NetMapper User Guide

Eric Malloy and Kathleen M. Carley

March 2018

NetMapper is a tool that supports extracting networks from texts and assigning sentiment at the context level. Each text is processed separately and a network representation of that text is exported. NetMapper is interoperable with ORA. Given a text, NetMapper extracts both concepts and links among them. NetMapper can be rapidly customized to support extraction of just terms of interest

The output from NetMapper is in CSV and in the xml format read by ORA. NetMapper is lexicon based and an extensive set of thesauri, translation files and delete lists in over 40 languages. In addition, it also supports the use of user-generated domain thesauri and delete lists. Hence, users who are working in a specialized area or on a specialized topic can fine tune what concepts are extracted using the domain files.

Input: Types of Texts

Currently, to use NetMapper the user must put each text to be processed into a .txt file. Generally, images should be removed. Examples of types of texts include:

- News documents
- Journal articles
- Blog posts
- tweets

NetMapper can accept data in multiple formats:

- US-ASCII
- UTF-8
- UTF-16
- UTF-32

Concept

A concept is a word on phrase that serves as a single ideological idea. Examples are president and John F. Kennedy. When concepts are not categorized into an ontological category they are treated as being of type knowledge. Alternatively, they can be categorized into a set of ontological categories. These categories are based on the ORA ontology.

Ontological Categories

If meta-networks are generated, NetMapper uses a pre-defined ontology and automatically classifies concepts into this ontology. The ontological categories and types are described below:

- Agent – specific, generic
    - Individual actors
    - Specific – unique often with first and last name - Jamie O'Connor
    - Generic – non-unique and often a role - haberdasher
- Organization – specific, generic
    - Groups, corporations, populations
    - Specific – unique - IBM
    - Generic – a type - Non-government organization
- Location – specific, generic
    - Places things can be at
    - Specific – unique with lat and lon or place on map – United States of America
    - Generic – may be at multiple locations – hill
- Event - specific, generic
    - Major happenings that impact groups
    - Specific – occur once – World War I
    - Generic – multiple occurrences – Tornado
- Knowledge
    - Branches of knowledge
    - Topics of interest
- Resource
    - Things that are not purely mental – disease, food, wire
- Task
    - Activities – eat
- Belief
    - "isms" - Catholicism
    - Sentiment – positive, negative
    - Belief statements – right to bear arms

Types of Networks Supported

Two types of networks can be extracted: semantic networks and meta-networks. The semantic networks are concept to concept networks. From an ORA perspective, all concepts are treated as being of type knowledge. The meta-networks are concept to concept except

Operation

- Potential to remove
    - Stop words
    - Punctuation
    - Numbers
    - 27 languages are supported

- Many concepts can be classified into common terms
    - Thesauri based
    - Special thesauri for disease, sports, numeric expressions in 27 languages

Support for user construction of domain thesauri.

Thesauri format

The purpose of a thesauri is two fold: First, it specifies how a concept should be referred to. Thus it provides information about what to translate (conceptFrom: the item in the raw text) to what (conceptTo: the concept that will be visible in the output file). This provides the user with a way of reducing complexity and so the number of concepts in the networks by: 1) by converting a set of synonyms to a common word; 2) overcoming common typos; and 3) clustering words into topical areas based on user choice. This also provides the user with a way of adding attributes such as the default valence for sentiment calculation.

The format of a thesauri file is a tab separated file with a set of columns specifying relevant information. Row 1 must contain the header for that column using the name specified below with exact spelling and case. All and only the following fields can be included.

Thesauri row 1 headers:

1. conceptFrom
    - This is a required field
2. concetpTo
    - This is a required field
3. ontology
    - This is a required field
    - Only allowed values are:

agent,organization,location,event,knowledge,resource,task,belief

4. nodetype
    - This is a required field
    - Only allowed fields are generic or specific for agent, organization, location and event
    - For other ontological classes this =must be blank – so blank for knowledge, resource, task or belief
5. Category 1
6. Category 2
7. Category 3
8. Country
9. First Name
10. Last Name
11. Gender
12. Suffix
13. Language
14. Acronym

15. Valence
16. Evaluation
17. Potency
18. Activity
19. Affect Mean
20. Military Role
21. Political Role
22. Religious Role

Not all the columns need values for every entry. What is required is conceptFrom (this is what you want search for), conceptTo (this is what you want it replaced with when its found), metaOntology the ontology it has to be either agent, event, organization, location, or knowledge, and nodeType which is either “specific” or ”generic”.

Each line after the header row contains information on a concept.

The file must be saved as UTF-8 (without BOM). To do this, do the following. In excel save the file as unicode. This creates a tab separated file that is UTF-16. Then using another tool like Notepad++, VIM, Emacs, etc .. re-save as utf8 without BOM.

NetMapper has a set of pre-defined thesauri in a large number of domains. The user can choose to use these or not. By default they are all applied. In addition the user can choose to create and use a domain thesauri.

In domain thesauri there must me at least four columns. These are conceptFrom, conceptTo, Ontology, nodetype.

Delete List

A delete list defines a set of concepts that should be deleted and not included in the resultant coded network. NetMapper has a set of pre-specified delete lists. By default, all will be applied. The user, however, can choose not to apply any or all of these delete list and/or can add a customized domain delete list.

The set of default delete lists contain concepts for:

- Time
- Measurement
- Symbols
- Stop words
- Numbers
- Regular expressions

The format of a delete list is a csv file with only one column. Each concept to be deleted is in its own line. The files are single concept per line. A concept may contain more than a single word most all Unicode characters are accepted with the exception of tabs.

Application of Thesauri and Delete Lists

Thesauri and delete lists are applied in the following order:

1. Domain Thesauri
2. Default Thesauris
3. Domain Delete List
4. Default Delete Lists

Link Generation

NetMapper creates networks through link generation. After the thesauri and delete lists are applied then NetMapper extracts the networks by identifying links among concepts. A link is placed between two concepts when they are within the window of operation. The window of operation is defined by either or both number of concepts and syntactic structure (e.g., number of clauses, sentences, or paragraphs). There exist default choices which have been found empirically to lead to the best results for the type of texts being examined. However, the user can choose to change the defaults by specifying:

- Window size based on number of words
- Window size based on number of syntactic units
- Whether or not deleted terms are "counted" in defining the size of the window

Finally, NetMapper auto detects language and moves correctly either from left to right up to down, or the reverse. The user may define the languages in the input file(s). This is on a per file basis, however the UI allows for languages selected for a given file to easily be selected for all other files. There is also the option to have language detection on or off. If off and no languages are selected for a given file, the default is English. If there are languages defined and language detection is enabled, NM will still search the file for additional languages and add any to the list of languages to be used that score above a certain threshold

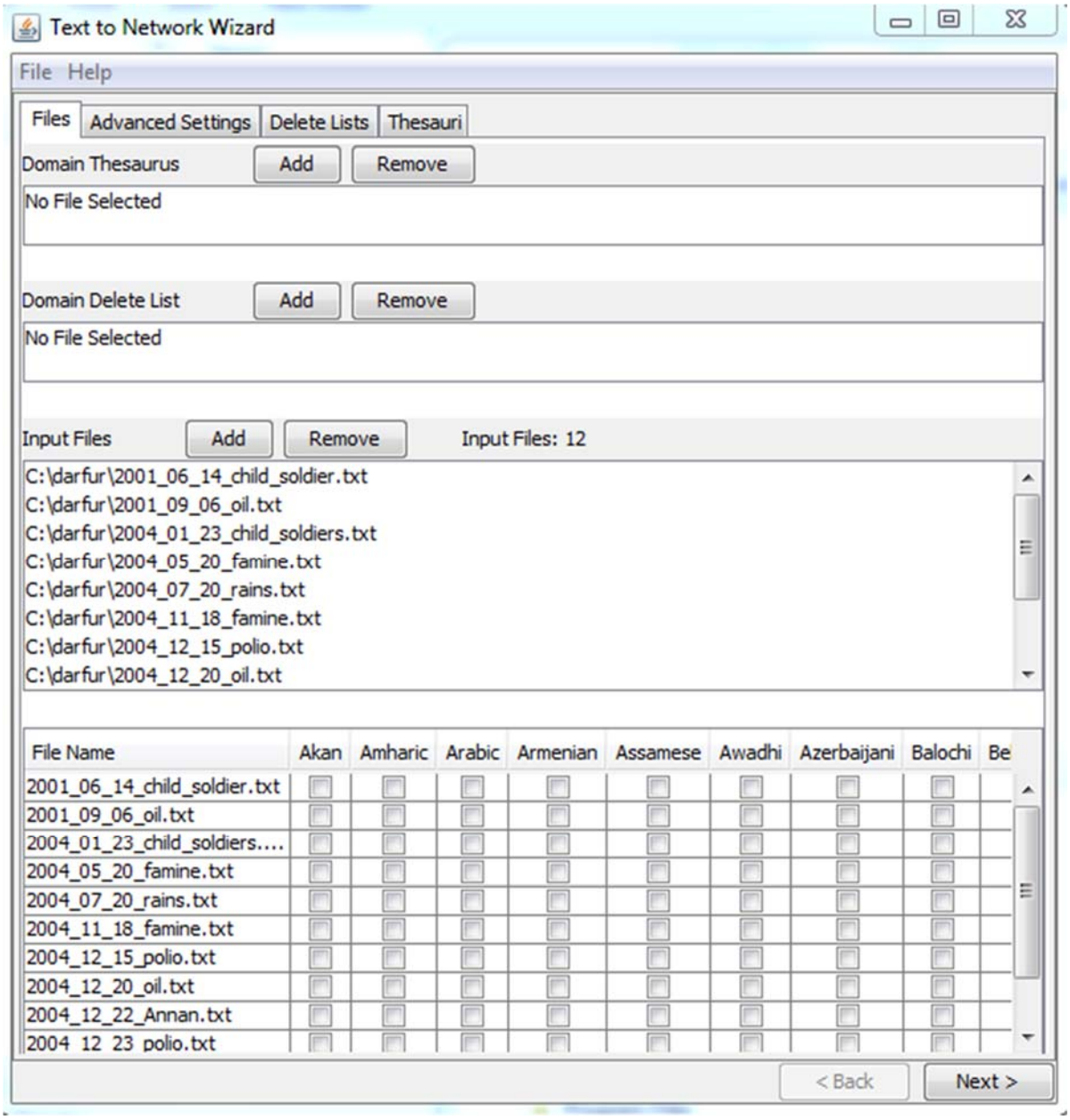
Outputs

- NetMapper generates the following outputs. DyNetML files for import to ORA
    - Meta Network (With or Without Unknowns)
    - Semantic Network (With or Without Unknowns)
- Original Text with modifications
    - Just UT Concepts in the Text
    - UT and DT Concepts
    - With or Without Deleted Concepts
- CSV files containing sentiment scores

Description of NetMapper Operations

The following pages describe the set of screens in NetMapper and the different parameters that the user can set.

Files Page



- **Domain Thesauri**
  - These are user supplied files.
  - You may add multiple domain thesauri using the add button. If you select any domain thesauri in the list and click the remove button, it will be deleted from the list.
- **Domain Delete List**

    - These are user supplied files. The files contain terms that will be deleted from the text and will therefore not show up in the generated networks.
- **Input Files**
    - Lists all the files to be processed.
    - Use the add button to bring up a file dialog box. Multiple files or a directory can be selected.
    - The remove button is used to remove individual files from the list. This is done by selecting a file or files and clicking the remove button.
- **Language Selection**
    - Allows the user to select additional (to English) languages to use for processing each input file.
    - Column Popup Menu (Figure 1)
        - The column popup menu has two items, “Select For All” and “Deselect for All”.
            - “Select for All” Will select that language for all the files.
            - “Deselect for All” Will deselect that language for all the files.
    - Row Popup Menu (Figure 2)
        - The Row Popup menu has three options “Select All Languages”, “Deselect All Languages” and “Clone Entry to All Others”
            - “Select All Languages” will select all the languages for the row that was right clicked on.
            - “Deselect All Languages” will unselect all the languages (except for English) for the row that was right clicked on.
            - “Clone Entry to All Others” will select the same languages for all the files that are select for the current row.

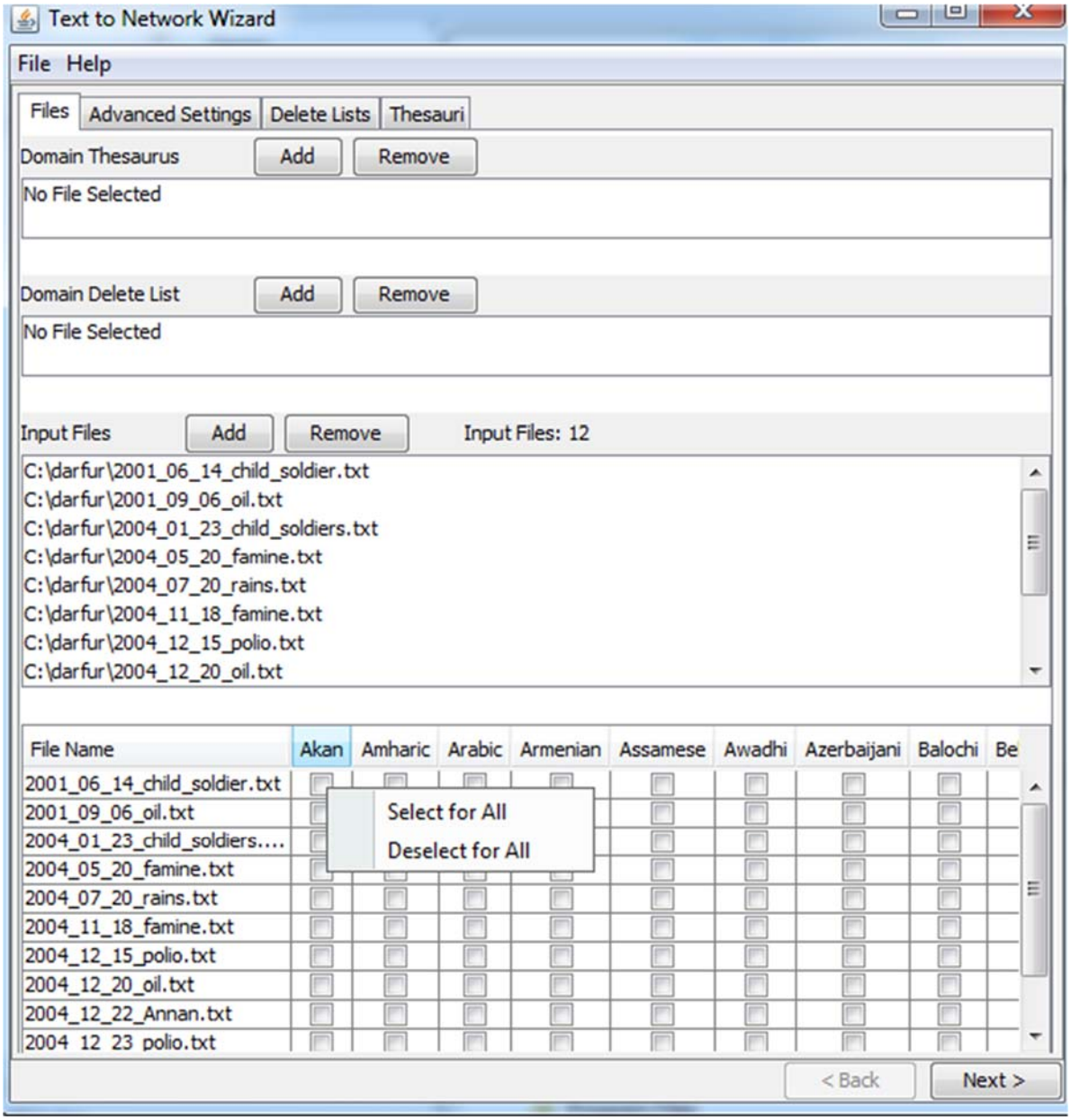

Text to Network Wizard
File Help
Files
Advanced Settings
Delete Lists
Thesauri
Domain Thesaurus
Add
Remove
No File Selected
Domain Delete List
Add
Remove
No File Selected
Input Files
Add
Remove
Input Files: 12
C:\darfur\2001_06_14_child_soldier.txt
C:\darfur\2001_09_06_oil.txt
C:\darfur\2004_01_23_child_soldiers.txt
C:\darfur\2004_05_20_famine.txt
C:\darfur\2004_07_20_rains.txt
C:\darfur\2004_11_18_famine.txt
C:\darfur\2004_12_15_polio.txt
C:\darfur\2004_12_20_oil.txt
File Name
Akan
Amharic
Arabic
Armenian
Assamese
Awadhi
Azerbaijani
Balochi
Be
2001_06_14_child_soldier.txt
2001_09_06_oil.txt
2004_01_23_child_soldiers....
2004_05_20_famine.txt
2004_07_20_rains.txt
2004_11_18_famine.txt
2004_12_15_polio.txt
2004_12_20_oil.txt
2004_12_22_Annan.txt
2004_12_23_polio.txt
Select for All
Deselect for All
< Back
Next >

Figure 1

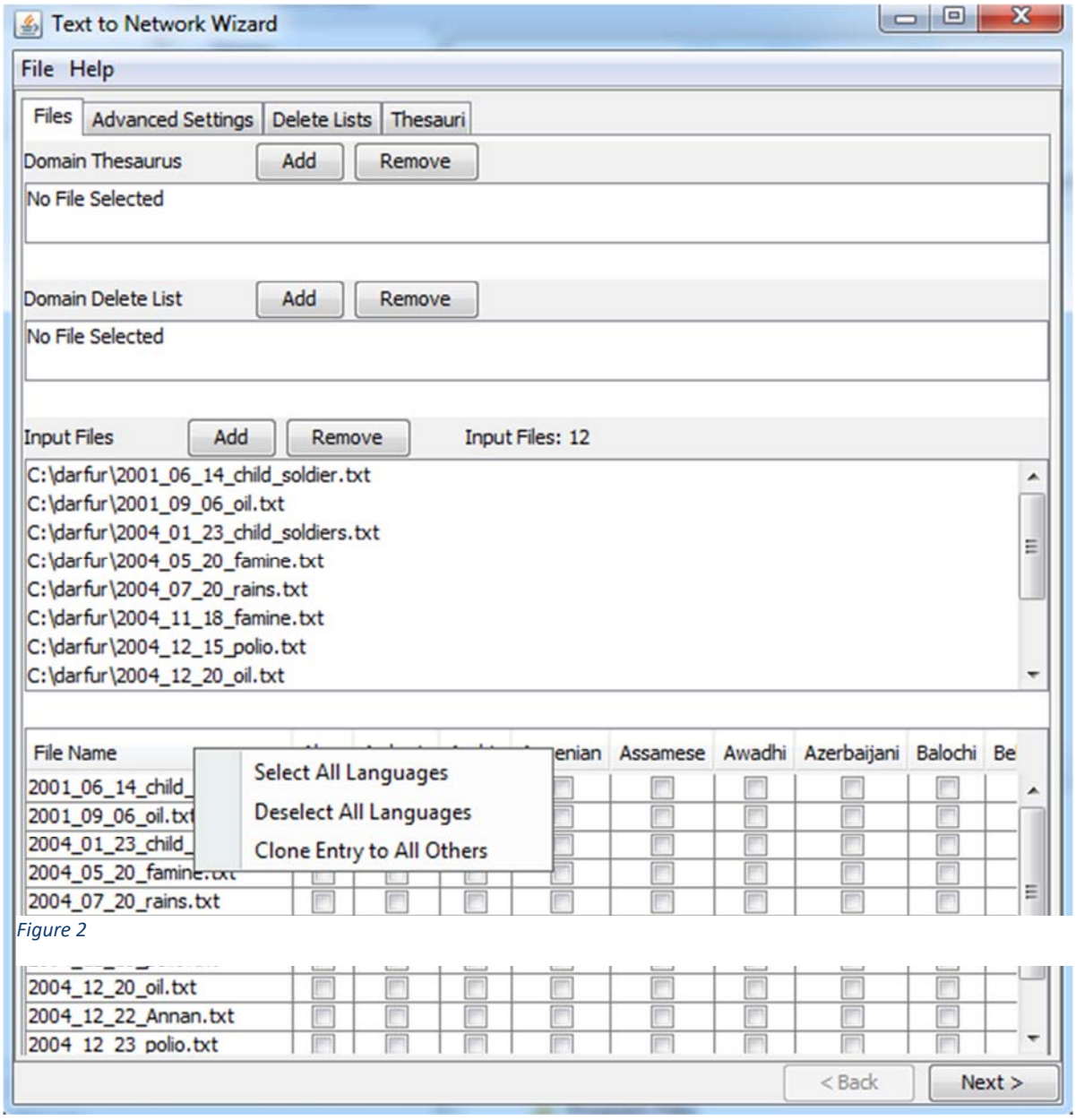
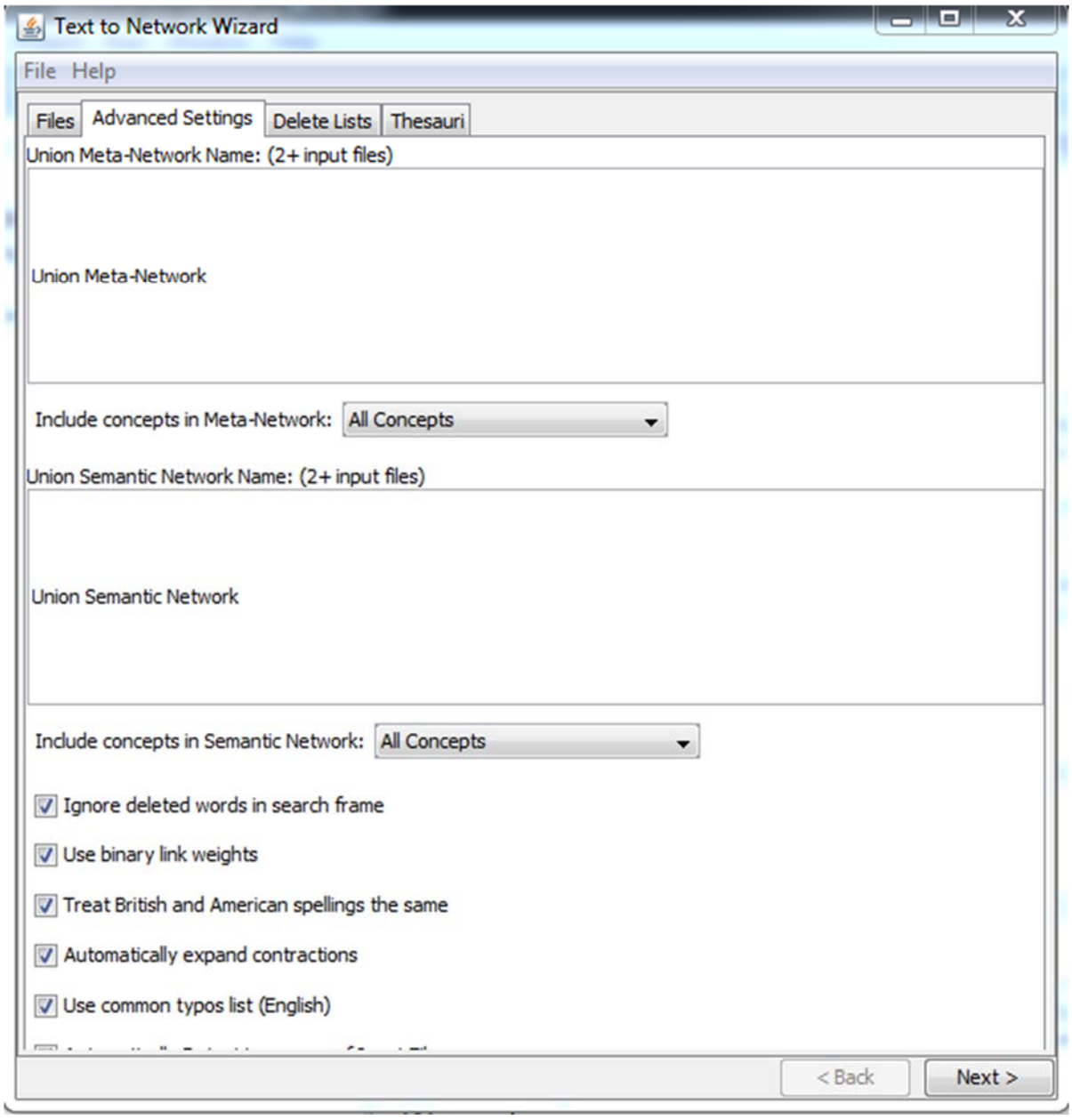

Text to Network Wizard
File Help
Files
Advanced Settings
Delete Lists
Thesauri
Domain Thesaurus
Add
Remove
No File Selected
Domain Delete List
Add
Remove
No File Selected
Input Files
Add
Remove
Input Files: 12
C:\darfur\2001_06_14_child_soldier.txt
C:\darfur\2001_09_06_oil.txt
C:\darfur\2004_01_23_child_soldiers.txt
C:\darfur\2004_05_20_famine.txt
C:\darfur\2004_07_20_rains.txt
C:\darfur\2004_11_18_famine.txt
C:\darfur\2004_12_15_polio.txt
C:\darfur\2004_12_20_oil.txt
File Name
Assamese
Awadhi
Azerbaijani
Balochi
Select All Languages
Deselect All Languages
Clone Entry to All Others
2001_06_14_child_
2001_09_06_oil.txt
2004_01_23_child_
2004_05_20_famine.txt
2004_07_20_rains.txt
2004_12_20_oil.txt
2004_12_22_Annan.txt
2004_12_23_polio.txt
< Back
Next >

Figure 2

Advanced Settings Tab

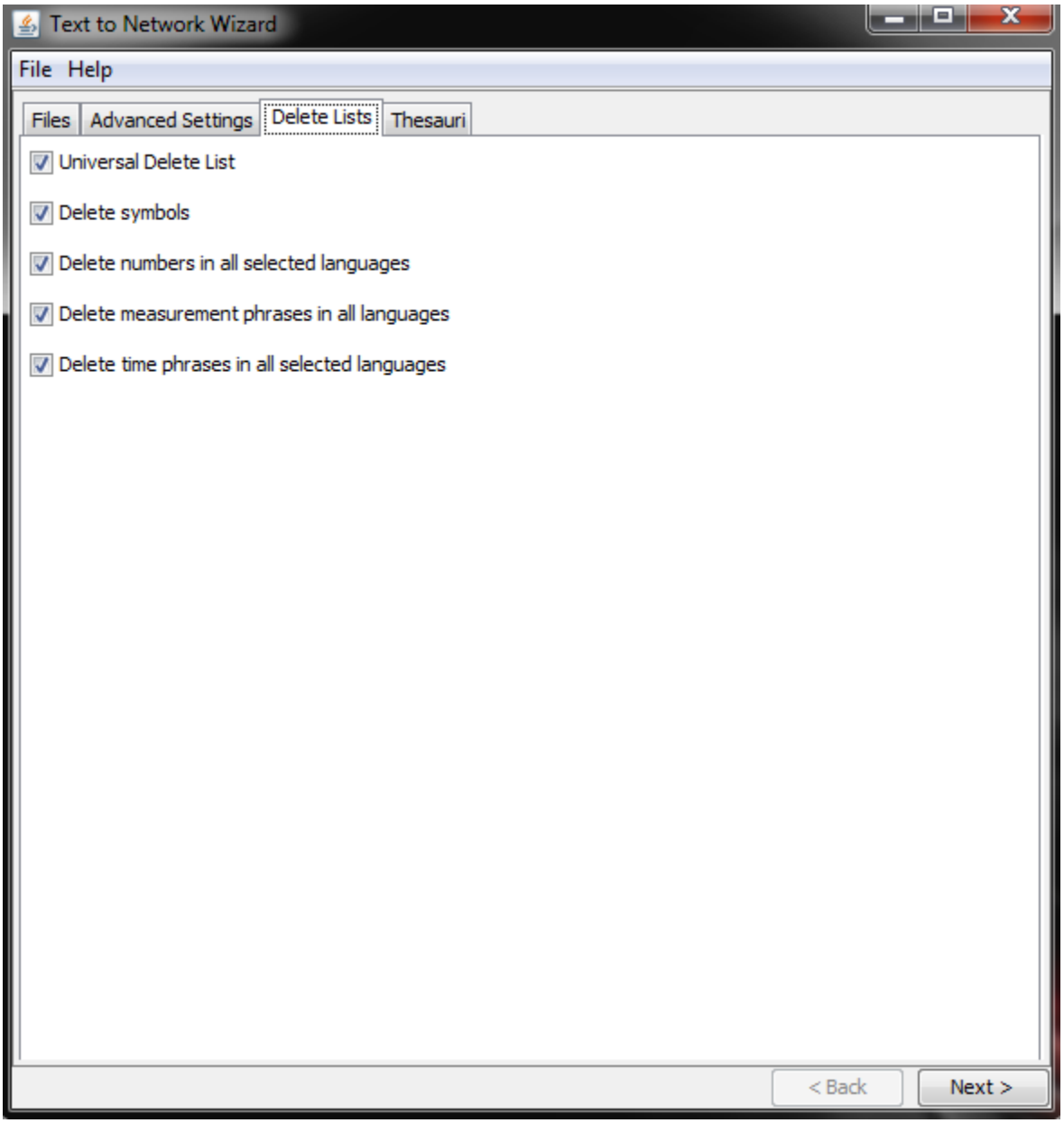
Note you do not need to use this tab. You can just use the defaults and never open this tab.



- Include concepts in Meta-Network
    - All Concepts – Everything
    - Only From Universal – Only include terms from the universal thesaurus
    - Only From Domain – Only include terms from the domain thesaurus

- Include concepts in Semantic Network
    - All Concepts – Everything
    - Only From Universal – Only include terms from the universal thesaurus
    - Only From Domain – Only include terms from the domain thesaurus
- Ignore deleted words in search frame
    - Means that NM will not count terms that have been deleted when counting words within the search window to create links
- Use binary link weights
    - All edges will have a weight of 1
- Treat British and American Spellings that same
    - British spellings will be treated as their American equivalent as opposed to being processed as a different word.
- Us Common Typos list
    - NM has the ability to fix some basic typos that often occur. This selection tells NM to make those assumptions.
- Automatically Detect Language of Input Files
    - If selected NM will attempt to determine the language of each individual input file. This may cause other translation thesauri to be loaded to process any given input file.

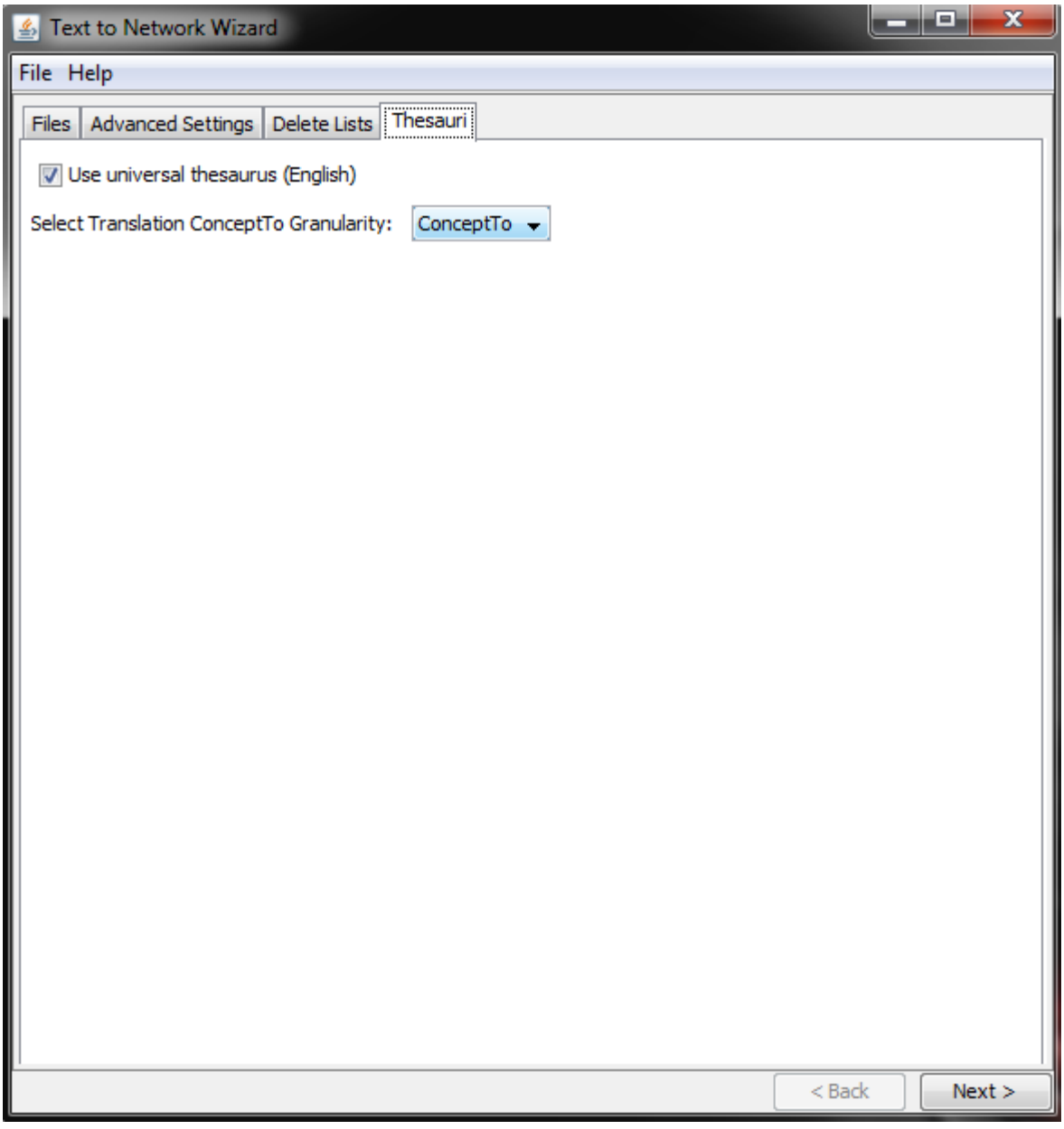
Delete Lists Tabs



- Universal Delete List
    - By selecting this option all words in the universal delete list will be deleted from the text and ignored for a good deal of NMs processing.
- Delete symbols
    - Many non alpha numeric symbols that are also not punctuation will be deleted
- Delete numbers in all select languages
    - Numbers will be removed, that means 1,2,3 will be deleted as well as one, two, and three.

- Delete measurement phrases in all languages
    - Selecting this options will remove terms like inches, centimeters and so on
- Delete time phrases in all selected languages
    - Similar to delete measurement this option will delete terms like hour, minute and second.

Thesauri Tab

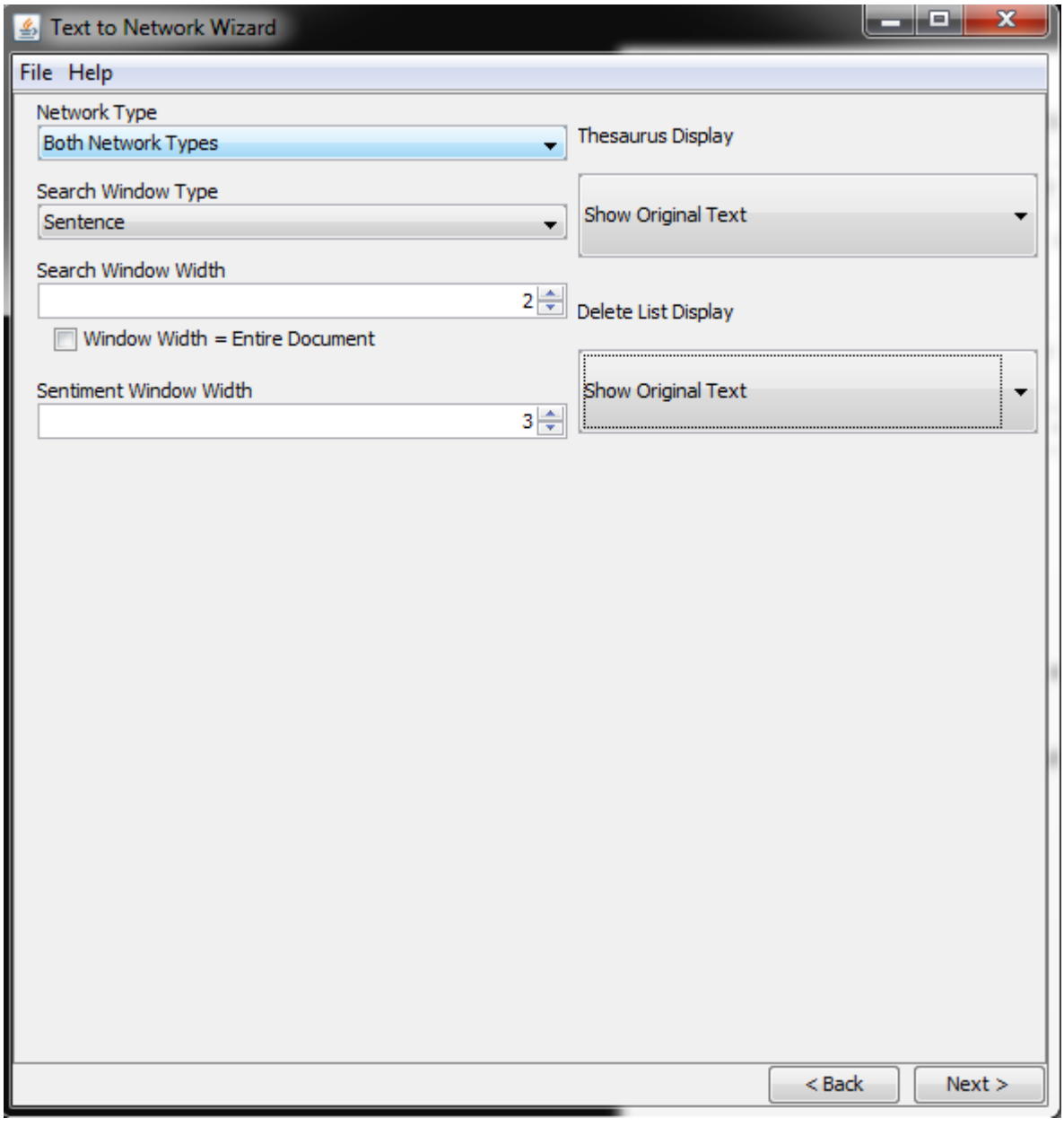


- Use universal thesaurus (English)

    - This will mean that a large number of predefined thesaurus entries will be used in processing text. This also includes all the translation thesauri. The general scope of the thesauri to be included by selecting this option covers many well known agents, locations, events and knowledge.
- Select Translation ConceptTo Granulatiry
    - There are several levels of granularity that a term can be translated too. In order of most specific to most general they are ConceptTo, Category 1, Category 2.

Network Generation Page

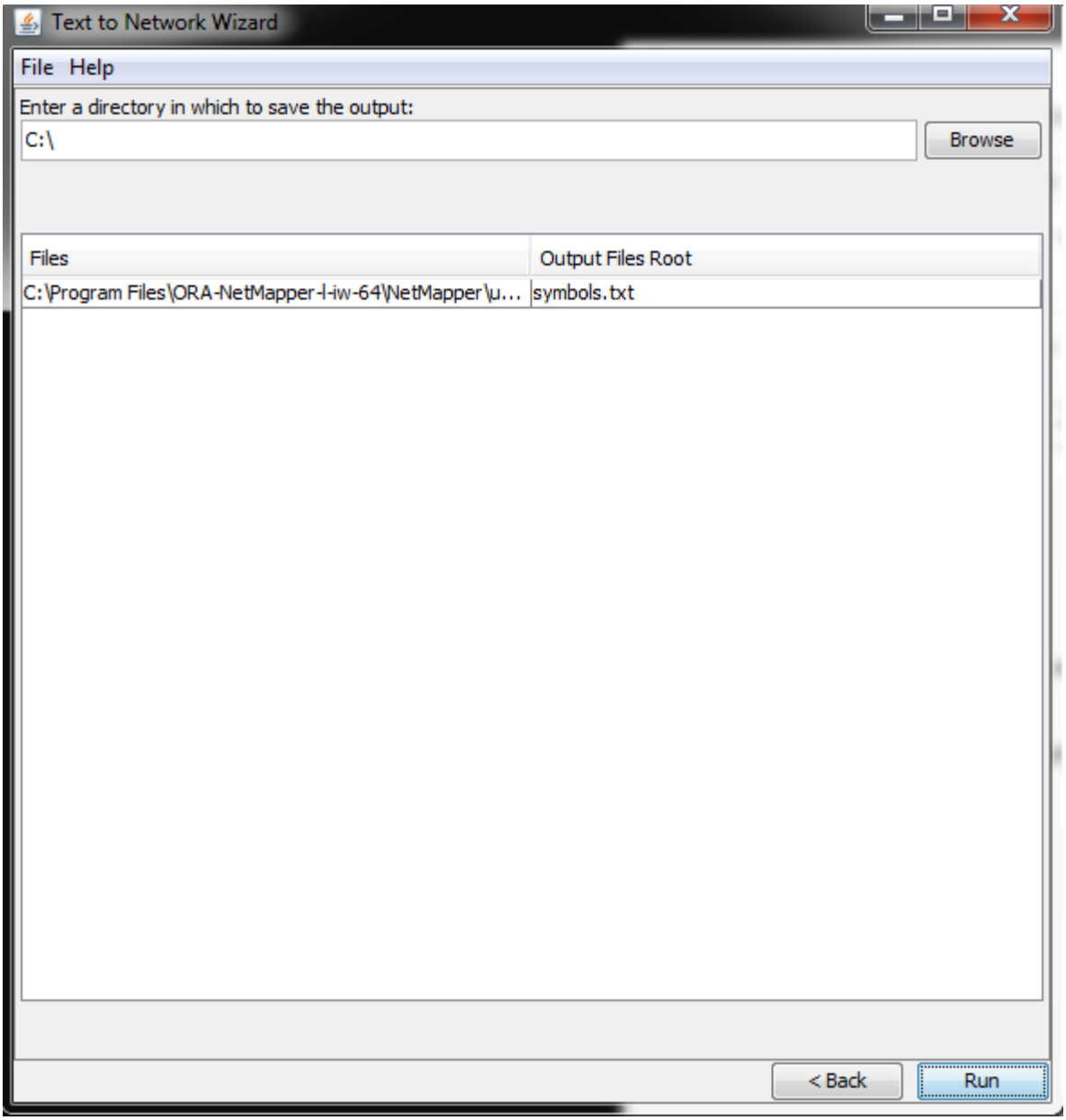
To generate network you use the Text to Network Wizard. This walks the user through a series of choices about what to generate and any constraints on how the links are generated.



- Network Type
  - Both Network Types will generate both Semantic and Meta networks.
  - Meta will only generate a meta network
  - Semantic will only generate semantic network

    - Concept List will generate only a text file with a list of concepts found
- Search Window Type
    - Word will determine the search width by number of words in the window.
    - Sentence will determine the search width as number of sentences in the window.
- Search Window Width
    - The number of words or sentences that should be used in the window when determining links between terms in the network.
- Sentiment Window Width
    - The window width to be used for the sentiment network.
- Thesaurus Display
    - Currently unimplemented
- Delete List Display
    - Currently unimplemented

In the Text to Network Wizard you will be asked where you want to put the output.



- Output Page
  - Output Direction is the location to put all the output files
- Files list
  - In this list you can choose to change the root naming convention of a given input file. This means that if you have a file named symbols.txt, all ouput files will be named symbols.<whatever extensions are necessary>. However if another file in the list has the same root name, the results from one will overwrite the other. By manually changing the root, that can be prevented.