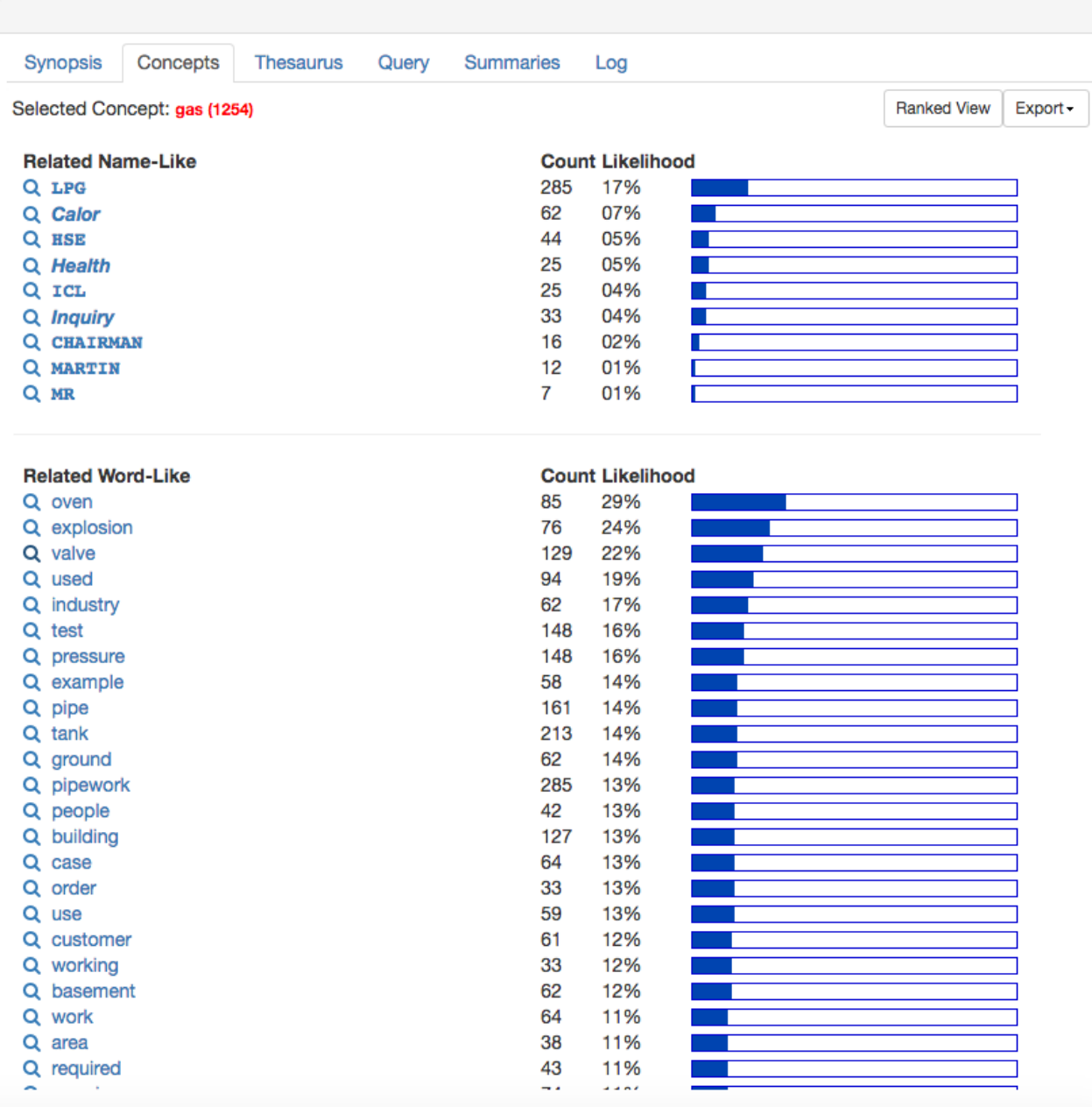


Fig. 22: Concept Co-occurences (click magnifying class to see text blocks)

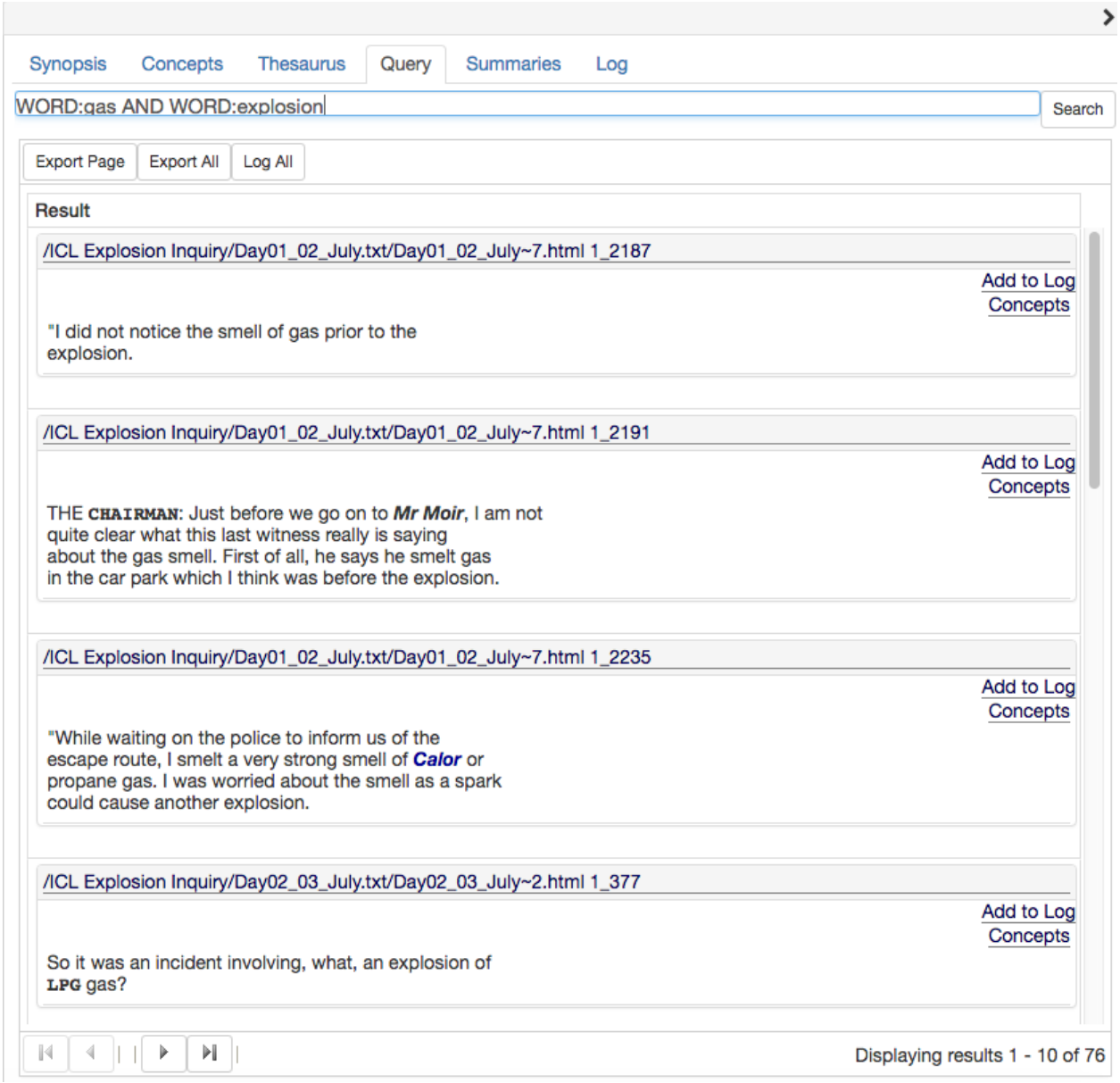


Fig. 23: Concept Co-occurence Query Panel

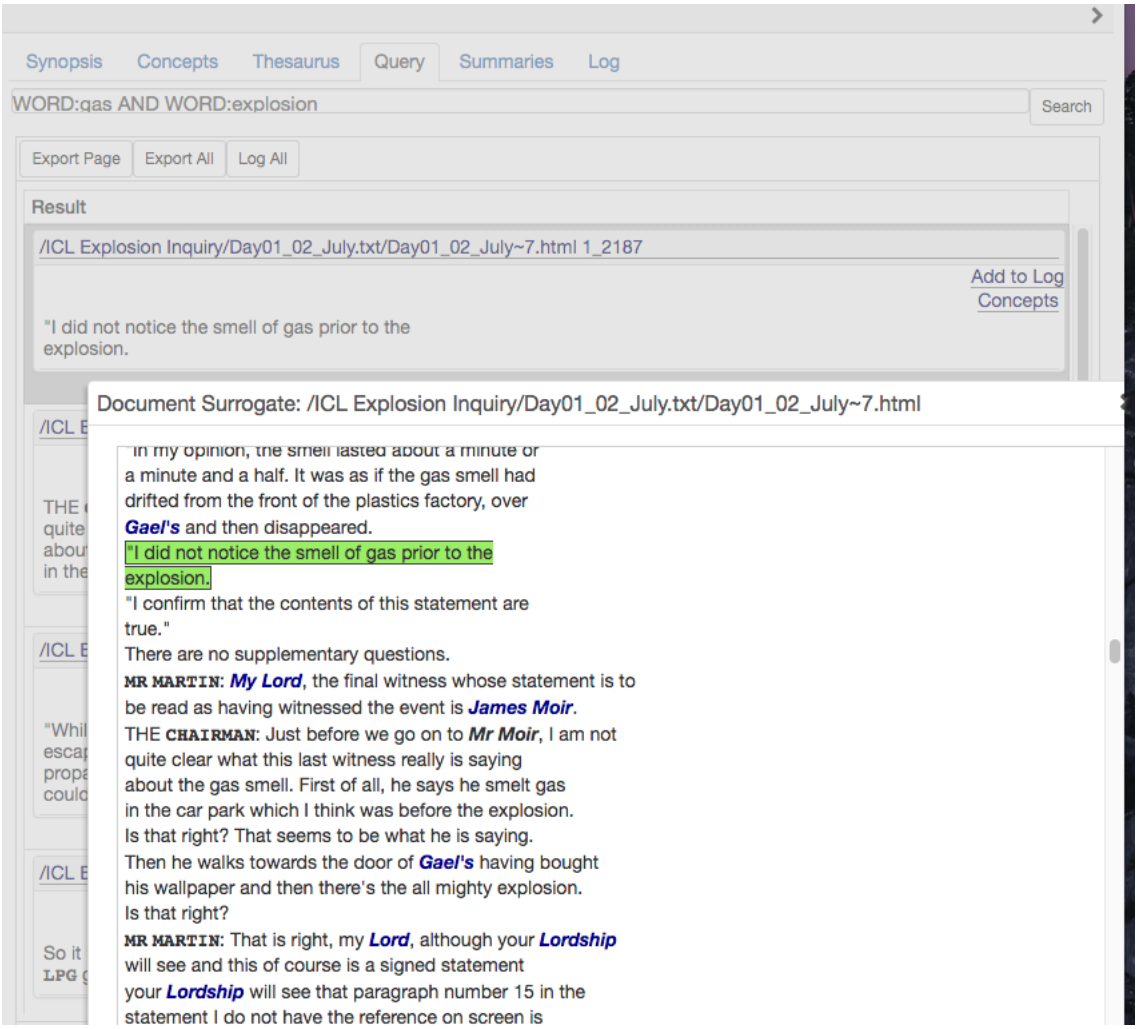


Fig. 24: Text Excerpt shown in Context

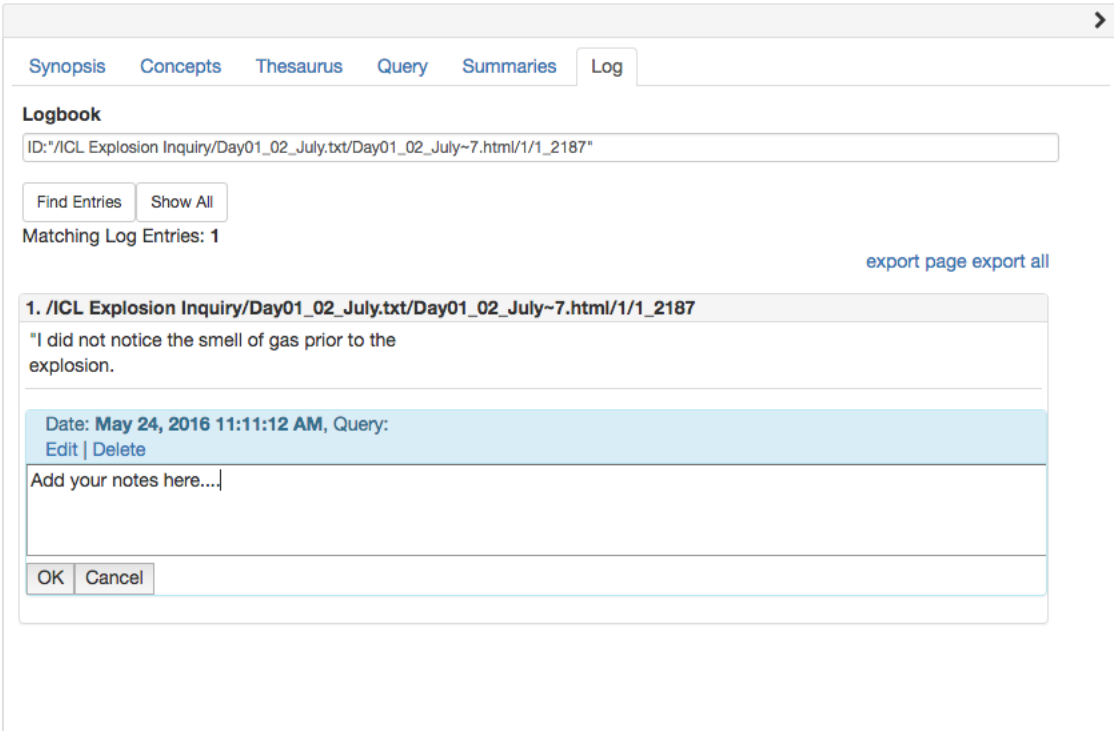


Fig. 25: Project LogBook

the corpus was reread and coded with evolving concept definitions before a stable classification result was achieved.

You can click on the Evidence button (the magnifying lens icon) to the left of a thesaurus item to browse text excerpts where that thesaurus term appears as evidence for a concept of interest (`thesaurus-query`):

2.5.4 Query

We have already discussed one of Leximancer's query functions: querying one concept against another:

Leximancer allows for more specific queries involving concepts and the raw terms (keywords) within the source documents (Fig. 27).

In the previous section, the thesaurus query made use of this to allow for a search for the keyword term *collapsed* (**WTERM:**collapsed) only where it co-occured with the concept *building* (**WORD**:building).

The full list of query modifiers Leximancer supports is listed below:

NAME:[concept] searches for a name-like concept.

WORD:[concept] searches for a word-like concept.

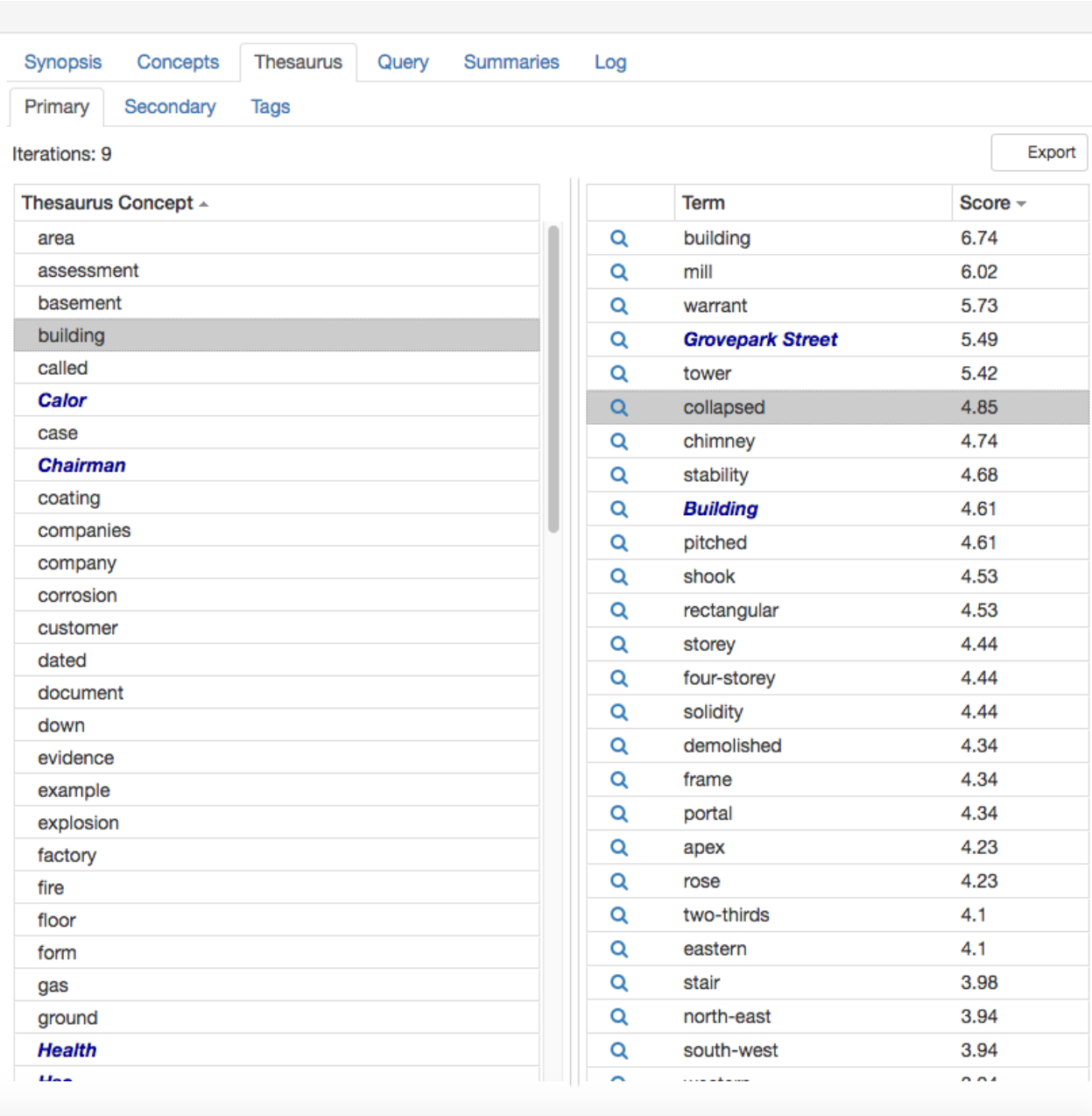


Fig. 26: Leximancer Thesaurus

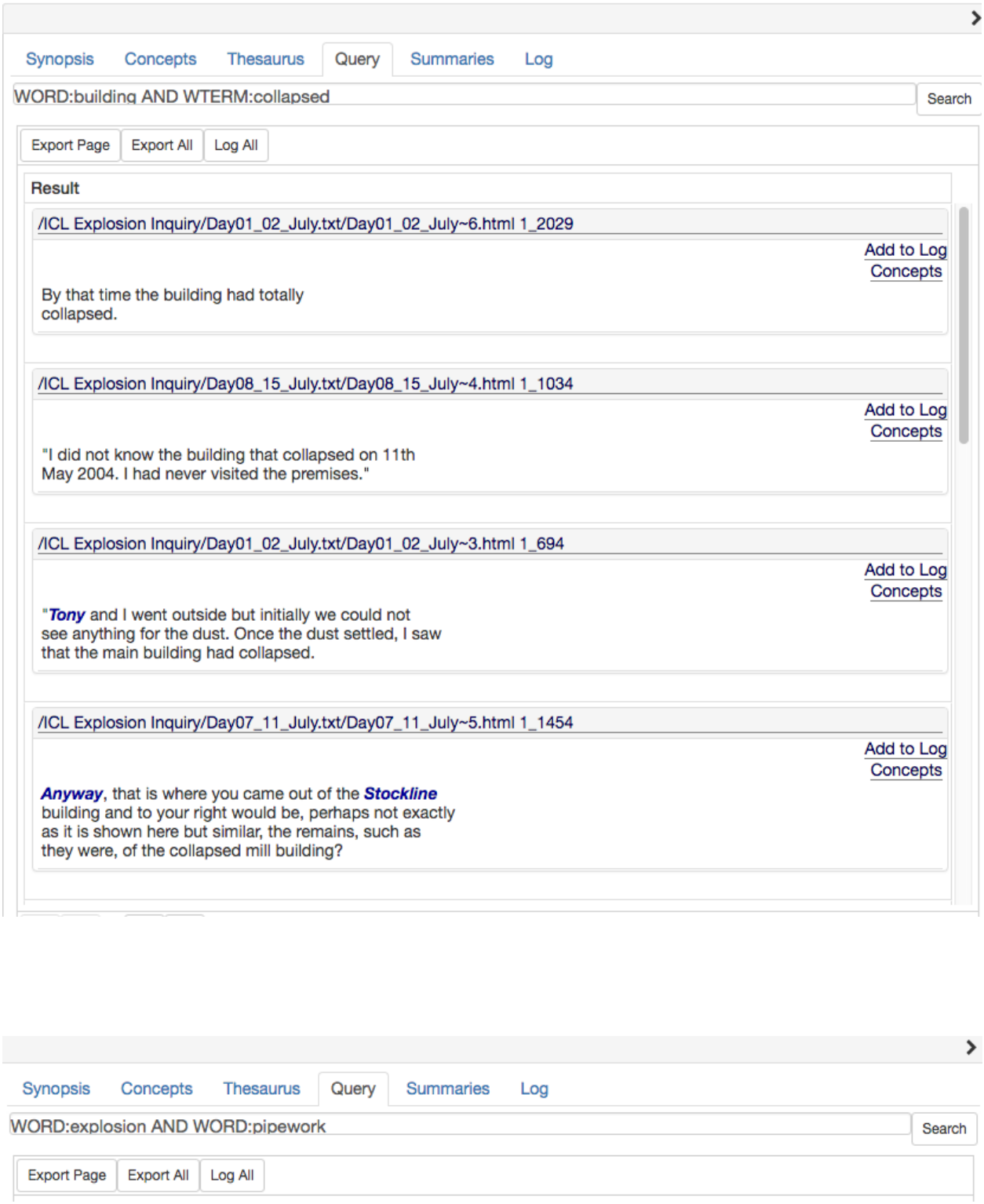


Fig. 27: Text Query Input Box

TAG:[file, folder, or tag] searches for a pre-defined tag.

WTERM:[term] searches for a regular keyword term. This is not a *concept*, just a word that appears in the text.

NTERM:[term] searches for a name-like keyword term. This is not a *concept*, just a word that appears in the text.

TERM:[word] searches for any keyword term, name-like or word-like. This is equivalent to (WTERM:[word] OR NTERM:[word])

These terms can be used in conjunction to search for co-occurrences of any number of specified concepts, tags, and / or keywords. For example, you could search for excerpts that mention the keyword term 'circumference' (WTERM:circumference) with the concepts 'pipework' (WORD:pipework) and 'explosion' (WORD:explosion) by entering this syntax into the Query box (Fig. 28):

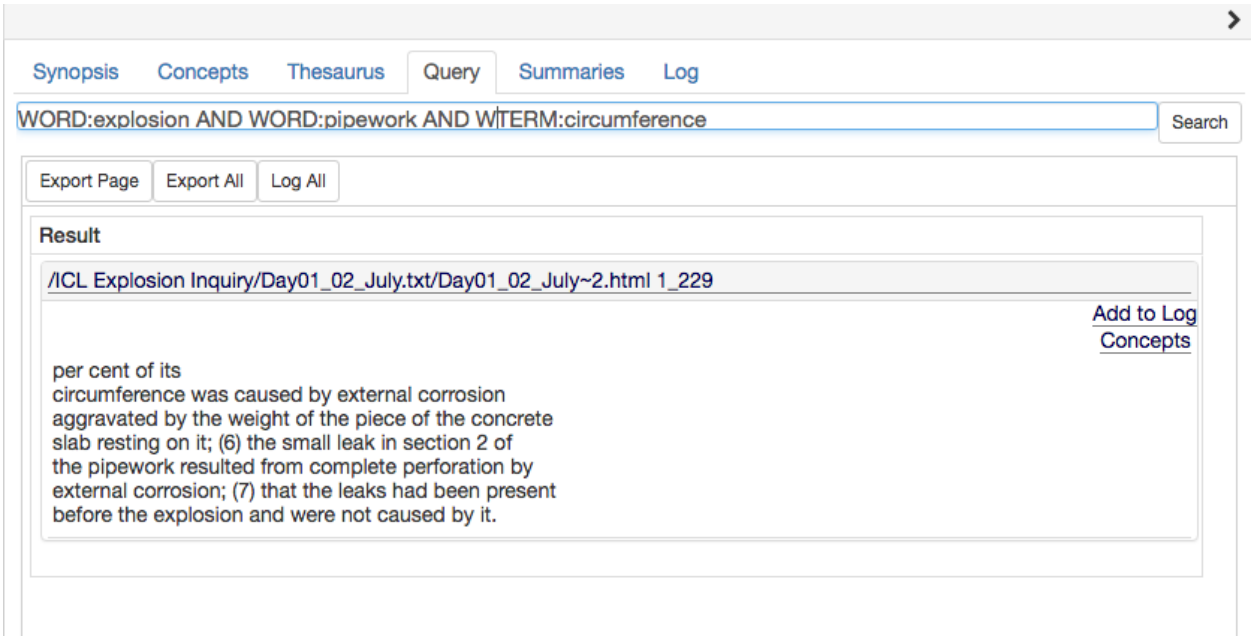


Fig. 28: Complex Query Example

In the project used for the above example, *circumference* is a term in the text but not a concept. It can still be used in queries by use of the **WTERM** query modifier.

There are also other, more specialized, query searches available:

ITAG:[file, folder, or tag] searches for your specified ignored tag. For example, a spreadsheet column with more than 500 values is classified as an ignored tag. Although ignored by Leximancer, you are still able to use it in your query to narrow results.

SECTION:[section number] searches for the section number of the text block

There are also some additional new tools to assist with query searches. These are query parsers that allow for more complex queries:

NAME:[partial concept]* adding a star to the end of a concept search will cause the query to be applied to the letters present and any derivatives. You must provide at least one character for this search. NAME:* alone will not produce results.

For example, a query of NAME:environment* will search for environment, environments, environmental, etc.

NAME:[concept] OR NAME:[concept]^2 searches for either concept specified, but with the latter concept to be considered more important

for example, a query of NAME:greenhouse OR NAME:emissions^2 will search for both 'greenhouse' and 'emissions' concepts, but 'emissions' concepts will be considered more important.

+NAME:[concept] +WORD:[concept] this is shorthand code for the search NAME:[concept] AND WORD:[concept]

+NAME:[concept] +WTERM:[concept] searches for the name concept co-occurring with a keyword

2.5.5 Summaries

The Summaries tab displays extracts containing the most important concepts discovered from the text. The list contains characteristic text segments that illustrate the relationships between key concepts (Fig. 29):

When you have finished exploring the Concept Map project tabs, close the Map Explorer tab using the Close button (X) in the top right-hand corner.

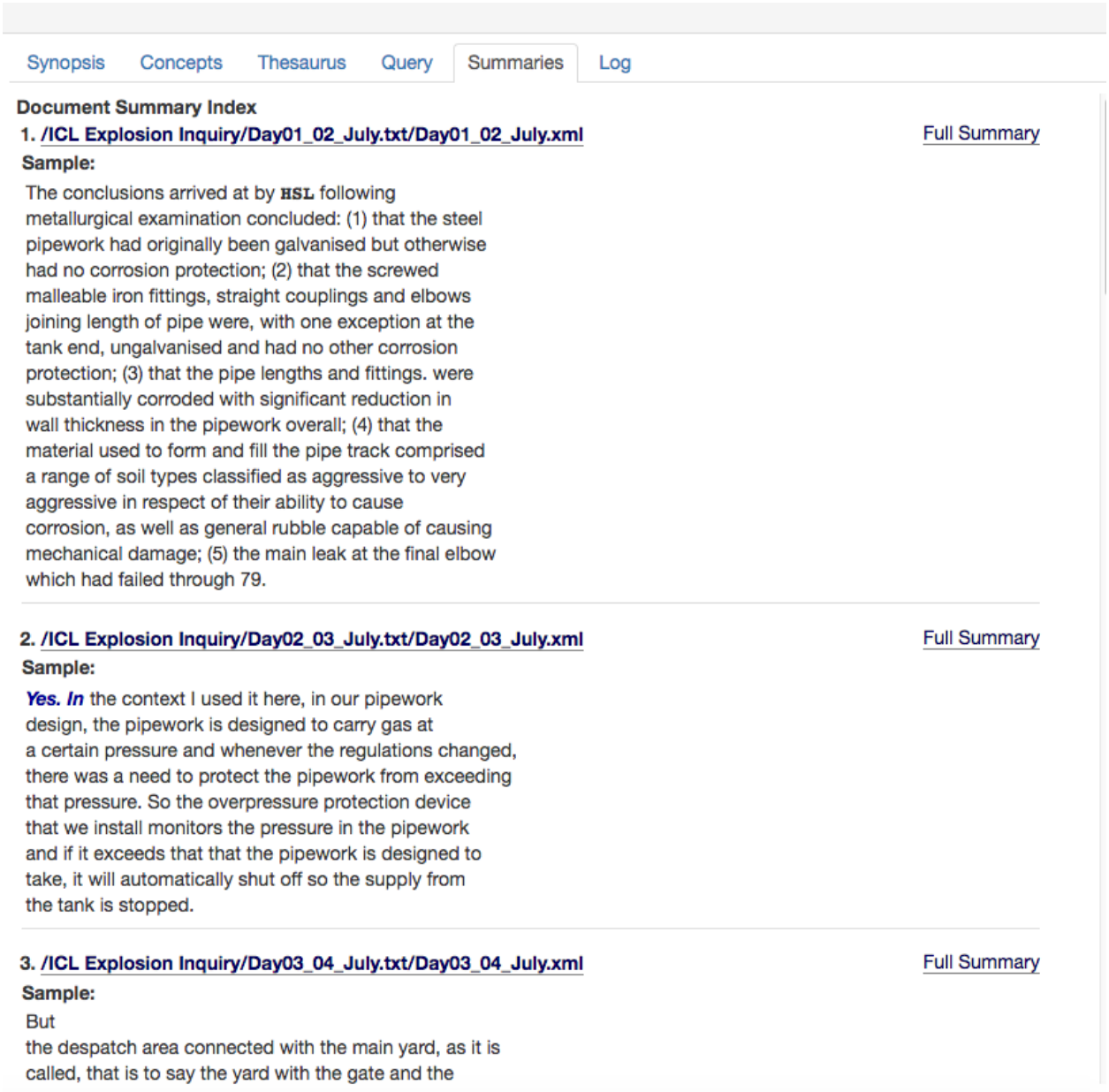


Fig. 29: Document Summaries

CHAPTER

THREE

CREATING AN AUTOMATIC/EXPLORATORY MAP

The aim of this chapter is to give you an overview and introduction to using Leximancer. In this chapter, you will be performing an exploratory analysis of a data set comprised of online media about climate change.

Exploratory maps involve minimal input from the user and are the starting off point for analysis using Leximancer. They are a means of gaining an overview of your data before adjusting Leximancer settings for more tailored results. Manually adjusted maps will be explored in the next chapter.

3.1 Creating an Automatic Concept Map

3.1.1 Supported File Types

Currently .doc, .docx, .pdf, .html, .txt, .tsv, and .csv files are supported. If there are other formats of text data you wish to process, consider converting them into plain text (.txt) or csv (.csv) first.

3.1.2 Desktop Installations

If you are working with a desktop version of the software, click on the desktop shortcut, or select Leximancer 4 from your Start menu. The first time you run Leximancer, you will be prompted to select your licence file (Fig. 30). If you have downloaded your licence key file, clicking in the box with the dashed outline allows you to navigate to the location where it's saved. You can also drag the key file icon into the box from your desktop.

A Leximancer icon will appear in your system tray or dock when the program is running. To exit the application, right click on the icon and select Exit Leximancer.

Leximancer Portal Accounts

If you are using the online Leximancer portal, open your internet browser and navigate to the Leximancer 4 url (http://www.leximancer.com/lexiportal/). Login using your username and password to start a session.

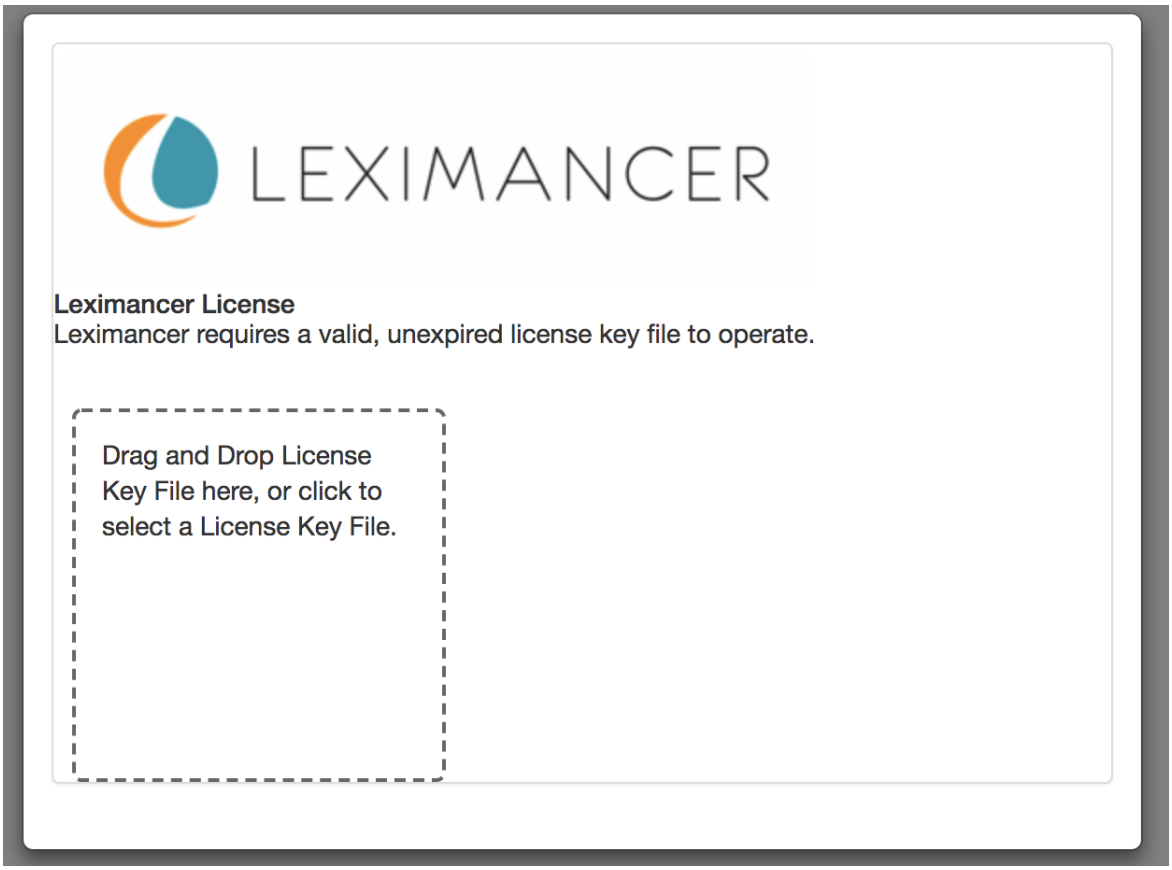


Fig. 30: Leximancer License Selector

Fig. 31: Leximancer in the Mac OS X Dock

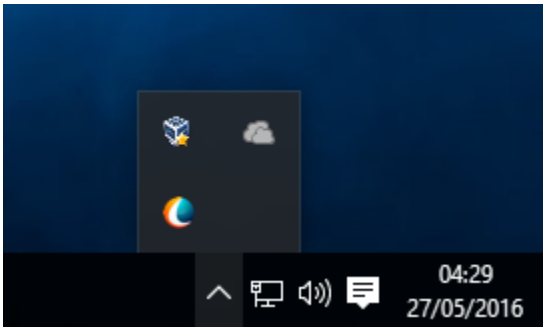


Fig. 32: Leximancer in the Windows 10 System Tray

3.1.3 Getting Started

When the program starts, you're presented with the Welcome Panel, and the Project Manager (Fig. 33):

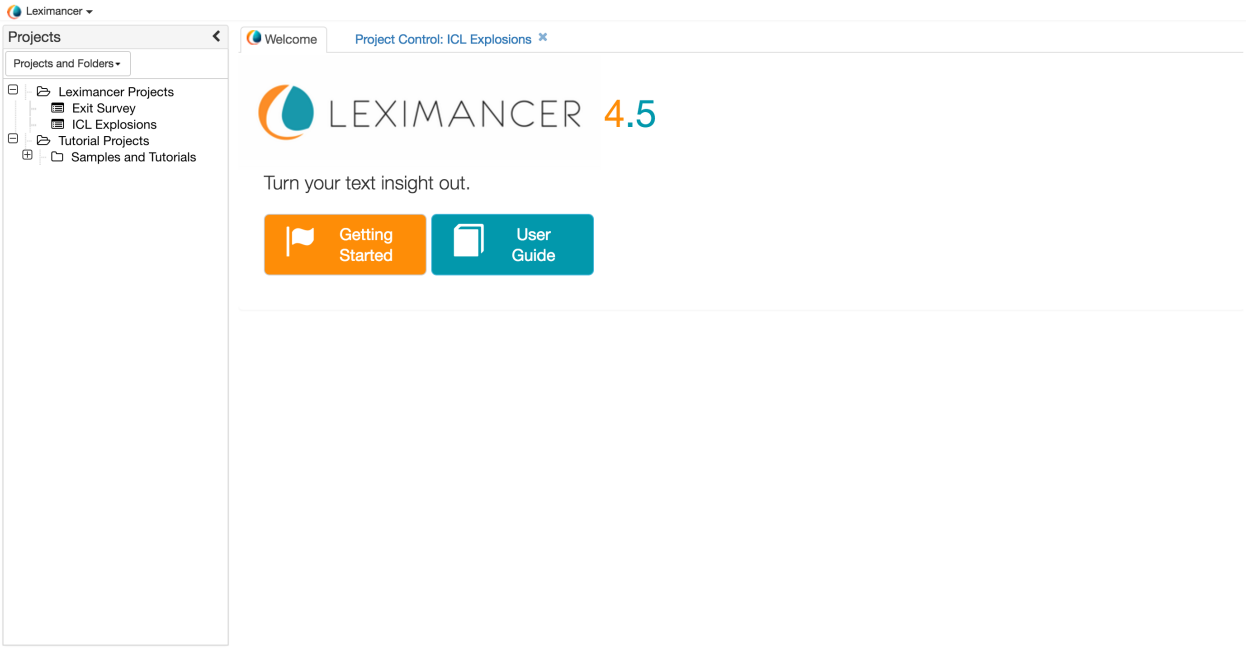


Fig. 33: Leximancer Welcome Panel with Project Manager

Two top-level folders are visible:

For Leximancer **Desktop**:

- *Tutorial Projects*, A folder containing a few small pre-loaded example projects.
- *Leximancer Projects*, A folder to house your personal projects.

For Leximancer **LexiPortal**:

- *Tutorial / Shared Projects*, A public folder accessible by all LexiPortal users that contains shared tutorial and other public projects.
- *User Projects*, Within this folder you will see a single folder with your user name, this is your **private** Leximancer project folder, which should be used for all of your personal projects. No other users have access to this folder.

LexiPortal Usage Note: The Tutorial / Shared Projects folder is shared and viewable by **all** users of the LexiPortal. Do not put your personal projects in this folder. Always put your projects in your folder within the User Projects folder.

3.1.4 Working with Project and Folders

Click on the plus sign adjacent to a folder to expand the tree and view any folders or files within. In this case, there is a project called 'ICL Explosions' inside the Leximancer Projects folder (Fig. 34):

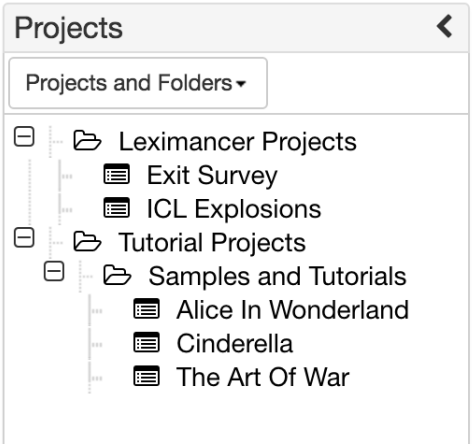


Fig. 34: Project Manager

- Double clicking on a project's name opens that project.
- Right clicking on a project allows you to Rename or Delete an existing project.
- Right clicking on Leximancer / User Projects, then selecting Create Project allows you to create your own new project.
- Create a hierarchy to organise your projects by housing New Projects in New Folders.

3.1.5 Creating a New Folder and Project

New Folder

Right click on Leximancer Projects (Desktop) folder, or open the User Projects folder and then the folder by your name (Portal), and select the New Folder option. You can then name (and optionally describe) the New Folder (Fig. 35).

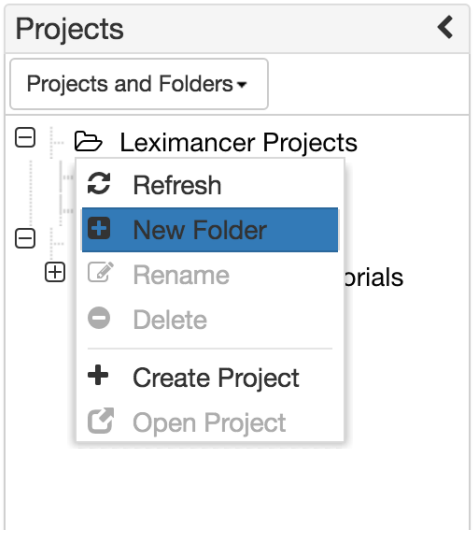


Fig. 35: Project Folder Menu

As a project file or folder is created using this name, try to use typical conventions for naming files (for instance, avoid including characters such as ".", "\", "*", or "/") (Fig. 36).

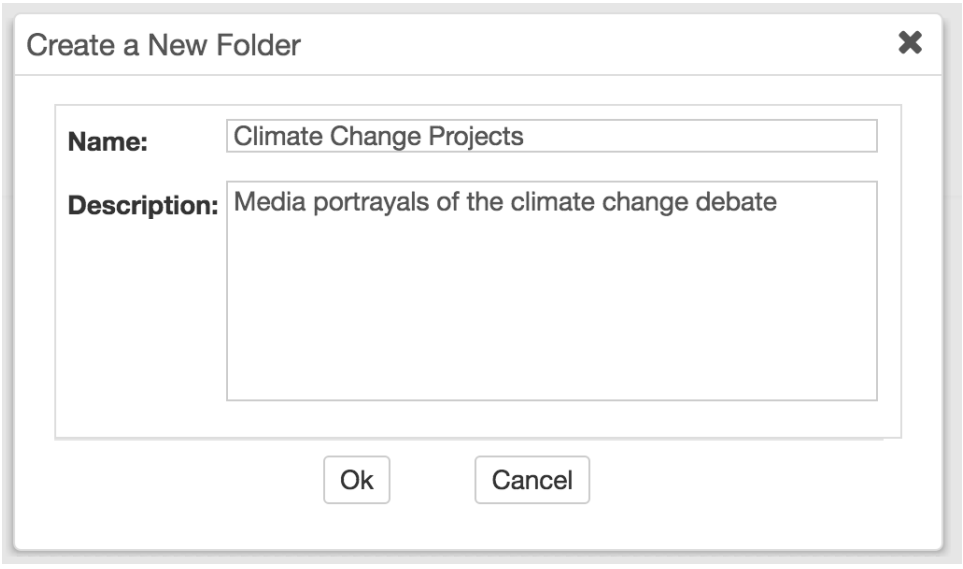


Fig. 36: Create New Folder

After you have named the folder, click 'OK'. This creates an empty project folder by this name under Leximancer / User Projects.

New Project

Right click on your new folder under Leximancer / User Projects, and select 'Create Project' to name (and optionally describe) a new project (Fig. 37).

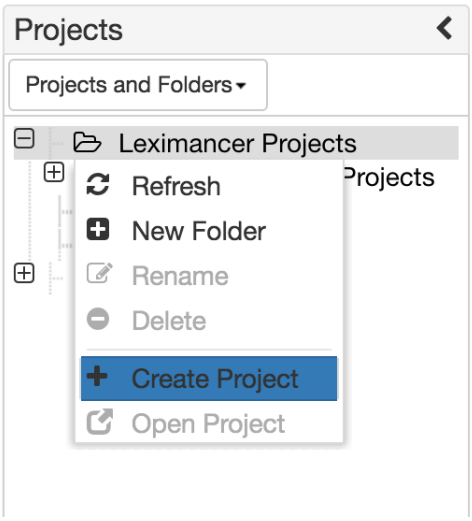


Fig. 37: Create a Leximancer Project

It helps to be descriptive when naming your projects, so that later, when you have multiple projects, you don't need to open them to check their content. As this example concerns creating an Automatic Map, the project is named to reflect that (Fig. 38):

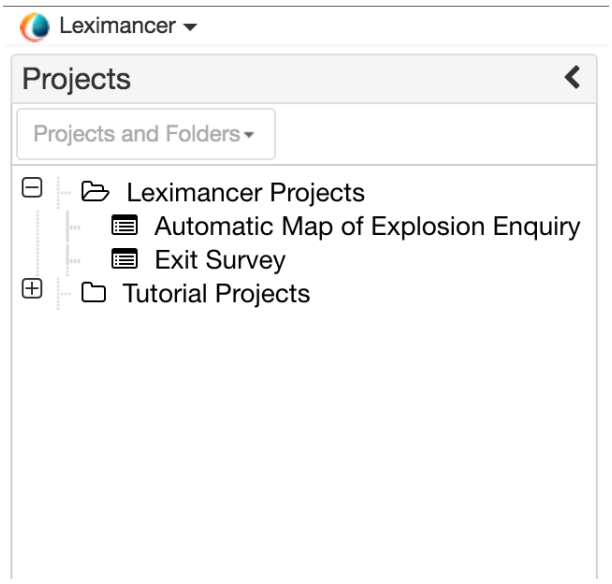
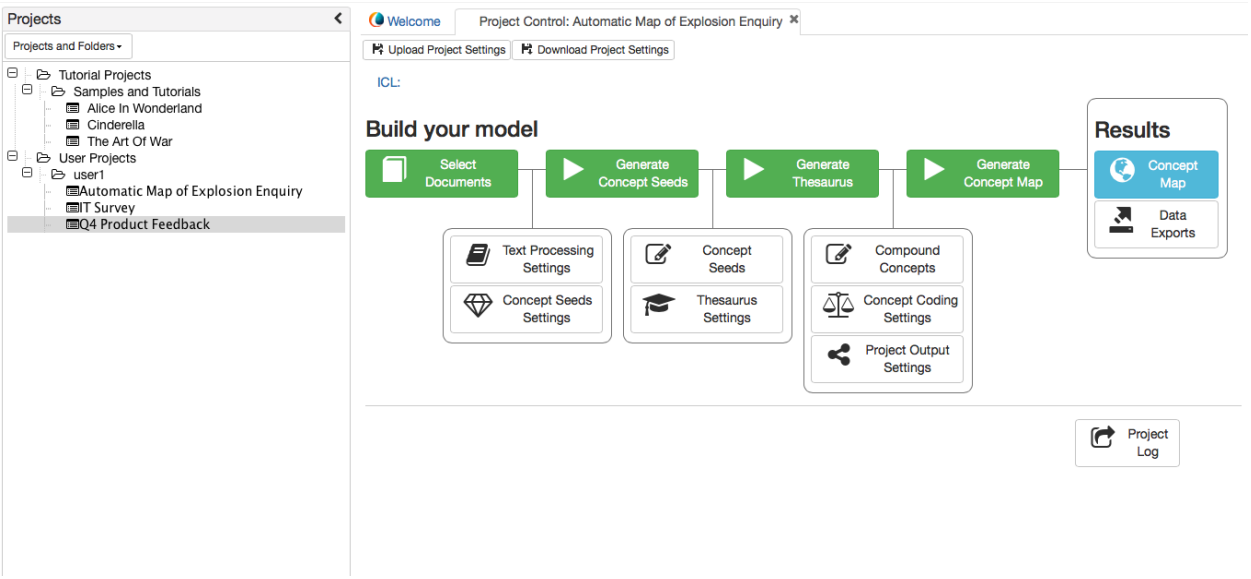


Fig. 38: Descriptive Project Name

This opens the newly-created project, and displays the main user interface in the right hand panel (`project-control`):



You can open several Leximancer projects at once. They will be displayed in separate tabs on the right hand side of the Leximancer display. When you've opened one or more projects, you can also collapse the Project Selection interface using the arrows in the top corners.

3.1.6 Leximancer Project Control

Having created a project, you are now ready to use the Main Leximancer User Interface.

The control panel is focused on defining the processing model for your project to generate the concept map and other useful outputs.

The panel is organised to show the Leximancer processing phases to generate a conceptual map.

The top of the panel shows the processing phase buttons:

- *Green:* The processing phase is complete
- *Blue:* The processing phase still needs to be run
- *Red:* A problem prevented this stage from completing successfully, clicking on a red phase will give further information about the problem.

A project may be run up to any processing stage by clicking on it. Leximancer will also run any necessary preceeding stages (Fig. 39).

The top-right portion of the project control panel gives access to the project results. The buttons in this section are disabled until the project phases have been all run and are *green* (Fig. 40).

The middle of the panel gives access to advanced settings that affect how each of the processing stages is run. In the illustration below The *Concept Seeds* (Editor) and *Thesaurus Settings* buttons

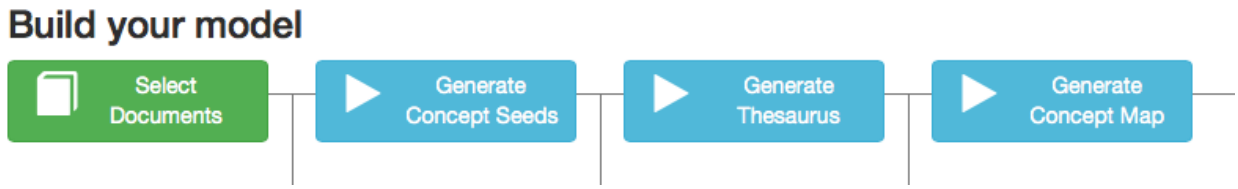


Fig. 39: Project Stages

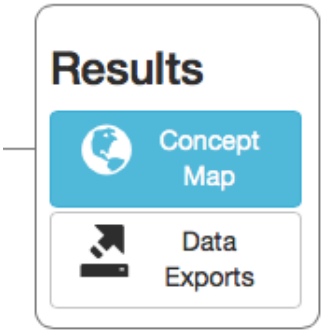


Fig. 40: Project Results

allow configuring parameters that affect the stage following these settings, in this case the *Generate Thesaurus* stage (Fig. 41). These settings will be discussed in detail later in the manual.

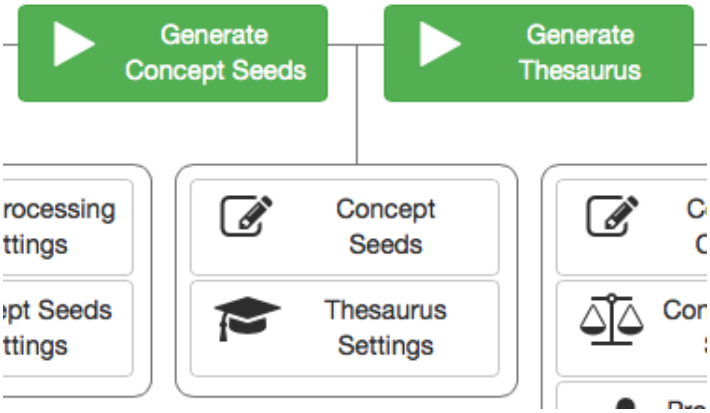


Fig. 41: Advanced Project Settings

The bottom of the panel has a status column that keeps track of the completion that updates as the project runs through phases. The *Project Log* button will open a more detailed text log of the processing steps for a project (Fig. 42).

For an Automatic Map, we only need to complete the *Select Documents*, and then click *Generate Concept Map*. Both of these buttons initally appear *blue*.

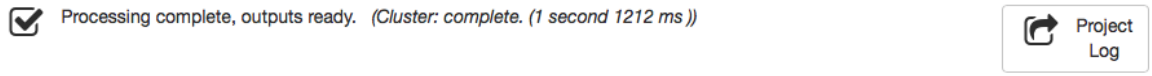


Fig. 42: Project Status and Log

When you click *Select Documents*,

If you are using *Leximancer Desktop*:

- You will see your local computer drives listed in the Available Documents panel on the left.
- Expand the drive directories to see the files and folders within them.
- Drag and drop the desired files and folders into the Document Set area on the right to link them to the project, then click OK (Fig. 43).

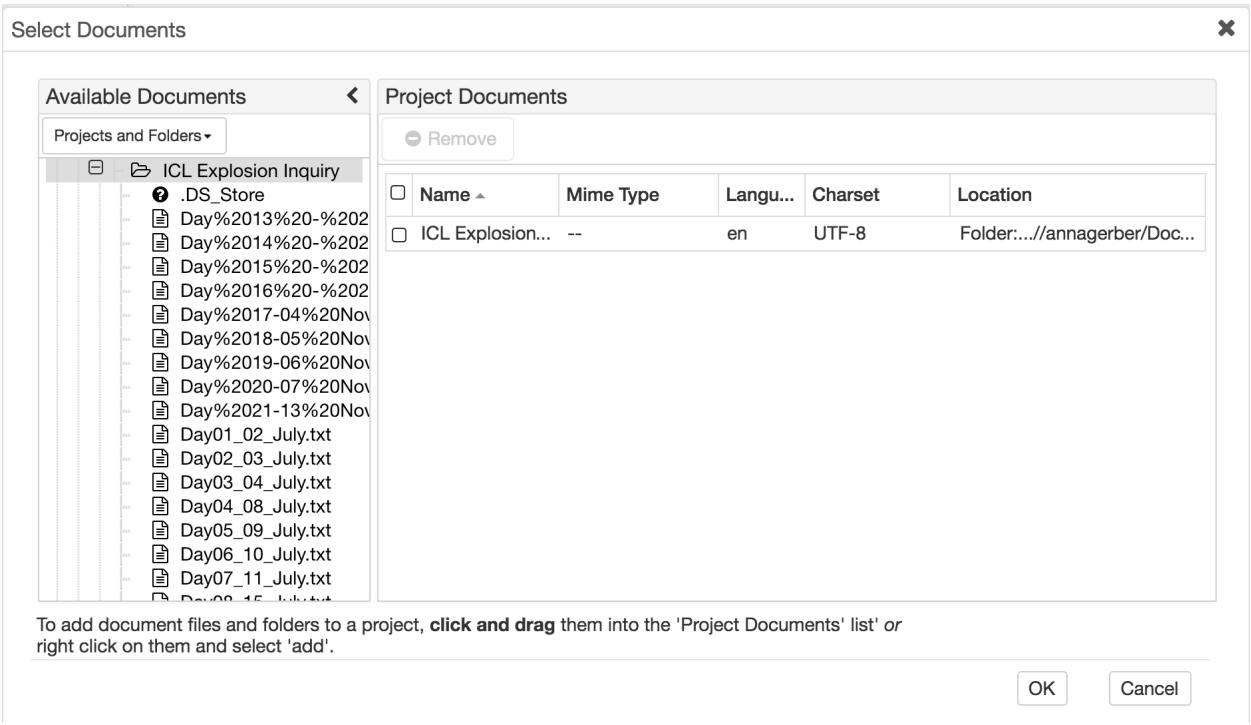


Fig. 43: Project Document Selection

If you are using *LexiPortal*:

Text documents must first be uploaded to the LexiPortal before any analysis:

- Right click on your user data folder, in the Available Documents panel on the left, and select *Upload Documents*. Your user data folder has the same name as your Leximancer user account (Fig. 44). Your user data folder is **private** and only accessible to you.
- Browse to locate the text documents on your local machine. If you wish to upload many documents at once, place them in a zip archive outside of Leximancer and select the zip file

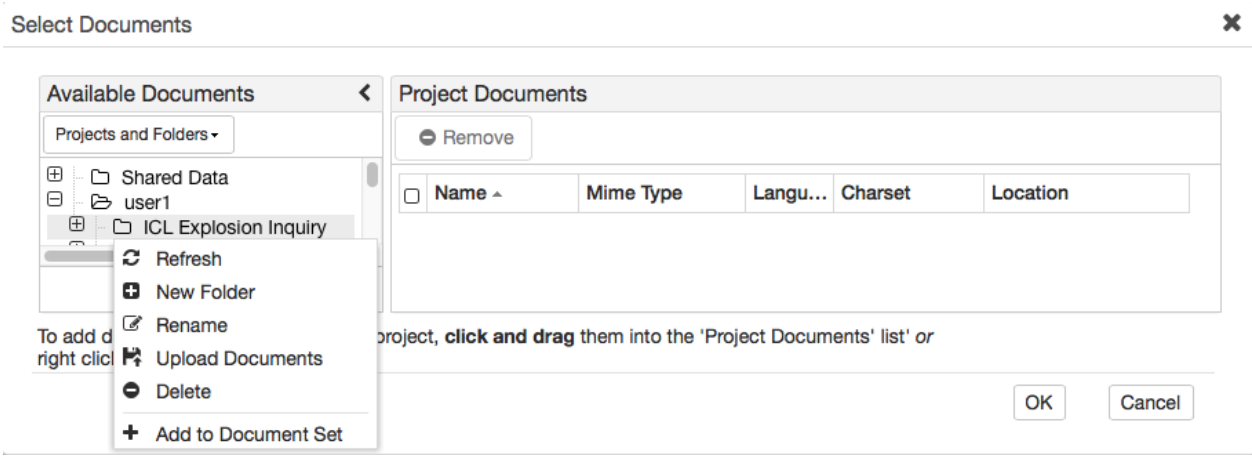


Fig. 44: LexiPortal Upload Documents Menu

for upload. The files and folders will be extracted from the zipped archive automatically on upload to Leximancer.

In the Text Document Upload window (Fig. 45), click in the blue box to open a file selection dialog. This lets you select multiple files from your local computer for upload. Alternatively, you can drag multiple files at once from your desktop onto the blue box. Once the file selection or drag is complete, the files will be uploaded immediately, so make sure you only select the files you want to upload. There is a limit on the total size of an upload. :

While your files are being uploaded, a progress bar show the percentage completed.

Once the upload is complete, the documents will appear in the left-hand Available Documents panel.

In this example, we have uploaded a single zipped folder containing a transcript for each day of a hearing after an explosion at a plastics factory.

Once the data is uploaded, you can expand the parent folder to see the files and folders within (Fig. 46):

You could choose to analyse just one of the documents within a folder (e.g., a single day's transcript) by dragging and dropping an individual file into the Document Set area on the right. Alternatively you can drag the whole parent folder into the Document Set area on the right to analyse all its contents.

Selecting the parent folder instructs Leximancer to analyse all the files and folders within. This allows you to analyse multiple files and folders at once, and facilitates comparisons between groups of documents using the software's automatic tagging options. See the Folder Tagging feature for more information.

Once you have some files and/or folders listed in the *Project Documents* area:

You have the option at this stage to specify the type of document(s) you wish to analyse (`select-doc-type`).

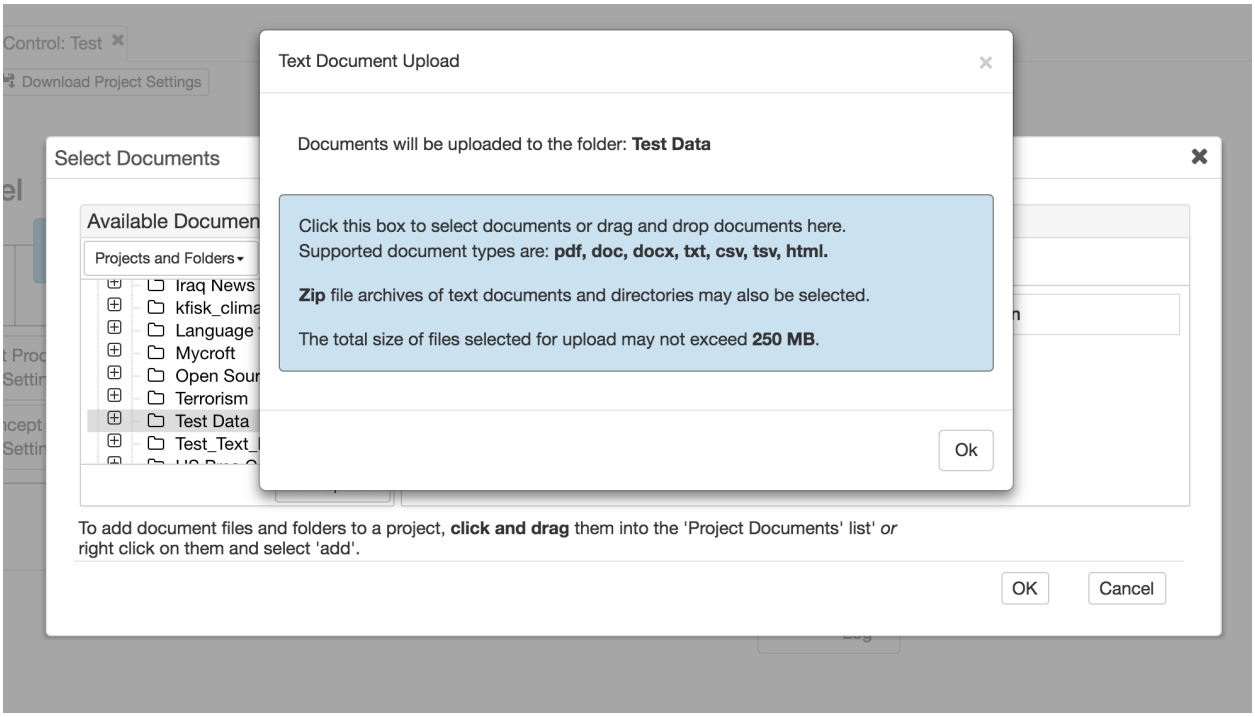


Fig. 45: LexiPortal Upload Documents Panel

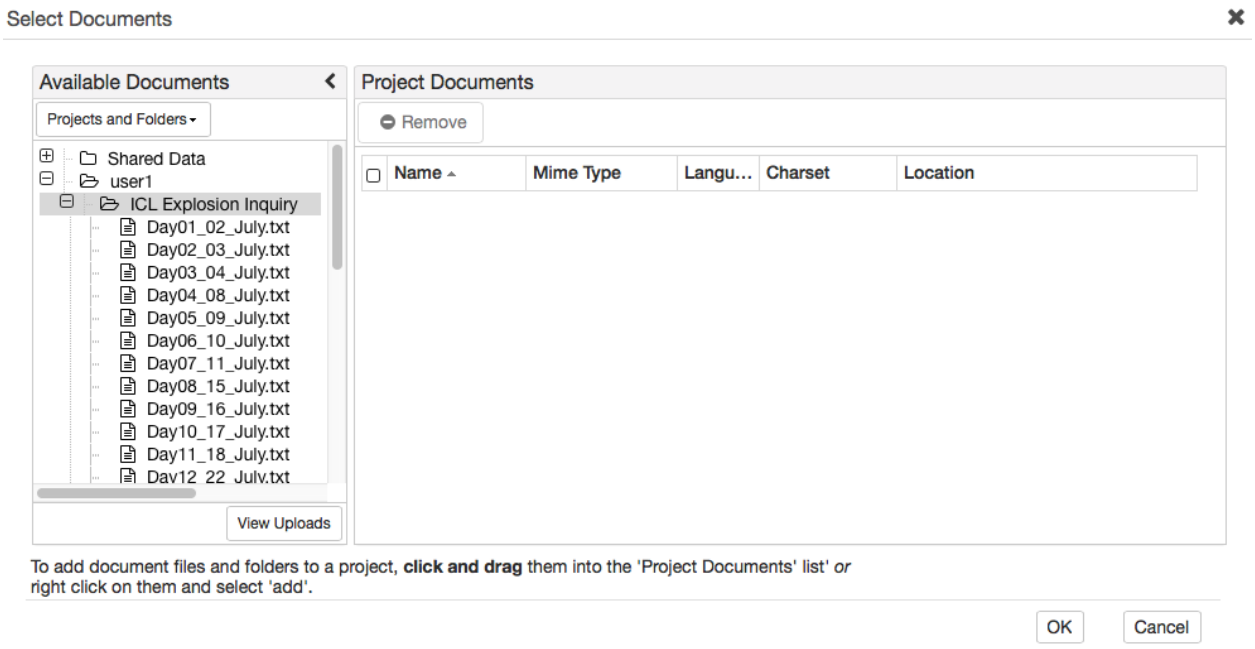
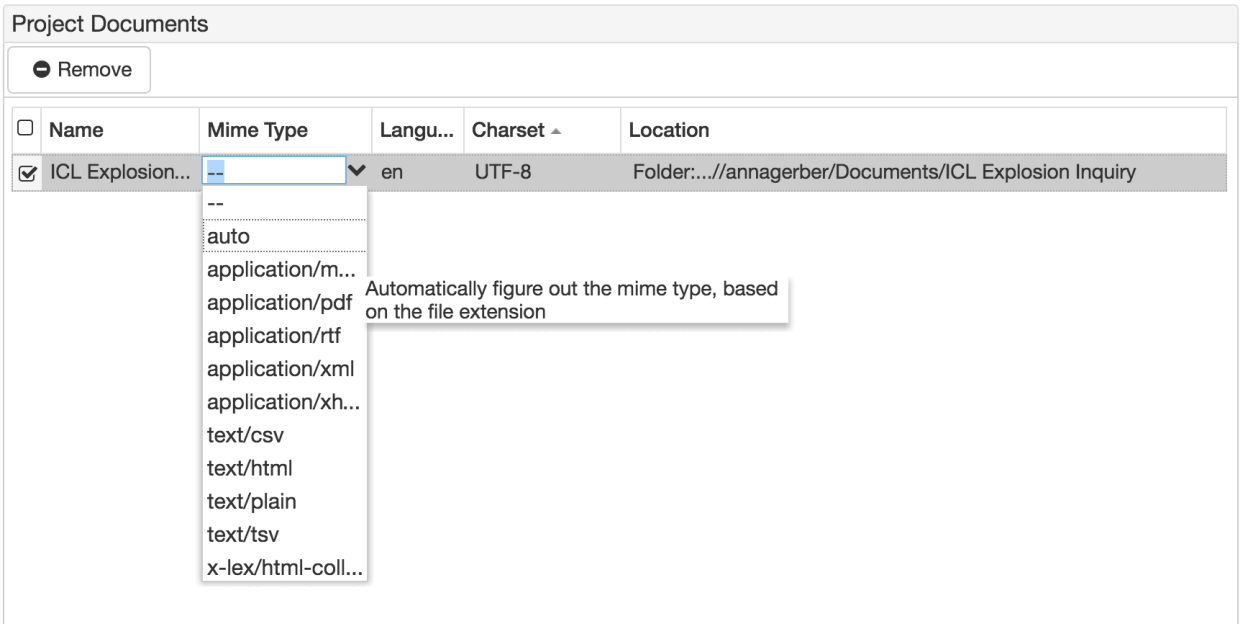
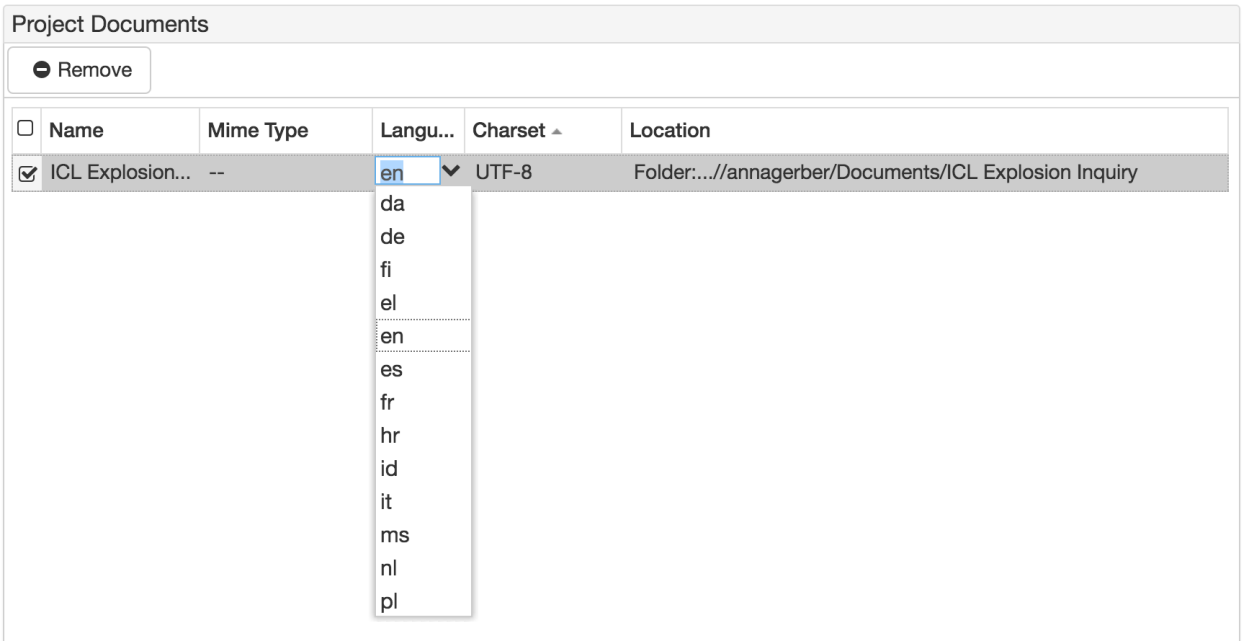


Fig. 46: Uploaded Documents Available for Selection



You may also specify the language of your documents (`select-lang`):

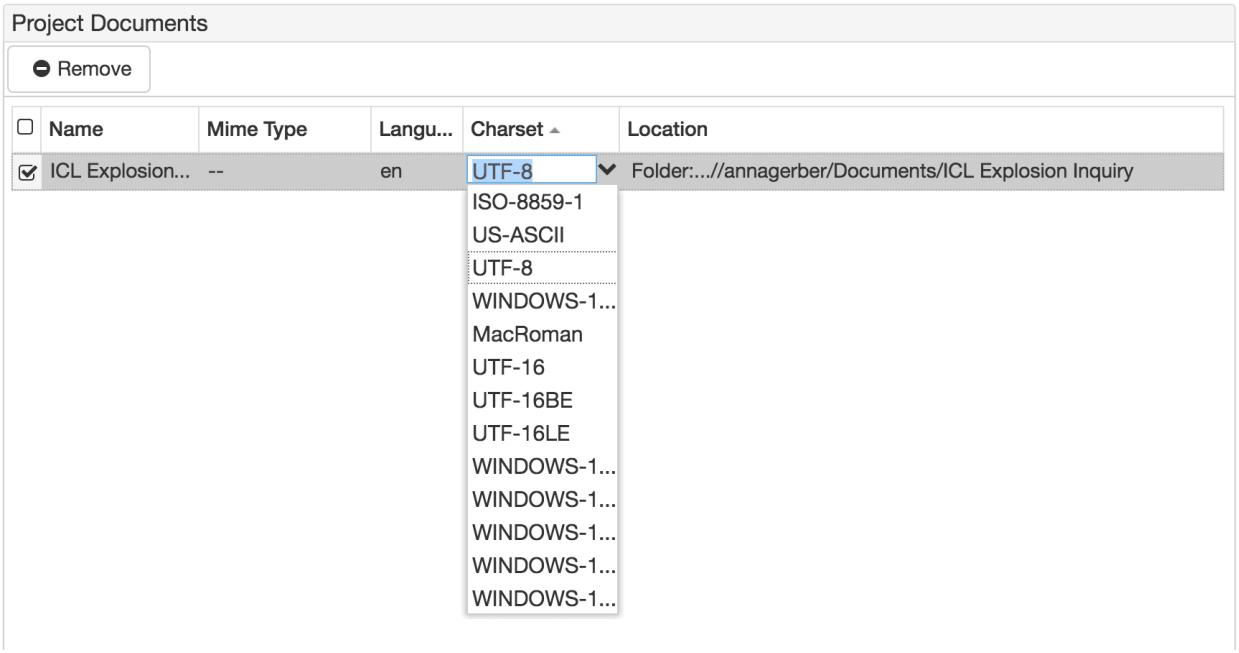


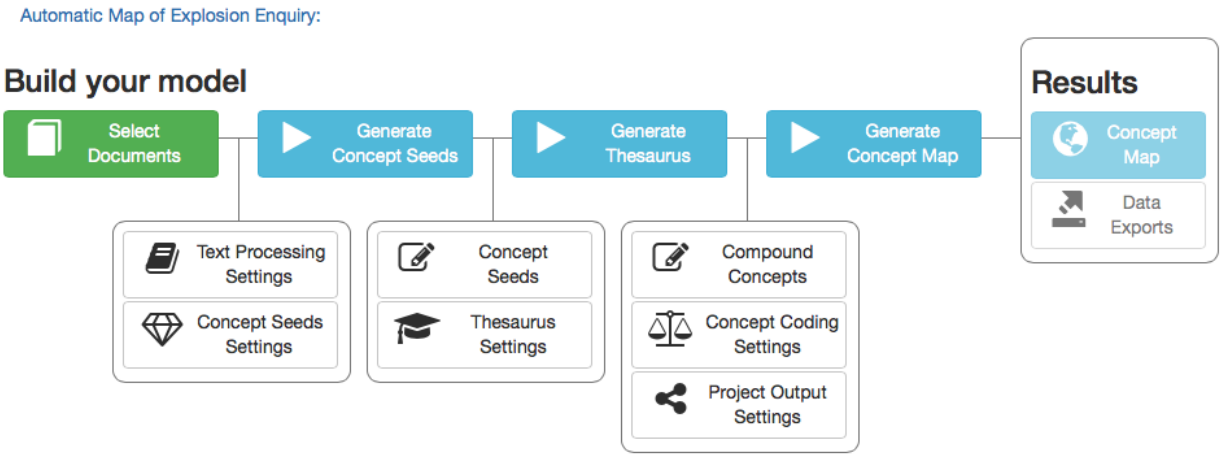
And the type of Character encoding used (`select-char-encoding`):

Finally, click OK to link the data to the current project.

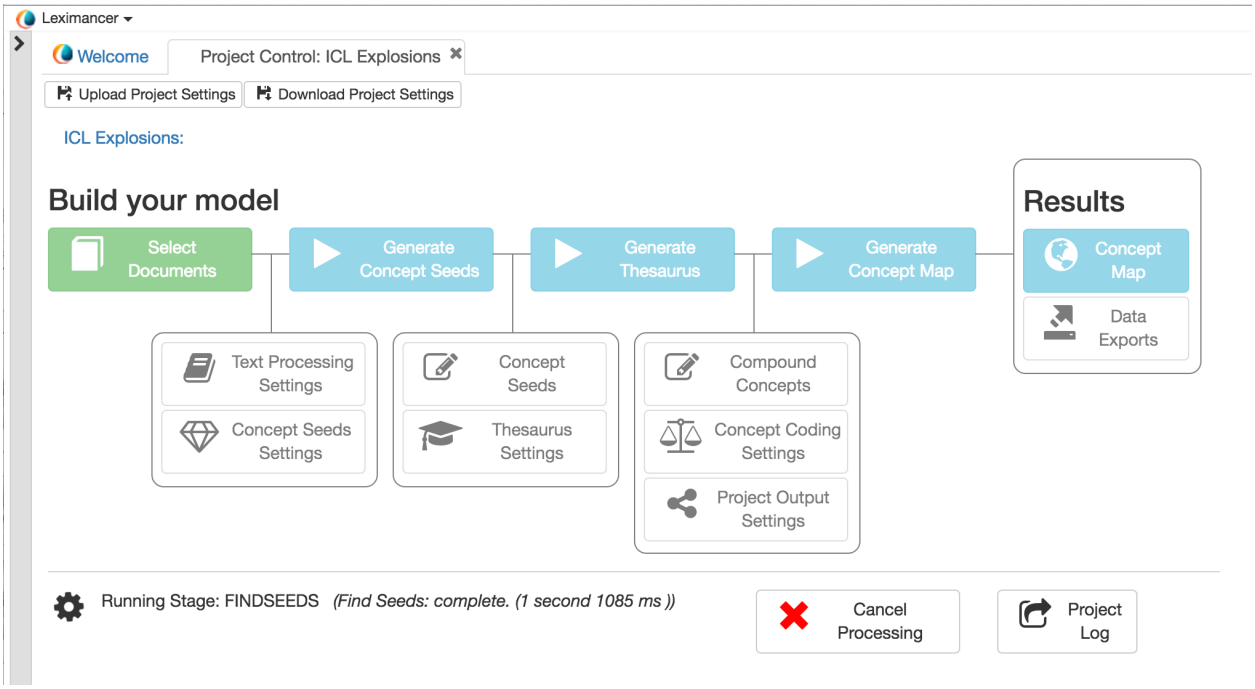
This returns you to the project control panel, where the status of the *Select Documents* stage will be green, as this step has now been completed (`select-docs-ready`):

To perform an automatic or exploratory analysis, click the *Generate Concept Map* in the upper right of the Project Control Panel, left of the Results buttons. This will run the project using default settings.

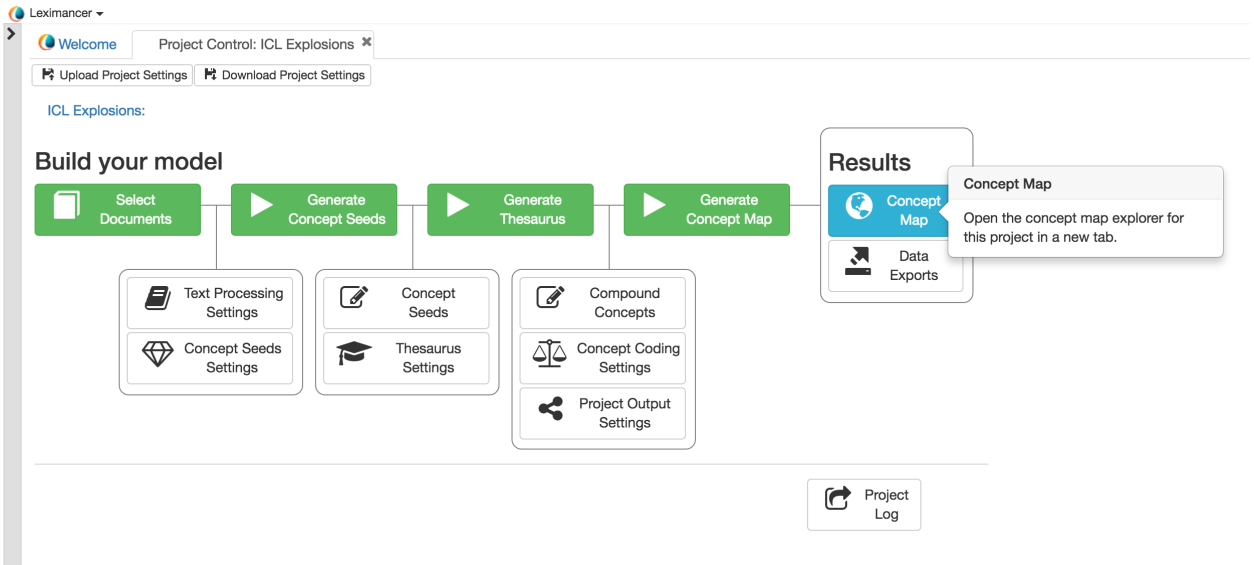

Project Documents
Remove
Name
Mime Type
Langu...
Charset
Location
ICL Explosion...
--
en
UTF-8
Folder:...//annagerber/Documents/ICL Explosion Inquiry
ISO-8859-1
US-ASCII
UTF-8
WINDOWS-1...
MacRoman
UTF-16
UTF-16BE
UTF-16LE
WINDOWS-1...
WINDOWS-1...
WINDOWS-1...
WINDOWS-1...
WINDOWS-1...


Automatic Map of Explosion Enquiry:
Build your model
Select Documents
Generate Concept Seeds
Generate Thesaurus
Generate Concept Map
Results
Concept Map
Data Exports
Text Processing Settings
Concept Seeds Settings
Concept Seeds
Thesaurus Settings
Compound Concepts
Concept Coding Settings
Project Output Settings

While the stages are running, the project control panel buttons will be disabled. As the stages complete, they will turn green (`project-running`). More detailed progress information is also visible at the bottom of the panel. The bottom of the panel also has a *Cancel Processing* button to use while the project is running. If you press cancel, the project will halt, *after* it has finished the current processing stage.



When all phases of processing are complete, click the *Concept Map* button in the Results section (`done-open-map`).



The Project Control Panel also includes:

- A Close tab button (x) that allows you to close the current project (current settings are saved) to exit or load other projects;
- *Import Project Settings* and *Export Project Settings*, in which you may save the settings of your project, or apply settings from an earlier project. (These import and export the overall project configuration but do not include edits that have been made to concepts or source document locations.)

CHAPTER

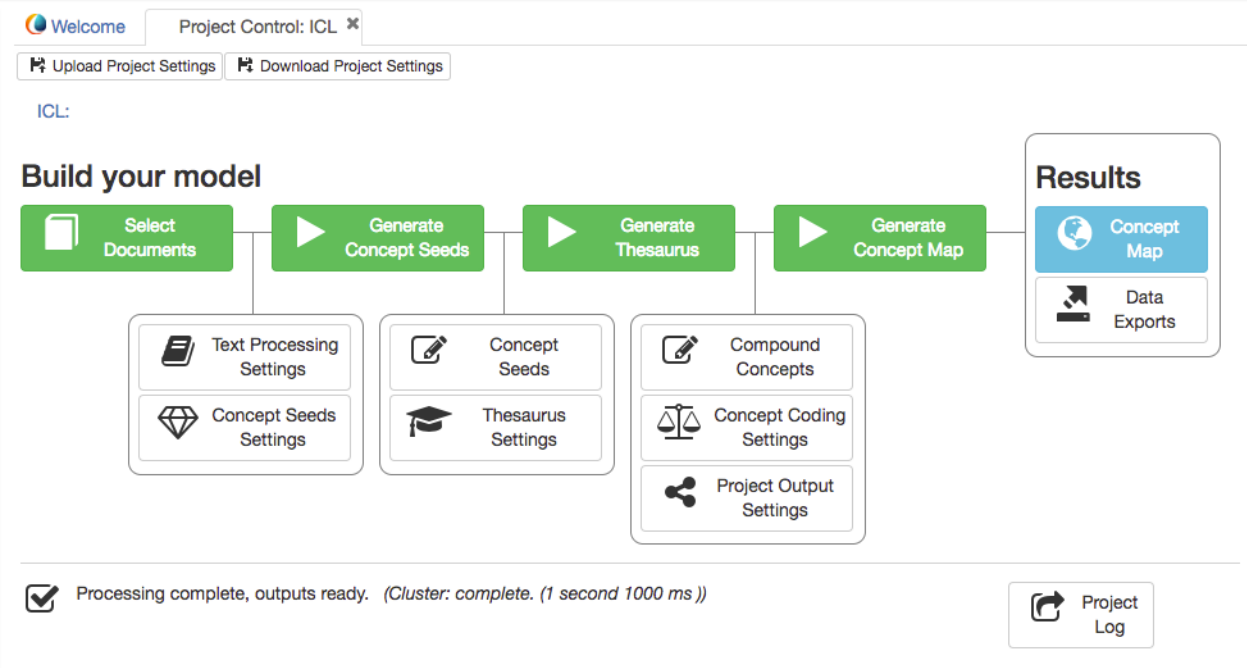
FOUR

CREATING A MANUALLY ADJUSTED MAP

Note: This chapter of the manual has not yet been fully updated for V4.5. The techniques presented still apply but the layout of the user interface has changed.

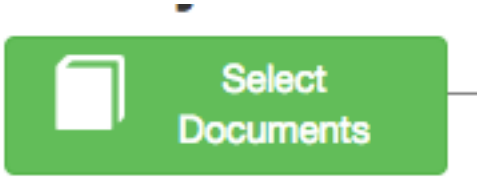
The manual will now discuss the ways in which the user may alter the settings of their project to create a concept map. This section will follow the sequence of the stages as represented in the control panel (below).

However, the 'Select Documents' stage will not be covered in this chapter. Please refer to the start of the previous chapter, Creating an Automatic/Exploratory Map, for information on how to start Leximancer and load data.



Stages of Processing

1. Document Selection:

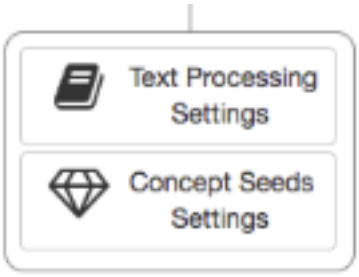


Please refer to the previous chapter (page 38) for information on how to load data in Leximancer.

2. Generate Concept Seeds:

You can run this stage of processing using default settings by clicking the Generate Concept Seeds button.

Options that affect the initial text processing and concept seed discovery are available from the two settings buttons that precede the "Generate Concept Seeds" stage button.



4.1 2a. Text Processing

This is the first phase of processing that is run from Leximancer's main menu. This phase converts the raw documents into a useful format for processing. Preprocessing involves the following steps:

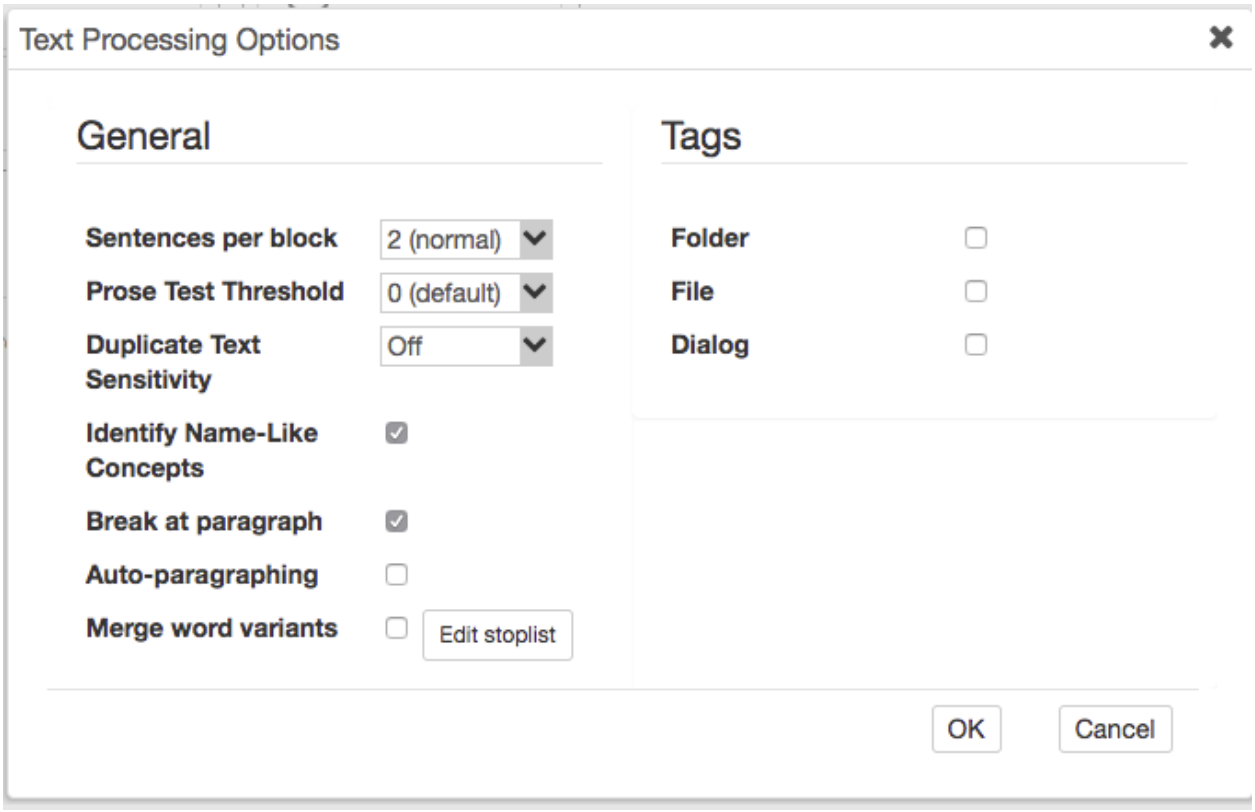
- Splitting the information into sentences, paragraphs and documents. These boundaries are important as they generally mark transitions in meaning. The conceptual map of the documents extracted by Leximancer reflects the co-occurrence of distinct concepts. To prevent concepts from being perceived to be related across changes in context (such as across different documents), the co-occurrence is only measured within (and NOT across) blocks typically containing 2 sentences.
- Removal of non-lexical and weak semantic information. Within each sentence, the punctuation is removed along with a collection of frequently occurring words (called the stop-list) that hold weak semantic information (such as the words 'and' and 'of'). Furthermore, for

documents extracted from Internet email and news groups, the headers are cleaned up and the non-text attachments are removed.

- Identifying proper names, including multi-word names. Often in documents the proper names (such as people, places or company names) depict important entities that should be mapped. For this reason, proper names are extracted as potential concepts. In Leximancer, words are classified as proper names if they start with a capital letter. As every word that starts a sentence falls into this definition, only start-of-sentence words that are not in the predefined stop-list are considered as names.
- Optional prose test of each sentence. To remove non-textual material from the text, such as menus and forms in web pages, sentences that are unlikely to be part of the specified language are removed. This is achieved heuristically by removing sentences that contain less than 1 (or 2) of the stop-list words. If processing spoken language, this setting should be turned off.

Practical: Configuring Pre-processing in Leximancer

Clicking on the Text Processing Settings node reveals the following interface:



Manual Editing Options:

Sentences per Block (1-100): This option allows you to specify the number of sentences comprising each context block (or text segment). A context block is said to contain a concept if the words therein provide sufficient cumulative evidence of its presence.

Commentary: The best value for this parameter depends on the nature of the data. It should almost always be two or three, though in some instances one sentence is sufficient (eg: abstract, press release, or verse)

Prose Test Threshold (0-5): The Prose Test feature examines raw text sentences to decide whether they are valid prose from the configured languages. This is achieved by counting the number of stop-words that appear within each sentence. If this number is high, it is likely to be a sentence from a configured language. This option allows you to specify the number of stop-words that are required for the sentence to be further processed.

Commentary: This feature is good for reports or other prose documents where you don't want to process tables of numbers or lists of words. It is almost essential for web pages or e-mail messages which often contain menus or signatures. This sort of repeated data can potentially contaminate your automatic seeds and machine learning. If the data is not prose from a supported language, or if it is composed of transcribed speech or other colloquial matter which does not obey the rules of prose style, then this feature should be weakened or disabled. You should also disable this if you need to analyse absolutely every bit of text in the data

Duplicate Text Sensitivity (Off|Auto|1-8): This setting suppresses the processing of duplicated text. This option is especially useful when analysing email data or blogs and reviews, where cross-posting and quoting is common.

Identify Name-Like Concept (Yes|No): This setting is important if you would like words that seem to be names (i.e., non-stop words, starting with a capital letter) to be stored as potential concepts.

This setting requires text data which uses upper and lower case, where upper case designates proper names. This doesn't work in many languages, or in some text data where case is missing, but it is very useful much of the time for tagging proper names. Note that it binds compound names into one token.

Break at Paragraph (Yes|No): This setting is to prevent context blocks from crossing paragraph boundaries. Only if the majority of paragraphs in the text are shorter than 2 sentences should you consider ignoring paragraphs.

Auto-Paragraphing (Yes|No): This setting identifies whitespace, particularly line-breaks and paragraph-breaks, if the document is consistent in its spacing, to identify new paragraphs. If there is a document with no reliable spacing for paragraph boundaries, this setting should be turned off.

Merge Word Variants (Yes|No): This option employs a stemming algorithm to identify the head-word for initial thesaurus items. For instance if stemming is turned, the initial thesaurus terms for the stem look in the Concept Seed Editor may include looked and looking. If you don't like the results, you can ungroup the thesaurus items in the seed editor. Lemmatisation is off by default.

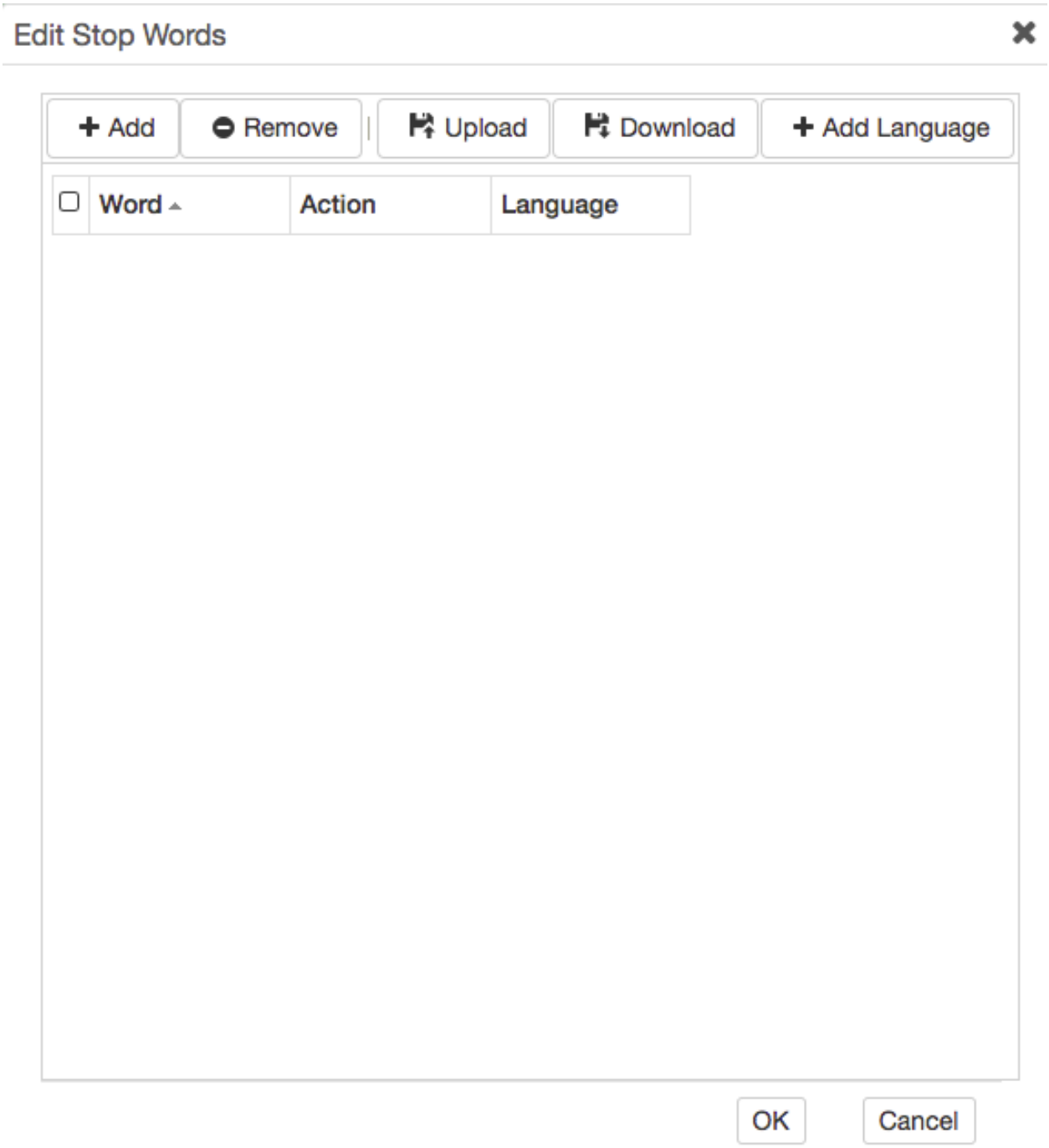
4.1.1 Stopword Removal

During Preprocessing, words with low semantic-content (meaning) are removed from the text data using a predefined Stopword List. An example of an English Stopword is 'and'. This word occurs frequent in English text, but would not constitute a useful or clearly-defined concept. If you are using an unsupported language, you can update this list (e.g. by translating the contained words into your language). Note that stop words are removed from the text to analysed, and cannot be selected as manual seed words.

Commentary: Stopwords are frequent words in a language that are rather arbitrarily designated as having little semantic meaning. Leaving stop words in the data has an obvious effect on the automatic seed selection. If you leave the stop words in, some will be chosen as automatic concepts. This can be annoying, depending on what you are looking for. The presence of stop words also impacts on the machine learning of thesaurus concepts, since almost everything can be correlated with words such as 'is' or 'and'.

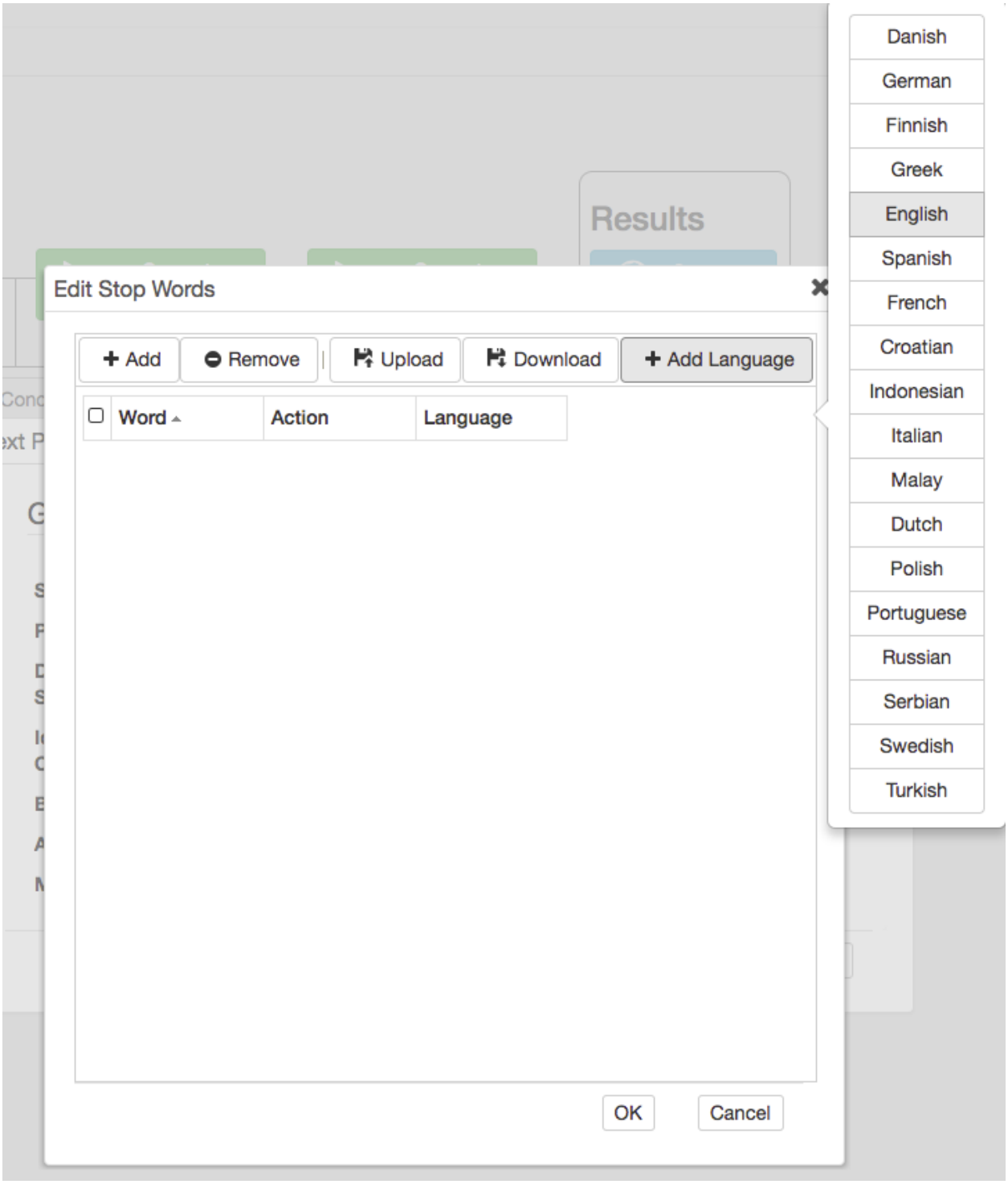
You can edit the words that are considered stop-words by clicking the "Edit stoplist" button:

The following interface will appear:

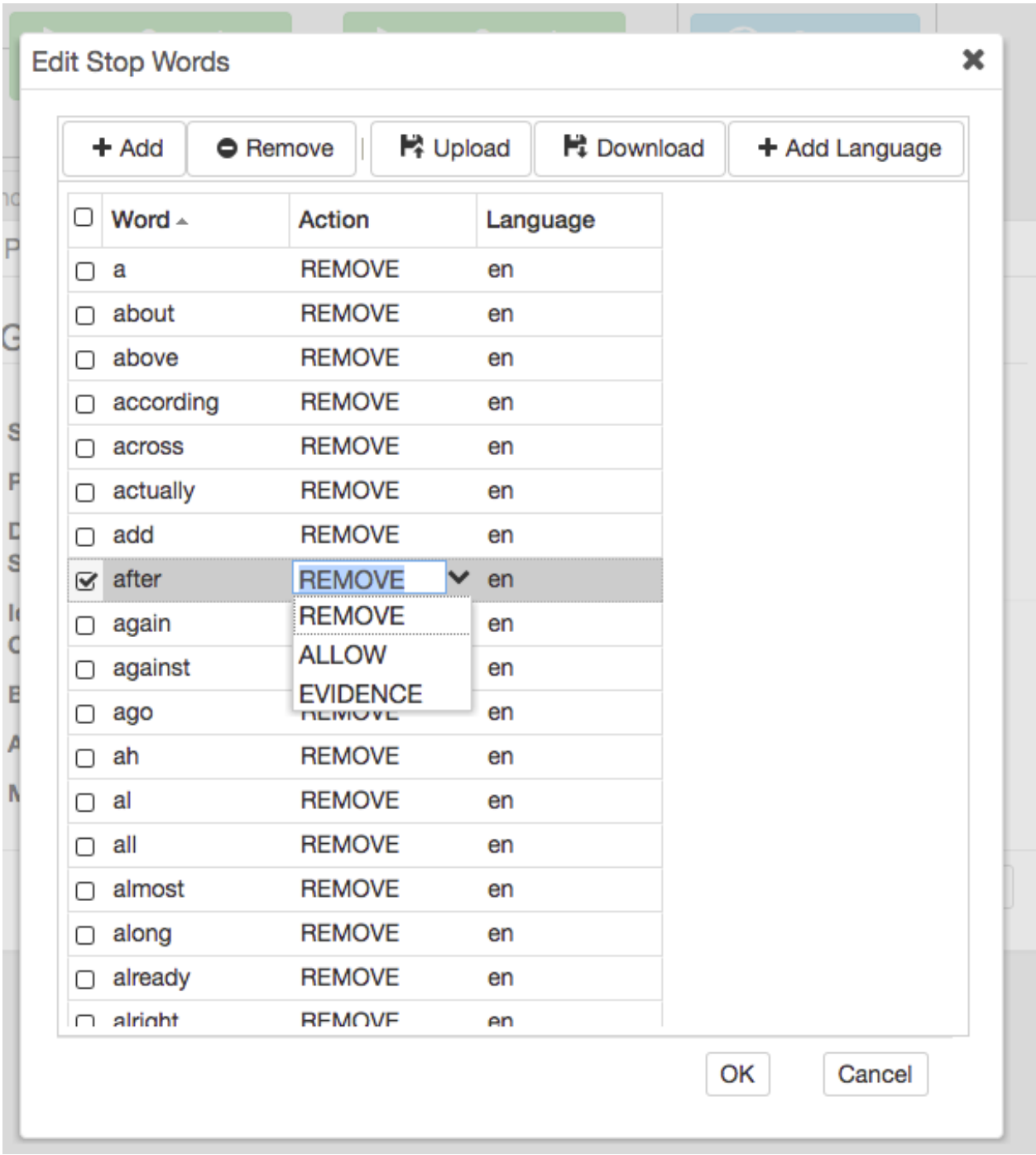


By default, the stop-list display will be blank. In this case, Leximancer will apply the *default* stop-list for the language of a document being processed.

If you wish to add or remove words from the default stop-list, click on "Add Language", and select the language of choice.



The stop-list words for that language will then appear in the stop-list editor panel.



You can browse through this alphabetical stoplist and see if you find any words that you would rather include in the analysis. If so, click on the *Remove* option next to the word, and change the status of the word to *Allow* using the dropdown menu. Set the word's status to *Evidence* if you want to allow the word to form part of a concept thesaurus, but not to be considered as a possible concept candidate itself.

You can edit words in the list by clicking on them and typing in the text box that's revealed. You can also Add words to the stoplist, or Remove them from the list. The Download button lets your

download the current .xml stoplist file, and the Upload option lets you upload another .xml stoplist file.

Note: Stoplist Downloads: If the stoplist has been edited in any way, including adding a default language stop-list, it must first be *saved* by clicking *Ok*, before it can be downloaded.

The Language column lets you know the language from which particular stopwords come in case you are analysing documents in different languages. You can change this setting so that stopwords are only removed from text in the appropriate language. The language abbreviations in the list are ISO 639 country codes.

You may also add the stoplist of another language. This includes high-frequency, non-semantic words from a wide selection of languages. Note that the relevant stoplist will automatically load after selection of the language in the earlier document selection window.

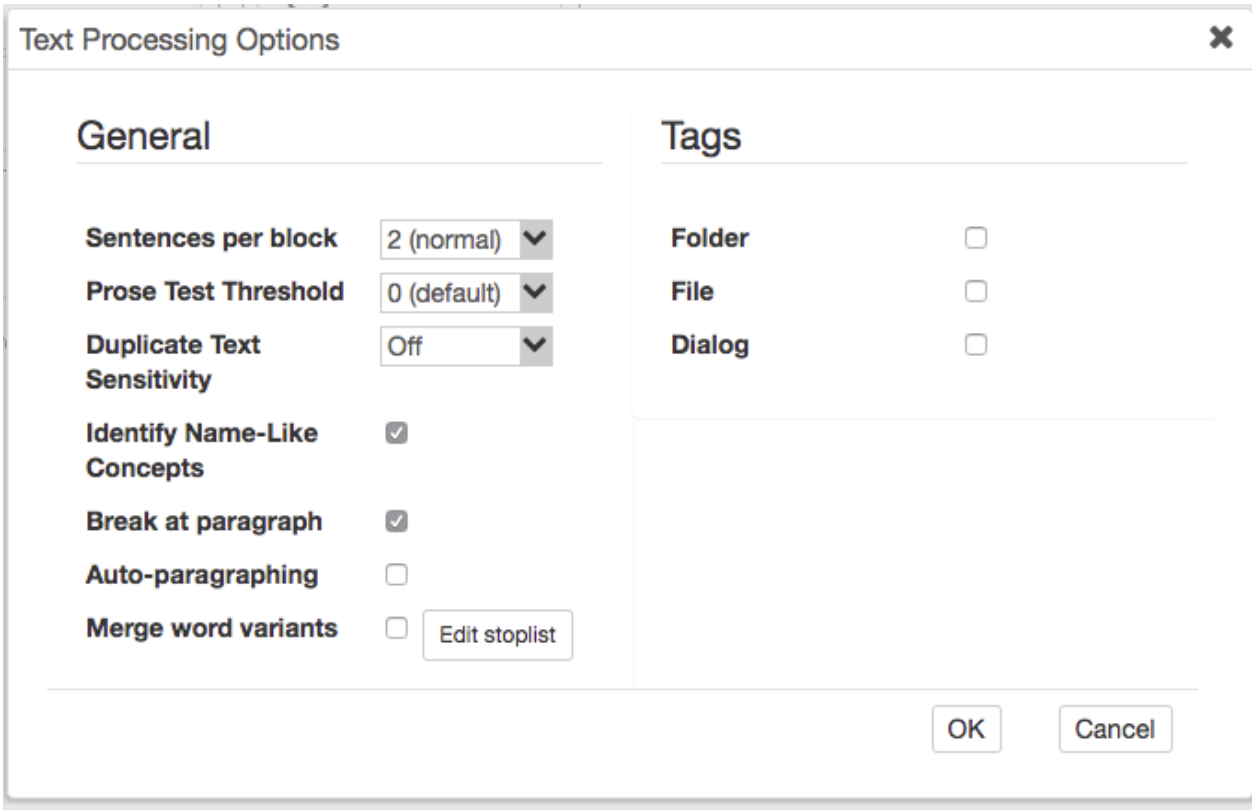
Stop-lists are language specific, and only one stop-list language is applied to a document. If needed, words from any language can be added to a specific language stoplist, if the language settings for the stopword is set correctly.

Press 'OK' to save the stoplist edits and return to 'Pre-process' dialogue, then click 'OK' to return to the Control Panel.

Note: If the stop-list displayed for a language in the stoplist editor, Leximancer will not apply the default stop-list.

Tagging Options

Tags are important for comparing different documents based on their conceptual content, for example, different speakers in transcript documents, or for a comparison between different text sources. At this stage in processing, you can instruct Leximancer to pay attention to certain tag categories so that you may analyse them later.



The Folder Tags (Yes|No) and File Tags (Yes|No) options can cause each part of the folder path to a file, and optionally the filename itself, to be inserted as a tag on each sentence in the file. In our example project, we can use the File Tagging facility to compare the content of the transcripts on different days of the hearing. Source document tags can then be included as concepts on the map.

Commentary: This is a powerful feature that lets you code all the sentences in each document with categorical tags just by placing the files in folders, possibly within other folders etc. The tags can then be included on the map, among the topical concepts. This is useful for performing document clustering in a semantic context, or for making a discriminant analysis between two categories. For example, if you had a set of political speeches from a debate on some issue, you could give each speech file the name of the politician, and place all the speech files from each political party in a folder named as the name of the party. Then, if you had several sets of these speeches from different years, you could place each year's set of folders in its own folder named with the relevant year. Applying folder and filename tags will then insert the name of each politician as a tag on each sentence, and the name of the containing political party as a separate tag on each sentence, and also the name of each year. When you map this data, you will find a concept for each politician, each party, and each year in the Tags collection. You can then choose which of these dimensions to cluster together on the map, so you could view the issues by year, by party, by politician, by year and party, by politician and party, by year and politician, or all of them co-varying at once if you are adventurous. You will need to deliberately add your desired concepts to the map - they won't appear automatically. See the Mapping Concepts topic for details.

The Dialogue Tags (Yes|No) function is designed to utilise speaker labels or headers in the text data. This setting identifies speaker labels that start with upper-case, end with a colon and space,

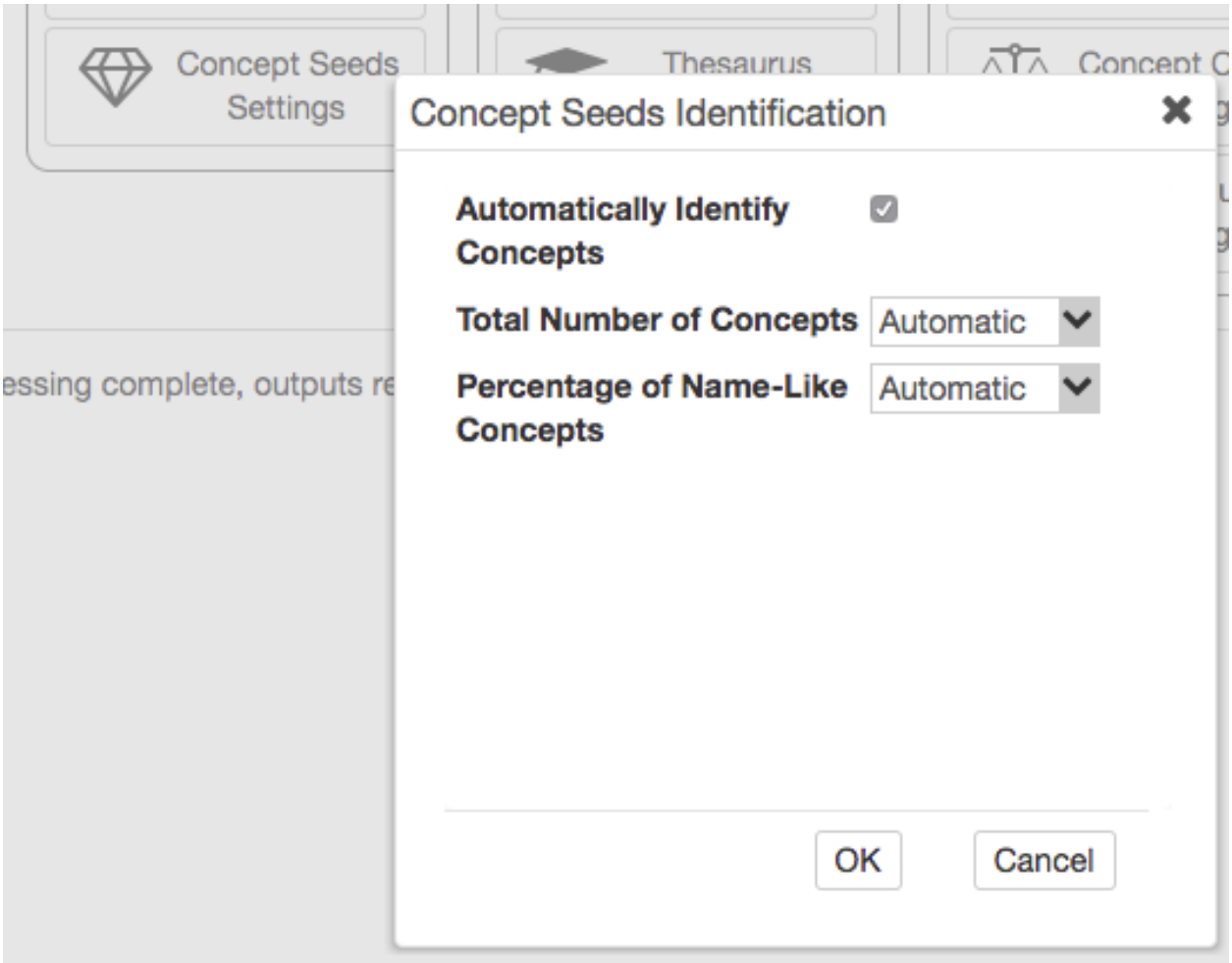
and are located at the start of a line. Each speaker label is appended to the end of every subsequent sentence until a new label is found. This can be useful for analysing focus group or interview transcripts.

4.2 2b. Concept Seeds Settings

This the phase of processing in which seed words are identified as the potential starting points of concepts. Concept seeds represent the starting point for the definition of concepts. They are single words (such as 'violence') that are the potential central keywords of distinct concepts. In this optional phase, the user may indicate whether they want Leximancer to automatically identify seed words (and the configuration of those seed words), or whether the user will manually provide seed words.

Practical: Configuring Concept Seeds Identification

The Concept Seeds Settings allow you to choose the number of concepts (if any) that you would like Leximancer to automatically extract from the text. To modify these settings, click on the Concept Seeds Settings node in the main interface, and the following dialogue will appear:



Automatically Identify Concepts (Yes|No): Turn off automatic identification of concepts if you would like only concepts that you define yourself on the map.

Total Number of Concepts (Automatic|Number): This sets the number of concepts to be extracted automatically. More diverse content requires more concepts, but less than 100 is recommended. Leaving this setting on Automatic allows Leximancer to extract the naturally emergent number of concepts from the data.

Commentary: The larger the data set, or the more conceptually diverse, the more concepts you should look for. As a rough guide, think of the number of concepts increasing logarithmically with the number of documents. However some data, such as magazine article collections, are very diverse. Note that the selected set of concepts starts at the top of a ranked list, so you should always get the most important and general concepts. You need to decide how deep you want to go into the more specific concepts. Be aware that if you force more concepts than are really found in the data, you can start getting junk concepts from among the noise.

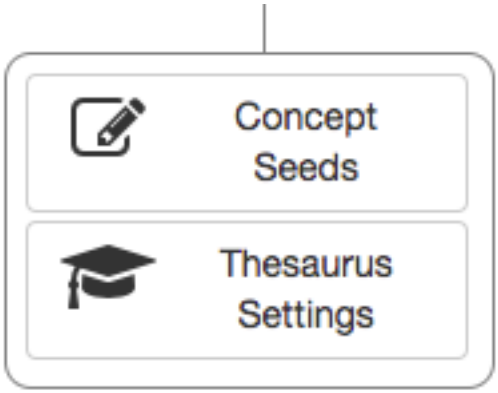
Percentage of Name-Like Concepts (Automatic|Number): This setting allows you to set what proportion of the automatically-extracted concepts should be forced to be names. Leximancer identifies names by looking for words that do not begin a sentence but start with a capital letter.

Commentary: This setting's default is Automatic, which creates a natural mixture of words and names by not forcing any names into the list. If you are not interested in names at all, you can set this to 0%. Increase this number if you are particularly interested in names.

4.3 3. Thesaurus Generation:

You can run this stage of processing using default settings by clicking the Generate Thesaurus button.

Alternatively you can configure this stage via the Concept Seeds node and the Thesaurus Settings node:



Starting with the seeds automatically extracted by Leximancer, the Concept Seeds editing phase (3a) allows users to edit, add or remove concept seeds from the list.

The following phase, Thesaurus Settings (3b), then generates the thesaurus of terms associated with each seed. As mentioned earlier, concepts are collections of correlated words that encompass a central theme. Once such lists of words have been identified for each concept, the concept map can be generated to illustrate the relationships between the concepts in the text.

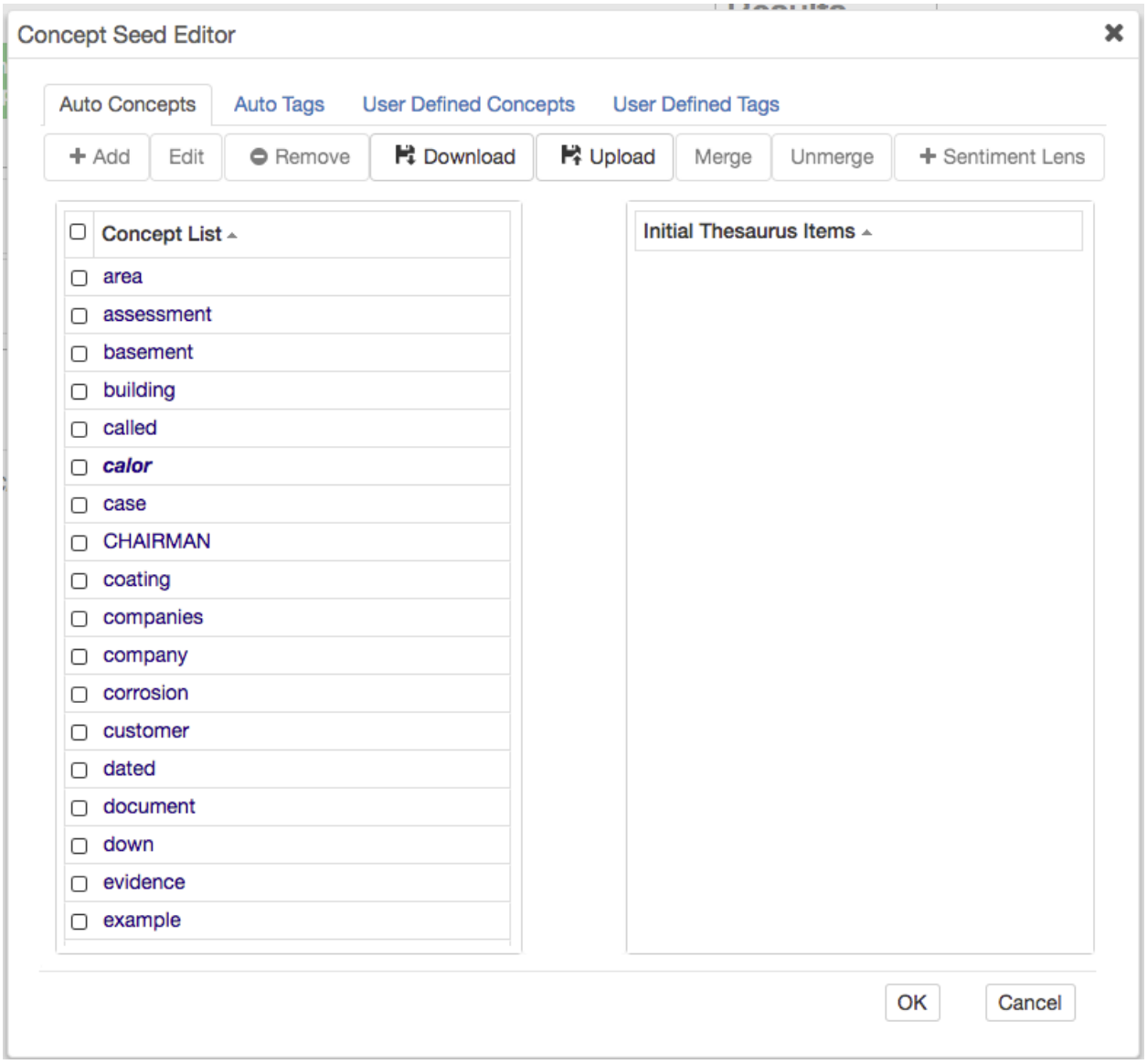
The learning of the thesaurus associated with each concept is an iterative process. Seed words are named as such because they start out as being the central terms of a concept definition - related keywords are collected during learning. During learning, seeds can also be pushed to the periphery if more important terms are discovered.

4.4 3a. Practical: Configuring Concept Editing

Clicking on Concept Seeds opens an interface that allows you to edit, add or delete concepts. This is important for a number of reasons:

- automatically extracted maps may contain concepts (such as think and thought) that are similar, or other concepts that are not of interest to you. In the Concept Seeds interface you can merge similar-looking concepts into a single concept, or delete concepts that you do not wish to see on the map
- you may wish to create your own concepts (such as violence) that you are interested in exploring, or create categories (such as dog) containing specific instances of terms found in your text (such as hound and puppy).

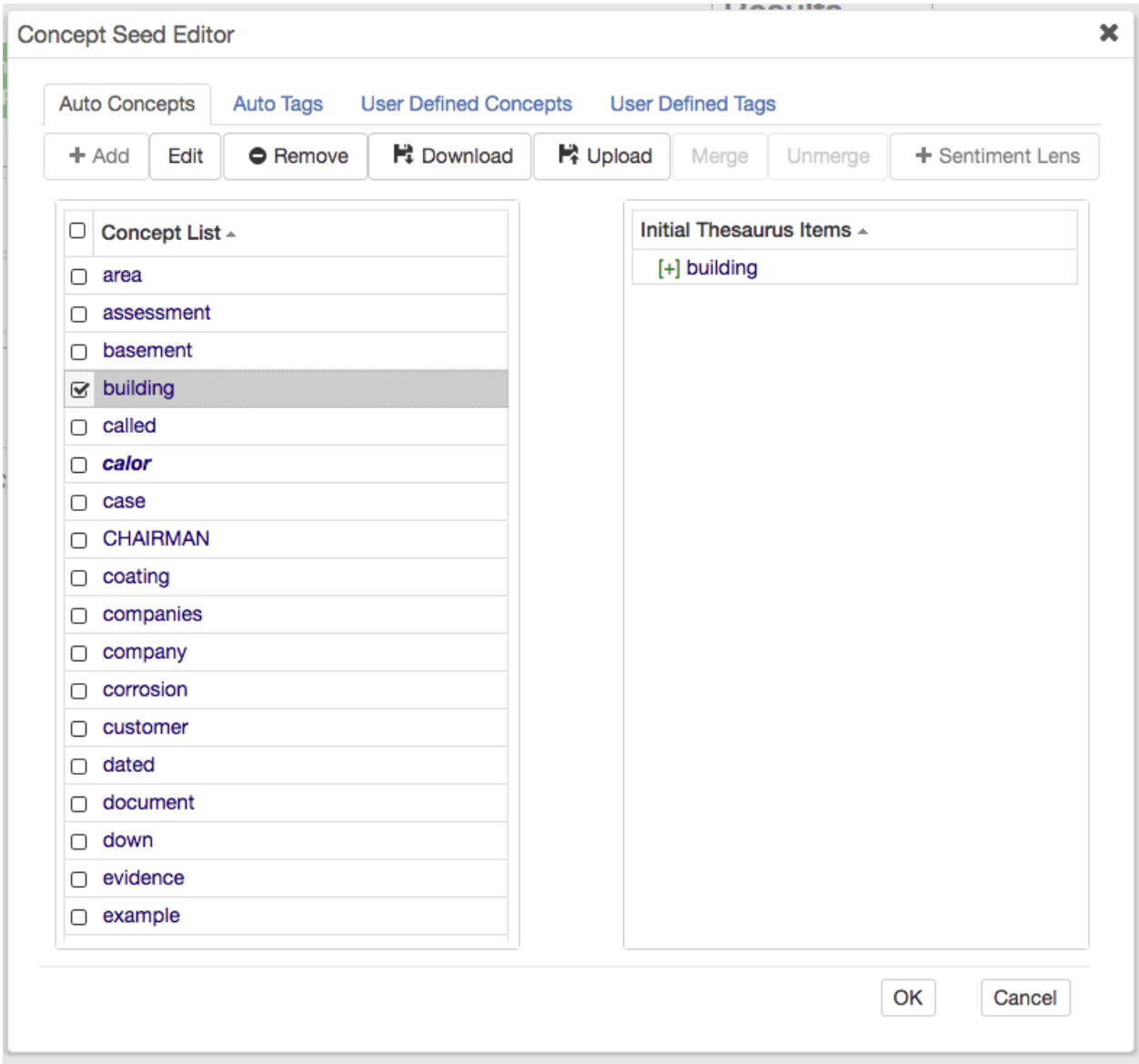
Once the Generate Concept Seeds stage has been run, you can check and modify the discovered concepts by opening the Concept Seeds node. The following interface will appear:



Here you can edit the concept seeds extracted automatically by Leximancer in the Auto Concepts tab, and create your own manual concept seeds in the User Defined Concepts tab.

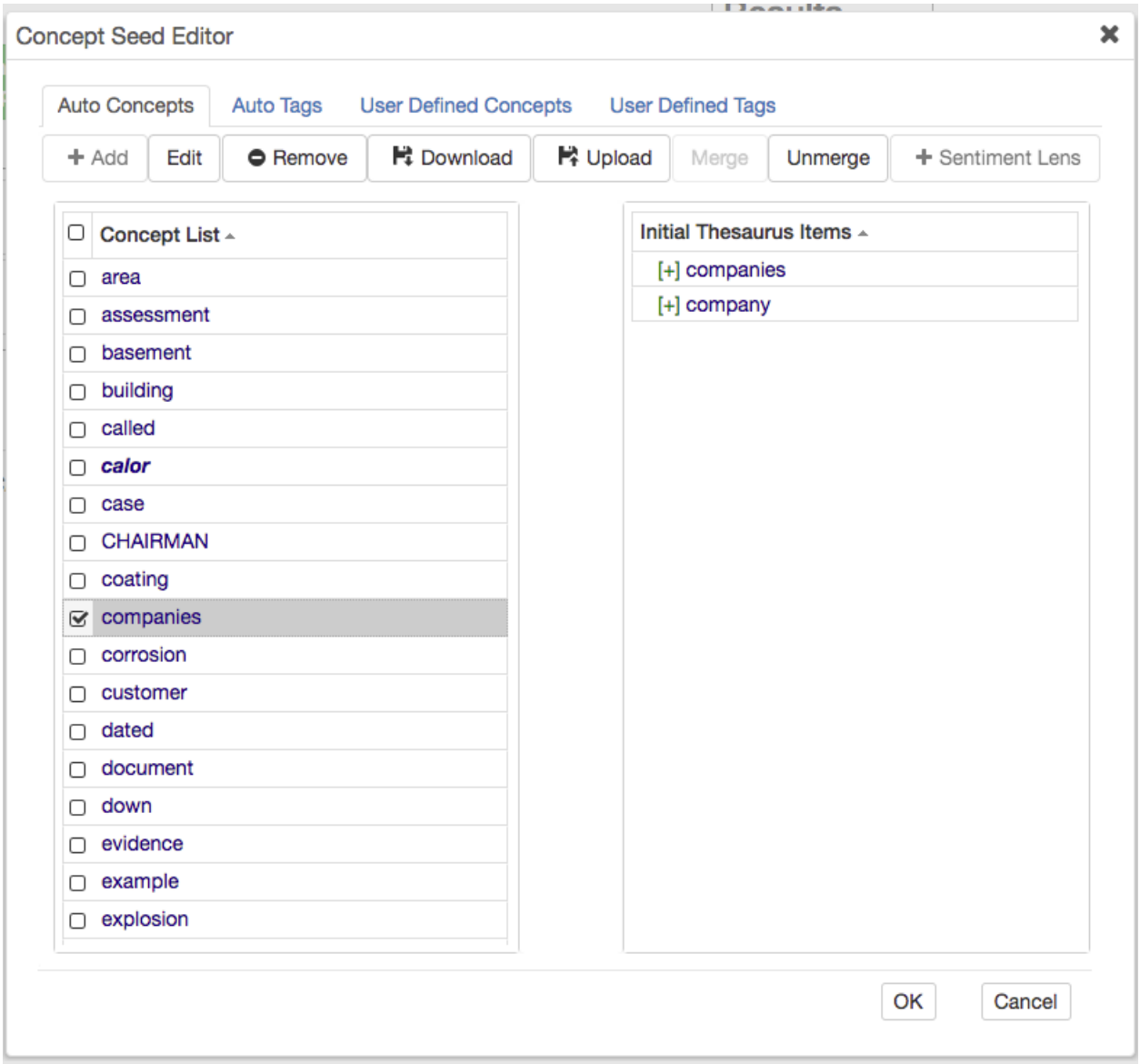
Name-like concept seeds appear in bold and italicized font in the alphabetical list.

Prior to thesaurus learning, only the central seed term for each concept has been identified. In the dialogue above, clicking to select the concept 'building' reveals a single seed term (identical to the concept name) comprising the thesaurus for this concept at this stage:



Single click on concept seeds to select them. Use the Remove button to remove any unwanted automatic concept seeds. Use the Select All button, or hold down control <ctrl> while clicking, to select multiple items.

You can merge similar concept seeds by selecting two concepts and clicking the 'Merge' button. If you do so, the merged concept takes its name from one of the concept seeds, and the concept then has two thesaurus items. For example, if you merge the concept seeds 'company' and 'companies', the following will result:

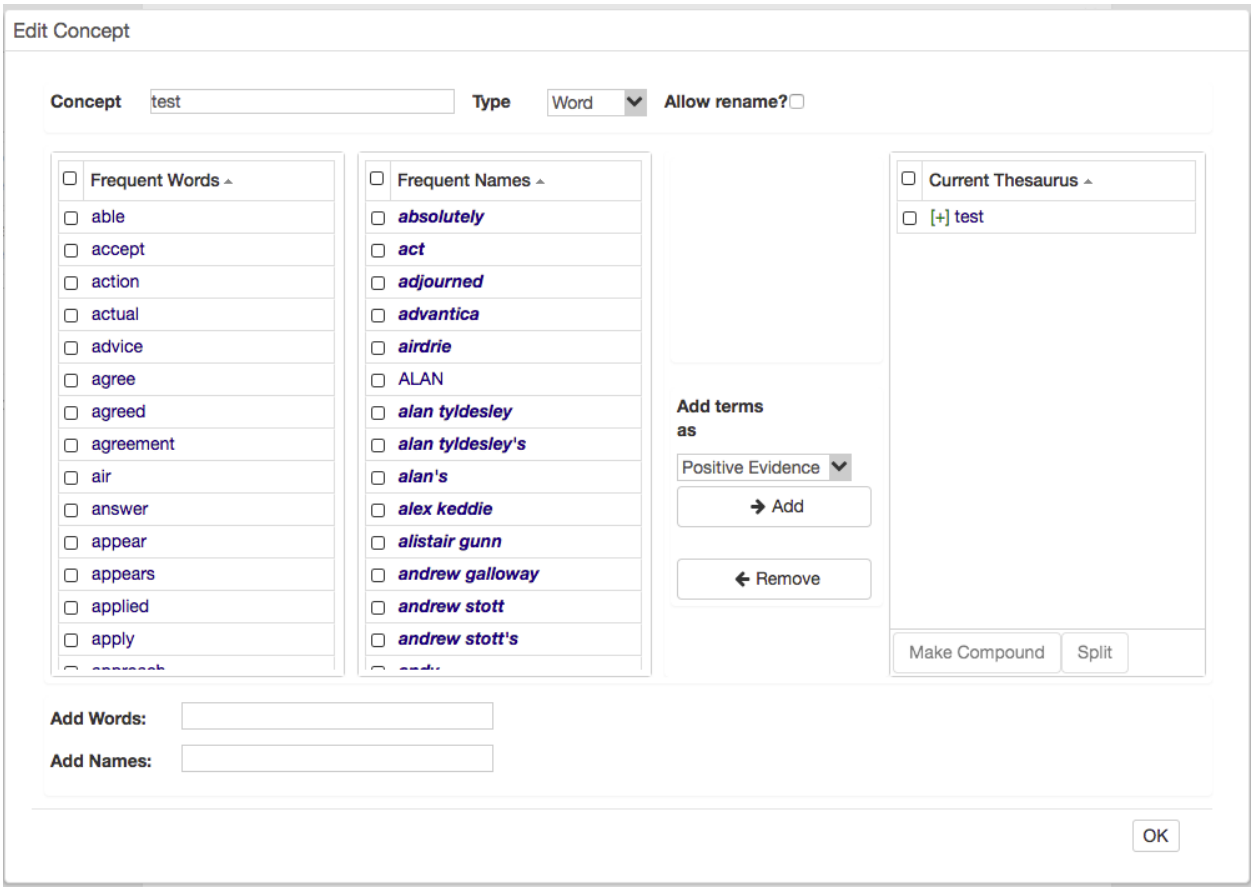


Do not forget to untick the merged item before moving on to work on other concepts. If you do not untick first, the following action will affect the merged item as well.

If you change your mind, you can select the merged concept in the list and use the Unmerge button to separate the two original seeds.

You can also edit automatically extracted concepts. For instance, if you wish to add additional thesaurus items, select the concept from the list and click Edit.

The following dialogue will appear:

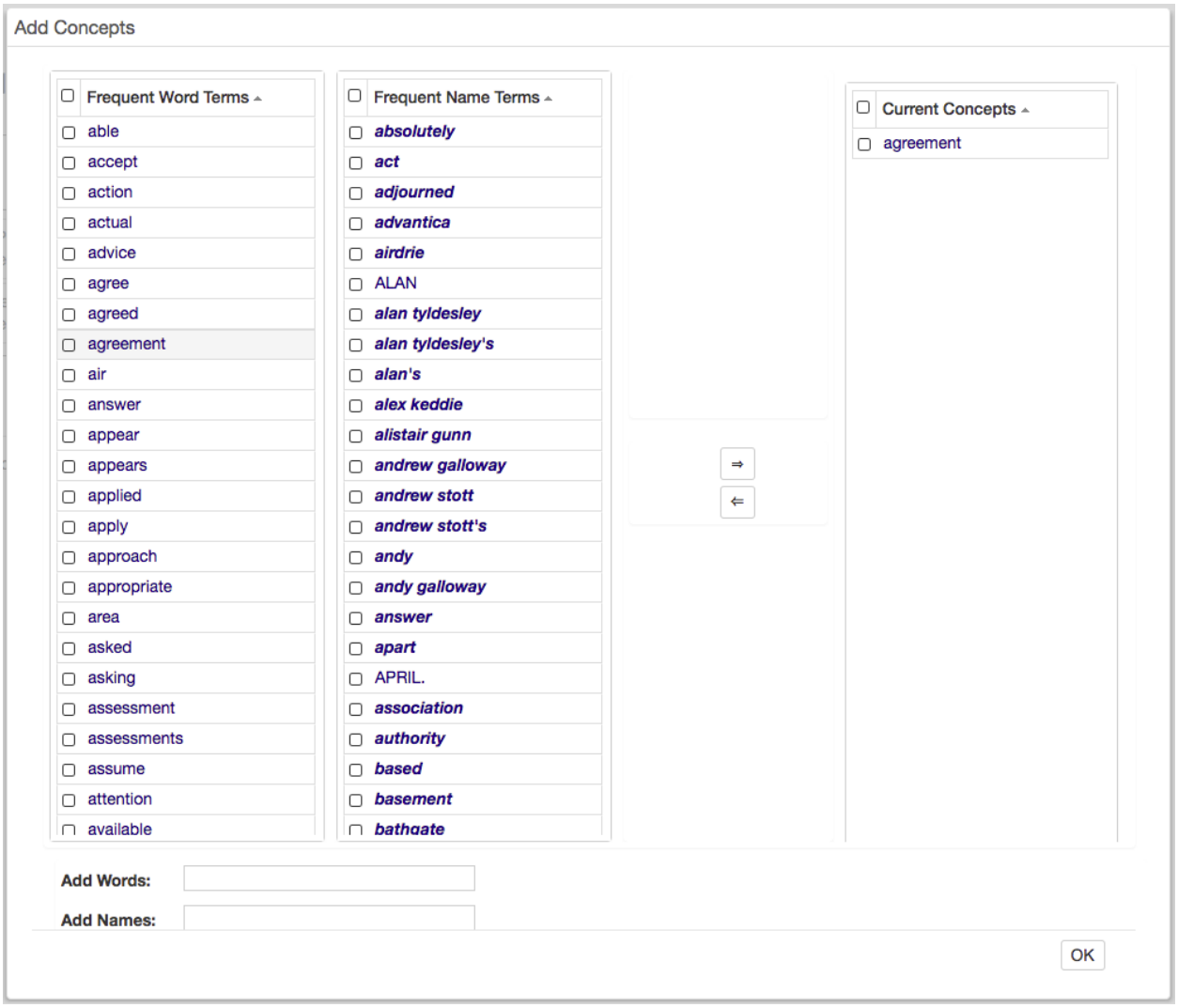


This interface allows you to change the name of the concept, identify it as a name-like or word-like concept, and choose whether to allow Leximancer to rename the concept during subsequent learning.

You can use the arrow buttons to Add or Remove terms related to the concept. Add terms if you believe that there are other words that predict well the presence of the concept in a section of text. You can choose words from the Frequent Words or Frequent Names lists, or enter your own words in the Add Words or Add Names text boxes. You can also identify terms that constitute negative evidence, or evidence that a concept is absent, but use this option with care. Leximancer will automatically learn the weightings for these words from the text during the Thesaurus Learning phase.

If you wish to create your own concept(s), close this dialogue and return to the Concept Seeds interface. Click on the **User Defined Concepts** tab and then click on Add. This opens the Add Concepts interface, where you can define new concepts yourself.

Name the new concept using the lists of frequent words and names, or type a concept name into the text box. Use the right arrow button to move the name of the new concept over to the Current Concepts list:



Click OK to close this window to see your new concept under the User Defined Concepts tab. The new user-defined concept can now be edited in a similar fashion to automatic concepts.

If you wish to rerun the project from a prior stage and retain your edits to the **auto** concepts, you should click OK to save your edits. Then reopen the Edit Concept Seeds interface, and Download your edited list of concepts somewhere on your local drives.

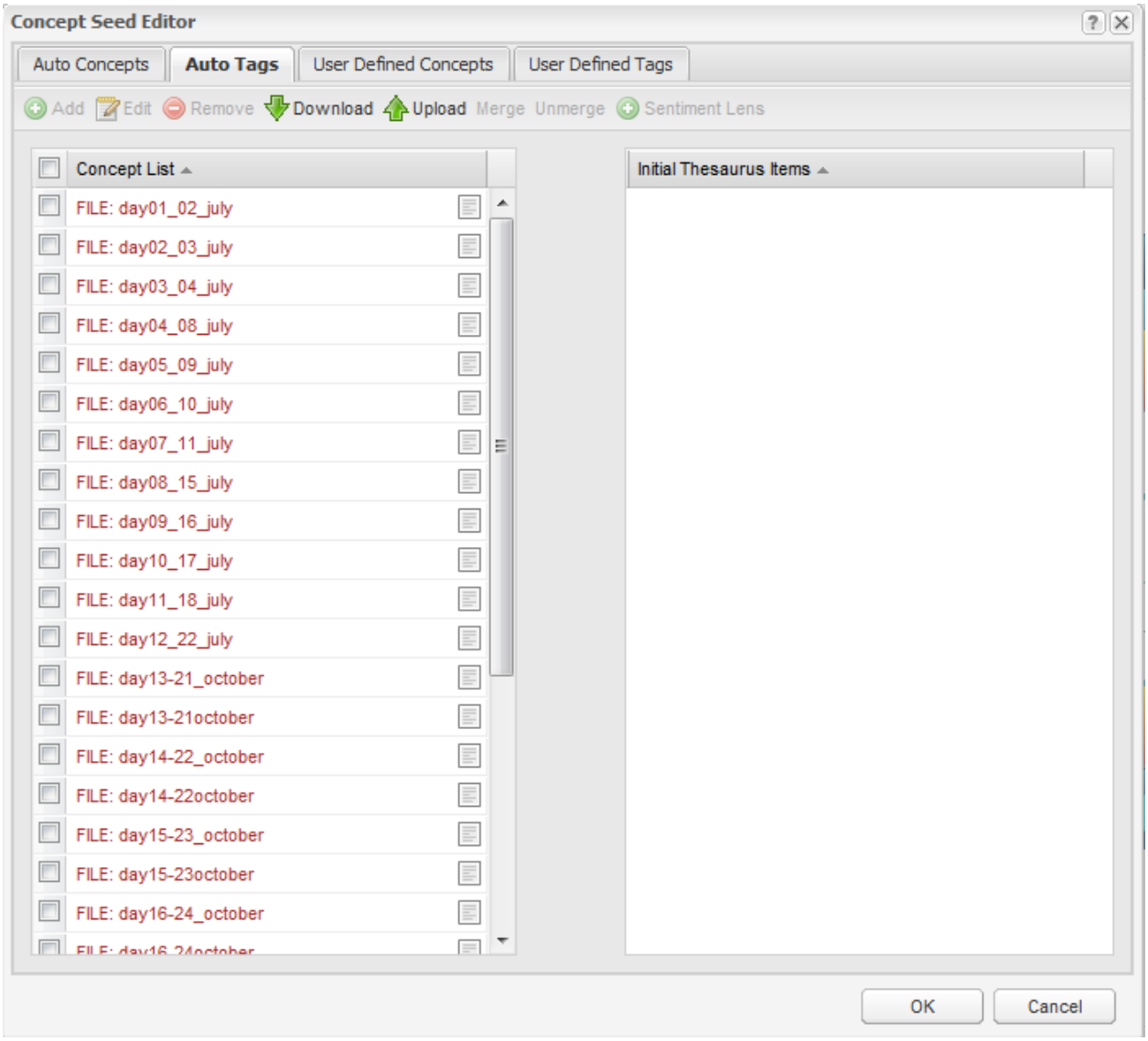
If you run the Generate Concept Seeds stage again, Leximancer will clear and regenerate its list of **auto** concepts.

If you wish to use your saved list of edited concepts, go into the Concept Seeds interface again, and Upload your saved concepts seeds file in the Auto-concepts tab, or one of the other tabs if desired.

You must save the concept seeds in each of the Auto- and User-defined tabs separately. The separation affords greater control by allowing you to reload individual seeds lists if you wish.

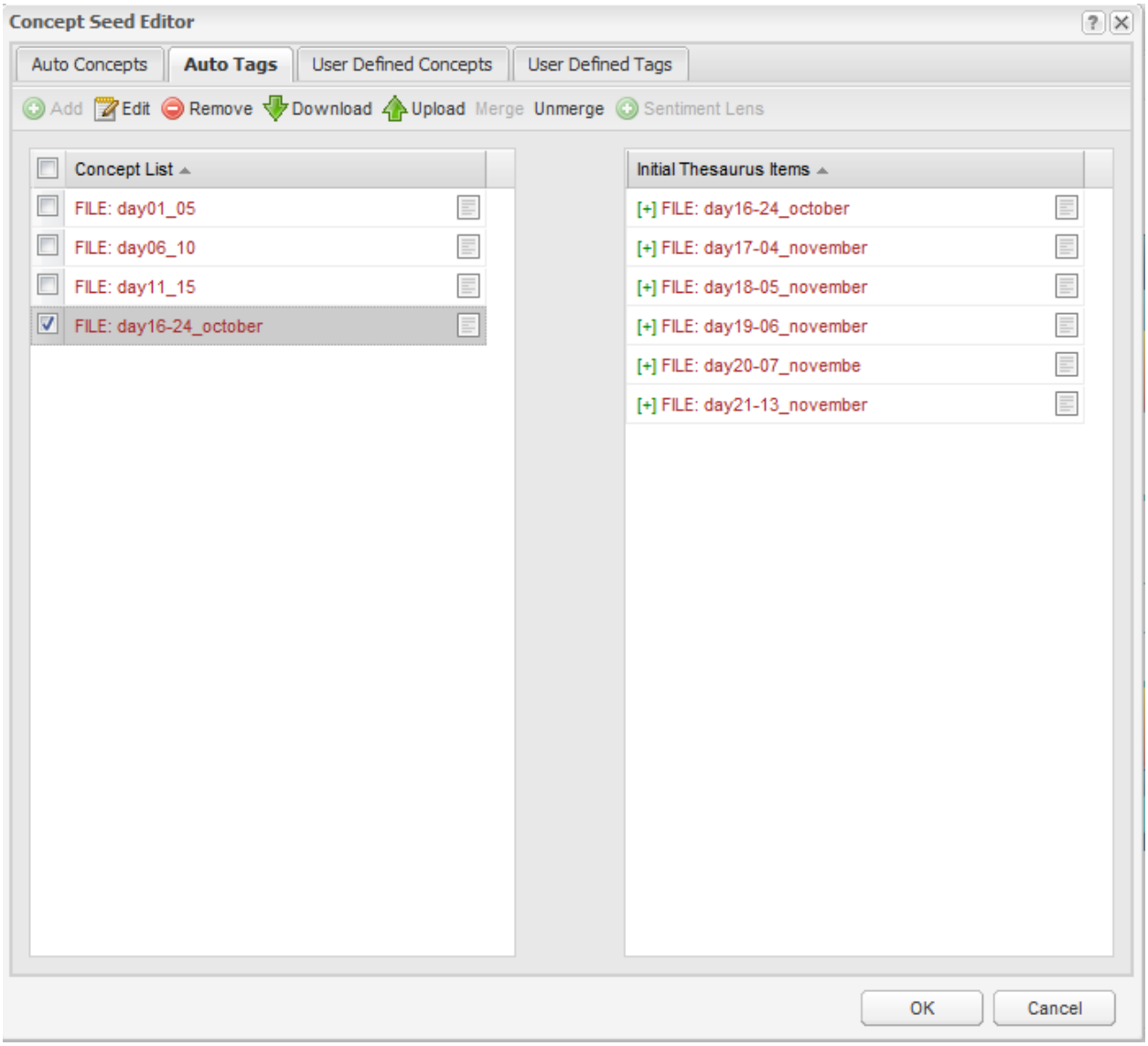
4.4.1 Using Tags

If you have opted to Apply File Tags in the Text Processing settings, then you should see a 'tag' representing each of your source documents in the Auto Tags tab in the Concept Seeds interface:



Tag concepts are concepts for which no associated terms will be learned by Leximancer (unless otherwise instructed). They are useful if you want to make comparisons among groups within the data.

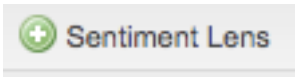
You can aggregate the tags using the Merge button, similar to merging concepts. In this case for example, we could merge the document tags to create 4 weeks of hearing transcripts for comparison:



If you have created Folder Tags, the numbering lets you know in which level of the hierarchy a folder resides (Level 1 is the top level).

You can also create User-Defined Tags to perform a simple keyword search for particular terms of interest.

4.4.2 The Automatic Sentiment Lens

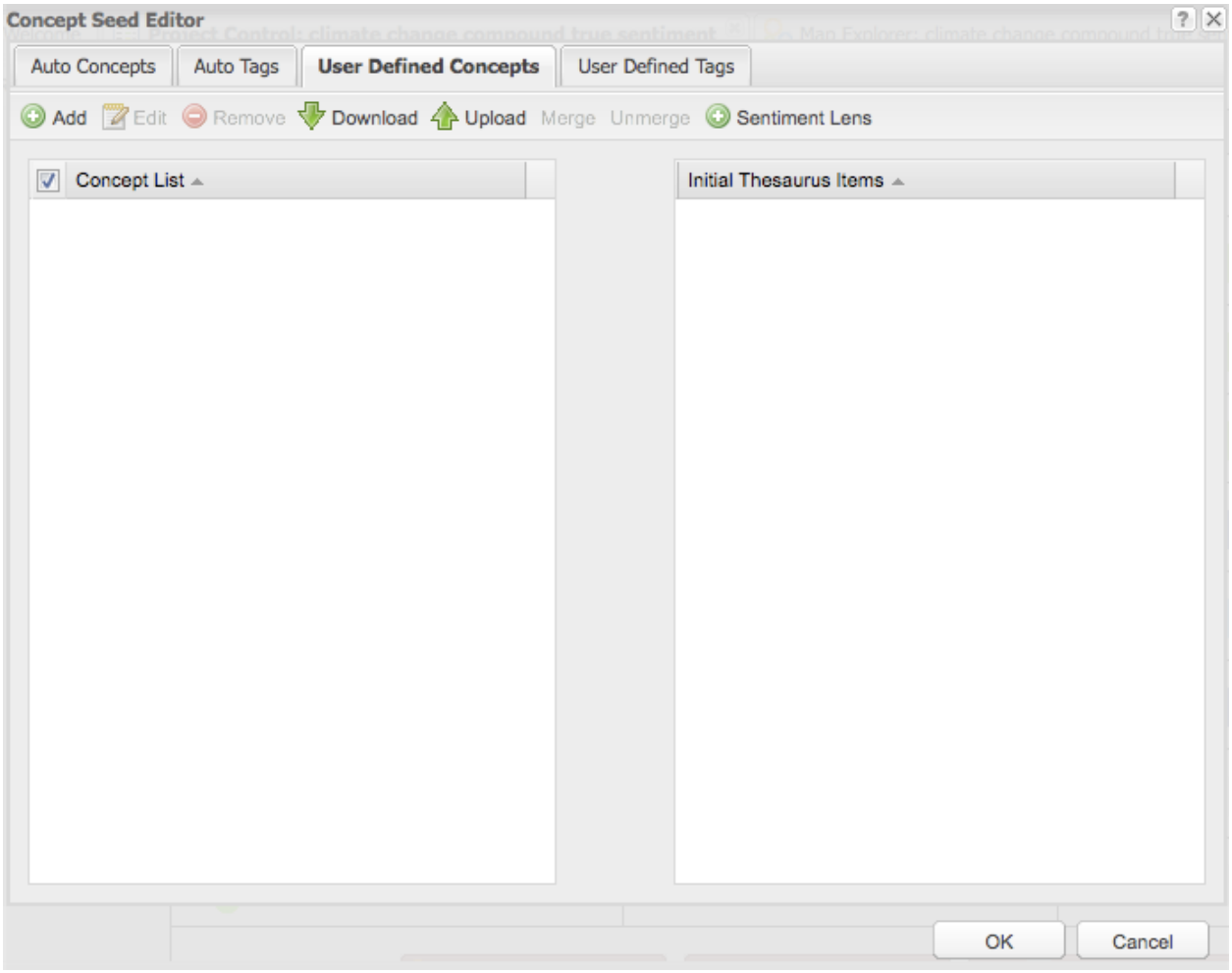


The Sentiment Lens automatically generates insight into positive and negative sentiment in your text. A default set of sentiment concept seeds and their terms are added to your user-defined list. Sentiment Lens will only apply the sentiment terms that are identified as relevant and used

consistently within your document set during processing. Sentiment Lens increases both the ease and accuracy of sentiment analysis

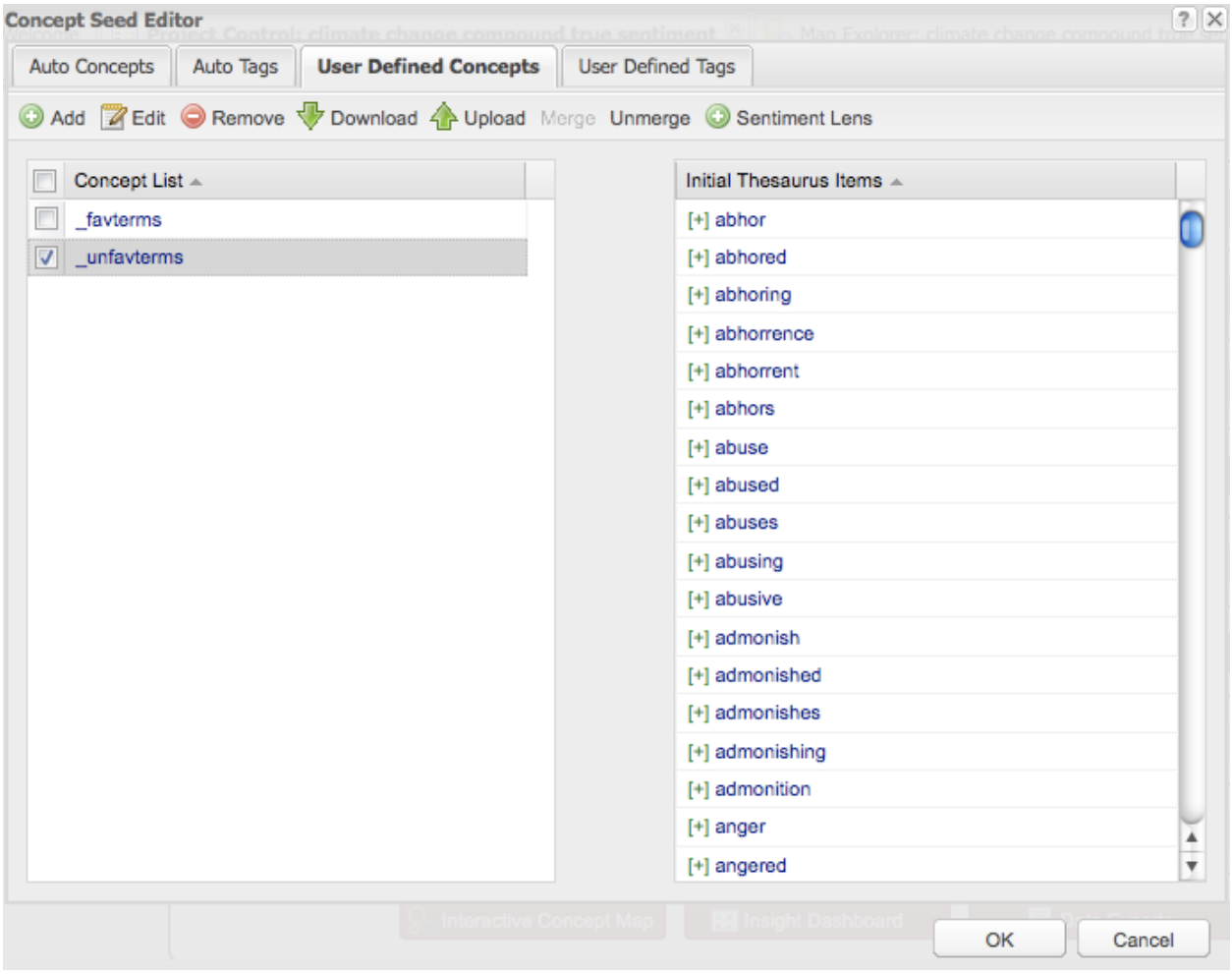
Configuring Sentiment Lens

Under the 'User Defined Concepts' tab, you will notice a button on the top right: Sentiment Lens. Clicking this button will merge a pre-defined list of sentiment seeds into your editor:



**Please note: if you make other changes to your user seeds, including tags, you must make these changes FIRST and then save them by clicking 'ok'. THEN you may re-enter the 'Concept Seeds' stage and use Sentiment Lens. Leximancer will display a warning dialog if you attempt to run without saving or discarding changes.

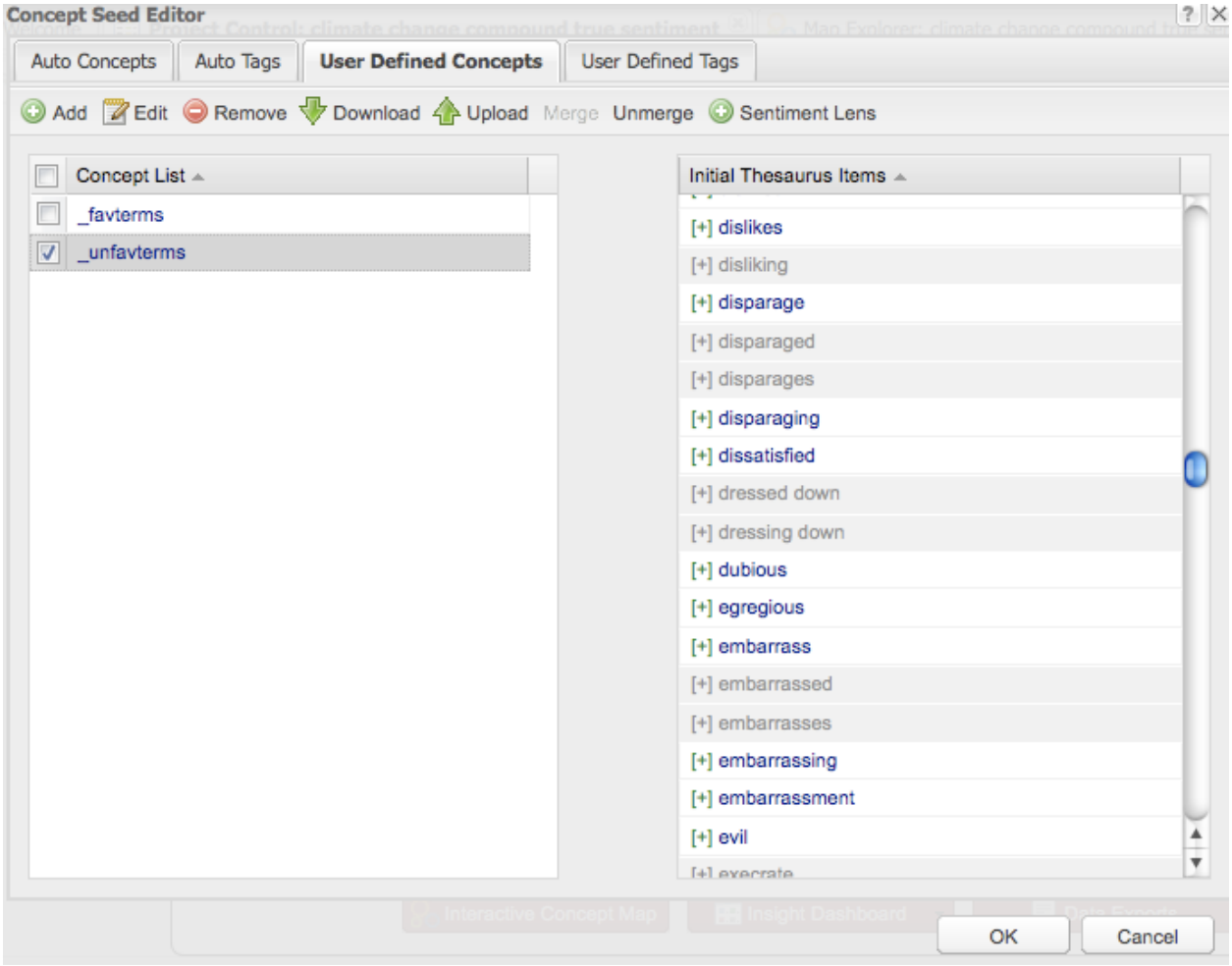
A list of sentiment seeds should appear under the categories '_favterms' and '_unfavterms'. Under the 'User Defined Tags' tab is '_negationterms'.



- '_favterms' is a list of commonly used favourable sentiment terms: approve, best, commend, favour, lauded, praise, great, etc
- '_unfavterms' is a list of negative sentiment terms: abhor, anger, blame, denounce, frown, revolt, etc
- '_negationterms' is a short list of words causing the following word to be negated: not, nothing, etc.

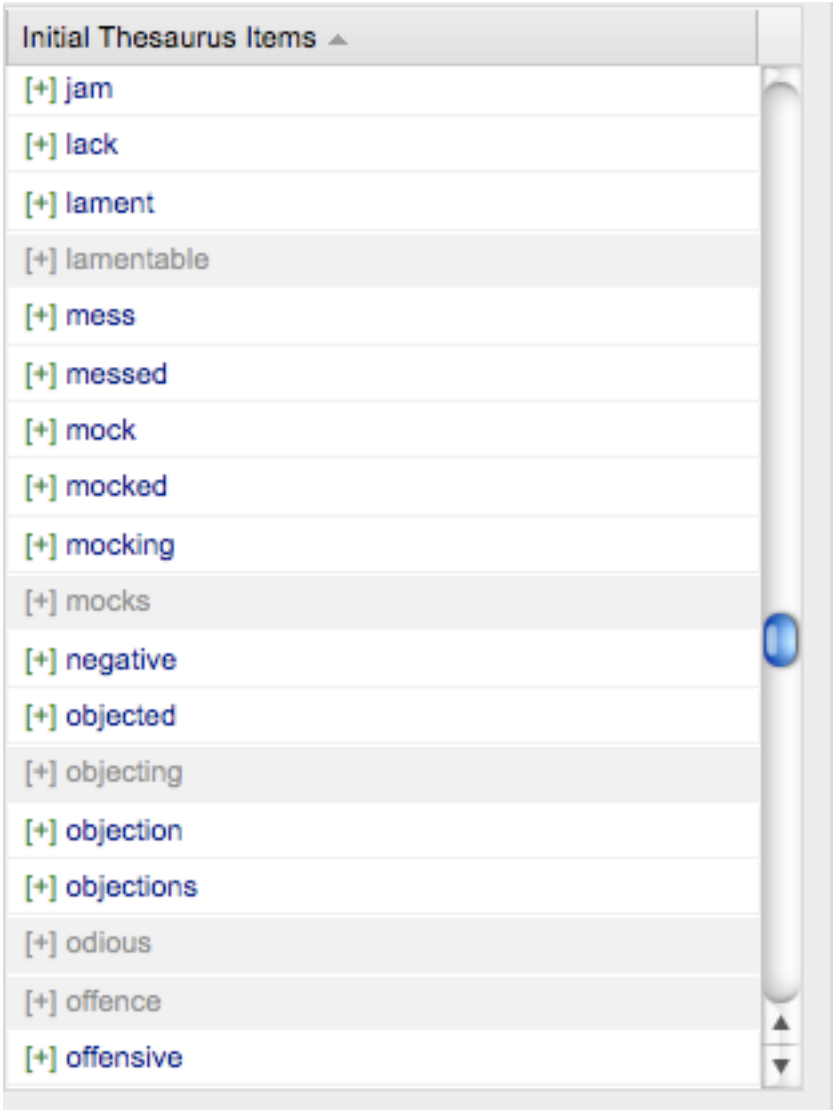
Click 'Okay' to return to the control panel.

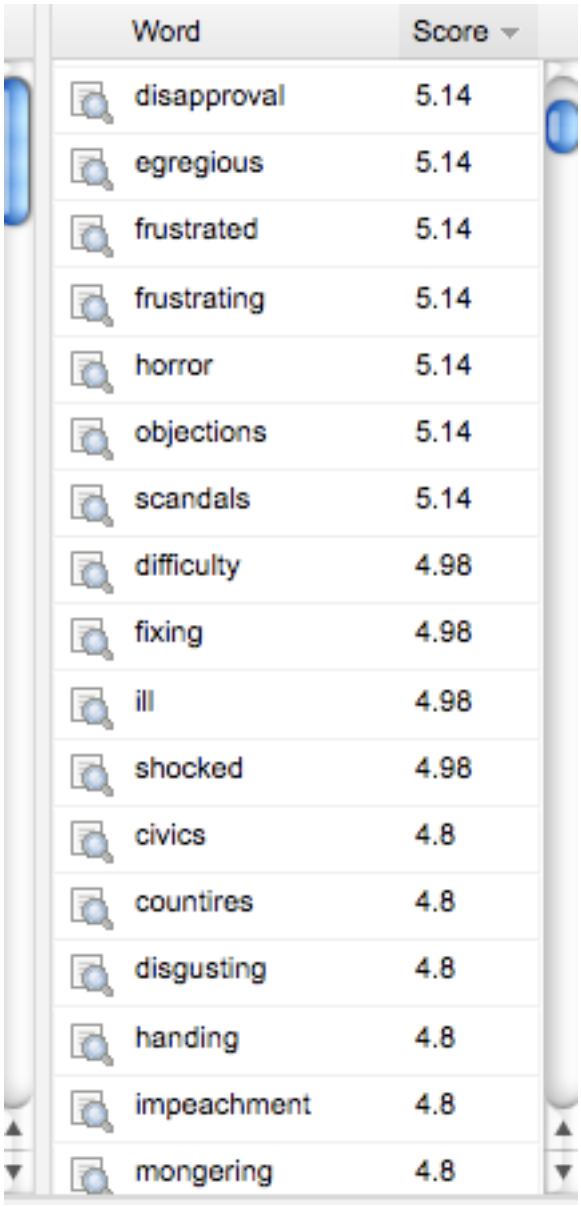
If you run the Thesaurus Learning stage, and then return to the 'User Defined Concepts' tab in the 'Concept Seeds' stage, you can observe the effect of Sentiment Lens. Sentiment thesaurus terms that are irrelevant or inconsistent in your text will be grey. Those left coloured are suitable for analyzing sentiment in your text and will be used as thesaurus items to develop sentiment concepts.



Once you reach the Concept Map, you may also observe new Sentiment terms that have been automatically added to the Thesaurus.

When analysing news articles about climate change for example, the term 'mongering' does not appear in the original seed list under '_unfavterms'. Yet once Sentiment Lens is applied and run, it appears in the list of Thesaurus items for '_unfavterms' (next page):


Initial Thesaurus Items
[+] jam
[+] lack
[+] lament
[+] lamentable
[+] mess
[+] messed
[+] mock
[+] mocked
[+] mocking
[+] mocks
[+] negative
[+] objected
[+] objecting
[+] objection
[+] objections
[+] odious
[+] offence
[+] offensive



Word	Score
disapproval	5.14
egregious	5.14
frustrated	5.14
frustrating	5.14
horror	5.14
objections	5.14
scandals	5.14
difficulty	4.98
fixing	4.98
ill	4.98
shocked	4.98
civics	4.8
countires	4.8
disgusting	4.8
handing	4.8
impeachment	4.8
mongering	4.8

Would appear here.

Concept Seed Editor stage Thesaurus tab in Concept Map stage

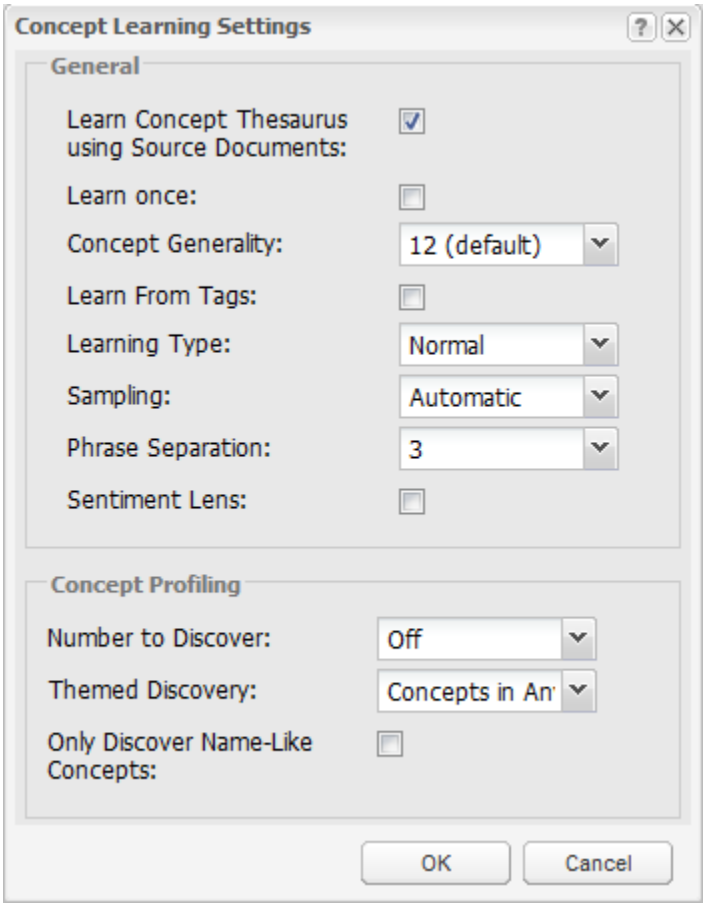
This is because in the climate change literature, the term ‘fear mongering’ is used with negative connotations to describe climate change science. Hence, the automatic Sentiment Lens has picked it up as a term that contributes evidence to unfavourable sentiment, even in the absence of the word ’fear’ preceding it.

4.5 3b. Generating the Thesaurus

The Thesaurus Learning phase generates a thesaurus of terms associated with each concept. Concepts are collections of correlated words that encompass a central theme, for example, the concept *client* may contain the terms customer, customers, client, clients, subscriber and subscribers. The learning of the thesaurus associated with each concept is an iterative process. Seed words start out as being the central terms of a concept, collecting related keywords over time. Through learning, the seed items can be pushed to the periphery if more important terms are discovered.

Configuring Thesaurus Learning

Clicking on the Thesaurus Settings node reveals the following interface:



The Learn Concept Thesaurus From Source Documents (Yes|No) option allows you to turn off the thesaurus learning and prevents Leximancer from adding additional items to the concept definitions. This will result in searches for concepts as keywords, rather than using a weighted accumulation of evidence. This may be essential for data sets shorter than a few pages. In such cases, thesaurus abstraction is less useful due to a smaller vocabulary. Few sensible emergent concepts would be produced.

Concept Generality (1-21): This setting allows you to control the generality of each learned concept. This value is inversely related to the *relevancy threshold* required for a word to be included

in the concept thesaurus. Raising this value will increase the fuzziness and generality of each concept definition by increasing the number of words that will be included in each concept. After you have run the learning phase, examine the log to see how many iterations of thesaurus learning took place to arrive at the final concept definitions. This number should ideally be between 3 and 8. If the number is more than 8, consider lowering the learning threshold. Conversely, if the number of iterations is very low, consider raising this threshold.

Commentary: This setting controls how easy it is for the thesaurus learning to grow a concept to a broader set of words. The easier it is, the more iterations of learning will occur, the more words will be added to the thesaurus, and the higher the risk of the concept growing into a more general concept. This can result in the concept being renamed or being subsumed into another concept which already exists. If you examine the log file after learning you can monitor this behaviour. If the data does not consist of natural language, you should probably disable thesaurus learning, as described above.

Learn From Tags (Yes|No): You can use this option if you have any tags, either automatically extracted from tables, file or folder names, or speaker tags, or manually entered user tags. Turning on Learn From Tags will treat tags like concept seeds, learning a thesaurus definition for each. This setting is important if you are conducting Concept Profiling (discussed below) where you wish to extract concepts that discriminate between different folders or files (such as extracting what topics segregate Liberal from Labour party speeches).

Learning Type (Normal|Supervised): There are two forms of learning that are supported by Leximancer: Automatic and Supervised. Automatic is the default behaviour, and in this a concept thesaurus is learned to characterize a list of seed words. For example, the initial seed word ‘dog’ be included in the thesaurus for the Dog concept, which also contains a collection of other dog-like terms discovered from related text.

Supervised classification, in contrast to Automatic learning, is possible when you have a complete definition of a category (such as a folder tag, speaker label, table category, or human coding tag) already embedded in some training text. In this case, you want to build classifiers that attempt to faithfully match human classification decisions, rather than discover an extended thesaurus from seed words. The learned concept should not include the initial seed item from its thesaurus definition. You are effectively giving Leximancer examples of a concept that you want learned. For example, if you were training the system to learn the concept ‘violence’, you might write the code ‘Violence’ in locations of the text that you think are examples of this concept. However, you do not wish the term ‘Violence’ to be a crucial term that is required to trigger the concept. Instead, a concept will be created, encompassing other discriminating terms from those contexts, that does not include this supervised term in the extracted classifier.

Sampling (Automatic|1-10): Sampling during learning speeds up the learning process by only reading every nth block of text. The automatic setting is normally fine, but you can override this, if necessary, by choosing n.

Commentary: The automatic sampling setting looks at the size of the total text data set to decide an appropriate value. The actual sampling number is increased by 1 for every 15 Mb of text data, so for anything under 15 Mb, sampling of 1 is used, which means every context block is examined for learning. For 15 to 30 Mb of text data, a sampling of 2 will be used, which means that every second

context block is examined for learning. Note that classification never uses sampling, and classifies every context block. Tests have show that classification performance only decreases marginally if the automatic schedule is followed.

4.5.1 Concept Profiling

These settings allow the learning process to discover new concepts that are associated with selected user-defined and automatic concepts. This is useful for profiling concepts or names, for doing discriminant analysis on prior concepts, or for adding a layer of more specific concepts which expand upon a top layer of general concepts. Profiling also allows you to ignore large sections of text that are not relevant to your particular interests.

Once the initial concept definitions have been created, words that are highly relevant to these concepts can be identified as potential seeds for new concepts. For example, if you profile the initial seed 'flowers', a concept definition is grown around this word as usual. Then new concepts are developed from the 'flowers' definition that would produce more specific topics, such as 'roses', 'daffodils', 'petals', and 'bees'.

This process is useful if you are trying to generate concepts that will allow segregation between various document categories. For example, if you are trying to discover differences between good and bad applicants, simply place exemplars of each in two separate folders (one for each type), and Folder Tags in the Preprocessing stage. This will create a tag for each folder. In the Concept Editor, only retain these folder tags in your Automatic Concepts and Tags lists. Switch on Learn From Tags in the Thesaurus Learning phase, and use the profiling settings described below to extract relevant segregating concepts.

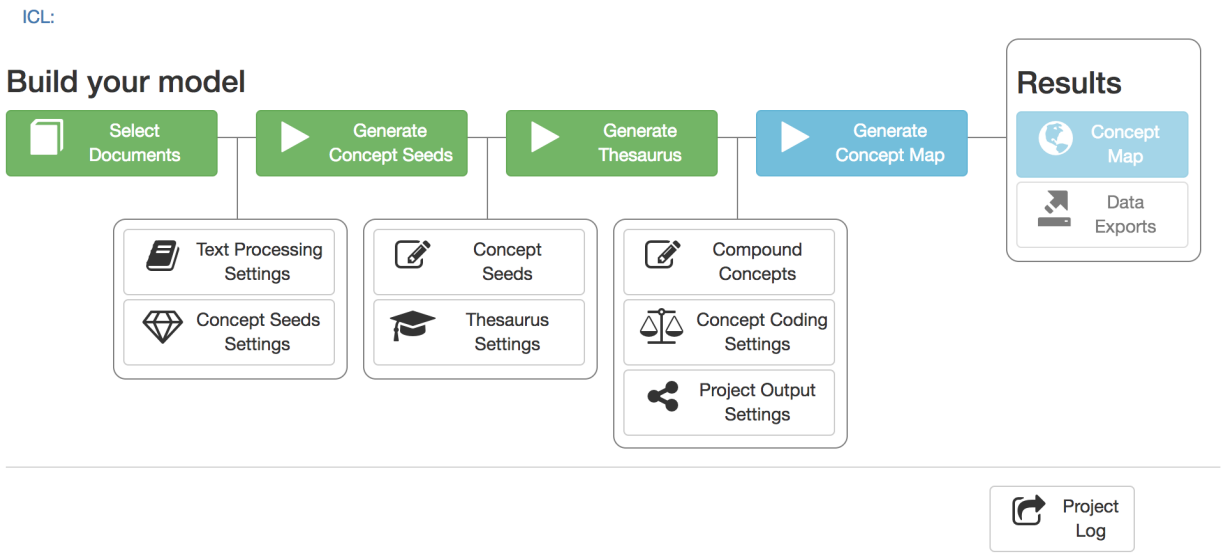
Number to Discover (Off|10-1000): This parameter specifies how many concepts should be profiled or discovered from the pre-defined concepts. More pre-defined concepts normally require more discovered concepts, but less than 100 is recommended. As a guide, select between 3 and 10 discovered concepts per pre-defined concept to give a reasonable coverage.

Themed Discovery (Concepts in ALL| Concepts in ANY| Concepts in EACH): Choose how you want the discovered concepts to be related to the pre-defined concept set: ANY gives the Union (related to concept1 OR concept2 ...), EACH gives the Exclusive Disjunction (XOR: related to concept1 OR concept2 but not both), and ALL gives the INTERSECTION (related to concept1 AND concept2). Choosing the intersection of the concepts will only extract concept seeds that are highly relevant to all or most of the learned concepts. For example, conducting a themed discovery from the concepts sun, surf, and sand may lead to concepts more relevant to beach scenarios than using words relevant to only one of these concepts (ie: the union of the concepts). XOR is designed for strong discrimination of target classifications (ie: finding specific concepts that segregate between the predefined concepts).

Commentary: If the pre-defined concepts surround some theme, such as say beach life or environmental issues, you probably want the discovered concepts to follow the theme, so choose the AND (intersection) option. If you want to discover concepts that strongly discriminate between the pre-defined concepts or tags, choose the XOR (disjunction) operator.

Only Discover Names (Yes|No): This option lets you discover name concepts only when profiling. This is useful for discovering social networks of association.

4.6 4. Generate Concept Map:



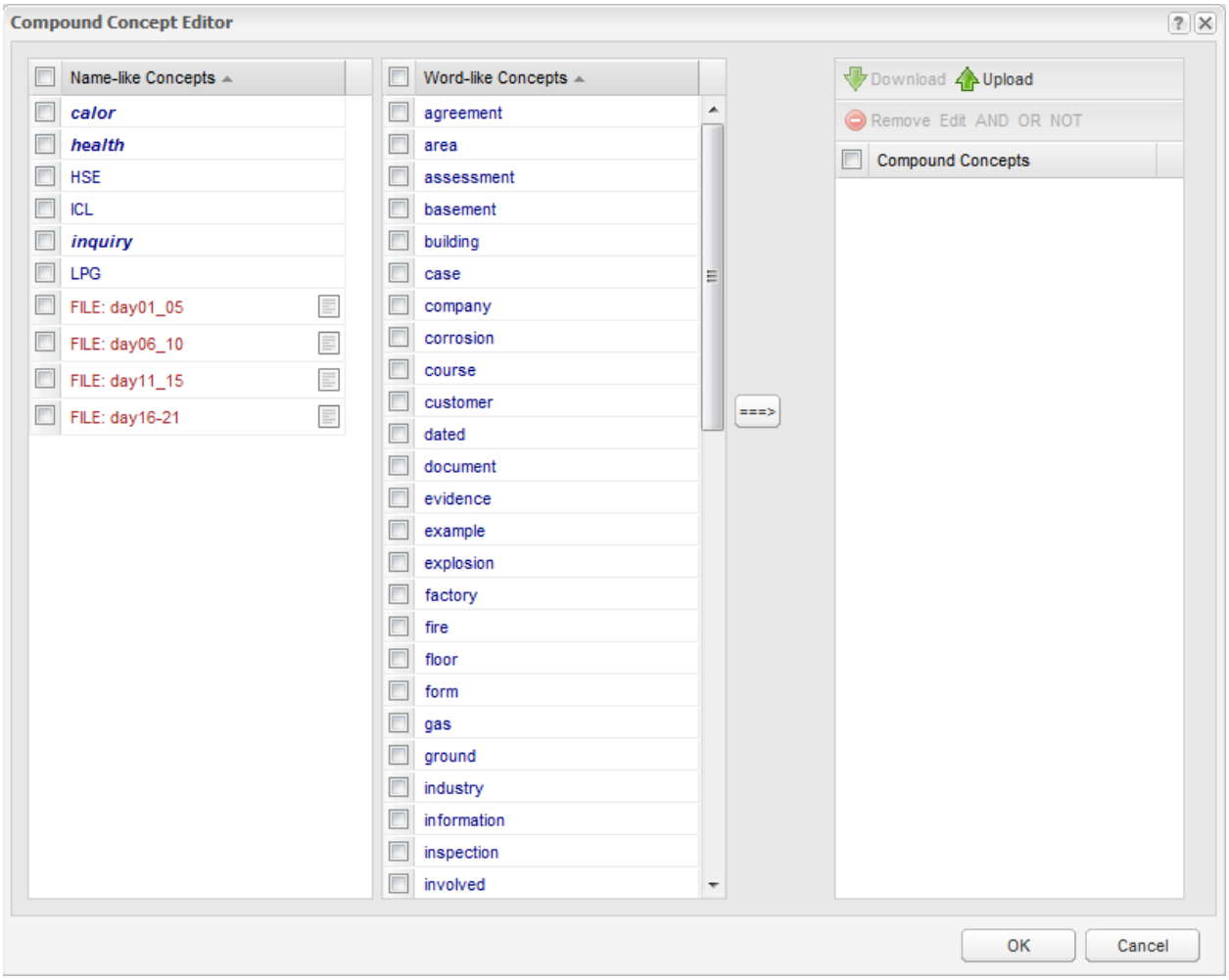
You can simply run this stage of processing (and all of these preceding it) using default settings by clicking the *Generate Concept Map* button.

Optionally, you can Edit the settings for three sub-stages using the boxes below and preceding the *Generate Concept Map* button.

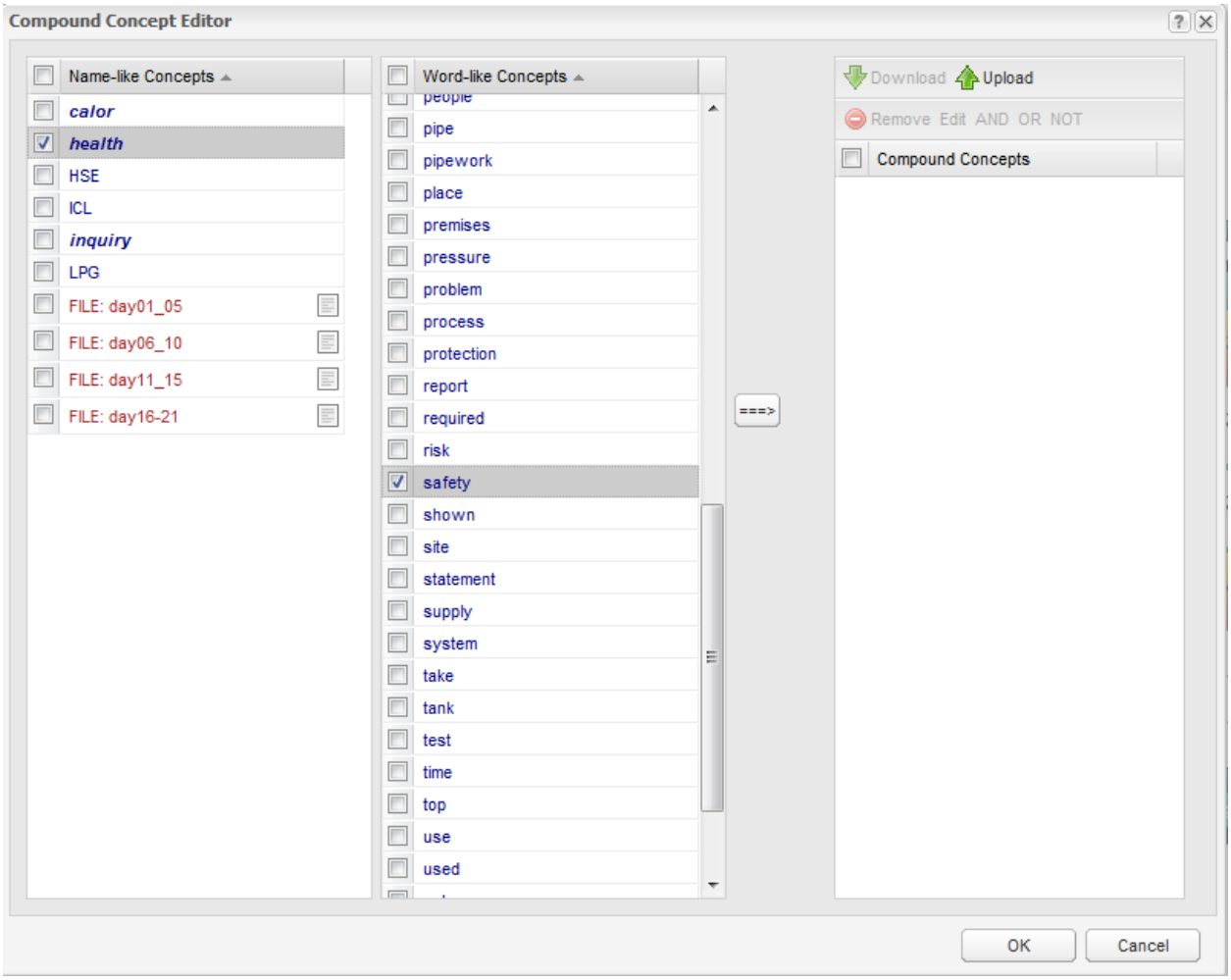
4.7 4a. Editing Compound Concepts

Manually compound selected concepts via Boolean operators to obtain deeper and more meaningful analysis.

Clicking the 'Compound Concepts' node opens this interface:

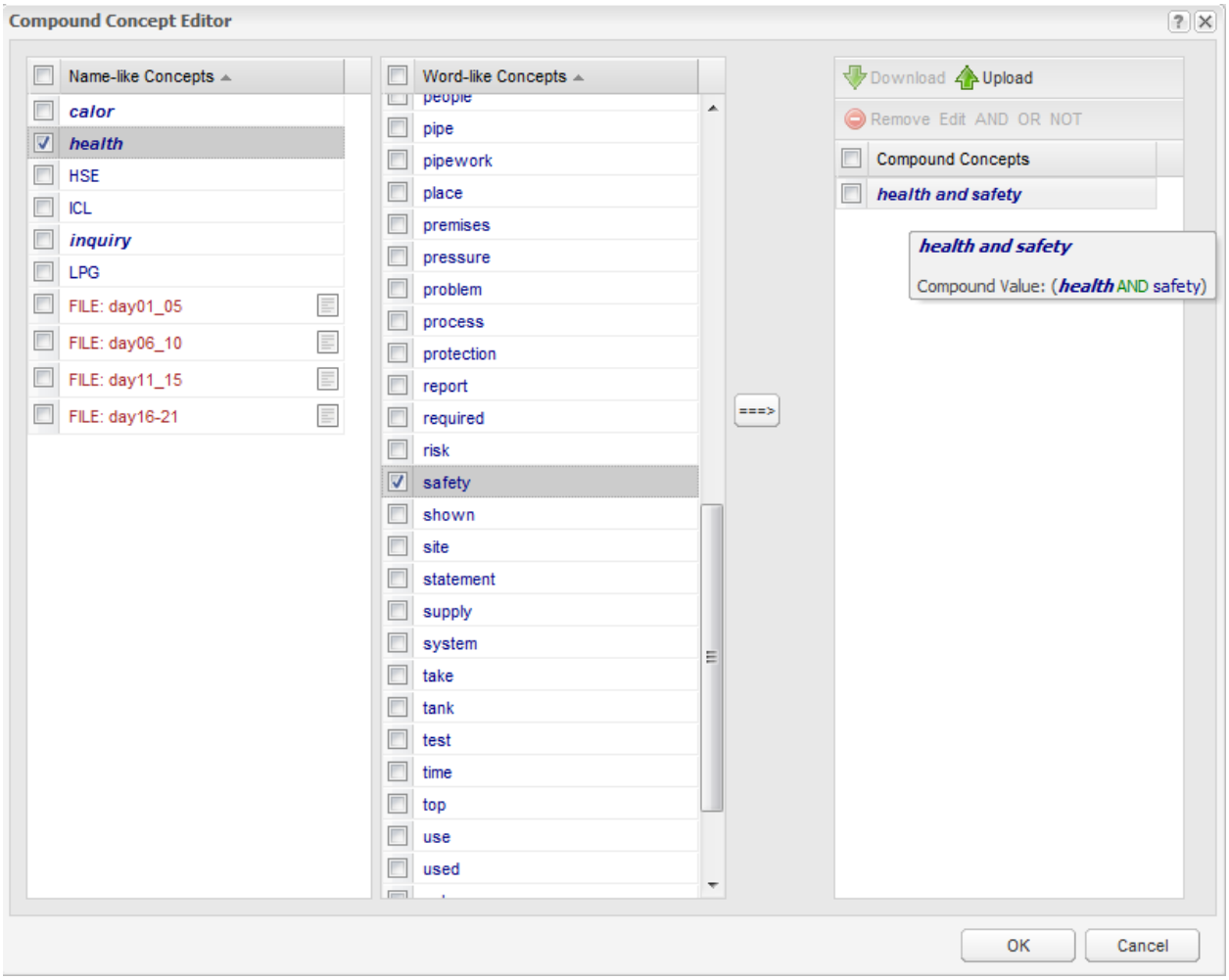


Select the boxes for any concepts you wish to combine into compound concepts. There are lists of all tags, name concepts and word concepts in the left and centre columns. Move them across to the compounding workbench using the right-hand arrow:



Combing the concepts using the AND operator requires both concepts to appear in the same (2-sentence) piece of text for the compound to be coded.

After moving the concepts you wish to combine into the right column, tick both of them again and click the ‘AND’ button in the header:



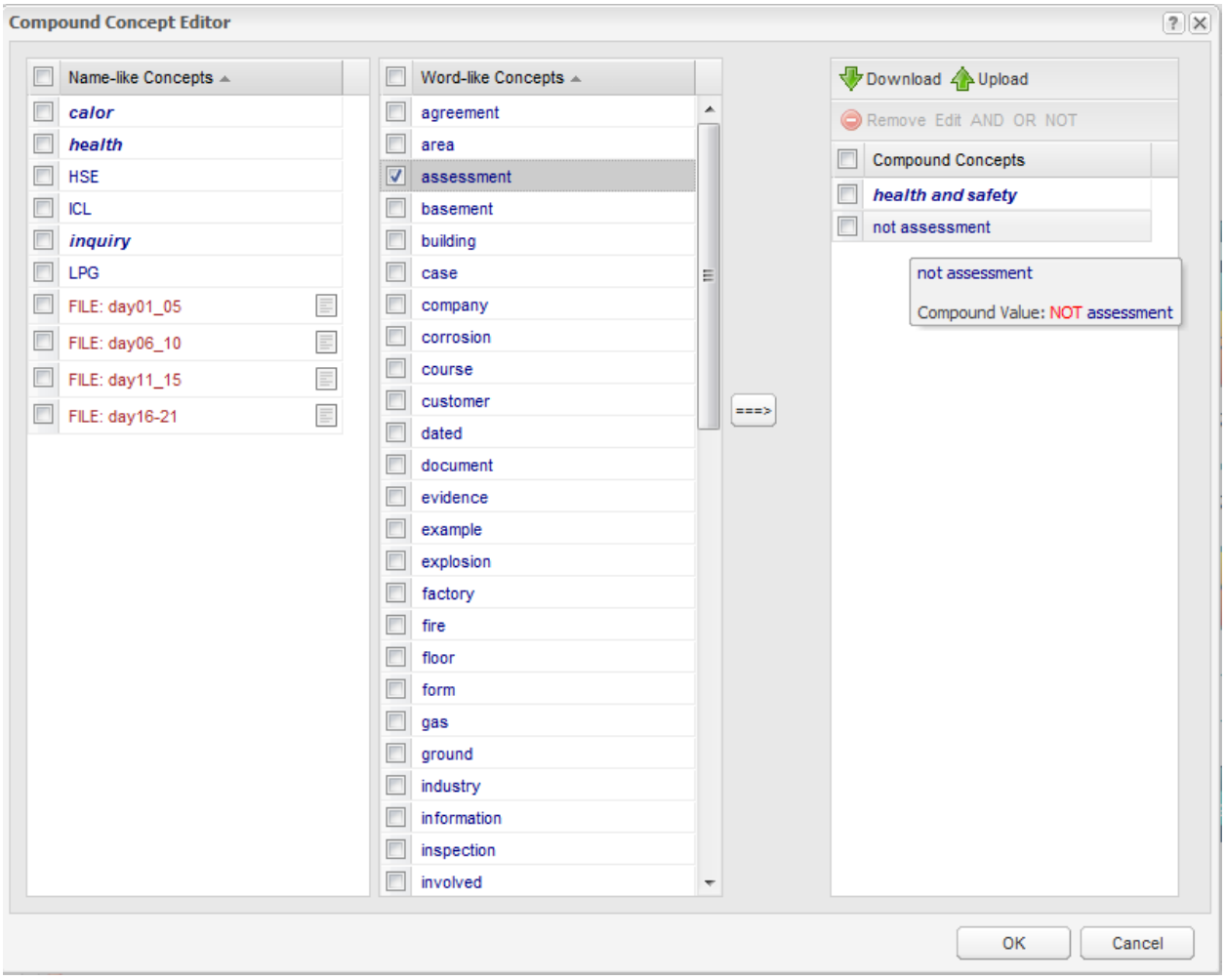
Hover your mouse over the compound concept to see the equation defining it.

Note: Concepts used to build a compound remain available as singular concepts in the project results as well (unless you exclude the constituent concepts from the Mapping Concepts list in the Concept Coding Settings).

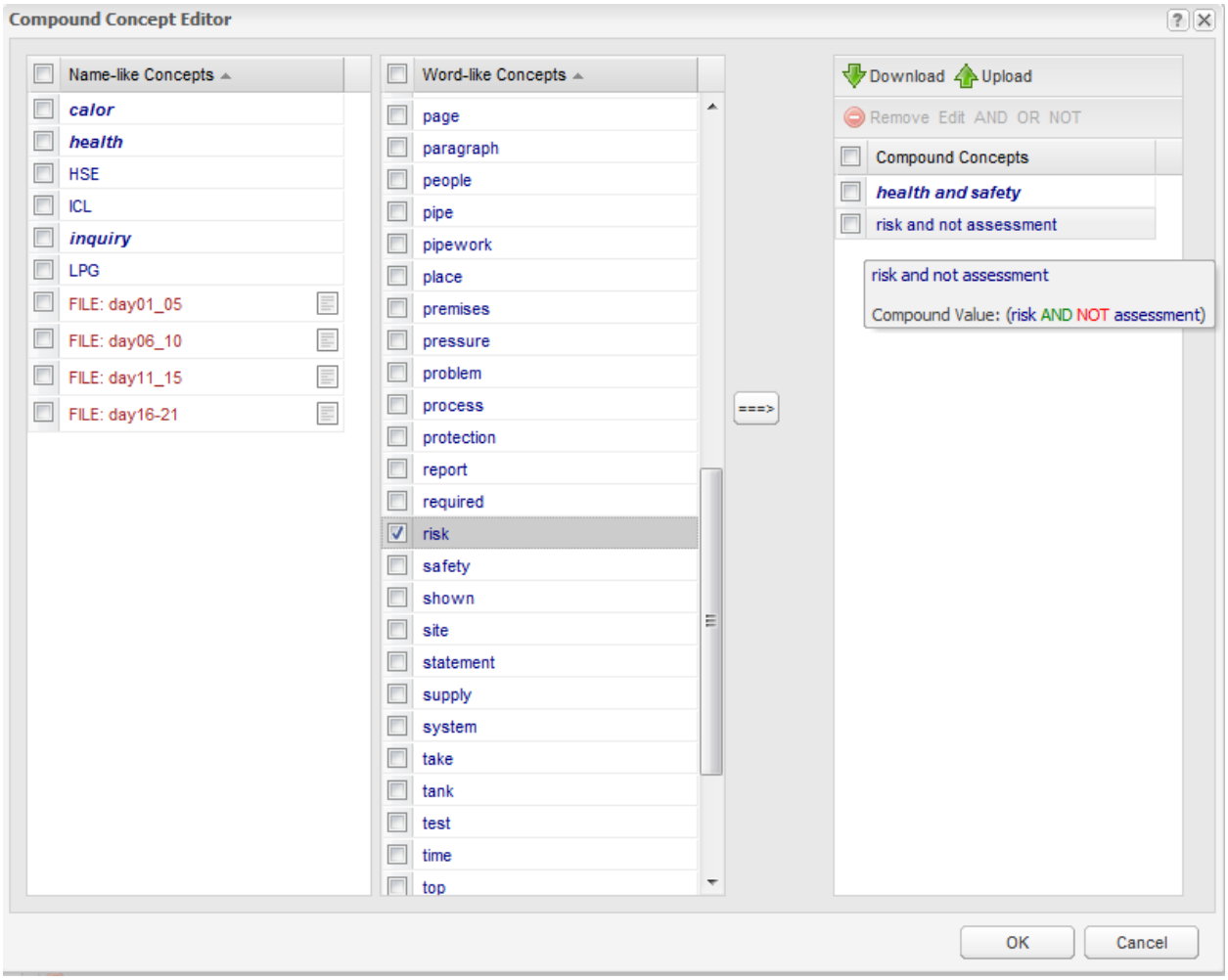
Use the same method to combine concepts with the Boolean operator 'OR'. Using the 'OR' conjunction means that evidence for your compound concept will be calculated to include evidence for either of your concepts.

Compounding concepts using the OR operator is similar to Merging concepts in the Edit Concept Seeds interface.

To make a compound concept using the 'NOT' operator, first select the concept you wish to negate. The 'NOT' button at the top of the column will become available. Clicking the 'NOT' button will negate the concept you have selected:

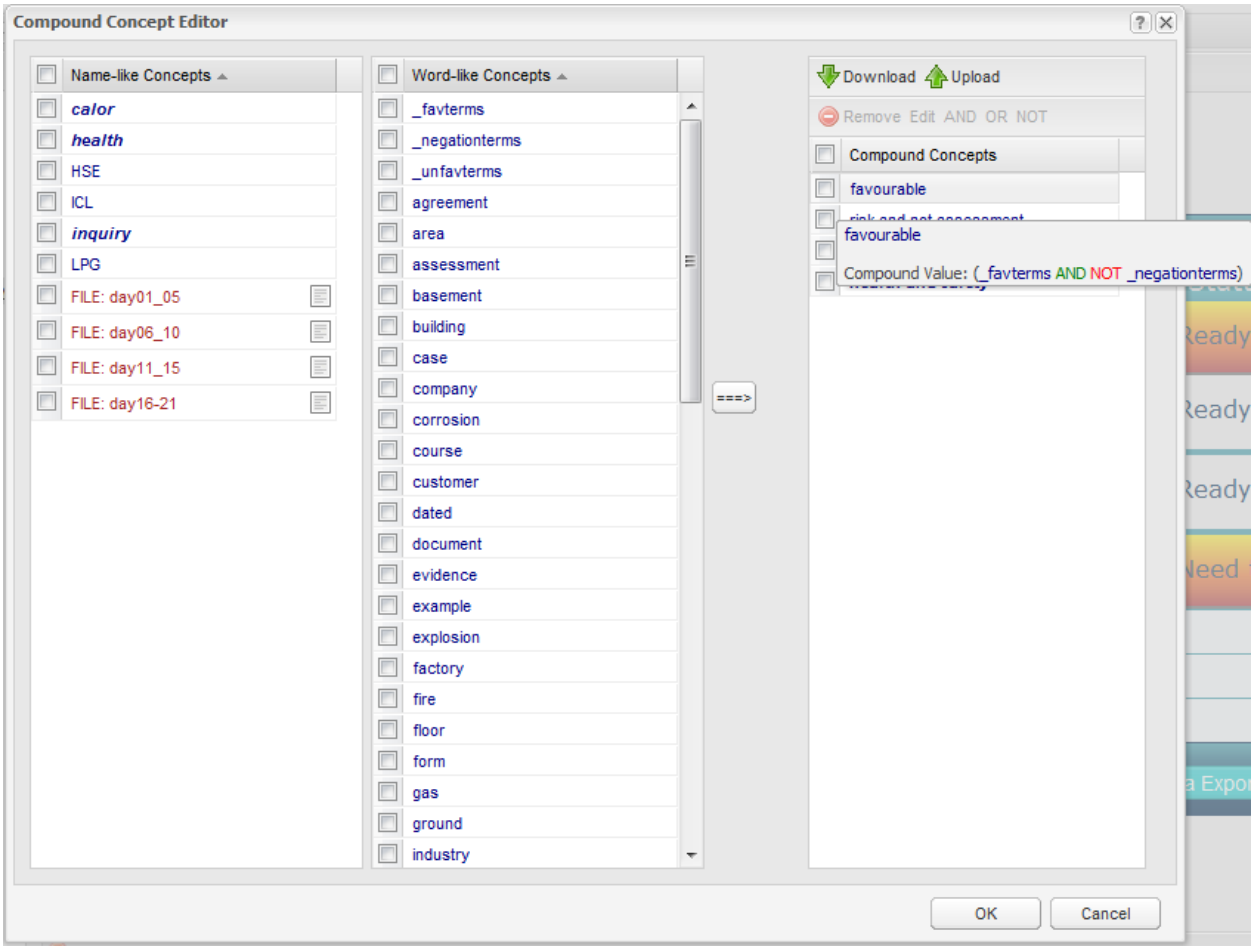


You can now combine the negated concept with another concept by following the steps for combining concepts with 'AND' as outlined above. Doing so will include instances of the positive concept, and exclude instances of the negated concept:



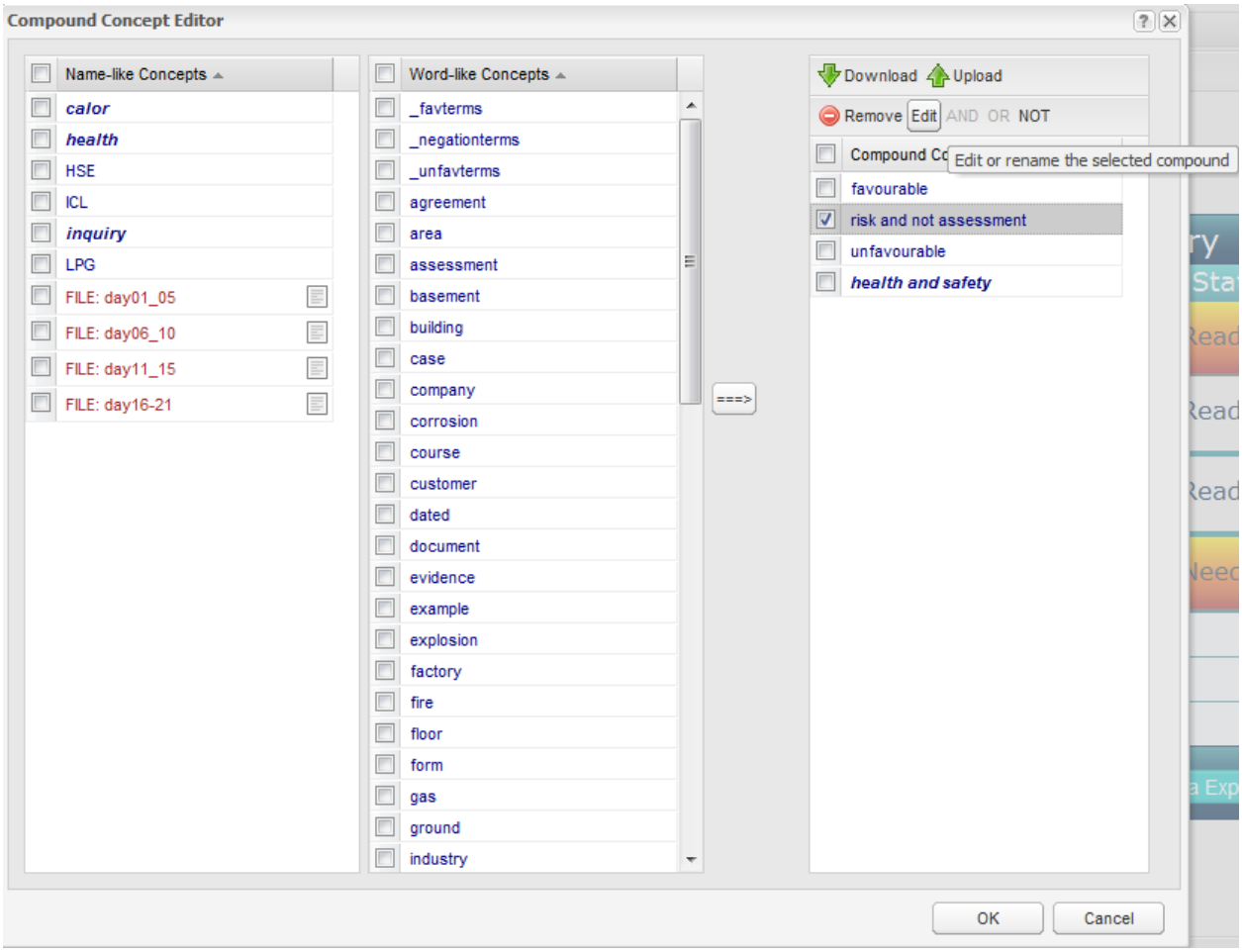
The same procedures outlined above can also be used to build more complex Sentiment concepts.

This is done automatically when you click Sentiment Lens in the Concept Seeds Editing interface. For example, the Sentiment Lens creates a compound concept for Positive Sentiment that includes favourable terms and excludes negation terms:

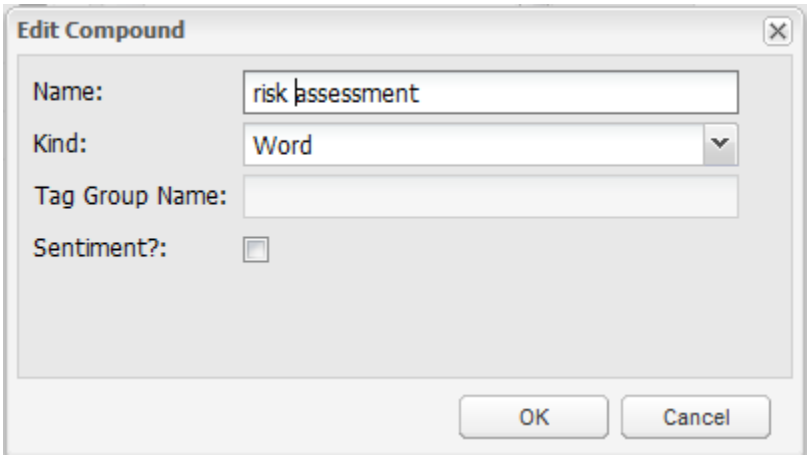


You can specify different rules for coding sentiment by combining the concepts used as building blocks for the Sentiment Lens ('favterms', 'unfavterms', 'negationterms') in different ways if you wish.

Finally, once you have created a compound concept you may edit it. Do so by selecting a compound, and then clicking 'Edit':



This allows you to rename your new compound concept, specify whether it is a name or a regular word, and whether or not it should be treated as sentiment:

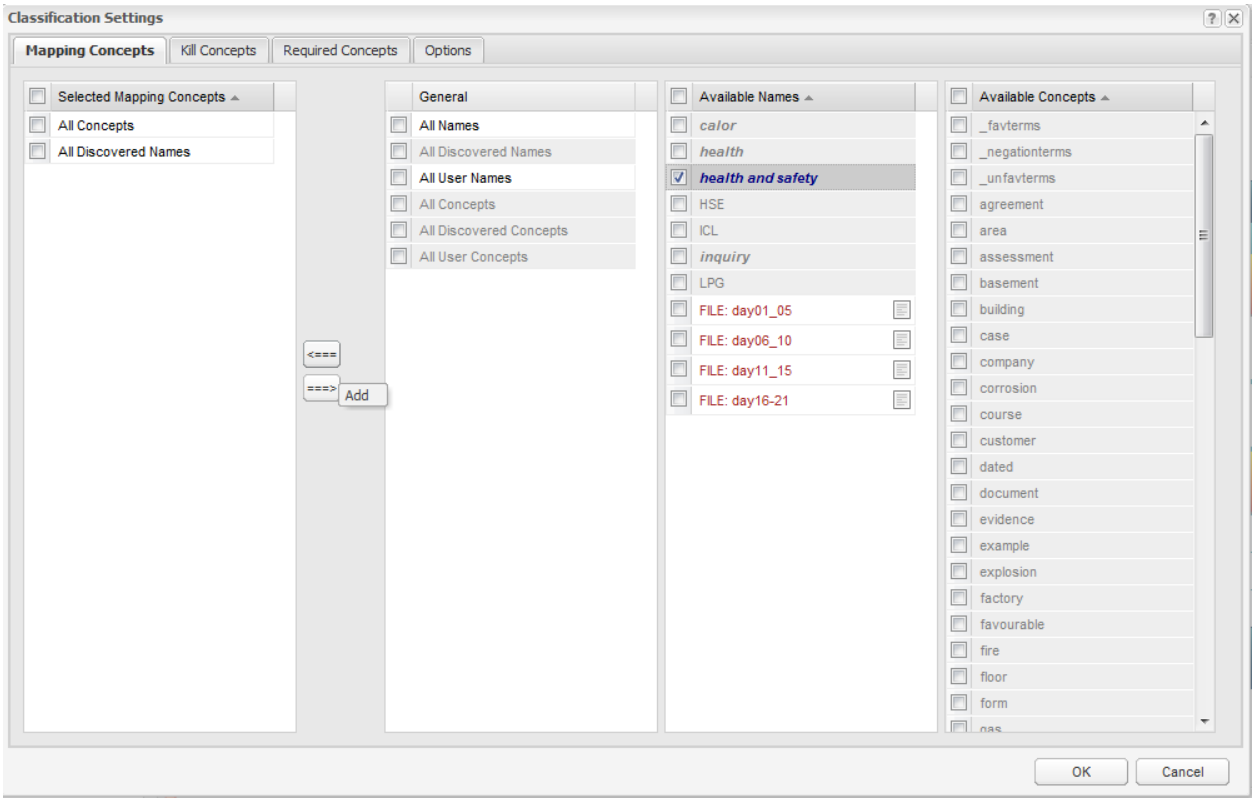


(NOTE: Sentiment terms will not appear on the map, but will be present in the report tabs to the right of the map).

Compound concepts will automatically appear on the map if they are regular word concepts. Name-like compounds need to be specifically added to the Mapping Concepts list in the Concept

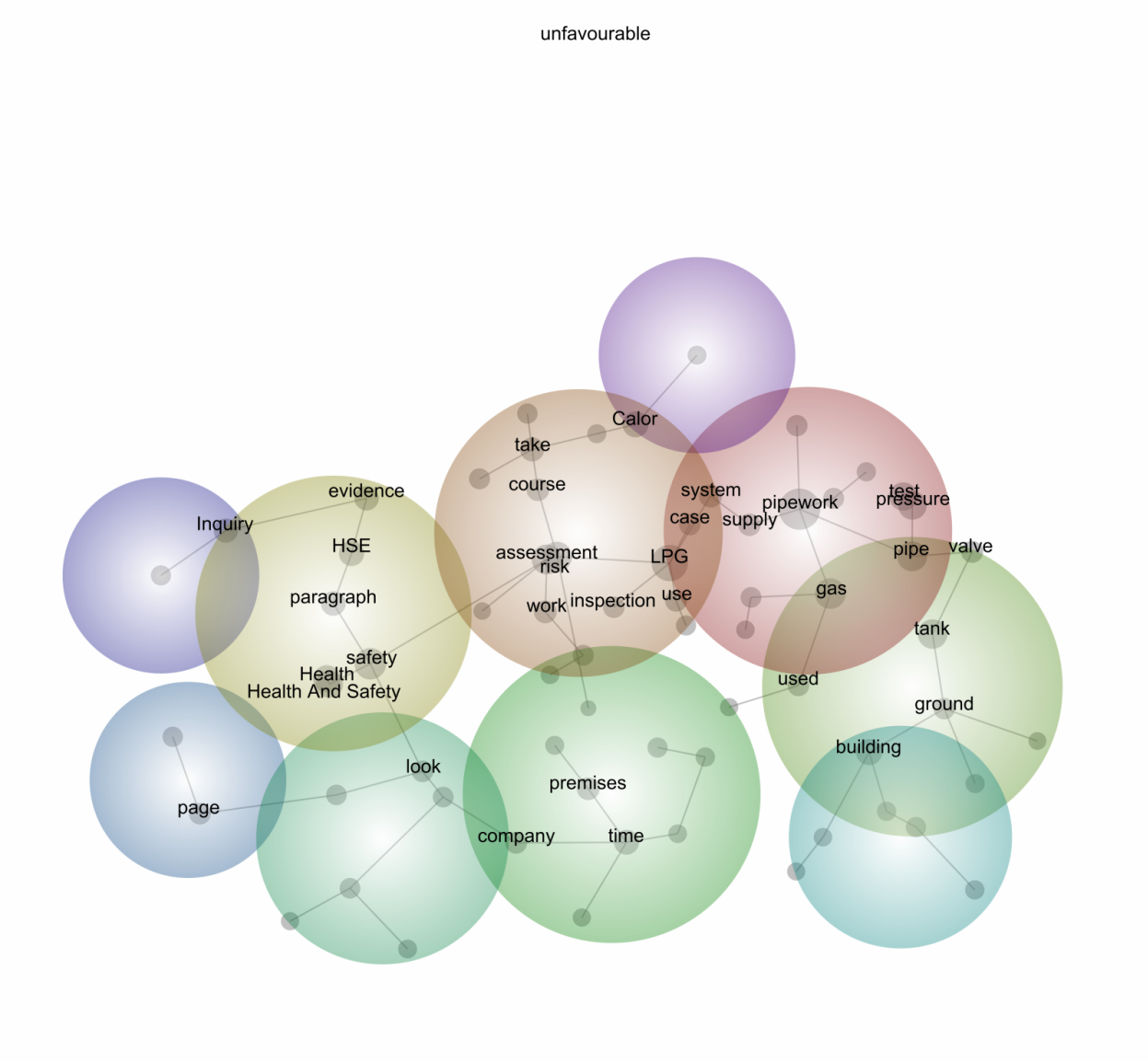
Coding Settings.

After Clicking Ok to leave the Compound Concepts interface, click the Concept Coding Settings node to open this interface:



The left-most tab is called Mapping Concepts. In the Mapping Concepts tab, select the compound concept you wish to add to the map in the Available Words column on the right, and move it to the left hand column using the left arrow.

Now click Ok, and then click on Generate Concept Map. When you open the concept map, your newly created compound concept will appear:



4.8 4b. Concept Coding Settings

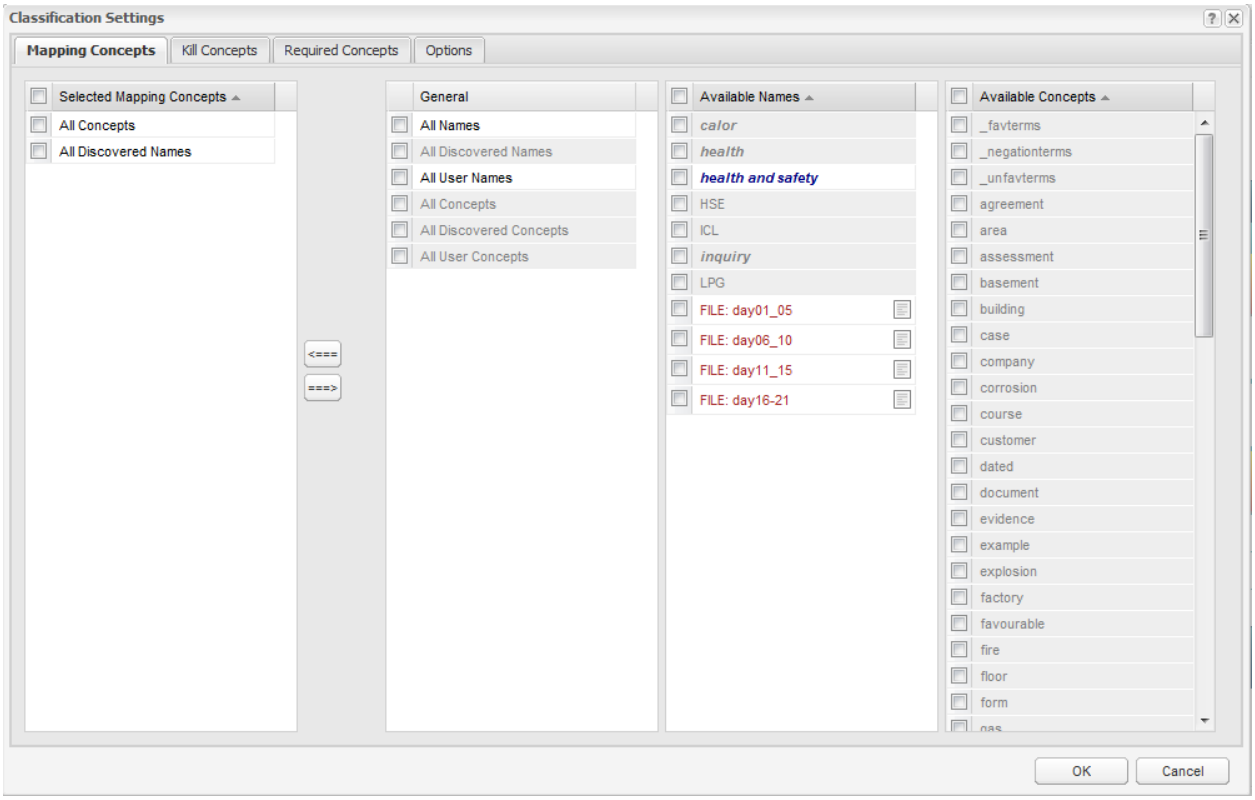
If you open the *Concept Coding Settings* dialogue, you will find four tabs, as follows:

4.9 Mapping Concepts

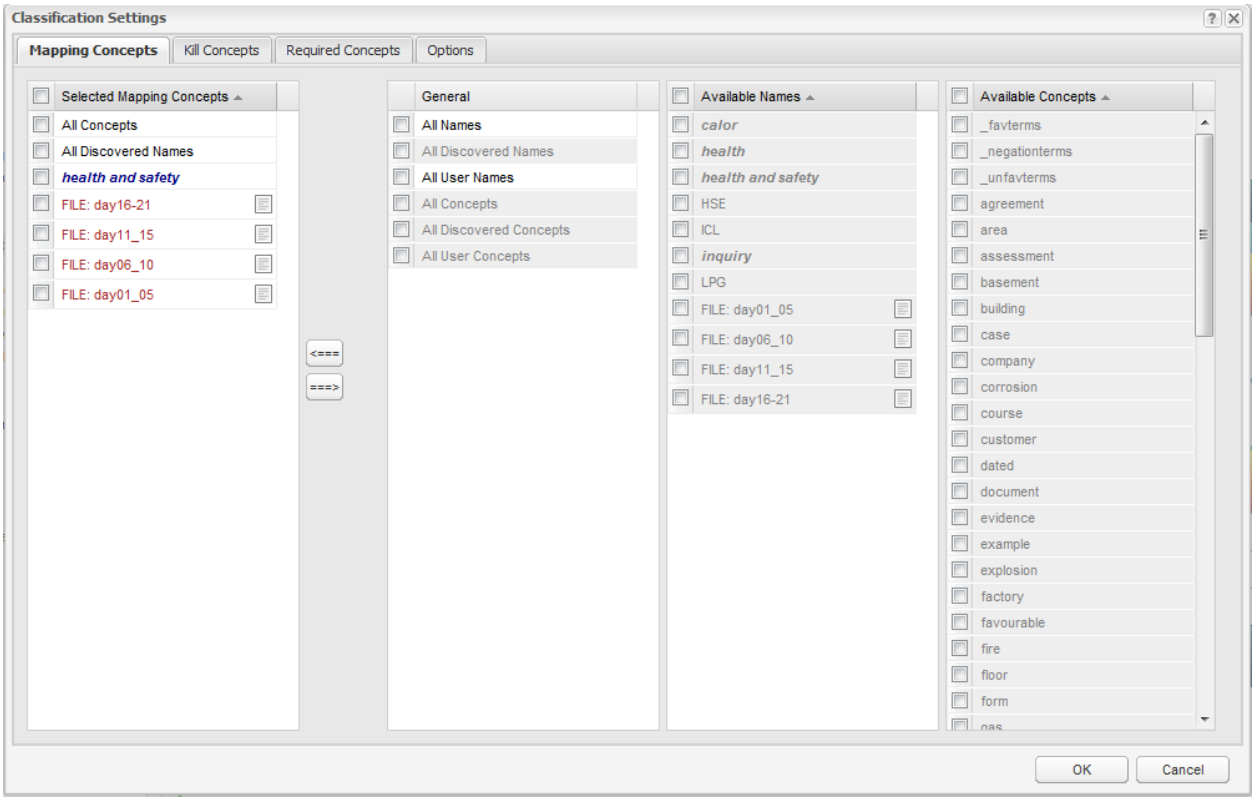
Mapping Concepts are the concepts that appear on the conceptual map, and represent the top-level of classification of the text. Generally all concepts can be used as mapping concepts, but there are some cases in which you may want to map only a subset of the concepts. For example, you could map only the concepts discovered automatically by Leximancer in once instance, and then only the names seeded by you the user in another. You can also choose to map specific names and concepts from the lists provided if you wish.

Configuring the Mapping Concepts:

You use the Mapping Concepts tab to specify which of the available Names, Concepts, and Tags you wish to include on your map. Clicking on *Concept Coding Settings* phase automatically opens the Mapping Concepts tab:



By default, All Concepts and All Discovered Names appear in the Mapping Concepts tab. This means that all word-like and name-like concepts discovered by Leximancer will appear on the concept map. Tags, compound concepts and name-like user-defined concepts must be added to the list manually. Using the arrows to replace these wildcards with others from the General list allows you map other groups of concepts:



Discovered names and concepts refer to those automatically-identified by the program, and User names and concepts refer to those created by you the user. Tags are treated as User-defined.

Instead of using the option in the General list, you can choose to map particular names (including tags) and concepts using the lists on the right. Simply use the arrow buttons to add or delete concepts from each of the lists.

This interface also allows you to filter records in and out of the analysis by specifying Kill Concepts and Required Concepts. Simply click on the appropriate tab and move the desired concept(s) or tag(s) across using the arrow buttons.

4.10 Kill Concepts and Required Concepts

The second and third tabs in the *Concept Coding Settings* dialogue are for filtering text data in or out of your analysis.

Kill concepts are concepts that if found in a classified block of text, cause all other classifications of that block to be suppressed. For example, you could use this option to suppress the processing of questions asked by an interviewer to focus on the responses of the interviewee. If you identified the dialogue spoken by the interviewer using the correct form (speaker name starting with a capital letter on a new line, and followed with a colon and a space, eg: Alan:), Leximancer has a setting in the Preprocessing Options called Apply Dialogue Tags which will automatically identify the dialogue tags. These will be presented to you in the Auto Tags tab in the Concept Seed Editor.

You can then set the interviewer dialogue tag as a Kill class to suppress the processing text spoken by the interviewer.

Required concepts by contrast, are classifications that must be found in blocks of text, or else the blocks are ignored. That is, at least one of the required classifications must be found in any context block for it to be included in the concept map.

4.11 Options Tab

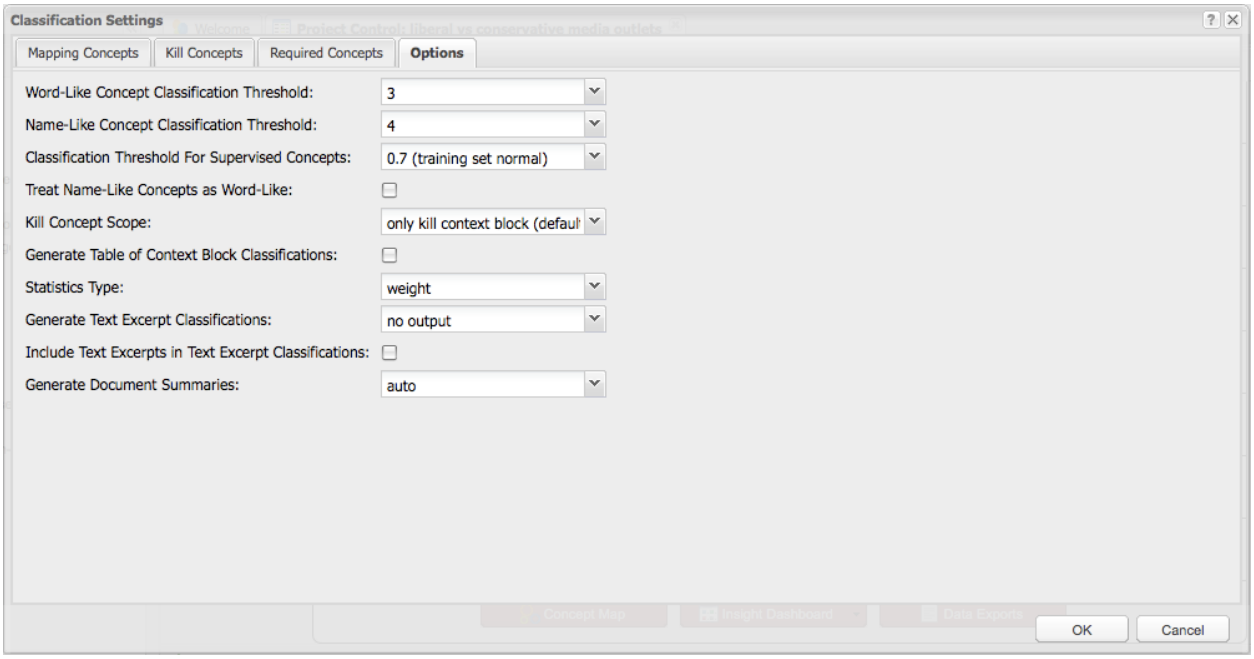
The fourth tab in the *Concept Coding Settings* dialogue is for changing the classifier/coder behaviour, and also for enabling two of the data export options.

Once the concept definitions have been learned, each block of text is tagged with the names of the concepts that it contains. This process is similar to manual coding used in 'content analysis'. However, the benefit of using Leximancer for this task is that it is fast (compared to the time taken for human coders), and is more objective (as opposed to humans, where there is much variability in coding performance).

In Classifying the text document, the following steps are taken:

- The text is broken into context blocks of n sentences
- For name-like concepts, all the associated terms that are present in the block are noted. The block is said to contain the concept if the word with the highest relevancy to the concept is above a set threshold.
- For word-like concepts, the relevancies of all the associated keywords that are present in the block are summed. The block is said to contain the concept if this sum is greater than a predefined threshold.

Clicking on the **Options** tab opens the following dialogue:



Word Classification Threshold (2-4.9): For word-like concepts, the relevancies of all the associated keywords that are present in the block are summed. The block is said to contain the concept if this sum is greater than a predefined threshold. This threshold specifies how much cumulative evidence per sentence is needed for a word concept classification to be assigned to a context block.

Commentary: Note that the actual threshold used is this value multiplied by the average number of sentences per context block found in the data, not the number of sentences found in any particular segment of text. This means that a fixed threshold is applied for all context blocks. It may seem that the actual threshold should be calculated from the number of sentences found, but this would mean that less evidence would be required to trigger a classification in some places than others. After all, one sentence is less evidence than three. The units of this threshold are the same as the relevancy standard deviation values shown for terms in the thesaurus, so you can get a feeling for how much cumulative evidence you need by looking at the learned thesaurus which can be viewed through the map interface.

Name Classification Threshold (2.6-5): For name-like concepts, all the associated terms that are present in the block are noted. The block is said to contain the concept if the word with the highest relevancy to the concept is above the threshold value specified here. That is, this value gives the minimum strength of the maximally weighted piece of evidence to trigger a name classification.

Commentary: The idea behind this threshold is that one strong piece of evidence is enough to indicate the presence of a named thing such as company. For example, the name of the CEO or the stock exchange code would be enough to indicate the company is involved. However, lots of weak evidence for the company is not sufficient, as it could be describing the same area of the market but actually a competitor company. This is related to the notion that a named thing is an instance of a class, rather than a class. If you want cumulative tagging of named things based on a similar lexical context, use the Treat Names as Words option. Like the Word classification threshold, the units of this threshold are relevancy standard deviation units as shown in the thesaurus.

The next step in processing is to measure the co-occurrence of the identified concepts in the text (indexing) for the generation of the conceptual map. The positions of the discovered concepts and groups of important co-occurring concepts are also noted for use with the query browser.

Classification Threshold for Supervised Classifiers (0.7-1.4): Supervised Concepts are used to find instances of a particular concept in the text. Generally, such concepts are 'trained' by tagging exemplars with a code word (such as 'violence'). They are trained to be sensitive to the vocabulary surrounding such tags, without including the tag itself within the definition. Thus, generally there is weaker evidence compared to normal concepts such as 'dog' in which the key word is present in the script. For this reason, supervised concepts require a lower classification threshold. This threshold specifies how much cumulative evidence per sentence is needed for a supervised classification to be assigned to a context block.

Treat Name-Like Concepts as Word-Like (Yes|No): This setting forces name-like concepts to be classified using the same system as word-like concepts, allowing more intuitive coding. This option allows tagging of named things based on similar lexical context, rather that similar identity.

Kill Concept Scope (Kill Whole Document|Only Kill Context Block): This option lets you choose whether to suppress the classification of the entire document should a killed class be present, or just the context block in which the kill class is located.

Generate Table of Text Segment Classifications (Yes|No): If this option is enabled, a delimited text output file listing the concepts tagged in each section of text through the data file is created. This output file, called Table of Text Segment Classifications, is accessible under the Export tab once processing is complete.

Statistics Type (Count|Weight): This setting affects the type of statistics produced in the High Resolution Data Output. The output file lists the concepts tagged in each section of text through the data. If the Generate Table of Text Segment Classifications option is enabled, this file is created and accessible under the Export tab. Normally an assigned tag is simply counted. You can change this setting so that tags are assigned a confidence weight. This results in weighted sums rather than count statistics.

Generate Document Section Classification (No Output|Document as Vector|Document as Matrix): Some advanced applications require classification metadata for each document.

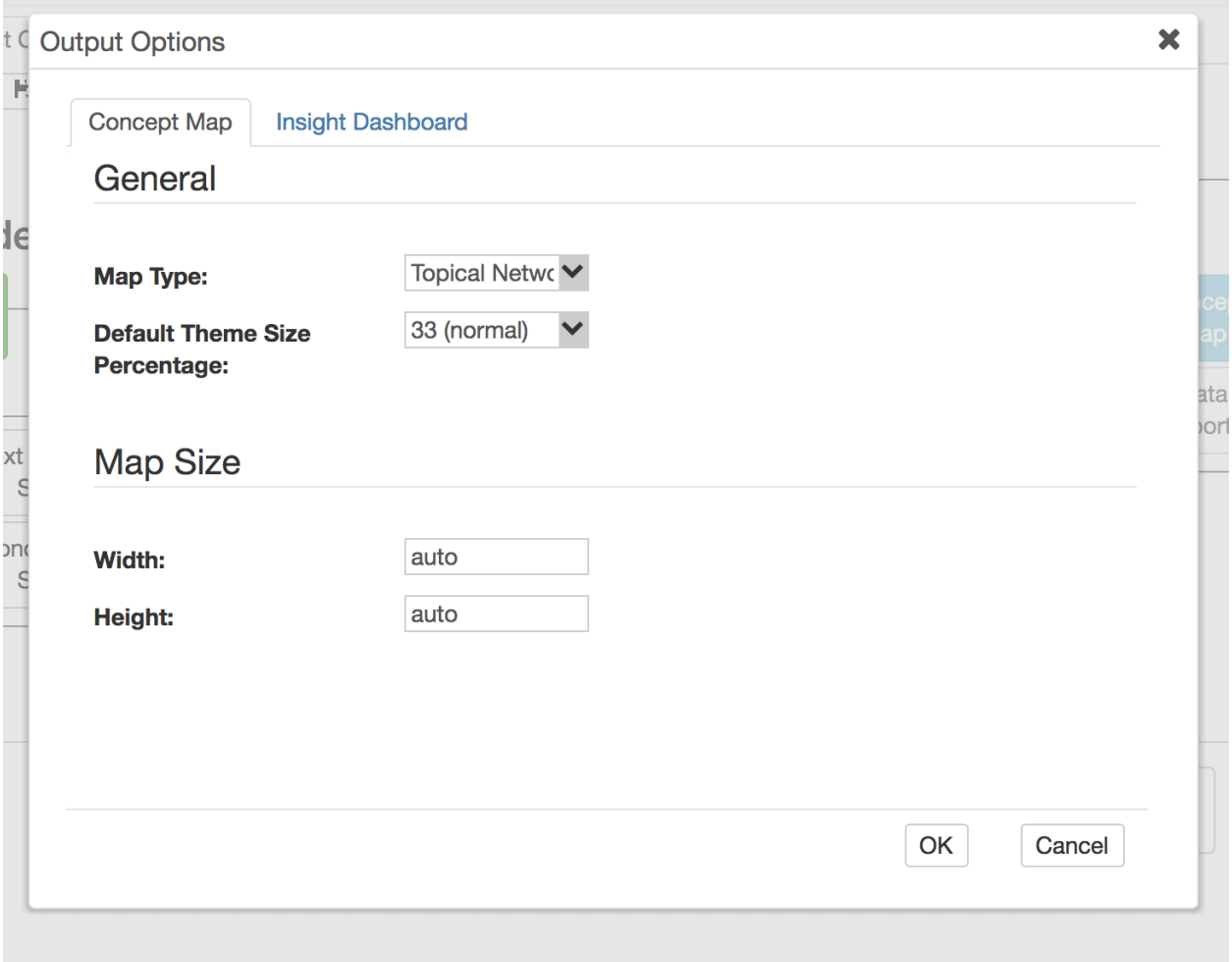
If Document Metadata is set to Document as Vector, an XML file is created that lists the concepts tagged in each document or document section. This output enables analysis of the conceptual content of the data by document section.

If Document Metadata is set to Document as Matrix, an XML file is produced that reports the matrix of concept co-occurrences within each document or document section. Matrix classification enables more accurate document comparison.

If either of these outputs is created, it is called Document Section Classification, and is accessible under the Data Exports button below the main Project Control Panel, or via the Exports tab in the concept map interface.

4.12 4c. Project Outputs

The last settings that can be edited through the main interface are Project Outputs. In this phase, the map displaying the relationship between variables is constructed. Clicking the *Project Output Settings* opens the following interface:



4.13 Generating the Concept Map

One of the principal aims of Leximancer is to quantify the relationships between concepts (i.e. the co-occurrence of concepts), and to represent this information in a useful manner (in a concept map) that can be used for exploring the content of the documents. The concept map can be thought of as a bird's eye view of the data, illustrating the main features (i.e. concepts) and how they interrelate.

The mapping phase generates a two dimensional projection of the original high dimensional co-occurrence matrix between the concepts. It must be emphasised that the process of generating this map is stochastic. Concepts on the map may settle in different positions with each generation of a new map.

In understanding this, consider that concepts are initially scattered randomly throughout the map space. If you imagine the space of possible map arrangements as a hilly table top, and you throw a marble from a random place on the edge, the marble could settle in different valleys depending on where it starts. There may be multiple 'shallow valleys' (local minima) in the map terrain if words are used ambiguously and the data is semantically confused. In this case the data should not form a stable pattern anyway. Another possibility is that some concepts in the data should in fact be stop words, but aren't in the list. An example of this is the emergence of the concept 'think' in interview transcripts. This concept is often bleached of semantic meaning and used by convention only. The technical result of the presence of highly-connected and indiscriminate concept nodes is that the map to loses differentiation and stability. The over-connected concept resembles a mountain which negates the existence of all the valleys in the terrain. To fix this, remove the over-connected concept.

The practical implication is that for a strict interpretation of the cluster map, the clustering should be run several times from scratch and the map inspected on each occasion. If the relative positioning of the concepts is similar between runs, then the cluster map is likely to be representative. Note that rotations and reflections are permitted variations. If the map changes in gross structure, then revision of some of the parameters is required. The concept map display supplied with Leximancer allows easy re-clustering and map comparison using the buttons above the map.

4.14 Topical versus Social Mapping

In concept mapping, there is one setting to choose, namely whether to use a Topical or Social map. The Social map has a more circular symmetry and emphasises the similarity between the conceptual contexts in which the words appear. A Social map is best when entities tend to be related to fewer other entities, such as a map made up of many name concepts.

The Topical map, by comparison, is more spread out, emphasising the co-occurrence between items. It tends to emphasise differences and direct relationships, and is best for discriminant analysis. The Topical map is also much more stable for highly connected entities, such as topics. The most common reason for cluster instability is that the concepts on the map are too highly connected, and no strong pattern can be found. The Topical variant of the clustering algorithm produces more stability in maps of this kind, so switching to this setting will often stabilise the map. However, the most important settings which govern the connectedness of the map are the classification thresholds and the size of the coded context block. If the coded context block is too large, or the classification threshold is too low, then each concept will tend to be related to every other concept. If you have some highly-connected concepts which are effectively bleached of meaning in your data, removing from the concept lists in the Concept Seeds Editor will often stabilise the map. Words such as 'sort', 'think', and 'kind' often appear in spoken transcripts and may be used as filler words which are essentially stopwords. Inspect the actual text locations to check the way words like these are being used before removing them.

In summary, the Topical clustering algorithm is more stable than the Social, but will discover fewer indirect relationships. The cluster map should be considered as indicative and should be used for

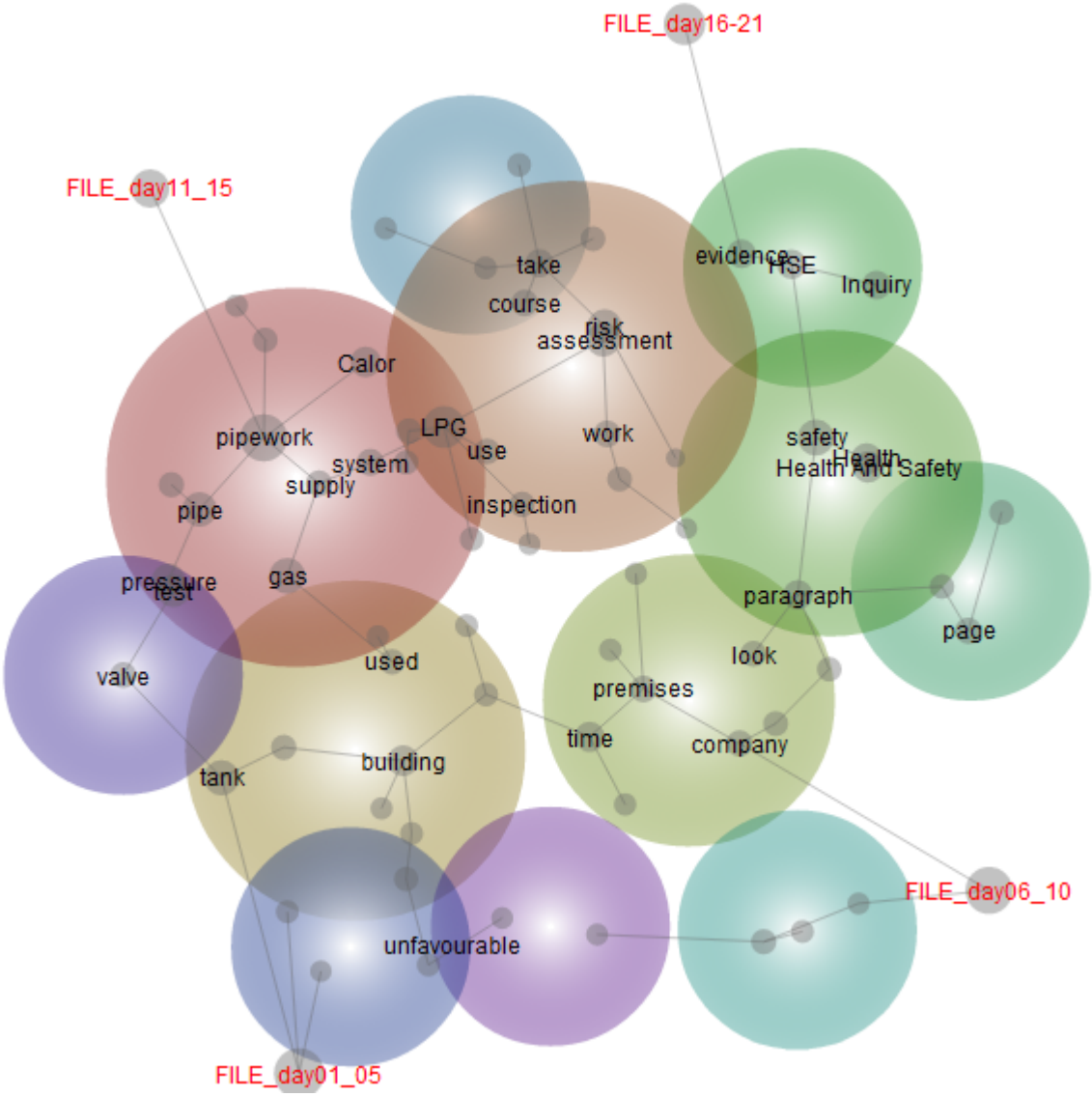
generating hypotheses for confirmation in the text data. It is not a quantitative statement of fact.

Default Theme Size Percentage (10-65): This option sets the size of themes (concept groupings) visible when the map interface opens initially. The theme size slider beneath the concept map also allows this parameter to be adjusted through the map interface.

You can also specify the Size (Width and Height) of the concept map to be produced in this interface.

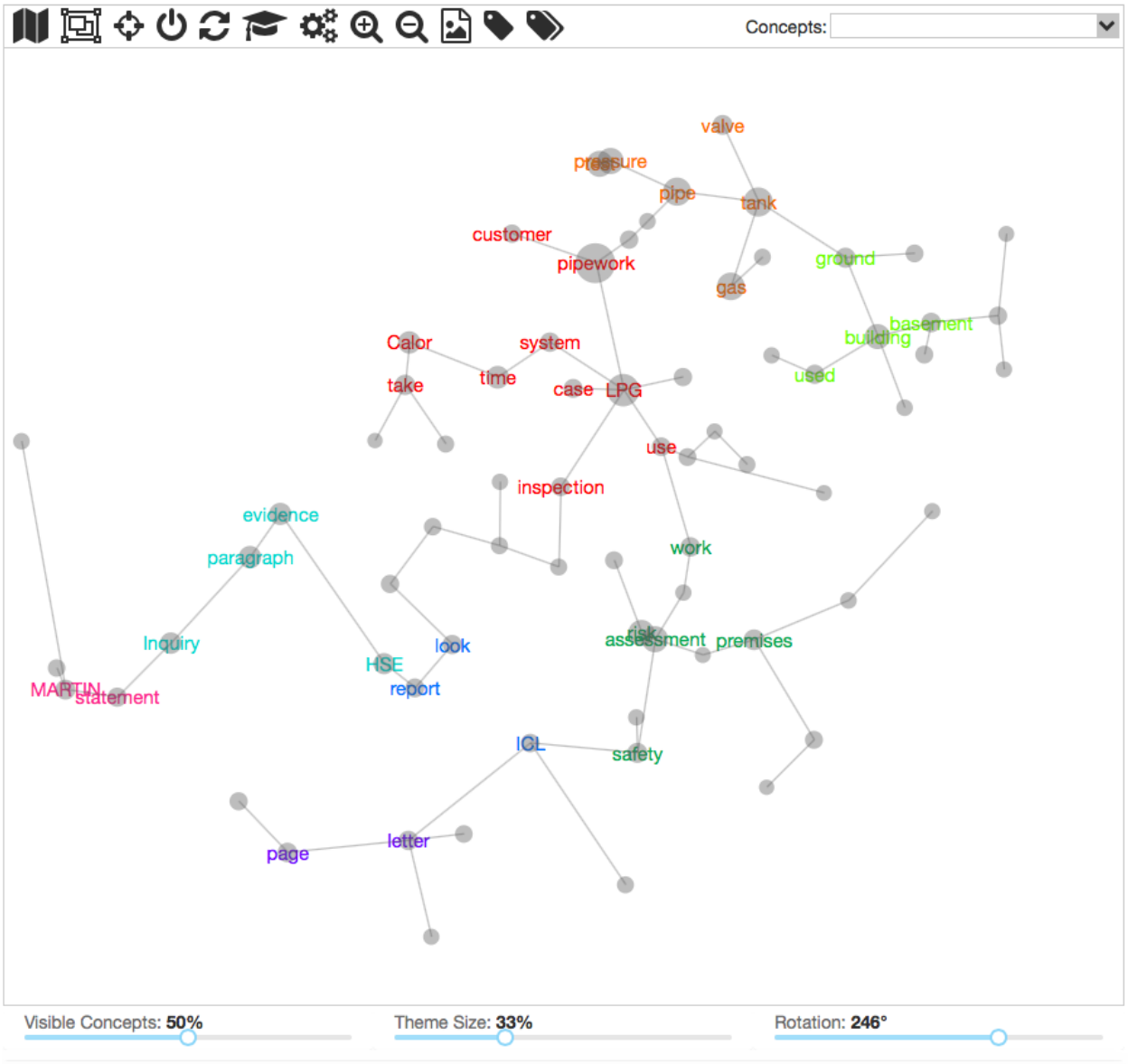
4.15 Final Outputs

1. The Concept Map



See page 8 for an explanation of the Concept Map.

2. The Concept Cloud



See page 15 for an explanation of the Concept Cloud.

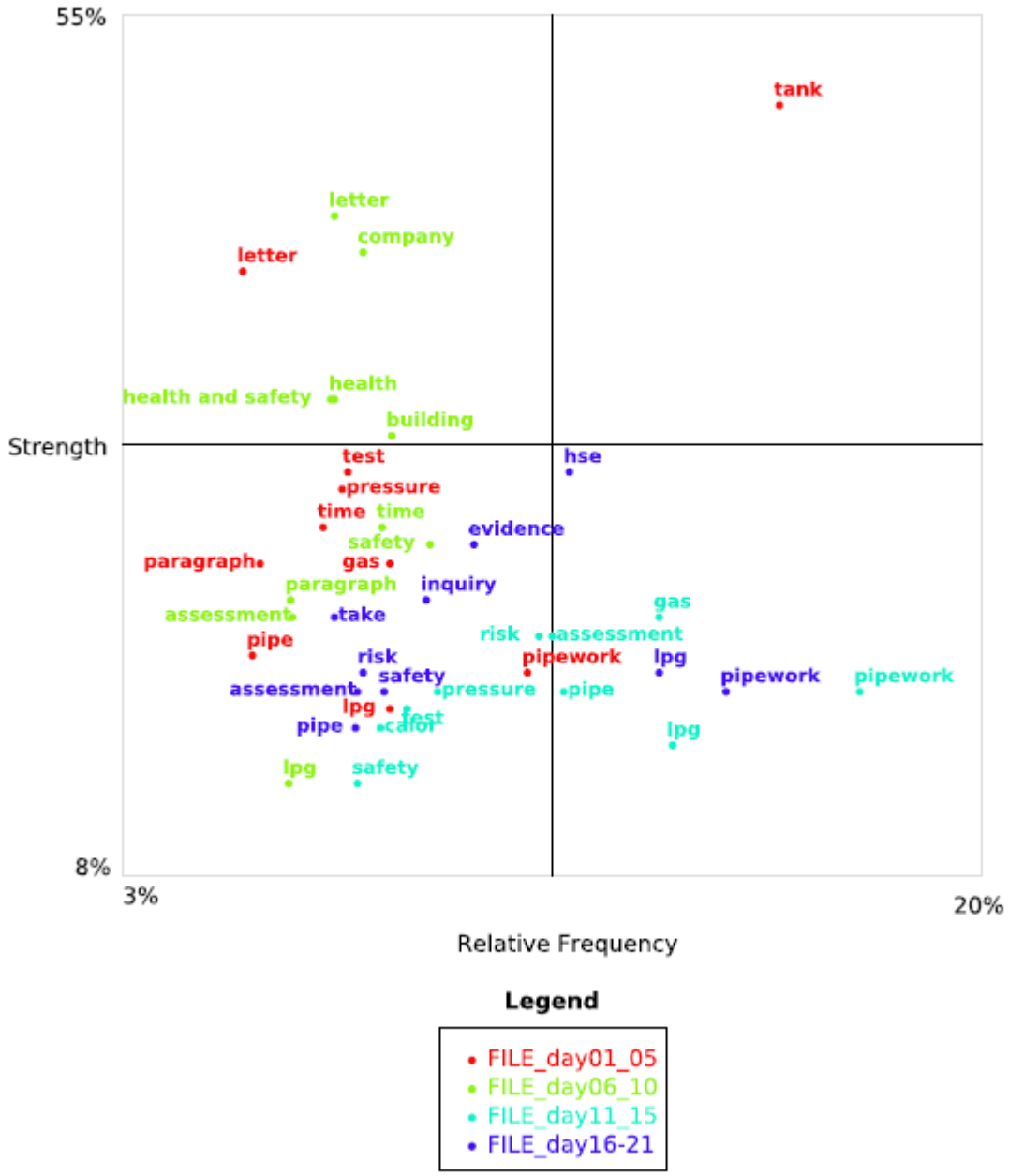
3. The Insight Dashboard Report

The Insight Dashboard is a standalone pdf or html report that can be produced within a Leximancer project. The Dashboard report adopts a more quantitative focus than the concept map, and is designed to provide a quick understanding of the project results.

The report is not set up to be created automatically when you run a project, as not all projects are suitable for the dashboard report format. The Dashboard is designed for some sort of comparison, or difference analysis. Usually you must have created some (source document, tabular, speaker or folder) tags to use as categories for comparison in the Dashboard report. You can then use the Report to investigate the Attributes (independent variables) associated with certain Categories

(dependent variables) in the data.

In this example, File tags were aggregated to create a category for each week of hearing transcripts after an explosion at a plastics factory. The Dashboard quadrant below shows the concepts most relevant to the discussion in each week (trends). The concepts towards the upper, right-hand (magic) quadrant were talked about most frequently in each week, and are most characteristic of the discourse during that time:

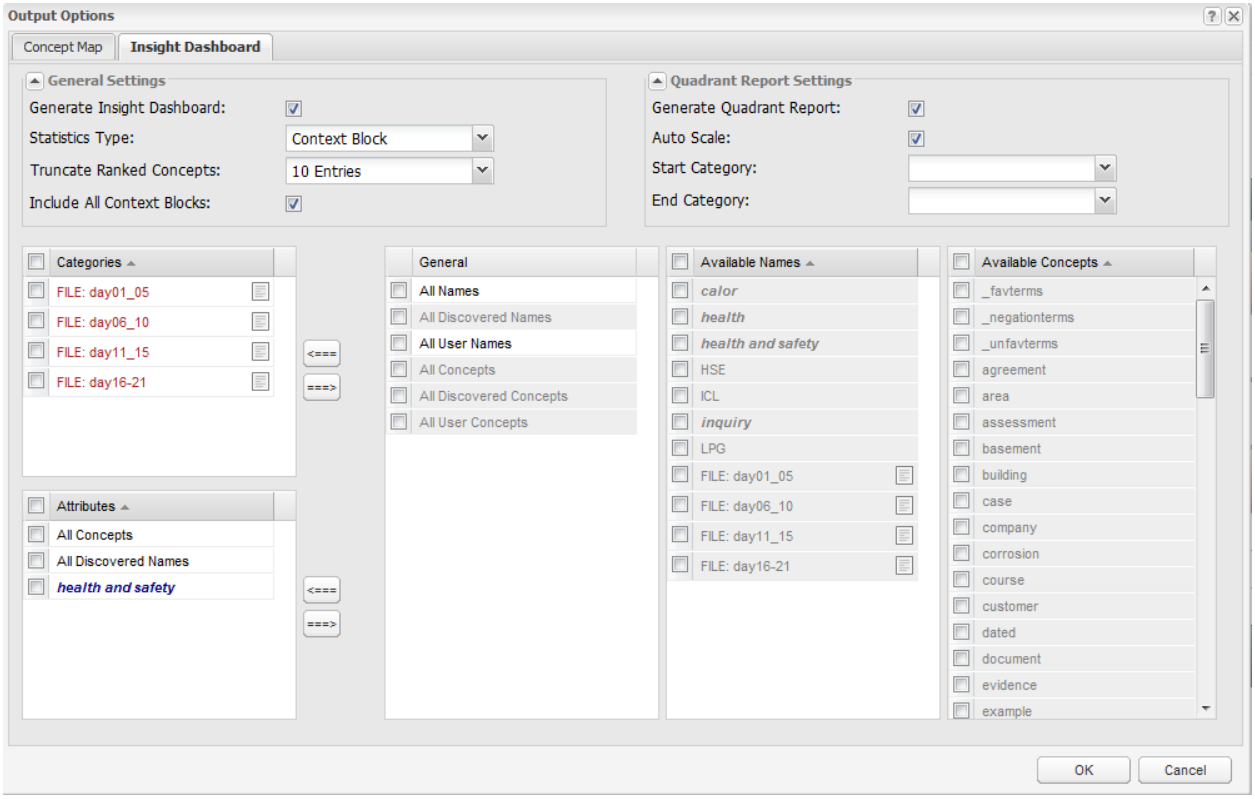


The Dashboard is produced as a standalone document so that it can be shared with others. Once created, it can be Downloaded from the Project Control header, or from the Data Exports beneath the main Project Control Panel. It can then be printed or sent to colleagues, collaborators and clients. The report does not require any knowledge of Leximancer to be understood.

The Dashboard is intended as a general purpose reporting tool, and the user can configure the options to suit their particular application or research question.

4.16 Configuring the Insight Dashboard Report

To configure a Dashboard Report, click on the Project Output Settings node. Click on the Insight Dashboard tab to reveal the following interface:



Tick the Generate Insight Dashboard option.

Tick the Generate Quadrant Report option. This includes a 'magic quadrant' graphic in the Report for identification of key issues.

Set the Statistics Type setting (Segment|Segment)

Using the Segment Statistics setting, concepts are coded and classified at the level of the text segment. You can define a text segment (the coding resolution) using the Sentences per Block setting in the Pre-processing Options.

Using the Section Statistics setting, concept codes are applied to sections of the data. The definition of a document 'section' depends on the type of data:

- If you are processing Microsoft Word or pdf documents, then a document section is usually an individual file.

- If you are processing delimited spreadsheet data, however, then a section is a single free text cell. This setting can be used to report the number of responses (whole comments) coded with particular concepts.

The Truncate Ranked Concepts setting allows you to specify the number of Attributes reported for each Category. You can control the level of detail in the Report using this setting. The Do Not Truncate default does not apply any cut-off, but presents all the Attributes associated with each Category.

The Auto Scale setting scales the axes in the Quadrant graphic. This spreads the Attributes in the Quadrant space for improved visibility.

- If this option is enabled, the Prominence statistics (described in the preamble to the Report) are expressed as relative probabilities.

The Include All Text Segments setting introduces an optional final section to the Report that presents all text segments matching each concept query.

Specify the Categories (or dependent variables) of interest. You can specify any number of Categories, though using upwards of 10 can clutter the Quadrant graphic.

Often the Categories will be (auto- or user-defined) Tags identifying different levels of variables, or groups for comparison, within the data. Gender (male or female) and tone (favourable or unfavourable) are examples of possible Categories.

- Select the Tags (or concepts) of interest from the Available Names (or Concepts) list(s), and use the appropriate left arrow to add these to the Categories list.

Note: If you wish to use Tags as Categories, you must add these to the Mapping Concepts list by hand in the Select Concept to Locate phase. This codes the data with the Tags so that they can be used in the Dashboard Report. It will also cause them to clustered on the map among the topical concepts.

Specify some Attributes (or independent variables) of interest.

You might use the emergent concepts as Attributes in the Dashboard Report. In this case, the Report will compare the concepts associated with each of your Categories

- Select the All Concepts wild card from the General list, then click the Attributes left arrow to use all the word-like concepts as Attributes. This wildcard includes all the entries in the Available Concepts list on the right. Alternatively, you can select individual concepts from the list and add them as Attributes by hand.
- Select the All Discovered Names wild card from the General list, then click the Attributes left arrow to use all the word-like concepts as Attributes. This wildcard includes all the entries in the Available Names list in the centre. Alternatively, you can select individual names from the list and add them as Attributes by hand.

When the settings have been configured, click Ok and run the final stages of processing to produce the Dashboard Report.

After clicking the Run Project button, you can download the Dashboard Report from the button beneath the main Project Control Panel.

The Dashboard can be downloaded in pdf of html format.

- The pdf version can be viewed in Adobe Acrobat Reader. The Quadrant graphic requires at least Acrobat version 9 to be displayed.
- The html version is useful if you wish to make edits to the Report. This version can be saved as a zipped folder (or archive) on your local machine. You may need to rename the folder (changing the extension from insight-dashboard-zip to insight-dashboard.zip) to allow it to be extracted or opened. Click on the insight-dashboard.html file to view the Report in a browser tab, or right click and select Open With to view the Report in another application (such as Microsoft Word).

The Dashboard is named after the project in which it was created. The header provides counts of the Total text Segments or Sections coded in the Report. It also presents counts for the number of Concepts and Categories specified.

Note: The preamble explains the various sections of the Report, and is included as part of Dashboard to allow others to understand its contents.

The Dashboard is designed to be interactive, and the Table of Contents includes clickable links, as do most many other sections of the Report. In the html version, you may have to hold down the <control> key and click to navigate the Report links.

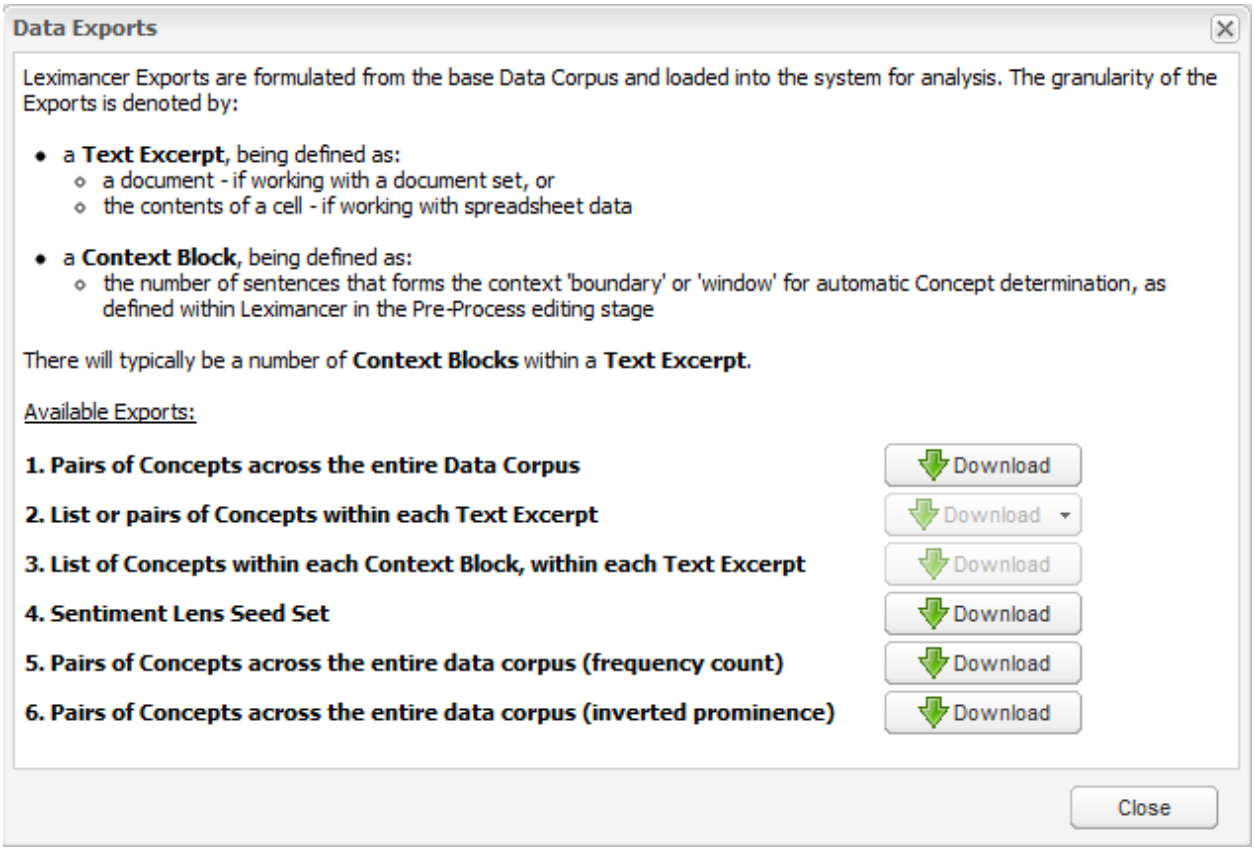
Explanation of the Statistics in the Insight Dashboard Report

- The Frequency axis on the Quadrant graphic represents a conditional probability. Given that a text extract comes from a particular Category, it gives the chance that the Attribute is coded in this text extract. This measures frequency of mention in the data, and is affected by the distribution of comments across the Categories. The frequency score is in fact a log scale (so that it can be mapped on the quadrant).
- The Strength score is the reciprocal conditional probability. Given that the Attribute is present in a section of text, it gives the probability that this text comes from that Category. Strong concepts distinguish the Category from others, whether or not the Attribute is mentioned frequently.
- The percentages in the Ranked Concept for Categories lists match the quadrant coordinates. They reflect the same Strength and Frequency conditional probabilities.
- The Prominence score combines the Strength and Frequency scores using Bayesian statistics. Prominence scores are absolute measures of correlation between category and attribute, and can be used to make comparisons over time.

4.17 Data Exports

Among the projects results, Leximancer produces several statistical reports. These can be exported for reporting or to allow further analysis in other applications.

Several reports are available for Download from the Data Exports button beneath the Project Control Panel. These include: (1) the Pairs of Concepts across the Entire Corpus (the co-occurrence matrix); (2) the List or Pairs of Concepts within each Text Excerpt; (3) the List of Concepts within each Context Block, within each Text Excerpt; and (4) The Sentiment Lens Seeds Set:



Hover your mouse over the Download button for a description of the level of detail in the each of the reports.

Leximancer Exports are formulated from the base Data Corpus and loaded into the system for analysis. The granularity of the Exports is denoted by:

- a Text Excerpt, being defined as:
  - a document - if working with a document set, or
  - the contents of a text cell - if working with spreadsheet data
- a Context Block, being defined as:
  - the number of sentences that forms the context 'boundary' or 'window' for automatic

Concept determination, as defined within Leximancer in the Pre-Process editing stage

There will typically be a number of Context Blocks within a Text Excerpt."

Available exports include:

Pairs of Concepts across the entire Data Corpus (the co-occurrence matrix)

Downloads a comma delimited file displaying the matrix of co-occurrences between concepts. This file will open a spreadsheet program, including recent versions of Excel. It contains co-occurrence counts, listed for every concept pair combination, as well as x,y coordinates for each concept on the map, and the weight for each concept, which is the sum of its co-occurrences with all the other concepts:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	concept	x	y	weight	FILE_day0	pipework	FILE_day0	LPG	FILE_day1	FILE_day1	safety	gas	tank	pipe	assessme	risk	pressure	test
2	FILE_day01_05	0.58475	0.751729	13045	6522	557	0	429	0	0	205	428	900	329	95	87	391	396
3	pipework	0.397195	-0.3131	11772	557	2931	257	604	537	547	128	362	389	596	294	293	375	364
4	FILE_day06_10	-0.62752	0.716412	11721	0	257	5863	317	0	0	416	278	294	201	320	311	141	134
5	LPG	0.09467	-0.2616	9586	429	604	317	2409	473	382	211	406	319	193	197	196	138	136
6	FILE_day16-21	-0.45908	-0.83443	9413	0	537	0	473	4298	0	279	160	121	264	265	268	171	164
7	FILE_day11_15	0.488835	-0.76946	8177	0	547	0	382	0	3390	209	373	199	310	304	296	244	230
8	safety	-0.51635	-0.0999	7244	205	128	416	211	279	209	1546	106	69	25	186	171	35	37
9	gas	0.411764	-0.07581	7030	428	362	278	406	160	373	106	1687	239	220	150	146	193	182
10	tank	0.598727	0.23291	6869	900	389	294	319	121	199	69	239	1798	186	106	105	203	203
11	pipe	0.531192	-0.21937	6593	329	596	201	193	264	310	25	220	186	1633	129	123	295	282
12	assessment	-0.20232	-0.34456	6528	95	294	320	197	265	304	186	150	106	129	1428	1384	89	77
13	risk	-0.20993	-0.36558	6422	87	293	311	196	268	296	171	146	105	123	1384	1405	84	72
14	pressure	0.610098	-0.11463	5948	391	375	141	138	171	244	35	193	203	295	89	84	1326	1247
15	test	0.606003	-0.09522	5799	396	364	134	136	164	230	37	182	203	282	77	72	1247	1285
16	Health	-0.59202	-0.04492	5216	137	55	346	94	151	126	983	54	28	17	141	130	11	10
17	Health And Safety	-0.59118	-0.031	5185	135	55	344	94	150	122	983	53	28	16	139	129	11	10
18	building	0.291737	0.274931	5040	300	246	387	187	124	163	80	165	168	183	65	65	69	60
19	Calor	0.199357	-0.39914	4886	288	376	162	179	233	219	97	98	175	129	90	88	82	82
20	time	-0.02617	0.307556	4637	377	171	379	131	185	130	78	91	109	128	85	84	94	92
21	paragraph	-0.42566	0.153561	4484	334	238	318	160	241	112	86	71	37	113	45	45	122	121
22	HSE	-0.54317	-0.39587	4473	127	160	180	226	398	101	179	70	27	32	87	93	38	38
23	evidence	-0.46439	-0.43264	4153	148	176	169	128	331	152	101	60	56	79	62	60	48	49
24	Inquiry	-0.67567	-0.32863	4116	172	110	162	156	302	181	92	52	33	22	56	56	31	31
25	take	-0.12647	-0.49505	3948	151	190	147	123	254	158	98	100	78	96	96	93	68	69
26	unfavourable	-0.39444	-1.01994	3439	203	177	152	124	165	136	71	101	65	92	76	76	53	45

List or pairs of Concepts within each Text Excerpt

This CSV or XML file contains classifications of text excerpts. Each block (row) indicates:

- the file and section number for the text excerpt;
- the surrogate id for viewing this context block;
- the tabular input data, a field for each quantitative column in the input data
- for tabular input, a field whose value indicates which text column in the row is indicated. This is for situations where there are multiple text cells in a row of input data
- a nested block containing classifications assigned to this text excerpt, with occurrence counts and cumulative weights

This output is designed for import of document or text excerpt classifications into a database. This output file is not sparse.

List of Concepts within each Context Block, within each Text Excerpt

This tab delimited text file contains one row for each context block. Each row indicates:

- **the file, text excerpt, and starting sequence number of the context** block
- **the html surrogate link for viewing this context block in a** browser
- the presence or absence of a concept or tag in the context block.

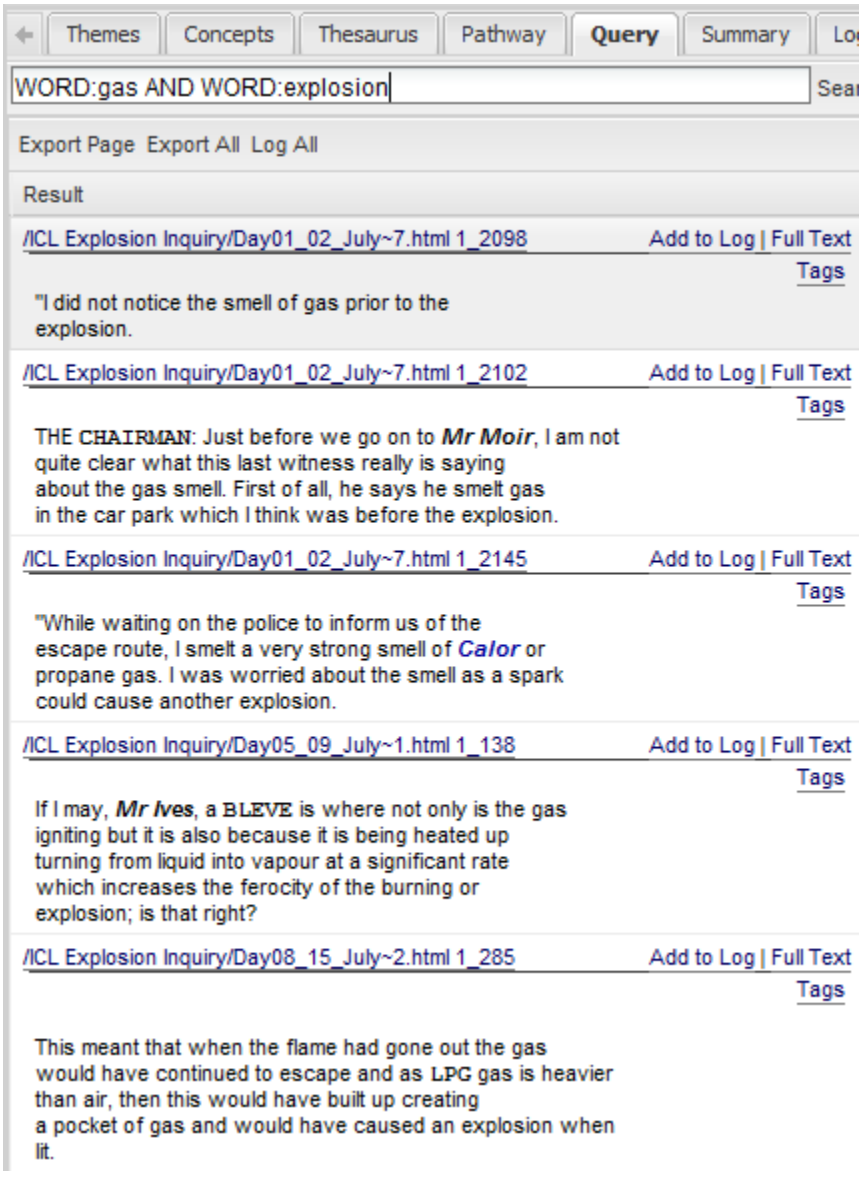
There is one column for each concept or tag class. As a result, this table is very sparse.

This import is specifically designed for input into statistics or data mining packages for building models such as: decision trees, rule sets, logistic regression, or market basket analysis. There is a setting in the Classifications Settings tab to generate either real valued or binary values.

This file can be Uploaded directly into other projects via the Concept Seeds Settings Edit interface.

Logbook Exports

In the map interface, Leximancer allows complex records of queries to be stored and exported. Find a particular query of interest, and its example text, and add it to the logbook:



Go to the 'Logbook' tab to review all the logged text:

From here, you can export either just the current page, or every entry in the logbook:

Your exported logbook entries will pop up in a new window of your browser. Pop-ups need to be enabled for this to occur. Leximancer will have logged the Document ID, Folder/s, Author, Date and Time, Query Terms, and the example text:

Logbook

Matching Log Entries Found: 3

1. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2098
"I did not notice the smell of gas prior to the explosion.

Author: julia
Date: Oct 23, 2011 5:56:47 AM

2. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2145
"While waiting on the police to inform us of the escape route, I smelt a very strong smell of Calor or propane gas. I was worried about the smell as a spark could cause another explosion.

Author: julia
Date: Oct 23, 2011 5:56:50 AM

3. /ICL Explosion Inquiry/Day08_15_July~2.html/1/1_285
This meant that when the flame had gone out the gas would have continued to escape and as LPG gas is heavier than air, then this would have built up creating a pocket of gas and would have caused an explosion when lit.

Author: julia
Date: Oct 23, 2011 5:56:52 AM

CHAPTER

FIVE

EXAMPLE ADVANCED TECHNIQUES

Note: This chapter of the manual has not yet been updated for V4.5. The techniques presented still apply but the layout of the user interface has changed.

This section is intended to be a guide for some of the styles of analysis that are possible with the Leximancer system. However, the system is designed to be a general-purpose tool, and the confident user is encouraged to experiment.

Specific tutorials include: Manual Concept Seeding; Profiling; Extracting a Social Network; Analysing Transcripts; Analysing Spreadsheet Data

5.1 1. Manual Concept Seeding

You can seed your own concepts. Press the Concept Seeds box and change to the User-Defined Concepts tab. If you create new manual seeds, thesaurus definitions will be extracted for these and any automatically-identified concepts during the Generate Thesaurus phase.

You can also create your own Manual Tags in the User Defined Tags tab. These act like lists of keywords, or fixed dictionaries – they are not modified by the learning process.

In many cases, an automatically generated map may contain concepts that are irrelevant to your interests or concepts that are close to each other (such as thought and think), or the map may be lacking concepts that you wish to locate in the text. You can 'seed' your own concepts prior to running the Thesaurus Learning phase by clicking on *Concept Seeds* in the Project Control Panel. There you can add, edit, merge or delete concepts in order to produce cleaner or more tailored concept maps. If you add (User-Defined) concept seeds manually, thesaurus definitions for these will be extracted from the text along with any automatically-extracted concept seeds. You can also create your own User-Defined Tags and these act like lists of keywords, or fixed dictionaries – they are not modified by the learning process.

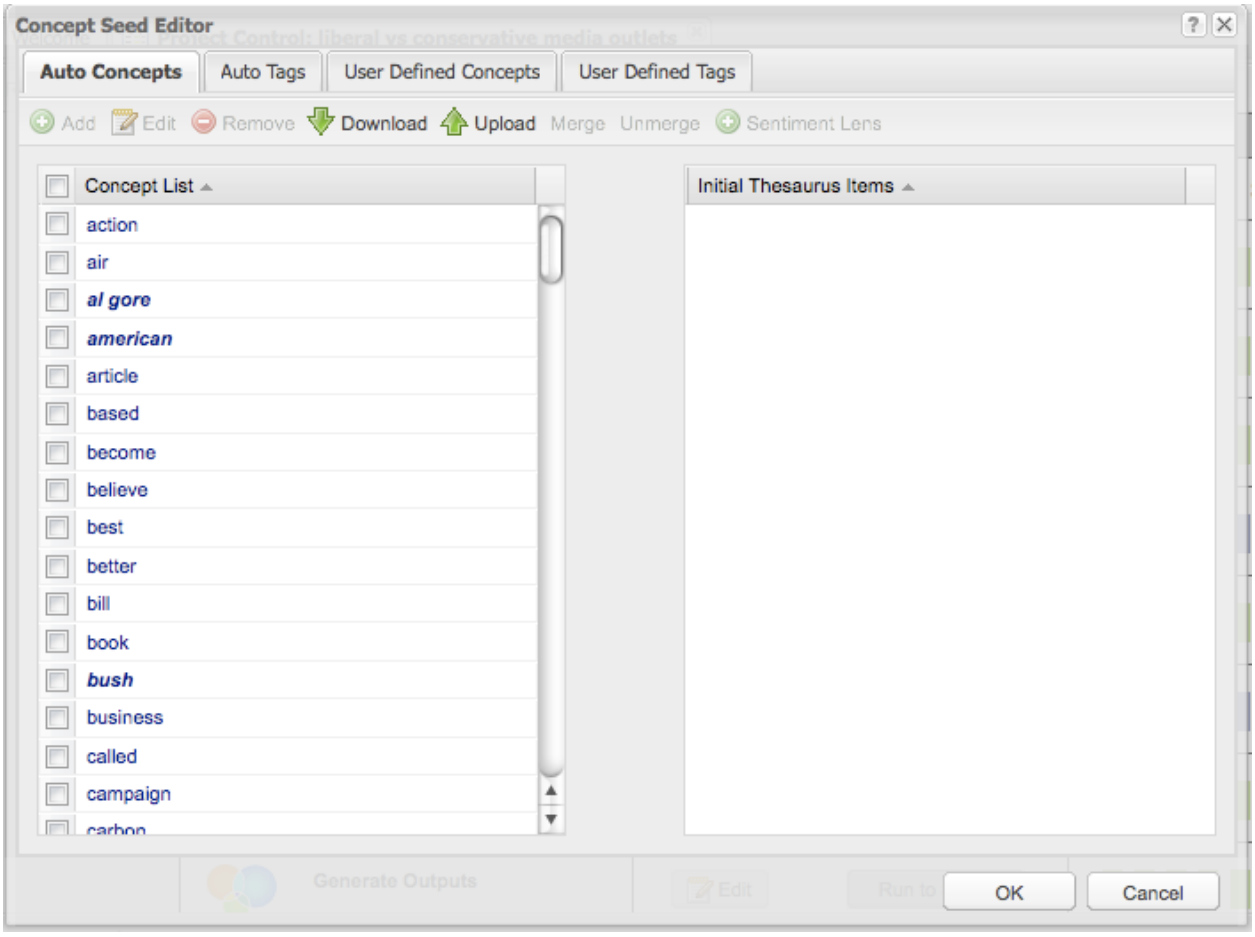
Note: Note that any edits to the Auto Concepts or Auto Tags tabs will be lost if you re-run the

Generate Concept Seeds stage.

5.2 Configuring Manual Concept Seeding

In order to modify concepts automatically extracted by Leximancer, Generate Concept Seeds (the node prior to Generate Thesaurus) needs to have been run. If this node has been run (e.g. if you have previously generated a map), it will be green and it's status will say Ready. If not, click on the Generate Concept Seeds button to run this stage.

After Generate Concepts Seeds has been run, clicking on Concept Seeds reveals the following interface:



There are tabs for editing the Automatic Concepts (concepts identified by Leximancer) and User Defined Concepts (concepts that you wish to define yourself).

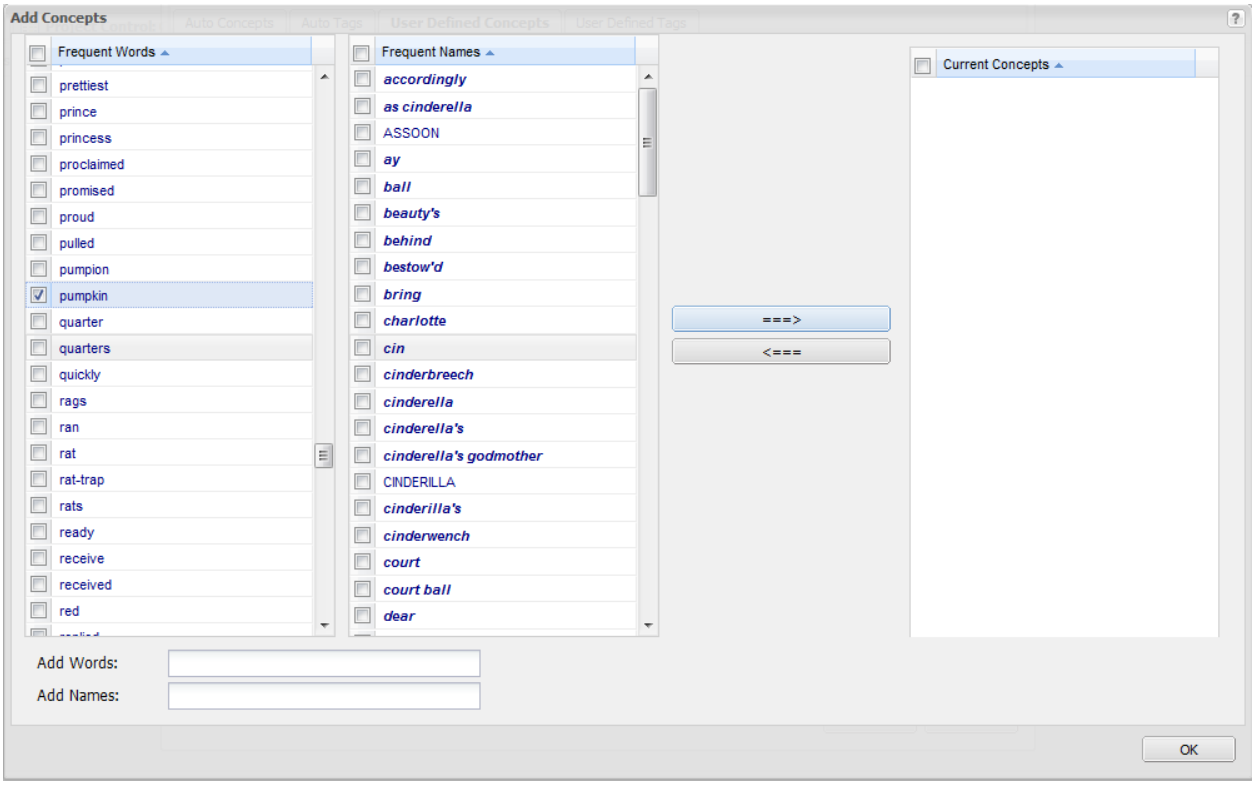
At this stage, only the central key word for each concept has been identified. The learning of associated terms and their weightings occurs in the following Thesaurus Learning phase.

In the interface above, you can select and Merge or Delete concept seeds (note: holding down

<ctrl> while clicking allows you to select multiple items).

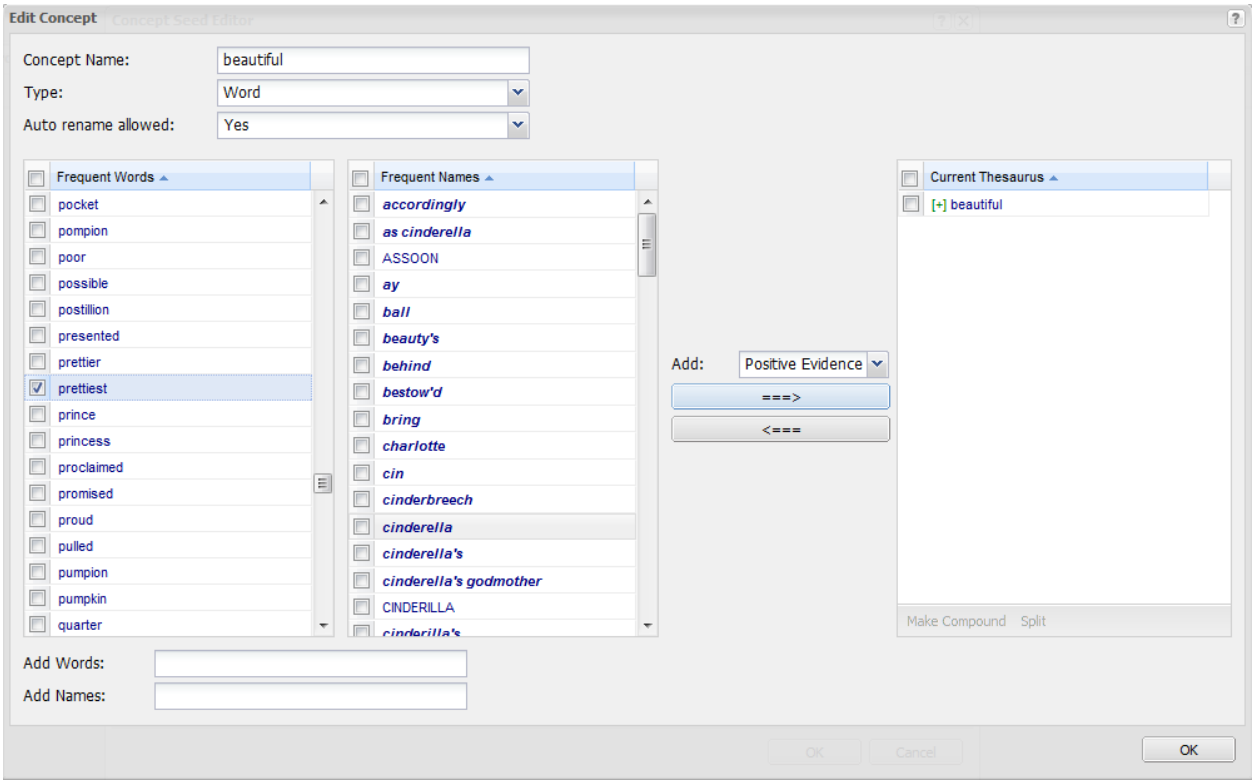
5.3 Adding Concepts

Open the User-Defined Concepts tab and click on Add to open the Add Concepts interface:



This dialogue lets you seed multiple concepts easily, one concept per selected word. You can select terms from the lists of Frequently-occurring Words or Names on the left, or type in your own seed words or names for the new concepts using the text boxes beneath the lists. Use the arrow buttons to move the term to the right to seed your concepts. Click on Ok to return to the Concept Seed Editing interface.

To add thesaurus terms to a concept's early definition, select the concept in the Auto- or User-Defined concepts tab, and click on Edit. The following interface will open:



Here you can add terms strongly-related to your concept. For example, if you are interested in finding sections in your text containing violence, create the concept 'violence' and add any terms from the Frequent Words or Names list above that you think indicate a violent act. Only use words that fairly unambiguously indicate this concept in the text. Leximancer will automatically find additional terms from the text during the Thesaurus Learning phase, so you don't have to know all the relevant words in advance. As it is difficult to know whether two words are in fact used in similar ways in the text, it is best not to combine too many seed words into one concept. Any errors in manual seed combining at this stage can seriously distort the concepts and the map. It is generally better to use one seed word per concept, and see what lies near to what on the map. You can then always come back and merge nearby concepts. Click on Ok to save your changes and exit.

When you return to the Concept Seed Editing interface, you can also create and edit Tag categories. These are concepts for which no associated terms will be learned by Leximancer. This useful if you want to compare groups in the data (using file or folder tags) or perform a simple keyword search for terms.

Click on OK to close the Concept Seed Editor and return to the main Project Control Panel. Click on Generate Concept Map to run the remaining phases on default settings. Click on Concept Map to view the map containing the new concepts.

5.4 2. Profiling

This function is not the same as automatic concept discovery. The aim here is to discover new concepts during learning which are relevant to the concepts defined in advance, either in the Automatic- or in the User-Defined Concepts tabs. For example, this setting would allow you to extract the main concepts related to stem cell research from a broader set of documents.

Concept profiling settings can be found under the **Thesaurus Settings** prior to the Generate Thesaurus stage.

Note that Tags do not take part in this process automatically. If you have Tag categories, folder tags for example, which you wish to profile, you must turn on the *Learn From Tags* option in the Thesaurus Settings.

The profiling function has three alternative behaviours, called **Themed Discovery**: ALL, ANY, and EACH. You can ask for the related concepts to be relevant to most of the prior concepts, and thus follow a theme encompassed by all the prior concepts – this is the ALL option, and resembles set intersection. Alternatively, the discovered concepts need only be related to at least one of the prior concepts – this is the ANY option, which is similar to set union. The EACH option discovers equal fractions of profile concepts for each predefined concept, and these concepts show what is peculiar to each predefined concept. The EACH option is very useful for enhanced discrimination of prior concepts.

If you wanted to extract the main concepts related to stem cell research from a broader set of documents, for example, you could disable Automatic Concept Identification in Concept Seeds Settings prior to the Generate Concepts stage. Then you would *create user-defined seeds* for multiple simple concepts that encompass the scenario. You might seed concepts such as 'research', 'ethics', 'debate' and so on. Keep the seed words for each concept simple, and don't try too hard to restrict the seeds of each concept to just the topic you are after. In this instance we will be considering the intersection of all these elements. Click on Thesaurus Settings. Specify a quota of concepts to be discovered in the Concept Profiling options. Choose to discover several profiled concepts per prior concept. Select *Concepts in ALL* as the **Themed Discovery** setting.

If you are attempting to discover the network of associated names around a name or a scenario, you can choose to only discover name-like concepts during profiling, by enabling the Profiling setting called *only discover name-like concepts*. You should try the Social mapping layout algorithm first for this style of map.

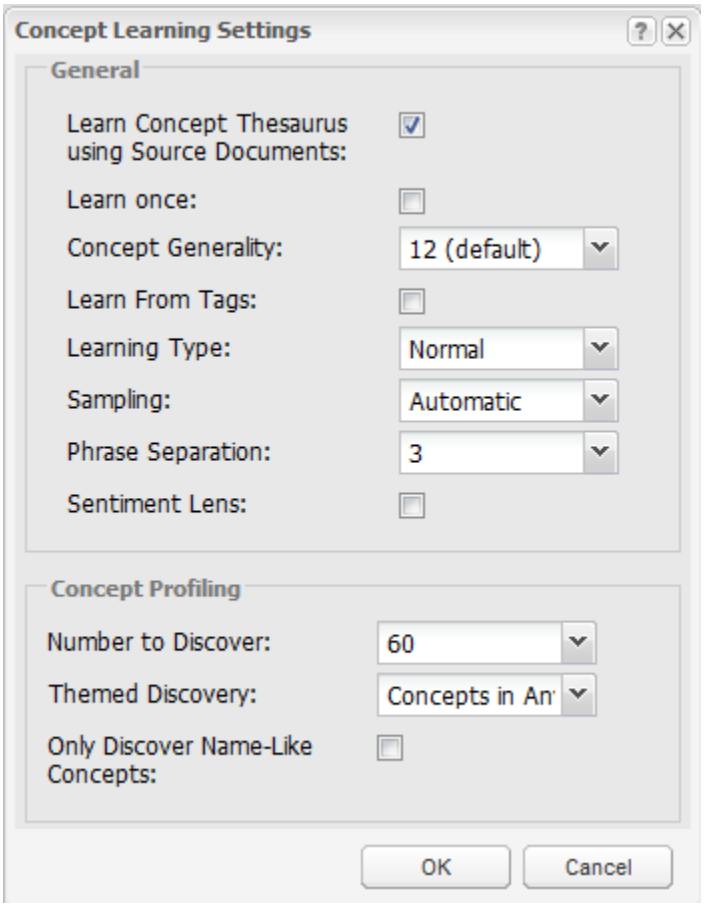
It is important to understand that although these discovered concepts are seeded from words that are relevant to the prior concepts, they are then learned as fully-fledged independent concepts. As a result, the map will contain some peripheral areas of meaning that do not directly contain the prior concepts. Contrast this with the Required Concepts function, which constrains all the coded text segments to be directly related to the required concepts.

5.5 Configuring Concept Profiling

Leximancer will generally extract the main concepts that occur within your documents. However, in some instances you may be interested in inspecting certain aspects of your text in more detail. For example, given 1000 newspaper articles, you may only be interested in the events containing references to violence. In such cases, you can use profiling to extract concepts that are specific to violence, rather than those that occur in all the articles that you have been given.

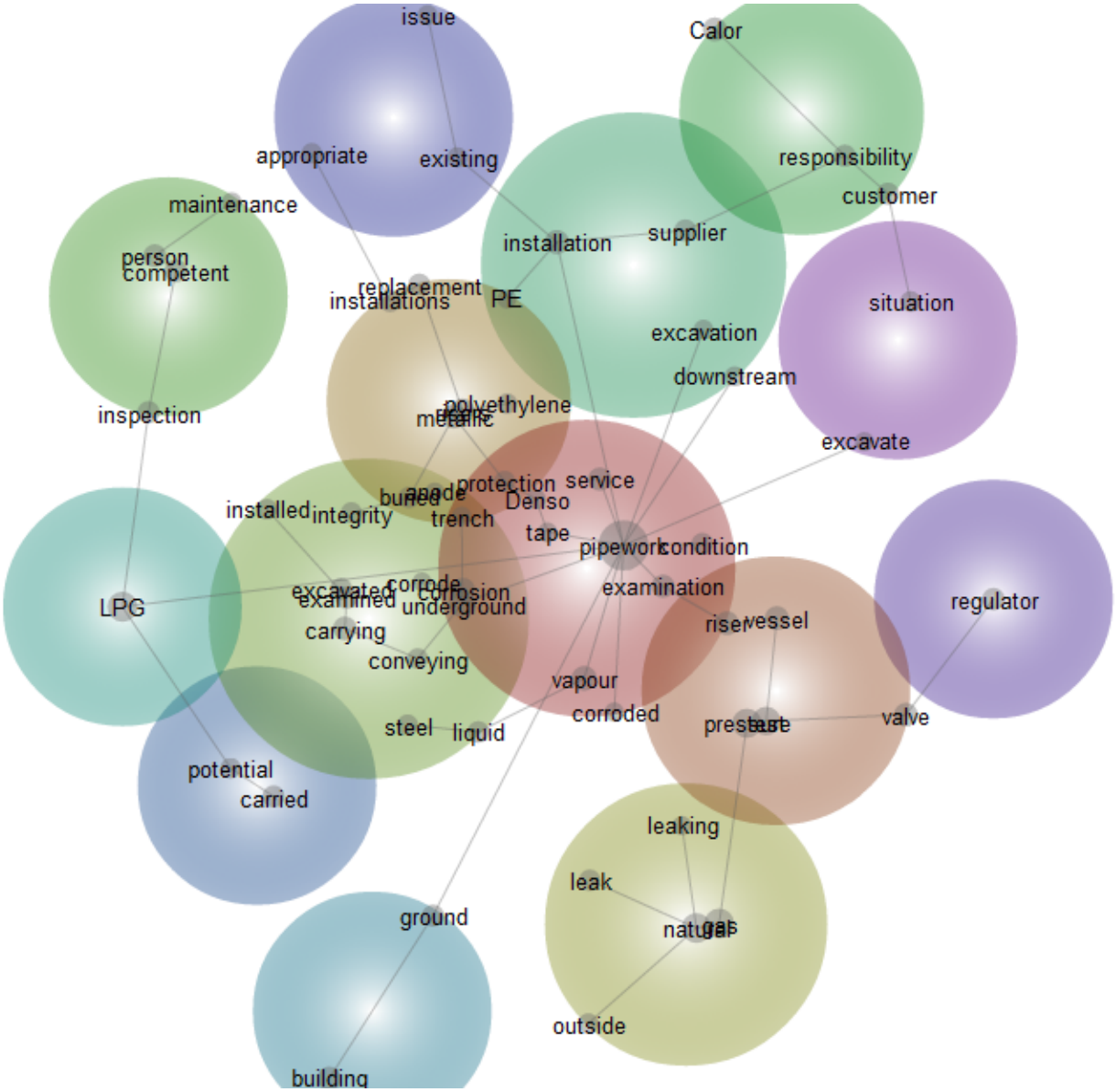
To profile an issue:

- Select your data files as usual, and expand the Generate Concepts Settings.
- Click on *Concept Seeds Settings* to *untick* the Automatically Identify Concepts box, because you don't want any concepts present other than the ones you are interested in profiling.
- Run the Generate Concepts stage.
- Click on Concept Seeds (as discussed above- Configuring Concept Editing) to create one or more used defined concepts to profile. In this case, a concept called pipework is of interest for profiling.
- Then click on Thesaurus Settings, and enter the number of concepts that you would like to see on the map related to this theme (60 in this example) **in the Concept Profiling section**:



- Click OK to exit the Thesaurus Settings. Click on *Generate Concept Map* in the main Project Control Panel to run all remaining stages of processing, then click Concept Map to reveal the profile map.

The concept map will contain profiled concepts specific to your area of interest. The map below is a profile of the Pipework concept in the Explosion Enquiry (hearing transcript) data set:



Concept profiling is very useful in cases where you have a large amount of text, but are only interested in exploring particular themes.

5.6 3. Configuring Folder and Filename Tags

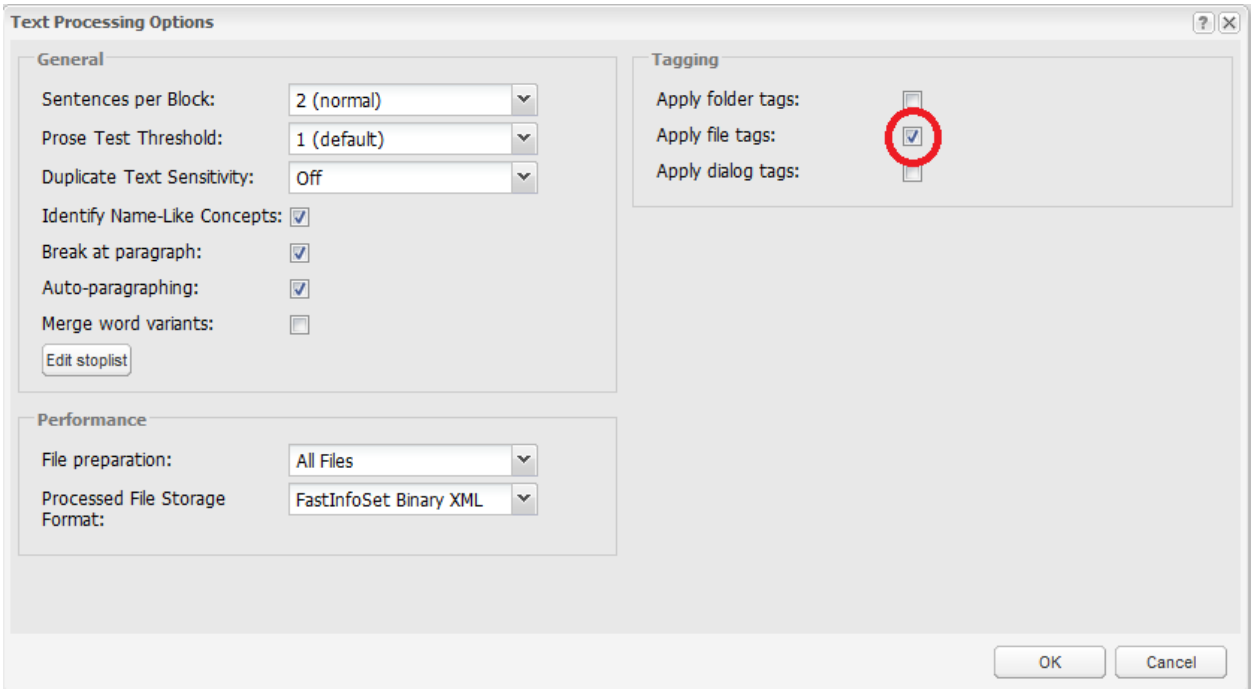
If you are going to analyse a set of multiple text files using Leximancer, you should consider making use of the Folder Tags function. Essentially, if the name of each file is a category of interest, or you can group the files into folders describing categories of interest, this feature is a simple of way to add a lot of power to your analysis.

For example:

- break up a book into various chapters (one file per chapter) to allow you to explore what topics or characters appear in each chapter,
- naming letters or reports - make each file name indicate the person or organisation who wrote the document to let you see each of these bodies on the map,
- group newspaper articles into folders by newspaper.

Leximancer can create a category called a Tag for each folder and / or file name in your data documents. You can create multiple levels of folders under your parent (top-level) data folder. For example, you can create a folder for each newspaper, and under each of those a folder for each month, and under each of those a folder for each journalist. You would place the text file for each article in the appropriate folder at the bottom of this tree. Leximancer can then create a Tag for each folder at each level of the tree.

To enable this function, click on Text Processing Settings to bring up the interface below. In the **Tag** options, select Folder tags or File tags:

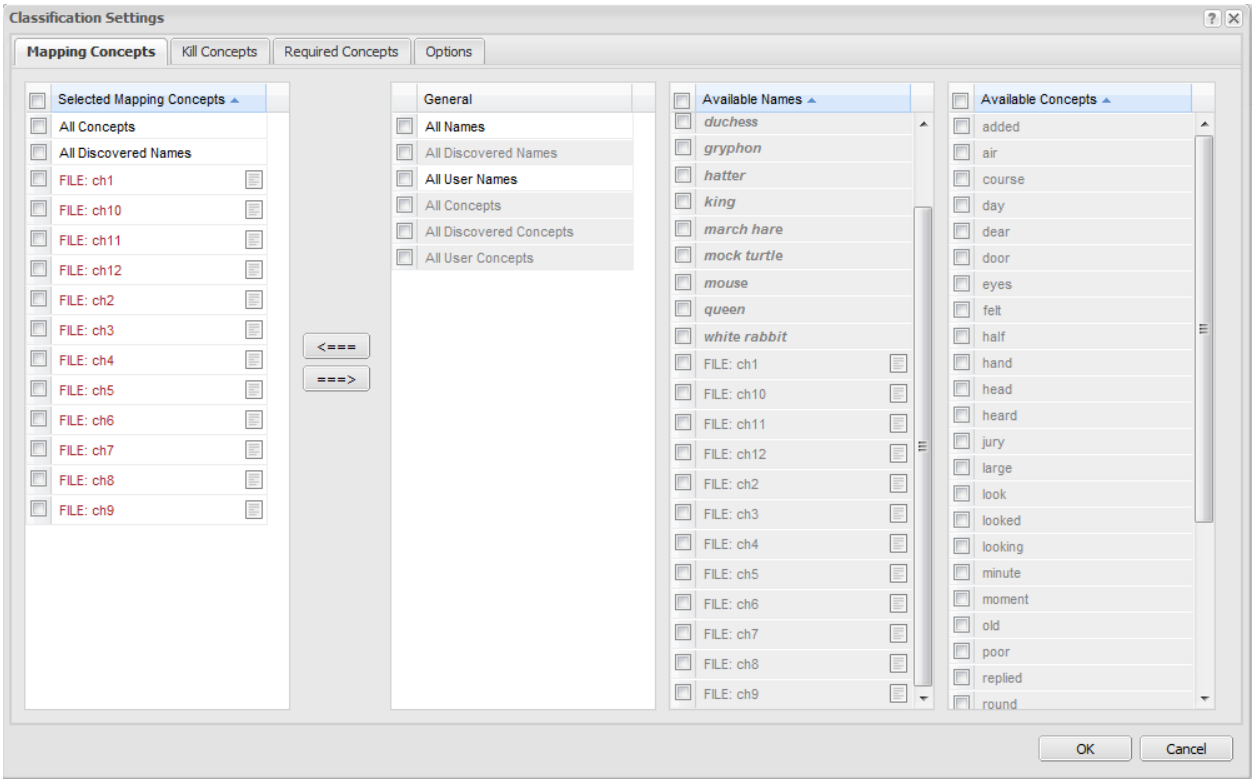


In this example, the chapters comprising the Alice in Wonderland story were loaded into Leximancer as separate documents. When the File tags option is enabled therefore, a tag is automati-

cally created to represent each the chapter in the story.

Click OK to exit the Processing Settings.

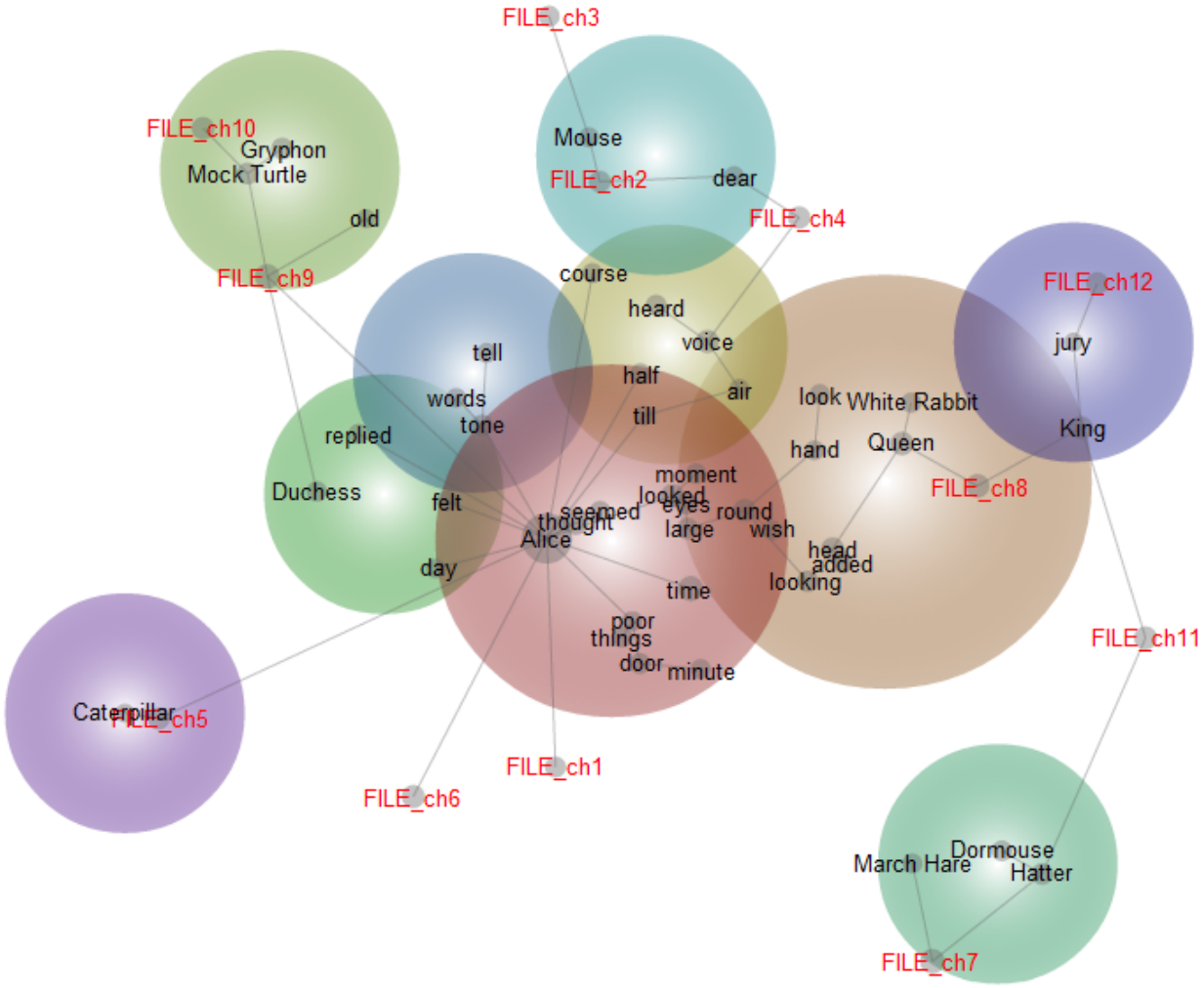
Run the Generate Concepts and Generate Thesaurus stages of processing. Edit the Concept Coding Settings. Use the left-hand arrow button to add the file tags to the list of **Mapping Concepts**:



Click Ok, and then click on Generate Concept Map in main Control Panel to complete the final phase of processing. Click on Concept Map when the project is complete.

The concept map now includes the chapter file tags, and the concepts are clustered around these according to their relationships. Concepts coming from the content of a particular chapter will tend to settle near that chapter's file tag in the map space.

You can explore the topics characteristic of a chapter by clicking on a file tag. A ranked list of related topics is revealed in the panel on the right. These are the concepts that are coded into the chapter frequently:



In the example above, the Mock Turtle and Gryphon are clustered near Chapter 10 file tag, indicating that they are fairly specific to this chapter.

Filename and folder tags are useful if you wish to explore similarities or differences between various conditions. For example, placing speeches by various politicians into folders according to the party to which they belong will allow you to explore the views of the party, whereas placing files into folders corresponding to different periods of time will allow you to explore temporal differences.

5.7 4. Extracting a Social Network

5.8 A multi-partite network of names and descriptors

The goal here is to extract a network of actors, or names, based on their interactions, and to characterize the conceptual nature of their interactions. An example of this could be finding a network of companies, employees, and cities from a set of industry reports.

Firstly, change the Concept Seeds Settings to extract only names, as many names as you think are warranted. Force all the automatic concepts to be names by setting the Percentage of Name-Like Concepts to 100%.

After running Generate Concepts, tidy-up the automatic concepts in the Concept Seeds Settings prior to Generate Thesaurus. For instance, you might remove any unwanted proper nouns, and add any missing names that you want to watch.

Next, go into the Thesaurus Settings and enable the Concept Profiling function. Select the *Concepts in ANY* themed discovery setting, and choose to discover one concept per prior name. You can increase the number of discovered concepts if you want a richer map.

Run the Generate Thesaurus phase, then go into the Concept Coding Settings. Make sure the tags, names and concepts of interest are included in the Mapping Concepts list. If not, use the left-hand arrow to add them to the list.

The resulting map will show a network of names intertwined with concepts that describe and mediate the relationships.

5.9 A unipartite social network contstrained by structural variables

For a more traditional and constrained social network which uses structural variables:

- Edit the Concept Seeds Settings to select only names as described above, then run the Generate Concept Seeds phase.
- In Concept Seeds, manually seed some user defined concepts as structural variables, such as family, economic, professional, etc.
- Run the Thesaurus Learning phase.
- Open the Concept Coding Settings, and put only the name-like concepts in the *Mapping Concepts* list.
- Place the desired structural concepts in the list of *Required Concepts*, then run the remaining stages.

The result will be that only text segments that contain one of the Required Concepts will be mapped. Consequently, the map will show a network of names based on relationships that involve at least one of the required concepts - the structural variables.

5.10 5. Analysing Transcripts

Transcripts of meetings, interviews and focus groups can be analysed in Leximancer as normal text, and if you group the interviews into files and folders, you can use Folder Tagging (Chapter 14) to enhance your analysis. Moreover, if your transcripts are in plain text or Microsoft Word and are suitably formatted, Leximancer allows you to select, ignore, or compare all the utterances of each distinct speaker. To allow the program to identify the speaker of any text segment, they must be identified in a certain way, and a new speaker label must be inserted whenever any new speaker begins. The format requires dialogue markers which are at the start of a paragraph; use upper case first letters for each constituent term; are made up of a maximum of three terms; and end in a colon followed by a space. For example:

Interviewer: So what does your son like to do for leisure, Susan?

Susan: Every Friday he plays uh ten pin bowling with the oldies. He's not bad either.

Alan: Oh yes, he excels at ten pin bowling, he's one of the better players there.

Interviewer: And do you have any plans for travel coming up?

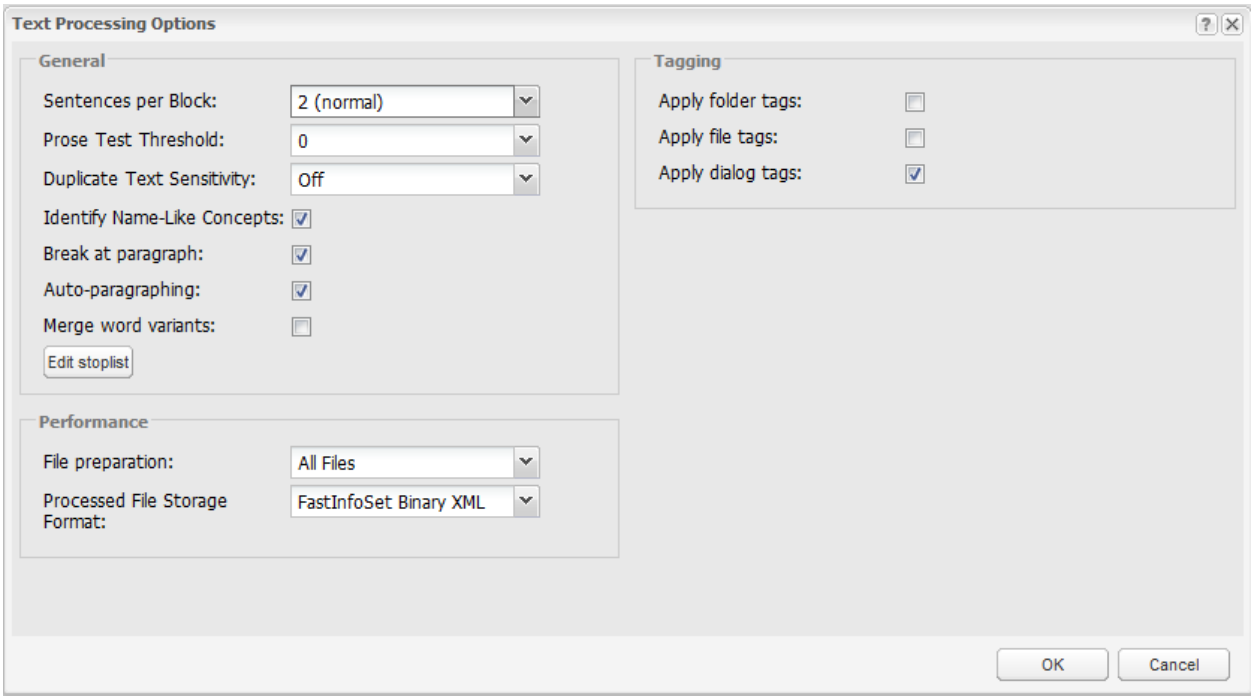
Susan: Yes actually we're going up to Thailand on the 13th of October, (B)'s coming with us for 14 days. My daughter in law is there, and they've got a little boy.

Alan: Yeah, so we'll show you a photo of that, she's very cute. Four boys, four grandsons, and one granddaughter.

Given text data in this format, Leximancer can extract the dialogue markers as tags and identify the speaker of every subsequent sentence until the next dialogue marker.

5.11 Configuring Transcript Analysis

Select your text data as usual. To create dialogue tags, open the Text Processing Settings. Tick Apply Dialogue Tags:

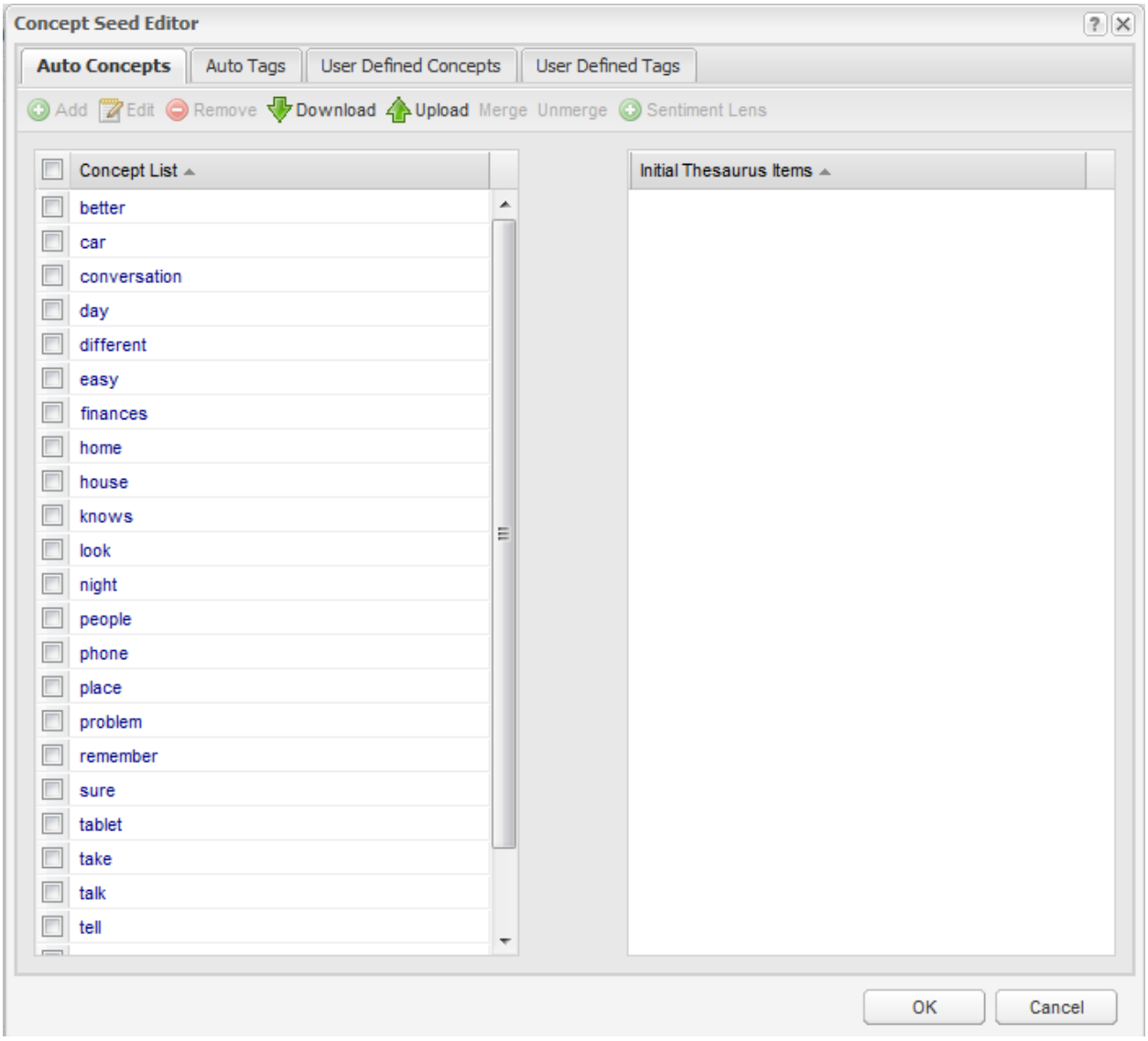


This transforms each dialogue marker in the text into a tag, which is then inserted into each relevant sentence and displayed for you under Auto Tags tab in the Concept Seeds interface.

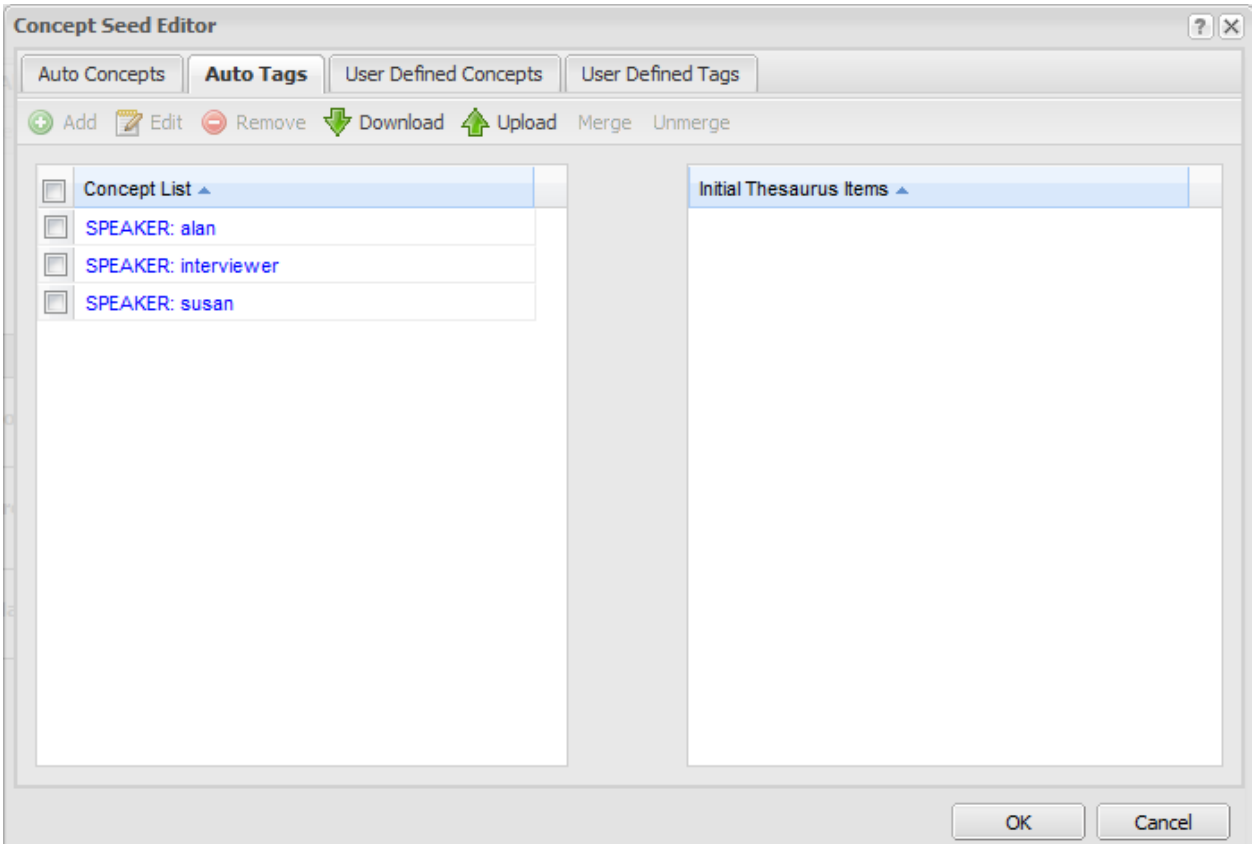
There is another setting in the Text Processing Settings called the Prose Test Threshold. If your interview text is quite colloquial and does not conform to standard stop-word usage, set the Prose Test Threshold to 0.

Run the Generate Concepts Seeds stage.

When this stage is complete, open the Concept Seeds node to inspect the extracted textual concept seeds in the Auto Concepts tab:

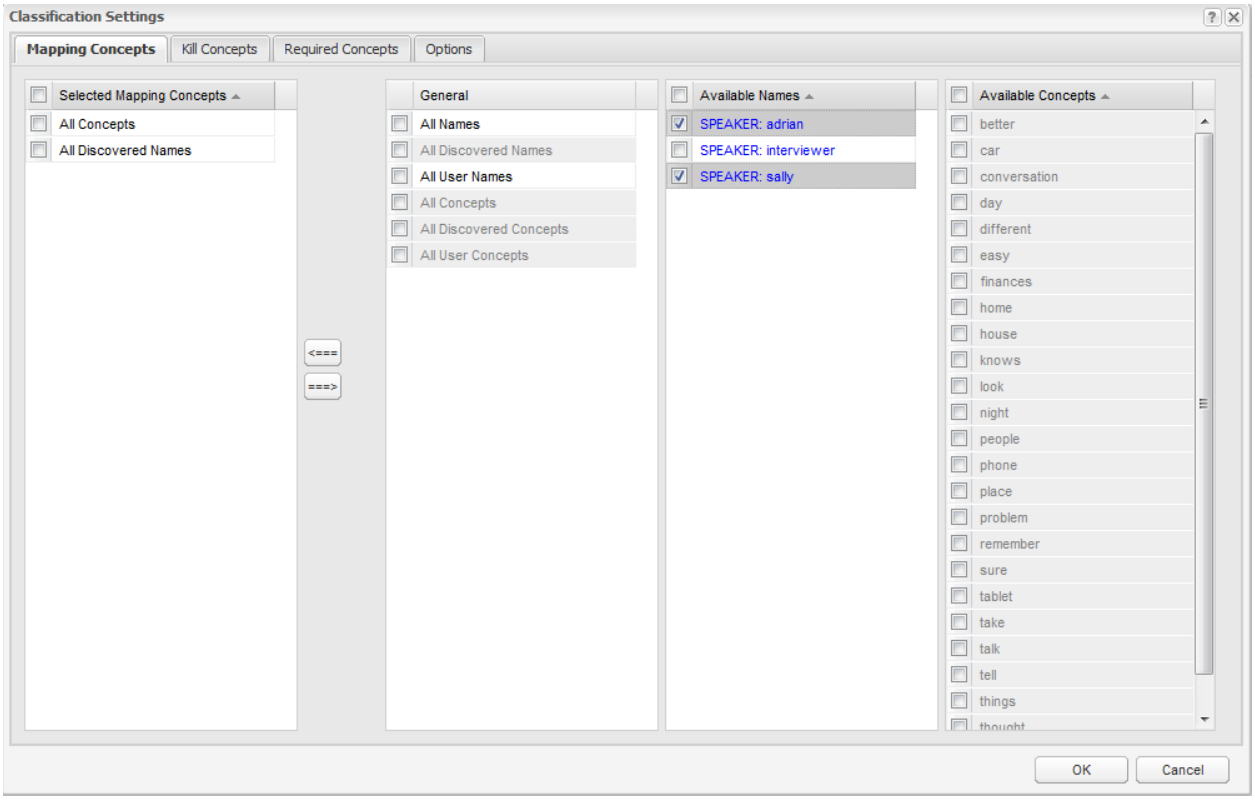


Then click on the Auto Tags tab to see the list of speakers identified by Leximancer:



Now run the Generate Thesaurus phase.

When this is done, you can choose whose utterances you wish to analyse, and whose you wish to leave out. You can also choose which items you wish to see on the map. These settings can be changed by opening the Concept Coding Settings:

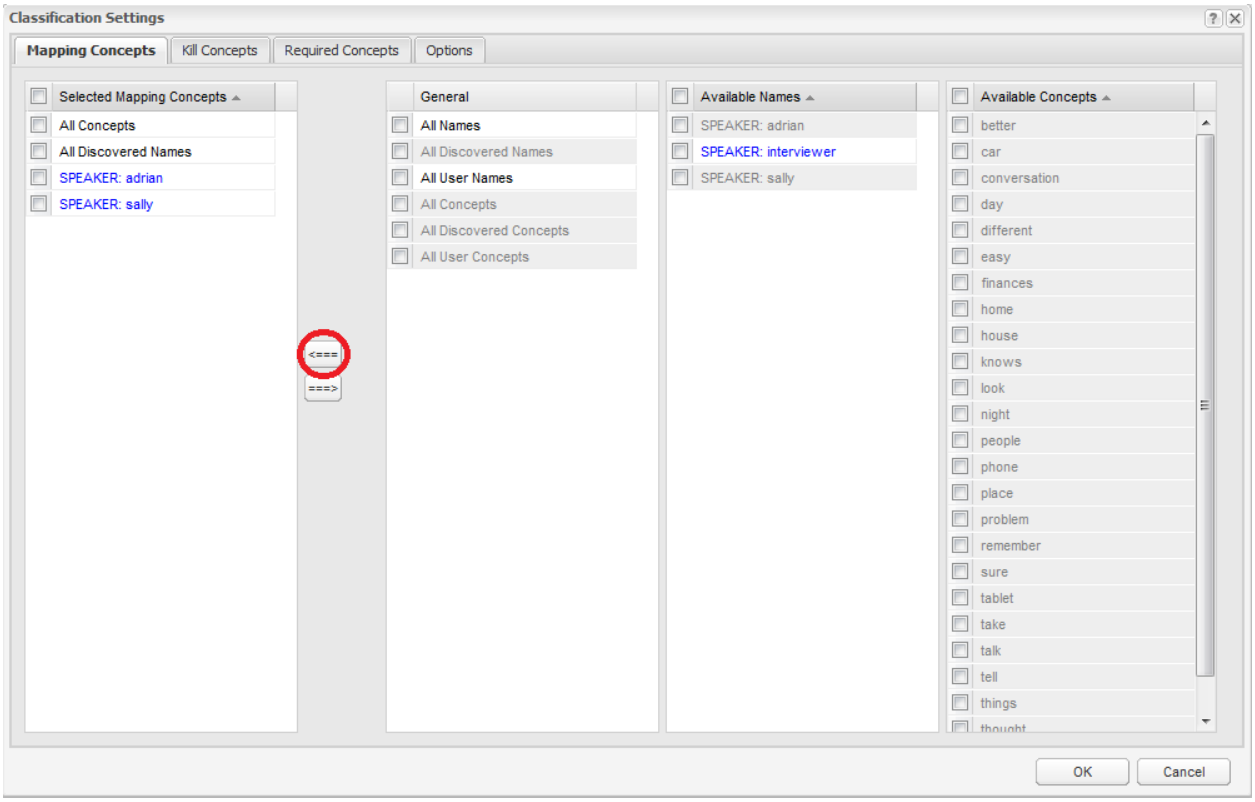


The Mapping Concepts list lets you chose which tags and concepts you wish to appear on the concept map. By default, there are two items in the list, All Concepts and All Discovered Names.

- The All Concepts wildcard represents all of the concepts identified for this project, be they automatically-discovered or user-defined.
- The All Discovered Names wildcard represents only those name-like concepts discovered automatically by the software.

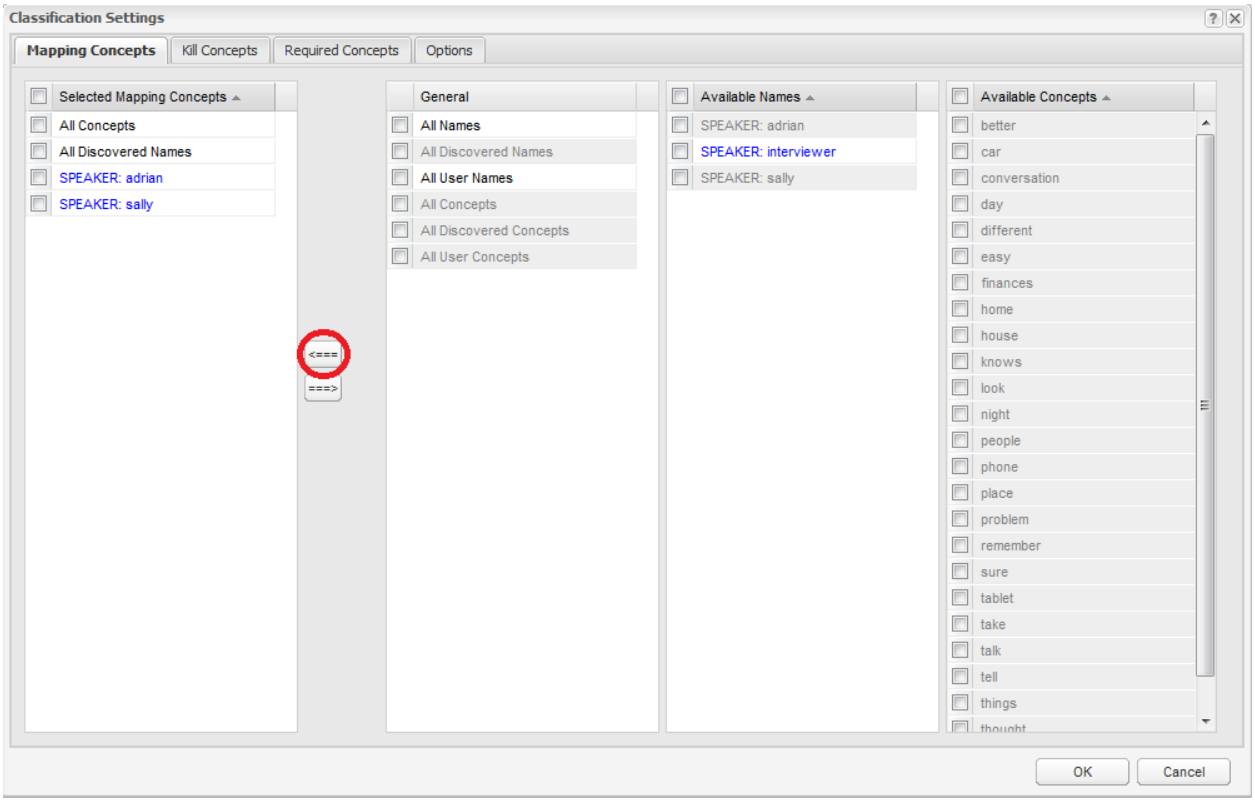
Full lists of the possible name-like and word-like concepts appear in the right-hand panels so that you can choose which entries you would like to see on the map.

If you wish to inspect the relative ownership of the textual concepts between your speakers, select the desired speaker tags and use the left arrow to add them to the Mapping Concepts list:

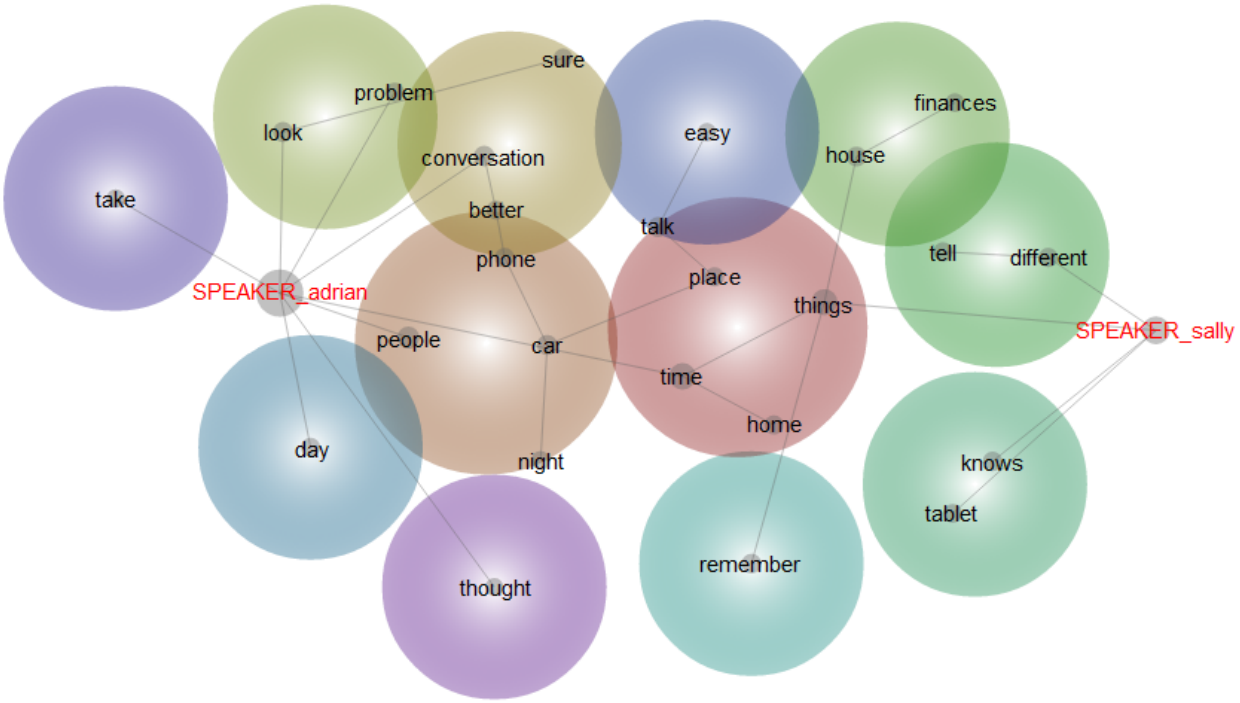


Since the full list of word-based concepts is of interest in this case, we can leave the All Concepts wildcard in the Mapping Concepts list.

The Required and Kill Concepts tabs allow you to select whose utterances you wish to analyse, and whose you wish to leave out from the analysis. For example, if you wanted to suppress all the utterances of the Interviewer, you would move the Interviewer speaker tag into the Kill Concepts list. This causes all the concepts coded into questions asked by the Interviewer to be removed from the analysis:



With these changes made, you can click Generate Concept Map to complete the final phase of processing and produce a concept map:



5.12 6. Analysing Spreadsheet Data

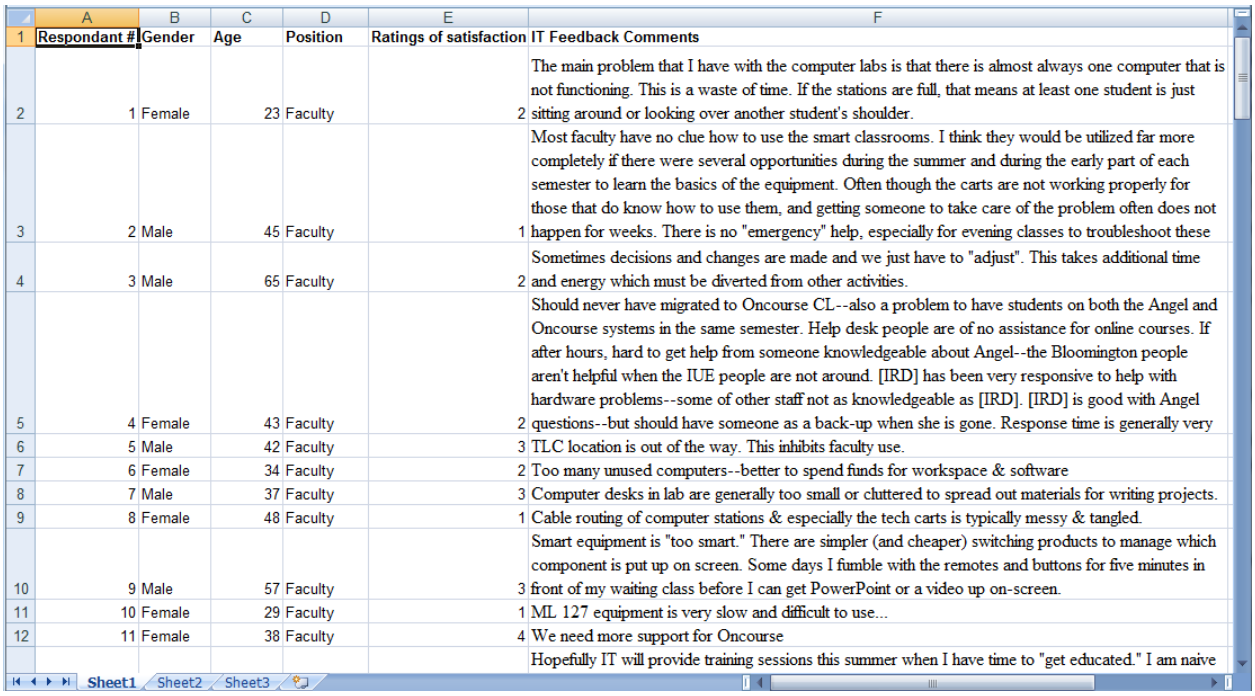
Leximancer can effectively analyse spreadsheets containing text fields and category fields (data in tabular format). You can analyse multiple text fields in each record, and also include the categorical fields as variables in your analysis. This enables very powerful text mining.

5.13 Practical: Analysing Spreadsheet Data

Your original data may have been in a spreadsheet application such as Microsoft Excel or in a database application such as Access. The required spreadsheet arrangement or layout for analysis is:

- The **first row** of the spreadsheet must contain headers for each column.
- There should be one respondent per row (e.g one survey respondent).
- There should be one response item (e.g. answer) per column.

To load the spreadsheet into Leximancer, you **must** export it first as a .csv file (comma separated values), or a .tsv file (tab separated values). This is easy in Excel, using the Save As function.



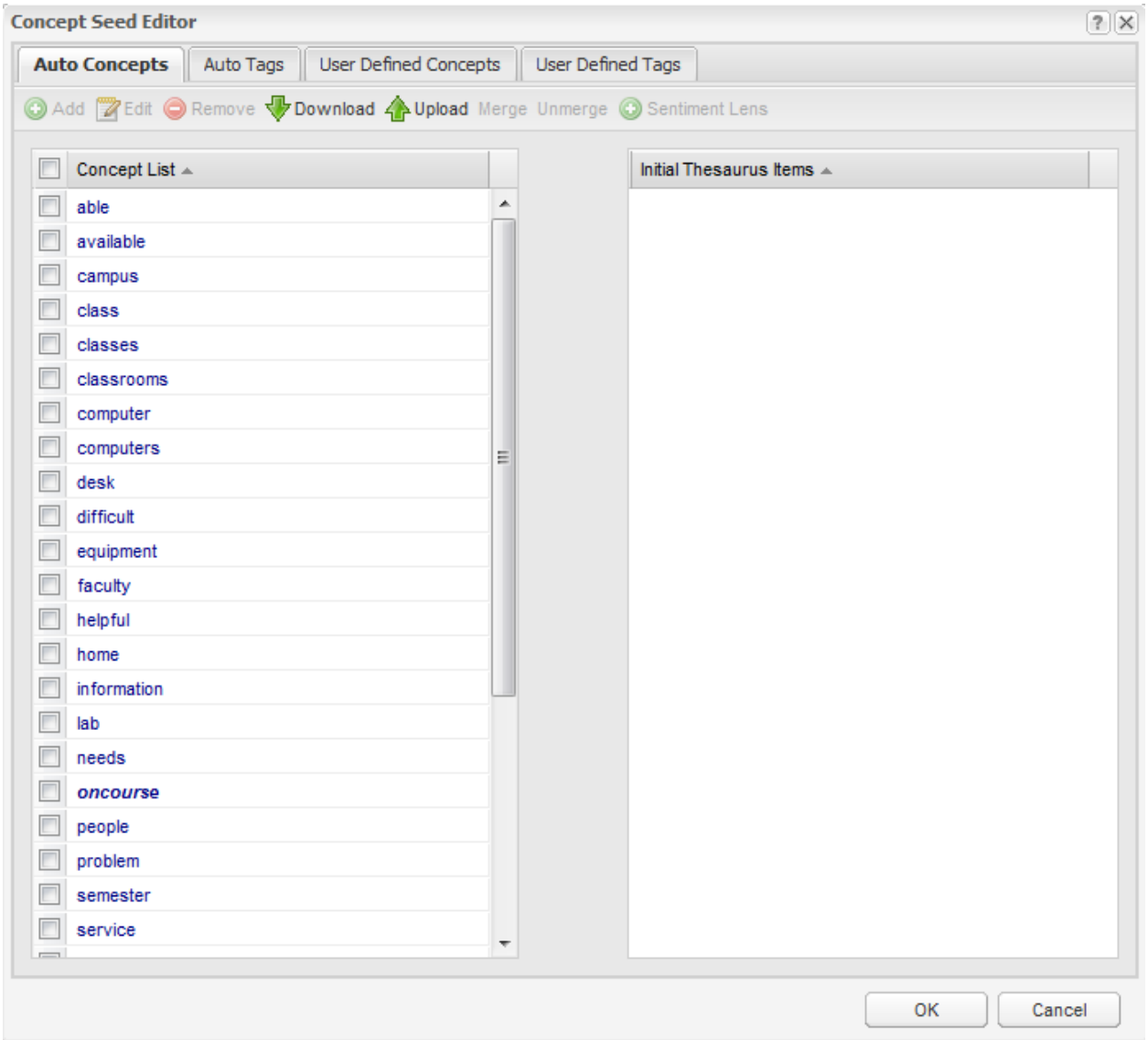
	A	B	C	D	E	F
1	Respondant #	Gender	Age	Position	Ratings of satisfaction	IT Feedback Comments
2	1	Female	23	Faculty	2	The main problem that I have with the computer labs is that there is almost always one computer that is not functioning. This is a waste of time. If the stations are full, that means at least one student is just sitting around or looking over another student's shoulder.
3	2	Male	45	Faculty	1	Most faculty have no clue how to use the smart classrooms. I think they would be utilized far more completely if there were several opportunities during the summer and during the early part of each semester to learn the basics of the equipment. Often though the carts are not working properly for those that do know how to use them, and getting someone to take care of the problem often does not happen for weeks. There is no "emergency" help, especially for evening classes to troubleshoot these
4	3	Male	65	Faculty	2	Sometimes decisions and changes are made and we just have to "adjust". This takes additional time and energy which must be diverted from other activities.
5	4	Female	43	Faculty	2	Should never have migrated to Oncourse CL--also a problem to have students on both the Angel and Oncourse systems in the same semester. Help desk people are of no assistance for online courses. If after hours, hard to get help from someone knowledgeable about Angel--the Bloomington people aren't helpful when the IUE people are not around. [IRD] has been very responsive to help with hardware problems--some of other staff not as knowledgeable as [IRD]. [IRD] is good with Angel questions--but should have someone as a back-up when she is gone. Response time is generally very
6	5	Male	42	Faculty	3	TLC location is out of the way. This inhibits faculty use.
7	6	Female	34	Faculty	2	Too many unused computers--better to spend funds for workspace & software
8	7	Male	37	Faculty	3	Computer desks in lab are generally too small or cluttered to spread out materials for writing projects.
9	8	Female	48	Faculty	1	Cable routing of computer stations & especially the tech carts is typically messy & tangled.
10	9	Male	57	Faculty	3	Smart equipment is "too smart." There are simpler (and cheaper) switching products to manage which component is put up on screen. Some days I fumble with the remotes and buttons for five minutes in front of my waiting class before I can get PowerPoint or a video up on-screen.
11	10	Female	29	Faculty	1	ML 127 equipment is very slow and difficult to use...
12	11	Female	38	Faculty	4	We need more support for Oncourse
						Hopefully IT will provide training sessions this summer when I have time to "get educated." I am naive

Sheet1 Sheet2 Sheet3

Note: To load the spreadsheet into Leximancer, you **must** export it first as a .csv file (comma separated values), or a .tsv file (tab separated values).

Create a Leximancer project as usual, and click on Load Data. Browse and select the spreadsheet file (the .tsv or .csv file). Drag and drop it into the Document Set area, then click Ok.

Run the first phase of processing (Generate Concepts), then open the Concept Seeds node. You will see the concept seeds automatically extracted by the software in the Auto Concepts tab:



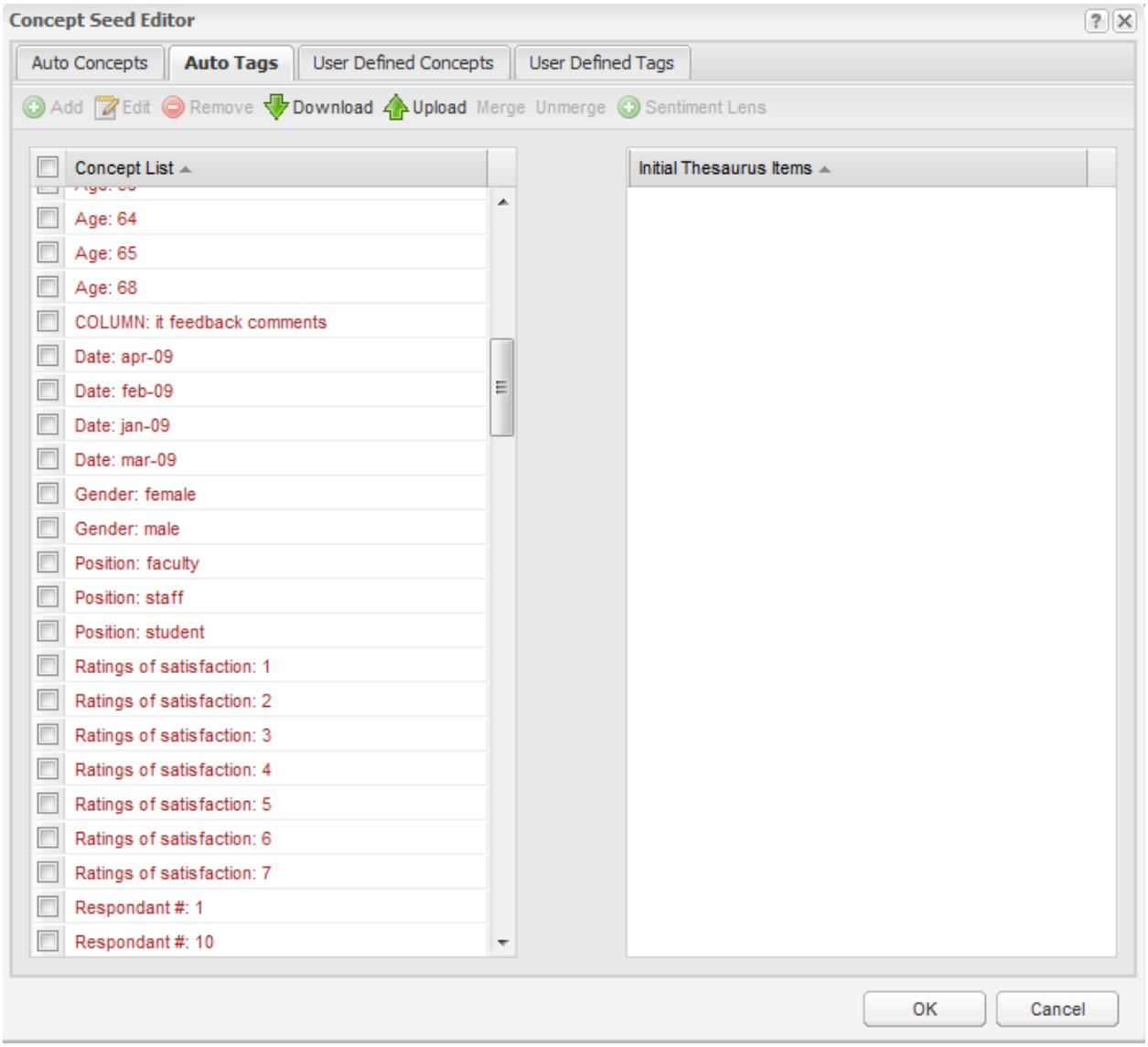
You can seed your own concepts in the User Defined Concepts tab.

If you want more (or fewer) automatic concepts, just go back one node to the Concept Seeds Settings and change the Total Number of Concepts to be suggested by Leximancer.

If you process a spreadsheet that uses the layout described above, Leximancer will automatically create a tag to represent each free-text column, and a tag for each of the levels (or possible responses) to the categorical variables in the data (to a limit of 500 unique responses). Entries that read '@none' represent null entries or empty response cells.

You can review what tags Leximancer has extracted by clicking on the Auto Tags tab. The tags

take their names from the column headers and categorical responses in the data. There is no hard limit to the number of free-text and categorical variables that can be analysed in Leximancer. The Auto Tags can be used for data mining correlations with textual concepts, and for selecting which text column(s) you wish to map at any time:

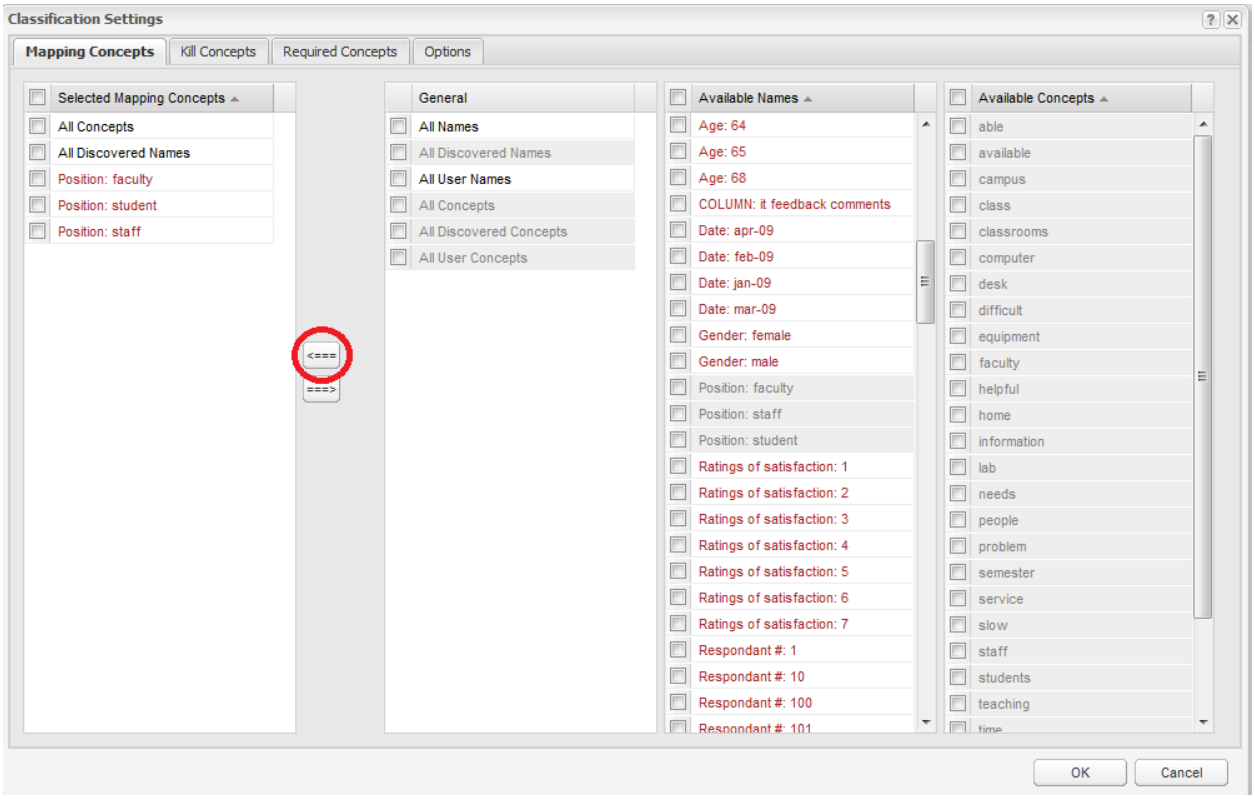


Run the Generate Thesaurus phase to extract a thesaurus from the data describing each concept seed.

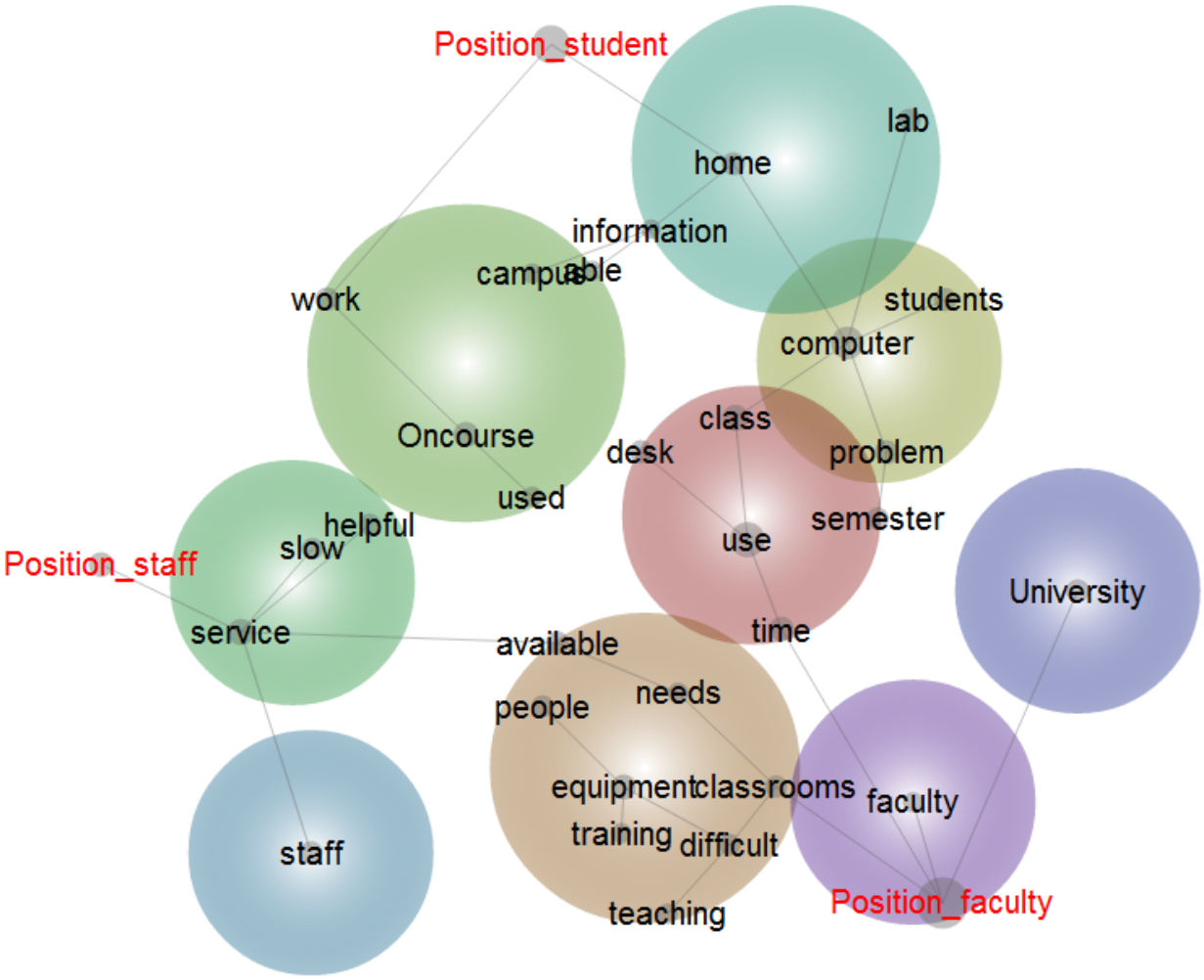
Open the Concept Coding Settings to access the data mining options.

The Mapping Concepts list lets you select what variables you want on the concept map, like choosing the columns you want in a database query. For example, let's say you wish to examine the correlations of the three position types (faculty members, staff and students) with the comments from the text column called 'IT feedback comments' in the data. You would like all the textual concepts on the map, so retain the All Concepts wildcard in the Mapping Concepts list. You would also need to select and add the three respondent type tags (Position: Faculty, Position: Staff and

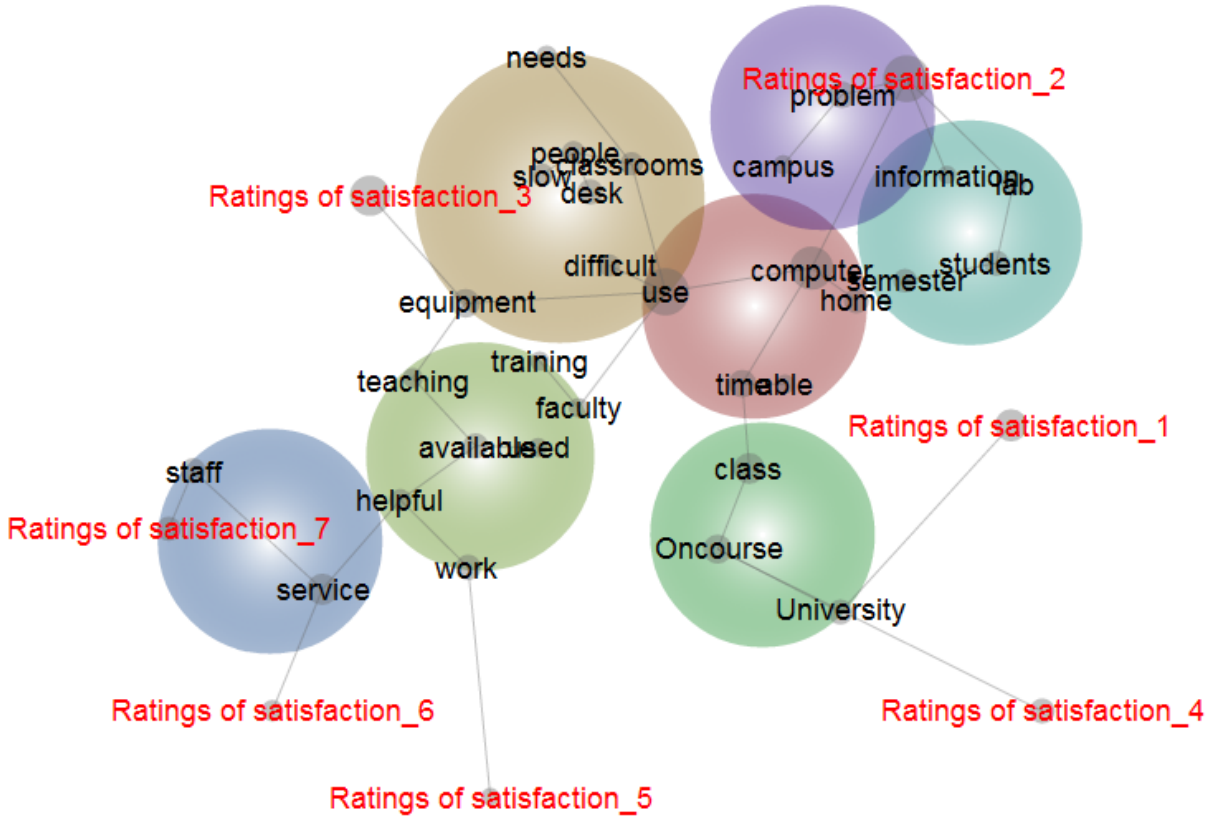
Position: Student) to the Mapping Concepts tab using the left arrow:



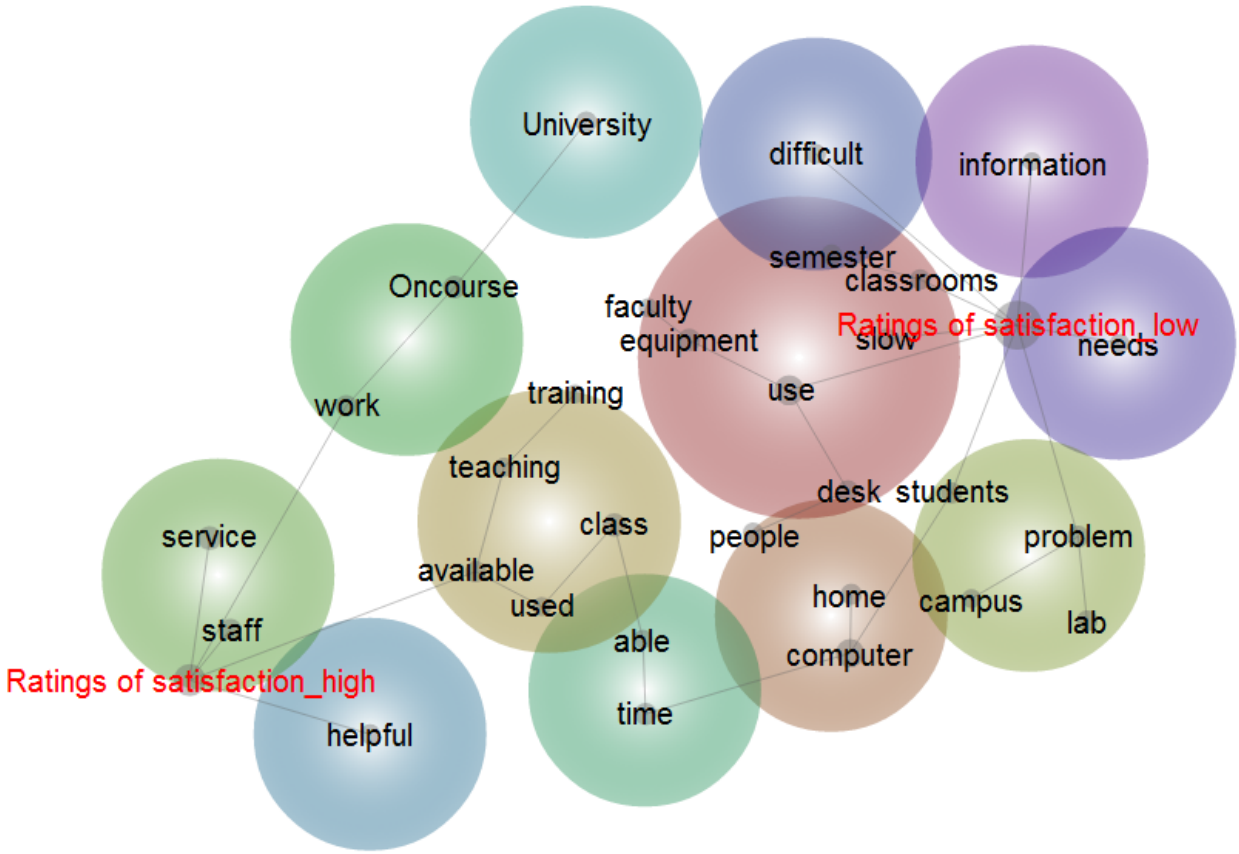
Once you have configured the data mining settings, run the last phase of processing and inspect the resulting concept map:



You could choose to correlate the textual concepts with the satisfaction ratings instead of the position categories. To do so, return to the data mining settings in the Concept Coding Settings, and remove the position tags from the Mapping Concepts list. Replace them with the satisfaction score tags. Rerun the final phase to produce this new view of the data very quickly:



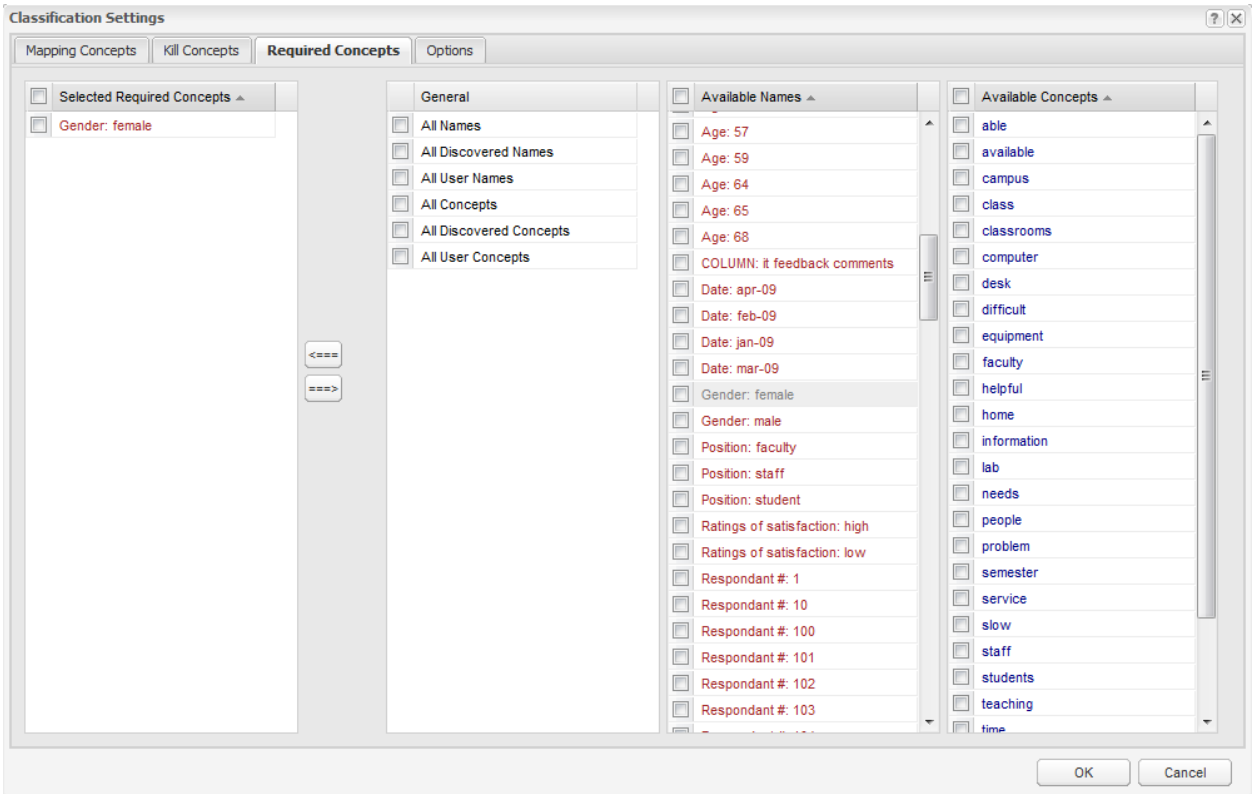
You could also aggregate the satisfaction tags in the Concept Seeds dialogue to produce 'Low' and 'High' satisfaction tags to place on the concept map. For instance, you could Merge the satisfaction scores of 1-3 and Edit this to rename it as tag as 'Low'. Then Merge the 4-7 tags and Edit to rename them as 'High'. Adding the 'Low' and 'High' tags to the Mapping Concepts list in the Concept Coding Settings produces the following map:



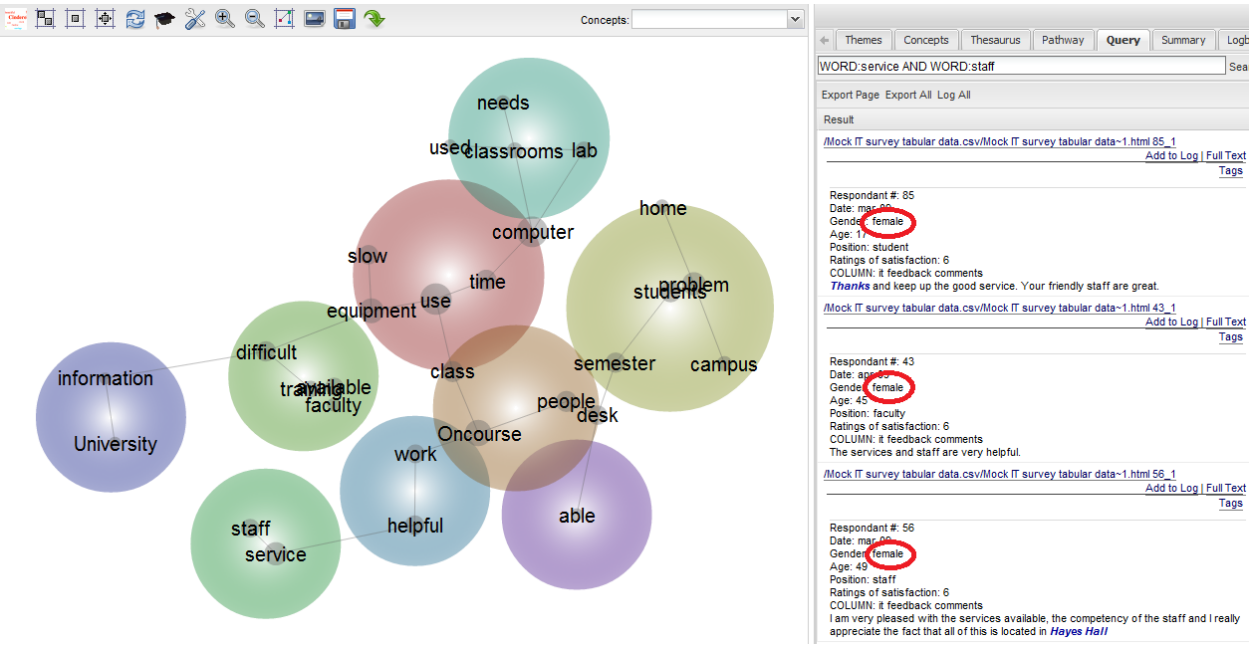
You can also use the Required Concepts and Kill Concepts tabs in the Concept Coding Settings to filter records in or out of your analysis. For instance, if you had more than one text column in your spreadsheet, you could choose to examine the concepts associated with a particular text response. You would do this by moving the tag denoting the text column of interest into the Required Concepts tab. In this case, if a data cell does not come from that text column, it will not be coded for concepts and mapped.

A Kill Concept is the opposite of a Required Concept. If a text segment matches a Kill Concept tag or concept, it will not be coded with any classifier. For example, you could suppress the analysis of all text segments which match the concept 'available' by identifying 'available' as a Kill Concept.

By way of example, if we wished to map only the comments made by women in this spreadsheet, we could either add the Gender: female tag to the Required Concepts tab, or by add the Gender: male tag to the Kill Concepts tab:



If we remove all the tags from the Mapping Concepts tab (so that only the default All Concepts and All Discovered Names wildcards remain), the resulting map reflects all the responses made by women in the data:



This concludes the Leximancer 4 Manual. Visit our website at info.leximancer.com

http://info.leximancer.com/.