

TextQuest 4.0 – Text Analyse Software

Ein detaillierter Überblick

August 2011

Autor: Dr. Harald Klein

Entwicklung, Consulting, Training und Vertrieb:

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

Tel.: 0541 969 4806
email: info@textquest.de
<http://www.textquest.de>

1 Überblick und Designprinzipien

TextQuest ist ein Programmpaket für die Analyse von Texten und wird in den Geistes- und Sozialwissenschaften benutzt. Es erfordert Windows 98 oder neuere Versionen von MS-Windows oder Apple Mac OS-X ab Version 10.4.

Einsatzgebiet sind viele Wissenschaften wie Medienwissenschaften, Soziologie, Politikwissenschaften, Literaturwissenschaften, Recht, Pädagogik, Medizin, Religion und viele mehr.

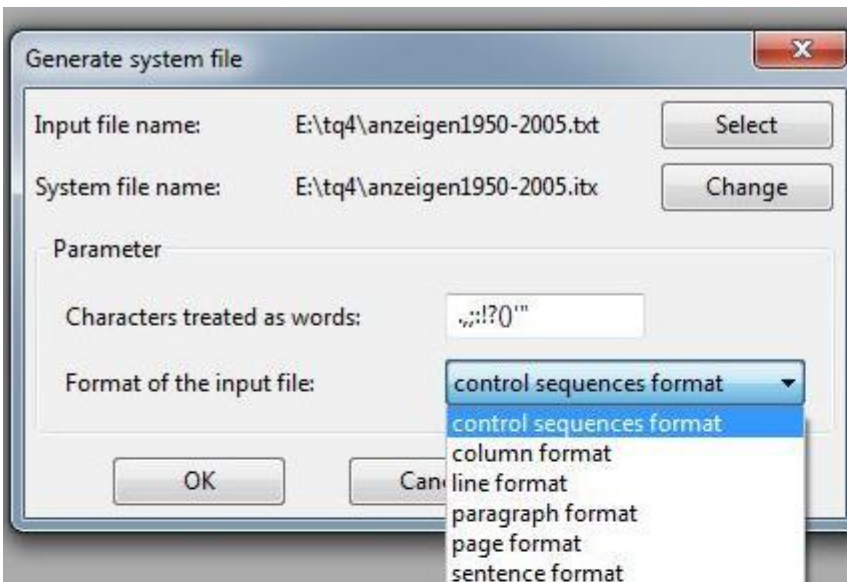
TextQuest ist sprachunabhängig. Zur Zeit können alle buchstabenbasierten Sprachen analysiert werden, zukünftige Versionen werden silbenbasierte Sprachen wie Chinesisch, Japanisch oder Koreanisch verarbeiten können. Sprachabhängige Teile wie die Sortiertabelle, die Tabellen für Negationsindikatoren oder Vokabellisten für Lesbarkeitsformeln sind reine Textdateien und können betrachtet und geändert werden.

Das Design von TextQuest ist benutzerfreundlich. Benutzer haben volle Kontrolle darüber, was mit welchen Dateien gemacht wird. Die Dateinamen werden von TextQuest auf der Basis der Projektkennung generiert. Alle Dateinamen werden angezeigt und können bei Bedarf geändert werden. Dieses Konzept hat den Vorteil, dass man nicht durch den sonst üblichen Dateidialog muss. Die Logdatei enthält alle Dateinamen, Optionen und Datumsangaben, so dass eine komplette Kontrolle über alle durchgeführten Analysen erfolgt, die auch einfach in ein Textverarbeitungsprogramm kopiert werden kann.

Die folgenden Anwendungen kann TextQuest durchführen:

- Häufigkeiten von Wörtern, sortiert nach Alphabet oder Häufigkeit, auch mit Ausschlusslisten (STOP-Wörter) für Deutsch, Englisch und Französisch
- Häufigkeiten von Wortsequenzen, so dass Phrasen entdeckt und ausgezählt werden
- Liste von Wortpermutationen, diese zeigen, welche Wörter zusammen in einer Texteinheit vorkommen
- KWICs - key word in context mit variabler Zeilenlänge
- SITs - Suchbegriffe innerhalb einer Texteinheit (erweitertes KWIC)
- seit Version 3.0: Vokabularvergleiche mit beliebig vielen Vokabularen
- Inhaltsanalyse mit mächtigen Möglichkeiten wie interaktives Codieren, Kontrolldateien und Erkennen von Negation
- seit Version 3.0: Editor zum Erstellen und Bearbeiten von Kategoriensystemen
- Kontrolle von mehrfachen Suchbegriffen
- Lesbarkeitsformeln mit 68 Formeln für Deutsch, Englisch, Spanisch, Französisch, Holländisch/Flämisch, Dänisch und Schwedisch
- seit Version 4.0: Texte in Latin-1 oder UTF-8 Codierung

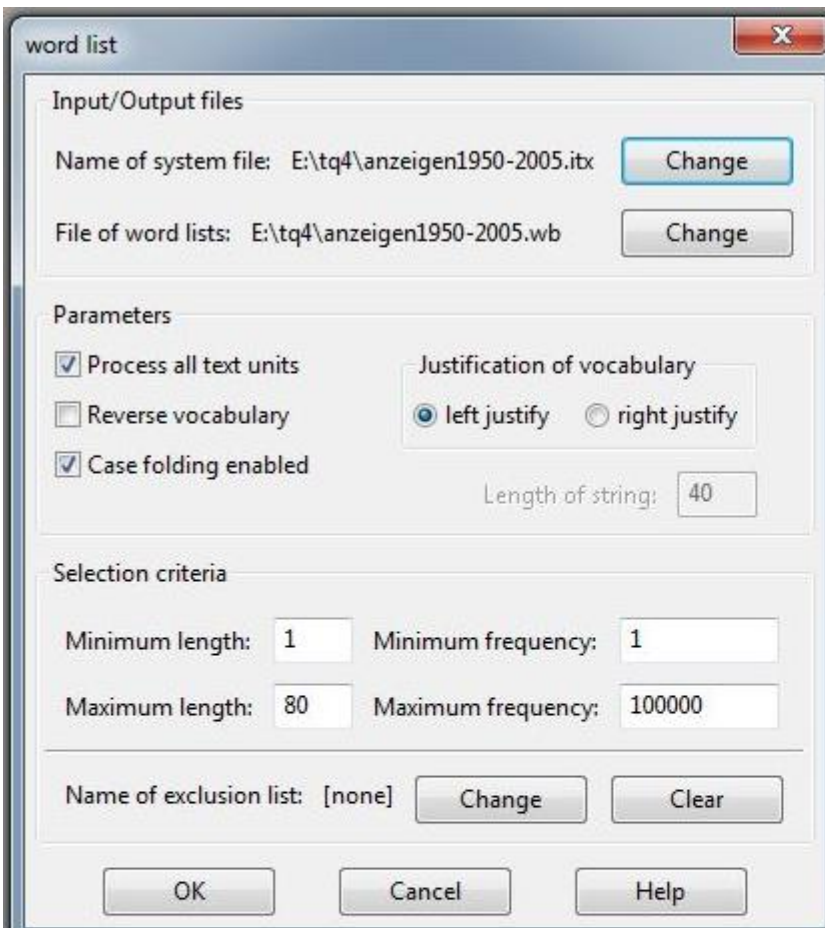
2 Eingabeformate



Jeder digitalisierte Text kann einfach für TextQuest verarbeitbar gemacht werden, auch die Konvertierung von Text aus WORD, WordPerfect und anderen Textverarbeitungsprogrammen geht einfach und schnell. Es gibt 6 verschiedene Eingabeformate (2 mit und 4 ohne Steuerungssequenzen). Zusätzlich können externe Variablen des Textes definiert werden, bis zu 50 sind möglich. Dies können Zahlen oder Zeichenketten mit einer maximalen Länge von 10 Zeichen sein. Beim Satzformat zerlegt TextQuest den Text in grammtikalische Sätze.

Die Werte der externen Variablen dienen zur Identifikation und zur Filterung von Textteilen. Bei einer Inhaltsanalyse können die externen Variablen auch für statistische Analysen benutzt werden. Abhängig vom Eingabeformat können die Werte der externen Variablen auch automatisch zugewiesen werden, z.B. Zeilenzähler, Absatzzähler oder Satzzähler.

3 Wörterlisten, Wortsequenzen und Wortpermutationen



Eine alphabetisch sortierte Wörterliste gibt einen Überblick über alle Zeichenketten und ihre Häufigkeiten im Text, Sortiertabellen können benutzt werden, so dass Umlaute, Zeichen mit Diakritika oder Akzenten richtig sortiert werden. Auch Unterschiede in der Groß-/Kleinschreibung können berücksichtigt oder nicht berücksichtigt werden. Ebenso können Zeichenketten auf der Basis ihrer Häufigkeit (absolut oder relativ) und/oder seiner Länge (in Zeichen) ausgeschlossen werden. Die Verwendung von Ausschlusslisten (STOP-Wörter) ist ebenfalls möglich, auch in Kombination mit Häufigkeit/Länge. Die Statistiken enthalten den TTR, und auch das Vokabularwachstum auf der Basis des TTRs wird berechnet, für den gesamten Text oder eine Stichprobe.

Wörterlisten können bei einer Inhaltsanalyse als Basis für ein Kategoriensystem benutzt werden, dies wird auch durch den seit Version 3.0 eingebauten Kategorieneditor revolutioniert, mit dem sich einfach und effizient arbeiten lässt. Wörterlisten werden in einer Inhaltsanalyse dazu benutzt, um Zeichenketten zu finden, um Kategorien zu finden. Wörterlisten enthalten nur einzelne Wörter, aber keine Kombinationen von Wörtern oder Phrasen. Diese können von TextQuest erzeugt und als Suchbegriffe in einer Inhaltsanaly-

se verwendet werden. Wortsequenzen besteht aus Sequenzen von mindestens 2 Wörtern, so dass diese gezählt werden können. Wortpermutationen sind Zweiwortsequenzen, wobei jedes Wort einer Texteinheit mit jedem ihm folgenden Wort kombiniert wird, so dass man sehen kann, welche Wörter innerhalb einer Texteinheit gemeinsam vorkommen.

4 Inhaltsanalyse

Content Analysis

Input files

Name of system file: E:\tq4\anzeigen1950-2005.itx Change

File of search patterns: E:\tq4\anzeigen1950-2005.dic Change

File of category labels: E:\tq4\anzeigen1950-2005.lab Change

Numeric results

File of codes as counters E:\tq4\anzeigen1950-2005.tab Change

File of codes in sequences E:\tq4\anzeigen1950-2005.vec Change

Parameters

Process all text units

Protocol files and interactive (i) modes

| | | | |
|---------------------------------------------------------------|----------------------------|------------------------------|---------------------|
| <input checked="" type="checkbox"/> all search patterns | <input type="checkbox"/> i | E:\tq4\anzeigen1950-2005.clg | Change |
| <input checked="" type="checkbox"/> ambiguous search patterns | <input type="checkbox"/> i | E:\tq4\anzeigen1950-2005.ctx | Change |
| <input checked="" type="checkbox"/> negated search patterns | <input type="checkbox"/> i | E:\tq4\anzeigen1950-2005.ntx | Change |
| <input checked="" type="checkbox"/> overlapping text passages | <input type="checkbox"/> i | E:\tq4\anzeigen1950-2005.otx | Change |
| <input checked="" type="checkbox"/> uncoded text units | | E:\tq4\anzeigen1950-2005.rtx | Change |

Options

command setup for SPSS overlapping text passages block

all search patterns ambiguous distance of negation: before 2 after 2

TextQuest wurde ursprünglich für die computerunterstützte Inhaltsanalyse entwickelt. Suchbegriffe für ein Kategoriensystem können Wörter, Teile von Wörtern, Wortsequenzen und Wortstammfolgen sein, die aus maximal 6 Zeichenketten bestehen können. Man kann angeben, ob diese innerhalb einer Texteinheit direkt hintereinander, hintereinander mit beliebigem Abstand oder gemeinsam ohne Beachtung der Reihenfolge vorkommen sollen. Auch Jokerzeichen können benutzt werden. Mehrdeutige und negierte Suchbegriffe können interaktiv am Bildschirm codiert werden, und der Codiervorgang kann mit verschiedenen Kontrolldateien überwacht und validiert werden:

- Datei der Texteinheiten, die potenziell mehrdeutige Suchbegriffe oder codierte Texteinheiten enthalten
- Datei der nicht codierten Texteinheiten
- Datei der Texteinheiten mit Negationen
- komplette Kontrolle des Codiervorgangs

TextQuest arbeitet mit Kategorienetiketten und zwingt somit den Benutzer, die Kategorien zu dokumentieren und so komfortabler zu arbeiten, weil die Etiketten in der interaktiven Codierung benutzt werden genauso wie als Variablenetiketten bei der statistischen Auswertung und in der Logdatei.

Negation von Suchbegriffen ist oft ein Problem und kann in zuvielen Codierungen und damit einer Überzählung einer Kategorie resultieren. TextQuest kann negierte Suchbegriffe erkennen, wenn Indikatoren für Negation vor und/oder nach einem Suchbegriff gefunden werden. Die Negationsindikatoren befinden sich in reinen Textdateien und können geändert werden (z.B. auf andere Sprachen angepasst werden). Auch mehrfache Negationen werden erkannt.

Die Unterschiede zwischen automatischer und interaktiver Codierung kann mit dem ICRC (interaktiver Codier-Reliabilitäts-Koeffizient) gemessen werden. Für die weitere statistische Auswertung werden Syntaxdateien für SAS, SimStat und SPSS erzeugt.

5 Lesbarkeitsanalyse

Readability

Input files

Name of System file: E:\tq4\anzeigen1950-2005.itx Change

Rapport files

Name of syllable control file: E:\tq4\anzeigen1950-2005.sco Change

Name of foreign words control file: E:\tq4\anzeigen1950-2005.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

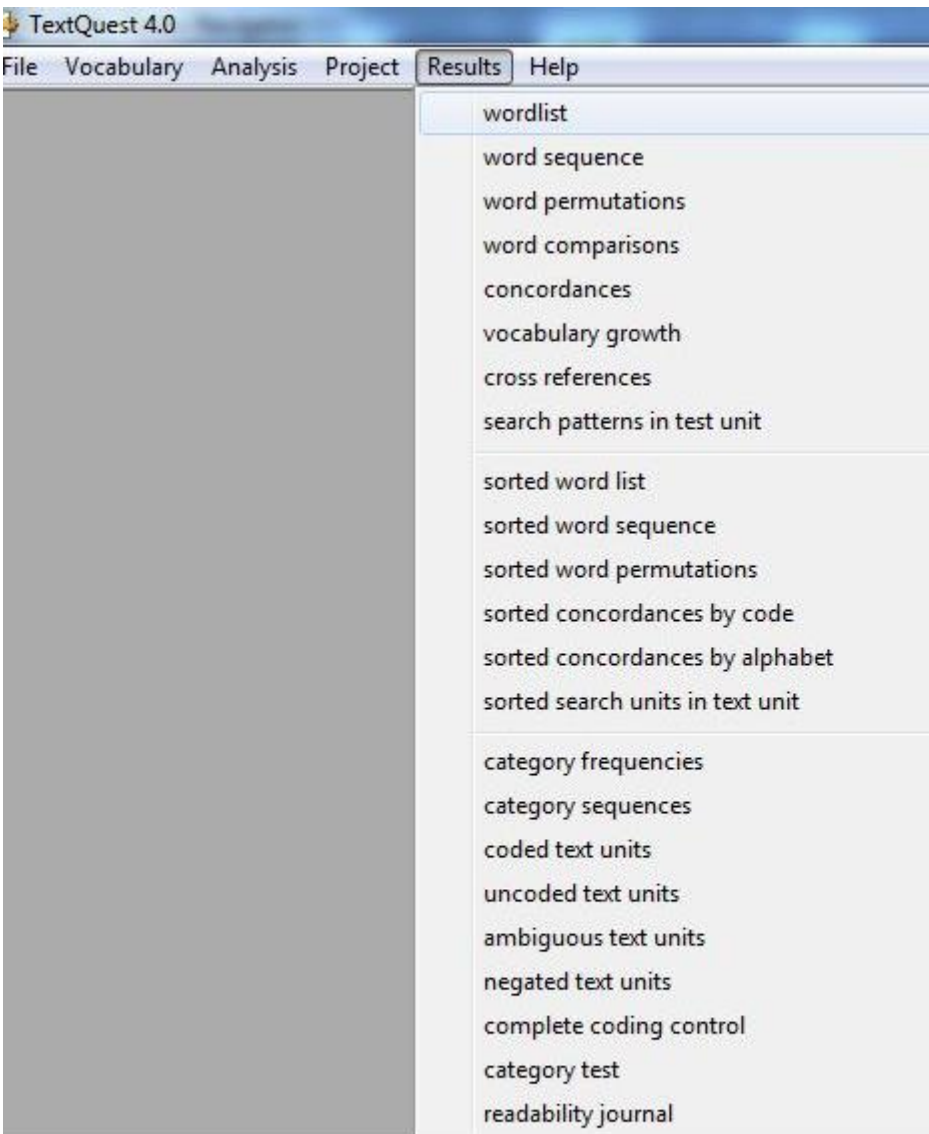
Too complex sentences sentence markers

OK Cancel Help

Die Lesbarkeitsanalyse berechnet Lesbarkeitsformeln und gibt als Ergebnis Leseklasse, Lesealter oder einen Index aus. Zur Zeit sind 68 Formeln für 7 Sprachen implementiert. Benutzer können Werte für Textmerkmale angeben, die den Text schwerer zu verstehen machen wie zu lange Wörter, zu lange Sätze oder zu komplexe Sätze. Diese werden in eine Datei geschrieben, so dass man auch besonders bei eigenen Texten feststellen kann, wo genau verständlicher formuliert werden kann.

6 Ergebnismanager

Der Ergebnismanager ermöglicht einen einfachen Zugang zu den Ergebnissen jeder Analyse. Ein Editor wird aufgerufen, damit man die entsprechende Datei bearbeiten kann, und wegen des Dateinamenvergabesystems muss man sich nicht durch Dateiöffnungsdialog hangeln.



7 Hilfesystem

TextQuest hat ein kontextsensitives Hilfesystem, und es enthält ein Tutorial. Teile des Handbuches sind im Hilfesystem integriert. Das Hilfesystem kann in jeder Anwendung aufgerufen werden und erklärt die verfügbaren Parameter, Optionen und deren Bedeutung.

8 Warum TextQuest? – Einige einmalige Eigenschaften

Dieser Abschnitt beschreibt Eigenschaften von TextQuest, die anderen Programme gar nicht oder nur teilweise haben.

- Eine der zeitaufwändigsten Arbeiten einer computerunterstützten Inhaltsanalyse ist die Vorbereitung des Textes. TextQuest stellt verschiedene Eingabeformate bereit, so dass Editierarbeiten auf das notwendige Minimum reduziert werden. Ein besonderes Merkmal ist, dass TextQuest Texte in grammatikalische Sätze zerlegen kann; das wird in der Lesbarkeitsanalyse gebraucht.
- Vokabulare können alphabetisch aufsteigend sortiert werden, dabei können Sortiertabellen verwendet werden, damit Wörter mit Akzenten oder Diakritika korrekt einsortiert werden. Auch eine Sortierung nach Häufigkeit absteigend ist möglich, so dass die häufigsten Wörter zuerst genannt werden.
- Wortsequenzen zeigen oft Phrasen, die aus mehreren Wörtern bestehen und als Suchbegriffe für eine Inhaltsanalyse geeignet sind. Auch für Wortsequenzen können Sortiertabellen und dieselben Ausschlusskriterien wie für Vokabulare verwendet werden.
- Das gemeinsame Auftreten von Wörtern innerhalb einer Texteinheit zeigt eine Liste der Wortpermutationen.
- Mehrere Bedeutungen eines Wortes können durch die Inspektion von KWICs (keyword-in-context) Listen erfolgen. TextQuests Implementation bietet eine variable Zeilenlänge zur Bestimmung des relevanten Kontextes.
- Ganze Texteinheiten können durch SITs (Suchbegriffe in Texteinheit) als Kontext angegeben werden.
- Der Wortschatzvergleich ermöglicht den Vergleich beliebig vieler Vokabulare, das kann sonst kein anderes Programm.
- Das Hauptmodul ist die Inhaltsanalyse mit vielen einzigartigen Möglichkeiten:
 - Die Definition von Suchbegriffen: als Suchbegriffe kann ein Wort oder jeder Teil davon definiert werden, aber auch ganze Phrase bis zu 200 Zeichen Länge. Auch die Definition von Wörter (oder Teilen davon), die gleichzeitig in einer Texteinheit vorkommen müssen, ist möglich und einzigartig.

- Die Logdateien bieten die komplette Kontrolle über den Codierprozess und ermöglichen so die Validierung des Kategoriensystems.
 - der Kategorieneditor erleichtert die Konstruktion von Kategoriensystemen erheblich. Auch bereits existierende Kategoriensysteme können bearbeitet werden, dabei können aus Vokabularen oder der Datei der nicht codierten Texteinheiten Suchbegriffe interaktiv gebildet werden.
 - Die interaktive Codierung erlaubt das Codieren von mehrdeutigen und/oder negierten Suchbegriffen am Bildschirm. Eindeutige Suchbegriffe können so automatisch codiert werden, während problematische Suchbegriffe durch den Forscher codiert werden können.
 - Mehrdeutige Suchbegriffe können zu falschen und zu vielen Codierungen führen. Diese Suchbegriffe kann man im Kategoriensystem markieren, und die entsprechenden Texteinheiten können in einer Ausgabedatei protokolliert werden. Auch die interaktive Codierung von mehrdeutigen Suchbegriffen ist möglich.
 - TextQuest ist das einzige Programm, das negierte Suchbegriffe findet und entsprechend codieren kann: interaktiv und/oder Protokollierung in einer Logdatei. Der Algorithmus kann auf andere Sprachen als Deutsch und Englisch angepasst werden.
 - Das Kategoriensystem kann daraufhin getestet werden, ob Suchbegriffe mehrfach vorkommen oder ob Suchbegriffe in anderen Suchbegriffen enthalten sind und so zu zu vielen Codierungen führen können.
- Die Lesbarkeitsanalyse benutzt Formeln um zu testen, ob ein Text verständlich ist. TextQuest berechnet die Werte für 68 Formeln für 7 Sprachen, so das nahezu jeder Text – abhängig von Sprache und Genre – analysiert werden kann. Eine Logdatei enthält alle Textteile, die schwer verständlich sein könnten, die Kriterien dafür kann man selbst definieren (Anzahl der Wörter pro Satz, Anzahl der Zeichen pro Wort). Andere Programmen bieten weit weniger Lesbarkeitsformeln an und bieten dazu auch keine Standardisierung an, damit beliebig große Texte analysiert werden können.

9 Mehr Information

Im Internet unter <http://www.textquest.de> finden Sie weitere Informationen über TextQuest, Bildschirmfotos und eine kostenlose Testversion. TextQuest gibt es mit Menus und Hilfesystem in Deutsch, Englisch und Spanisch; Handbücher sind in Deutsch und Englisch verfügbar.

Wenn Sie generell an Informationen über Textanalysesoftware interessiert sind, besuchen Sie <http://www.textanalysis.info> und folgen Sie dort den Links zu Herstellern, kostenlosen Test- oder Demoversionen und einigen Kommentaren. Die Beschreibungen der Software sind in Kategorien unterteilt, so dass Sie sehen, welche Programme eine gleiche oder ähnliche Funktionalität haben.