DATA ANALYSIS, Lugano, 21st February, 2011
Notes by: Bojana Ćulum

Aims:

- methods and techniques for early stage data analysis: coding and qualitative content analysis
- role of interpretations in early stages of data analysis (many legitimate interpretations, not one valid:)
- develop ideas about securing comparability of qualitative data analysis in cross-country collaboration

- we are talking about the early stages of data preparing

- critical discourse analysis - proučiti malo

Research questions

- grounded theory - your research questions is whatever you find the answer on your data; you have data and answer from which you derive research questions

1. why are they important?
- focus
- need to be reminded on what you are looking for
- having it in mind while you conducting interviews in order to know why are you doing this, to direct it, as well as to direct to content analysis
- 1 question mark - very important to wrap up your research around one basic and simple theoretical question!
- having a good theoretical question helps you discover new insights in your data
- have to be careful and not let theory get in our way in analyzing data
- you get categories and codes from your question/from previous research and theory (at least part of it)

2. what are our research questions?

EUROAC: How do academics view major developments around and within HE as potentially relevant for them, and how do they interpret and eventually shape their professional roles under these given circumstances?

This project intends not only to address the ways in which the AP perceives and handles these major societal challenges, but also, how it responds to changes in the *organisational fabric of the HE system* which are related to these societal changes, and are more clearly visible in the daily life of HE.

RHESI: How do authority relations shape conditions for the emergence of intellectual innovations?
TRUE: How do new governance instruments affect the universities as organizations?

3. What do we want to explain?

Explanatory strategies in the social sciences

1. **Identifying causal conditions**
- social, institutional, financial
- conditions that are necessary or sufficient for a specific effect to occur
- most systematic approach: **Qualitative comparative analysis** (Ragin) - Ragin is one of the first who addressed the systematic issue for analyzing data
- identifying conditions leading to effects (mechanism are left in the black box)
- conditions - ??mechanisms?? - effects (A produces B, but how?)

2. Identifying causal mechanisms
- frequently occurring sequences of causally linked events that are triggered by certain conditions and produce a specific effect
- '**process tracing**'
- identifying mechanisms (from the black box) leading to certain effects
- conditions - causal mechanisms - effects

QUALITATIVE COMPARATIVE ANALYSIS (QCA)

Basic idea - comparing cases according to:
- conditions present or absent
- effects
- truth table (Boolean algebra)

Problems:
1. QCA works only if all combinations of occurrences of conditions are represented by the cases
- readings - Lieberman (Small N's and Big Conclusions) - drunk driving does not cause accidents

accident	drunk driving	car entering from right-hand direction	driver speeding	runs a red light
yes	yes	yes	yes	yes
no	yes	no	no	yes

- it does not know chain of events
- could running a red light be consequence of drunk driving?

2. QCA works only if all causal conditions that vary between cases are explicitly included
3. QCA requires to reduce empirical data to the dichotomy of "condition present" - "condition not present" (fuzzy set QCA avoids that to a certain degree, 30%...50%...) - strength of presence, but eventually you have to end up with basic boolean 0/1 presence or not

PROCESS TRACING (finding causal mechanisms)

- no literature
- George and Bennet, 2005: 210-212:
    - write a detailed historical narrative
    - transform it into a theoretical explanation

Q: How would we approach the 'drunk driving' cases with process tracing?

- police reports from reconstruction of the events/accidents
- reports from the drivers
- sequences of events and linking them - which event let to another? what happened? what happened next because of that?
- large number of cases
- detected mechanism of chain reactions/events
- try to reconstruct the chain of events and their interlinks

Problems:
- data on the whole uninterrupted causal path required
- equifinality - several causal mechanism may produce the same outcome, some of them even from the same initial conditions - they find them, but they can confuse your system of tracing process

Working backwards

1. Linking raw data to the research question
(identifying, locating, structuring)

2. Consolidating raw data

3. Empirical data in easily manipulated form

4. Identifying patterns (of conditions and sequences of events)

5. Integrating patterns (of conditions and sequences of events)

Results:

Conditions triggering causal mechanism

Causal mechanisms ← Conditions affecting the operation of causal mechanisms

Effects

Integrating patterns (of conditions and sequences of events)

1. Linking raw data to the research questions

- text contains raw data and irrelevant information (relevant vs. interesting - have to be disciplined, guided by the research question - that why the clear research question is important, otherwise you will not know how to recognize relevant data/empirical research questions could help you; very interesting might be irrelevant from the point of view of specific research questions; not that new

ones should be ignored, but rather noted for some other research opportunities, while this one has to be addressed)
- solution: read the text/assess the content/if you find relevant info (=raw data) in a text segment you can:

a) stick a note to it that describes what's in there (code it)
b) extract it - take the information away and store it somewhere else

- basically, you can not avoid being subconsciously selective, because we always are...
- the decision about the relevance of information is about to exclude irrelevant info in the first step (data vs. noise)
- the first thing in mind should be the theoretical one (more priority; empirical questions will help, as well as categories, variable models)
- reading: first independently and then compare it to other team members (looking for relevant info first)
- research question is basis for deciding what is data and what is noise

TYPOLOGIES

- reduces irrelevant/redundant data, reduces complexity to something we can process
- usually done from 2/3 variables / creating dimensions
- 4x4 popular (but researchers rarely explain the process of extracting variables and creating dimensions)

- What is a type?
Abstract construct that represents a sub-class of empirical objects by expressing the combination of properties distinguishing the sub-class from the others.

A type...
- requires a more general concept (type of what - general smoker/non smoker/occasional smoker)
- requires at least one dimension in which properties vary
- never comes alone
- 2 dimensions x 2values - 4 types - you can easily end up with large number of types

Consolidating raw data

- only possible and necessary if you extract data
- compress information, cleans it up to an extent that information is kept but further compressed
- correct errors
- you can simply notice that several respondents gave the same/similar answers/easier to look for patterns

Databases for finding patterns

- if we want more than just a description of a case (one interview/transcript can be a case, but doesn't have to be; thematic parts of interviews can be cases, actors/institutions/relations can be cases, all of them can be a case - depending on the research question, one again we can see its relevance)
- you have to play with your data, it requires the data base (paper or electronic)
- play around with data: sort them, combine them - until we see something interesting (it is not satisfactory to go and read transcripts until something comes up...their critic to grounded theory...something interesting could come up, but without any empirical or theoretical relevance

- ideas for rearranging data come from the research question and the data
- you go through text, find something interesting, creating empirical typology in line with this interesting/relevant data - everything that follows should be connected with the principle of your typology
- data must exist in a form that support rearranging: coded text + tables (two basic forms)

Typologies - is RHESI a type of project?

To what types of projects does RHESI belong to? just for the illustration...
- internationally collaborative
- grant funded
- three years (e.g. code - duration of project)
- social science
- multi lingual

What typologies are we currently exploiting in the project?
- societal challenges
- professional roles

- EUROAC: we could do typology on governance structure/development stages of academic career/ types of professionalization/new professional roles - and then link data to the various types discovered/defined within these typologies
-

for RHESI...innovations, fields, switching costs

What other typologies could be interesting? type of university / career stage / fields, disciplines

What could mechanisms we search for look like?

- important to have in mind that there are more then 20 definitions of mechanisms (the discussion on sociology of science is relatively new, 15-20 years all) - still no consensus on how should a mechanism look like!
- hypothesis - what is possibly going on? (a causes b is boring, not interesting for this kind of research and analysis)

RHESI - interest in innovation/hight cost/funding opportunity = negotiates investment (success or failures)

EUROAC - ???

Coding as an initial step of qualitative data analysis

- we are linking raw data to the research question by identifying and making categories
- identification and processing of relevant information by indexing raw data
- making codes is not the theoretical question in its nature, but question of better organizing your codes and data (from a theoretical perspective it is not necessary to divide codes into sub-codes, it is a matter of data organization, and sub-codes/codes/family codes could be merged again later if it become more logical)

Coding:

a) as part of a Grounded Theory
- Glaser&Strauss, 1967 - "The discovery of grounded theory" (Glaser 1978, 1992) + Strauss/Corbin (1990) - different approaches to theory
- it is totally forbidden to draw codes from theoretical considerations (Glaser/Strauss 1967)
- theory roots: analyzing interaction between nurses and dying patients
- having a theoretical research questions is already too much:)

b) as an independent analytical process
- Miles and Huberman 1994 (great description - must read!!!) + Patton, 1990

Procedure:
1. read a text sequence (line, paragraph)
2. interpret it
3. decide which code to assign (indexing a word, line, paragraph) - short descriptors of what it's in the paragraph

What is code?
- categories (keywords, phrases, mnemonics, numbers) which are assigned to a text segment

Where do codes come from?
a) from the empirical data
b) from theoretical considerations

Why is this an important question?
a) principle of openness - the empirical research process must be open for unexpected information
b) principle of theory-guidedness - we must proceed from the existing theoretical knowledge; only this way we can contribute it

2 MAIN APPROACHES
1. step - open coding, then
2. theoretical coding = codes arrive from epistemology, general theory are allowed (Glaser 1978)
Examples: limit, extent, goal, social norms

1. open coding, then
2. axial coding = relating of categories by using a general model of action (Strauss/Corbin 1990)

Miles and Huberman 1994:
- you should always start with theory!
- use all prior knowledge about your empirical object/case, about the subject - don't let the prior knowledge blinds you for the new knowledge
- theories are storages of knowledge
- exploiting what other people had done on the subject

Jochen and Grit
- start with theory and prior knowledge on the subject!
- use middle range theory (theory about universities, organization theories, middle range organizations theory) - something you can build a theoretical framework from
- the less theory exists, the more open the process itself becoming

What is a good code?
- related to a research question
- sufficiently precise
- clearly defined
- distinctive
- it should enable a large part of the data material to be subsumed under it
(Kelle/Kluge 1999)
Example - reading material (manual)

Coding list:
- research question
- variables/theoretical model
- empirical research questions
- interview guide
- empirical knowledge

Advise: examine 5 interviews and do a list of codes that will be applied to the others interviews - use part of your materials for codes identification!

- separative codes can be grouped into family (code family manager)
- by coding each sentence/small lines, you can easily lose the context, which is very important!!! these way you can easily manipulate with data;
- by coding small/larger paragraph you can try to keep the context, but can end up with coding great amount of text

1. indexing of textual units (coding)
2. synopsis of all textual units which have certain categories and possibly other characteristics in common
3. identification of structures and patterns in the data material which can lead to new categories

Further Strategies after Initial Coding

a) collect all text segments that are tagged with the same code and compare them
b) co-occurrences of codes (collect all text units that are tagged by the same two, three codes)
c) reorganizing data by axial coding or selective coding (identifying core categories, find relations between codes)

Use coding as a tool for indexing large amount of text, read those selected part of text again and again and search for interesting patterns! Do not use coding for interpreting or for making conclusions!

- in qualitative analysis you have to explain all 100% of variance!
- in quantitative you have statistical "drop out" for which you don't have to worry

QUALITATIVE CONTENT ANALYSIS

- you take the info away from the text, reformulated it and in future deal only with this extracted info; the original text stays in your files, you can always come back, but you are processing only info you find relevant!

- in order to extract info from the text you have to have a very clear idea what you are looking for
- method is as open to changes as is coding, but while you can start coding without any idea what you are looking for, you need to know what you are looking for, because you are using a tool for extracting only the info you find relevant (following your theoretical considerations)
- it extracts what was actually said (coding explains what was the specific part of the interview about)

1. **define the variable** - a definition that gives us an idea what we are looking for

e.g. funding rules of the university - all rules of the university and its units which prescribe the distribution of funding for research purposes

Dimensions of the category:

Time: for what point in time/period of time is there information about a rule (IMPORTANT - the sequence of events should always be reported!)

Scope: to whom does the rule apply?

Subject matter: what is governed by the rule?

Content: what does the rule say?

Causes: what causes for the existence of the rule are reported?

Effects: what effects of the rule are reported by the interviewees?

- mark paragraph - **paragraph is a unit of analysis**
- create a macro for extraction rules related to the funding of university
- creates a table and each time we extract info from the text, it will be placed into table (dimension/category)
- we have to translate the relevant info into our analytical language (e.g. "at the moment"...we could translate it as a time of interview; "in our school" = school of biosciences; subject matter=automatic recurrent funding for researchers
- we have to be precise while "decoding" word/phrases from interview into our analytical language to fulfill the categories
- you can put a sign for your interpretation, e.g. causes = ((bad financial situation)) - stavimo oznaku, recimo s duplim zagradama za nešto što je za sada naša interpretacija, nije jasno istaknuto u intervjuu, ali iz konteksta možemo pretpostaviti; ovaj se moment razvija dalje kroz cijeli intervju, ali u momentu prve analize i impresije, obavezno trebamo staviti oznaku da je to naša interpretacija!

time	scope	subject matter	content	causes	effects	source (link with paragraph)
time of interview	school of biosciences	automatic recurrent funding for researchers	none	((bad financial situation))		- important if you want to consult the original text - very important to keep the link with the source

- proces se može/treba ponavljati dok ne iscrpimo sve informacije iz odlomka!
- this is how you gradually get the info from the interview and play around with your data

- we have to link all interviews with this categories
- it is a huge reduction of information (but usually not in terms of numbers of papers because table ends up with a lot of empty spaces)
- you can/have to sort the table, remove sufficient rows (reducing information) if they have the same/very similar info; it is very important in that case to cut-paste the source, and to have all the links to original source kept!
- it is easier to search for the patterns
- you can add additional things in the table, it does not affect the macro!!! (because table works with default values)

- very important - keep your backup!!!

SUGGESTIONS FOR THE INTERNATIONAL COMPARISONS
- agree on the initial set (common) categories - inform other if need to change or add the category
- think about the subject matters - maybe to agree on the analytical language that will be used
- this way the table can be produced in english language, suitable for every team

Meiden (??) - pogledati - offers plenty of techniques for content analysis

- still the technique used the most - printing and reading carefully interviews+making notes +subconsciously coding/making categorizations

!!! important to make a protocol of your own coding rules/extracting decisions !!! (what categories to use for certain and ambiguous cases/decisions) - e.g. when having a time dimension use CAREER STAGE as variable/code family

TUESDAY, 22nd Feb, CODING EXERCISE (Amy, Elke, Luminita, Boba)

1. extract all information (be as specific as possible)
2. use your own values (don't be lazy!)
3. change categories/dimensions (don't be lazy!)
4. log your decisions! (your own rules and protocol)

PROBLEMS AND SUGGESTIONS

- TIME DIMENSION - use any time marker from which you can piece together chronological info (while i was at the institute, before i got my first child, during my postdoc...)
- no multidimensional possibilities in coding (coding means attaching individual code segments to the text) - while life is multidimensional:)
- number of codes - you can easily end up with having a lot of codes (it is connected with inability to create a structure before the process of coding) - you can use a network of codes later
- lot of interpretation problems
- !!! very important that analysts analyze interview they conducted themselves !!! (the easiest way to transform the context knowledge)
- connection between data and theoretical model
- rules become very important when more people are included into the process of analysis - there is rarely THE BEST way to extract, you have to be consistent, especially within the team - you have to solve ambiguous information

- it is impossible to formulate definitions and rules that have unambiguous covered whole cases (suggestion: let each member code or extract one or two same interviews and layer discuss the commonalities and differences, reach and agreement on the codes/categories/dimensions/ definitions - it is very important to correct interpretations and to identify categories that are not sufficiently clear - it is very important to discuss these issues within the team working on the analysis)
- very important to be critical about the instruments we use and the procedures we use to interpret data
- you have to check and validate your own work, specifically while doing qualitative content analysis (if you do mistake in coding, not a problem; but with extractions the effect your mistakes might have on the interpretations is much influential)
- tag/extract the quotes that best describe your interpretations immediately while coding/extracting the part of the text, don't wait until the end
- keep notes about your impressions and implicit info related to the interview (e.g. interviewee was in the very bad mood/has a huge problems...)
- both methods are disruptive, need to handle our general/specific impressions

How to achieve comparability in international projects:

- have to work together!!!
- develop categories and mutually develop theoretical framework, methodology, categories/ dimension
- it has to be used by analyzing the same interview (at least one interview from each country has to be translated in english - money issue!!! - would it be easier to conduct one interview in english?)