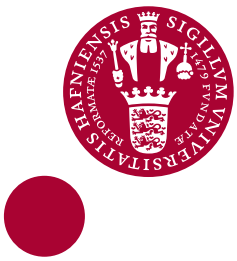

The data cycle

Mortensen, Janus; Hazel, Spencer

Published in:
Kansai University International Symposium

Publication date:
2012

Document version
Publisher's PDF, also known as Version of record

Mortensen, Janus and Spencer Hazel. 2012.
The data cycle. In Keiko Ikeda and Adam Brandt
(eds.) Kansai University International Symposium:
Challenges and new directions in the micro-analysis
of social interaction. Osaka: Kansai University. 22-29.

The data cycle

JANUS MORTENSEN SPENCER HAZEL
jamo@ruc.dk spencer@ruc.dk

Roskilde University / CALPIU Research Center

Abstract
This article is of a very practical nature. It presents an outline of some of the central steps involved in research concerned with social interaction in naturalistic settings. We focus specifically on the technical challenges involved in research of this kind – the craft of data collection and data handling – with the aim of providing an overview of the choices that need to be made at various stages in the process of planning and executing a research project. The article is perhaps particularly useful for newcomers to the field, but more experienced researchers may also find one or two items of interest.

Research projects concerned with the analysis of social interaction in real-life settings, for instance in the fields of Conversation Analysis (CA), Interactional Sociolinguistics and some forms of Social Psychology, are complex processes that involve a number of distinct, interrelated tasks. In idealized form, the central tasks can be listed as follows:

1. Briefing participants and obtaining informed consent
2. Selecting recording equipment
3. Recording
4. Storing recordings
5. Preparing data for transcription and analysis
6. Transcribing data
7. Analyzing data
8. Presenting results and sharing data

In practice, the steps are not ordered in neat succession, nor are they as clearly delineated as this list implies. But for our purposes here, it makes sense to treat them separately because each step requires distinct choices to be made. In the following, we will discuss the steps in turn, and highlight some of the considerations that we believe should go with each of them.

1. Briefing participants and obtaining informed consent

When participants are briefed about the recordings they are about to be part of, it is obviously necessary that all relevant rules and regulations, from the level of applicable national law to local university guidelines for obtaining informed consent, are adhered to. As the details concerning this vary from place to place, we cannot provide specific advice here.

However, for an example of documents used in this connection by a large-scale research project in Denmark, see calpiu.dk.

It is also an important point to remember to set out clear principles for how the data will be stored and used. If you want to share the data with other researchers, this should be part of the agreement that you enter into with the participants. Here it will usually be important to address questions of participant anonymity, especially if you are working with video data. See more about this issue under point 8.

Some of the other central questions to keep in mind when informing participants include how much you want them to know about the details of the study - and how much they are entitled to know. You might also want to consider how much *you* would like to know about *the participants*. In some research traditions (particularly CA) background information about the participants is considered largely irrelevant, while in others (like for instance sociolinguistics) things like the age and social background of participants will often be considered important because this information may be used as variables in the analyses of the collected data. This information can either be obtained through questionnaires, or by face-to-face interviews. You may also need to take into consideration how much time such procedures require of the participant(s). In certain settings, for example service encounters, the recorded interaction may last only a few seconds, a few minutes at the most; in such circumstances, if obtaining informed consent is compounded with lengthy questionnaires or interviews to elicit additional background information, participants may be more reluctant to agree to their participation. Prior ethnographic observation of the setting should help you develop some sensitivity to how much time you can ask of each participant, and what might be the right way to go about these procedures.

2. Selecting recording equipment

When selecting recording equipment, there are (at least) three basic choices to make. The first is whether to produce video or audio recordings; the second concerns what media should be used for the recording and the third concerns the file format (the latter is only indirectly applicable to non-digital recordings).

The use of audio recordings has been standard practice for many years in CA and similar research traditions, but with the advent of affordable (digital) video recording equipment, the use of video has become more and more common. In general, we would recommend that researchers use video recordings as far as possible, even if the research they are pursuing only relates to 'what people say'. A recording, whether audio or video, will always be a highly imperfect representation of the recorded event. However, all other things being equal, an audio recording will result in a greater loss of information than a video recording, and it is therefore recommendable to use video in order to optimize the information captured in the data. Exceptions occur, of course, for instance in cases where the recording is only approved by the participants on the condition that it is audio-only, or when the use of cameras is considered too intrusive.

Audio and video recordings can be stored on various types of media (tapes, DVDs, computer hard drives and flash memory). For research of this kind, digital recording straight to a hard drive or, even better, flash memory media like SD cards is much to be preferred, simply because these media types are much quicker to work with than other

storage facilities. Working with tape recordings (even digital ones), for example, entails a considerable amount of additional post-production work, because the recordings will have to be transferred into a format and onto a medium compatible with present day computer technology.

It is difficult to point to 'the best' digital format for audio and video recording, partly because there are already many good formats available, partly because new formats continue to appear, and partly because the very concept of 'a format' is quite complex in this context. However, it *is* possible to point to some general guidelines for best practice.

In terms of audio, it is generally a good idea to opt for a *lossless* format, i.e. a format that does not reduce the quality of the recording in order to minimize file sizes. The Waveform Audio File Format (also called WAVE or .wav after its file extension) is probably the most common format used for lossless recording, and most quality audio recorders will be able to store audio in this format, or similar formats like Apple's AIFF. The popular MP3 format is the most common *lossy* format, and many recording devices will allow you to record in this format. However, following the principle of minimizing information loss in the act of recording, we would generally advise against using MP3. WAVE recorders and MP3 recorders are not far apart in terms of price, and though WAVE files tend to be bigger than MP3 files, this rarely presents a problem given how cheap digital storage is today. Another reason for opting for WAVE recordings is that they usually integrate seamlessly with the kind of software used later in the research process (see below) whereas MP3 files often need to be converted to work properly, and this takes up valuable time.

Most video formats are lossy because raw, uncompressed digital video is simply too large to handle with today's consumer level technology. However, as in the case of audio formats, there are certain file formats that tend to be preferred over others. The two most popular file formats for this kind of research are probably .mov files which are associated with Apple's QuickTime and .mpg files that are more generic. The problem with video files is that the data they contain can be encoded by a seemingly infinite number of different *codecs*. So having a .mov file does not necessarily mean that you will never encounter any compatibility issues running the file with the software tools you are using. But in our non-expert experience .mov files and .mpg files are generally quite reliable.

A general piece of advice is to avoid proprietary formats, i.e. formats that are developed by individual manufacturers and more or less tied to their devices and software. Another piece of general advice is not to rely on formats that require too much work in post-production. Here modern HD video cameras are not actually very user-friendly, because it requires multiple steps to get from the recording on the camera to a file that you can use as input for the software mentioned below under steps 6 and 7.

3. Recording

When it comes to the actual recording, there are also a number of choices to be made. Some research traditions prefer to optimize the technical quality of the recording and for that reason issue participants with head-worn microphones, and follow them around with handheld cameras.

Other traditions are less concerned with the technical quality of the recording and may want to optimize what could be called 'the quality of interaction' instead, for instance

by using unmanned cameras and microphones placed on tables or walls in an attempt to make the recording process as unobtrusive as possible (though it will obviously never pass unnoticed unless the recording is clandestine, which is generally not considered ethical, and in some cases even illegal). It is worth anticipating these choices at an earlier stage (step 2 above), as the type of data you choose to collect will depend on the right equipment being available to accomplish this. For example, you may aim to impact as little as possible on the setting where the recording is being carried out. To achieve this you may opt for smaller, pocket cameras that can be taped to a wall or placed on a shelf, rather than larger, better quality video cameras that need to be placed on tripods.

It is generally a good idea to have at least two recording devices running at the same time (for instance a video camera and audio recorder) since it is not uncommon for things to go wrong one way or the other during the recording process, and it is a very unfortunate thing not to capture anything at all when you have spent considerable time getting participants to sign consent forms and have set up all the recording equipment.

For some research purposes it is beneficial to use multiple cameras, for instance in order to capture interaction in a classroom. In that case, it is worthwhile to get hold of a clapperboard, or get into the habit of clapping your hands at the start of each recording session, to assist the synchronizing of picture and sound in post-production (see step 5 below).

4. Storing recordings

Once the recordings have been made, they should be stored by following carefully designed procedures. These procedures should involve rules for file naming, keeping a general log of recordings and participants, and secure storage.

In most cases, the researcher promises the participants that their participation in the project is anonymous (cf. step 1). This means that data must be stored in way that ensures that outsiders cannot get access to it. In many cases, it will be possible to use a university server for this purpose, but depending on the amount of data that is generated, this might not be a viable option. In that case, it might be an idea to use portable storage devices like external hard drives. However, it should be remembered that such devices can be lost – forgotten on trains or planes – and therefore any piece of data stored on them should be encrypted. Another storage option is to sign up for a 'cloud' solution like DropBox. However, here the question of anonymity and data access/security also needs to be carefully considered.

Under all circumstances it is vital that the data is backed-up consistently, and preferably to a remote server. Losing data is a very costly affair (in time as well as money), so money spent on a robust back-up solution is money well spent.

5. Preparing data for transcription and analysis

In some (rare) cases, it is possible to work straight from the raw audio and/or video files. However, in most cases, it requires some additional work to prepare the recordings for subsequent steps. The choices made in steps 2 through 4 should ideally minimize the work needed in step 5, but it is rarely possible to do away with it altogether.

In many cases, the high quality audio or video that has been produced as part of the recording session will simply be too big and heavy to work with for transcription and

analysis purposes. In addition, the data may also come in a format that is not specifically geared towards the software that is needed later in the cycle. So, step 5 will often involve *compression* and/or *conversion* of the recorded data. Additionally, it may also involve splicing files from different recording devices, merging multiple camera angles into a single file and the like. It can be tempting to delete original files and replace them with compressed or edited ones. However, in general, we would recommend that originals are kept alongside the derivatives. Again, the underlying principle is that you should never throw information away that you – or someone else – might want to use later.

There are various software tools available for post-production of the sort discussed above. At the end of the article, we have included a list of the tools that we have found useful, but you should be warned that lists like these tend to become outdated very quickly. Our general guideline is that the software that is used should be as simple as possible while still being able to do the thing it needs to do. Using very powerful (and expensive) packages like Adobe Premiere or Final Cut Pro is rarely necessary for research of this kind and is always highly time consuming for novice users. We have found the (inexpensive) pro-version of Apple's QuickTime to be very useful for most of our purposes, in combination with Handbrake, which is a (free) conversion/compression tool.

Finally, when working with large data sets, step 5 may also involve a careful sifting through of the recorded material in order to identify passages or sections that are of special interest to the researcher, and which will then be selected for further analysis. The traditional way of doing this is to watch the video/listen to the audio while making notes on a piece of paper. However, this procedure can be radically improved by using linking software like ELAN, CLAN or EXMARaLDA (see more below, and Hazel, Mortensen and Haberland, this volume).

6. Transcribing data

For the transcription of the data, there are a number of interrelated choices that need to be made, concerning what tool you wish to use to carry out the transcription, what level of detail should be included in the transcript, and what transcription conventions you will employ.

Transcripts have been and are still produced using such wide-ranging tools as the humble pencil and paper, the typewriter, the word processor, and more recently, specially developed transcription software programs. What sets the latter tools (for example, ELAN or Transana) off from the other three is the way the transcribed text can be linked to the recorded data, allowing instant access from transcript to media file. A frequently voiced concern is that to learn to use one of these tools entails an investment of valuable time when you may already be competent in word-processing tools such as Microsoft Word. Our experience is, however, that this investment very quickly pays dividends not only in terms of time saved at later stages of the research (see below), but also in terms of the quality of the eventual analysis. With the data digitally aligned to the transcript, relevant sequences identified throughout large tracts of transcribed recordings can be made available almost immediately for listening and reading, without you needing to manually scan through the audio or video to find this or that particular item of interest. This also has the

consequence that the researcher is always encouraged to attend to the *recorded data*, rather than to the *transcriptions of the data*.

If you do opt for transcription linking software, you will need to consider which of the programs best suits your needs. In some cases, this decision may have been made for you, as you may be part of a research group that aims to create a uniform shared database between its members. If that is not the case, one choice you will need to consider is the particular electronic environment provided by the different programs. Generally, one choice you will need to make will be between a *partitur editor* and a *transcript editor*. The main difference here is how the text is represented on the screen, with the *partitur editor* (for example EXMARaLDA) representing the interaction in a musical score format, and the *transcript editor* (for example CLAN) working with the conventional format of the printed page. Each environment has its advantages and indeed both can be used in conjunction with one another at particular stages of the research.

In transcribing interactional data, it is wise to be consistent with the level of detail that one includes in the overall transcripts. If the levels of granularity fluctuate within a transcript, it can lead to uneven readings, where it may appear that more seems to be happening in some section of the data than in another, whereas in truth what has changed is simply the level of detail included by the transcriber. In much the same vein, you need also to consider which transcription conventions are to be included in the transcribing, and be clear and consistent about the conventions you use. This is especially true in situations where more than one transcriber is employed to generate the body of transcripts.

7. Analyzing data

The way data is analyzed differs considerably from one research tradition to another other. However, for most types of analysis, it is now the case that there is a range of software tools available that can assist the researcher in the analytical process. For some approaches, qualitative data analysis software packages like NVivo, Atlas.ti, or Transana are very popular. However, the linking software tools mentioned under step 6 above can indeed also be utilized for a range of analytical purposes, including coding of various sorts, although they may not be as sophisticated and/or user friendly as the packages that specialize in qualitative analysis.

Under all circumstances, it is important to be aware of the fact that many of the choices that have been made in steps 1 through 6 will to some extent limit or at least affect the range of packages that are (immediately) available for use. For one thing, the file formats can be an issue, though maybe not so much today as five years ago. However, if you know that you are going to use CLAN or NVivo for your analysis, or possibly both, it is crucial that the recording equipment that you use actually produces files in a format that fits these programs. Similarly when devising a matrix for naming your files, make sure not to include characters that are not accepted for files to be loaded into your analysis program of choice.

Finally, an important issue to consider when deciding which software package to use (if any) is *interoperability*. The selection of software tools should be made not only with a view to what the individual tool can do, but also with a view to its compatibility with other packages, i.e. the extent to which it allows the researcher (or other researchers

in the same research team) to pursue specific interests by means of other/additional tools than the ones originally selected.

8. Presenting results and sharing data

Once the analytical work is done, researchers will typically want to share their findings. This may be done in a number of different ways, including delivering talks at seminars or conferences, publishing articles in scholarly journals, posting notices on research blogs and so on. At this point, a number of choices made earlier in the data cycle will become particularly pertinent.

For one thing, the type of consent obtained from the participants in step 1 will be crucial in determining in what shape and form the data can be presented to an audience. Have participants agreed to video clips being shown at conferences? If not, the absence of the video might make it more difficult to present a convincing argument. If the participants *have* agreed to something like this, the use of linking software in the transcription phase will prove very useful, since the linked transcript can either be played back as part of the presentation, or alternatively be used as a means for producing subtitles for the video. Presenting data in this way is much more consolidated and time-efficient than using a combination of handouts with transcripts and video projected on a screen.

As far as print publication is concerned, screen shots would usually be the relevant option, however, in the not so distant future it will probably be common to have data recordings available in conjunction with written analyses. This is already the case in certain journals, for instance *Language & Communication* (cf. the editorial for issue 31, 2011). This means that the choice of recording equipment and data format in step 2 should also be made with the standards of potential publishers in mind, or alternatively be taken into consideration under step 5 when preparing the data for analysis.

The possibility of publishing not only transcripts but also the recordings which the transcripts are based on also means that the use of linking software becomes even more attractive. For one thing, the pressure to produce highly detailed transcripts will be somewhat alleviated by the fact that the recording will be freely available for anyone who might want to check particular details. In essence this means that sharing a linked transcript along with its associated media file allows colleagues to scrutinize analyses at a level of detail that would otherwise be impossible, or at least unrealistic, simply for practical reasons. While a check-function of this sort will probably provoke a certain measure of anxiety in many of us, there can be no doubt that it will improve the scientific rigor and reliability of the work that is produced in our respective fields.

Finally, when the research has been published and the research project is done and dusted, the question of what to do with the data may arise. Assuming that the necessary permission has been granted by the participants in step 1, the researcher may at this point wish to go back to step 6 and consider whether the data might be better off stored in a communal database of some sort rather than on an individual hard drive. If the researcher decides to share the data, all the choices that have been made during the data cycle will have a bearing on how useful the data will be for other researchers. So if data sharing is something that should be promoted – and we think it should, because it will help further

our joint research enterprise – the choices that the individual researcher makes as part of the data cycle should not just be seen as individual concerns, they are important choices also for the larger scientific community.

Final remarks

In closing, we should like to emphasize that although the choices we have discussed above have been treated in relation to specific 'stages' of the unfolding research process, it has in fact been our intention to demonstrate how each choice is inevitably part of a matrix of choices and decisions. We hope – and believe – that an awareness of this matrix may be of help to researchers working with data cycles of the kind treated here. We would like to stress that our argument is not meant to be prescriptive. Each research project will have its own individual fingerprint, which will necessarily entail specific demands and lead to particular choices. However, we believe that a general awareness of how choices in the data cycle are part of a bigger web of possibilities may act as a guide in handling the choices in practice.

A SELECTION OF RESOURCES

Amadeus Pro
www.hairersoft.com
Audio editor
Mac OS only
$60

Audacity
http://audacity.sourceforge.net/
Audio editor
Windows and Mac OS
Free

Quicktime 7 Pro
http://www.apple.com/quicktime/extending/
Audio and video editor
Windows and Mac OS
$30

CLAN
http://childes.psy.cmu.edu/clan/
Transcript editor
Windows and Mac OS
Free

ELAN - Linguistic Annotator
www.lat-mpi.eu/tools/elan/
Partitur editor
Windows and Mac OS
Free

EXMARaLDA Partitur-Editor
www.exmaralda.org/partitureditor
Partitur Editor
Windows and Mac OS
Free

Handbrake – video transcoder
http://handbrake.fr/
Video converter
Windows and Mac OS
Free

Alive Video Converter
www.alivemedia.net
Video converter
Windows only
$55