Co-occurrence Tools

The Cooccurrence Explorer allows to ask a different type of questions. Using this tool, you can ask ATLAS.ti to show you all codes that co-occur across all of your primary documents. The result is a cross-tabulation of all codes.

As compared to the Query Tool where the user has to determine and select codes or code families and the appropriate operator, the Co-occurrence Explorer by default looks for all codes that co-occur in the margin area combining the operators WITHIN, ENCLOSES, OVERLAPS, OVERLAPPED BY and AND.

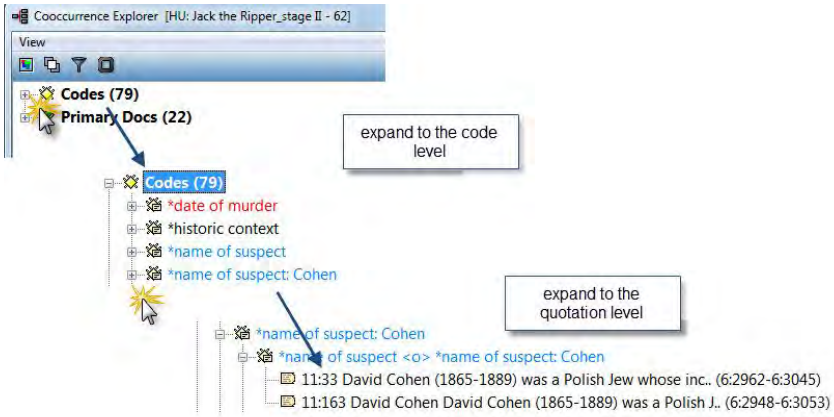
Instead of cross-tabulating all project codes, it is often more meaningful to apply filters for certain codes and documents in order to concentrate on a more specific set of concepts. The output of the Cooccurrence Explorer can be displayed in a tree view or as a data matrix. Below you see an example for both.

How To Open The Co-occurrence Tools

Select **ANALYSIS / CODE COOCCURRENCE TREE** or **CODE COOCCURRENCE TABLE.**

The Co-occurrence Tree Explorer

When running the tree explorer, you only see the root objects when it opens. Open the branches by clicking on the **+** sign to see the the cooccuring codes on the first level and the associated quotations on the second level.



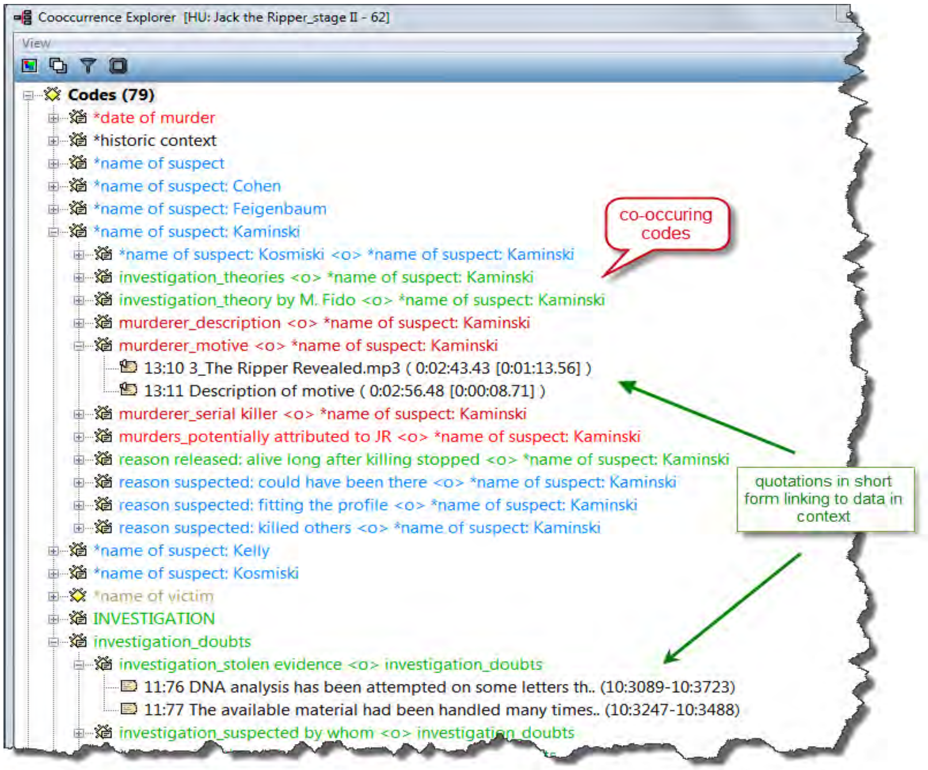
11:33 David Cohen (1865-1889) was a Polish Jew whose inc.. (6:2962-6:3045)

11:163 David Cohen David Cohen (1865-1889) was a Polish J.. (6:2948-6:3053)

Figure 241: Expanding to code and quotation level in the tree explorer

The same option is available for primary documents. If you expand the branch for Primary Docs, you can see which codes have been applied to this PD. Further, you can expand to the quotation level to look at the material coded there.

Let's take a look at a potential question that the Cooccurrence Tree Explorer can answer. The example is based on the Jack the Ripper Stage II project that you find in the samples folder.



```
Codes (79)
    *date of murder
    *historic context
    *name of suspect
    *name of suspect: Cohen
    *name of suspect: Feigenbaum
    *name of suspect: Kaminski
        *name of suspect: Kosmiski <o> *name of suspect: Kaminski
        investigation_theories <o> *name of suspect: Kaminski
        investigation_theory by M. Fido <o> *name of suspect: Kaminski
        murderer_description <o> *name of suspect: Kaminski
        murderer_motive <o> *name of suspect: Kaminski
            13:10 3_The Ripper Revealed.mp3 ( 0:02:43.43 [0:01:13.56] )
            13:11 Description of motive ( 0:02:56.48 [0:00:08.71] )
        murderer_serial killer <o> *name of suspect: Kaminski
        murders_potentially attributed to JR <o> *name of suspect: Kaminski
        reason released: alive long after killing stopped <o> *name of suspect: Kaminski
        reason suspected: could have been there <o> *name of suspect: Kaminski
        reason suspected: fitting the profile <o> *name of suspect: Kaminski
        reason suspected: killed others <o> *name of suspect: Kaminski
    *name of suspect: Kelly
    *name of suspect: Kosmiski
    *name of victim
    INVESTIGATION
    investigation_doubts
        investigation_stolen evidence <o> investigation_doubts
            11:76 DNA analysis has been attempted on some letters th.. (10:3089-10:3723)
            11:77 The available material had been handled many times.. (10:3247-10:3488)
        investigation_suspected by whom <o> investigation_doubts
```

co-occuring codes

quotations in short form linking to data in context

Figure 242: The Co-occurrence Explorer Tree View

With one click you can see which codes were used when the code"name of suspect: Kominiski" was applied: the description of the suspect, his potential motive and a list of reasons why he was suspected. If you expand the tree one more level you gain access to the full context with a click on the quotation link.

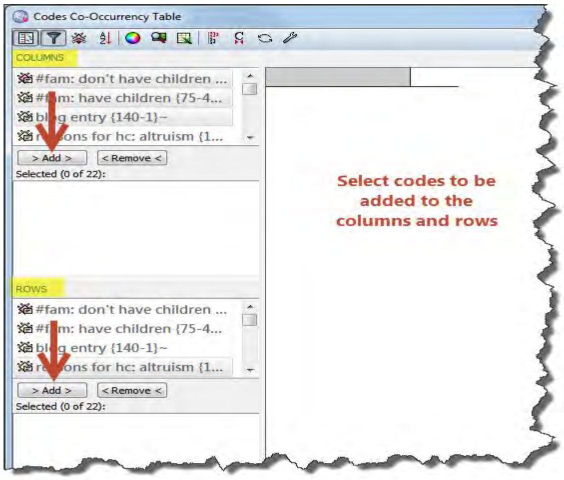
In the section "Explaining frequency count and number of quotations listed" on page 287 it is explained how to interpret the listed quotations. If you want a count of the number of quotations that co-occur, you need to run the table explorer (see below).

The Co-occurrence Table Explorer

The Co-occurrence Table Explorer in comparison to the Tree Explorer shows the frequencies of co-occurrence in form of a matrix similar to a correlation matrix that you may know from statistical software.

To produce such a table, select Analysis / **Code Cooccurrence Table.**

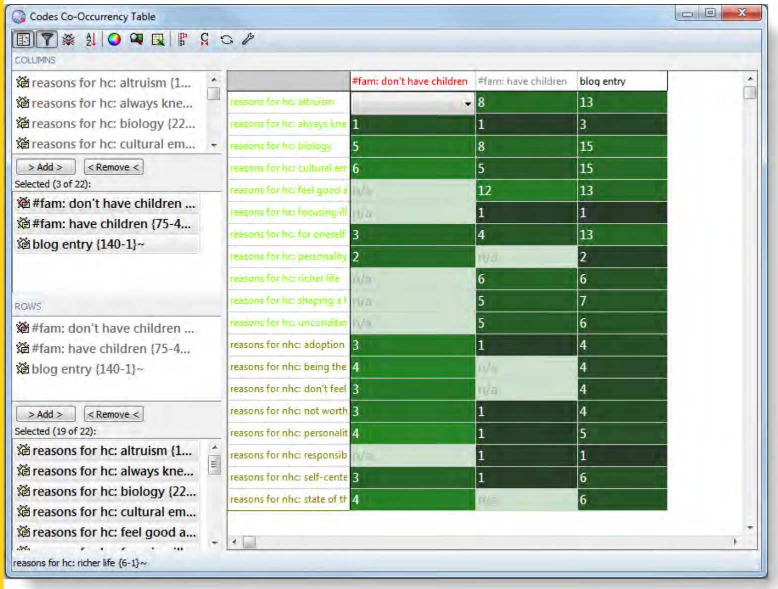
Next you need to select the codes for the column and for the rows as shown in Figure 243:


Codes Co-Occurrency Table
COLUMNS
#fam: don't have children ...
#fam: have children {75-4...
blog entry {140-1}~
reasons for hc: altruism {1...
> Add >
< Remove <
Selected (0 of 22):
ROWS
#fam: don't have children ...
#fam: have children {75-4...
blog entry {140-1}~
reasons for hc: altruism {1...
> Add >
< Remove <
Selected (0 of 22):

Select codes to be added to the columns and rows

Figure 243: Selecting row and column codes

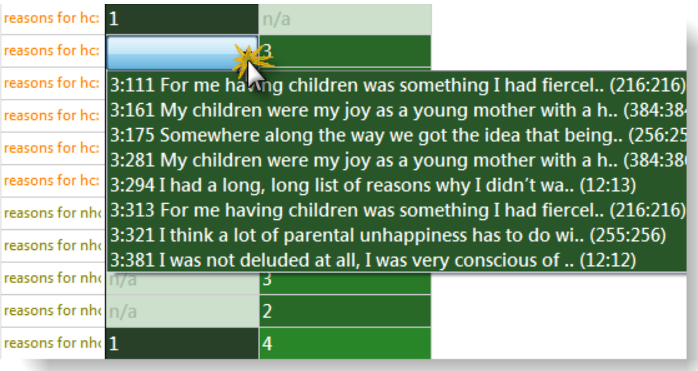
The results are displayed immediately:


Codes Co-Occurrency Table
COLUMNS
reasons for hc: altruism {1...
reasons for hc: always kne...
reasons for hc: biology {22...
reasons for hc: cultural em...
> Add >
< Remove <
Selected (3 of 22):
#fam: don't have children ...
#fam: have children {75-4...
blog entry {140-1}~
ROWS
#fam: don't have children ...
#fam: have children {75-4...
blog entry {140-1}~
> Add >
< Remove <
Selected (19 of 22):
reasons for hc: altruism {1...
reasons for hc: always kne...
reasons for hc: biology {22...
reasons for hc: cultural em...
reasons for hc: feel good a...
reasons for hc: richer life {6-1}~

	#fam: don't have children	#fam: have children	blog entry
reasons for hc: altruism		8	13
reasons for hc: always kne	1	1	3
reasons for hc: biology	5	8	15
reasons for hc: cultural em	6	5	15
reasons for hc: feel good a	n/a	12	13
reasons for hc: focusing ill	n/a	1	1
reasons for hc: for oneself	3	4	13
reasons for hc: personality	2	n/a	2
reasons for hc: richer life	n/a	6	6
reasons for hc: shaping a h	n/a	5	7
reasons for hc: unconditio	n/a	5	6
reasons for nhc: adoption	3	1	4
reasons for nhc: being the	4	n/a	4
reasons for nhc: don't feel	3	n/a	4
reasons for nhc: not worth	3	1	4
reasons for nhc: personalit	4	1	5
reasons for nhc: responsib	n/a	1	1
reasons for nhc: self-cente	3	1	6
reasons for nhc: state of th	4	n/a	6

Figure 244: Results of a cooccurence query

Double-click on a cell, then the list of coded quotations opens. In case of overlaps, the list shows two quotations for one cooccurence. This is further explained in the section “Explaining frequency count and number of quotations listed“ on page 287.



3:111 For me having children was something I had fiercel.. (216:216)
3:161 My children were my joy as a young mother with a h.. (384:38
3:175 Somewhere along the way we got the idea that being.. (256:25
3:281 My children were my joy as a young mother with a h.. (384:38
3:294 I had a long, long list of reasons why I didn't wa.. (12:13)
3:313 For me having children was something I had fiercel.. (216:216)
3:321 I think a lot of parental unhappiness has to do wi.. (255:256)
3:381 I was not deluded at all, I was very conscious of .. (12:12)

Figure 245: Viewing the co-occurring quotations

Click on an entry to see the quotation displayed in context.

The entry **n/a** indicates that the pair of codes does not co-occur anywhere in the data material (= not applicable).

Explaining Frequency Count And Number Of Quotations Listed

The co-occurrence frequency does not count single quotations it counts co-occurrence „events“. If a single quotation is coded by two codes, this would count as a single co-occurrence. The complications arise when we take overlapping quotations into account. In such a case when each of the two quotations is coded by one of the codes, this also counts as a single co-occurrence. However, in the cell drop down list you will find both quotations. In fact there are currently no means to discriminate between a single quotation's „strong“ co-occurrence and the „weak“ case for two quotations in close proximity. The drop down list will display an ordered list of all quotations for all co-occurrence events for the pair of codes.

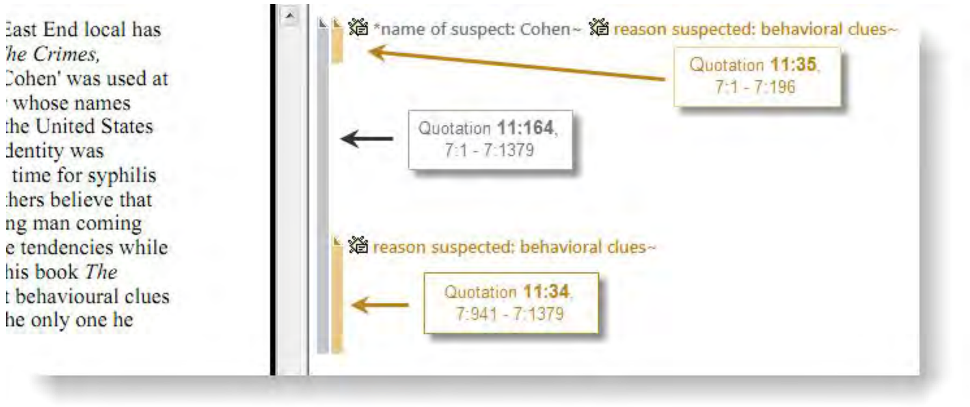
Take a look at figure 257 above. Quotation 11:34 and quotation 11:35 (the two codes for “reasons suspected” are embedded within the larger segment, quotation 11:164, coded with “name of the suspect: Cohen”. This is shown by the quotation references and if you look at the quotation in the context of the data (see figure 246 below). The references indicate that the quotes are from a PDF document and can be found on page 7:

Quotation 11:35 starts at character 1 and ends at character 196.

Quotation 11:34 starts at character 941 and ends at character 1379.

Quotation 11:164 starts at character 1 and and at character 1379.

If you want to read more on quotation references see page 163.



East End local has
he Crimes,
Cohen' was used at
whose names
the United States
dentity was
time for syphilis
hers believe that
ng man coming
e tendencies while
his book *The*
t behavioural clues
he only one he

*name of suspect: Cohen~
reason suspected: behavioral clues~
Quotation 11:35, 7:1 - 7:196
Quotation 11:164, 7:1 - 7:1379
reason suspected: behavioral clues~
Quotation 11:34, 7:941 - 7:1379

Figure 246: Explaining the relation between frequency of cocccurence and number of quotations

Thus, there are three quotations, but only two co-occurrences that are counted for the frequency count.

Toolbar

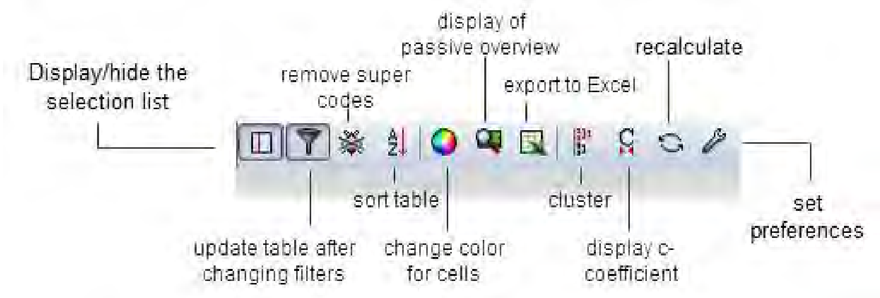


Figure 247: The Codes Co-occurrence Table toolbar

Cell Colors

Coloring helps in detecting co-occurrences. The following options are provided:

> You can choose among three colors for the table cells: blue, red and green. To select a different color, click on the color button in the tool bar.

ɔ-occurance exist are colored with the selected color

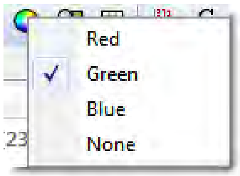


Figure 248: Set colors for table cells

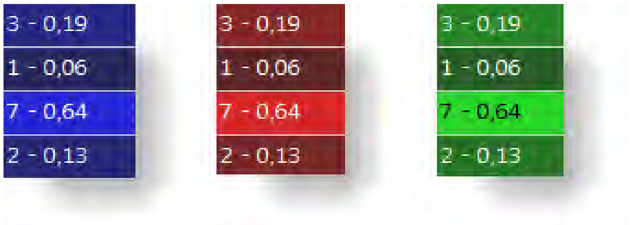


Figure 249: Three alternative colors and their shades

Passive View

If you just want to get feeling for potential patters in your data, try the passive view, which is a detached image that only shows colors.

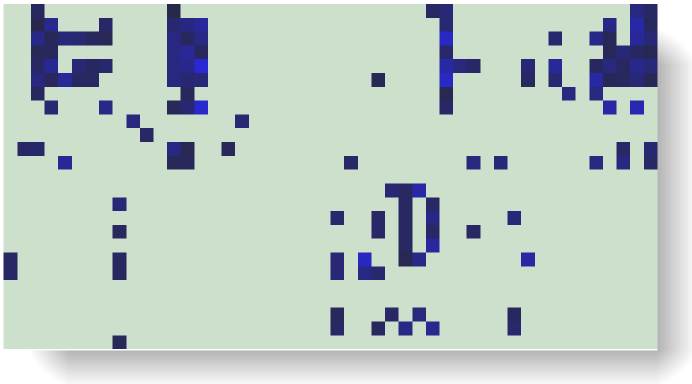


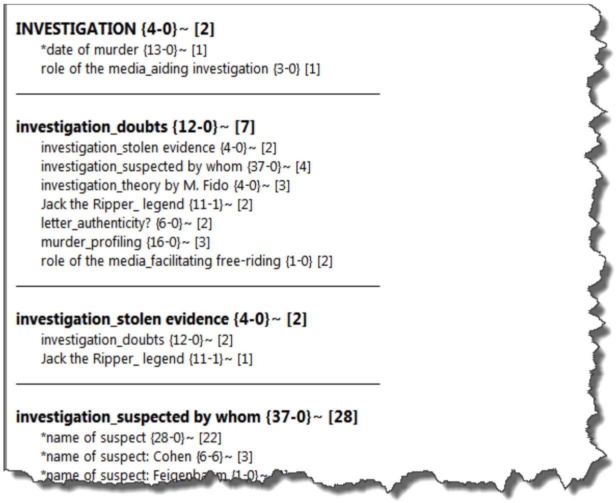
Figure 250: Example of a passive overview as an aid to detect pattern

Data Export

RTF (QUALITATIVE)

You can either export a list of the co-occuring codes, or the list including quotation IDs and names in form of an rtf file. The the full content of the quotations cannot be exported as this potentially results in very large outputs.

- To export the list of co-occuring codes, select **Codes / Output / Cooccurring codes.**
- Next you are asked whether to include the quotation lists (= ID plus name of the quotation.
- Next, select the output destination (Editor, File, Printer).



```
INVESTIGATION {4-0}~ [2]
    *date of murder {13-0}~ [1]
    role of the media_aiding investigation {3-0} [1]
________________________________________________

investigation_doubts {12-0}~ [7]
    investigation_stolen evidence {4-0}~ [2]
    investigation_suspected by whom {37-0}~ [4]
    investigation_theory by M. Fido {4-0}~ [3]
    Jack the Ripper_ legend {11-1}~ [2]
    letter_authenticity? {6-0}~ [2]
    murder_profiling {16-0}~ [3]
    role of the media_facilitating free-riding {1-0} [2]
________________________________________________

investigation_stolen evidence {4-0}~ [2]
    investigation_doubts {12-0}~ [2]
    Jack the Ripper_ legend {11-1}~ [1]
________________________________________________

investigation_suspected by whom {37-0}~ [28]
    *name of suspect {28-0}~ [22]
    *name of suspect: Cohen {6-6}~ [3]
    *name of suspect: Feigenbaum {1-0}~
```

Figure 251: List of cooccuring codes in rtf format

EXCEL (QUANTITATIVE)

You can either export the frequency counts of the c-coefficient. If the c-coefficient is activated, then the coefficient is included in the output; if not, the Excel table shows the frequency of coccurrence.

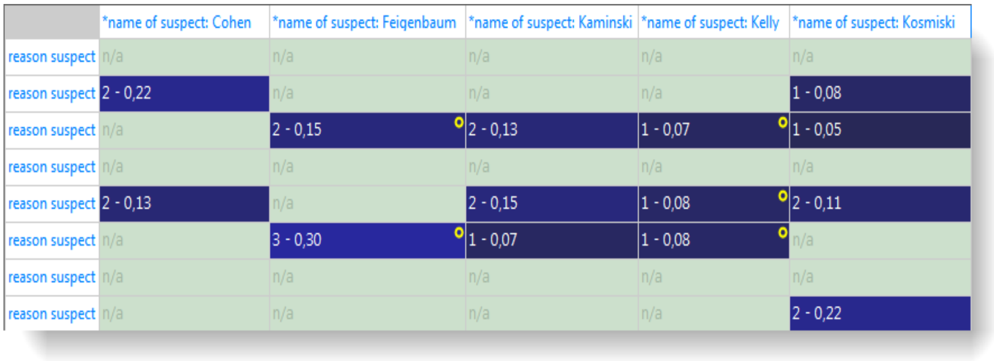
- To create an Excel table of either the frequency counts of the c-coefficients, click in the Excel button in the tool bar.
- As output select destination **File & Run**.
- Save the file and wait for Excel (OpenOffice Calc) to be opened. Confirm the conversion of the data.

Clustering Quotations

If you want to count embedded quotations as only one count (compare “Embedding Operators“ on page 258), select the *Cluster Quotations button in the tool bar (s*ee left).

C-Coefficient

In addition to the frequency count a so called c-coefficient can be displayed. You can display or hide it (see button to the left). The c-coefficient indicates the strength of the relation between two codes similar to a correlation coefficient.



	*name of suspect: Cohen	*name of suspect: Feigenbaum	*name of suspect: Kaminski	*name of suspect: Kelly	*name of suspect: Kosmiski
reason suspect	n/a	n/a	n/a	n/a	n/a
reason suspect	2 - 0,22	n/a	n/a	n/a	1 - 0,08
reason suspect	n/a	2 - 0,15 ○	2 - 0,13	1 - 0,07 ○	1 - 0,05
reason suspect	n/a	n/a	n/a	n/a	n/a
reason suspect	2 - 0,13	n/a	2 - 0,15	1 - 0,08 ○	2 - 0,11
reason suspect	n/a	3 - 0,30 ○	1 - 0,07	1 - 0,08 ○	n/a
reason suspect	n/a	n/a	n/a	n/a	n/a
reason suspect	n/a	n/a	n/a	n/a	2 - 0,22

Figure 252: Co-occurrence Table displaying c-coefficients

The calculation of the c-coefficient is based on approaches borrowed from quantitative content analysis (see Garcia, 2006). Thus, interpreting such a coefficient is only meaningful with a sizable data set and not for an interview study with 10 respondents. Given the possibility to work with survey data to analyze open-ended questions, it however is a valuable addition to the other more qualitative oriented analysis tools that ATLAS.ti provides.

The c-coefficient should vary between 0: codes do not co-occur, and 1: these two codes co-occur wherever they are used. It is calculated as follows:

c := n12/(n1 + n2) – n12

n12 = co-occurrence frequency of two codes c1 and c2, whereby n1 and n2 are their occurrence frequency

What you may experience is the following:

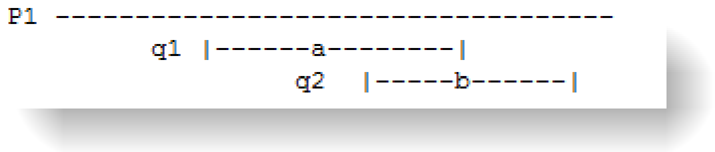
- Out of range. The C-index exceeds the 0 - 1 range it is supposed to stay with.
- Colored circles. Cells can have additional visual cues, e. g., a red, yellow or orange circle.

OUT OF RANGE

The c-index (structurally resembling the Tanimoto and Jaquard Coefficient, which are similarity measures) assumes separate non-overlapping text entities. Only then can we expect a correct range of values.

However, ATLAS.ti's quotations may overlap to any degree. Overlaps would only then bear no problem if there wasn't any „coding redundancy“ (the ones you can eliminate using the Coding Analyzer, see page 369 for further detail). Let's look at a few scenarios.

Case 1: Two differently coded quotations overlap, we assume no more quotations available. Let P1 be a textual document, q1 and q2 be quotations and a,b be codes. q1 is coded with a, q2 is coded with b.



```
P1 ------------------------------------
         q1 |------a---------|
                  q2  |-----b------|
```

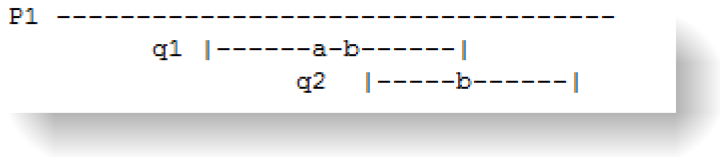
Figure 253: Out of range example 1

Using the formula: c := n_ab/(n_a + n_b) – n_ab, we get:

n_ab = 1 one co-occurrence of a and b
n_a = 1, n_b = 1 a and b each code exactly one quotation.
c = 1/(1 + 1) – 1 = 1

Such a scenario results in the maximum co-occurrence of 1 !

Case 2: q1 is coded with both codes a and b, the overlapping quotation q2 is coded with b.



```
P1 ------------------------------------
         q1 |------a-b------|
                  q2  |-----b------|
```

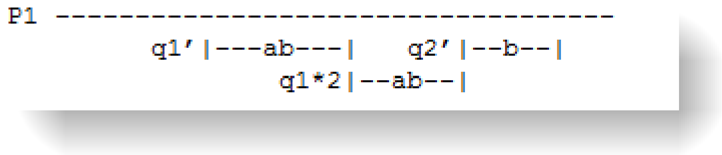
Figure 254: Out of range example 2

n_ab = 2. q1 alone counts for a co-occurrence event and the overlapping q1*q2 for another.

n_a = 1, n_b = 2
c = 2/(1 + 2) – 2 = 2!!

This results in a value of twice the allowed maximum. Thus, the C index is not appropriate to correctly represent co-occurrence in redundantly overlapping texts. If the c-coefficient exceeds 1, you need to do some cleaning up and eliminate the redundant codes. ATLAS.ti currently does not correct such redundancies automatically.

Correcting the redundant overlaps, could for example look like this:



```
P1 ------------------------------------
         q1'|---ab---|   q2'|--b--|
                q1*2|--ab--|
```

Figure 255: Out of range example 2 normalized

We get three quotations. q1' coded with a and b, q1*2 coded with a and b, q2' coded with b:

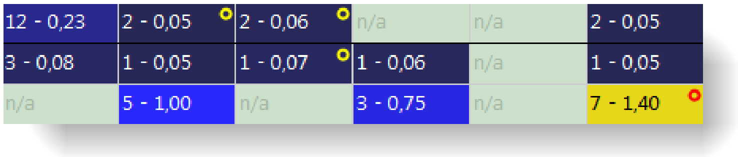
n_ab = 2, n_a = 2, n_b = 3
c = 2/(2 + 3) – 2 = 2/3 = 0.67

The result is within the allowed range and it correctly takes into account that of the three possible co-occurrence events only two apply.

To detect and correct redundant coding, select **Tools / Codings Analyzer**. See page 369.

COLOR INDICATORS

In order not to present a misleading image, all cells displaying an out-of-range number (> 1) are colored in yellow.



12 - 0,23	2 - 0,05 ○	2 - 0,06 ○	n/a	n/a	2 - 0,05
3 - 0,08	1 - 0,05	1 - 0,07 ○	1 - 0,06	n/a	1 - 0,05
n/a	5 - 1,00	n/a	3 - 0,75	n/a	7 - 1,40 ○

Figure 256: Examples of yellow and red circle markers

Circles with different colors are painted into a cell's upper right corner when certain conditions apply.

Red circle: When the c-index exceeds 1 (see “Out of range“ on page 291). In addition to the red circle, the entire cell is highlighted in yellow.

Yellow circle: An inherent issue with the C-index and similar measures is that it is distorted by code frequencies that differ too much. In such cases the coefficient tends to be much smaller than the potential significance of the cooccurrence. For instance, if you had coded 100 quotations with code "depression" and 10 with "mother" and you had 5 co-occurrences:

n_dep = 100, n_mother = 10, n_dep-mother = 5
c = 5/(100 + 10) - 5 = 5/105 = 0.048

A c index of only 0.048 may slip your eye easily, although code "mother" appears in 50% of all its applications with code "depression". Looking from code "depression" only 5% co-occurr with code "mother".

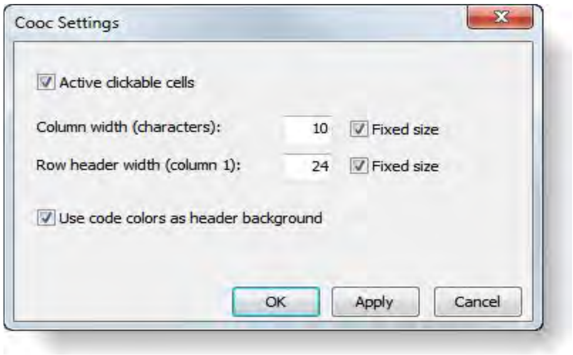
If the ratio between the codes frequencies exceeds a certain threshold (currently 5 but will be user definable in the future) the yellow light goes on in the cell. So whenever a cell shows the yellow marker it should invite you to look into the co-occurrences of this cell despite a low c-index.

When the mouse hovers over a cell with a yellow mark, a pop-up displays the ratio of the two codes.

Orange Circle: The orange circle is simply a mixture of the red and yellow conditions.

Preference Settings

The preference settings allow you to set the column and row header width and to set the code colors as header background.



☑ Active clickable cells

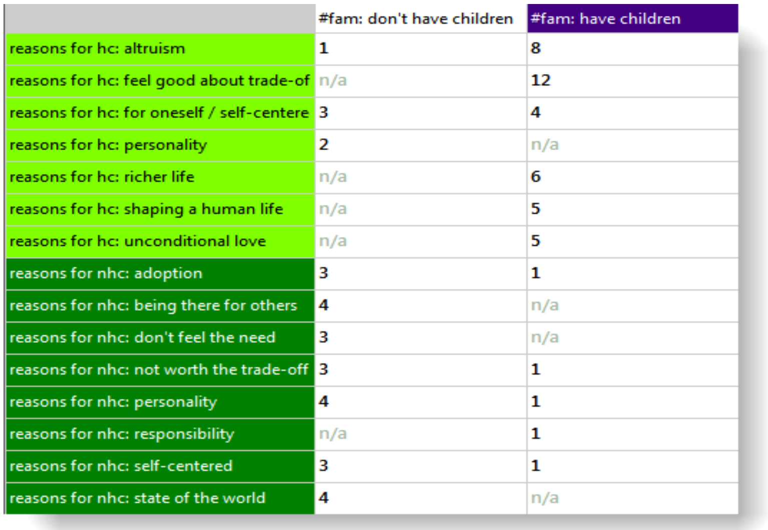
Column width (characters): 10 ☑ Fixed size

Row header width (column 1): 24 ☑ Fixed size

☑ Use code colors as header background

Figure 257: Code Cooccurence Table settings

The following table shows how you can alternate the look of the table by changing the settings:



	#fam: don't have children	#fam: have children
reasons for hc: altruism	1	8
reasons for hc: feel good about trade-of	n/a	12
reasons for hc: for oneself / self-centere	3	4
reasons for hc: personality	2	n/a
reasons for hc: richer life	n/a	6
reasons for hc: shaping a human life	n/a	5
reasons for hc: unconditional love	n/a	5
reasons for nhc: adoption	3	1
reasons for nhc: being there for others	4	n/a
reasons for nhc: don't feel the need	3	n/a
reasons for nhc: not worth the trade-off	3	1
reasons for nhc: personality	4	1
reasons for nhc: responsibility	n/a	1
reasons for nhc: self-centered	3	1
reasons for nhc: state of the world	4	n/a

Figure 258: Settings: No color for table cell; use code color as header background

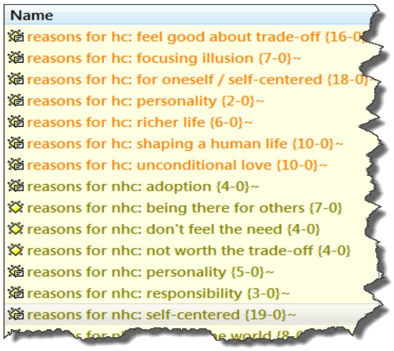
An Example Query

Let's take a look at the code-code matrix for this Happiness sample project. A special code family as been prepared for this exercise that helps us to gain an overview of the responses regarding the question why or why not having children voiced by parents and non-parents. The family contains the two attribute codes (#fam: has children and #fam: no children) plus all sub codes of the two categories "reasons for having children" and "reasons for not having children".

First, set a family as filter to reduce the list of codes to select from:

> From the main menu select: **CODES / FILTER / FAMILIES**: "for Quick Tour: Coocurrence Example". Or set this family as global filter in the side panel of the Code Manager (Ctrl+Shift+Click).

After setting a filter, all effected fields are shown in a pale yellow color.



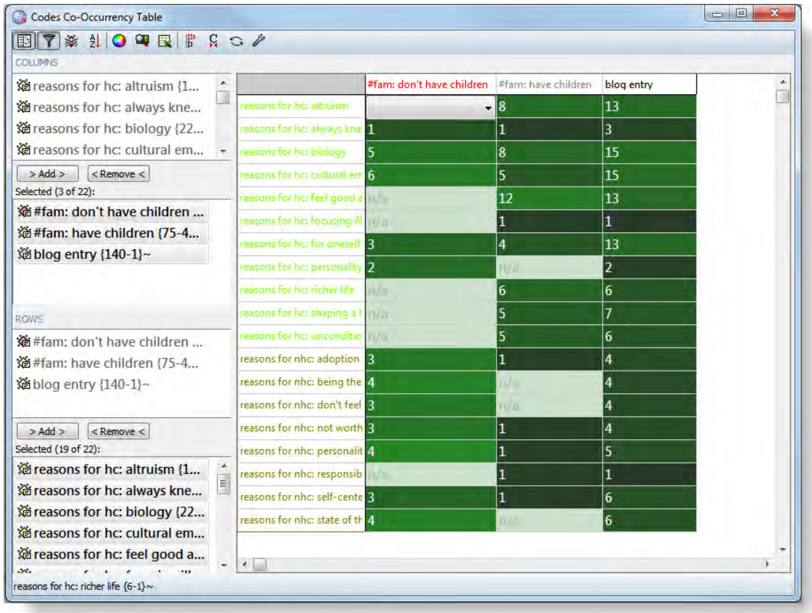
Name
- reasons for hc: feel good about trade-off {16-0
- reasons for hc: focusing illusion {7-0}~
- reasons for hc: for oneself / self-centered {18-0}
- reasons for hc: personality {2-0}~
- reasons for hc: richer life {6-0}~
- reasons for hc: shaping a human life {10-0}~
- reasons for hc: unconditional love {10-0}~
- reasons for nhc: adoption {4-0}~
- reasons for nhc: being there for others {7-0}
- reasons for nhc: don't feel the need {4-0}
- reasons for nhc: not worth the trade-off {4-0}
- reasons for nhc: personality {5-0}~
- reasons for nhc: responsibility {3-0}~
- reasons for nhc: self-centered {19-0}~

Figure 259: Setting a code family as filter

From the main menu select **ANALYSIS / CODE COOCCURRENCE TABLE.**

Next you need to select which codes should be displayed in the rows and which ones in the columns (see Figure 243):

Select the #fam: don't have children / have children and the blog entry codes as columns and all other codes as rows.



Codes Co-Occurrency Table

COLUMNS
- reasons for hc: altruism {1...
- reasons for hc: always kne...
- reasons for hc: biology {22...
- reasons for hc: cultural em...

Selected (3 of 22):
- #fam: don't have children ...
- #fam: have children {75-4...
- blog entry {140-1}~

ROWS
- #fam: don't have children ...
- #fam: have children {75-4...
- blog entry {140-1}~

Selected (19 of 22):
- reasons for hc: altruism {1...
- reasons for hc: always kne...
- reasons for hc: biology {22...
- reasons for hc: cultural em...
- reasons for hc: feel good a...

reasons for hc: richer life {6-1}~

	#fam: don't have children	#fam: have children	blog entry
reasons for hc: altruism		8	13
reasons for hc: always kne	1	1	3
reasons for hc: biology	5	8	15
reasons for hc: cultural em	6	5	15
reasons for hc: feel good a	n/a	12	13
reasons for hc: focusing ill	n/a	1	1
reasons for hc: for oneself	3	4	13
reasons for hc: personality	2	n/a	2
reasons for hc: richer life	n/a	6	6
reasons for hc: shaping a h	n/a	5	7
reasons for hc: unconditio	n/a	5	6
reasons for nhc: adoption	3	1	4
reasons for nhc: being the	4	n/a	4
reasons for nhc: don't feel	3	n/a	4
reasons for nhc: not worth	3	1	4
reasons for nhc: personalit	4	1	5
reasons for nhc: responsib	n/a	1	1
reasons for nhc: self-cente	3	1	6
reasons for nhc: state of th	4	n/a	6

Figure 260: Results of the example cooccurence query

The results of this analysis have been visualized in the two network view “Reasons for having children” and “Reasons for not having children”. See “Network Views“ on page 301 for further detail.

Application

The two Co-occurrence Tools are very useful for many kinds of analysis. But not all options make sense for all type of data. If you have a smaller data set like a typical interview study with 10 to 20 respondents, then taking a look at the frequency count for exploratory purposes is likely to provide some new ideas and you may gain new insights. The c-coefficient is useful when working with larger amounts of cases and structured data like open-ended questions from surveys. If you use the c-index, pay attention to the additional colored hints. As your data base is qualitative, the c-coefficient is not the same as for instance a Pearson correlation coefficient and therefore also no p-values are provided.

In any case, co-occurrence measures need to be clearly understood, not only for the mechanical but also for semantic issues involved in their meaningful interpretation (e. g., mixed application of codes with different level like broader and sub terms). Furthermore, you need to be aware of the artifacts enforced by a table approach like being reduced to a pairwise comparison. Higher order co-occurrences which would take more than two codes into account need more elaborate methods.

References: Garcia (2004) http://www.miislita.com/semantics/c-index-1.html

Codes-Primary Documents Cross-Tabulation

Even though a bit hidden, a further analysis tool with an emphasis on quantitative output is the **CODES-PRIMARY-DOCUMENTS-TABLE.** You find this option under the **ANALYSIS** menu and under **CODES / OUTPUT.**

The table is available as internal report within ATLAS.ti in text format, or can be exported to Excel. The internal report displays all PDs as columns and the codes as rows.

The table contains either a frequency count for each code or code family per document or document family, or a word count of the coded segments per code and primary document.

A useful application is a comparison across different groups of documents for a particular category of codes. Thus, you are likely to create such a table if you have a certain research question in your mind. This will guide you to create the code and PD families you need to construct your query.