Leximancer v 5.0 : distinguishing features

This document could be read in conjunction with the 'Choosing a CAQDAS Package Working Paper' which provides more general commentary of common CAQDAS functionality. Leximancer sits within a very specific "Content Analysis" paradigm. The review does not provide an exhaustive account of all the features provided by Leximancer but is designed to highlight some of its distinguishing elements. The Comment section at the end details our opinions on certain aspects of functionality and usability. We thank Andrew Smith, University of Queensland, and Steve Wright, University of Lancaster for checking the accuracy of this review.

Background and philosophy **https://info.leximancer.com/** **Leximancer** is available for Windows, Mac or as a web-based application (LexiPortal). It was developed by Andrew Smith and Michael Humphreys in 2000 at the University of Queensland, in Brisbane, Australia. More recently they have released a low budget software, mentioned here at times, **LexiReader**, which provides a smaller subset of the same tools. Leximancer is a text mining and content analysis software. The developers' purpose was to let the data generate a transparent model in order that the researcher can efficiently examine and make sense of vast amounts of text. The tool is used for determining the presence of words or concepts in collections of textual documents. It is used to break down the material into a finite number of categories and identify a semantically derived relationship between them. Tools both visualize and quantify text. The model building process (see Figure 1) processes the flow of automated work in a matter of seconds depending on the size of the dataset and the work put in by the researcher to vary settings. Results are displayed textually and interactively by **topic** in **Topic Guide** or as a visual **Concept map** with accompanying statistics and charts. **LexiReader** – the smaller version of the software - consists of the **Topic Guide** but does not include the **Concept map** tool.

Minimum System Specifications (recommended by developer) Windows: 7, 8, 10 (64 bit) ■ Mac Intel OS X 10.11+ (El Capitan, Sierra, High Sierra) 64 bit. ■ Minimum of 2GB ram with at least 750MB dedicated to Leximancer. ■ 20GB free disk space for source text documents and Leximancer project data-store files. ■ For LexiPortal (web version) IE 11, IE Edge, Firefox 45+, Chrome 50+, and Safari 10+

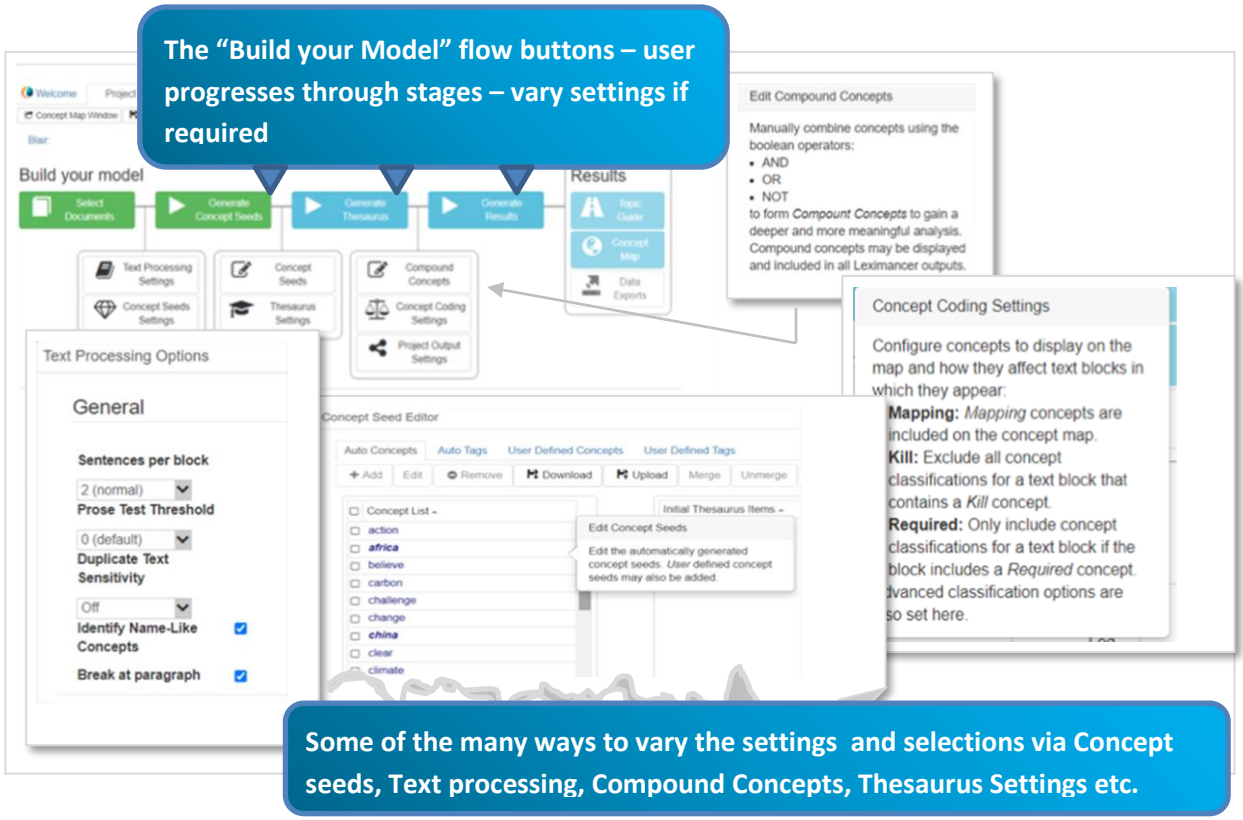

Figure 1. The central *Build Your Model* flow chart in Leximancer – with variable settings

The structure of work in Leximancer The user can influence selections, settings and views but most work is machine-based on the identification of content rather than interpretation followed by any 'manual' coding by the user. Importantly however, the user can define at the outset exactly what content Leximancer looks for and what is relevant (in LexiReader, the user must accept the default settings). ■ **Select Documents** After creating a project, the documents required for the current analysis are moved across to the project's selection list (at the first flow chart button). There remain three stages of the flow chart (see Figure 1). In fact the researcher can jump straight to clicking on the **Results** button if the default settings are acceptable or if only the text results in **Topic Guide** are required. **LexiReader** will also speedily perform this task. As the process can be so quick, creating multiple projects with different selections of data is easy ■ **Vary the Settings** Researchers taking more control of Settings in Leximancer will follow the four flow chart buttons and the Settings options between each, to define how the software completes the *text-mining* process. Options include those listed at Figure 1. (keywords/concept seeds can be user-defined) ■ **Generate Concept Seeds** (at the second button). Text-mining begins - a little like coding, concept seeds are created on the basis of strongly correlated keywords ■ The optional **Thesaurus** process (at the third button), generates **Concepts** based on the likely co-occurrence of concept seeds within chunks of text (the default being two sentences). This underpins the classification and later retrieval of data segments at **Topics**. The default (but optional) differentiation of ***word-like*** words and ***name-like*** words constitute an important aspect of automatic interrogation of the data in that Names are combined with Words where they co-occur to create **Concepts** and **Topics** (e.g. *Africa Development*) to classify or 'code' data ■ The **Results** (final button with two options) executes a "Cluster" process usually calculating findings to be shown in two main ways: a) **Topic Guide** which provides a list of the most interconnected pairs of co-occurring Concepts to form **Topic** titles with associated textual segments or b) via a varied interactive statistical and visualized **Concept Map** area (Concept map is not available in **LexiReader**) (See Figure 3). ■ **Floating Help windows** appear at every stage to assist the user make choices.

Types of data Leximancer and LexiReader can accommodate Plain text, Microsoft Word DOC or DOCX, PDF, CSV (Excel export) and HTML. The developers state that text corpora of up to 2GB have been analysed so far, but there is no theoretical limit other than the utility of the results.

Closeness to data – Interactivity in Leximancer From the **Topic Guide** there is good one-click interactivity from lists of **Topics** to the 1 or 2 sentence passages in which they occur (proximity context can be user-defined). In addition, double clicking within a passage opens the highlighted passage within the source file ■ From the **Concept map** there is flexible interactivity between map and statistics and between text passages from tabs such the **Synopsis** (a statistical and interactive overview) and the **Summaries** (displaying exemplar text segments per data file). Initially a 'top example' is shown, from which topics can be viewed in more detail. In other tabs the nature of interactivity is related to quickly recalculated and re-ordered frequencies, based on user choices. See Figure 3.

Machine-based (auto) coding – Thesaurus – Concepts **Auto Tags** are created by default in Leximancer. Unless the user opts out, these are created by the auto-identification of regular structures e.g. the "Variables/Values" of speakers in a CSV file or they can also be based on File names or folders. However they are created, Auto Tags are mainly used to enable interrogation within e.g. files or speaker sections ■ **Concept seeds** (in both **Leximancer** and **LexiReader**) are automatically identified keywords based on frequencies and word correlation ■ The **Thesaurus** (also described as **Concept Learning**) goes beyond the concept seeds to learn from the data on the basis of likelihood of word co-occurrence with concept seeds in order to produce a weighted accumulation of evidence ■ **Concepts** are created therefore by the Thesaurus not on raw frequencies, which might include unimportant words, but on the more complex basis of the likely co-occurrence of important terms with other keywords; for instance in the example – the Blair Speeches dataset – the Concept seed *Climate* might combine with other seeds to become two new single Concepts, *Climate Emissions* and *Climate Conference*, but may also remain as the discrete concept *Climate* based on its importance. Similarly *Emissions* will also be a Concept. However, each Concept, whether apparently combined or discrete, continues to retrieve data based on other respective co-occurring **Concept seeds** generated in Thesaurus, even if not all remain

explicit in the Concept label. (See Illustration in Figure 2). ■ Importantly in the **Results/Topic Guide** dominant co-occurring pairs of Concepts are listed as the main **Topic** titles; below these the constituent parts are listed, retrieving respective data if clicked. These may, depending on frequency, include those Concept seeds combined by the Thesaurus stage (but not explicitly featuring in the Concept label) (see Figure 2) ■ The optional use of ***name-like*** concept seeds will be matched with ***word-like*** seeds to form other Concepts (if frequent) e.g. *Africa Development* now listed as a new more tightly focused combined Concept ■ **Machine *learning*** is affected by a whole range of settings - see more in Figure 1. ■ **Sentiment Lens** evaluates terms frequently indicating negative or positive sentiment and considers any negations (e.g. *that's bad* versus *that's not bad*) and their co-occurrence with extracted concepts.

Human intervention in "coding" In Leximancer and LexiReader 'coding' happens automatically by machine identification of frequent or recurring content, rather than the user creating and applying codes to data segments. However there are key ways that users can *affect* coding: first by varying settings at each stage and second by creating Compound Concepts ■ **Varying Settings** (see also Figure1). Settings include: limiting finds in Concept/Classification Settings, e.g. define the threshold of frequency whereby a concept seed or concept is counted or ignored; control the generality of finds, the detail of how statistics are calculated and how Project Output is presented. In the Concept Seeds Editor setting can be varied to identify concepts that occur at any frequency. Text Processing Settings determine how Leximancer should look for co-occurring concepts ■ **Compound Concepts** are created by the user out of combinations of Concept Seeds and are crucial to the role the user has in instructing Leximancer how to loosely collect data together or conversely make tighter connections. This happens via the use of Boolean operators during the Thesaurus stage. Unlike the automatic Thesaurus creation of combined concepts which always happens using the AND operator, Compound Concepts allow the generation of broader concepts based on e.g. the OR operator. Each Compound Concept is then defined as a single concept e.g. "Climate Or Emissions". (See Figure 2)

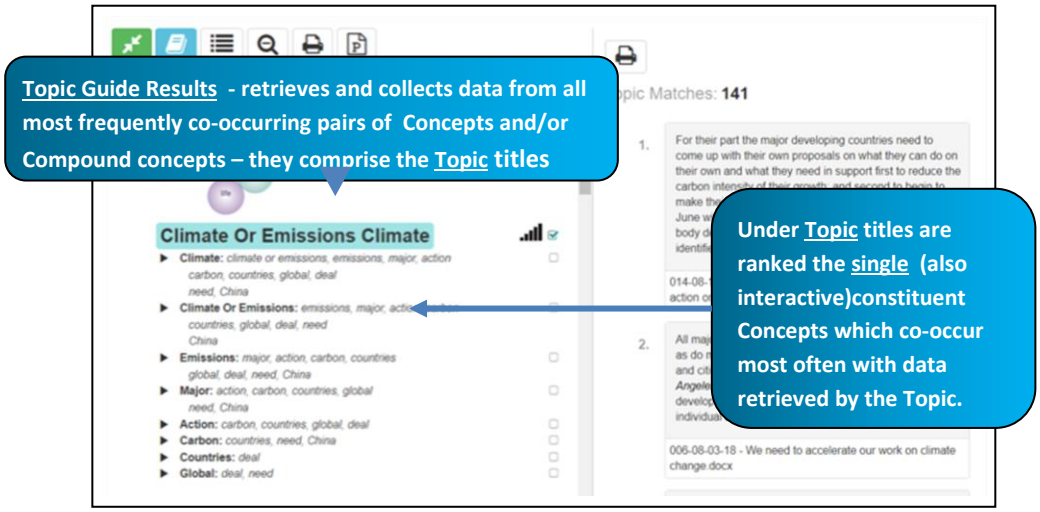

Figure 2 Topic Guide: showing dominant Topics (i.e. pairs of dominant co-occurring Concepts)

Retrieval & Interrogation in Leximancer – Results in the Topic Guide The **Topic Guide** is the main way to see how data has been categorized (see Figure 2) and is selected at the first Results type button – **Topic** titles comprise pairs of dominant co-occurring Concepts. The user can inspect a ranked list of words which make up each concept, and can drill down to interactively retrieved text to inspect the validity and nature of the induced abstract relationships. In this way, the Topic Guide creates a space, organized by topics in overarching frequency order, so that further sense-making by the researcher can take place ■ **Interrogation of the data** has already taken place in various ways: e.g. the matching of data chunks in which *name-like* words and *word-like* words have co-occurred and the collation of data chunks together because of frequent co-occurrence ■ **In Survey-type data** and with use of **Auto tags** (created from quantitative / descriptive fields or speaker identifiers,

files or folders) the user can make use of Compound Concepts by combining auto tags with Concepts – e.g. *Females and Work* will find passages where frequent co-occurring concept seeds e.g. *work, jobs, career, employ* etc., appear in that group. As usual in Leximancer, the threshold of required frequency and other settings can be user-defined.

Interrogation in Leximancer - Results in the Concept Map Visualisation The Concept Map tool is both visual and statistical and is the main way to interrogate frequency, relevance for different name-like words, auto tags and co-occurrence in the data ■ **Visualisation and Themes –** in the map view - concepts are clustered into higher-level 'themes'. 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 ■ **Clusters** - Concepts that co-occur often within the same (default) two-sentence coding block are visualized as closely linked. Clusters of concepts are grouped by theme circles. For example, in Figure 3 a cluster of conceptually related concepts is grouped by a most connected concept of *World*. The user has clicked on this theme to see its links emphasized. The themes are heat-mapped from reds to blue to indicate reducing connectedness to other concepts ■ Good access to relevant text passages from the right-hand pane ■ **The map** auto-arranges and can be visualized in various different views depending on researcher choices ■ **Bookmarks -** each view can be bookmarked to preserve it as part of the project management process.

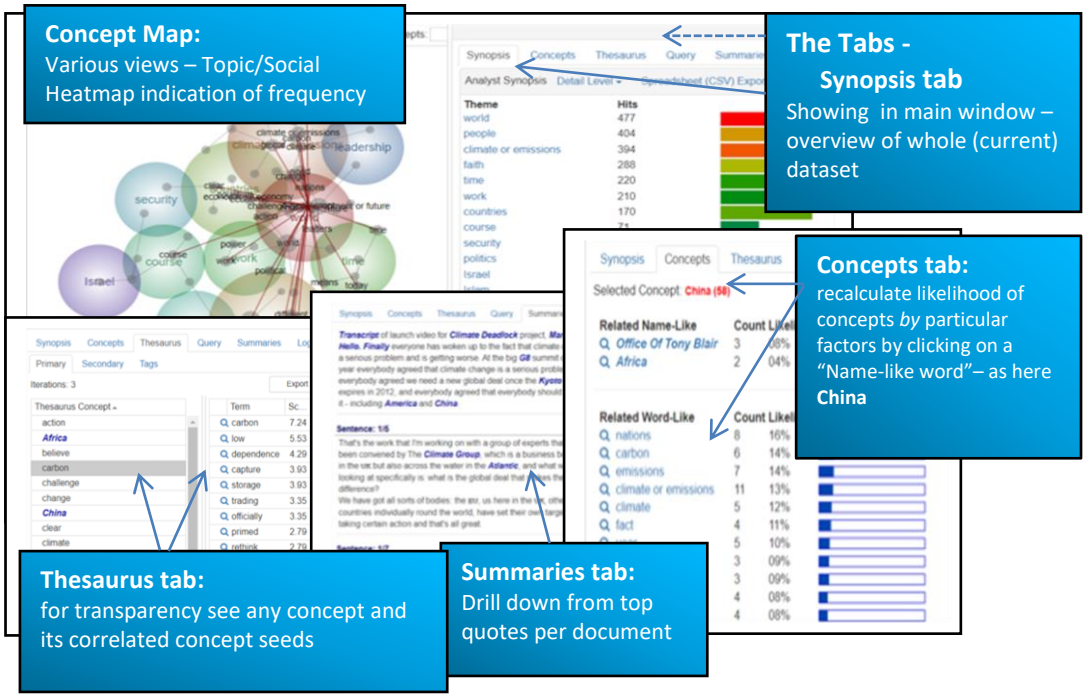

Figure 3. The Leximancer Concept Map – tabs provide different views and information

Statistics – the right hand view alongside the maps comprises a wide range of interactively relevant statistics and information featured at **Tabs.** They include: The **Synopsis** tab provides a statistical interactive overview with bar chart, the **Concepts** tab allows the researcher to click around the name-like Concepts (in this example, e.g. China or Africa) to use as factors by which to vary the ordering of *Likelihood* counts of each concept – instantly recalculable, the **Summaries and Thesaurus** tabs keep all connections transparent and provide extra ways to access text passages, the **Query** tab allows further complex content searches to take place.

Comments on Leximancer

- ***Content analysis*** *– researchers vary in their understanding of "content analysis"* [i]*. The developers are clear that Leximancer improves access to data and topics within the data but does not seek to enable, on its own, an interpretation of phenomena. However, the more the user becomes familiar with ways to vary the setttings the more the researcher can determine relevance and exactly what content the text mining process should be based on. Based on those decisions the researcher can use the ways content is presented, and the access to the data, to make sense and carry out further analysis and interpretation. That said, the researcher needs to have a clear idea of the basis and manner in which Results are presented or it may appear confusing.*

  i **Content Analysis in the context of what Leximancer does -** The tool is used for determining the presence of words or concepts in collections of textual documents. It is used for breaking down the material into a finite number of categories and relationships in order to quantify and visualise text and used for building a reproducible and computable model of a "complex and intuitive information space".

- ***In a straightforward, qualitative sense****, content and its dominant references are made very accessible, very quickly (in both Leximancer and LexiReader). Leximancer also provides a sophisticated combination of organized content-based and quantitative information about qualitative data. Certain tools, like the Sentiment Lens might be more relevant in a business environment, and in any case would require the researcher to be watchful and intervene where words have misleadingly positive or negative connotations.*

- ***Speed of processing –*** *the results of each run through the data are produced very quickly. Since basic processing is so fast, alternate selections of data and separate Models can help to develop comparisons across subsets of data.*

- ***Project management*** *– the user must establish clear procedures to track, log and save separate model building exercises and their settings.*

- ***Size of dataset*** *– clear advice is given that the Thesaurus stage (to go beyond the creation of Concept seeds) can be missed out with a very small dataset of just a few files. The user might usefully experiment with* ***LexiReader*** *but must do without the more statistical and visual Concept Map tools of* ***Leximancer****.*

- ***Topic Guide*** *–dominant correlations are the basis of the presentation of results with text here. This could be an easy fast way to explore important topics at the outset with any large dataset.*

- ***The Concept map*** *is based on relationships, frequency and correlation. Though part of it is visual its functionality is mainly statistical. The software determines the appearance and arrangement for each type of visualisation. The interactive range of possibilities to compare and contrast frequencies and counts in different ways is remarkable. The context of "Content Analysis" is always the dominant paradigm in Concept Map.*

- ***Transparency -*** *the set of related co-occurrent words in the Thesaurus for each Concept influences retrieval and is transparent. The user can check all related Thesaurus terms*

- ***Sense-making -*** *The sense-making opportunities for the researcher could be wholly qualitative – or quantitative or a triangulated mixture of both.*

- ***Writing tools*** *– These could be more extensive. There is an Add to Log facility near text segments. You can then choose to annotate the log to build a journal.*

- ***LexiReader*** *– is an efficient way to examine content. See Results in Topic Guide. There is no way to vary settings as there is in Leximancer, and no Concept Map Results. As a low budget option it is worth trying just to see frequent co-occurring content. The packages can be leased on a monthly basis*

References and Further Reading

Andrew E. Smith, Michael S. Humphreys *Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping* Behaviour Research Methods 2006 38, (2):pp 262-279

Cretchley J, Rooney D, Gallois C. *Mapping a 40-Year History With Leximancer: Themes and Concepts* Journal of Cross-Cultural Psychology. 2010;41(3):pp 318-328.

Leetaru, K. *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content,* 2012. London, Routledge.