Title: The transcription as an archival document
Author: Sandra Orihuela
Translator: Salvatore D'Errico
1. Introduction

When speaking of transcriptions in an archive, one generally thinks of the transcription that converts a relatively old handwritten text into another text adapted to a more current and intelligible notation system. The transcriptions that will be discussed in this guide, however, are those originating from sound and audio-visual documents; those texts that copy into a written form what has previously been stated orally. The challenges presented by this type of transcription have to do with the transition from oral language to written language, with the change of medium and code in documents in which both form and content are relevant. This transition from one language to another can raise several questions:

- What exactly is a transcription and how important is it in an archive?
- Can it be considered as just a copy of an original document?
- Is there a universal and correct way to transcribe?
- How can those elements that are typical of orality be put into words?
- Can the transcriptionist’s intervention have any effect on the final document?
- Is it possible to replace the manual transcription with an automatic transcription?

We will answer to these questions, and analyse what a transcription is, what transcribing means in the context of an archival institution, which general guidelines must be considered, and what are the possibilities of current technology in relation to automatic transcription.

2. Transcription and its characteristics

In general terms, one can argue that the transcription is a document (although not necessarily an archival document) that is a written copy of another text or of a spoken discourse. This copy of the spoken speech may be made from a recording or from notes taken at the time the speech was heard or recorded. Additionally, the transcription can be performed using a particular character or writing system, which allows representing certain elements for which the traditional writing system is not sufficient, as in the case of phonetic elements.

The transcription as an archival record, however, introduces new issues when dealing with the document because of its particular use. In this context, the transcription is intended as a source by the researchers. Although it is the original document that has ultimate proof and evidence value, it is in fact the transcription that becomes the main reference that is consulted to produce evidence. In this sense, it is paramount that, when transcribing, the desire to faithfully reflect what happened in the original recording is respected.
The transcription can be considered as an archival document because;

- Helps researchers to assess quicker whether the document is useful and relevant to their research, because it requires less time to read a text than to watch and listen to a recording.

- Facilitates the conservation of the original support, which tends to be unstable. In the case of old documents, magnetic tapes’ damage is avoided, for instance.

- Facilitates the preservation of information, which is contained in media affected by constant changes in technology and that are, therefore, very fragile.

- Facilitates the access to information, because it allows to clarify the ambiguities and rectify possible errors by the interviewee (in dates or locations, for instance). In this sense, the transcription can be more than a mere copy. It can contain new layers of information that the original document did not include.

In general, the original document is the sound and audio-visual document. However, its written version can be much more practical for study and analysis. A corrected and indexed transcription not only facilitates conservation, preservation and access, but can also add value by clarifying ambiguous passages and correcting possible errors in names, places, dates, etc. Yet, it must be noted that transcriptions can also entail some inconvenience; above all, the loss of part of the information included in the recording, since not all the characteristics of oral discourse can be translated into writing. This is because in sound and audio-visual documents one can find also other components present in the context of the recording, such as those understood by the participants. The information is not only in the language but also in the omissions, in the looks, in the intonation, in the pauses and in facial or body gestures.

In short, the information emerging from oral sources is relevant, but the way in which the speaker decides to convey the information is equally important. The transcription must be able to transfer, as much as possible, those characteristics of oral and gestural language into written language, which has its own code and is insufficient to recreate everything that is present in the original recording. By acquiring the written form, language loses spontaneity and must follow new rules. A good example of this is what happens with variations in intonation, which in a text are reduced to changes in punctuation – commas, periods, hyphens – that are usually insufficient to express all the changes in tone.

However, there is not a single way of doing transcriptions; each institution must decide what to prioritise and which elements should be turned into words and, from there, designate which notation system is going to be adopted, and what information is kept.

Broadly speaking, we can determine three main types of transcriptions:

1) A condensed transcription, in which the elements of orality are deleted, such as doubts, hesitations, and false starts of the speaker. In this type of transcription, the attention is focused mainly on the information provided by the interviewee. This is the option that is usually used in the publications.
2) A transcription trying to convey the complexity of oral language through a notation system that indicates, thanks to the use of different symbols, particular aspects, such as differences in intonation, emphasis, and pauses. This form is the one that is usually used in disciplines, such as conversation analysis, in which importance is given to the way things are said, and in which the transcription responds to very specific scientific needs. The symbols used are, therefore, many and very complex, and interpretation is difficult for those who are not familiar with these disciplines. Furthermore, as in any transcription, the meaning of the symbols is not universal, which makes understanding even more difficult.

3) A transcription combining the two previous options, in which some of the oral elements are deleted, but others are preserved to give an idea of the “texture” of the spoken speech, and of the linguistic particularities. In this type of transcription, one can find symbols that usually are not present in the most common texts, but that do have a certain written tradition and can be easily recognised by the reader. This option is usually the most recommended for files, because it allows to include both aspects that respect the content and preserve features of orality that may be relevant, without reaching the complexity of the second option.


In reality, it has little consequence if one opts for a transcription that eliminates certain aspects, such as false beginnings, or if, on the contrary, one prefers a transcription that respects as much as possible all the elements present in the original document. The ultimate goal is to obtain a written text allowing a clear understanding of what happened and has been recorded. The resulting document must be both faithful and understandable.

Whatever perspective is taken as a starting point, the transcription must represent what is heard, or what is heard and seen, clearly and with respect to what is found in the sound or audio-visual document. The decisions regarding what remains or what is eliminated...
are determined by the use that is going to be given to the document, and by the approach that each project takes as a starting point. If the institution considers that the most important thing is the information that emerges from the speech, it may not be necessary to make a transcription that collects most of the elements of orality. If the archive is more interested in also recreating the environment in which the recording took place and the linguistic particularities of the speakers, the transcription should be, therefore, more detailed.

All of these decisions can be considered interpretive interventions by the transcriptionist, whose work has, therefore, an impact on the final document. Every decision about what to keep or what to delete or what symbol to use is an interpretive intervention, that is, a way of mediating between the original document and the researcher who consults it at a later stage. These interventions can be summarised as:

- Corrections and clarifications regarding possible errors in places, dates, names, etc.
- The inclusion of certain implicit elements in the dialogue that are not fixed by the language, such as those already mentioned.
- The inclusion of facial and body gestures.
- The elimination or not of repetitions and hesitations.
- The elimination or not of dialectal peculiarities.
- The use of punctuation marks and separation in paragraphs.

Transcribing, therefore, is not a neutral practice. It always involves a point of view, a perspective that interprets and constructs the story, which is eventually going to be read by someone else. It is based on decisions that, on many occasions, can be taken in an unconscious way. The key is to act responsibly before these interpretive decisions, which are inevitable, but also necessary, are taken. In this sense, it is advisable to define from the beginning what methodology the institution is going to follow to perform transcriptions, while being transparent about it, and providing the users with guidelines.

3. General guidelines for the creation of a transcription guide

As already mentioned, there is no universal and adequate way of transcribing, although there are guides that propose a specific method. Some general best practices can be established from the consultation of some guides. In this way, each institution can develop its own internal manual, based on their objectives and needs.

The first step, therefore, is to decide what information is deemed relevant, and what will be kept in the written version of the recording: i.e. prioritising the mere content or the different elements appearing in the original document (such as non-verbal language, facial and body gestures or notes on the environment in which the recording took place).

From here, some minimum methodological principles or basic criteria must be taken into account:

- **The elaboration of categories**: it is necessary to define the categories that are going to be transcribed, and assign them a symbol or spelling that will correspond only to that phenomenon. For example, if it is decided that the transcriptionist's comments (such as corrections) are to be indicated in square brackets, the same
symbol should not be used for other types of annotations. The important thing is not what symbol is used, but that it has a unique and clear meaning.

- **Legibility and economy**: the transcription must be legible and, therefore, it is necessary to use graphs, symbols and other resources that are easy to interpret. In other words, care must be taken not to use too many symbols, and to use those which people are already familiar with, because they already have a relevant written tradition: italics, bold, underlined, parentheses, square brackets, etc.

- **Accessibility and robustness**: symbols must be easy to identify, learn, and use. Both the transcriptionist and the researcher must be able to learn or intuit the symbols’ meaning. Furthermore, the transcriptionist must have it easy to utilise them: that is, the resources must be easily accessible from the keyboard of any computer.

- **Computer processing**: it is advisable that the transcription facilitates computerised searches. This can be more or less difficult to respect, depending on the amount of speech particulars that one wants to add to the transcription. If it is decided that variations in pronunciation are going to be graphically recreated, for instance, the automatic search will be difficult; there will be some words that, being in fact the same, will not be detected because their form will be expressed in a different way.

Taking these basic criteria into account, one can define the script with the main elements that the transcription must have. These elements are based on different aspects of formatting, editing, reviewing, and style and are part of those interpretive decisions that must be taken.

**Format:**

- Choose font and font size, page margins, indents and spacing. The use of Times New Roman, size 12, with double line spacing and with page margins of about 2.5 centimetres is generally proposed. In practice, one may use any font that is easily readable, and line spacing and margins that are also visually agreeable and easy to read.

- Define what should be included in the transcription’s heading: names of the project and of the institution involved, full names of the interviewee and of the interviewer, place of the interview, etc.

- Decide how tape changes or interruptions for technical issues in the recording are indicated.

- Decide whether to include a report that accompanies the transcription, and define the elements needing to be added: comments regarding the environment in which the recording was conducted, notes on speech peculiarities, general information about the participants, justification of the interview method and type of transcription used.
Editing and revision:

- Decide how the transcriptionist's corrections and other annotations are indicated. This can be by means of symbols (brackets or parentheses, for instance) or in separate notes, in the footer, or at the end of the transcription. Normally one or the other option is chosen depending on the length of the transcriptionist's comment. If it is only a matter of completing a name or correcting a date, it is usually noted in the same text of the transcription. If, on the other hand, it is a somewhat more extensive clarification, it can be noted at the bottom of the page or at the end of the transcription.

- Decide whether to indicate the implicit elements in the dialogue that are not fixed by the language, such as understatements or irony. Should one opt to include them, one must define how it will be done.

- Decide if the elements of orality are indicated; changes in intonation, pauses, facial and body gestures, etc. It is important to bear in mind that some of these aspects can only be noted in the case of audio-visual documents, in which the sound is accompanied by the image.

- Should one decide to include them, one also has to define how it will be done: whether to include all of them or not, and what symbols are going to be used. An example of this could be the use of taglines by the participant in the recording. Depending on the type of transcription, one can decide to eliminate those taglines completely, keep them all (although this could make reading difficult) or keep only some (one in four, for instance) to give an idea of some peculiarities of the spoken language.

- Decide whether to include or eliminate false starts, hesitations, and repetitions. Should they be included, define to which extent; that is, if all of them are kept or just part of them. There are institutions that prefer to eliminate them completely because they hinder reading; others opt to keep those that provide certain information, such as a certain nervousness or difficulty in discussing certain topics.

- Define how unclear or unintelligible passages are indicated.

- Decide whether or not dialect or pronunciation particularities are included. Should one decide to include them, clear limits and the form to be used need to be established: for instance, it is not possible to transcribe all the pronunciation variations, but the most striking or significant ones. On the other hand, certain dialectal peculiarities can be noted, including the meaning to ensure correct interpretation.

- Decide whether or not to include certain aspects of the recording’s context: background noise, sudden interruptions, etc. In the same way as with false starts and hesitations, these elements can be eliminated completely if it is considered that they do not contribute anything. Some projects, however, include them with the desire to reflect as faithfully as possible the context in which the recording took place.

Style:

- Clearly define what symbols are going to be used and what they mean: bold, italics, underline, hyphens, commas, periods, parentheses, brackets or any other resource that one wants to use. Assign each symbol to a certain category and, therefore, to a unique meaning.
Establish minimum principles regarding separation into paragraphs: whether it is done or not and, if so, what guidelines are to be followed.

Define the criteria for the footnotes and for the notes at the end of the transcription.

Define how abbreviations and acronyms are used.

Define how to indicate the letters (in the case of spelling, for instance).

Define how to indicate numbers: Roman numerals, dates or time.

Decide whether to use symbols or words to indicate percentages, currencies, etc.

Define how to indicate the quotes and paraphrases speakers can enunciate, regarding works or regarding what other people have said.

4. Other guides
The above guidelines have been collected from the comparison of existing transcription guides, created from the perspective of oral historians, and from other texts proposed by linguists and other social scientists.

All the articles consulted are detailed in the bibliography. There are two main manuals that are worth commenting on in more detail, the *Style Guide: A Quick Reference for Editing Oral History Transcripts*, published by the Baylor University Institute for Oral History (BUIOH), and the *Oral History Transcription Guide*, developed by the Columbia University Center for Oral History Research (CCOHR). Both manuals are the most recent guides (2018), usually recommended by English-speaking oral history associations, and are a good introduction to transcribing oral interviews.

The Baylor University Institute for Oral History (BUIOH) manual follows a quite simple structure, consisting in a brief introduction, a note on formatting, general spelling and transcription correction rules, and a guide on alphabetically arranged style. The Columbia University Center for Oral History Research (CCOHR) guide, on the other hand, presents the information in more detail, adding the table of contents to facilitate the research, and appendices with templates and examples.

The two guides focus on the same topics. However, the CCOHR does so by breaking down the topics further; introduction, process, format, reviewing, editing, style, appendices, and bibliography. Although one may consider that the CCOHR guide is more complete, in general, both manuals coincide in most of the guidelines, both in form and content. That is, the two institutions largely use the same symbols to indicate different aspects, and also follow similar principles regarding what to remove and what not to remove from transcriptions.

As mentioned, both are good introductions and can be a good starting point when developing your own guide without having to start from scratch. The two guides are good sources to bring your own adaptations, especially if you choose to make transcriptions that are halfway between the simplicity of publication and the complexity of discourse analysis.
Despite this, it is important bearing in mind that these guides are in English and, therefore, much of the advice they offer refers to the ambiguities and problems that this language may cause. In the case of making transcriptions in other languages, it will be necessary to adapt the guidelines to the specific characteristics of the language concerned.

5. Automatic transcription

Transcribing from scratch can be a long and costly process and, for this reason, in recent years the possible use of automatic transcription has been considered as a substitute for manual transcription.

Automatic transcription is made possible by speech recognition technology, or automatic speech recognition (ASR), which allows a computer programme to process the human voice, and translate it into written text. For this process to take place, the software must be able to recognise different aspects of the language, such as grammar, syntax, and structure, as well as the elements of voice signals and their composition. The ASR system, therefore, not only captures what is said, but also interprets and transcribes it; it divides the sounds into parts that it can recognise, analyse, and put into words. This technology, which can only be developed by merging different disciplines (such as linguistics, mathematics, and statistics), is programmed and improved as more data are provided, and the algorithms are perfected. That is, the more data have been entered, the more likely it is to make correct assumptions regarding speech patterns or words that appear related. ASR technology, therefore, not only analyses what one can hear, but also provides solutions for fragments that one may not have captured correctly, based on algorithms and previous data. This is why the system can be improved and adapted; for instance, its accuracy can be increased if there are words on a specific topic that appear frequently.

ASR systems work thanks to different components that allow to divide the tasks needing to be developed; recognising the voice input, extracting its characteristics, decoding it and giving an output to the transcription by using appropriate words. Decoding is also possible thanks to acoustic models, pronunciation dictionaries, and linguistic models of the languages to be transcribed and that have been previously introduced so that the software can recognise them. On the other hand, software in the field of artificial intelligence does not only work by recognising data, but also through ‘supervised learning’, which implies that the programme is capable of giving a certain output to an input that it has never seen or heard, because it has been found with other models of similar input-output pairs.

The problem with automatic transcription is that the language is highly variable, even among people who speak the same language, and that can increase the error rate, or word error rate (WER), which is used to assess its accuracy. The particularities in the pronunciation or in the accent, as well as the tone, the volume, the speed, and the possible background noises are decisive and may increase the error rate considerably. In the case of background noise, the software can be ‘trained’ to be able to discern what is not part of human language, and filter or eliminate it. In the case of the particularities of speech, it implies a great effort on the part of the people programming the ASR system,
as they are required to collect the greatest possible variety of speech so that the system can recognise and learn from it.

Refining ASR systems is highly dependent on the quantity and quality of data that can be provided, so that errors can minimised. In this sense, progress in these types of software are largely determined by those languages that have a greater number of speakers or that are geographically widespread. Minority languages, therefore, are relegated to a secondary position, and progress in this case depends on the research groups that are interested in offering solutions and reducing the error rate of these languages. As a result, the development of this type of technology is unequal and dependant on the starting language, not because it is technically more difficult, but because the development of technology is usually determined by the market.

As has already been mentioned, ASR systems are usually limited by certain drawbacks that, in the case of sound and audio-visual documents, are usually accentuated:

- The general quality of the recordings, which can be highly variable because the origin of the documents is also variable.
- Possible background noises, which, for the same reason, are more or less present depending on where the recording was made and with which medium.
- The particularities in pronunciation and dialectal peculiarities, increased by the great variety of people who participate in sound and audio-visual documents.
- The tone of voice, volume, and speed, which are also very personal and depending on each speaker.
- Context information, which the ASR system cannot recognise, and that may be relevant to the correct interpretation of certain passages.

To tackle these problems, it is usually proposed to personalise and adapt an ASR system to the needs of the institution and its documents.

- If it is a matter of improving the quality of the audio, there are two possibilities of adapting the acoustic model: 1) modifying the recording signal in such a way that noise is reduced and compensated (Speech Enhancement); 2) training the system by introducing audios with a wide variety of noises, so that it can recognise them in other future audios and not consider them as an analysable part of speech (Multi-Condition Training).
- If it is about solving language problems, it is necessary to focus on adapting the linguistic model and training the ASR system through the use of audios with a wide variety of types of speakers showing differences in: gender, age, provenance, etc.

For the system to be accurate, it is necessary that the training is carried out through the manual transcription of a large number of recordings. In other words, many human and economic resources have to be devoted to a task that does not always gives satisfactory results. This is so because the results vary a lot depending on the type of document.
needing to be transcribed: if it is about transcribing videos and audios from a television channel, for instance, it is more likely that the customisation of the ASR system will give good results, because speakers tend to speak in a more standard language. On the other hand, if you intend to automatically transcribe documents from disciplines such as oral history (interviews or life histories), the accuracy of the ASR system will be reduced and many more manual corrections will have to be made.

In general terms, automatic transcription is still in the process of development. The various acoustic and linguistic factors prevent the error rate from being reduced to the maximum; noise, overlapping conversations, diversity in pronunciations or the variety of dialects are challenges that are still being addressed. On the other hand, there are certain elements that are present in the recordings that can only be recorded by a transcriptionist, and which are not taken into account in ASR systems. These can be aspects related to the environment in which the recording took place, comments regarding intonation, facial and body gestures or those moments in which human interpretation that is capable of contextualising is needed. Despite this, the need for these last elements is very much determined by the type of transcription that is to be carried out.

In the same way as with the manual transcription guide, each institution must assess, depending on the type of documents it holds, what are the advantages and disadvantages of automatic transcription, and if that can be really useful. In the event that a merely informative transcription is required, in which only the words spoken are recorded, the automatic transcription can be useful once the software has been properly trained. In this context, it can be used as a first approximation, and then one can correct manually any errors that might have occurred. On the other hand, it is also necessary to assess whether the efforts and resources, economic and human, required to personalise or adapt ASR systems are really worth it in each specific case. The training of the software is only possible thanks to a large amount of data that can be useful in collections with abundant hours of recording. If the institution holds only a small number of documents, manual transcription may be more appropriate.

In conclusion, the complete substitution of manual transcription for automatic transcription is still in progress. Each institution must assess, based on its reality and needs, whether it is worth spending time and resources on a type of technology that is not entirely accurate and is still under development.
Bibliography


Comparison of transcription guides and general guidelines


**Automatic Transcription**


