Concordances and semi-automatic coding in qualitative analysis: possibilities and barriers

Graham R Gibbs,

Centre for Applied Childhood Studies

School of Human and Health Sciences

University of Huddersfield

Coding in Qualitative Research

- Identify chunks of text
- Give these a label
  - Inductively - create new concept grounded in the data
  - Deductively - codes derived from theory before start of analysis.
- Label stands for the concept or idea that stands for the collection of similarly coded chunks of text
- Key = reading the text and identifying its meaning

Problem

- This involves human judgement
- And hence
- Is very time consuming
- OK with small data sets, but problematic with large.

Ways to speed up coding

- Suggestions for coding
- Qualrus s/w uses AI
- After you have open coded some text, it examines the words coded and suggest other codes you could use on that text (based on similar texts)
- The amazon technique (people who bought this book…)

Text search tool

- Works the other way around.
- Use the search tool in CAQDAS to find similar passages that could be coded the same.

Problems

- Key terms not always used by speakers
- Words used for other purposes in passages that will be coded in different ways.
- Usually needs human checking

I.e. can be an assistance with exploration.

Text search tools in CAQDAS

- Like search tools in word processors

BUT

- Finds all occurrences of the text
- Can use wildcards (and in some cases GREP)
- Can auto code - found terms (and some surrounding text) is coded.

How successful can this be as a way of identifying the content of text (and coding)?

Success of coding prediction

Depends on:

1. Term used uniquely (not used elsewhere)
2. Relevant text uses the term
3. We know what terms to use

What terms to look for

The answer to Q 3 =

- Look for terms in text already coded that way.

To test this used Climbié data corpus.

The Victoria Climbié Corpus (VCC)

- The Victoria Climbié Inquiry (Laming Report)
- Major review of the child protection system in England and Wales ➔ Green Paper ‘Every Child Matters’
- Inquiry investigated circumstances surrounding the death of Victoria Climbié
- Took evidence on wider aspects of the child protection system through a series of seminars
- Reported to both the Home Office and the Department of Health

Testimony already on the Web

- http://www.victoria-climbie-inquiry.org.uk/
- *64 days of the verbatim cross-examination of witnesses (up to 200 pages per day). About 2 million words.*
- *Written submissions (image pdf – not part of current study)*
- *Evidence about state of child protection services in late 1990s*
- *detailed testimony about:*
  - day-to-day practice
  - decision-making
  - inter-agency working
  - the context of service delivery
  - policy making across all agencies: social services, police, health, voluntary groups.

Three stages of our research

- Identify themes & topics in cross-examination required by or of use to a range of professional & educational users. Done using Delphi technique
- Catalogue and code data & establish system of data management & retrieval. (Using Atlas.ti) Doing it now
- Establish an online data corpus available for future research outside the University. (Using XML output from Atlas.ti.)

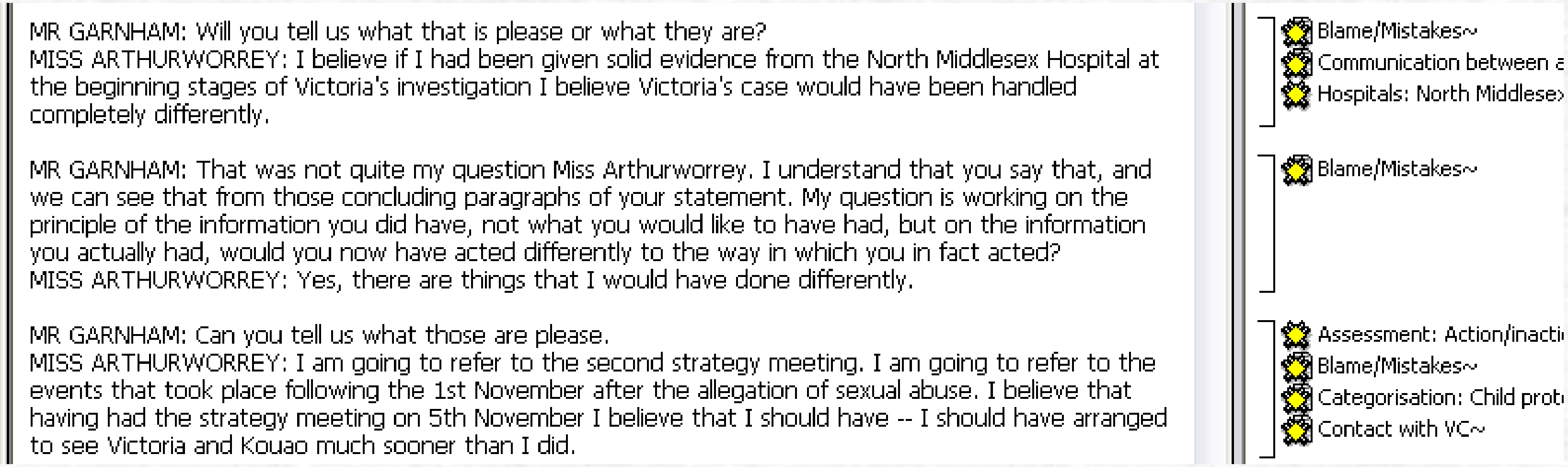
Themes for coding

- From preceding and other info. from Delphi,
- 108 codes used to code the data thematically e.g.

Assessment: Action/inaction	Categorisation: Sick Child case
Assessment: Decision plan of action	Communication between agencies
Assessment: Exchanging information	Communication within agencies
Assessment: General	Contact with Victoria Climbié
Blame/Mistakes	Family Status
Categorisation: Child in need case	Files/Records
Categorisation: Emotional abuse case	Mangmt: Responsibilities and direction
Categorisation: Housing/homeless/subsistence case	Management: Roles

Used Atlas.ti

- **Question from lawyer and answer from witness combined into single quotation**
  - **47,352 quotations**



MR GARNHAM: Will you tell us what that is please or what they are?
MISS ARTHURWORREY: I believe if I had been given solid evidence from the North Middlesex Hospital at the beginning stages of Victoria's investigation I believe Victoria's case would have been handled completely differently.

MR GARNHAM: That was not quite my question Miss Arthurworrey. I understand that you say that, and we can see that from those concluding paragraphs of your statement. My question is working on the principle of the information you did have, not what you would like to have had, but on the information you actually had, would you now have acted differently to the way in which you in fact acted?
MISS ARTHURWORREY: Yes, there are things that I would have done differently.

MR GARNHAM: Can you tell us what those are please.
MISS ARTHURWORREY: I am going to refer to the second strategy meeting. I am going to refer to the events that took place following the 1st November after the allegation of sexual abuse. I believe that having had the strategy meeting on 5th November I believe that I should have -- I should have arranged to see Victoria and Kouao much sooner than I did.

Coding work

- Being undertaken by a Research Assistant
- Using given codes and definitions
- Quality checking by other member of research group.
- So we have
  1. Some testimony coded early on
  2. Some testimony coded later
  3. Some testimony still to be coded.
- Can use 1 and 2 to assess usefulness of search for future coding.

Procedure

- Chose some key codes from the VCC
- Use early version of project
- Retrieve coded data for that code
- Produce wordlist (using Concordance s/w)
- Eliminate ‘open class words’ (and, but, the, mine etc.)
- Extract words that seemed to capture meaning of the code and were not obviously going to be common in text coded in other ways.

Example word list

MR	799
GARNHAM	628
HAVE	454
NOT	444
ARTHURWORREY	332
MISS	329
DR	319
YES	289
HAD	247
ROSSITER	242
WE	242
WOULD	232
DO	216
THERE	198
DID	185
PAGE	181

Example terms used in search for Files/Records

CP1|CP2|CP3|CP4|CP5

address*|amendment*|annotate*|application|arrow*|book*|box|bundle*|case*|column*|copy|copie*|data|database|date*|details|diagram*|document*|draft*|entry|entries|evidence|fact*|fax*|file*|form*|history|information|initial*|input*|investigation|letter*|log*|margin|meeting*|memo*|minutes|note*|page*|paragraph*|point*|record*|reference*|referral|relating|report*|response*|section*|sheet*|stamp*|statement*|summar*|tick*|time|volume

handwrit*|handwrote

write|writing|written|wrote

Procedure, 2

- Refine list of terms
- Allow for variations (write*|wrote|writing)
- Include some synonyms (used Thesaurus and WordNet)
- Use search tool in Atlas.ti and autocoding feature to code text to new codes (called ‘Auto Race’ if original was ‘Race’)
- Compare text coded this way with text coded in second stage human coding (An Atlas.ti code search retrieval).

Files/Records - Number of quotations

	Early coding	Late coding
F/Recs	100% (766)	100% (439)
Agreement F/Recs & AutoF/Recs	82%	81%
Disagreement (F/Recs & not AutoF/Recs)	18%	19%
Disagreement (Not F/Recs & AutoF/Recs)	237%	683%
Auto F/Recs	319%	764%
Total All quotations	4466	5514

Files/Records (Only high frequency terms) - Number of quotations

	Early coding	Late coding
F/Recs	100% (766)	100% (439)
Agreement F/Recs & AutoF/Recs (HF)	77%	72%
Disagreement (F/Recs & not AutoF/Recs (HF))	23%	28%
Disagreement (Not F/Recs & AutoF/Recs (HF))	214%	601%
Auto F/Recs (HF)	291%	674%
Total All quotations	4466	5514

Resources Code & Auto version - number of quotations

	Early coding	Late coding
Resources	100% (63)	100% (74)
Agreement (Res & AutoRes	83%	66%
Disagreement (Res & not AutoRes)	17%	34%
Disagreement (Not Res & AutoRes)	968%	1227%
Auto Res	1051%	1293%
Total All quotations	4466	5514

Race - Number of quotations

	Early coding	Late coding
Race	100% (49)	100% (41)
Agreement Race & AutoRace	94%	80%
Disagreement (Race & not AutoRace)	6%	20%
Disagreement (Not Race & AutoRace)	53%	83%
Auto Res	147%	163%
Total All quotations	4466	5514

Good and Bad

- Generally the technique did not work well
- Codes that shared terms and ideas with others did not work well
- E.g. Files/Records overlapped with:
  - Communications between agencies
  - Assessment - exchanging information
  - Assessment - general
  - Workplace practices
- Need to use with codes that share less with other codes

Outcomes

- Can be good at capturing what is coded later
  - Most of existing and new coding is captured
- But tends to code many other passages that are not coded later
  - many type 2 errors
- Less frequently fails to code text that is coded later.
  - A few type 1 errors
- Works better on some codes - more distinctive - little shared vocabulary - distinctive terms used.

Conclusions

- Procedure still needs work
- To reduce type 2 errors (text is coded but should not be)
- May have a useful role in supporting exploratory work (to help find new passages to code)
- Help as quality check after coding is finished. Type 2 errors can be used to check if text should be coded in other ways

Example concordance programs

- Conc v. 1.76 (for Mac) free
- Concordance (for PC) £55
- Concorder (for Mac) free
- Intext (for PC) free (manual on CD 20€
- MonoConc Pro (for PC) $85
- TextSTAT 2.4 (for PC, Linux, Mac OSX) free
- WordSmith (for PC) £50

And see *onlineqda.hud.ac.uk*