

TextQuest 3.1 – Text Analysis Software

A detailed overview

April 2008

Author: Dr. Harald Klein

Development, consulting, training, and distribution:

Social Science Consulting
Dr. Harald Klein
Lutherstr. 2
49082 Osnabrück
Germany

Tel/Fax.: +49 541 1819492
email: info@textquest.de
<http://www.textquest.de>

1 Overview and design principles

TextQuest is a program package for the analysis of texts in the humanities and the social sciences. It runs under all Windows versions newer than Windows 98.

It is used in many sciences like media science, sociology, political science, literature, law, education, medicine, religion, and many more.

TextQuest is language independent. Currently all letter based languages can be analysed, later versions will also be able to analyse syllable based languages like Chinese, Japanese, or Korean. Language dependent parts like the sort order table, tables for negation indicators, or tables of vocabularies used by readability formulas are plain texts and can be viewed and altered.

The design of TextQuest is user friendly. The user has the complete control on what is done with which files. The file names are generated by TextQuest using default file extensions. All file names are shown and can be altered if necessary, and this concept has the advantage that the user does not have to go through the open/close file dialogues. The log file contains all file names, options, and dates, so a complete control of all analyses done with a text is documented and can be easily pasted into a word processor.

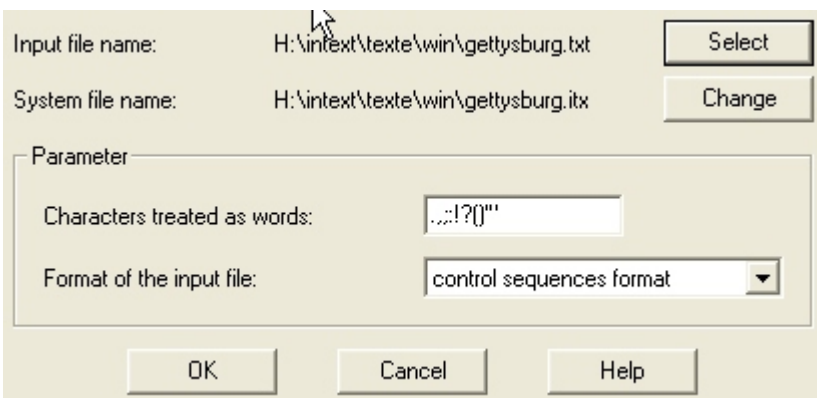
The following applications can be performed by TextQuest:

- list of words, sorted by alphabet or by frequency, also with exclusion lists (STOP-words) that come with the program for English, German, and French
- list of word sequences so that phrases can be detected and counted
- list of word permutations, this shows which words co-occur within a text unit
- comparison of vocabularies, these can be more than just 2
- KWICs - key word in context with variable line length
- SITs - search patterns in text unit (extended KWIC)
- content analysis with powerful features like interactive coding, control files, and negation detection
- category system manager
- control of multiple search patterns
- readability analysis with 68 formulas for English, German, Spanish, French, Dutch, Danish, and Swedish texts

2 Input formats

Each digitized (machine readable) text can be easily converted for TextQuest, also converting texts from text processors like MS-WORD, OpenOffice, and other text processors can be done easy and fast. There are 6 different input formats (2 with and 4 without control sequences) available. Additionally external variables of the text can be defined, up to 50 are possible. These can be numbers or strings up to 10 characters each. Using sentence format TextQuest is able to split text into grammatical sentences.

External variables can be added to the text, up to 50 are possible, each having a maximum length of 10 characters. The values of external variables can be used for the identification and the filtering of parts of a text. If a content analysis is performed, the external variables can be used for statistical analyses also. According to the input format, the values of external variables can be assigned automatically, e.g. line numbers, paragraph numbers, or sentence numbers.



The image shows a dialog box for configuring input parameters. It has a light beige background and a standard Windows-style layout. At the top, there are two rows of text labels followed by text boxes and buttons. The first row is 'Input file name:' with the path 'H:\intext\texte\win\gettysburg.txt' and a 'Select' button. The second row is 'System file name:' with the path 'H:\intext\texte\win\gettysburg.itx' and a 'Change' button. Below these is a section titled 'Parameter' enclosed in a rounded rectangle. Inside this section, there are two rows: 'Characters treated as words:' with a text box containing '.,:;!?"' and 'Format of the input file:' with a dropdown menu showing 'control sequences format'. At the bottom of the dialog box, there are three buttons: 'OK', 'Cancel', and 'Help'.

3 Word lists, word sequences, and word permutations

The screenshot shows the TextQuest software interface with the following sections:

- Input/Output files:**
 - Name of system file: H:\intext\texte\win\gettysburg.itx (Change button)
 - File of word sequences: H:\intext\texte\win\gettysburg.ws (Change button)
- Parameters:**
 - Process all text units
 - Reverse vocabulary
 - Case folding enabled
 - Sorted by: First string Last string
 - Justification of vocabulary: left justify right justify
 - Length of string: 40
 - Number of strings: 2
- Selection criteria:**
 - Minimum length: 1
 - Minimum frequency: 1
 - Maximum length: 80
 - Maximum frequency: 100000
- Name of exclusion list:** [none] (Change button) (Clear button)

At the bottom are buttons for OK, Cancel, and Help.

A word list sorted by alphabet gives an overview of all strings occurring in the text and their frequency, sort order tables can be used so that umlauts, characters with diacritics or accents are sorted properly. Also case folding can be enabled or disabled. One can also exclude strings due to their frequency (using absolute or relative values) and their length (in characters). Also using a list of exclusion words (STOP-words) is possible, these strings are excluded from further processing. Statistics provided include the TTR, also calculating the dynamics of the TTR is available for the whole text or for a sample of it.

The use of the word list in a content analysis is to find strings that can be used for the building of categories. The word list contains only single words, but no combinations of words. Word sequences and word permutations can be generated with TextQuest, thus allowing to define search patterns that consist of more than a word or any part of it. Word sequences consist of sequences of words (at least 2) with a variable number of words, so phrases can be counted. Word permutations are two-word sequences where each word of a text unit is combined with each word that follows, and this allows you to see which words co-occur within a text unit.

4 Content analysis

The screenshot displays the TextQuest software interface, which is organized into several sections:

- Input files:** Contains three fields with corresponding 'Change' buttons:
 - Name of system file: H:\intext\texte\win\gettysburg.itx
 - File of search patterns: H:\intext\texte\win\gettysburg.dic
 - File of category labels: H:\intext\texte\win\gettysburg.lab
- Numeric results:** Contains two checked checkboxes with corresponding 'Change' buttons:
 - File of codes as counters: H:\intext\texte\win\gettysburg.tab
 - File of codes in sequences: H:\intext\texte\win\gettysburg.vec
- Parameters:** Contains one checked checkbox:
 - Process all text units
- Protocol files and interactive (i) modes:** Contains five unchecked checkboxes, each with a corresponding file path and a 'Change' button:
 - all search patterns: H:\intext\texte\win\gettysburg.clg
 - ambiguous search patterns: H:\intext\texte\win\gettysburg.ctx
 - negated search patterns: H:\intext\texte\win\gettysburg.ntx
 - overlapping text passages: H:\intext\texte\win\gettysburg.otx
 - uncoded text units: H:\intext\texte\win\gettysburg.rtx
- Options:** Contains several dropdown menus and text input fields:
 - command setup for: [dropdown]
 - overlapping text passages: block [dropdown]
 - all search patterns: ambiguous [dropdown]
 - distance of negation: before 2 [text input] after [text input]

At the bottom of the window are three buttons: OK, Cancel, and Help.

TextQuest was originally developed for computer aided content analysis. Search patterns for a category system can be words, parts of it, word sequences, and sequences of (parts of) words (so called word root chains). These are strings - up to 6 - that must co-occur within the same text unit, one can specify their sequence and their distance. Also wild cards may be used. Ambiguous and negated search patterns can be coded interactively on the screen, the coding process can be controlled with several log files:

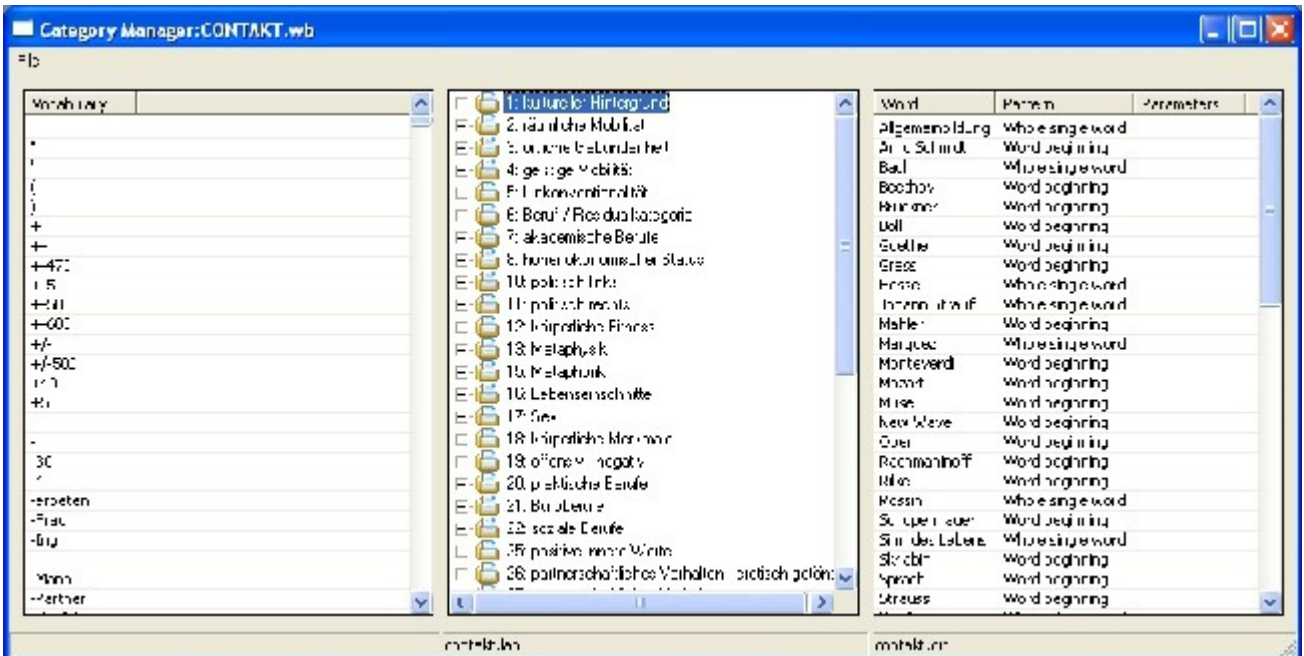
- file of text units containing potentially ambiguous search patterns or coded text units
- file of uncoded text units
- file of text units with negations
- complete control of the coding process

TextQuest works with category labels that forces you to document your categories and makes the usage more comfortable because these labels are used in interactive coding, for the variable labels in the syntax file for statistical data analysis software, and in the log files.

Negation is often a problem, and can result in too many codings and thus overestimate a category. TextQuest can detect negated search entries, if indicators for negations are found before or after the search pattern. The negation indicators are stored in files and can be altered (e.g to adapt them to other languages). Also multiple negations are detected.

The differences between automatic and interactive coding is measured with the ICRC (interactive coding reliability coefficient). Also the generation of syntax files for SAS, SPSS, and SimStat enlighten the statistical analysis.

5 The category system manager



The category system editor is new since version 3.0 and allows the easy constructing and maintaining of a category system. The left windows shows you a base file, e.g. a word list, a phrase list, or the uncoded text units of a content analysis (this is a very convenient for coding answers of open ended questions). The middle window is the working window that allows you to add, change, and delete categories and/or search patterns.

6 Readability analysis

Input files

Name of System file: H:\intext\texte\win\gettysburg.itx Change

Rapport files

Name of syllable control file: H:\intext\texte\win\gettysburg.sco Change

Name of foreign words control file: H:\intext\texte\win\gettysburg.fwp Change

Parameters

Process all text units

Too long sentences words Too long words characters

Too many brackets brackets Too many foreign words foreign words

Too complex sentences sentence markers

OK Cancel Help

The readability analysis calculates the results of readability formulas in the form of reading class, reading age, or an index value. Currently 68 formulas for 7 languages are implemented. The user can specify criteria that make texts harder to read like too long words, too long sentences, or too complex sentences. A log file consists of all sentences that meet at least one of the criteria mentioned above.

7 Result manager

The result manager provides easy access to all results of each analysis. An editor is invoked to display and edit the selected result file, and due to the default file name system the user is not bothered with entering or selecting file names.



8 Help system

TextQuest has a help system that is context sensitive, but there is also a part that serves as a tutorial. Parts of the manual are integrated in the help system. The help system can be invoked in each application and explains the available parameters, options, and their meaning. Also many parts of the manual can be found in the help system.

9 Why TextQuest? – Some unique features

This section describes features of TextQuest that other programs do not have or only partially have.

- One of the most time consuming processes in computer aided text analysis is the preparation of the text. TextQuest provides several input formats so that editing task are limited to the necessary minimum. A unique feature is that texts are splitted into grammatical sentences, a feature that is very useful for readability analysis.
- Word lists can be sorted by alphabet ascending using sort order tables, and sorted by frequency descending so that the most frequent words occur first.
- Word sequences show you how often phrases consisting of several words occur within a text. Also the features for a word list can be used like a sort order table.
- The co-occurences of words within a text unit can be detected with the word permutation list.
- The vocabulary comparison module can compare multiple vocabularies, e.g. 5 speeches on the same subject and inspecting common words and/or exclusive words.
- Several meanings of a word can be inspected using a KWIC (key-word-in-context) list. TextQuest's implementation offers a variable line length which allows the specification of the required context.
- The whole text unit can be inspected using the SIT-feature. Each search pattern is shown in the whole text unit.
- The main module is the content analysis with many unique features:
 - The definition of search patterns: one can define single words or any part of it as a search pattern, or a whole phrase up to 200 characters, or specify a number of words (or any part of it) that must co-occur within a text unit. The latter is a feature no other content analysis program has.
 - The log files give you the complete control of the coding process and are a tool to check the validity of the category system.
 - Interactive coding allows you to code ambiguous and/or negated search patterns on the screen. So unique search patterns are coded automatically, whereas problematic search patterns are coded by the researcher.
 - Ambiguity of a search pattern can lead to false codings. These search patterns can be marked in the category system, and the text unit where they occur can be written to an output file. Also interactive coding is possible.
 - TextQuest is the only program that can detect negated search patterns and offers adaequate handling of this problem: interactive coding and/or a log file. The algorithm can be adapted to other languages than English or German.

- The category system can be tested. The test checks if search patterns occur more than once in a category system or whether a search pattern is part of another one (this can result in the overestimation of the category).
- The category system manager allows you to construct and maintain dictionaries easily. It is very useful to code answers to open ended questions.
- Readability analysis is using readability formulas to test whether a text is understandable. TextQuest has 68 different formulas for 7 languages implemented, so that nearly every text – dependent on language and text genre – can be analysed. A log file consists of parts of the text that might be more difficult to understand, the criteria can be specified by the user (length of sentences and words, complexity of sentence).

10 More information

The website <http://www.textquest.de> provides more information on TextQuest, screenshots, and a free test version. TextQuest is available with menus and information messages in English, German, or Spanish.

If you are looking for information on text analysis software in general, choose <http://www.textanalysis.info> as your starting page. It gives you links to software providers, free test or demo versions, and some comments. The descriptions of the software are classified into categories, so you can see what programs provide the same or similar functionality.