

# Leximancer as a content analysis tool

<http://www.leximancer.com/>

# Leximancer and content analysis

- Leximancer does conceptual analysis and relational analysis
- Conceptual analysis (thematic analysis)
  - The detection & quantification of predefined concepts within the text
- Relational analysis (semantic analysis)
  - Measurement of relationships between identified concepts within text

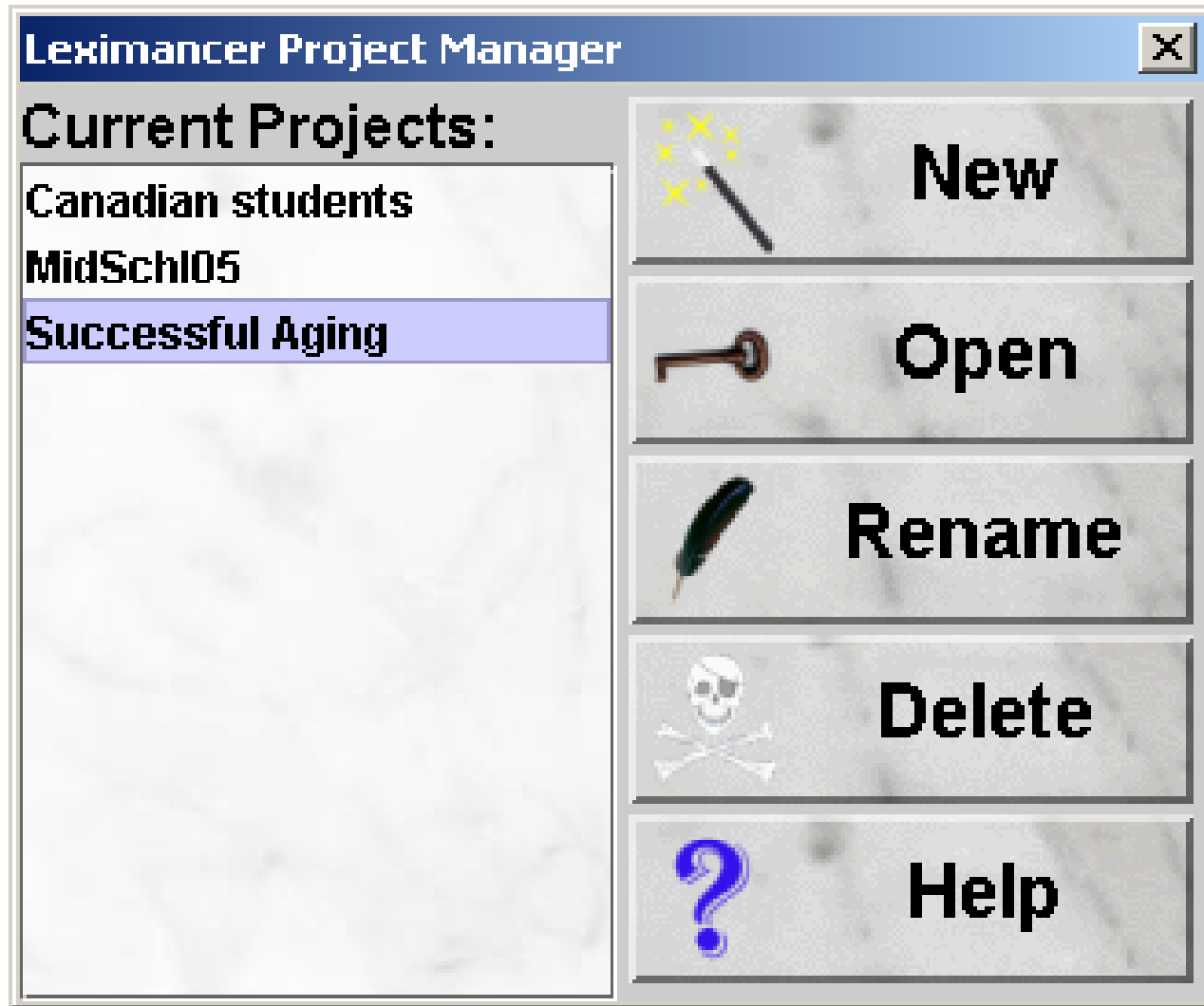
# Steps in conceptual analysis

- Frequently used terms used (concept seeds) to generate a thesaurus of terms
  - Names identified (i.e., usually start-of-sentence)
  - Non-lexical and weak semantic information removed (e.g., 9, &)
  - Nontextual material excluded (e.g., menus)
  - Can be turned off to use self-defined concepts instead
- Terms around which other terms cluster identified
  - (e.g., The terms, “fleas, bites” cluster around “dog, hound, puppy”)
- Iterative process ensues in which some of the potential concepts eliminated
- This process converges on stable state containing most highly relevant concepts only

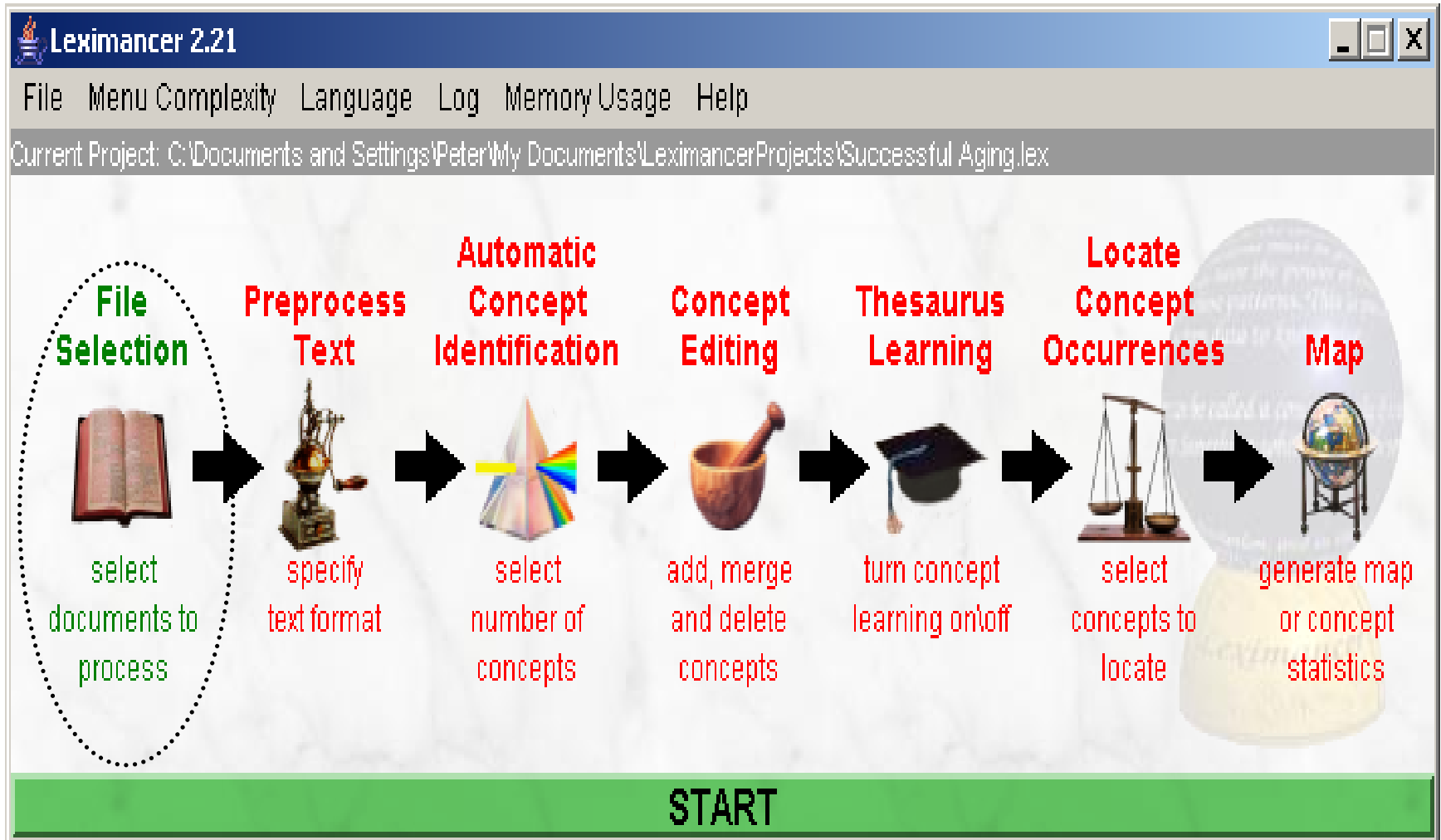
# Steps in relational analysis

- Proximity analysis measures co-occurrence of concepts within text
  - Length of words or sentences specified (called window)
- Window moved sequentially through text, noting co-occurring concepts (usually in three sentence blocks)
- Result stored in co-occurrence matrix, which stores frequency of co-occurrence of all concepts against all others
  - Leximancer stores information in spreadsheet (spreadsheet.txt)
- Third stage of relational analysis (cognitive mapping) represents the information visually for comparison (concept map)

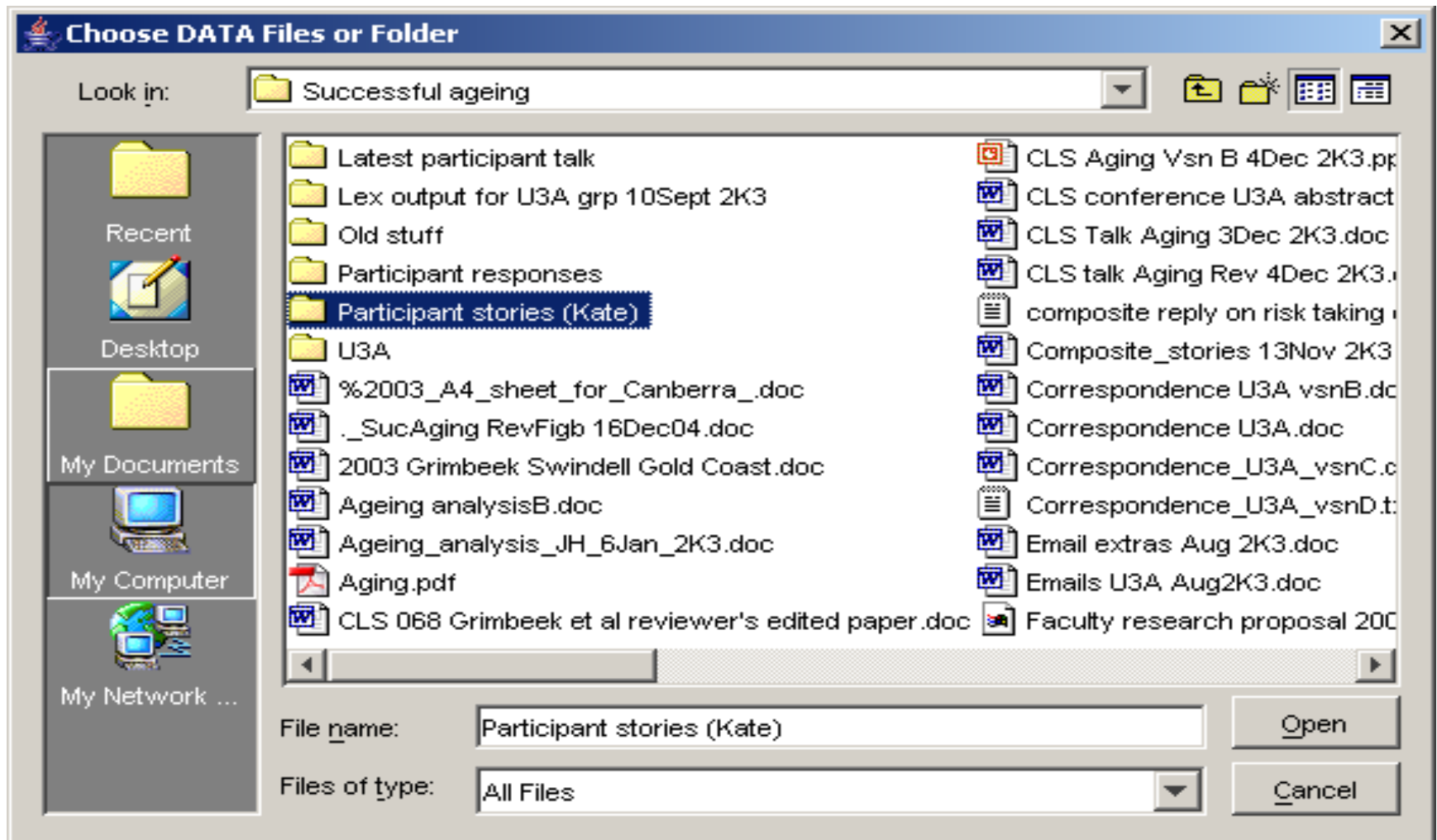
# Using Leximancer on default settings: Starting Screen



# Main menu: Leximancer 2.21

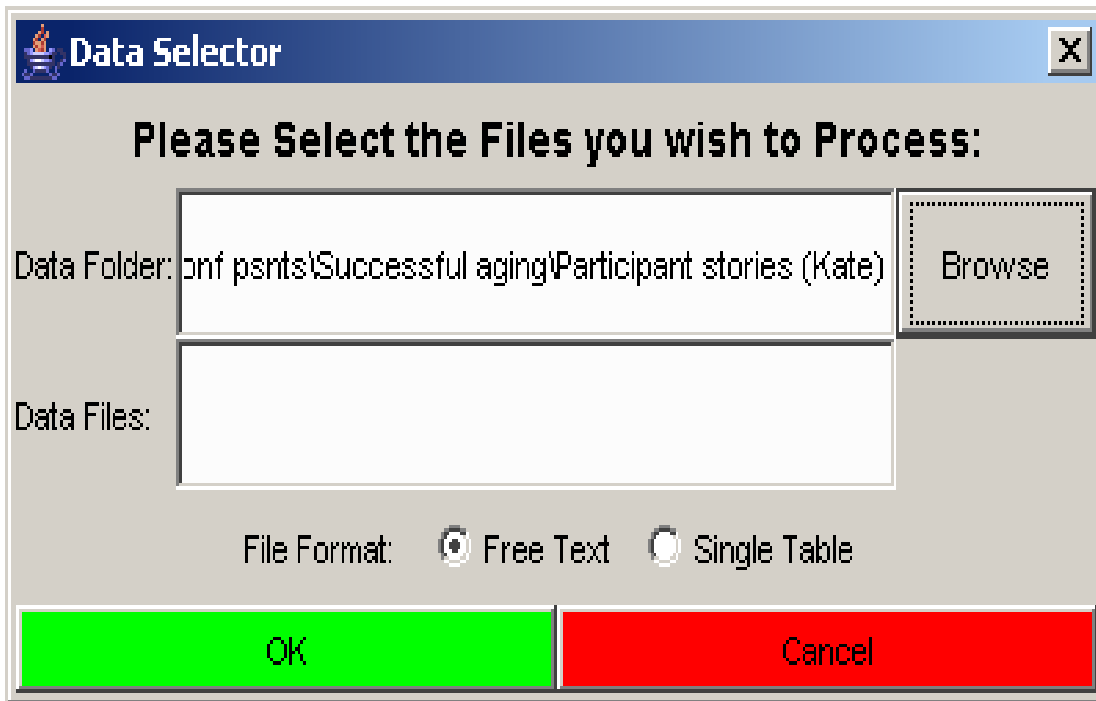


# File selection

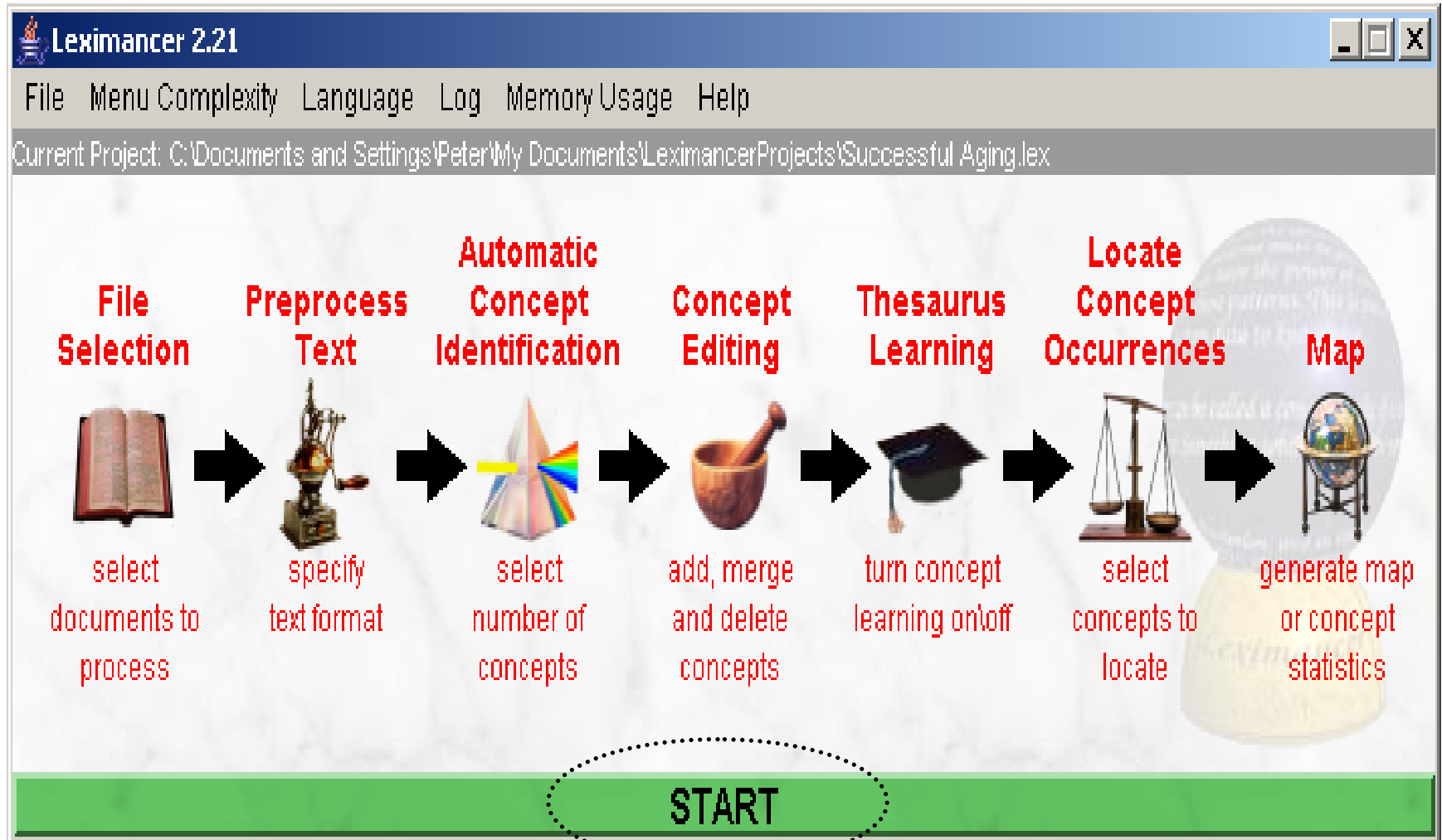


# File selection

- Leximancer can process files of the type .doc, .html, .htm, .txt, .xml and .pdf.
- Files with no extension are also supported, but are treated as .txt formatted.
- Generally, if there are other formats of text you wish to process, it is recommended that you look into pre-converting them into HTML.
- Here all the files in a data folder are being selected simultaneously.




# Default setting operation



# Initial display of outcomes (default settings)

The screenshot displays the Leximancer Map application window. The main area shows a concept map for the word "books" with 1000 iterations. The map features several concepts represented by colored circles of varying sizes and brightness, connected by lines. The concepts include "communication", "rev", "online", "internet", "time", "computer", "found", "courses", "e-mail", "information", "people", and "interest". The "found" concept is the largest and brightest, indicating its high frequency. The "Map Instructions" panel on the right provides a welcome message and directions on how to use the concept map.

**Map Instructions**

Powered by   
**Leximancer**

Welcome to the Text Explorer. This system allows you to explore a concept map of a document collection.

**Directions**

**What the Concept Map means:**

- The brightness of a concept is related to its frequency (i.e. the brighter the concept, the more often it appears in the text).
- The brightness of links relate to how often the two connected concepts co-occur closely within the text.
- Nearness in the map indicates that two concepts appear in similar conceptual contexts (i.e. they co-occur with similar other concepts)

**How to use the Concept Map:**

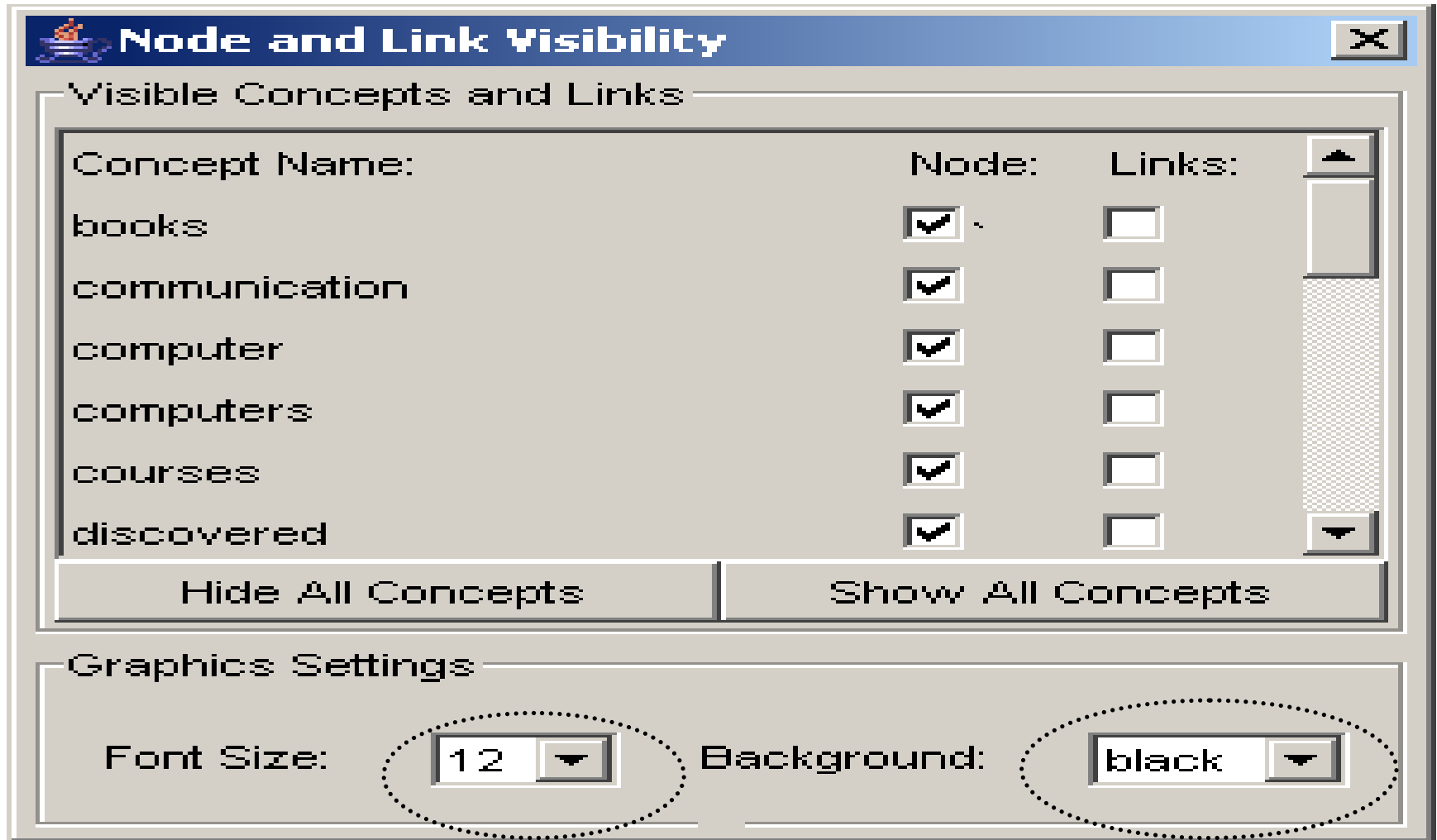
- Left-Click on a concept to reveal its links.
- Left-Click on a vacant position to hide all the visible links.
- Drag on the map to scroll (Right-Click to centre it again).
- <Shift>-Click to zoom in and <Ctrl>-Click to zoom out.

Copy to Clipboard

# Concept map instructions

- What the Concept Map means:
  - The brightness of a concept is related to its frequency (i.e. the brighter the concept, the more often it appears in the text).
  - The brightness of links relate to how often the two connected concepts co-occur closely within the text.
  - Nearness in the map indicates that two concepts appear in similar conceptual contexts (i.e. they co-occur with similar other concepts)


# Changing concept map font and background



# Concept map after changing settings

The screenshot displays the Leximancer Map application window. The main area shows a concept map on a grid with various terms in colored circles: 'communication' (green), 'found' (blue), 'people' (pink), 'information' (orange), 'computers' (yellow), 'online' (cyan), 'internet' (light blue), 'time' (purple), 'computer' (purple), 'e-mail' (blue), 'courses' (blue), 'interest' (red), and 'know' (red). The interface includes a 'Concept List' dropdown set to 'books', 'Iterations = 1000', and a 'Settings' button. Below the map are controls for 'Learn', 'Stop', 'Reset', and 'Max', along with sliders for '# of Points: 50%', 'Theme Size: 33%', and 'Rotation: 0'. The 'map type' is set to 'Linear'. A 'Map Instructions' panel on the right provides a welcome message and directions.

**Map Instructions**

Powered by   
**Leximancer**

Welcome to the Text Explorer. This system allows you to explore a concept map of a document collection.

**Directions**

**What the Concept Map means:**

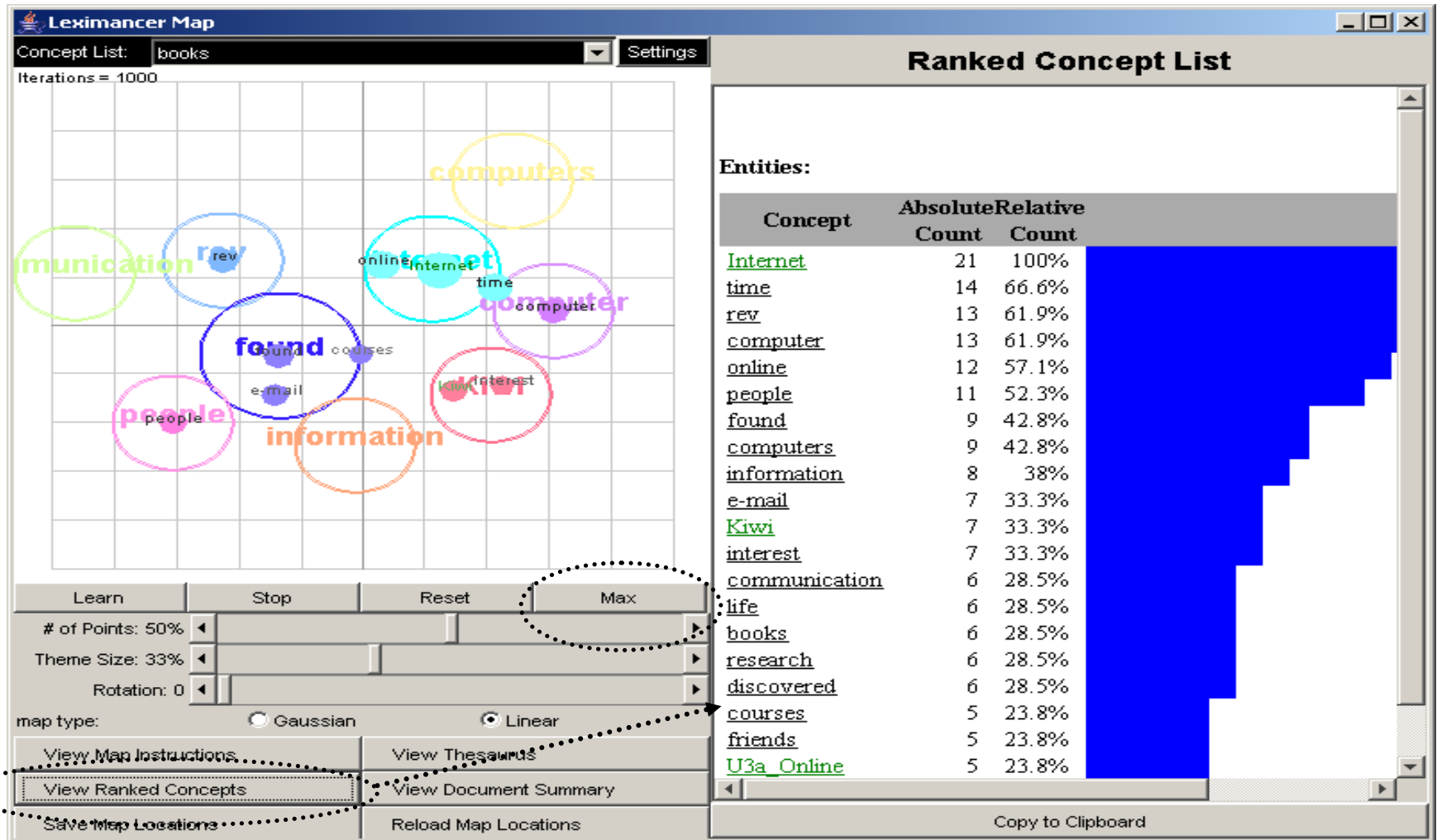
- The brightness of a concept is related to its frequency (i.e. the brighter the concept, the more often it appears in the text).
- The brightness of links relate to how often the two connected concepts co-occur closely within the text.
- Nearness in the map indicates that two concepts appear in similar conceptual contexts (i.e. they co-occur with similar other concepts)

**How to use the Concept Map:**

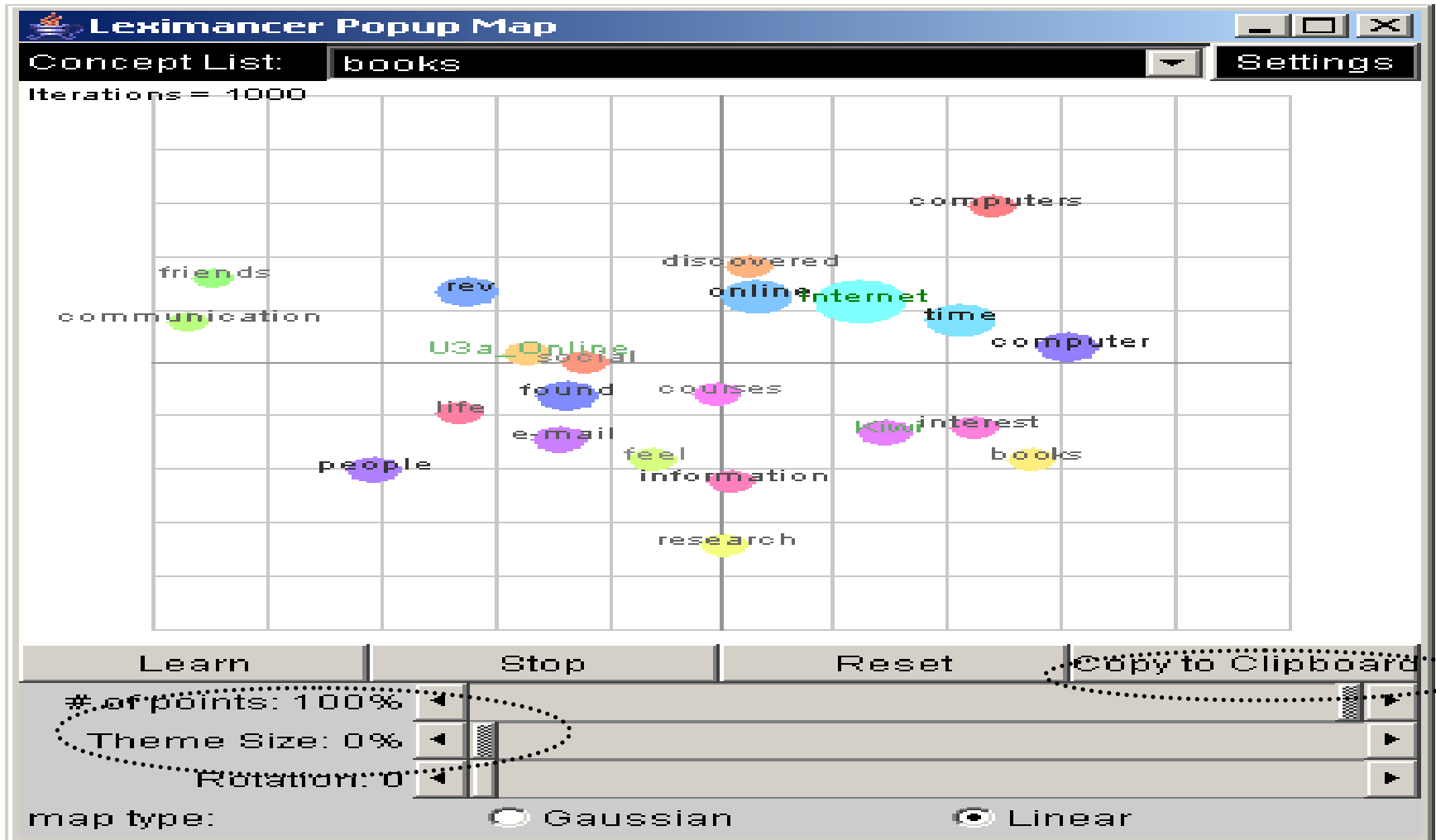
- Left-Click on a concept to reveal its links.
- Left-Click on a vacant position to hide all the visible links.
- Drag on the map to scroll (Right-Click to centre it again).
- <Shift>-Click to zoom in and <Ctrl>-Click to zoom out.

Copy to Clipboard

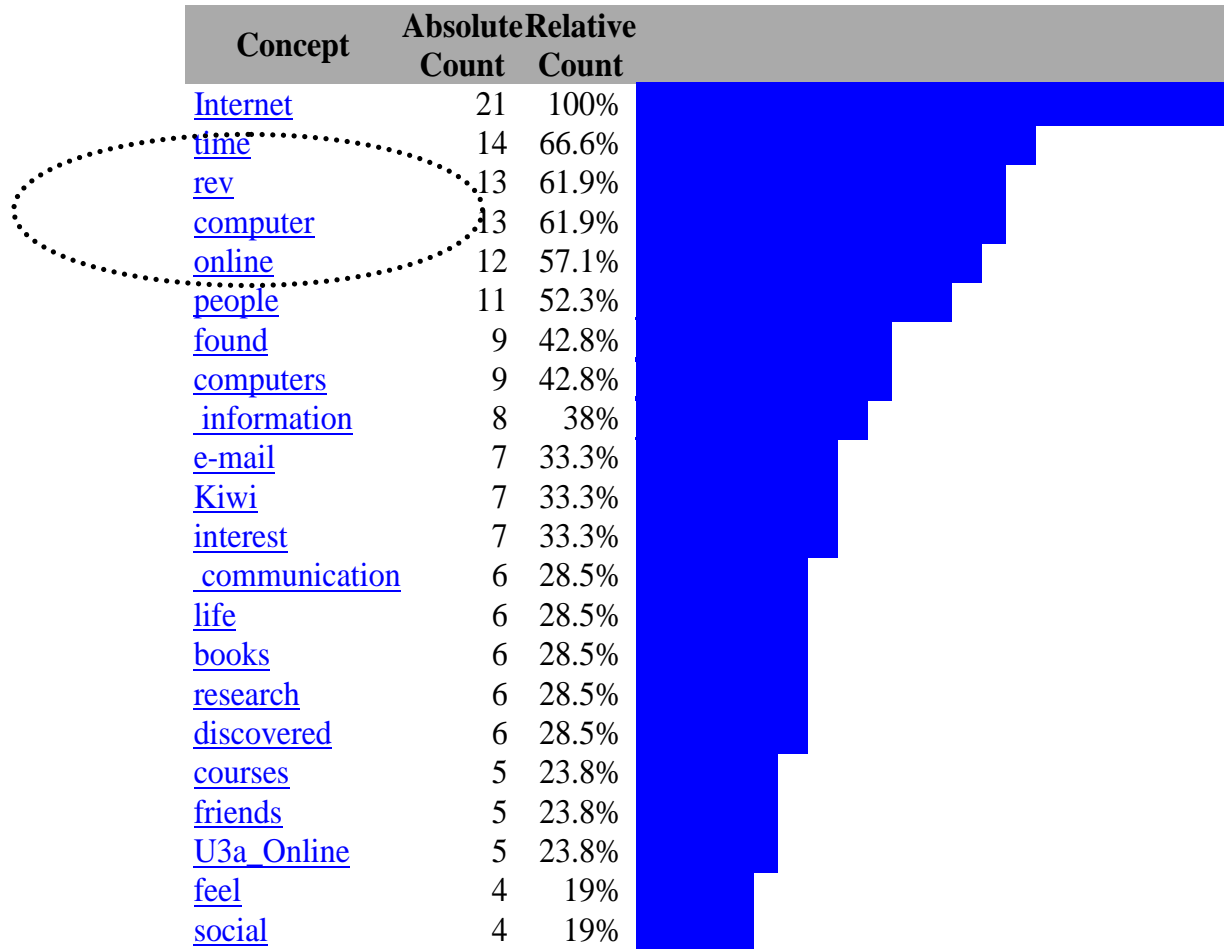
# Ranked concept map



Clicking maximise produces pop-up map that can be copied to clipboard (and Word); Here the number of points have been set at 100% and the theme size at 0%

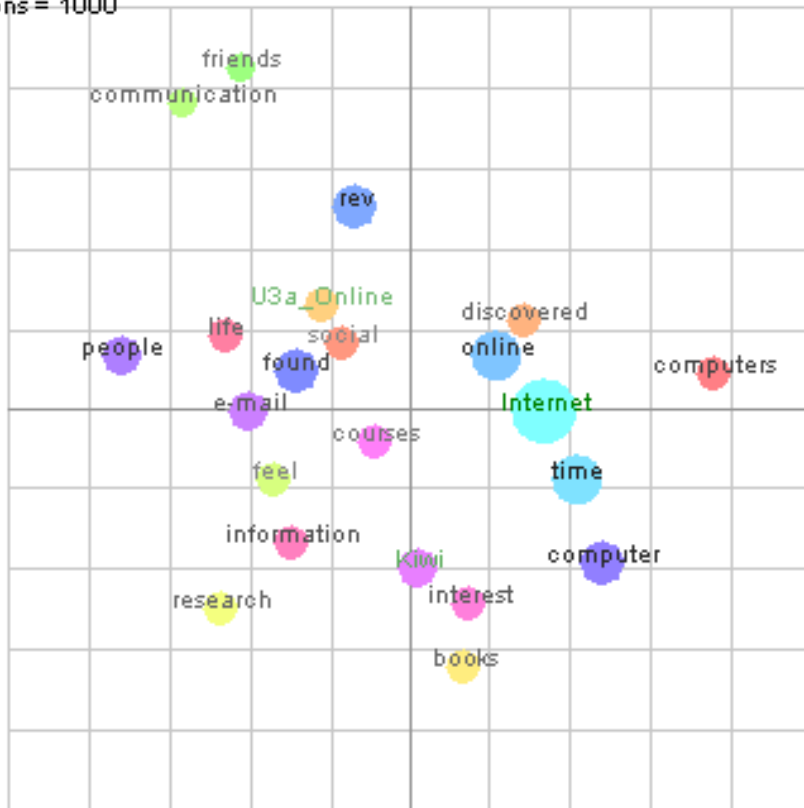


# Ranked concept map provides information about frequency of use of terms (concepts)

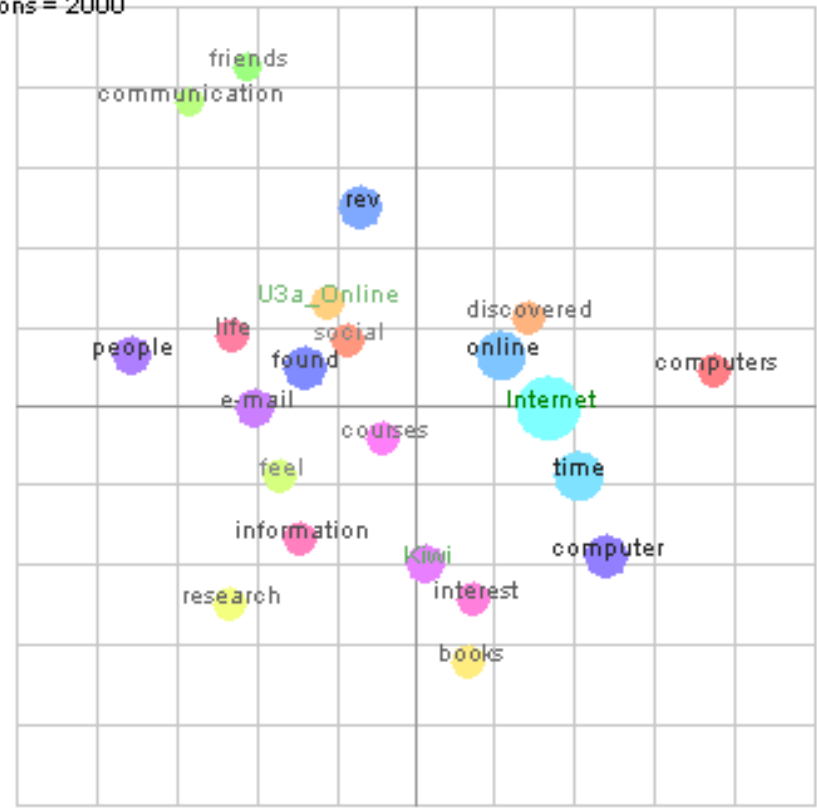


# Popup map with most frequent term (Internet) rotated to align with horizontal axis after 1000 & 2000 iterations

Iterations = 1000

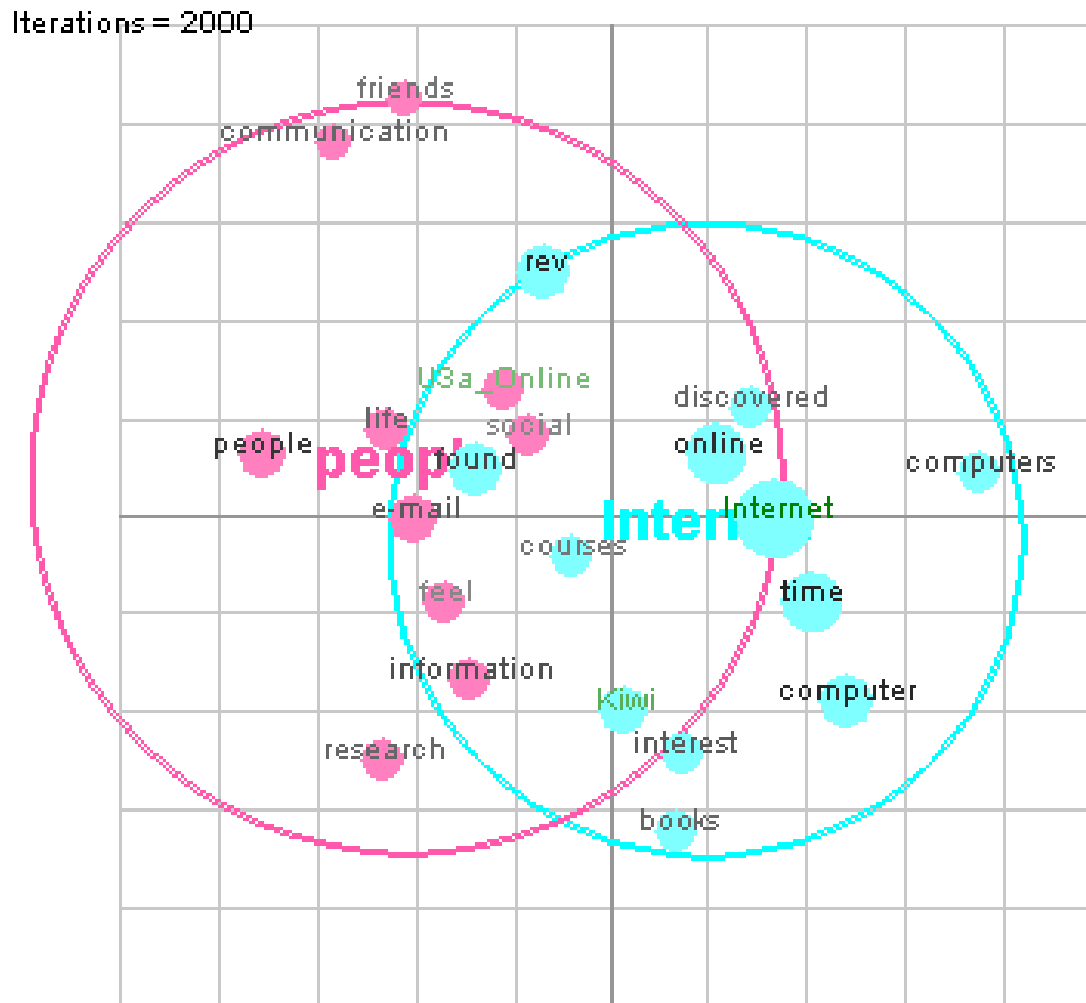


Iterations = 2000



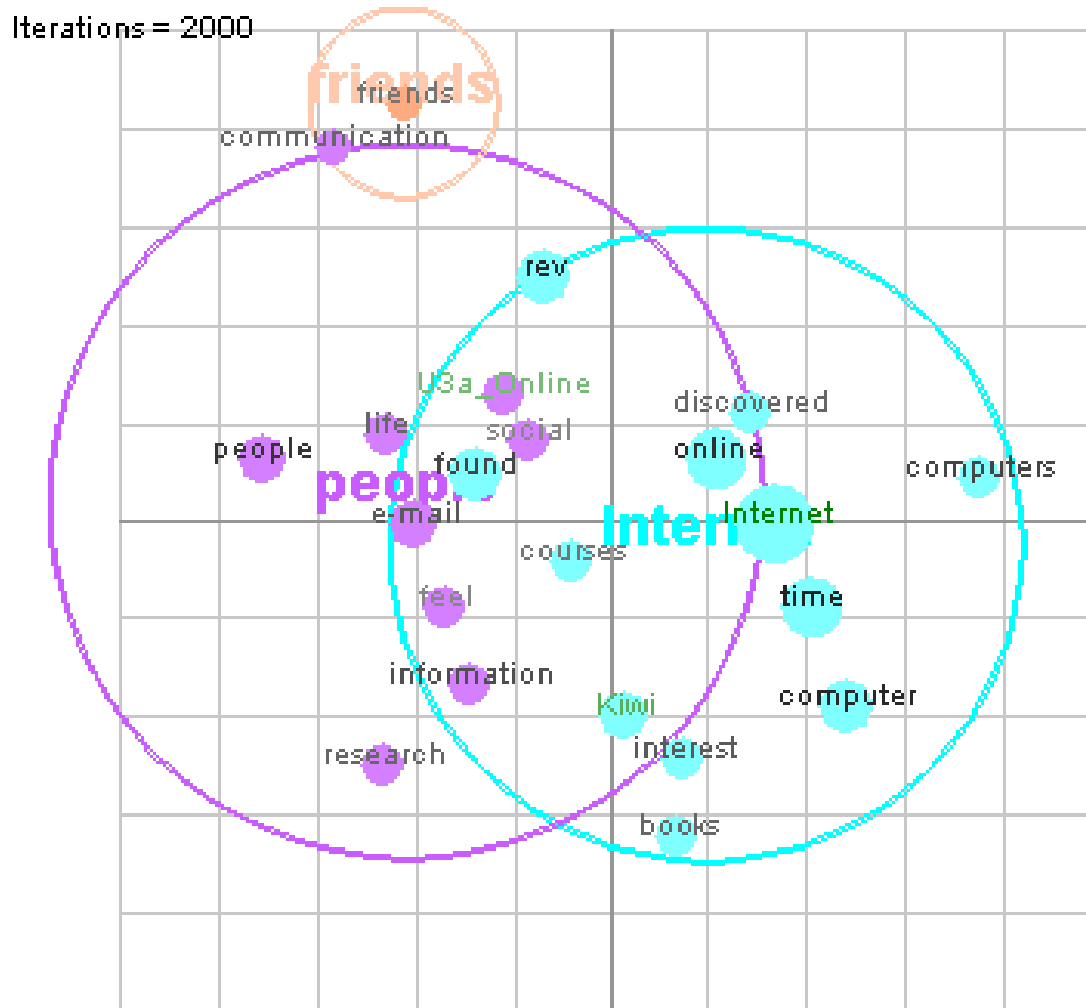


# Popup map with adjusted theme size

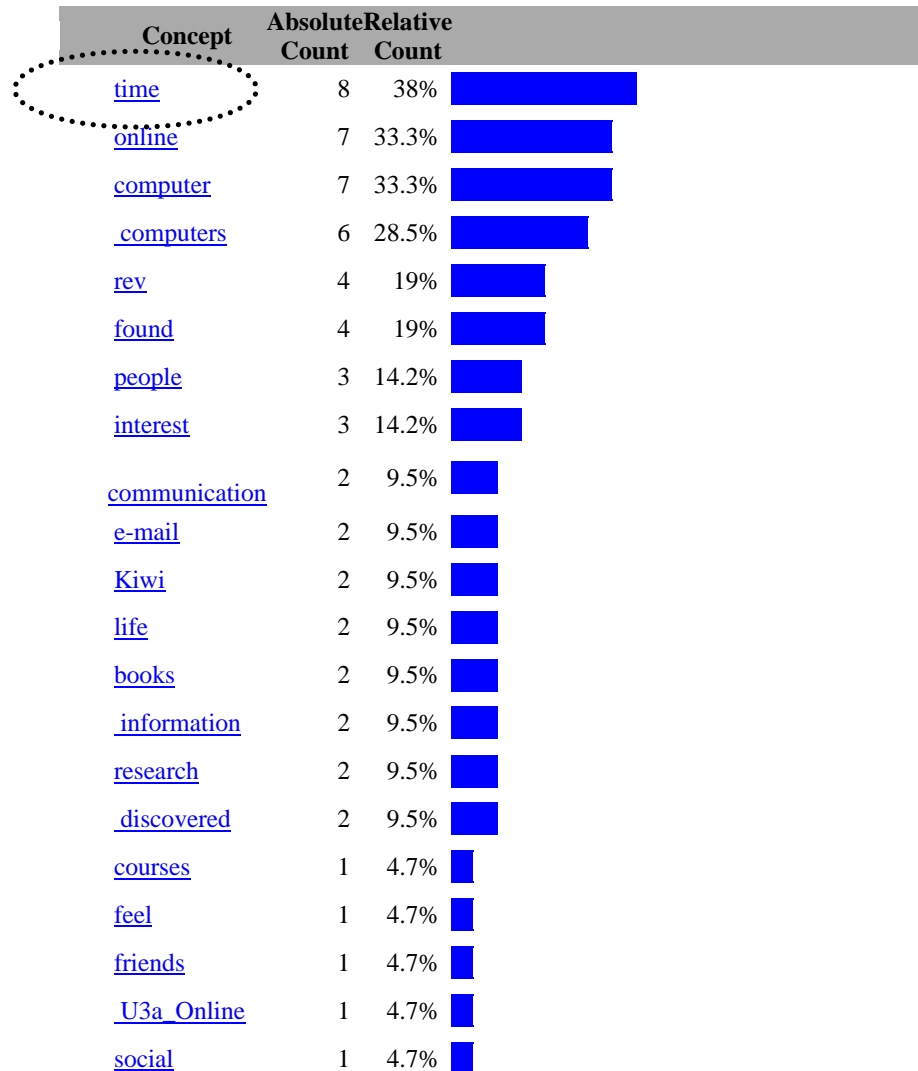


First steps

# Popup map with adjusted theme size

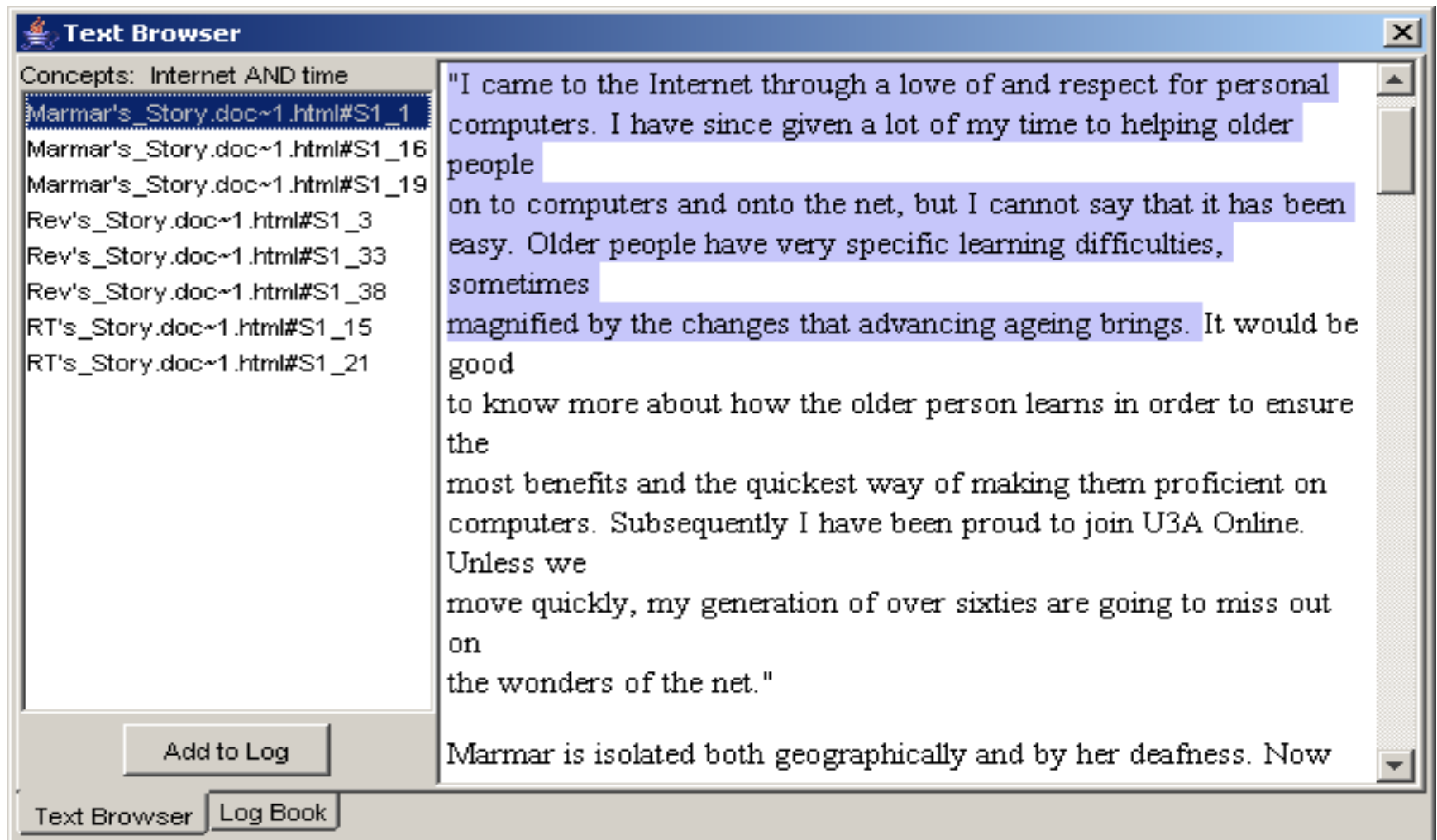


# Terms associated with Internet



First steps

# Text associated with Internet & time (first example)



The screenshot shows a window titled "Text Browser" with a search filter "Concepts: Internet AND time". A list of documents is shown on the left, with "Marmar's\_Story.doc~1.html#S1\_1" selected. The main area displays a text block with several lines highlighted in blue. At the bottom, there are buttons for "Add to Log", "Text Browser", and "Log Book".

Concepts: Internet AND time

- Marmar's\_Story.doc~1.html#S1\_1
- Marmar's\_Story.doc~1.html#S1\_16
- Marmar's\_Story.doc~1.html#S1\_19
- Rev's\_Story.doc~1.html#S1\_3
- Rev's\_Story.doc~1.html#S1\_33
- Rev's\_Story.doc~1.html#S1\_38
- RT's\_Story.doc~1.html#S1\_15
- RT's\_Story.doc~1.html#S1\_21

"I came to the Internet through a love of and respect for personal computers. I have since given a lot of my time to helping older people on to computers and onto the net, but I cannot say that it has been easy. Older people have very specific learning difficulties, sometimes magnified by the changes that advancing ageing brings. It would be good to know more about how the older person learns in order to ensure the most benefits and the quickest way of making them proficient on computers. Subsequently I have been proud to join U3A Online. Unless we move quickly, my generation of over sixties are going to miss out on the wonders of the net."

Marmar is isolated both geographically and by her deafness. Now

Add to Log

Text Browser Log Book

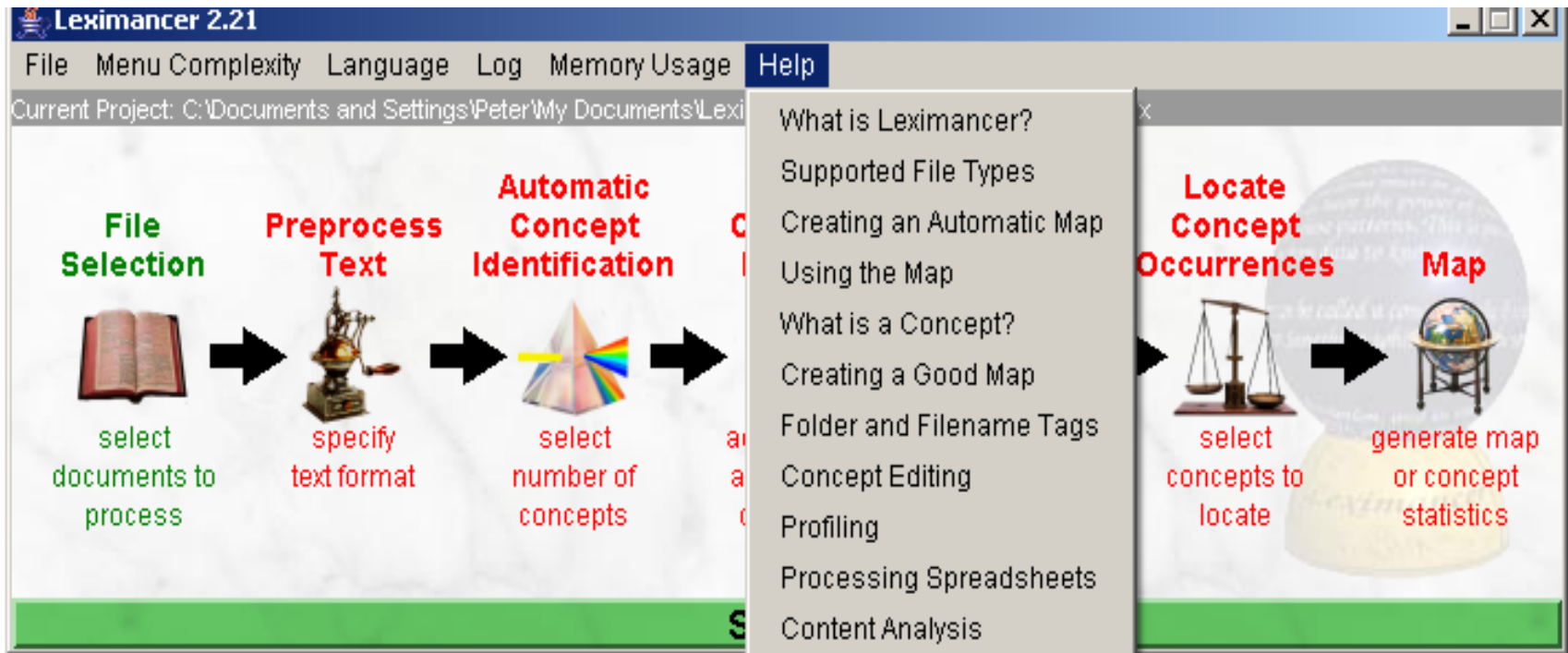
# Internet & time related text can be copied into logbook

The screenshot shows a window titled "Text Browser" with a table containing the following data:

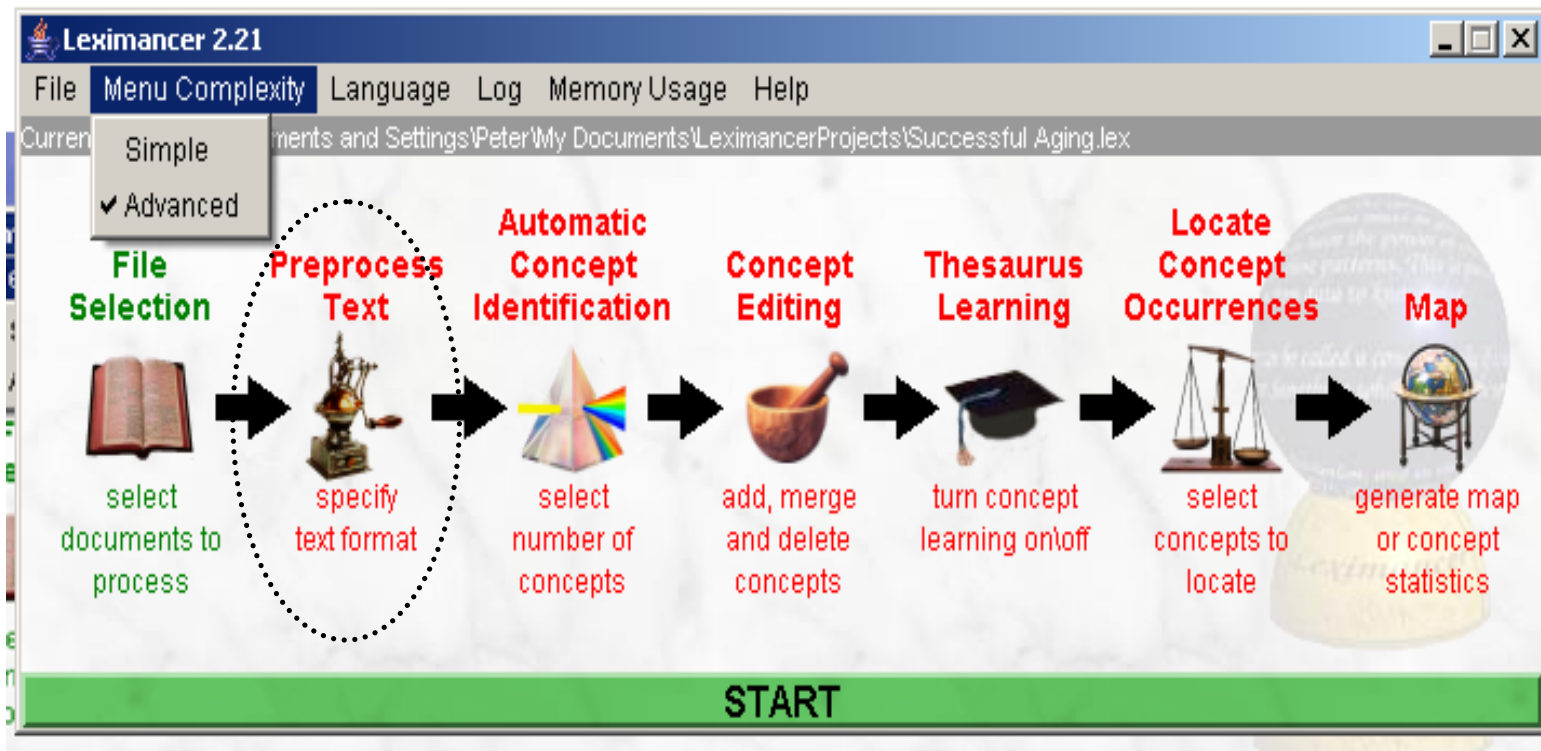
Location	Text	Comment
Marmar's_Story.doc~1.html#S1_1	"I came to the Internet through a love of and respect for personal computers. I have since given a lot of my time to helping older people on to computers and onto the net, but I cannot say that it has been easy. Older people have very specific learning difficulties, sometimes magnified by the changes that advancing ageing brings.	Concepts: Internet AND time

At the bottom of the window, there is a "Copy to Clipboard" button and a "Log Book" button.

# Help menu



# Advanced menu settings



First steps

# Some advanced menu settings

- Preprocess text
  - Adding folder and file name tags
- Automatic concept identification
  - Varying the number of concepts to be extracted
- Concept editing
  - Merging and removing concepts
- Thesaurus learning
  - Turning off thesaurus learning for small texts
    - (Note: Can allow for paragraph structures in press releases)
    - Note: Can counteract over-connectivity in larger texts
- Concept occurrences
  - Including (required classes) or excluding (kill classes: e.g., interviewer comments) blocks of text
    - (Note: Can allow for sentence structures in plays and poems)
- Map settings
  - Generating concept statistics
    - Notes on types of maps; alternative outputs (co-occurrence spreadsheets)

# Adding folder and file name tags: Text pre-processing

**Text Preprocessing**

**Stopword Removal**

Remove Stop Words:  yes  no ?

Edit Stopword List: Edit List ?

Stopword Pattern: ?

**Performance Options**

File Preparation:  All Files  New Files Only ?

HTML:  Normal  Simplified ?

**Language Options**

Make Folder Tags: do nothing ▼ ?

Identify Names:  yes  no ?

Language Testing: weak ▼ ?

Label Identification: none ▼ ?

**Options for Plain Text Documents**

Sentence Boundaries:  Automatic  New Line ?

Start of Document: ^===== ?

Start of Paragraph: blank line ▼ ?

or: ?

OK Cancel

# Varying text pre-processing to make folder tags (e.g., tag folder & files names)

**Text Preprocessing**

**Stopword Removal**

Remove Stop Words:  yes  no ?

Edit Stopword List: Edit List ?

Stopword Pattern: ?

**Performance Options**

File Preparation:  All Files  New Files Only ?

HTML:  Normal  Simplified ?

**Language Options**

Make Folder Tags: make folder and filename... ?

Identify Names:  yes  no ?

Language Testing: weak ?

Label Identification: none ?

**Options for Plain Text Documents**

Sentence Boundaries:  Automatic  New Line ?

Start of Document: ^===== ?

Start of Paragraph: blank line ?

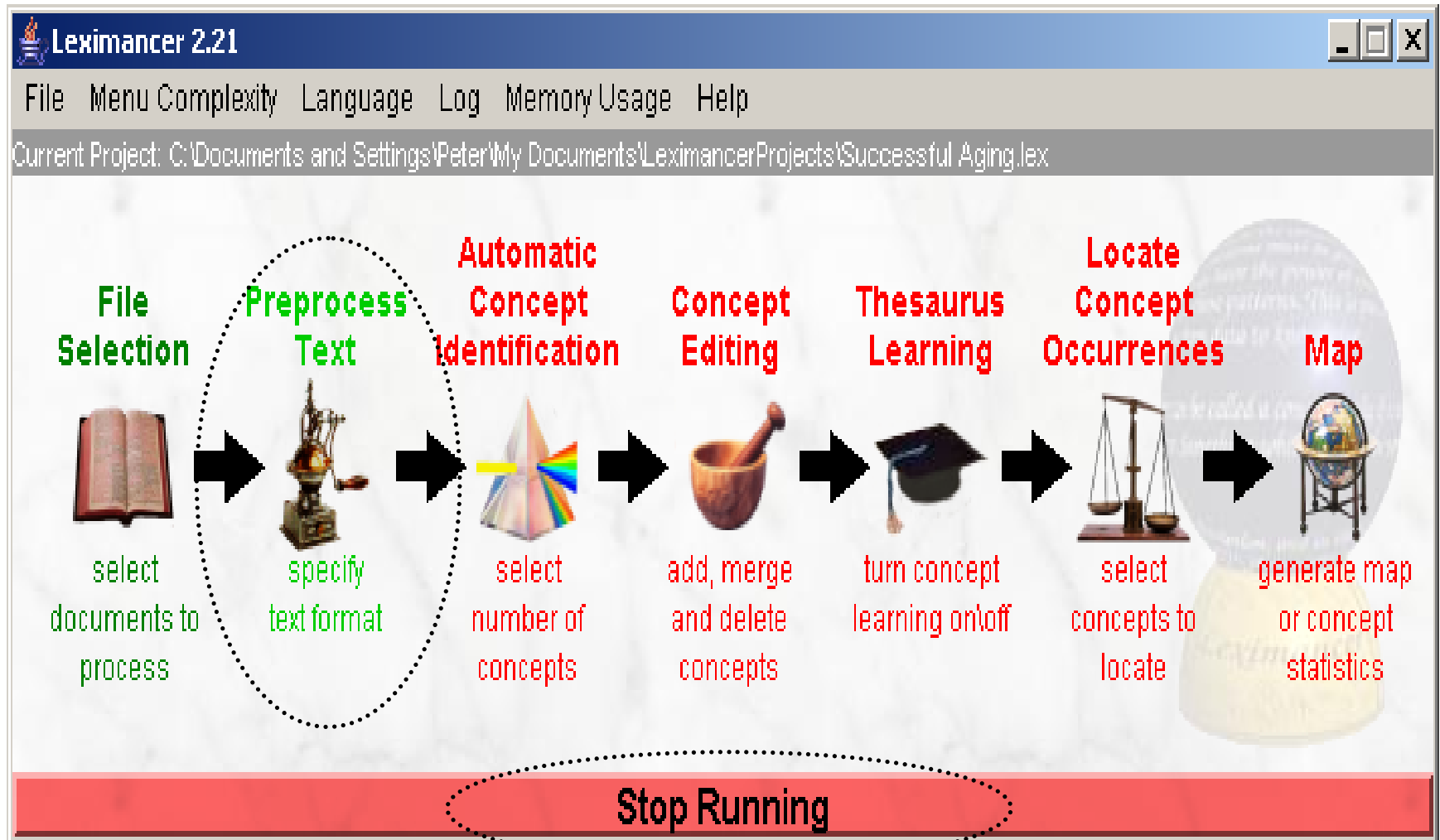
or: ?

OK Cancel

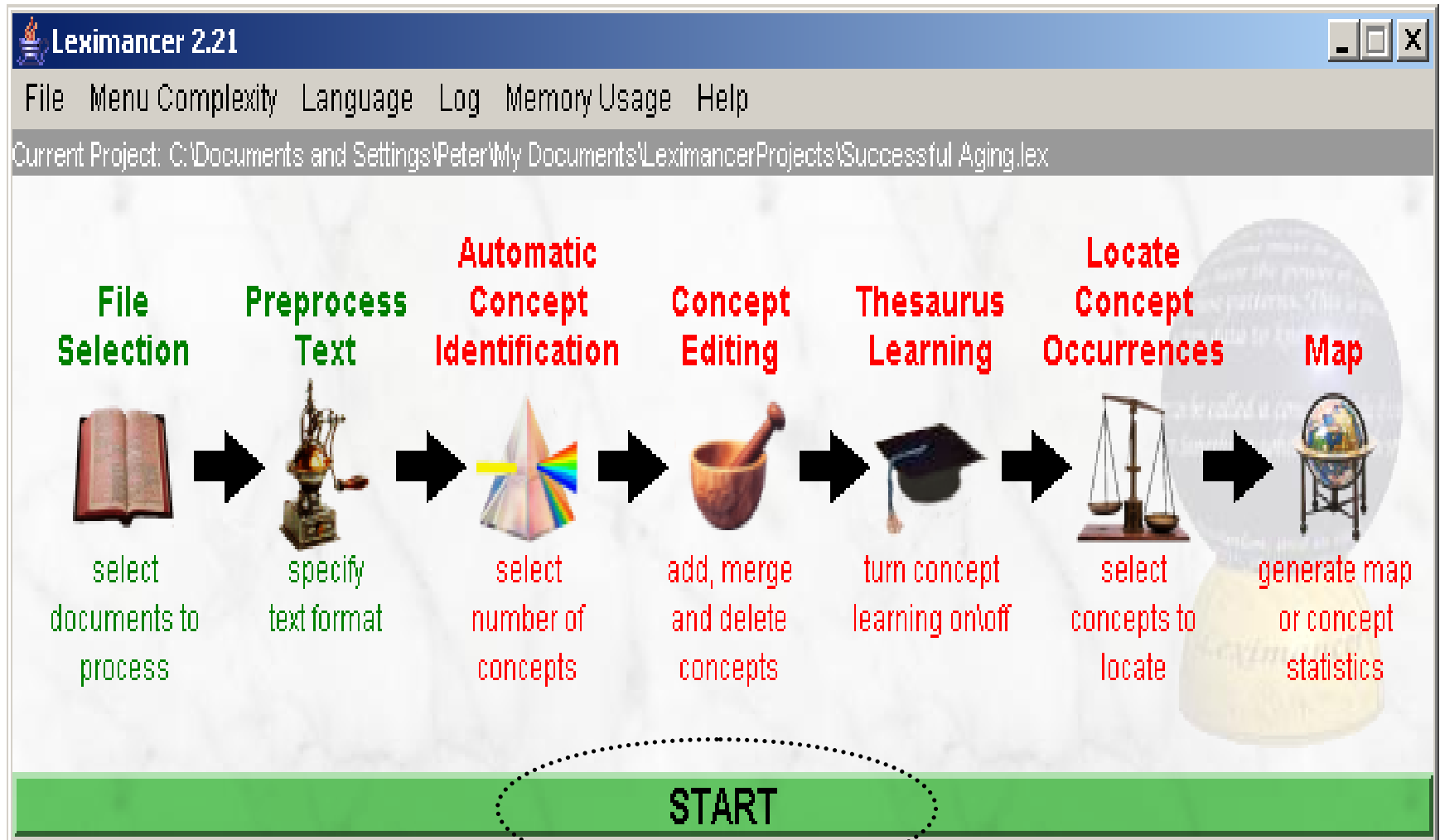
# Folder tags (off | make folder tags | make folder and filename tags)

- This parameter is very important if you are comparing different documents based on their conceptual content.
- This parameter, if set, causes 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.
  - For example, a file called “patient1” inside the folder 'control/' below the data folder would have separate tags `[[control]]` and `[[patient1]]` inserted in each sentence (if folder and filename tags are set).
- These tags will be included as concepts in your map.
- Thus, inspecting the links formed with the other concepts can allow you to compare the content of the various folders.

# Using folder tags



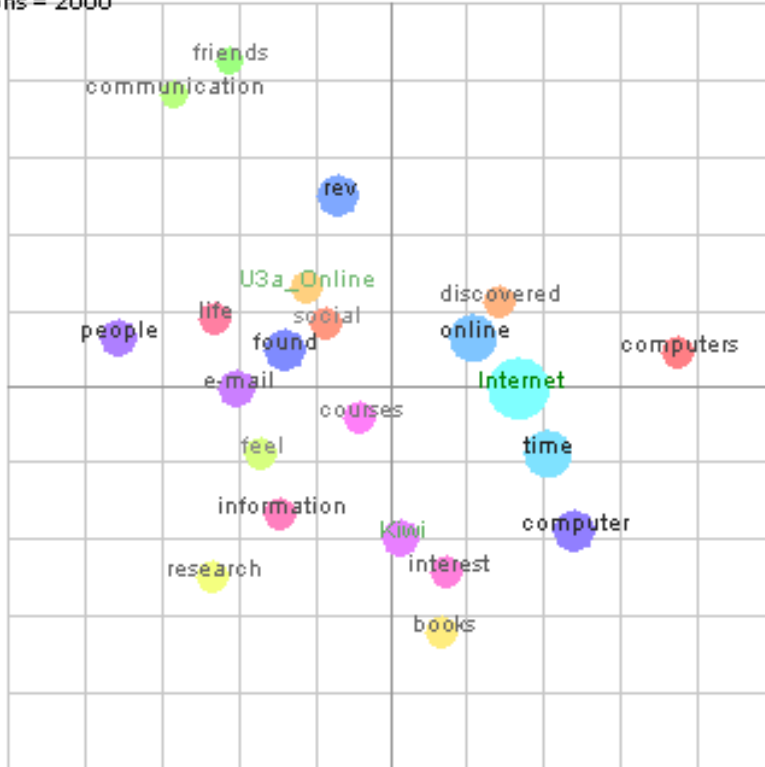
# Using folder tags



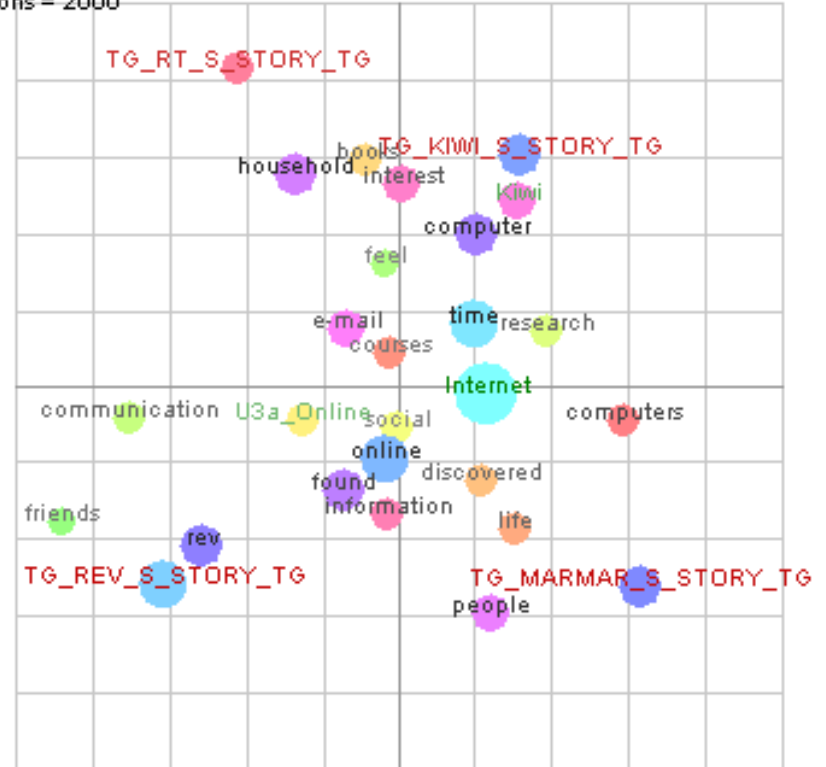
First steps

# Using folder tags

Iterations = 2000

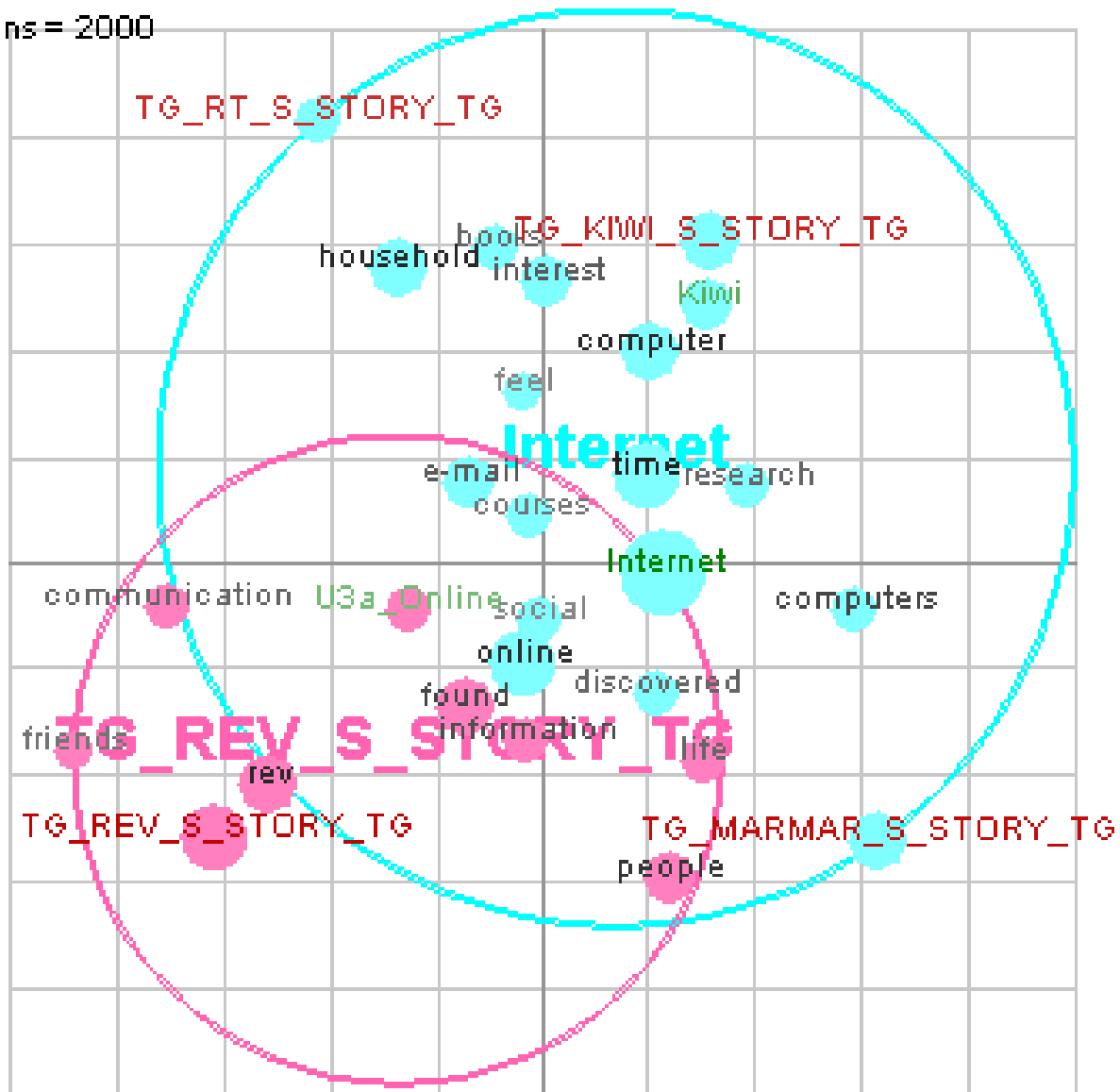


Iterations = 2000



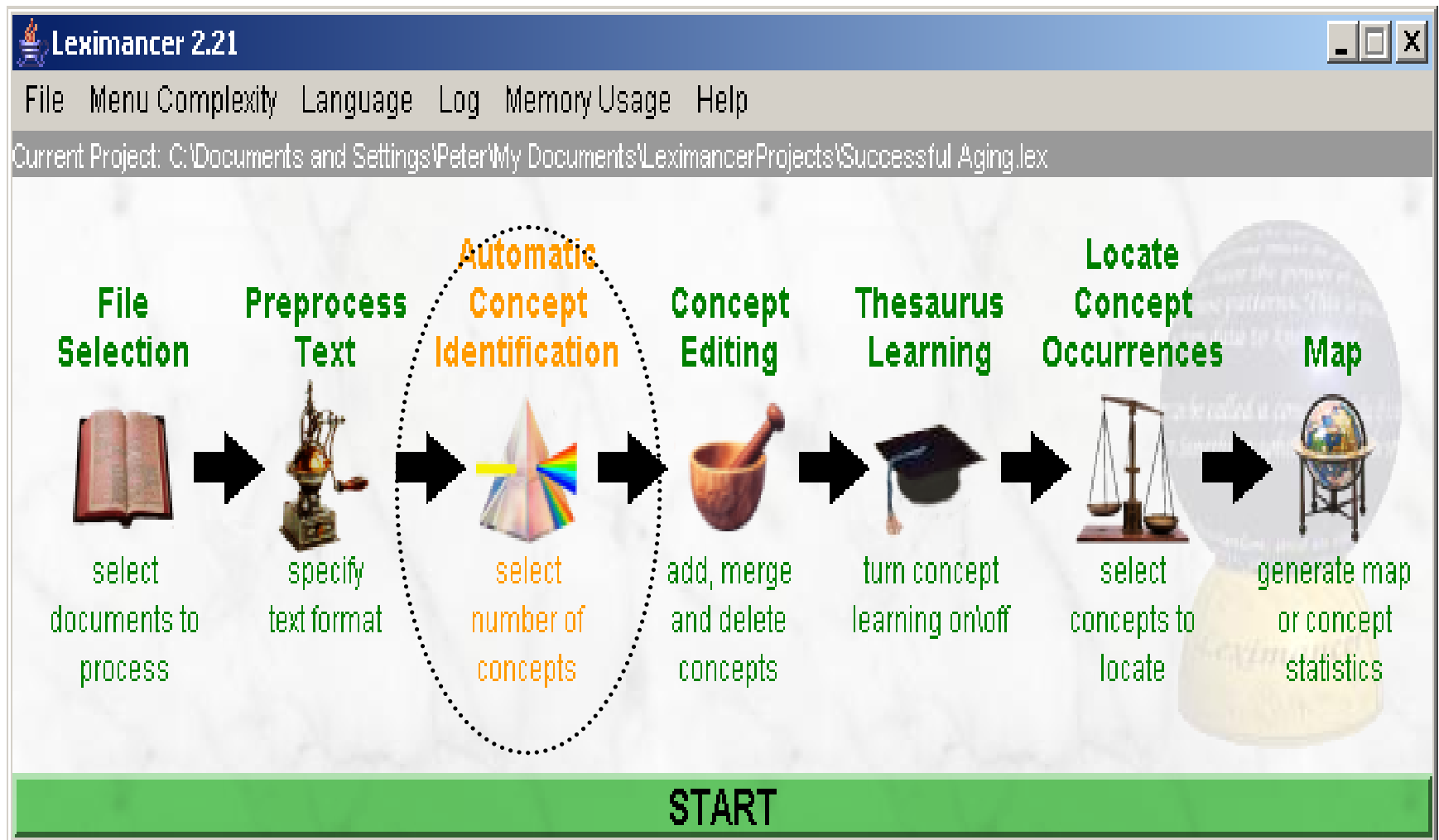
# Using folder tags

Iterations = 2000

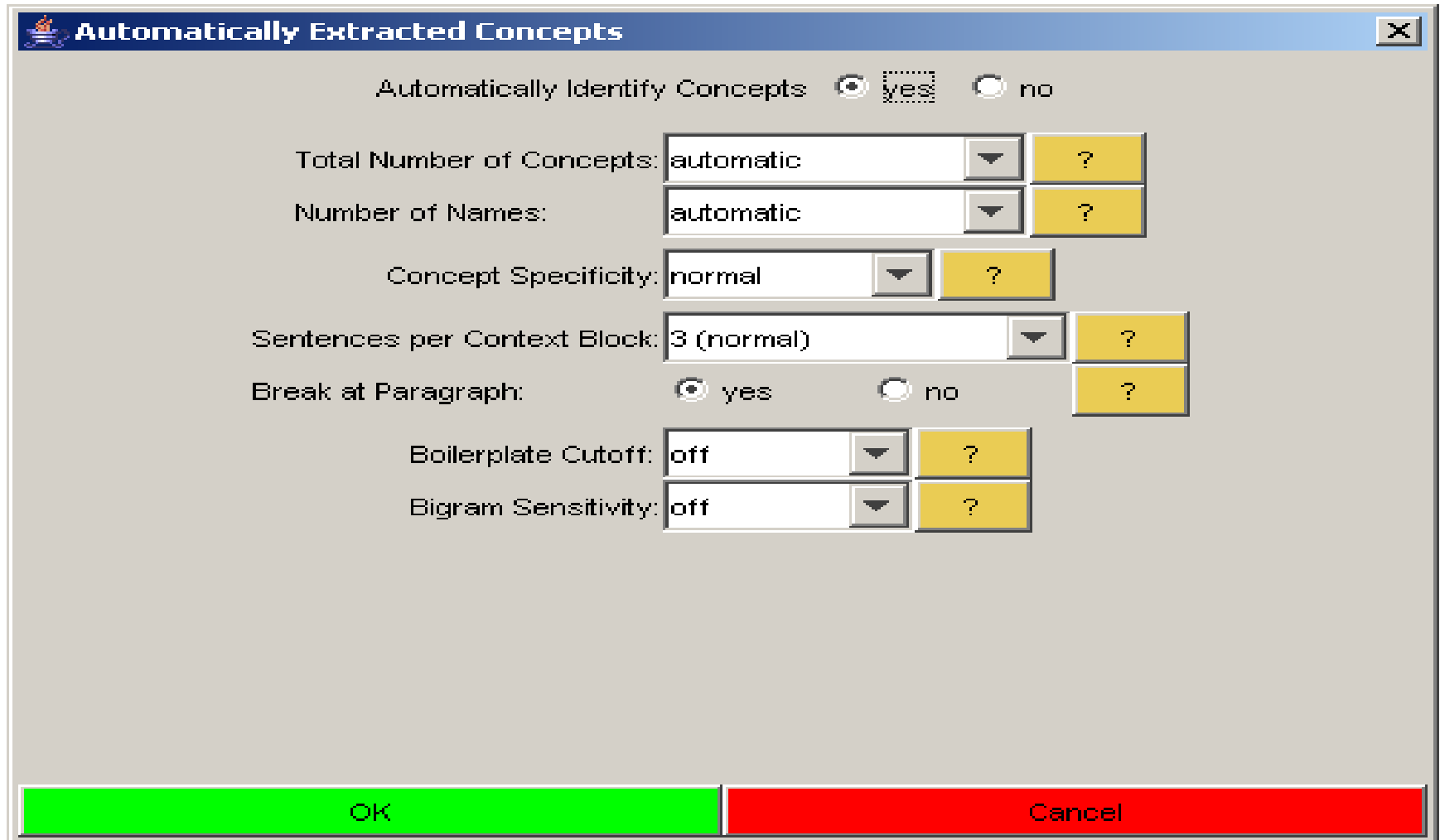




# Setting number of concepts manually via automatic concept identification (after running Preprocess text)



# Automatically extracted concepts



**Automatically Extracted Concepts**

Automatically Identify Concepts  yes  no

Total Number of Concepts: automatic [?] ?

Number of Names: automatic [?] ?

Concept Specificity: normal [?] ?

Sentences per Context Block: 3 (normal) [?] ?

Break at Paragraph:  yes  no [?] ?

Boilerplate Cutoff: off [?] ?

Bigram Sensitivity: off [?] ?

OK Cancel

# Sentences per concept block

- This setting specifies how many sentences are included in each context block.
- This setting is shared by the Thesaurus Learning phase, and can also be set there.
- This should be three (3) in almost all circumstances.

# Setting number of concepts manually

**Automatically Extracted Concepts**

Automatically Identify Concepts  yes  no

Total Number of Concepts: 30

Number of Names: automatic

Concept Specificity: normal

Sentences per Context Block: 3 (normal)

Break at Paragraph:  yes  no

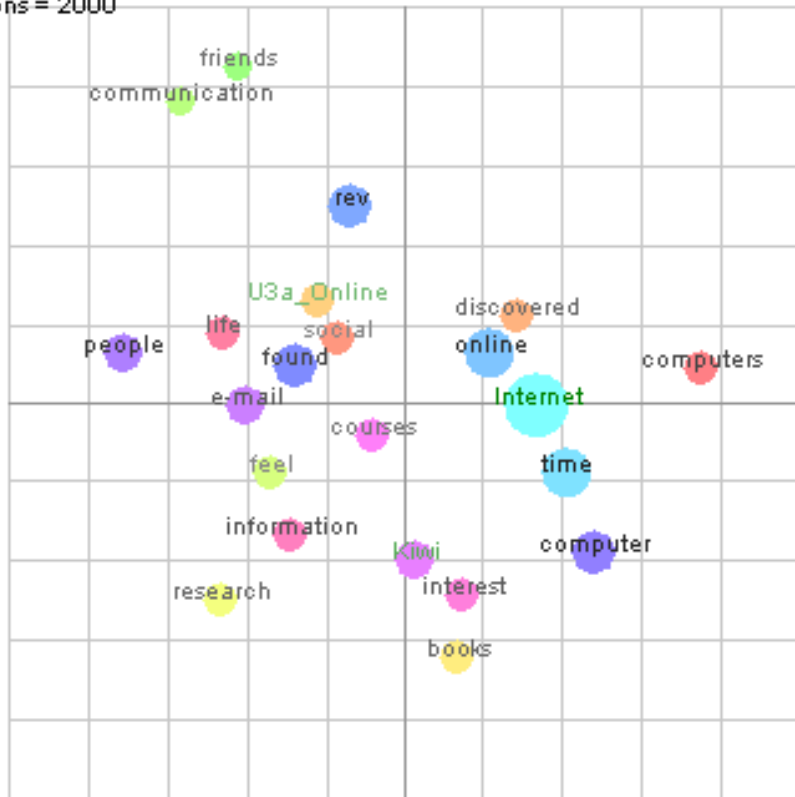
Boilerplate Cutoff: off

Bigram Sensitivity: off

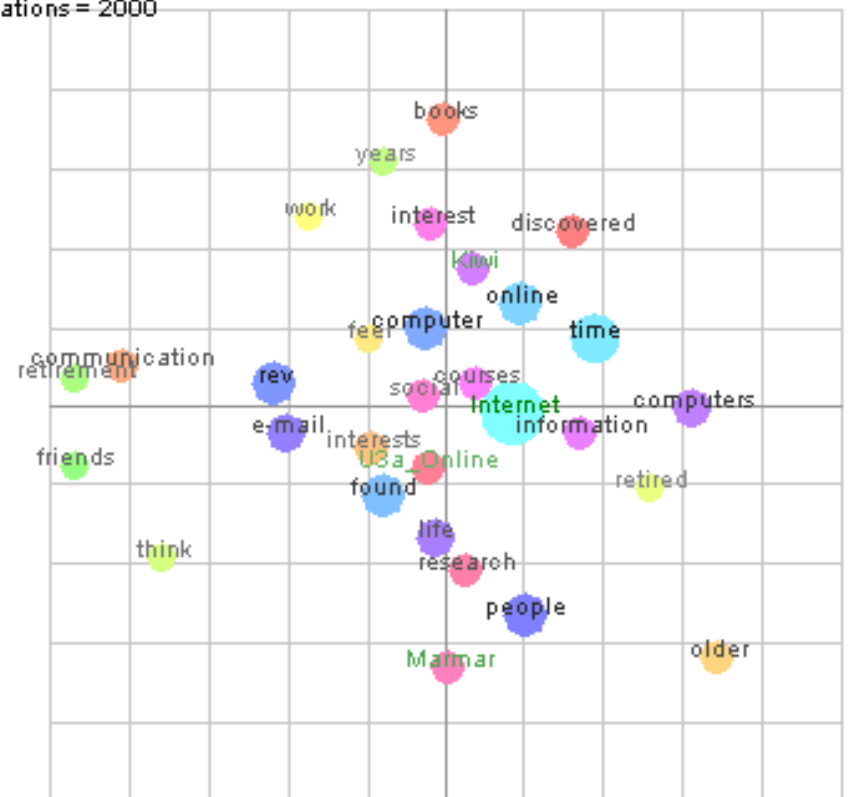
OK Cancel

# Setting number of concepts manually

Iterations = 2000

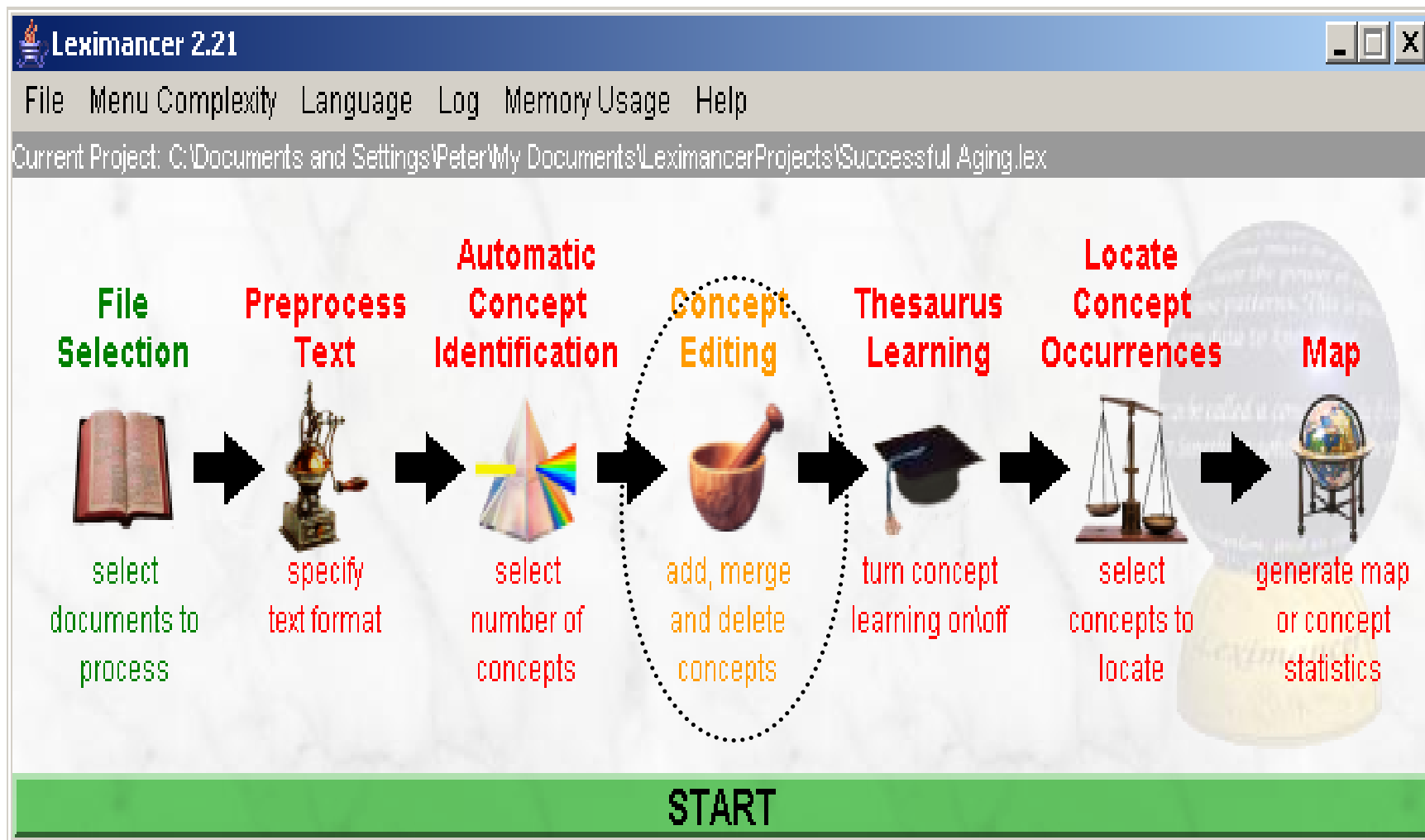


Iterations = 2000



First steps

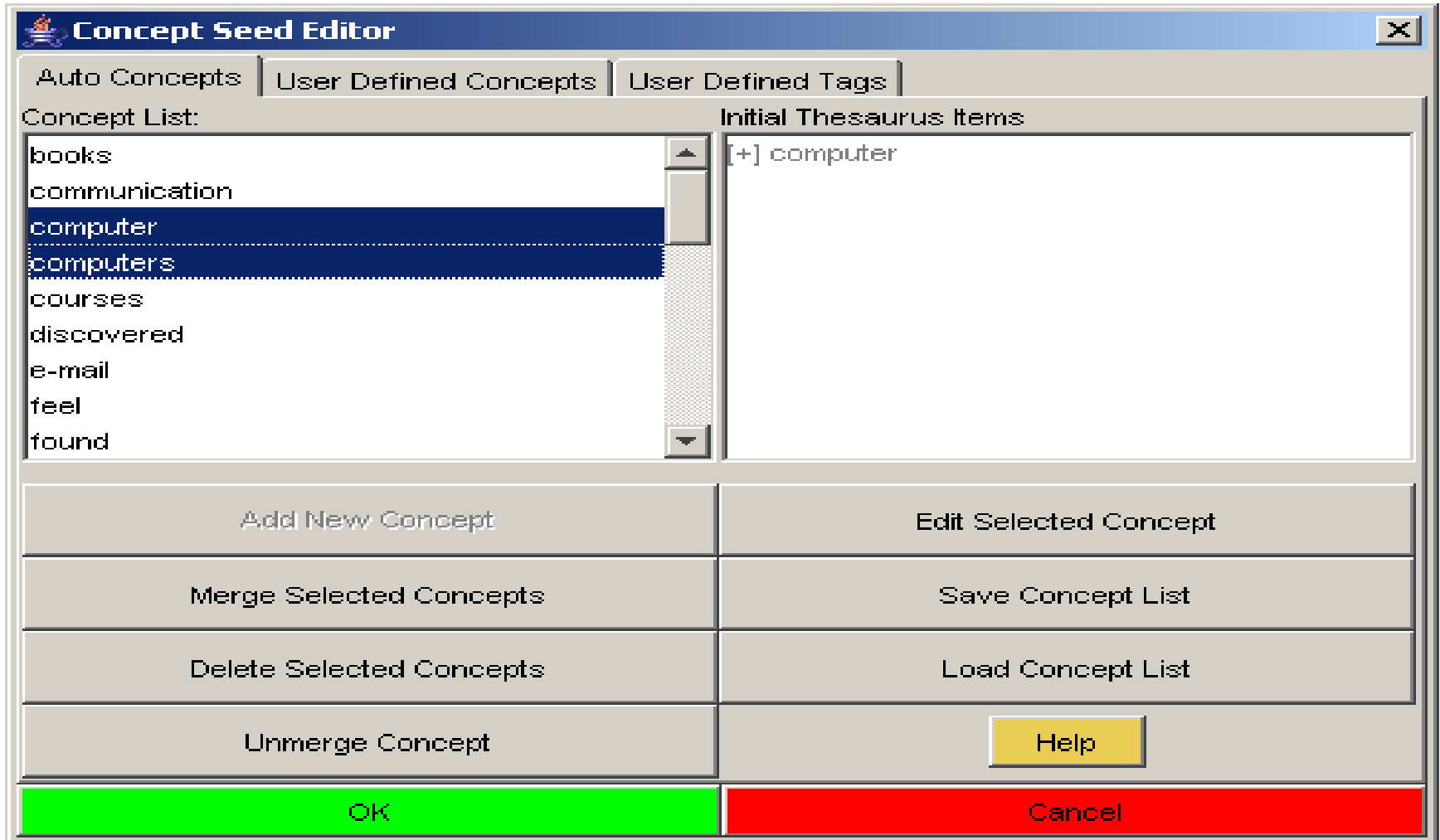
# Merging and deleting concepts via Concept editing (after running automatic concept identification)



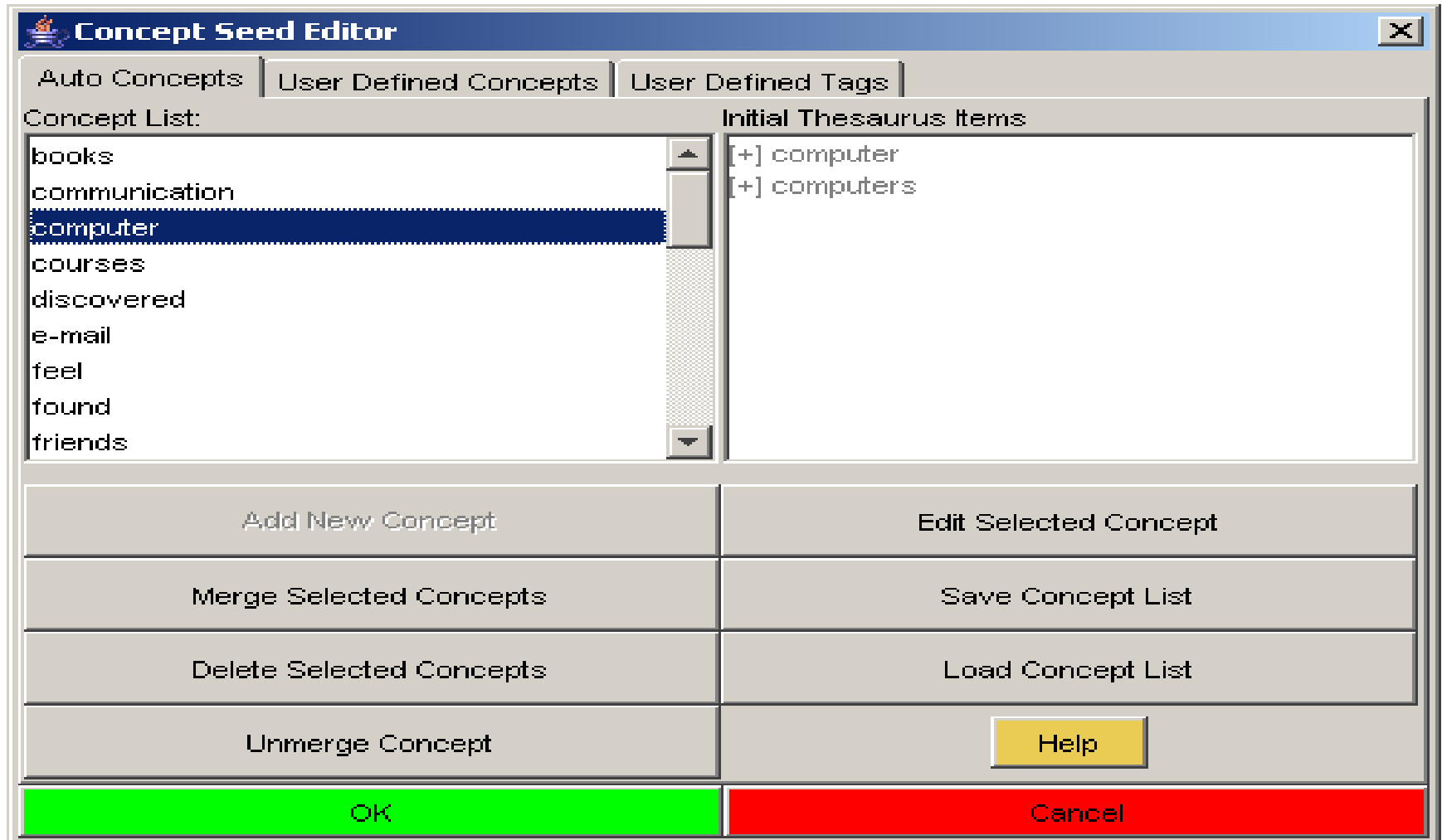
# Hints to creating a useful map

- Run processing up until the automatic concept identification node by single clicking on this node and hitting “Start”
- Open up the Concept Editing dialog (by double-clicking on the concept editing node)
- Remove any concepts that you think are unimportant
- Merge any concepts that you think refer to the same thing
- Close down the editing dialog, and generate a map (i.e. click on Start).

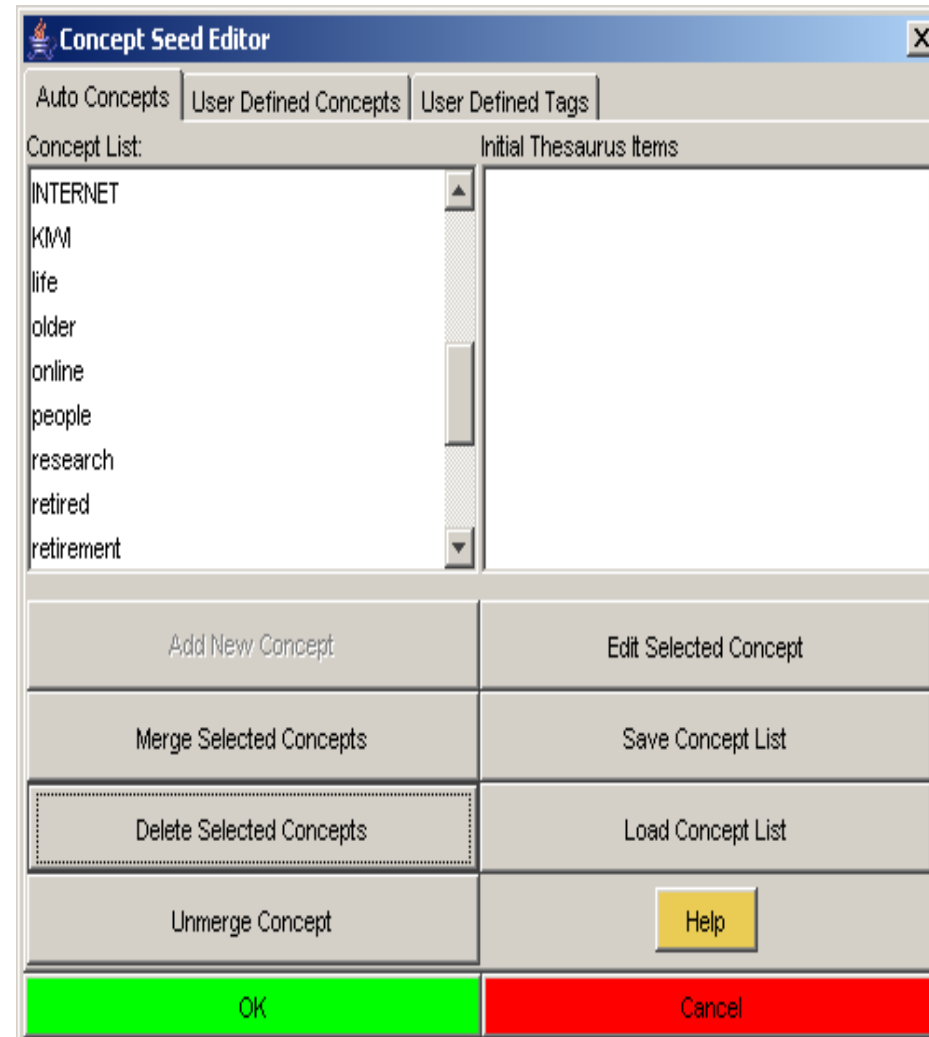
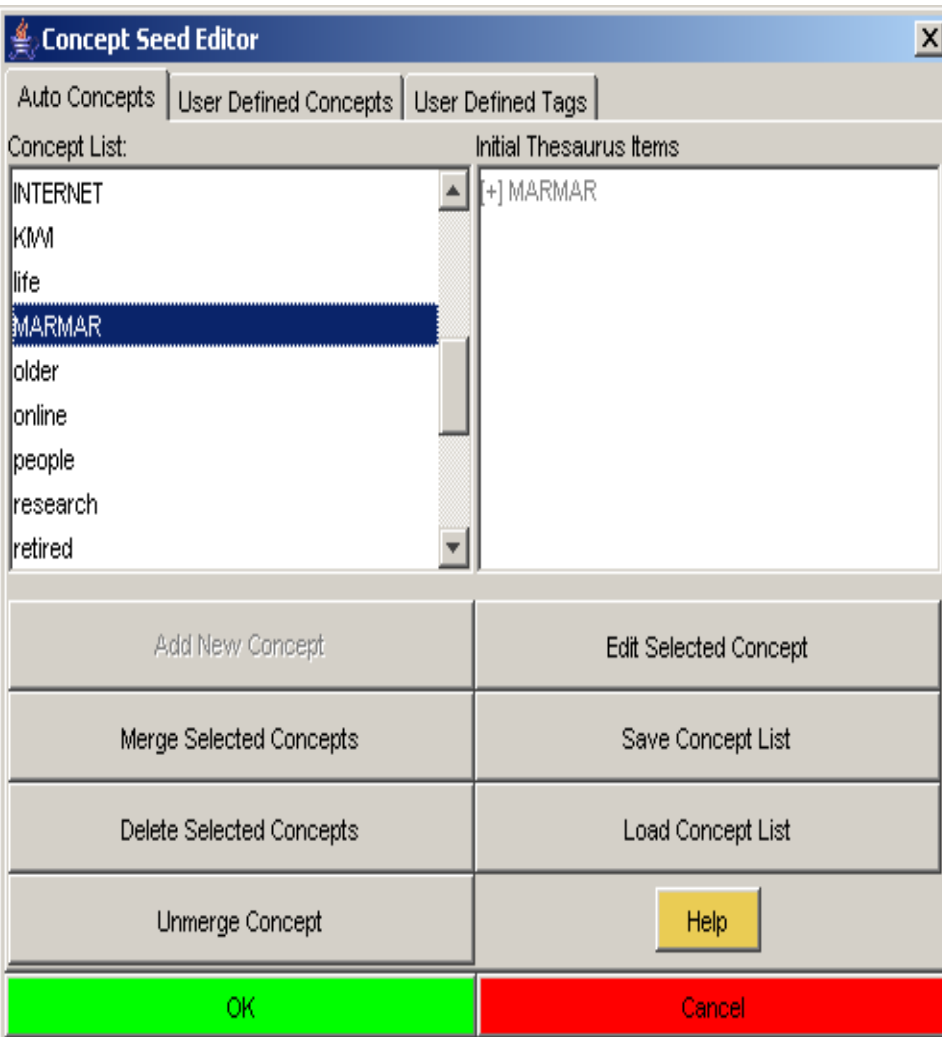
# Concept seed editor



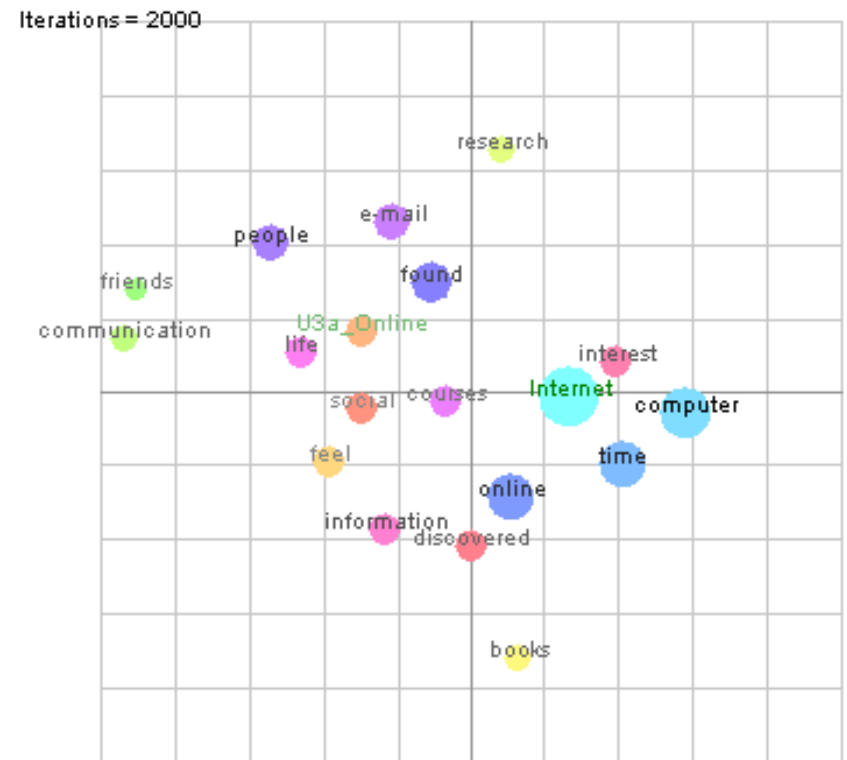
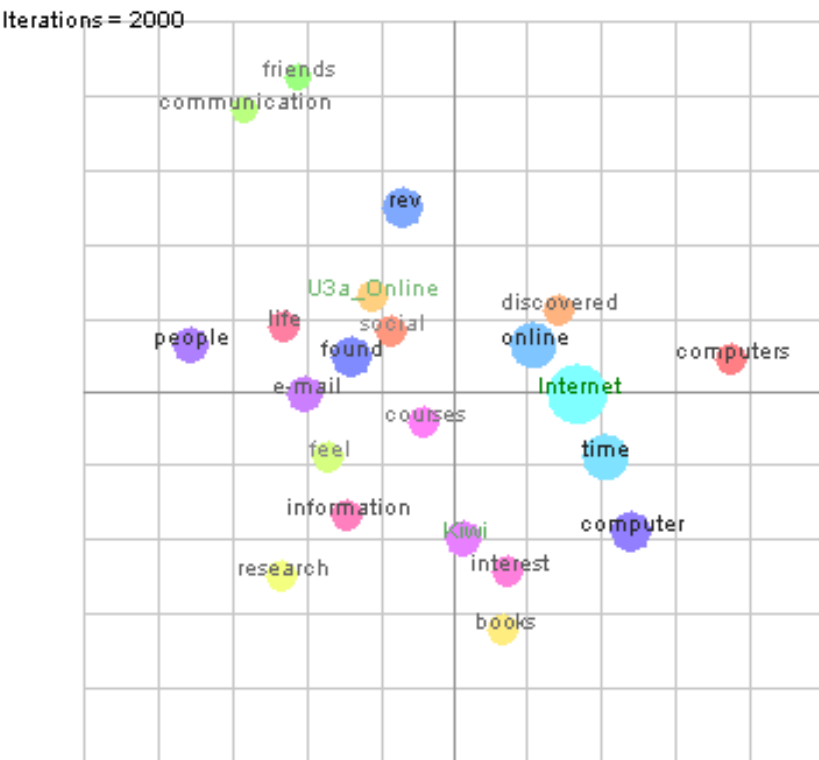
# Merging selected concepts



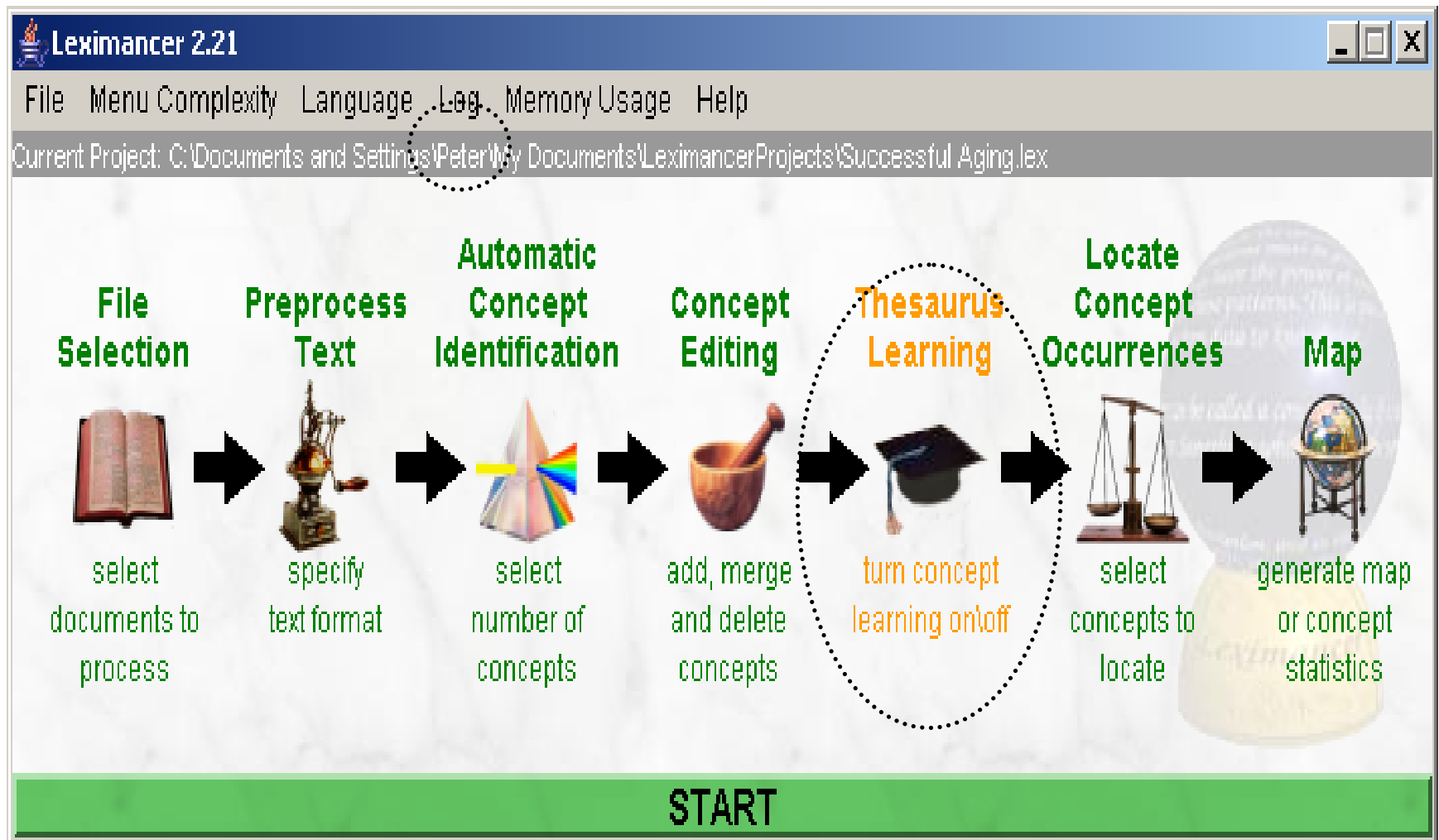
# Deleting selected concepts



# Result of editing concepts (excluding Rev & Marmar, merging interest/interests & computer/computers)



# Turning off thesaurus learning (after running concept editing module)



# Thesaurus learning

Concept Learning Settings

Learn Concept Thesaurus:  yes  no

Learning Threshold: 14 (normal) ?

Sentences per Context Block: 3 (normal) ?

Break at Paragraph:  yes  no ?

Learn Tag Classes:  yes  no ?

Learning Type:  Automatic  Supervised ?

Clear Lexicon:  yes  no ?

Follow-on Learning:  yes  no ?

Sampling: automatic ?

Phrase Separation: 3 ?

**Concept Profiling**

Number to Discover: 0 ?

Themed Discovery:  Concepts in ALL  Concepts in ANY  Concepts in EACH ?

Only Discover Names:  yes  no ?

OK Cancel

- Learn concept thesaurus: Turn off for small texts prevents Leximancer from adding items to concept definitions

# Varying thesaurus learning for press releases

**Concept Learning Settings**

Learn Concept Thesaurus:  yes  no

Learning Threshold: 4 (normal) ?

Sentences per Context Block: 3 (normal) ?

Break at Paragraph:  yes  no ?

Learn Tag Classes:  yes  no ?

Learning Type:  Automatic  Supervised ?

Clear Lexicon:  yes  no ?

Follow-on Learning:  yes  no ?

Sampling: automatic ?

Phrase Separation: 3 ?

**Concept Profiling**

Number to Discover: 0 ?

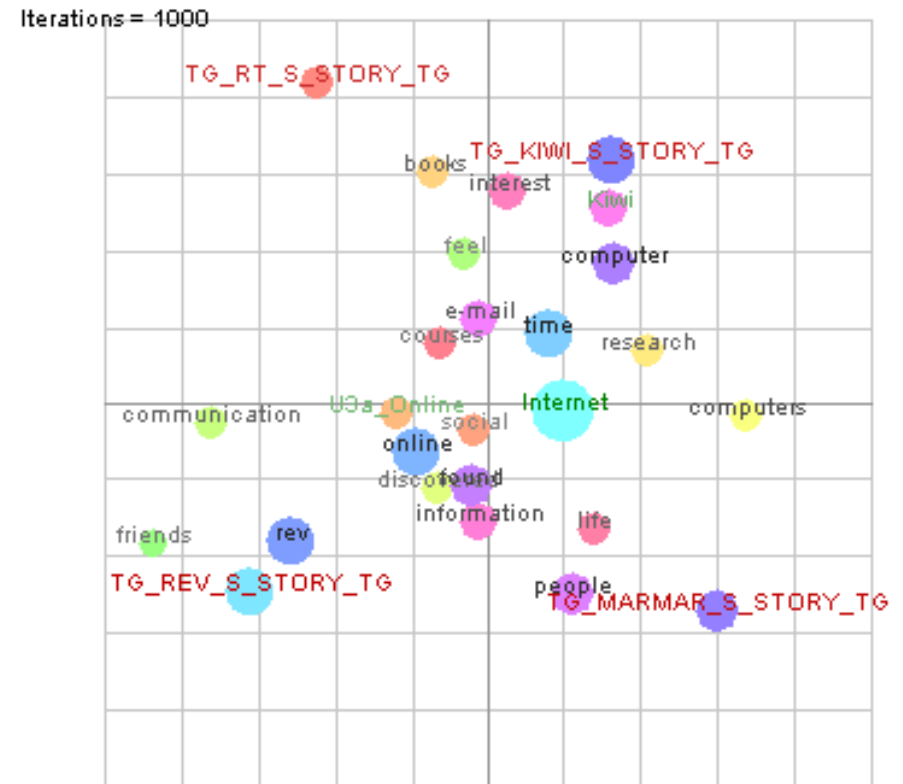
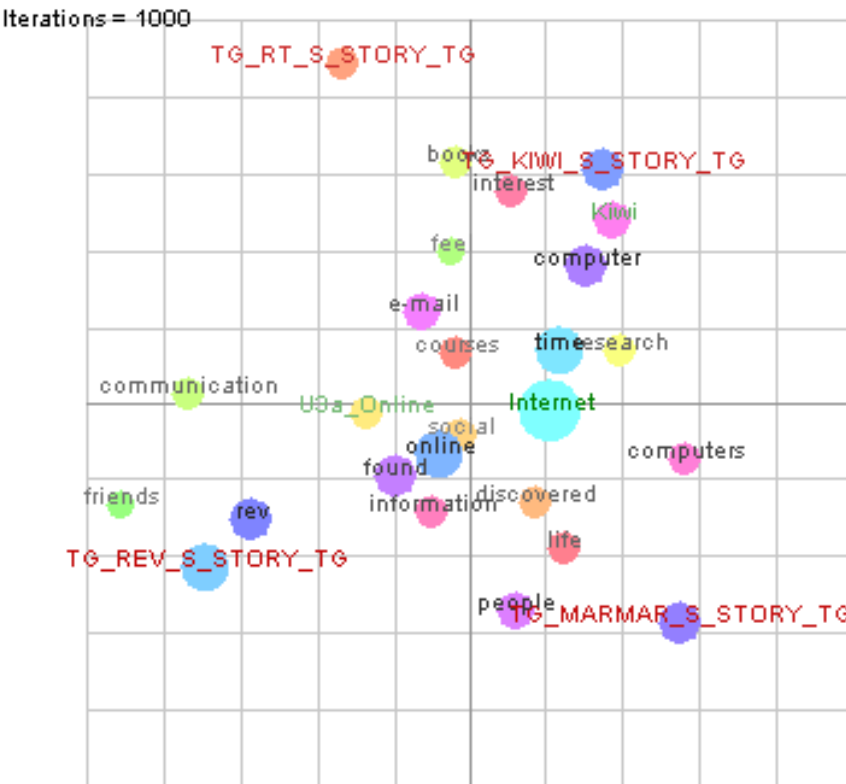
Themed Discovery:  Concepts in ALL  Concepts in ANY  Concepts in EACH ?

Only Discover Names:  yes  no ?

OK Cancel

- Learning threshold controls generality of each learnt concept
  - Increasing generality increases No. of words included in each concept
  - Left-hand panel of thesaurus lists iterations to finish (ideally 6-11)
- Break at paragraphs and scan three sentences except for press releases

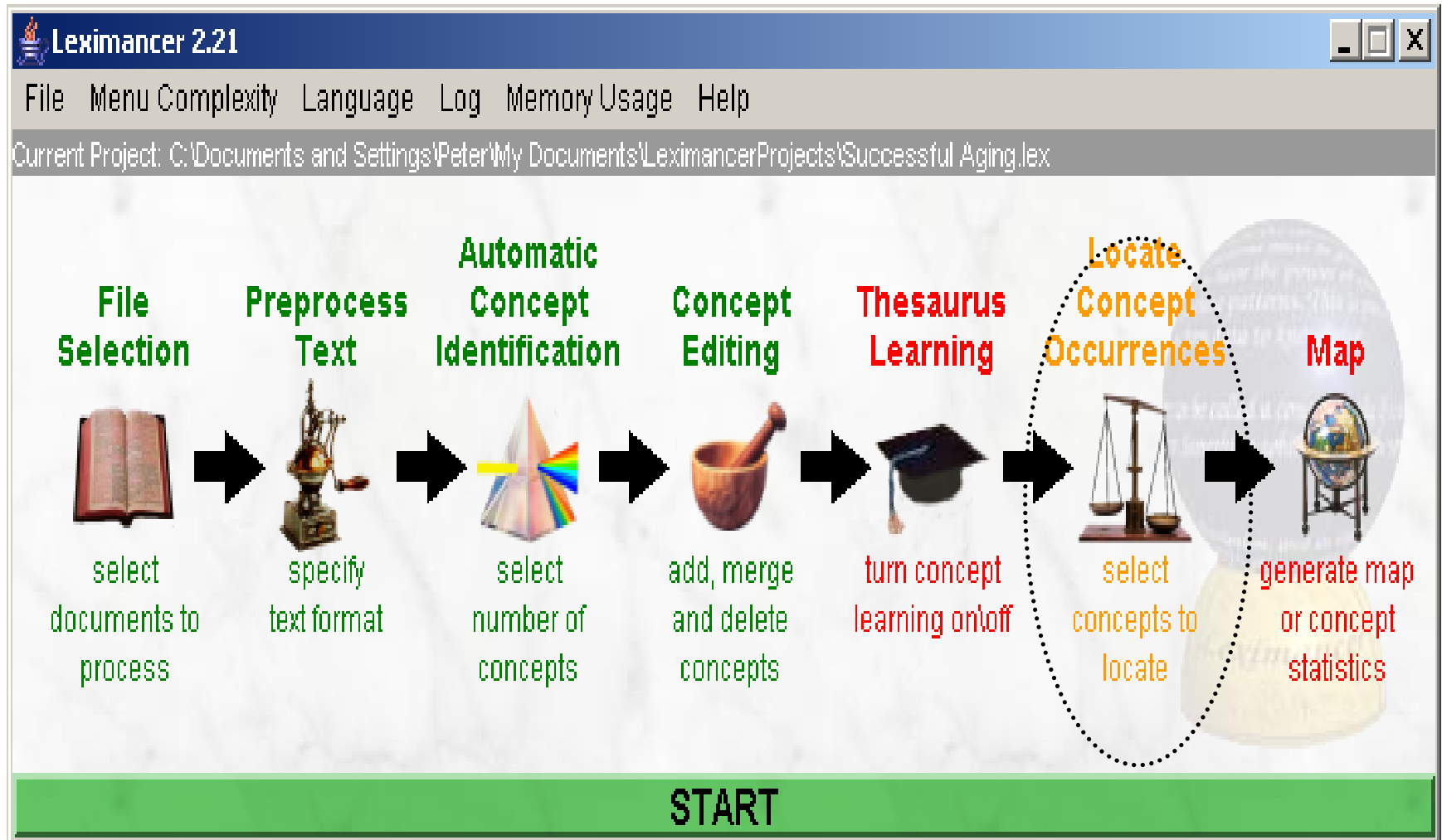
# Outcome of turning off thesaurus learning



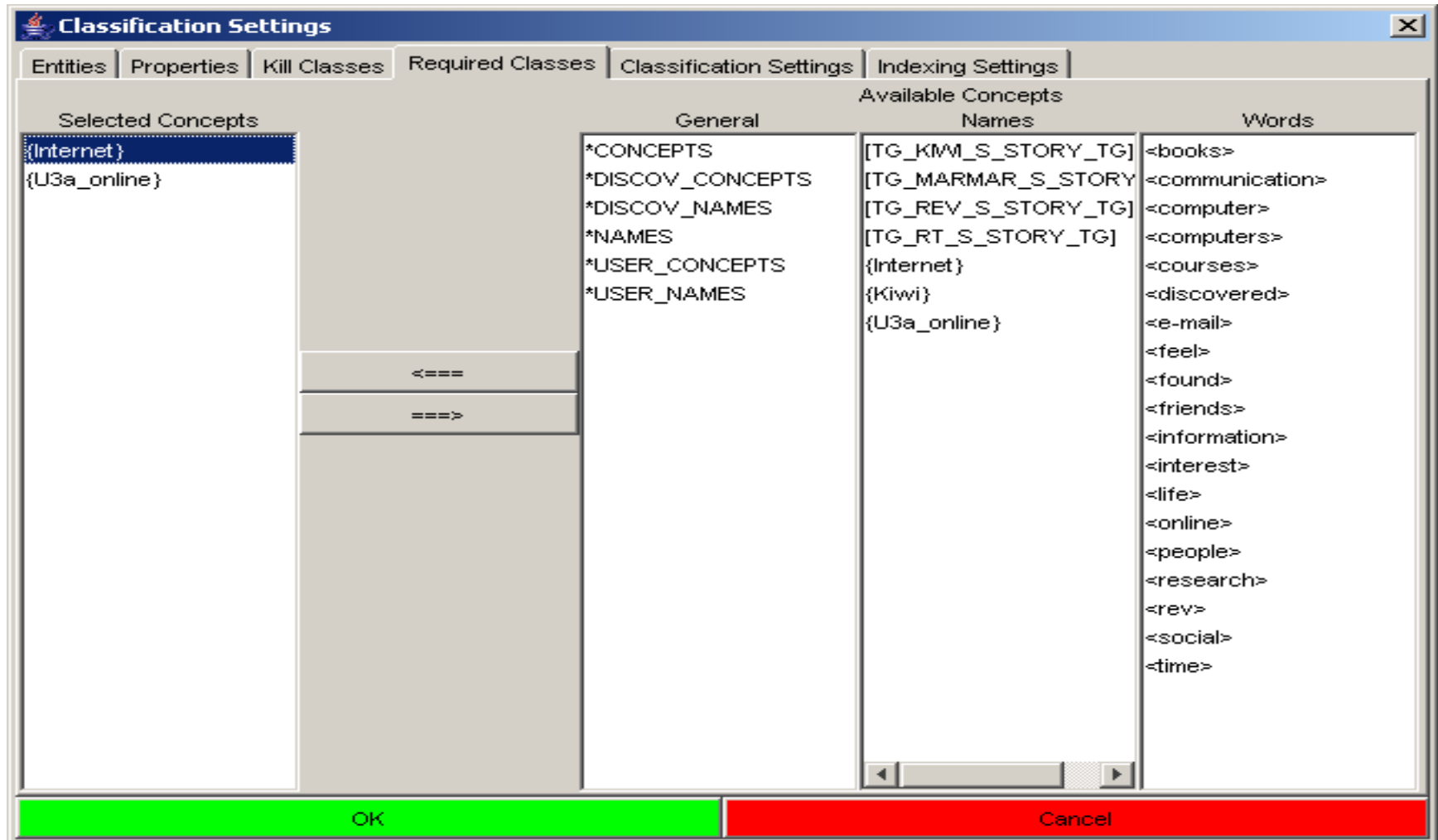
# Note: Concept profiling (for larger data sets)

- Allows learning process to discover new concepts associated with user-defined and automatic concepts
- Generates concepts that segregate document categories
- Normally set 3-10 concepts to be discovered per pre-defined concepts
- Themed discovery (All, Any, Each) options relate discovered to predefined concepts

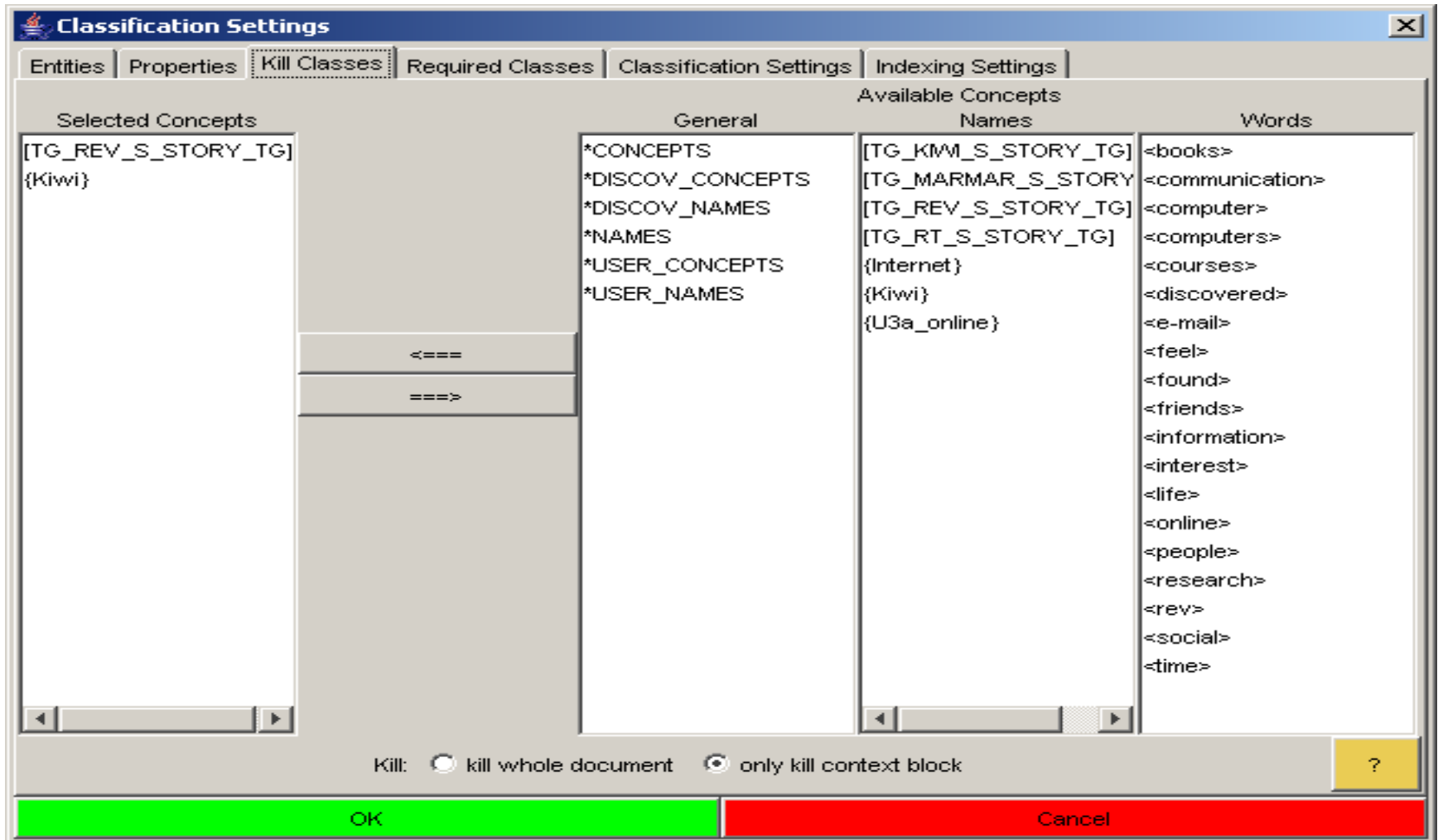
# Excluding or including blocks of text via *Locate concept occurrences*



# Defining concepts to be present by moving them to Required Classes



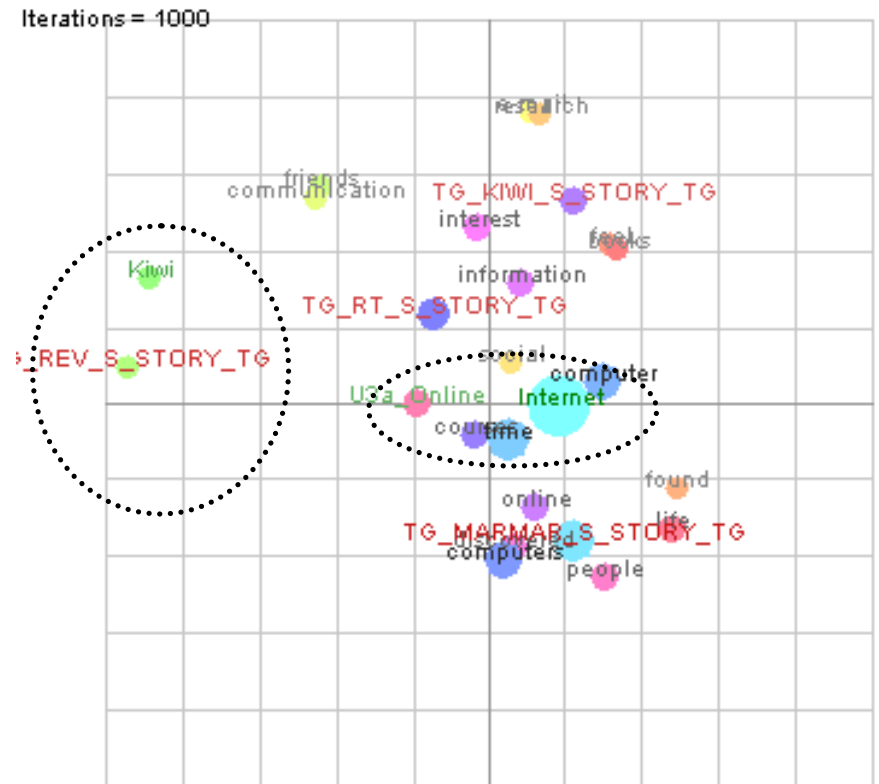
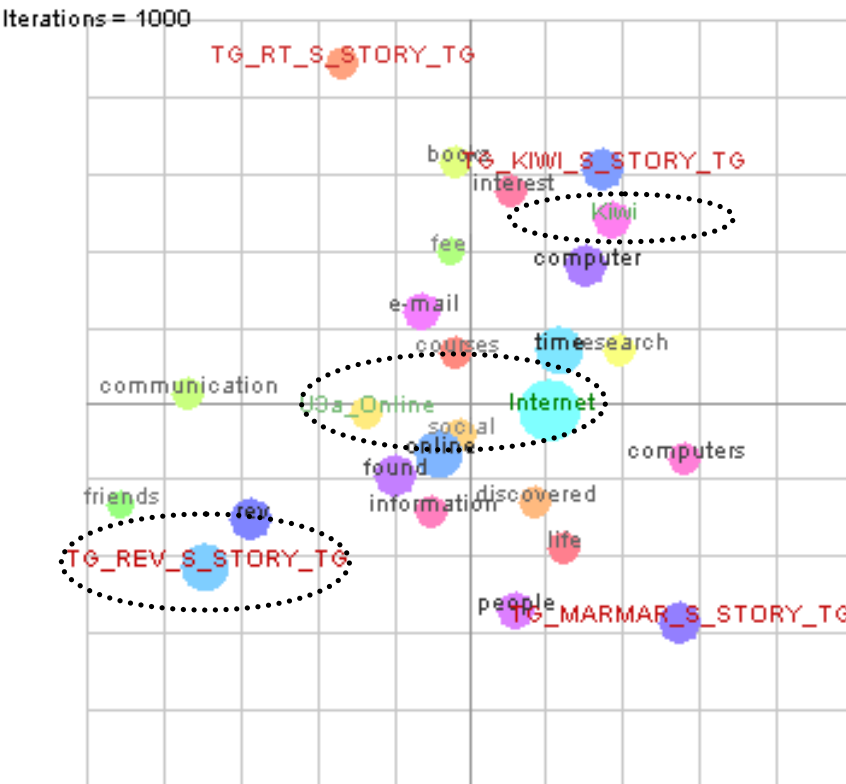
# Define concepts to be suppressed by moving them to Kill Classes



# Required and Kill Classes

- Apart from specifying concepts to use as entities and properties, can define Required Classes and Kill Classes.
- Required Classes must appear in a block of text for it to be analysed.
- Kill classes , in contrast, are concepts that if found in a classified block of text, remove the block from the analysis.
  - Kill Classes are useful for removing unwanted material such as interviewer's comments from a transcription.

# Outcome of using Kill & Required Classes



# Note: Varying classification settings for poetry and plays

Classification Settings

Entities Properties Kill Classes Required Classes Classification Settings Indexing Settings

Sentences per Context Block: 3 (normal) ?

Break at Paragraph:  yes  no ?

Word Classification Threshold: 2.4 (normal) ?

Name Classification Threshold: 4.5 (normal) ?

Classification Threshold for Supervised Concepts: 0.7 (training set normal) ?

Treat Names as Words:  yes  no ?

Blocks per Bucket: 1 ?

Statistics type:  count  weight ?

Document Metadata: no output ?

OK Cancel

- These settings control relational analysis
  - Increasing number of sentences or crossing paragraph breaks not recommended (concept map is more connected but also less stable)
  - These choices become difficult for poetry and plays (e.g., Blake, Shakespeare)

# Note: Varying classification settings to counteract over-connectivity in larger texts

Classification Settings

Entities Properties Kill Classes Required Classes Classification Settings Indexing Settings

Sentences per Context Block: 3 (normal) ?

Break at Paragraph:  yes  no ?

Word Classification Threshold: 2.4(normal) ?

Name Classification Threshold: 4.5(normal) ?

Classification Threshold for Supervised Concepts: 0.7(training set normal) ?

Treat Names as Words:  yes  no ?

Blocks per Bucket: 1 ?

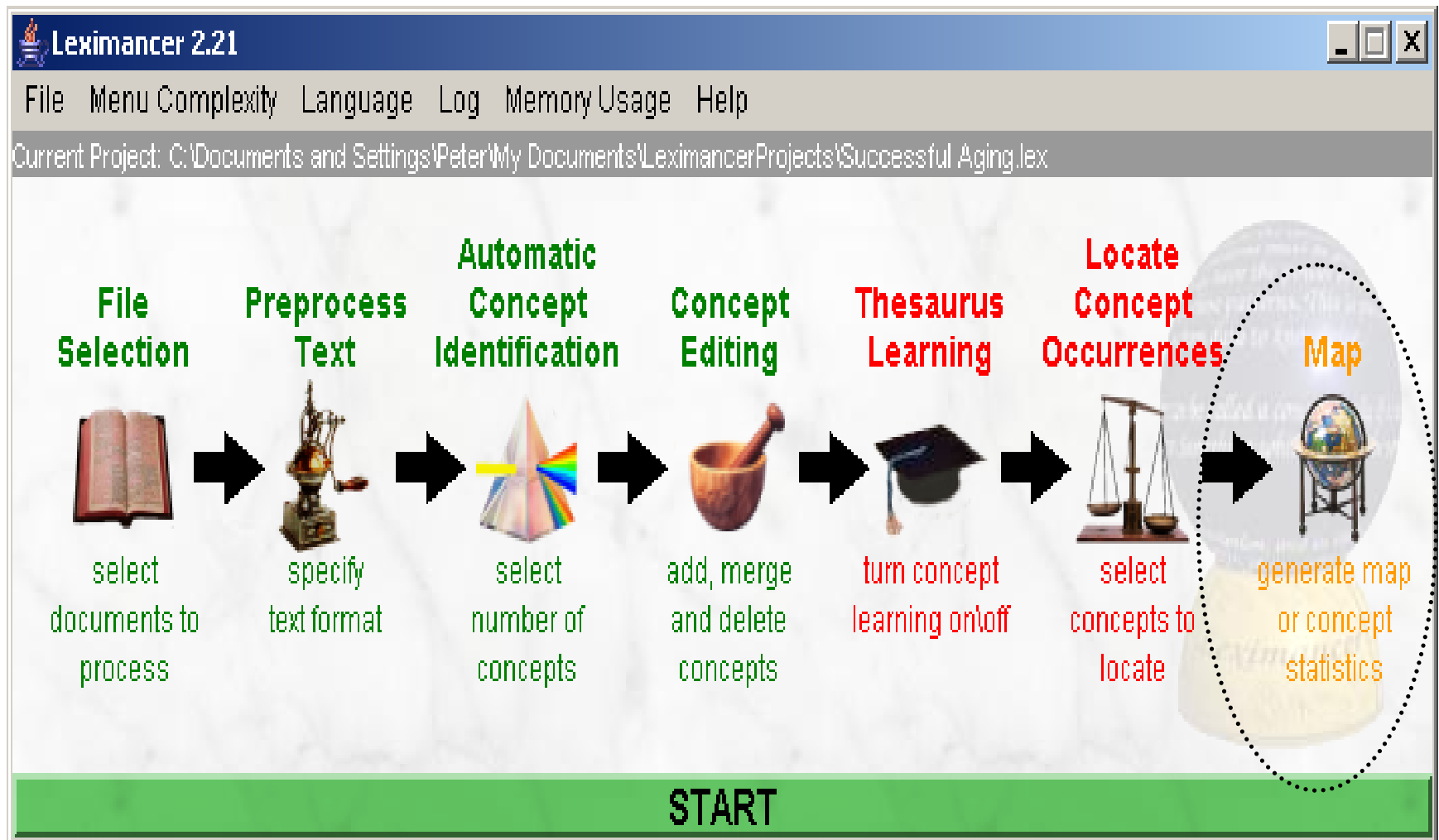
Statistics type:  count  weight ?

Document Metadata: no output ?

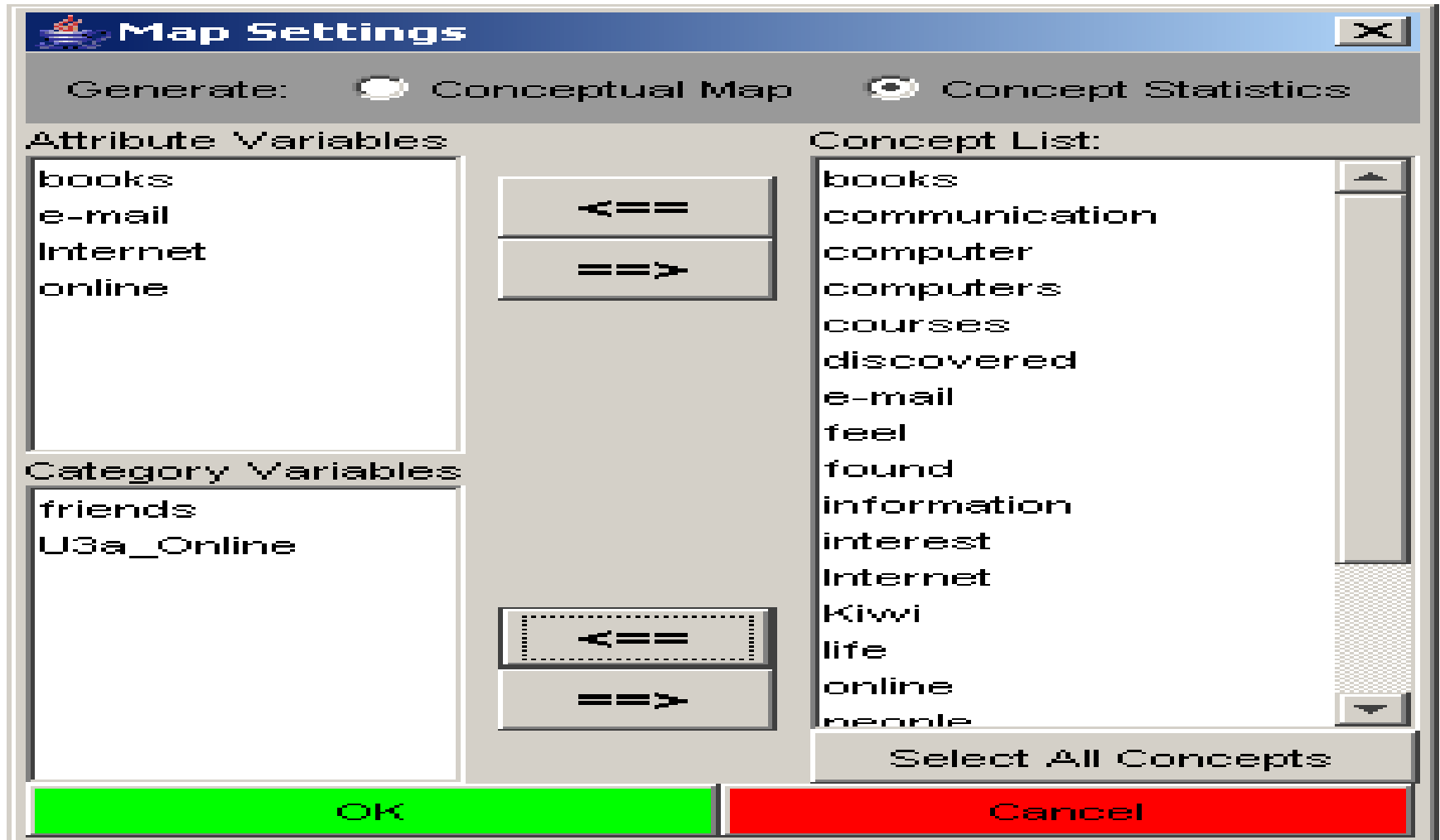
OK Cancel

- Buckets contain one or more context blocks
- Higher settings (more bits per bucket) counterbalance over-connectivity in larger blocks of text
- May need to exclude concepts that don't reach certain relevancy thresholds
  - Words such as sort, think, and kind increase connectedness at expense of meaning and can be removed

# Using the Map menu to generate concept statistics





# Map menu: Concept statistics



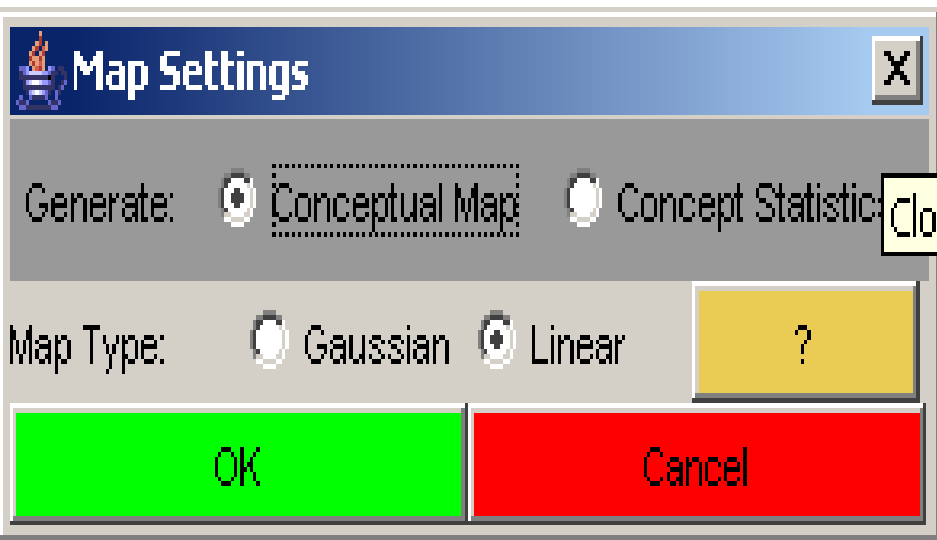
# Map menu



# Map menu: Concept statistics

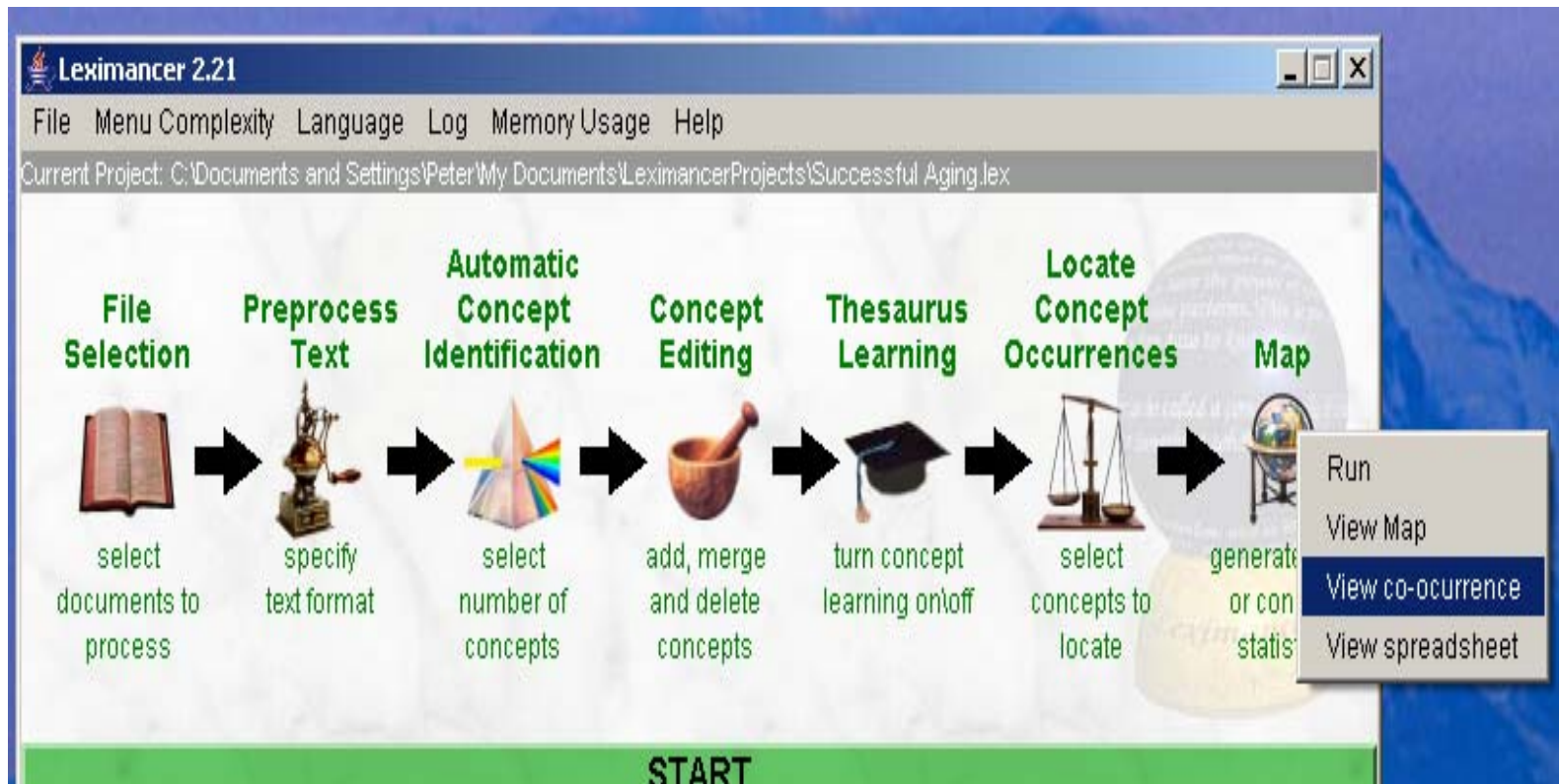
Category Variable	Attribute Variable [ P(attr cat) ]	Attribute Variable [ P(attr cat) ]
friends		books (0.0%) e-mail (20.0%) Internet (20.0%) online (20.0%)
U3a_Online		books (0.0%) e-mail (20.0%) Internet (20.0%) online (20.0%)

# Notes on concept maps



- Gaussian map has more circular symmetry and emphasises similarities in conceptual contexts
- Linear map is more spread and emphasises co-occurrence between concepts
- Mapping procedure is stochastic (random), so requires several runs to establish the relative positioning of concepts
  - Cluster instability due to overly-highly connected concepts
  - Linear map more stable under such conditions

# Notes on additional output options revealed by right clicking the Map icon



# Viewing the co-occurrences

Entity	x-ord	y-ord	weight	Internet	computer	time	online
Internet	0.617469	0.597788	21	21	13	8	7
computer	0.651225	0.738823	20	13	20	9	4
time	0.544265	0.710388	14	8	9	14	6
online	0.459876	0.576136	12	7	4	6	12
found	0.626026	0.352031	9	4	3	1	2
household	0.281678	0.784752	13	3	4	4	1
people	0.630349	0.104731	11	3	1	2	1
interest	0.445653	0.709988	7	3	2	3	1
e-mail	0.656022	0.256502	7	2	3	1	0
courses	0.464781	0.460004	5	1	1	2	2
life	0.510239	0.247768	6	2	1	1	2
information	0.283563	0.452343	8	2	1	2	2
discovered	0.355635	0.570878	6	2	2	2	3
social	0.469704	0.362548	4	1	1	1	2
U3a_Online	0.421436	0.326748	5	1	1	1	1
books	0.179719	0.682646	6	2	1	1	2
feel	0.353262	0.303284	4	1	1	0	1
research	0.779809	0.417774	6	2	1	1	0
communication	0.379408	0.01476	6	2	0	0	1
friends	0.25364	0.064906	5	1	0	0	1

First steps

# Viewing the spreadsheet

file	section	sentence	Internet	U3a_Online	books	communication	computer
data/./Kiwi's_story.doc	S1	1	22.3632	0	19.71	0	21.7263
data/./Kiwi's_story.doc	S1	4	0	0	0	0	24.34995
data/./Kiwi's_story.doc	S1	7	0	0	0	0	27.80921
data/./Kiwi's_story.doc	S1	10	0	0	18.24	0	0
data/./Kiwi's_story.doc	S1	13	23.8027	0	0	0	0
data/./Kiwi's_story.doc	S1	16	18.783	0	0	0	0
data/./Kiwi's_story.doc	S1	19	0	0	0	0	0
data/./Kiwi's_story.doc	S1	21	0	0	0	0	19.73261
data/./Kiwi's_story.doc	S1	24	24.7895	0	21.49	0	0
data/./Kiwi's_story.doc	S1	27	0	17.9137936	0	0	0
data/./Kiwi's_story.doc	S1	30	0	0	0	0	19.87441
data/./Kiwi's_story.doc	S1	33	0	0	0	20.27431873	0
data/./Kiwi's_story.doc	S1	36	0	0	0	0	0
data/./Kiwi's_story.doc	S1	37	20.1496	0	0	0	21.95705
data/./Marmar's_Story.doc	S1	1	22.2243	0	0	0	21.96525
data/./Marmar's_Story.doc	S1	4	0	17.8695271	0	0	21.49474
data/./Marmar's_Story.doc	S1	7	0	0	0	0	0