Spoken Corpora Good Practice Guide 2006

Olivier Baude, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, Laurence Goury, Michel Jacobson, Isabelle de Lamberterie, Christiane Marchello-Nizia, Lorenza Mondada

CORPUS ORAUX

Guide des bonnes pratiques 2006

CORPUS ORAUX

Guide des bonnes pratiques 2006

coordonné par **Olivier BAUDE**

PRESSES UNIVERSITAIRES ORLÉANS

CNRS EDITIONS

Délégation générale à la langue française et aux langues de France
6, rue des Pyramides75001 PARIS
http://www.dglflf.culture.gouv.fr

ISBN 2-271-06425-2 (CNRS ÉDITIONS)
ISBN 2-913454-30-5 (PUO)
EAN 9782271064 257 (CNRS ÉDITIONS)
EAN 9782913454 309 (PUO)

This book /work is the result of a working group brought together by Isabelle **de LAMBERTERIE**.
It was coordinated by Olivier **BAUDE**.

Olivier **BAUDE** (*DGLFLF et CORAL – Orléans University*)
Claire **BLANCHE-BENVENISTE** (*EPHE and the University of Provence*)
Marie-France **CALAS** (*DMF*)
Paul **CAPPEAU** (*Poitiers University*)
Pascal **CORDEREIX** (*BnF)*
Laurence **GOURY** (*CNRS – CELIA*)
Michel **JACOBSON** (*CNRS – LACITO*)
Isabelle **de LAMBERTERIE** (*CNRS-CECOJI*)
Christiane **MARCHELLO-NIZIA** (*CNRS-ILF and ENS-LSH-Lyon*)
Lorenza **MONDADA** (*ICAR, CNRS,Lyon2 University*)

With the collaboration of:

Gilles **ADDA** (*for the COPTE LIMSI-CNRS*), Michel **ALESSIO** (*DGLFLF*), Alain **CAROU** (*BnF*), Ibrahim **COULIBALY** (*CDF – Grenoble University*), Valérie **GAME** (*BnF*), Fabrice **MOLLO** (*CNRS-CECOJI*), Michel **RAYNAL** (*INA*), Jean **SIBILLE** (*DGLFLF*), Dominique **THERON** (*BnF*), Luc **VERRIER** (*BnF*).

Traduction : Caroline **SARRE**

LIST OF CONTRIBUTORS

OLIVIER BAUDE
Senior Lecturer in Language Sciences at Orléans University, member of the *Centre Orléanais de Recherche en Anthropologie et Linguistique (EA-3850).* Secretary of the scientific committee of the *Observatoire des pratiques linguistiques, Délégation générale à la langue française et aux langues de France.*

CLAIRE BLANCHE-BENVENISTE
Emeritus Professor, *École Pratique des Hautes Études* in Paris and Université de Provence. Researcher in the field of French linguistics: written and spoken language, syntax, morphology, collection of oral language corpora.

MARIE-FRANCE CALAS
Chief Curator of the national heritage. General Inspector of French Museums, *Direction des Musées de France.* Specialist in the field of oral documents, which is a wide pluridisciplinary field comprising the history, management, conservation and promotion of spoken language, music and environmental sound recordings, all considered today as parts of the immaterial heritage.

PASCAL CORDEREIX
Chief librarian, in charge of the oral document section at the audiovisual department of the *Bibliothèque nationale de France.* He is also the vice president of the *Association française des détenteurs de documents audiovisuels et sonores* (French association of owners of audiovisual and sound documents). Most of his work deals with the problems of sound archiving.

LAURENCE GOURY
Research fellow at the IRD *(Institut de Recherche pour le Développement),* member of the CELIA *(Centre d'Etude des Langues Indigènes d'Amérique),* field linguistics and typology (Creole languages in particular).

MICHEL JACOBSON
Computer engineer in the « *Langues et Civilisations à Tradition orale* » research team at the *Centre National de la Recherche Scientifique.* specialising in the management of oral corpora.

ISABELLE DE LAMBERTERIE
Research supervisor (*Directrice de recherche)* at the *CNRS*, in charge of the team *« Normativité et société de l'information »* at the *Centre d'études sur la coopération juridique internationale* (CECOJI – UMR 6224), member of the CNRS Committee on Ethics.

CHRISTIANE MARCHELLO-NIZIA
Professor in Language Sciences at the ENS-LSH (Lyon), Director of the *Institut de Linguistique Française (CNRS)*: historical linguistics, history of the French language, theories on the evolution of languages.

LORENZA MONDADA
Professor in Language Sciences at Lyon 2 University, member of the research team ICAR (UMR CNRS 5191). Specialist of interactional linguistics, she works on corpora of oral languages in interaction, as well as on the multimodal analysis of video corpora.

PREFACE BY XAVIER NORTH, GENERAL DELEGATE FOR THE FRENCH LANGUAGE AND FOR FRENCH LANGUAGES

Rare are the moments, in the history of science or cultural politics, where an ensemble of raw data and uncertain material becomes an object of knowledge. The publication of this guide does just that, since it offers to every researcher the tools, a guide for "good practice" which will allow him to proceed in this metamorphosis: The transformation of verbal productions into oral corpora, likely to be studied and kept and resultantly to take its place in the Nation's cultural heritage.

Language productions in their written form, being both fixed and definitive, literary works or historical documents, have always been at the heart of the politics put into place by the Minister for Culture, whether in the form of books or archives. But it is only recently that we have started to become interested in the living aspect of language, in its spontaneous springing, in its ordinary daily enunciation, and in the extraordinary variety of its parlances… For the first time, it has become possible to make real archives of the spoken word based on solid ground. Technological advances should contribute to this.

Indeed, an oral corpus is not just a simple collection of recordings of human speech, but it is a tangible object that has been "constructed": processing data (digitisation, transcription, and indexation) allows us not only to conserve it but also to give it a new status, i.e. that of research material and promotion. But this implies using the prescriptions of methodology that are coherent and easy to put into place.

Thanks to "A Guide to good practice", a new and vast domain has now become available to researchers. Through its *Observatoire des pratiques linguistiques,* the *Délégation générale à la langue française et aux langues de France* initiated this work, and then strived to gather and coordinate the various resources, both human and material, which produced this book, whether they originated from the world of research or the different horizons of the Ministry for Culture involved in this initiative.

Ensuring the development of oral corpora, their distribution and their preservation is also making the French linguistic heritage available to listen to in its diversity, richness and truth. It is also creating a precious tool of knowledge of language use which is necessary for the definition of linguistic politics as well as the politics of education and sociology.

For several months, this research brought together lawyers, linguists, librarians and computer experts all working conjointly with the common goal of making it possible to explore new areas of culture and research while respecting the law. It is the result of a common effort that we present in this book today, in the hope that, in its turn, it will generate numerous works.

PREFACE BY BERNARD MEUNIER, PRESIDENT OF THE CNRS

The spoken word and the written word. These two elements possess a powerful evocative force. We think about the way in which civilizations became structured by their oral practices and then by the creation of writings which led to a better transmition in space and time of the words spoken by one or another of us.

As a researcher looking at the respective roles of the spoken and written word in the dissemination of scientific knowledge, I don't fail to remember that, far beyond the essential role of the written word, delivering an oral presentation in front of our peers or a wide audience is always essential for circulating, convincing or sharing ideas. The spoken word maintains a power of conviction, allowing the largest number of people to be reached provided that it can be recorded and transmitted with the help of current audio-visual means.

The collection and use of oral corpora should be done in compliance with a code of "good practices", in the same way as it is done for the collection of written corpora. We all know how a sentence which has been taken out of context and broadcast without reserve can become dangerous for the person who produced it, for a group of people or a community.

The authors of this outstanding book have examined in depth all the legal aspects involved in the collection and use of written corpora. I hope that this book will get the best possible circulation among the actors and users of oral corpora, something that we all are at one time or another.

PREFACE BY JEAN-NOËL JEANNENEY, PRESIDENT OF THE *BIBLIOTHÈQUE NATIONALE DE FRANCE*

The Bibliothèque nationale de France is happy to have contributed to the elaboration of this *Guide*. Indeed, it has had a long-lasting and close link with spoken languages their preservation and their distribution. Its audiovisual department ensues from Ferdinand Brunot's Spoken Archives (*Archives de la Parole)*, created as early as 1911. From then on, our institution's constant concern has been to ensure the best conditions possible for recording and preserving oral expressions of every type, as well as their distribution to as large an audience as possible.

Today digital technology is reinforcing this historical and scientific link. In terms of conservation, an ambitious plan to digitise our collections has been initiated, from which sound and audiovisual documents benefit in particular. Furthermore, the distribution of these precious resources within our walls and from a distance is further enhanced by the rapid expansion of our on-line digital library, "Gallica", which allows every internet user, wherever they may be based or whatever the purpose of their research or interest may be, to have access to fundamental sources of knowledge.

The fruit of a faithful collaboration, this *Guide* shows how complementary knowledge is between linguists, lawyers, librarians, computer scientists, sound and picture technicians: I am delighted that the *Bibliothèque nationale de France* has been able to contribute to this innovative and fruitful undertaking.

Words highlighted by an assterisk can be found in the legal glossary at the back of the book.

1 PRESENTATION

1.1 OBJECTIVES

There is currently a vast amount of fundamental or applied research, which is based on the exploitation of oral corpora (organized recorded collections of oral and multimodal language productions). Created as a result of linguists becoming aware of the importance to ensure the durability of sources and a diversified access to the oral documents they produce, this *Guide to good practice* mainly deals with "oral corpora", created for and used by linguists. But the questions raised by the creation and documentary exploitation of these corpora can be found in numerous disciplines: ethnology, anthropology, sociology, psychology, demography, oral history notably use oral surveys, testimonies, interviews, life stories. Based on a linguistic approach, this *Guide* also touches on the preoccupations of other researchers who use oral corpora (for example in the field of speech synthesis and recognition), even if their specific needs aren't consistently dealt with in the present document.

The *Guide* that we have put together for you primarily aims at providing the necessary *information* for making corpora of oral or multi-modal data, and at offering useful *suggestions* concerning the judicial as well as the material aspects involved just as much with the collection, structuring and transcription of data, as with the exploitation, communication and preservation of the data.

The second objective of this guide is to help researchers who are making or contributing to oral corpora to *anticipate* certain "delayed difficulties" which could seriously jeopardize the exploitation and the future of their corpus. Certain choices made at the beginning, certain missing elements can turn out to be important at later stages of the process once it is too late to make any changes.

The third objective is to encourage the definition of *common practices* in order to fulfill the current requirements of conservation and interoperability of corpora, of evaluation and ethics as much in the constitution as in the use of the data.

1.2 CONTEXT OF ELABORATION

The scientific committee of the *Observatoire des pratiques linguistiques (Délégation générale à la langue française et aux langues de France)* has wished to strongly encourage all measures of preservation, constitution and promotion of oral and multimodal corpora for the following reasons:

- To allow for the maintenance of a rich national heritage about language uses in France;
- To help develop large reference corpora, for research, teaching, the language industries and also for the linguistic heritage;
- To help in the development of computer tools for processing, enriching and promoting corpora;
- To encourage the availability and accessibility of corpora;

1.3 LEGAL ASPECTS

It quickly became apparent that the judicial aspects linked to the consitution and use of oral corpora represented a recurrent and major obstacle.

These legal aspects mainly concern questions of moral and property rights, and data ownership, which arise in each of the four main stages of corpus work:

- Collecting data and recording it (the right to one's image and voice, interview situations, authorisations…);
- The use and computerized exploitation of data (archiving, use of database for research, teaching, engineering…);
- The distribution and publication of data (rights, the right of quotation, online publication of data…);
- The conservation of data.

In view of the fact so many domains were involved, the DGLFLF initiated the creation of a committee made up of experts for a diversity of fields. This committee set up a working group whose objective was to help research groups to standardise their ways of collecting and exploiting data in compliance with the law whilst taking into account the many constraints inherent to the research. This guide is the result of fifteen months working in this group.

Of course, this working group had to include legal experts in research law, but that wasn't enough. We needed collaborators with specialized skills in collecting, using and preserving corpora, hence why the working group took on linguists working in the field of corpus linguistics and oral data, representatives of the most important organizations for the preservation of the national heritage *(INA, INSI, BnF)*, and computer scientists specialized in corpus management, alongside the legal experts.

To achieve its goal this working group gave itself the following objectives:

- To look at current practices and as a matter of priority to define the methodological constraints and theories bound to research;
- To circulate a synthetic document about existing legislation;
- To make recommendations;
- And if necessary, where there is a gap or something unclear in the law, to formulate suggestions for the creation of legal norms and rules (notably those in Europe).

In order to do this, it was first necessary to:

- Review the judicial domains concerned;
- Identify and quantify the risks;
- Work out the existing responses;
- And then to formulate the responses in the form of a series of recommendations for good practices (both legal and ethical).

For this purpose, the group decided to work closely together with many control teams who were collecting or had previously collected oral or audiovisual data. In this way, the goal was to come up with a "typology of situations", and to examine all the practises and solutions already being used in France as well as elsewhere.

1.4 OTHER ASPECTS

Whilst working towards these goals, the group realised that making a simple list of recommendations or solutions of a legal nature wouldn't be enough to effectively overcome the difficulties that were encountered.

It actually became clear that the difficulties or the solutions were linked to the practices used when collecting or using the data and that certain solutions had to be found through examining technical measures which had an impact on the data itself (anonymisation or blurring). It also appeared that solving a legal problem at one particular stage rather than another did matter. In short, offering solutions for legal questions meant examining the very process of collecting, transcribing, circulating or using this type of data.

Finally, over and above respecting the legal rights of the people who had been recorded, the question of "the right of ownership" of this type of data arose: What rights do the people who collect this data have? Who is legally responsible, who has the right to disseminate it and in what form? As we can see, the legal aspects linked to scientific ownership or penal responsibility were also inseperable from the collection process and use of the data.

With this in mind, would it not be better to enlarge the field of the proposed "Guide" and deal not only with the legal practices but also with all the practices involved in this type of corpus? This is the choice that we made because it allowed all aspects to be intertwined as they are in reality.

1.5 METHODOLOGY

The methodology on which this group has agreed has the following characteristics:

- The conviction that you cannot let people believe that there are ready made answers to every type of situation;
- The eagerness not to hold back researchers (by prohibiting certain practices for example);
- The respect of the researcher's methodology and of the constraints linked with observation (researchers want to record situations that should not be altered by technical or legal constraints).
- The need to elaborate and compile this guide by bringing together the skills required at the different stages (linguists, lawyers, librarians);
- The display of a procedure founded on the respect of the law and ethics;
- The need to provide through this *Guide* a tool for risk assessment (pinpointing and also evaluating risks).

1.6 THE FRENCH JUDICIAL FRAMEWORK

A large number of questions and solutions revolve around the notion of *consent* of the interviewees but also around the responsibility of the *owning* institutions. It is

certainly a nodal point, but it is far from being the only thing in question and besides the answers to such a question proved to be complex.

Current practices for gaining consent and authorisation are very varied. No specific norms exist and there are multiple difficulties.

In the first instance, consent should be *informed* (framework, objectives, "risks" for the interviewee).

But it would appear that gaining consent can sometimes hinder the study (the observers's paradox) in formalizing a situation when what is desirable is to obtain » natural » data that is as close as possible to ordinary conversation.

In this way, for example, one practice which proved interesting and efficient (in addition to collecting authorisation) consists in handing out to the interviewees a document explaining the framework, the objectives, the risks, the accessibilty, and the details allowing the references of publications and results to be subsequently found.

The difficulty also comes from a *contradiction* between the need to specify the objectives of the study in order to 'inform' the consent, and the impossibility to anticipate all the objectives and *the future possiblities for the use of the data, considering the current concern with coming up with maximimum* interoperability.

Finally, it should be noted that certain spoken cultures (and not just on the other side of the world) don't offer the possibility to propose and keep a traceable written consent.

All other questions of a legal nature also have the same complexity: anonymisation, encryption, blurring, defining responsibilites, depositing, papers, etc., all the necessary practices linked to the constitution and existence of an oral corpus. None of these aspects rests on one specific practice which is clearly defined and accepted everywhere.

Each of these steps is closely linked to technical choices, to social or scientific practices, all these elements being very difficult to dissociate.

This is why the choice of the working group was to offer a Guide which would not only be a "judicial memento", but also a practical and reliable tool covering all aspects of the process.

1.7 A "GUIDE TO GOOD PRACTICE"?

Taking into account the existing legal framework in France (and more generally in different parts of Europe), this guide relies on the questions asked by researchers who participated in its elaboration. They tried to comprehend the foundations of the judicial rules to abide by and the stakes linked to the respect of these rules and to their implementation. *A dynamic vision of legal regulations* has therefore served as a framework for this guide, through the procedure used by researchers. The authors of the guide, involved themselves in the fields of research dealt with, were concerned with proposing practices and uses which respected the existing laws. For this, the research process should consist in knowing the existence of the laws and of the

constraints which surround them. Then, consequences of these constraints need to be identified as much in the stage of data collection as in that of data promotion.

To present such a procedure in a credible and rigourous way, it first has to be put in its context whether it be scientific, political, judicial or institutional. Throughout, the suggested uses and practices will be "clarified" by the context, with a view to better understanding what the implications of respecting or not these uses and practices are.

1.8 SOME FREQUENTLY ASKED QUESTIONS

The first objective of this guide is to provide information and elements to answer the questions asked by all researchers or people in charge of collecting, exploiting, conserving and circulating corpora.

To reach this objective, the guide includes numerous cross-references which make up many possible reading paths. The following questions are representative of the queries which traditionally arise at the beginning of a research project and in this way suggest a first example of reading paths.

FREQUENTLY ASKED QUESTIONS

1. *What types of written permission do I need to get the speakers I am recording to sign in order to be able to exploit the corpus and be able to:*
   a. quote from it in a university paper;
   b. quote from it in an article published in a scientific journal;
   c. quote from it in a book published for commercial use;
   d. make it available on an internet site;
   e. disseminate it on CD.

   Are these different types of exploitation subject to the same rules?

   Answer: Questions a, b and c fall within the province of the 'right to quote' (see legal documentation *Right to quote*). Elements of the answers to questions d and e are presented in chapters 2.1.5, 2.3 and 3.5. (see legal documentation *Consent* and *examples of written permission forms*).
2. *I have made a recording of people that I know well.*

   a. Under what conditons can I exploit this? (exploitation is understood in the same way as in question 1)

   b. Can they go back on their consent?

   Answer: All of chapter 3.4. is a reflection on the conditions of collecting data and aims to make people aware of the numerous problems which can arise during the course of collecting data. Being familiar with the people recorded doesn't lighten the legal requirements (that are owed to them), it's quite the opposite (it raises questions of trust which can give rise to rather complex situations). See legal documentation *Consent*.
3. *When I record children,*

   a. who can give their consent?

   b. when a child comes of age can they go back on their consent?

   c. if the recording takes place on school premises are particular authorisations needed?

   Answer: this case comes under the more general one about people for whom we need to ask for additional permission from people in charge and guardians (in this case parents and the school institution) (see chap. 3.3.2 participants section).

4. *In the case of a project carried out by a research team,*
   a. Who is considered as the author of the corpus?
   b. What right(s) does the work give to the researcher?
   Answer: See chap. 2.3 (copyright) and the Copyright section in the legal documentation.
5. *Who is considered as "responsible" for the distribution and processing of a corpus?*
   Answer: See chap. 2.3 and the Processing Manager section in the legal documentation.
6. *If I cover up people's proper names, is that enough for me to be allowed to use a transcript freely?*
   Answer: Anonymisation doesn't simply consist of erasing the proper names of the people involved. See chap. 3.5 Anonymisation and the *Personal data* and *Anonymisation* sections in the legal documentation.
7. *Under what conditions am I allowed to archive my corpus in the form of computer files?*
   Answer: Is is necessary to take the judicial aspects into account (privacy protection, *informatique* and *liberté* Act, asking for permission, see *Personal Data*, *Anonymisation* and *Processing Manager* sections in the legal documentation, and the technical aspects of conservation (see techincal documentation).
8. *If the people I have recorded (in the media or in private) have died, am I free to exploit these recordings?*
   Answer: Copyright exists up to seventy years after someone has died! As far a protecting someone's privacy is concerned, this cannot be invoked after someone has died unless the person has prohibited its diffusion during their life. In addition, the family members of the deceased can invoke their own personal right to protect their privacy. See chap. 2.3.1 and *Personal Data* and *Anonymisation* sections in the legal documentation.
9. *I discover recordings in a cupboard. I would like to exploit them. I have no record of who made the recordings or who was recorded.*
   a. Am I allowed to use these documents?
   b. What precautions should I take (what guarantees)?
   Answer: We cannot insist enough on being prudent and making the necessary research in order to identify the documents, including for purposes of scientific rigour. See chap. 2.3 and chap. 3.5.
10. *I am recording a radio programme (or television programme).*
    a. Am I allowed to freely use the transcript?
    b. Can I use the soundtrack?
    c. In terms of permission, is there a difference between public radio broadcasts and private ones?
    d. Is there a difference between recording famous people and people who are not famous? (people who are witnesses, people expressing themselves freely on air, listeners who are asking questions, etc.)?
    e. Are the rights of exploitation different if I buy a videotape, a DVD or a CD of the programme or if I record the programme myself while it is being broadcast?
    Answer: Radio programmes are protected whether they are public or private. See 3.3.1 about re-using media recordings and in particular news items.
11. *I would like to make a corpus of authentic data. What precautions should I take?*
    Answer: See chap. 3. which offers a reflection combining a methodology for field research and ethical and judicial problems that can be encountered in the process.

    And many other questions…

2 CONTEXT

SCIENTIFIC, POLITICAL, JUDICIAL AND INSTITUTIONAL

When we think about context, we think about putting things into perpective. Such is the objective of this chapter which presents what the linguist's scientific work on oral data is. Putting things into perpective also means considering the *political* and *judicial* aspects. The institutional context is of growing importance as the traceability and the continuation of research need to be ensured on a long-term basis. In guaranteeing the durability of the data which has served as the basis of a researcher's work, as well as that of the results that have been obtained, both researchers and institutions participate in the development of knowledge in the immediate future or one more distant.

2.1 LINGUISTICS AND ORAL CORPORA

For about twenty years, studies on spoken language corpora have completely revitalized language sciences. To be assured of this, all that needs to be done is to refer to recent bibliographies in France and abroad (for example the *Revue Française de Linguistique Appliquée* or the *Recherches sur le Français Parlé)*. These studies have allowed new hypotheses to be formulated about the normal and pathological functioning of language and have become an essential component of the discussions between linguists and computer scientists. In France, up until more recently, interest in spoken languages was essentially shown in domains where it held a central place "by default": in the first instance, studies about the sound components of language (phonetics, phonology and prosody), the language of young children, or everything that was classified as "languages without a written tradition", in France regional languages and local dialects, and outside France all languages referred to as "exotic". We can add to these a few isolated cases in the 1950s and 1960s which attempted to bring together models of spoken French in order to teach French as a foreign language, notably the *Français Fondamental* and the *Corpus d'Orléans*.

Descriptions of the French language, in particular in grammar books, remained based on written language data, be it literary or not, the "grapho-lects", as Ong used to call them (1988), or on data provided by the institution. This distancing from spoken language data has brought about two major consequences: on the one hand, the very negative image that the French have of their own language and, on the other hand, a considerable influence on the most common linguistic theories. New data brought to light by spoken language corpora hasn't had an impact yet on the image that the general public has of the language, but it has already contributed a lot to the evolution of theories amongst specialists.

New domains, which were first initiated in the 1970s in Great Britain (Sinclair & Coulthard, 1975 for the school of Birmingham), have emerged in France, as the development of interation models and conversation analysis (the original paper by Sacks, Schegloff, Jefferson in the USA in 1974, papers by Bange and by Quéré in France, in 1983 and 1984).

Spoken language data collected before the computer age cannot be compared to what we presently call "spoken language corpora". Each of these old collections, distributed according to research projects, followed its own rules in terms of choice, recording, transcription and preservation, in such a way that it is now difficult to have access to them and to pool them (recordings of the *Français Fondamental* have been erased, and those of the *Corpus d'Orléans* now have to be transcribed again). None of them could ever become very big (generally there are only a few hours of recordings) and searching for information within the data could only be done manually. From the 1980s and 1990s, advances in computer technology allowed the creation of modern corpora of spoken language throughout the world, in the first instance in Anglo-Saxon countries. A new discipline was born, that of corpus linguistics (G. Kennedy made a description of it in 1998 for English and Habert and his collaborators for French in 1997), which arouses interest among academics and in the language industries and which, as a *Language Resource*, now makes up part of the national linguistic heritage. France, which was ahead in the perfecting of written language corpora (in particular for FRANTEXT which is the source of *Trésor de la Langue Française*), has fallen behind in consituting spoken language corpora.

Numerous types of spoken language corpoa exist, intended for diverse objectives and in various disciplines. These always consist of recorded sound data, possibly with added visual data (recorded with a camcorder or on television), and nearly always complemented by transcripts and computerized processing. Without trying to cover everything here, we will deal with four of their aspects: the types of data and speakers, the size of corpora, the transcripts and a brief overview of the possible exploitations and results.

TYPE OF DATA AND SPEAKERS

Certain data is "contrived". For instance, speakers are invited to come and work, acting as Guinea pigs, with teams of phoneticians in order to provide different types of pronunciation and intonation under very good recording conditions. They are made to pronounce words or lists of words, numbers or lists of numbers, or to read texts or passages from texts. These documents serve for different exploitations, either to record and study pronunciations themselves, as is done by J. Durand, B. Laks and Ch. Lyche to study contemporary French pronunciation (PFC project), either to test for a particular language behaviour (as is done in hospital units where cases of aphasia are studied), or to carry out analyses which will be useful in speech synthesis, in text-to-speech systems, or in man-machine interactions (this is the objective of *SpeechDat Exchange*, which stores from 500 to 5 000 telephones recordings in 28 languages). In all these situations, speakers generally know that they are being recorded and they have an idea, be it precise or approximate, of what their production will be used for.

Another type of data is "continuous speech data", with varying degrees of spontaneity (this notion has been specially studied in an issue of the *Revue Française de Linguistique Appliquée*). Some data is collected in situations which have not been staged by researchers and which would have taken place without researchers having

anything to do with them. Other data is more or less "contrived" and orchestrated and organised by researchers. The ideal circumstance – that of total spontaneity – would be to record speakers without them suspecting anything (hidden microphones, pirate recordings), and to tell them afterwards or not that they have been recorded, the objective being to capture their language "freely", with as little control as possible. Legal requirements partly prevent this procedure. In any case, the presence of the interviewer and his equipment hinders this freedom (it's "the observer's paradox", an issue which was made popular by W. Labov). In practice various degrees of constraint can be identified, according to whether it is private or public speech, in front of familiar people or in front of strangers, with various forms of complicity or not, whether it relates to a face to face conversation or speech transmitted through a medium such as the telephone, an answering machine, the radio, television or other technical devices. A good ethnographic approach (repeated recordings) allows the problem of adjusting the microphone's sensitivity to be solved. But this requires researchers to spend a lot of time on it during the data collection stage.

Modern corpora are seldom composed of "run-of-the-mill" speech. The choice of speakers and situations is generally decided on according to the objectives set at the beginning of the research. Researchers propose collecting conversations between adults, negociations in professional contexts, interviews (prepared or not), talks in public organizations, electoral speeches, explanations between public services and their users, public lessons, sermons, political speeches, conferences (specialised or popularized), historical testimonies, narrations of trivial events, life stories (told by individuals, groups, group representatives, spokesmen), dialogues between mothers and young children, children's conversations recorded within a school context or outside (during games or conversations, when answering tests or otherwise, in a school environment or outside, while playing freely or playing directed games, with parody and role-play), sick people in hospitals, etc. Here is an example of these: the CLAPI *(Corpus de Langue Parlée en Interaction)* database is currently being set up in Lyon (ICAR research team) in order to bring together corpora of "speech in interaction" as diverse as possible, in situations unprompted by researchers: conversations at mealtimes, dialogues between solicitors, phone calls to emergency social services and in therapeutic consultations, etc. This database comprises 300 hours of audio and video recordings, transcripts and "metadata" giving details about the speakers.

Numerous disciplines are trying to study the correlations between spoken language productions and other phenomena. Correlations between language and socio-economic parameters were at the root of sociolinguistic research. In the USA, W. Labov had carried out famous studies about the black population in eastern American cities, by interviewing people in their homes, in the streets or in department stores (often in bad recording conditions). Studies about language development are done according to the age of the children, the activities observed, the instructions provided and family background. Taking into account different "genres" (as D. Biber defines them for English) enables correlations to be made with the place where the speech was recorded, the topics dealt with, the type of

speaker and of exchange (monologues, dialogues, group conversations). To be able to measure these correlations, the content and size of corpora are generally defined from the outset: this many types of situations and speakers (as was done by the Sankoff-Cedergren team in the 1970s to study social variation in the city of Montréal). In other cases, researchers define representative sub-corpora within existing corpora adapted to their study (which is what D. Biber suggested to do in order to sample the large British National Corpus). Here, we talk about "balanced" and "sampled" corpora.

Linguists, for their part, have often collected "open" corpora, which they modify as their work progresses, without beforehand defining a predetermined object of research, because they are certain they will discover new phenomena which can't be predicted from the start: the distribution of formal and informal language, the relationships between grammar and lexis, the links between the degrees of complexity of the syntax and the type of speaking situations, the use of oral morphology, the role of contexts in the construction of the meaning of utterances, the role of prosody in textual structure etc.

The technical quality of recordings obviously depends on the technical equipment that is used, but it also depends on the types of situations and speakers chosen (noisy places, an important number of speakers, speakers with a speech impediment). These various situations also have an impact on the speakers' consent: it is easier to get permission to record a public speech than a private conversation, the words spoken by someone who is self-confident than those of someone who is anxious and sensitive to what has been called "linguistic insecurity".

In all cases, it is somewhat difficult to explain the need for making recordings by the will to study the language. If this objective is presented, French speakers undoubtedly have the impression that they speak badly and that the study will ridicule them. Very few of them are relaxed about this. Nearly all researchers have come up with strategies to tackle the problem of bias by saying that they are interested in the content of what is being said, in the testimonies, in the explanations, in the speakers' particular knowledge (which could be knowledge about the language, in the case of research made about regional languages). In all the works done on speech in interaction, things are a little different: researchers are able to say that they are precisely interested in the way in which the participants interact with each other, in their coordination, in the outstandingly accurate adjustments they resort to through speech, gestures, facial expressions, looks and body language (multimodal resources which are difficult to control as a whole, even when speakers pay attention to them).

SIZE

The ideal size of a corpus and of the units which make it up varies depending on the study that is going to be made. Studies in phonetics, phonology and prosody can give good results with sound units of a fairly limited length. But if we want to study correlations between the language and other phenomena, or if we want to study lexis, much more developed units are needed, in a bigger quantity and in more

diverse domains. The size of spoken language corpora and of the elements they are composed of can be measured with two types of units. Units of time are used when the main focus of interest is recording sound, setting the transcripts aside. For example, fourteen-to-thirty-second long corpus elements are considered as very small (fourteen seconds being the average for a news item on the radio). However, even smaller sub-units are taken into account when studying speech overlaps between speakers or when measuring pauses (up to one tenth of a second). Small units are used for example by telephone companies which are currently setting up European directory inquiry services in every European language (EuroSpeech 2003). Ten-minute long elements fall into the big element category and sixty-to-ninety-minute long elements are considered as very big. By adding up all these elements, we can say, for example, that we have at our disposal 100 to 500 hours worth of recordings.

However, these measurements aren't very reliable for large corpus components, because the density of the recordings depends on the speakers' delivery. In French, people who speak slowly are considered to utter 110 words per minute while those who speak very quickly say an estimated number of 350 words per minute (in certain cases of aphasia, and under the influence of neuroleptics, the delivery drops to below 100 words per minute which is very difficult for the listener. Above 350 words per minute, listening and transcribing become very difficult). Density therefore varies from one to three, which is quite considerable. In relation to the two afore-mentioned deliveries, one hour of recording can be equivalent to 6 600 or to 21 000 words. In is in our interest to assess large corpora according to the number of words written in the transcripts. Large spoken language corpora collected in the world today have a size in the region of ten million transcribed words for British or American English. Unfortunately current spoken French corpora only have as many as a million words. With such a limited size, it is neither possible to do lexical search nor to draw reliable statistics on language uses.

TRANSCRIPTS

Transcripts of spoken language in use today are so different from one another that it is difficult to classify them under the same heading. In certain cases, when we only consider the content of the recordings and we make free changes to their form, it would be more appropriate to talk about *transpositions* or *adaptated versions*. This is what journalists do when they report what interviewees have said by summarizing their words and by generally using more normative forms (in the case of a politician saying *ça, je sais pas, pour pas que…,* they use *cela, je ne sais pas, pour que …ne pas…*). Historians and Sociologists sometimes use similar practices, when their main focus of interest is informative content: they select specific data, cut out passages which are of no interest to them and discard the specificities of oral language which are seen as a potential nuisance, such as repetitions, hesitations or repairs. In certain specific fields of activity, such as transcribing parlimentary debates, these tasks have even been codified by defining different levels of adaptation.

When taking an interest in the language itself, the choice of the type of transcription depends on the objective of the study (European and international projects have

come up with intructions for editing corpora) and, as E. Ochs pointed out in 1976, transcribing always rests on a theory. Certain studies require phonetic or phonological transcripts. The Unicode standard, synchronised with the ISO-10646 norm, already contains more than 96 000 characters in version 4.0, in particular those of the *International Phonetic Alphabet*. This is necessary for all the works involved in pronunciation, but also for all the cases where it is difficult to make out regular morphemes which could be transcribed with standard spelling: the language spoken by very small children (the CHILDES international model), the language spoken by foreigners in the process of learning it, the transcription of certain regional accents, the transcription of certain forms of aphasia such as jargons (Abou-Haidar 2002). These transcriptions, which can only be made for small parts of corpora, are often supplemented with line by line translations. The representation of prosody requires specific models whose development has been taken further in recent techniques (Martin 1987). Video recordings require special transcription techniques which can be more or less elaborate (Van der Straten 1998, Mondada 2006).

As far as large spoken language corpora are concerned, they are transcribed using standard orthography, in order to make them easy to read. Several options follow from this choice: standard ortography with or without adaptations, with or without punctuation, with or without indications for pauses, for lenghtening, for rythm, for stress, for hesitations, for coughs, for laughter, for gestures, etc. These points have been at the heart of many debates in order to define optimum conditions for transcription adapted to the objectives of the research. For instance, linguists studying syntactic units in spoken language are generally cautious about punctuation, as it imposes delimitations which are distinctive characteristics of written language and as it often turns out to be misleading when it is added before the texts are thoroughly analysed. But texts without punctuation irritate computer scientists whose automatic analyses require punctuation marks. Negociations between linguists and computer scientists sometimes take place (ICOR in the ICAR research team) in order to come up with transcription conventions which take these problems and international standards into account (GAT, TEI, Du Bois, Jefferson).

The transcriptions used by linguists meticulously include all the specificities of oral production: repetitions, hesitations, beginnings of words, repairs. They require transcribers to make sure that they don't show any signs of their own interpretation in the transcript. (adding or removing "ne" for negation in French, for example, or reconstructing a passage of the text following accepted stereotypes). This concern with detail demands specific training and practice for transcribers. It's a long and demanding task which is also full of pitfalls (Leech 1991). According to accepted estimations, a minimum of thirty minutes work is necessary to transcribe a one minute recording (the creators of the Dutch corpus consider that it amounts to one euro per written word!). Because of their very accuracy, transcriptions of spoken language displease the uninitiated: they see in them a number of "mistakes in French", of repetitions, of paddings. Showing a transcription of their speech to an uninitiated informant is often a cause for rejection. It is not a very good way of getting their permission to transcribe and publish the result of the research.

Computer tools have transformed the work of a transcriber, on the one hand through the help that it has brought them, and on the other hand, through the new constraints that it has introduced. Transcription tools (Anvil, Clan, Elan, Ite, Praat, Transcriber…) make manipulations easier and allow the portions of recording under study to be easily listened to repeatedly. The technique of *synchronised corpora* allows portions of written text to be read on a screen while listening to the same portions in their audio form (*Speech Communication* 33, a special issue about annotation and anaylitical tools for corpora). New requirements affect computerized annotation: morpho-syntatic labeling of all elements in the text, branching, metadata (about the circumstances in which the recording took place, the situations and the speakers). Several classification and coding systems allow the necessary lemmatisations and concordancers to be done in order to be able to make queries on the whole corpus. A great debate started in the years 2000 about the degree of sophistication of the annotation which seemed necessary (Sinclair, Teubert). Standardisation is now done on a European level (SpeechDat Exchange Format).

SPEECH PROCESSING

Contrary to many other areas of research about speech, automatic speech transcription, which is done from a continuous acoustic flux, requires the modelling of all the phenomena observed in the sound signal. It is therefore necessary to establish models, beyond the words with which a phonological representation is associated in the pronunciation dictionary, of extra-lexical phenomena: breathing, hesitations, word fragments etc.

According to the types of documents processed, automatic recognition systems obtain very different rates of error[1]. However, if there is a discrepancy between the models (basically, the knowledge) of the system and the corpora to be transcribed, the rates of error can quickly go up. In order to achieve the best results possible, transcription systems need to be adapted to the corpora to be transcribed.

Current research shows that automatic transcription is becoming a precious tool to help with the transcription and annotation of corpora. For example in Barras *et al.* (2004) the usefulness of automatic transcription is shown for the semi-manual generation of accurate acoustic transcriptions (that is to say comprising not only all the orthographical words but also the "disfluencies" and other extra-lexical events). Research in process also shows that automatic speech transcription can become a precise tool to explore and analyse corpora and quantify linguistic phenomena. More generally, we can imagine that in the future the opposition between the vision of linguists and that of computer scientists will have to abate. In this respect, the emergence of oral corpus linguistics as an area of research should rest on the training of linguists skilled in computer science and of computer scientists skilled in linguistics.

[1] Works done by the LIMSI research team (Barras 2004) show results ranging from 10 % to 30 % of word errors with systems which were optimized for a given task.

EXPLOITATION AND RESULTS

Current large corpora of spoken language are expensive. Certain corpora, notably in engineering, are exploited in association with industrial partners: man-machine interaction, speech recognition and synthesis, telephone calls etc. (organisations such as ELRA/ ELDA are specialised in the distribution of corpora and resources available in this area).

In the first instance, large corpora serve as a source of general documentation about the National language. Large *corpora of reference,* which are sampled taking into account the regions and the socio-economic and cultural data, allow large-scale linguistic policies to be defined. For example, the corpus of reference for spoken Portuguese, which contains recordings made in Portugal, in Africa, in Brazil and in Asia, allows the differences according to world geography to be assessed and it enables certain uses in school practices and even governmental decisions to be based on this examination. The *British National Corpus* served as a basis in the constitution of a famous grammar reference, the *Longman Grammar of Spoken and Written English*, created on very new foundations. A great deal of publications have been made in English, making the most of these materials. In this way, Collins publishers have used English corpora for the publication of a large number of books about didactics used in the teaching of English as a mother tongue and as a foreign language. Documentation about spoken language is sometimes the starting point to launch new activites: some spoken language corpora have served as foundations to diffuse languages which are hardly written (or not at all), as was done for the Maori language which served as a model to develop radio and television broadcasts (Kennedy 1998: 72).

The comparison between spoken languages belonging to the same linguistic group allows an *in vivo* evaluation of the similarities and differences within a large linguistic area.

An important exploitation is the one offered by multi-lingual corpora (also called parallel or aligned corpora), which are used by translators, for language teaching and contrastive studies. There are also multi-lingual corpora for written language:

- English/French at the University of Lancaster, the University of Oslo, in Mannheim, at the University of Gand in Belgium (Contragram, bank.ugent.be/contragram/newslet.html), and at the University of Montréal;
- French/ English/Dutch at the University of Courtrai,
- French/English/Spanish at the University of Pennsylvania.

A recent study, based on the recordings and transcriptions of four Romance languages (Italian, French, Portugese, Spanish), allows a comparison of prosody to be made (intonation, stress, rhythm), taking into account different situations and media (C-ORAL-ROM, Cresti & Moneglia).

This is how large spoken language corpora have renewed a number of linguistic problems. The data provided by these large collections has laid the foundations of new disciplines, such as conversation analysis and interaction analysis, analysis of negociations and codes of politeness. Research in pragmatics relies greatly on this

data. Some knowledge has evolved a lot, as for example studies focusing on oral production and how spoken language is perceived and, as a consequence, on the fragile nature of linguistic intuition (Blanche-Benveniste 1997). The degree of ordered and systematic organization in interactions has been shown. This has been used to question certain basic units such as *sentences*, and to introduce new ones like *macro-syntactic* units, now used by several teams of linguists (Blanche-Benveniste *et al.*, 1999, Scarano 2003, Nolke 2002). Intonation study has been taken very seriously into account in the delimitation of macro-syntactic units (Cresti & Moneglia 2005, Couper-Kuhlen & Selting, 1996). In interactions, it has been shown that many interwoven levels of organization exist (Turn-Constructional Units, Selting 1995, 1998, 2000, Auer *et al.* 1999, Ochs, Thompson & Schegloff, 1996). In different languages, the role of the characteristics of oral productions, i.e. discourse particles, repetitions, hesitations or "repairs", which are currently of interest to neurosciences, has been determined. Perspectives on the history of languages have even been modified, in the sense that we can now study the influence that different speaking situations have on the type of grammar used (Biber 1987). For example, it is possible to show that, in French, explanation and argumentation talks show highly embedded syntactic uses, which isn't the case in conversations, or that accident reports have complex chronological organisations. We know that themes reputed to be "sublime" (discussions about morality, religion, death) entail characteristics of "ceremonial language": for example, in French, overuse of links and of the negation *ne,* and even sometimes unexpected uses of past historic. Large corpora allow certain grammaticalization processes in development to be followed. They demonstrate the numerical importance of parenthetical utterances, focalizations and thematizations. They make it necessary to consider that fossilized phrases take up a very important place in relation to the free production of utterances, so much that the link between grammar and vocabulary now appears more clearly than before, many grammatical turns of phrase being used by speakers only for a small list of lexical words. The conclusion that has to be drawn from this is that, when we speak, we don't choose a "word" but preconstructed phrases (Sinclair, 1991).

This obviously questions linguistic theories which aimed at isolating syntax as an independent component of language.

These large corpora, when they exist, are extremely helpful: they serve as databases for all the comparisons concerning the language: to evaluate the language of children at different stages of its acquistion, to support diagnoses in language pathologogies, to evaluate the degree of profiency in the acquisition of a mother tongue and a foreign language, to assess the effects of group and professional languages (Gadet), to study forms of coordination in a team or in a group, to understand the specific characteristics of different types of activity and of the contributions which are appropriate to the activities in different institutional contexts, or to know the impact of regional influences. For example, before judging that a turn of phrase is characteristic of child speech at a particular age or of a particular origin, it is necessary to use a comparison database to find out whether the turn of phrase is characteristic or not (the most common errors made in the use of French relative

pronouns *dont* and *lequel*, first degree, are most commonly made by the most educated adults, and for quite a long time as far as we are able to judge).

CORPORA OF LANGUAGES WITH AN ORAL TRADITION

The problems encountered at the stages of constitution, exploitation, circulation and preservation of oral corpora in so-called societies "with an oral tradition", or "ethnic", or "exotic", are partly similar to those encountered while putting together large Occidental language corpora. The precautions to take (as they are recommended in the guide) to respect people are then to be adapted to the context in which field work is taking place.

In a society with an oral tradition, the permission obtained after informing the participants (on the model of the "informed consent" described in this guide and which, again, will have to be adapted to the situation) can, in certain cases, not be valid unless it is oral and given by the person who is entitled to do this (just as in a survey situation in a medical context where permission is only valid if it has been given by the Medical Association). In addition, informing the speaker isn't problem-free in societies where research itself and the objectives of making a corpus and of circulating it via networks (research papers, the internet) don't correspond to anything concrete.

Researchers also need to inquire about the law in force in the country where they are going to work. For example, French law doesn't recognise intellectual property rights or copyrights in the case of collecting tales or myths considered to be part of the national heritage and belonging to the public domain. However, in many African countries, copyright offices have been created to protect this type of production and the authors. Besides, certain communities don't acknowledge the national Law of the state they live in. This is the case, for instance, in certain Amerindian communities in Guyana who live in compliance with a collective law and not a private one (Tiouka 2005, for a reflexion on the integration of customary law into French and European Law). Certain authorisations will only be valid in these communities if they respect the customary law and, although researchers feel protected in respecting the national Law, they can find themselves in conflict with the customary authorities and be denied access to the field.

In most cases, the exploitation of the corpus requires the intervention of several people: the transcriber, who can be the speaker in the recording, but not always; the translator (id). These people's rights on the corpus once again are to be defined according to many parameters: national Law if it makes sense for the community or customary law.

The collection of oral corpora in certain societies differs from that of large corpora of national languages on two points:

1.size of the corpus

It is hardly conceivable to collect corpora of certain languages which could have the same size as large European language corpora comprising millions of words; The problem of the representativeness of corpora then arises under a different guise.

2. Feedback to the community

Up to about fifty years ago, the practice of fieldwork (anthropology at the beginning of the twentieth century, missionary linguistics etc.) has left its mark on the commmunities which felt fleeced and exploited without ever having been given access to the results of the research. These communities are now asking for research to have direct spin-offs in diverse forms, and these claims have been taken up by all the organisations which finance or organize research about endangered languages (UNESCO, the DOBES project by the Max-Planck Institute[2], etc.). Claims coming from the communities have nothing to do with the compensation for individual work of the different speakers taking part in the

[2] See the links to the organizations' sites in the bibliography.

collection and exploitation of corpora, and require researchers to become involved (participation in educational programmes, restitution of the recordings and materials collected, making archives accessible by the communities, etc.). The necessity to give back the materials collected to the community should, for that matter, encourage researchers to make databases and to archive their corpora.
However, it seems that these corpora are often still "personal tools" whose sole function is to serve as foundations for linguistic analyses done by researchers only. There are many reasons for this: as for the constitution of any oral language corpus, the technical aspect and the time neccessary for pre-processing (digitisation, and in certain cases, synchronization, etc.) discourage researchers, and even more so as this extra work is not given a high regard by scientific institutions. Furthermore, in certain fields, the collection of a corpus is the result of a relationship based on trust which has developed between the researcher as a person (and not as a representative of a scientific community) and the community or certain members of the community in difficult contexts. The decision to distribute the corpus within the scientific community or more widely (Internet accessibility) is based on the researcher's ethics and is no longer a matter of judicial framework. In all cases, it is preferable for the circulation of the corpus to take place once it has been given back to the society concerned.
Researchers therefore find themselves divided between the desire to preserve a priviledged relationship with their field and the growing need to make the resources on which the analysis and the results of their research are based available to the scientific community.

As the current large corpora of spoken language are developing, standardisation is advancing (since the recommendations by EAGLES in 1993) and the fields of research are becoming more and more interesting. In this perspective, making existing corpora of spoken French or those of any other language accessible and creating new ones is an important task for the "immaterial heritage". Legal aspects related to the protection of speech, which were wrongly considered as secondary for a long time, are currently acting as a very serious brake: many researchers won't distribute their corpora because they aren't sure of having "the right permissions". Many think twice before starting new ones because asking for permission seems to be a fundamental step but also a difficult task to achieve. This is why collective reflection on this point has now become essential.

2.2 POLITICAL FRAMEWORK FOR THE DISSEMINATION OF RESEARCH

DISSEMINATING RESEARCH RESULTS IS PART OF THE RESEARCHER'S ROLE.

> *"Public organisations should constantly be concerned about making the Nation benefit from the fruits of their labour in the best possible way…".*
>
> *"The politics of research and technological development aim at increasing knowledge, promoting research results, disseminating*

scientific and technical information and at promoting French as a scientific language"[3].

It is in these terms that the report appended to the law passed on 15th of July 1982 outlines the "promotion" of research. There is no doubt that these general principles are applicable to researchers whose works result in the development of oral corpora.

However, the conditions of the promotion and distribution will also depend on the possible existing rights on the collected contents and on the results of their anaylsis by researchers.

DYNAMICS OF EXCHANGE AND OPPORTUNITIES GIVEN TO PEOPLE OWNING RIGHTS IN ORDER TO FACILITATE FREEDOM OF ACCESS IN THE INFORMATION SOCIETY

Today we can probably talk about a new way of looking at everyone's connection in the exchange of information. This dynamics of exchange de facto entails new behaviours. Freedom of access, gratuitousness and the right to re-use seem to be self-evident when they are reciprocated.

On 22 October 2003, in Berlin, most of the General Directors of Public Institutions involved in Science and Technology signed the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities,* whose objective is to promote the Internet as "a functional instrument for a global scientific knowledge base and human reflection".

On signing this declaration, research policy makers, research institutions, funding agencies, libraries, archives and museums committed themselves to considering certain measures. These measures should allow "solutions that support further the development of the existing legal and financial frameworks in order to facilitate optimal use and access" of the internet to be found. The text also recognizes the existence of a possible contradiction between the demands for protection and for free access. Finally, it ensues from this declaration that free access requires everyone's commitment as developers of scientific knowledge or holders of the cultural heritage, this free access being granted "while respecting the copyright of the authors or right holders". Free access should, therefore, be regulated and modulated by right holders. Authors (or institutions) can conceed to a "free, irrevocable, worldwide right of access to the work" or even "a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship". "The right to make small numbers of printed copies for personal use" can be given on its own. Formalization of the authorisations could be made in the form of licenses of the **creative commons* type (authorisations for the use of data is given directly by authors without financial compensation. If need be, authors can

[3]Art 5 in Act n°82-610 (15 July 1982) on research orientation, planning of research and technological development, which is now Art. L 111-1 in the law on Research Code. *JO* 16-07-1982, p. 2273 and following.

nevertheless impose limits to this use by restricting it exclusively to educational purposes).

When applied to oral corpora, these licenses can be a means of placing the responsibility on prospective users for the respect of the commitments made by the researcher who created the corpus to those who contributed to its creation.

HERITAGE DIGITISATION PROGRAMMES

The context of our information society has led to numerous public initiatives aiming at assuring the perpetuation of cultural memory. In 2001 in Lund, Sweden, a group of national state representatives from the European Union interested in the issue of digitisation came up with a text which notably encourages putting into place standards of interoperability, the diffusion of good practices – among which the management of intellectual property rights –, the setting up of centres of expertise in digitisation for which information specialists are responsible.

The question of preserving research results arises today with just as much relevance as the results, but also the very materials used in the research, are on digital media. How can traceability of the different stages of the research process be assured? What data should be preserved? Who will be in charge of its preservation? Under what conditions? These questions should now be asked and elements to answer them should be found for each research project. Even though general recommendations can be given, those who initiate a research project whose objectives or stages comprise developing an oral corpus are not exempted from their responsibilities.

2.3 LEGAL FRAMEWORK

The intention of this guide, aimed at researchers, isn't to deal with all legal technicalitities to take into account (readers will be referred to a more detailed presentation of certain subjects dealt in the specific legal documentation in the appendix). It is about making readers more aware and encouraging them to ask themselves the necessary questions in order to understand not only their obligations, but also their rights.

What could the legal status of each of the oral corpora developed by researchers be? This question might *a priori* appear theoretical, but we cannot dismiss it since it will be possible to determine the conditions of exploitation and distribution of corpora thanks to the answers found. To answer this question, the context of construction of the corpus and its different components first have to be known. Is the corpus composed of information in the * public domain? Was it developed from one or several intellectual creations likely to be protected by *copyright law? Is the content of the corpus *personal data? What then are the rights of the speakers or the people concerned?

Once these legal statuses have been determined and once the ensuing rights are known, the modalities for the contractual management of these rights should be inquired about. Have right holders made their views known about the conditions of accessibility and reusability of corpora?

Finally, the issues related to the responsibility of all those who will be playing a part in the "life of the corpus" deserve attention: the responsibility of the creators, the corpus hosts, the diffusers, the archivers… (see appendix).

To make the researcher's task easier, an overview of four important questions which consistently crop up in the construction and life of corpora will be given here: What is the public domain, that is to say "the inappropriable"? When do we talk about copyright for corpora? How can personal data protection be ensured with regard to processing the information making up oral corpora? What are the responsibilities of the people in charge of circulating corpora on the Internet?

THE PUBLIC DOMAIN AND COPYRIGHT LAW

WHAT IS THE PUBLIC DOMAIN?

If the expression "public domain" is generally known by everyone, the legal acceptation of the term could be understood in different ways which must be clarified to avoid ambiguities or misunderstandings when developing oral corpora. In the legal sense, the public domain is a many-sided concept which can refer to a place, as well as to a system or to contents.

The public domain can thus be "the place where civil society makes every effort to influence the way in which collective assets are managed and distributed". In this way, the UNESCO is at the origin of a true policy about contents and is developing a promotional strategy for a strong public domain which is accessible on-line and off-line. The public domain not only covers the concepts of free access to and free use of data, but also the possibility for everyone to be able to exploit it. It is characterized, moreover, by the absence of a monopoly since the information which falls in the public domain becomes *de facto* part of "common things".

However, two types of information can be distinguished: that which was born in the public domain and that which "fell" into it. Ideas, language, legal texts and all the elements which are the basis of the common heritage of a given community make up, by their nature, the "common collection of resources" of the public domain. This common collection of resources remains, however, difficult to fix limits to. In this way, linguistic recordings cause many hesitations. Putting aside the rights of the person who made the recording, are the content of a language and its phonetic expression part of the public domain or not? The question can also be asked regarding traditions and customs. Moreover, is this common collection of resources universal or only common to a small community? Today, it gives rise to more and more identity claims which raise new questions.

After a certain time, works protected by intellectual property rights, notably through copyright or patents, end up falling into the public domain. Copyright, for example, protects works for sixty years after the author's death. In French law, when this time period has elapsed, other types of protection can remain valid for works of the mind: patrimonial rights on the one hand, imprescriptible attributes of moral rights on the other hand. As a result, certain works in the public domain can still be under the protection of moral rights.

These distinctions show two types of situations which seem in opposition to each other: either corpora are developed from works from the public domain which cannot be subject to appropriation (either by their nature or by the fact the deadline for their protection has expired) and thus are copyright-free, or corpora are subject to copyright and therefore under the obligation of getting the necessary authorisations. In reality, as we have seen it, there is an intermediate possibility when copyrighted corpora are given free accessibility within the framework of a license granted by their right holders and allowing the use and exploitation of the results. Although these corpora are not in the public domain, their developers have wished to make them freely accessible and usable. Nevertheless, if creators can renounce their patrimonial rights, it is not possible for them to renounce their moral rights which remain imprescriptible.

COPYRIGHT AND CORPORA

What are the conditions necessary for a corpus to be protected? There are three.

In the first instance, it has to meet the requirements of a *creative activity*: a work compiling information isn't protected by itself.

To be protected, it is necessary for the corpus to have a *defined form*. What is protected is not what the corpus contains, but its exterior and structure.

Finally, the form of corpus needs to meet the requirement of originality. What does the originality of a corpus mean? The originality of numerous creations in the digital age, such as software or databases, can only be assessed according to objective criteria. It appears to be the same for corpora, as they are very often comparable to databases. This is how, more often than not, the fact that the corpus is copied or not and shows a minimum of creative activity will be criteria to determine whether it is original or not (as opposed to only taking into account the mark of its author's personality).

"THERE IS NO ROOM FOR COPYRIGHT WHEN THERE IS NO AUTHOR"

The author is, in theory, the person (or people) under whose name(s) the work is distributed. Scientific work implies the participation of a number of actors, a large number of which are likely to claim authorship of the research results.

Certain oral corpora, like other products of research, can remain the work of one sole person, whereas others can be the work of many. In the case of many authors, the law makes a distinction between collaborative work and collective work. For the fomer, each co-author has the same prerogatives. Other works, such as databases or dictionaries, can be classified as collective works when they have been created

> *"under the initiative of a physical or moral person who edits publishes and distributes it under their own direction and their own name, and in which the personal contributions of the various authors merge as a whole"*[4].

In this last case, it is the physical or moral person who has initiated the work who owns copyright. Besides, the context of creation or the author's status can have

[4] Art. L. 113-2 of the CPI.

consequences on the determination of the copyright holder. Has the work been created in the context of a service assignment carried out by an employee or a state servant? What are the respective rights of the author and their employer? Even if the matter is most often settled by a work contract, it remains tricky when the author is a state servant. Indeed, two approaches have been in direct opposition for several years, one that recognizes the creator's rights on the one hand and one that only recognizes the right of the state over state servants' creations on the other hand. The transposition of the directive on copyright in the information society has encouraged authorities to come to a compromise which recognizes both the authors' rights and the rights of the "state" as an employer when the work is created while carrying out public service missions. If this text is voted by Parliament, copyrights could be granted to state servants.

In return, all rights to exploit the work for the needs of the state servant's mission would be given up to their state employer (the rights to present or to disseminate in the context of the mission). However, in the case of commercial exploitation, the author as a natural person will recover their rights with the obligation to grant a right of preference to their employer and the possibility of sharing profits from the commercial exploitation. This text didn't fail to cause much debating and questionning. How will the scope of the service mission be defined for researchers participating in the development of the corpus? How can the difference be made between exploitation for the service mission and commercial exploitation when – as we have seen it before – the researcher's missions include disseminating the results of their research and promoting them through publication?

WHAT COPYRIGHTS ARE APPLICABLE TO ORAL CORPORA COMPARABLE TO WORKS?

A difference should be made between patrimonial rights and moral right prerogatives. It should also be remembered that the law sets limits to the authors' exclusive rights.

Patrimonial rights consist in exclusive rights granted to the author (or the right holder) or their beneficiaries (beneficiaries of a transfer, heirs etc.) to allow or forbid the protected work's reproduction or distribution to the public. If the oral corpus is a work, any reproduction (digitisation is considered as reproduction from a legal point of view) and any attempt to make it accessible to the public (on an internet site or any other medium) require formal permission from the author or right holder.

As for the moral right prerogatives of the natural person who created the protected work, they are four in number: the *right of disclosure, the *right to retract, the *right of attribution, * right of integrity. Each of these rights is applicable to oral corpora. The corpus developer (in compliance with their right of disclosure) can decide when and how the corpus will be made available to the public, depositing it as an archive is not necessarily considered as disclosure. A corpus which has not previously been distributed thus cannot be made available to the public without its developer's permission. The researcher-developer who will not release the corpus they have developed is perfectly within their rights (in compliance with copyright laws), even if they could get a reprimand from their administration for not carrying

out their public service mission which consists in disseminating the results of their research. The right to retract also applies to oral corpora but the developer's second thoughts can only have to do with their work's intellectual content, not with the financial terms and conditions for the distribution of their corpus.

Even though the right of attribution is easily understandable, one can wonder what the right of integrity means when applied to a corpus. This right encompasses the respect of the work's form (no deletion, addition or modification is allowed) as well as the right to respect the work's spirit (changing the purpose of the corpus is not allowed either).

As it is the case in any monopoly, the authors' exclusive rights have limits. First of all, it should be reminded that they are limited in time and that works fall into the public domain after the time period has elapsed (see above). These limits can also be justified by the way the works are used. In this case, we talk about copyright exceptions whose justifications can be the objectives, the context or the general interest.

Finally, the right to copy works for personal use or the right to quote from them are directly applicable to oral corpora (refer to the Right to quote section in the legal documentation).

THE RESPECT OF PRIVACY

THE RESPECT OF PRIVACY IN THE DEVELOPMENT, EXPLOITATION, DISTRIBUTION AND PRESERVATION OF CORPORA

Developing a corpus, more often than not, involves collecting data. As the data collected can be personal information, the collection process has to comply with the *Informatique et libertés* act: lawfulness and loyalty, prior information, obtaining consent from the people concerned (refer to the Consent section in the legal documentation), respecting the objectives announced[5]… When talking about "research objectives", should the expression be interpreted in a restrictive way as a specific research project identified as such or should its interpretation be wider? The question arises, once the corpus has been developed and scientifically exploited by the researchers who initiated its development, when reusing the corpus and finding new scientific exploitations for it are considered. Today, scientific research is an exception to the general principle through the inforcement of what is called the *extension of objectives*. However, any new scientific exploitation will have to be carried out in the respect of the formalities required for processing data (new declaration or authorisation procedure) and of the rules enforced by law (information, consent, and other relevant garantees…).

Even when the distribution of corpora and new exploitations for them are done under the conditions required, the problem of personal data detention arises.

If the data is irreversibly anonymised, it is no longer within the bounds of the law and it can be kept (refer to the Personal Data and Anonymisation section in the legal

[5] http://www.cnil.fr/fileadmin/documents/approfondir/textes/CNIL-78-17_definitive-annotee.pdf

documentation). However, the need for traceability in research often requires to keep personal data.

And yet, in theory, the **right to oblivion* implies that personal data cannot be detained beyond the time period originally envisaged and whenever the original purpose, which was presented during its collection, has no grounds for existence, the data has to be destroyed. Does this mean that it is not possible to preserve corpora including personal data which hasn't had the chance to be anonymised? It doesn't, but such cases have to remain the exception whenever the detention of personal data can be justified on scientific grounds. In these cases, oral corpora – as public archives – could be entitled to a departure from the right to oblivion, which would allow them to be preserved beyond the time period envisaged in order to be processed for research, historical or scientific purposes. The laws applicable to archives will then determine the conditions under which they will be made freely accessible (the time period can vary between 60 and 150 years depending on the sensitivity level of the data contained in the corpus).

WHAT ARE THE RESPONSIBILITIES OF THE PEOPLE IN CHARGE OF THE DISTRIBUTION OF CORPORA ON THE INTERNET?

The distribution of corpora on the internet can be considered as "the online publication of a communication service to the public". It is thus crucial to understand the obligations and responsibilities of publishers of online communication services to the public (refer to the Processing manager section in the legal documentation).

3 THE PROCEDURE
CONSTRUCTION, EXPLOITATION, PRESERVATION, DISTRIBUTION

3.1 CLARIFYING THE PROCEDURE

The objectives, especially scientific ones, which are related to the construction, exploitation, preservation and the distribution of oral corpora are very diversified and respecting them, as well as their heterogeneousness, involves acknowledging the diversity of procedures which can be used by researchers and those responsible for distributing and preserving these corpora.

The *Guide to good practice* doesn't intend to impose a procedure by prescribing a set methodology but wishes to provide all the information necessary to pinpoint "sensitive" legal and ethical issues. Only accurate and detailed identication of the elements of the situation at stake, and specifically of the data type and medium, of practices in the field but also of the different stages in the processing of data will allow partial legal answers peculiar to the situation to be given as well as an assessment of possible risks. Finally, a reflective analysis of the procedure used in building and processing oral corpora is the first step in constructing a code of ethics acknowledged by the whole scientific community.

3.2 ELEMENTS OF THE SITUATION AT STAKE

Recordings which make up the primary data of a linguistic survey are far from being a uniform whole. In this way, a tale recorded on a magnetic tape at a traditional ceremony on a village square is a scientific object of national heritage which is quite different to the digital recording of a text read by a "paid informant" in the confines of a university laboratory, to the answers to a questionnaire recorded on a minidisc by a researcher at the interviwee's home or even to a spontaneous conversation unprompted by researchers taking place in a café and filmed by one or more cameras.

It is therefore advisable to begin with identifying the elements characterising the data collected in a given situation:

- the *type of data* which makes up the corpus and the data medium (of recording, of storage for its exploitation and also of preservation),
- the *different techniques* used by researchers to collect data,
- the definition of the *participants* and of their role,
- the categorization of the *places* of data collection.

3.2.1 CORPORA AND TYPES OF DATA

If the desire to "capture" speech dates back a long time, it is only recently that technological advances and research (notably in linguistics) have made it possible to consider recordings as real "data". In this way, the International Phonetic Alphabet is an example of an alphabetical recording system designed by linguists in order to standardise the encoding of phonetic and / or phonological transcription of speech. The modern history of audio and video recordings unfolds with the evolution of recording modes and recording media in use.

RECORDING MODES

The analogical recording mode was first to be used for recording and archiving sound. It encodes measured variations in the form of signals which follow the same variation law as that which governs their propagation in a natural environment. For several decades, the digital recording mode has been prefered. In this mode, punctual measurements of air pressure are taken on a regular basis (sampling). These measurements are then encoded in the form of numerical values expressed on a reference scale and are then represented on the storage medium in the form of an organized sequence of binary units.

RECORDING MEDIA

- *Physical media*

The first modern media to allow speech to be archived were physical media. This term comes from the fact that pressure variations measured by a piece of apparatus (a microphone) are physically written onto the medium's material. Among these, there are old cylinders, vinyl records etc. These media are able to keep data in the material they are made of (vinyl, wax, etc.) in the form of an undulating groove serving as an analogical image of the pressure variations measure. These media were used in the last century and have almost been abandoned. Today they show problems of access and preservation.

- *Magnetic media*

Magnetic media appeared later in the second half of the twentieth century. Different media to store data were and still are used today (wire, tape, disk) in different forms (reel, cassette, cartridge etc.). The principle here rests on the endurance of the magnetic particles spread out along the length of the medium (i.e. the property which these particles have to remain magnetic for a long time). Particle magnetization will, according to the different recording modes, be able to encode the information in a binary form (as for hard-disks, DAT cassettes, floppy disks, etc.) or even in an analogical form (as is the case for mini audio cassettes, VHS video cassettes, etc.). Part of these media is intended to be used with IT equipment and others with audio/analogical equipment. Here again, as for any medium, the recordings made deteriorate inexorably with time. As these media remain very popular, it is still easy to have access to tools to read and write them.

- *Optical media*

Optical media appeared more recently; they are mainly known in the form of Compact-Discs (audio CDs, CD-ROMs, etc.). Technology rests on the optical properties of the components, that is, in the case of audio CDs, the capacity of the pits they are composed of to reflect the light of a laser beam. These media are mainly used to store digital data (with the exception of certain unpopular types of laser disks and silver films which are hardly ever used for sound recordings). A very large part of these media is intended to be used with IT equipment, which facilitates the access, tranfer and processing of data. Preservation problems are the same as for all other types of media, even if these aren't susceptible to the same type of damage (caused by light, heat, magnetic fields, humidity, etc.). As it is the case for magnetic

media, optical media have the advantage of being recent and popular, which makes them easy to use today.

Other types of media combining, for example, optical and magnetic techniques exist. (See appendix *Media for recording and archiving sound*).

SELECTION CRITERIA

Media preservation raises, in any case, similar issues whatever the chosen type. That is why selection criteria for the right recording and archiving medium should rather rest on the quality of encoding and the ease of access and processing as well as on the possibility for its contents to be replicated without losing information. Digital media should therefore be preferred to analogical ones as they can be replicated identically and endlessly. Computer media should also be preferred due to the range of tools that computer technology offers for the management, access to reading equipment, the distribution and the processing of data (encryption, anonymisation techniques, etc.) while recognizing at the same time that these tools still raise a number of standardisation issues (for example when it comes to choosing software, formats and compression codecs). Finally, when trying to preserving data, a medium which cannot be erased is also, perhaps, a good safeguard to avoid unfortunate accidents.

The choice of a format which allows identical replication guarantees data durability. It questions the very notion of "master copy" which then refers less to the medium than to the data itself.

STANDARDISING ANNOTATIONS

Oral corpora are generally composed of audio or video recordings and their annotations.

- *Primary data vs. Secondary data*

A distinction is usually made between primary and secondary data:

- *Primary data* is made up of *recordings* and has as close a link as possible to the event documented. It is also composed of other objects collected in the context of the event, such as the documents read or written during the recorded event, the objects manipulated, the pictures referred to, etc. They are also comprised of the computer tracks left during the activity.
- *Secondary data* is made up of a series of descriptions, transcriptions and annotations which complement the primary data and which are often provided after the event and based on the primary data. It also includes metadata, transcription conventions, participants' permission, etc.

The distinction between primary and secondary data is especially useful to differentiate several levels of interpretation and underline the importance of referring back to primary data and of its availability. An analysis is thus based on the audio or video track and not exclusively on its transcription even if it is an important

supplement without which it would probably be impossible to make an analysis. It is in this direction that annotation tools allowing the sound/visual source and the text of the transcription to be time-aligned have been developed. However, this distinction between primary and secondary data has its limits: it should not be forgotten that every recording is the fruit of decisions which are both technical and theoretical – about, for example, the choice of the recording time and the beginning and end of the recorded event, of framing and video optics, of the position and orientation of the microphone for audio features – which rests on prior knowledge about the recorded event. Data is never "offered" or "collected" but is actively developed by researchers (Mondada, 2006).

- *Clarifying data structure*

When writing annotations, standardised procedures are used to encode the content of the comments as well as to specify the type of comments made. For example, in the case of relational databases, tables with fields containing identifications (i.e. POS for Part Of Speech) are used to store values (for example "verb") used in particular types of structure (chains of characters, numbers, etc.).

An alternative standard which is widely used in the field of text annotation is that brought about by the large family of markup languages. This standard marks the boundaries of each comment with formal marks (i.e. tags) specifying the type of comment made. Today, there is a fairly wide consensus in all disciplines about the adoption of the recent XML text markup language as a standard in data structuring and document sharing (see appendix *Encoding and Formats*).

- *Standardisation*

While choosing a standard that allows all annotations to be expressed and their structure to be specified is indispensable, it is not enough to allow the sharing or preservation of a document. In order to share and preserve a document, the language used to encode its structure as well as the content of the comments should be common between the participants (in the context of an exchange) or should remain understandable over time (in the context of a long-term conversation. In the case of a document using a text markup language, the names of structural elements (tags, attributes…) should be known and their definition agreed on and shared, as should be all the constraints (tag sequences, controlled vocabularies, optional or obligatory nature of certain structures…).

We speak about standardisation when a large number of people or an entire community manages to agree on a common language. This is what happened with the International Phonetic Alphabet for example (IPA). While standardisation is necessary for sharing, long-term preservation requires guarantees about the distribution of and access to the documentation on the common languages put into place. Standardisation organizations should therefore be able to ensure some sort of durability to the norms they put into place as well as independence to private interests. They should also be representative of the general interest. Under these circumstances, using norms for encoding and formatting data, wherever they are available, is an advantage. For example, character encoding for ISO 10646, better known as Unicode, is a character code which is claimed to be universal and takes

into account most forms of writing in the world, including the International Phonetic Alphabet. For linguistic analysis encoding, it should be interesting to read the recommendations outlined in the Text Encoding Initiative (TEI) which offers analyses for data structures such as dictionaries, poems or speech transcriptions. It should also be very useful to follow the progress of the ISO working group about the management of TC37 SC4 linguistic resources (see appendix *Encoding and formats*).

In this way, the guiding principles when choosing a technology rather than another can be summarized in the four questions below:

– Does the chosen technology allow *annotations to be encoded in an explicit way?*
– Is the chosen technology of a *proprietary nature* or does it include *legal limitations* which could prevent annotations from being shared with others (proprietary formats, techniques based on patents, etc.)?
– Has the chosen technology been *accepted by the community* with which the data exchange is envisaged?
– Has the chosen technology been *standardised*?

3.2.2 SURVEYING TECHNIQUES

COLLECTING AND DEVELOPING DATA

Linguistic surveys have not always produced recordings for technical reasons (the first speech recording tools have only existed for just over a century) as well as methodological and theoretical reasons. In this way, using written questionnaires, note-taking, the researcher's own intuition and / or observations were and still are description tools used by linguists. The possibility to record speech and the evolution of techniques (miniaturisation of equipment, quality of the recorded signal, digitisation and computer processing of audio and video data), have nevertheless allowed field studies to develop methods which are still very different due to the diversity of the scientific fields concerned (dialectology, sociolinguistics, conversation analysis, psycho-linguistics, oral linguistics, automatic speech processing, ethnolinguistics…). Nevertheless, research on survey methodology has led researchers to consider recorded data as being the product of the survey situation as opposed to viewing it as pre-existing data that has simply been collected (Cameron *et al.*, 1991).

Finally, surveying techniques play an important role in giving the possibility – or not – to control the data supplied to researchers by the interviewee. The remaining pages of this chapter deal with the different surveying techniques used when building oral corpora.

QUESTIONNAIRES

Recorded oral questionnaires can take on different forms; it is most often composed of closed or semi-open questions and of lists of lexical items or texts put together by researchers. The case of texts pre-existing the questionnaire can possibly raise the issue of intellectual property (as, for example, reading a copyrighted text or a piece

whose original content is protected by law). In all other instances, it is a matter of picking up, notably, the variations, regularities and perceptions of these regularities by the interviewee in reference to a common linguistic system.

The degree of sensitivity of the collected information is most often predictable since researchers are responsible for designing the questionnaire and can thus assess the risks according to the nature of the questions asked. However, apparently insignificant questions can also contain hidden stakes, unsuspected by researchers, which come to the surface in the particular context of the survey. Moreover, it should be noted that, more than any other technique, questionnaires contain markers of the act of questioning and of the researcher's *grasp* (Encrevé, 1983) and therefore potentially induce the feeling of being assessed, even if this is often alleviated by the possibility explicitly given not to answer all or some of the questions (see Achard 1991 for an analysis of the different situations in which questionnaires are used). Finally, attention should be paid to a point which concerns numerous survey situations but which is particularly relevant to questionnaires: one part of the questionnaire is often devoted to the collection of personal data (age, social-professional category…) with the aim to draw up the survey's sociological profile.

INTERVIEWS

Interviews are made up of open questions whose main objective is to collect a substantial amount of linguistic data. Interviews always imply guidance from the interviewer in varying degrees (from directed to semi-directed interviews or even non-directed ones; from more standardised to less standardised), thus becoming more similar to oral questionnaires or less constrained oral interactions. (Maynard *et al.*, 2002, Houtcoop-Streenstra, 2000). Although, when interviewing, researchers often introduce the categories and themes which they wish interviewees to deal with, their research methodology can also require for the object of research not to be stated in great detail before the interview in order to collect the most natural productions possible and therefore raises the issue of the choice of the time and content of the information provided to the interviewees (cf. Mondada 2001).

From a legal point of view, interviews are more often sources of data and information about the private life of the person interviewed or of the people mentioned in the course of the interview, and as such have to be protected.

THE COLLECTION OF TALES, SONGS…

The collection of tales, songs and oral productions from traditional cultures is common practice in the fields of the description of languages with an oral tradition and ethnolinguistics in particular. Besides the importance of contextualizing songs, tales and narratives (implicit meanings in a particular cultural context can go unnoticed or seem trivial in another), two elements should be taken into account: intellectual property rights of the traditional productions of a community, and the conditions of collection which are often related to social activities in a public or private setting.

LIFE STORIES

Life stories are often required in anthropological, historical and ethnoliguistic research, as well as in dialectology and many other domains (Guillaumou *et al.* 1997). These types of recordings are by nature an important source of personal data about the author of the narrative and third parties, who can possibly be associated to a particularly sensitive social or historical context, particularly when a personal story echoes an event experienced by one or many communities.

In this way, even in the case of research about exclusively linguistic phenomena, the points featured in life stories and the question of the impact of their distribution in the public space can't escape the researcher's responsibility as he/she requests and uses them.

Furthermore, the conditions for the exploitation and distribution of these narratives can be made in a social context which can be very different from the very particular one which has put its markers on the collection and which often takes place in a particular setting thanks to a special relationship between the researcher and the witness.

Finally, the issue of intellectual property rights of a life story and of inalienable moral right can prove to be particularly relevant to original life stories.

LABORATORY RECORDINGS

Laboratory recordings following an experimental protocol are used in language sciences particularly in the fields of psycho-linguistics, phonetics and automatic speech processing. In this way, certain corpora are of direct interest to applied research and to companies involved in linguistic engineering and are therefore sometimes subjected to partial or total private funding.

Just as for questionnaires, except in particular cases of copyrighted texts, the participants' productions follow an experimental protocol designed by researchers and don't seem to come under intellectual property right (except for particular cases). The particular situation of the person recorded relates, however, to all cases of experimental research carried out on humans.

RECORDINGS OF PROMPTED ACTIVITIES

These are mainly recordings of activities carried out in the ordinary context of the social actors concerned, even if the instructions are provided by researchers (activities assigned to children in a school environment, simulated tasks in a professional environment, etc.). This situation combines both the characteristics of recordings following an experimental protocol (which is the researcher's responsibility) and those of an ordinary context in an ecological environment; it thus offers a double framework that researchers can control. This explicit intervention of researchers (whose role can be clearly identified by the participants) makes it easier to obtain informed consent; however, particular attention should be paid to a professional environment which could restrain consent (confidentiality…).

RECORDINGS OF ACTIVITIES CARRIED OUT IN THEIR NATURAL ENVIRONMENT

Research in sociolinguistics, conversation analysis and analysis of the use of technology (Computer Supported Cooperative Work; man-machine dialogue), takes an interest in data collection in activity situations which are neither prompted by researchers nor induced by their instructions. These are activities as they take place ordinarily even in the absence of researchers. These activities can vary greatly: meetings, professional activities, requests for information, telephone interactions, etc. Data collection techniques are also very different. These range from participating observation to authorised recording and include the use of "key informants" chosen within the group of peers observed and in charge of carrying the recording device (a microphone and possibly a camera).

The common objective of these techniques is to research naturally occurring data and therefore implies a methodology which attempts to minimise the effects produced by the recording devices (Heath, 1997; Jordan & Henderson, 1995). The data is therefore very likely to contain sensitive information with regard to the protection of private life. The conditions for consent collection should take this into account and adapt accordingly.

THE REUSE OF RECORDINGS

Certain corpora which are made up of recordings produced by people other than the interviewers for non-scientific purposes or purposes other than those mentioned during consent collection can be reused in order to carry out linguistic research or to be made available to the public for patrimonial, memorial or political purposes (this is why, for example, the New York Fire Brigade in August 2005 made the radio communications that occurred during the 11 September 2001 attacks available). These corpora are therefore characterized by the absence of consent for their new purpose and by the fact that the archived utterances weren't produced with full knowledge of this purpose but in another framework and with other objectives in mind. During interviews or seminars – recorded for example with the aim of distributing their transcribed contents – permission for distribution can thus relate to the validated transcribed utterances, and not to the later reuse of the recordings.

THE REUSE OF BROADCAST RECORDINGS

The reuse of broadcast recordings is a particular case in the previous category and has the particular characteristic of dealing with data produced within the framework of public broadcasting.

Once again, if the content of the recordings is copyrighted (for example in the case of an original production), collecting consent is a requirement for exploiting it. However, an exception exists for a determined amount of time when referring to speeches intended for the public and delivered in public, as specified in the following lines:

> *Intellectual Property Code, article 122.5:*
> *Dissemination, even in their entirety, through the press or by broadcasting, as current news, of speeches intended for the public made in political, administrative judicial or academic gatherings, as well as in public meetings of a political nature and at official ceremonies*[6].

It should be reminded that the personal recording of a broadcast corresponds to a legal licence to make this copy for personal use only. Performances of the copy can be carried out exclusively within the "family circle". This also applies for commercial cassettes or DVDs, as the right to make copies (with its limits) doesn't grant any exploitation rights.

Finally, it should be specified that the public nature of the context of broadcasting doesn't imply restricted protection of personal data.

The diversity of the techniques used to collect data defines a variety of *situations* which give prominence to *participants* whose role is the first element of categorization.

3.2.3 ROLE OF THE PARTICIPANTS

The participants in the study and in the recorded activities can be categorized in different ways which all throw specific light on what they are doing and what they are saying (Sacks, 1972). This is how participants in a recording situation can be considered at the same time as surveyees (if we consider that the situation itself is an object of study) and as social actors - whose precise characterisation depends on the context, the activity, the forms of their involvement and participation, which involves both the social history of the people and the local fulfillment of their role, and of their identity during the encounter. According to the way in which researchers themselves deal with these multiple categories, there can be different consequences both for the object of the survey and for the assessment of the more or less sensitive nature of the activity.

PARTICIPANT CATEGORIES

The very varied terminology used in literature to define the categories of participants in a survey shows various ethical and theoretical implications (Cameron *et al.*, 1991). Here is a non-exhaustive list of the terms used in different research contexts to characterize participants according to their involvement in the survey:

– informants,
– speakers,
– subjects,
– 'Guinea pigs',
– natives,
– social actors,

[6] Article 122.5 in the Intellectual Property Code.

– participants,
– collaborators,
– partners,
– surveyees,
– witnesses.

These terminological choices most often stem from theoretical and political considerations which show the type of pre-existent relationships which have been built or have developed between the interviewee and the interviewer.

Even if here we can't develop the stakes of these theoretical considerations, it is nevertheless important to pinpoint the markers of a special *relationship* which is the basis for various types of developments of the interviewee/interviewer *pair*, with implied rights and obligations which depend on the characteristics of this relationship (Sacks, 1972).

Two elements in particular define this relationship: the proximity and distance of the participants and the roles in action and in situation.

- *Proximity and Distance*

The question of the accessibility of the surveyed situations for the researcher has always been raised and has led to different forms of *fieldwork,* ranging from immersion into a community completely unknown to the researcher to the exploitation of the researcher's links to the community he/she belongs to.

These problems have been dealt with in terms of the observer's paradox – a principle according to which the surveyed phenomenon is influenced by the researcher's observation (as was the case for the vernacular in Labov, 1972) – as well as in terms of symbolic violence between the interviewee and the interviewer (Bourdieu, 1993). They have also been dealt with in terms of *reflexivity* – by researchers who took their own presence and that of the surveying device into account when analyzing the surveyed object (particularly in anthropology, Clifford & Marcus, 1986; Mondada, 1998).

Surveys carried out among the researcher's "close friends and relatives", when he/she uses his/her own networks for a survey, make it easier to make contact and to access the field and at the same time make it more difficult to make distinctions between the relationships induced by the survey and personal relationships. These questions are not raised in the case of surveys made among *"distant people"* (an observed community, a sampled panel, witnesses who haven't been chosen by the researcher…) where the difficulties of access can be greater but where once confidence has been gained and a relationship established, the invetigator as such often has a more clearly defined and recognised status (Beaud & Weber, 1977).

Moreover, research in social sciences and the humanities has often used "captive populations", in the sense that researchers have easier access via institutions (schools, hospitals...) and that these populations have limited possibilities to refuse to collaborate (children, pupils…). Consequently, particular attention should be given to the way certain populations are approached. These include:

– the underprivileged,
– the handicapped,

– children,
– pupils and students,
– employees from companies or organizations who have been contacted through their superiors,
– etc.

It is therefore usual practice to get double permission – authorisation from the people surveyed and from a legal guardian (children represented by adults).

This particular case shows that signed authorisation cannot always be considered as enough and that certain surveyees need to be protected over and above their signed permission (researcher's responsibility).

- *Roles in situation*

In a survey situation and according to the techniques used, the interviewer/interviewee relationship can take on very different forms and can imply more or less direct involvements.

- *Interviewer's roles*

– *external observer,*
– *participant-observer,*
– *committed observer* (defending the community),
– *member of the community* taking part in action-research (the project comes from the community or takes its problems and objectives into account and intervention is performed through a specific procedure in the form of "action-research"),
– *a disguised observer* (*cross-dressing* in the ethnographic tradition) who fits in the community through their connections, their job or their position but who doesn't say that they are carrying out a survey.
– "*a Wizard of Oz"*: the researcher hides behind a technological device which is supposed to answer the informant.

- *Interviewee's roles*

– *surveyee*/informant/focused speaker
– '*peripherals'*: technicians, passers by, spectators...
– *people associated* to the official participants to the survey (ex. customers phoning a call centre, or even the interviewed woman's husband)
– the "*partner*": a priviledged informant who carries the recording device and who allows the interviewer to enter a group which the partner is part of or which he/she has access to.

These roles give an account of the possible variations between participation and observation, the tension between the two terms being illustrated in the expression "participant observation" (Becker, 1960; Platt, 1983; Spradley, 1980). According to these roles (Adler, 1987), the involvement in the survey and in the recordings will be very different, as will be the conditions for making contact to obtain informed consent.

3.2.4 PLACES

Information about the place of collection conditions the specific legal reponses related to its own characteristics and the role it plays in the survey situation.

A difference can thus be made between *public places,* in which the scientific activity of audio and video recording doesn't require other authorisation than that of the person being recorded, and *private places* which are subjected to preliminary authorisation of the owner/person responsible, a requirement which is distinct from the collection of the interviewee's consent.

Place can also be defined according to the relationship that the participants develop. Is it a place where the interviewee's presence is due to the interviewer (laboratory, recording room …) or does the interviewer go into the field and occupy the interviewee's own space?

Finally, the place where recording takes place can be included in the data (audio or visual characteristics present in the data) or it can only be a piece of information possibly part of the metadata.

3.3 FIELD PRACTICES

This section aims to show the omnipresence of the ethical and judicial stakes in the different stages which make up the field procedure for the contruction of corpora composed of oral, interactive and multimodal data. Attention should be paid to the *preparatory stages* of the survey, prior to the recording of data, where a relationship with the people concerned has to be established: these modes of approach are closely linked not only to survey methodologies (cf. *supra* 3.3.2) but also to the technical possibilities and limitations of the chosen recording device, on which specific constraints to get permission to carry out a recording depend. Once the survey is complete and the data analyzed, the *return to the field* has to be planned to give 'feedback' under different forms about the results and the experiments; this should be anticipated and configure the type of commitment made towards the people concerned.

3.3.1 MODES OF APPROACH

Surveys whose objective is the collection of recorded data are necessarily dependent on the quality of the relationship with the key informants – whether they are referred to as informants or partners (cf. *supra* 3.2.3). The mobilization of these people varies according to the chosen survey method: attention is drawn below to the temporality of the different approaches with the people directly concerned or their superiors, and on the issue of knowing how to plan the return to the field, compensation or possibly payment to these people.

TYPOLOGY OF RELATIONSHIPS AND MODES OF APPROACH

The way in which people are approached in the field – the way in which a personal and social relationship is established – can be considered as an act which has immediate ethical and judicial implications. Establishing a relationship with informants has, on the one hand, an impact on the quality of their collaboration and,

at the end of the day, on the quality of the data thus built; on the other hand it has consequences on the relationships of trust, of acceptance, or even of interest or scientific curiosity that informants develop towards interviewers.

A typology of the relationships established with informants can be outlined based on the point when they are approached in the survey process:

- When the survey uses a process of *personal convocation* of informants in the laboratory, the modalities of their involvement are generally made clear *beforehand* when the people accept to take part in the recording which will be carried out in places and at times agreed on in advance. People are then either selected and contacted by the researcher (or by an institution working for them), or they respond to a "call for volunteers". This call or the recruitement offer is the first act of communication which shows (or generates expectations regarding) the form of the contact, or even of the contract which is being established with the researcher.
- When the survey is carried out in the form of *fieldwork* requiring the interviewer to be present in the field for a period of time that can vary in length and to use forms of *participant observation* – usually dealt with in ethnographic methods borrowed by linguists and other researchers in the field of humanities and social sciences (Depperman 2000, Duranti, 1997, Hammersley, Atkinson 1995 & Moerman 1988) – the relationship with informants is established through the *duration* of this presence and is often associated with the development of personal relationships which imply, amongst other things, mutual trust. In certain fields, the researcher isn't the first to intervene and others may have done so before them. Depending on the way their predecessors behaved, their welcome by the community will be more or less warm, and the demands in terms of feedback (cf. 3.3.4), in particular, will be bigger or smaller.
- When the survey makes use of *interviews, "vox pops" or recordings of activities* carried out *randomly* in public places without targetting particular witnesses but passers-by chosen simply because they happen to be in a particular place at the time of the recording, a preliminary meeting with the informants is, by definition, impossible. Explaining the objectives and requesting authorisation therefore take place *just before, during* or *just after* the recording is made.
- In certain cases, it is possible to consider a contact *after* the recording has been made: this is the case for recordings carried out without some of the participants being aware of it as their stepping onto the recording set was unpredictable (this is the case for telephone conversations, for example, in which part of the people collaborate in the survey and others are not always aware of the recording taking place; they are contacted afterwards to give their consent).

The form of the contact, the involvement, the credibility and confidence varies enormously depending on whether the interviewee/interviewer relationship is

established beforehand or during fieldwork, in a long-lasting way or at the very time of recording, or even after this.

THE PEOPLE CONTACTED

In the points that have just been outlined, contact has been considered, for the sake of simplicity, to be established with the one person or the people directly concerned by the recording; and yet, these people often belong to a group or an organization - which implies making multiple contacts. It is therefore a matter of making a distinction between:

- The case where the informant acts *in their own name*, individually.
- The case where the informant *is contacted* in the context of his/her professional or institutional activities and therefore takes part as someone belonging to an organization. The superiors of the people targeted by the survey are also contacted beforehand: this can be the case for managers of a company, for the headman of a tribe, or for pupils' parents. It should be noted that the relationship between the person involved and their superiors isn't always plain and calls for distinguishing between what will be promised, explained, shown etc. to the people and their superiors.

PAYMENT

When people are approached to participate in the study, promises can be made, real contracts can be offered as well as compensation, payment and reimbursement. These agreements can be ethical, judicial, social or even financial. In any case, the issue of the informants' "compensation" is raised and is very different whether it is considered as "counter-gift", "payment", "compensation", "service rendered"…

Several cases are possible:

- *During or even before the survey:* payment can be promised as soon as the first contract is drawn up, as well as counter-gifts in kind, token counter-gifts and benefits for the community concerned;
- *After the survey:* acknowledgement of the informant's role can be done in forms ranging from thanking or acknowledgement to considering them as a co-author, a collaborator or even a research partner, dissemination of the results, restitution of the data/corpus in the form of archives, dissemination of skills, positive spin-offs for the community in the wide sense and in the long term (as it is the case for the benefits expected from medical research).

For a discussion of these forms of "returns", the reader can refer to the "feedback" in the field section (*infra*, 3.3.4). The question remains to know what we can/should promise to informants when establishing a relationship, bearing in mind the fact that:

- This relationship *changes with time* (especially if the field survey is due to last a long time).
- This relationship can more or less recognise the "informant" as a *partner* in the research project (and not only as an "object"), in participating

projects where "natives" bring more than their own performances (for example by collaborating in the transcription, translation or data annotation process).

– *Financial renumeration* can be less problematic for informants who are recruited (sometimes by specialised organisations) in the framework of a formal contract; it can be more problematic in the field as it implies competition not only between the possible participants but also between the researchers having access to these informants (such is the problem, for example, for linguists coming from universities providing little financial support faced with researchers coming from universities that are better funded – and, as a result, informants could prefer the latter and have demands that are difficult to meet). The practices of anthropologists and linguists differ on this point. In the case of participant observation, it can be delicate for anthropologists to pay the people who provide information as it could trigger an escalation of the price of the information. Nevertheless paying speakers and/or translators who spend several hours a day with linguists is fair compensation for real work and doesn't necessarily hinder the relationship of trust which has been established between the two people.

  Financial renumeration is only one case amongst others of possible "return" (or compensation, salary…), and is always done in a more or less implicit way in field surveys, as part of daily life and of the negotiation of mutual relationships.

3.3.2 Recording device

The choice of the device to record copora influences the way in which the people concerned will be treated, their consent will be obtained and the acceptance or acceptability of the recording will be negotiated.

A few aspects which may prove to be pertinent will be discussed here, from the choice of contexts in which the recording is carried out to the terms of the recording.

Context of the recording

By definition it isn't possible to record *everything* and researchers have to make choices. These depend on the target object of research and technical constraints (it is, for example, difficult to record a video at night or audio data in very noisy premises), and also on the respect of the people recorded.

Other factors to be considered:

– Choosing the *moment* to record: it is about finding the right balance between the moments that are interesting for the interviewer and the respect of the interviewee's privacy;
– Choosing the *activites* to record: these can be more public and social or even rather intimate and private;

– Choosing *where* to record: here again, there is a difference between public places disconnected from people's ordinary private lives and private places: the *laboratory* is a place which is totally disconnected from the informants' lives – and this is why researchers willing to work on situated social practices avoid it; people's *homes* are where people live and can themselves be divided into more "public" or more "private" places (a meal eaten in the dining room, in the kitchen or in bed doesn't have the same content, in the same way as an interview carried out in the living room or at the kitchen table); *workplaces* are also, although in a different way, structured by confidentiality issues which need to be respected; Failing to respect them could lead to an obligation of confidentiality of the data collected which means that it would be impossible to use it (see 3.4); *religious places* and sacred and/or taboo places also need to be respected. Overall, knowing the place well, as well as its geographical and social organisation is necessary before considering any recording (image or sound).

The right balance to be found is therefore somewhere between the contextuality and naturalness of the recorded data and voyeurism – the choice of the moments to record can have important consequences on the rest of the survey (on the permission to use the data and on the participants' *post hoc* right to retract).

Recording methods

Recording methods are often considered when choosing the contexts to record (cf. *supra*), the targeted activites as well as the acceptance or resistance modes of the people concerned. Different technical aspects can play a part in the acceptability of the recording by the people recorded:

– The fact that the recording is an *audio* or a *video* recording: for certain activities, the people concerned might prefer audio to video – which is considered to be more intrusive – even if it means switching from audio to video afterwards anyway once the recording methods and effects of the recording on the activities have been noted.
– The fact that the recording is carried out with the *interviewer being present*, by *technicians* or by a *pre-installed device* which doesn't require the researcher's presence has an effect on its acceptance: even if the camera or microphone are often viewed as "prostheses" or extensions of the researcher (for example, when participants speak directly to them), certain participants may prefer the researcher not to be present.
– The fact that the recording is carried out by the *researcher* or the *participants* themselves: on the one hand delegating the recording to the participants can be seen as a form of control on their part about what is recorded; on the other hand, this type of delegation can be turned down as it can be seen as a form of advanced collaboration which diverts participants from their activity.
– The fact that a recording is carried out by a device that is *visible* or *discreet*, or even hidden: there are numerous debates about resorting to a

hidden microphone and about the consequences this choice can have on the possible relationships with the participants (Mitchell, 1991, Mondada, to be published, Welland & Pugsley, 2002); moreover, even when participants are aware that they are being recorded, resorting to the use of a visible device can be perceived both as a guarantee of transparency and as a hinderance. Miniaturized devices can often be installed in a way which quickly makes them blend into their environment without having to conceal them at all.

– The fact that the recording relies on *technical devices which require human intervention at frequent intervals* (for example, in relation to the lifespan of the battery or the duration of the cassette) means that the researcher (or participants in charge of replacing the cassette) will disturb the activity, which is something that other devices with longer autonomy manage to avoid (for example, those that record directly onto hard disks). This can have repercussions on the witnesses' behaviour due to the disturbance caused, especially for certain activities (such as operating on a patient, carrying out a therapy session, discussing a tricky contract, being in the process of creating something).
– The fact that the recording does or doesn't include *blind spots* for participants who would like to take time out for a moment: for example, the frame and angle determined by a single camera allow the zones which are not covered by it to be inferred, whereas the sensitivity of a microphone as it is pictured by participants or the fact that the researcher resorts to several cameras on the same set could give the impression that it is a surveillance system which people can't get away from.
– The possibility to stop or to *demand the recording to be edited* can be seen as the materialization of the right to retract; the fact that erasing or editing the recording can be carried out by participants whenever they wish or carried out later, or by a third party, can give the impression of a greater or lesser latitude to intervene on the data and implies different relationships of trust.. This issue – as many others – is once again linked to the technical constraints of the recording and to how sophisticated the recording device is. This should be taken into account when choosing the recording medium which allows data to be erased immediately or not or which allows what has been recorded to be viewed on the premises or not.

These considerations (Mondada 2006) clearly show how technological and judicial questions overlap, as the personal, ethical and judicial respect of the participants is materialized in the technical choices put into place.

3.3.3 Asking for authorisation and informed consent

The definition of "informed consent" and its translation into forms of social relationship (contact with informants) and materialized forms (documents that are exchanged and signed) are attuned to the context and the objects of the survey, as

well as to the socio-cultural conditions of the group in which the survey takes place. Here, some of the points to reflect about are outlined, starting with the definition of what "informed consent" is, by examining the issues as to when these questions arise, which people need to be informed and asked for permission, what forms the information should take, which objects should people be informed about and what forms their consent itself should take.

Definition of "informed consent"

Authorisation forms to be filled in by informants are often discussed; it is important however to make sure this authorisation is closely linked to the preliminary information given to the people concerned: without *information*, *asking for authorisation* has no objective or sense. This is why we talk about *informed consent*, in the sense that accepting to be recorded closely depends on how well the objectives of the recording are understood. In certain fields, the difficulty encountered by researchers to make the objectives of their work be understood shouldn't however lead them to overlook the request for consent which should then be expressed according to the type of society in which the fieldwork takes place (for example, how can a signed form of individual consent be considered in a society with an oral tradition in which private law doesn't mean anything?).

Time for informing and asking for authorisation

Asking for authorisation depends on the way in which the people recorded are approached. It can be different depending on the time it takes place.

- information and request can be *prepared in advance* during a field trip and depend on how well interviewees know interviewers and how much they trust them,
- information and request can be made *just before* recording,
- information and request can be made *just after* recording,
- information and oral request can be made *before* recording and written request can be put into place *after* recording (with the possibility to retract).

The longer the researcher stays in the field, the more thorough the information is; it is more limited when the request for autorization occurs quickly before or after recording, without any other form of contact between the interviewers and the interviewees.

The time when informing and asking for authorisation take place can be chosen according to its planned effects on the way the recorded activity is structured: the time for informing and asking for authorisation is often chosen in a way that won't disrupt the activity from the participants' point of view (e.g. Asking a customer for their authorisation while buying something can be a cause for disrupting the sale and can thus be refused to the interviewer who wishes to document this activity), or from the interviewers' perspective (for example, requesting authorisation at the start of a conversation changes the sequential organization of this opening).

If the information and the request take place *after* recording, the information can be seen as an "unveiling" or as a "revelation", which means the recording is considered as "deceitful" after the event: this can bring about recategorizations of the participants and activities (the person introducing themselves as a tourist lost in the city and asking for directions becomes a researcher working on the description of space in requests for itineraries) (Mondada, to be published). Furthermore, this technique is not conceivable for many research fields. These cases of deceit are thus particularly unwelcome in certain communities and harm the scientific community as a whole and future researchers.

STATUS OF THE PERSON ASKING FOR AUTHORISATION

Even if the researcher is the person who informs and usually asks for the authorisation to record, there are different possible scenarios:

- The most frequent scenario is when the *interviewer* is the person who informs and asks for authorisation.
- Often, however, the researcher sends *students* or *collaborators* in the field who are spokespeople for the project.
- In certain cases it is posssible for the participants themselves to become spokespeople for the project: this is the case when the researcher asks a participant to inform other participants (eg. the host who invites his friends for a meal which will be recorded; the salesperson who asks his clients to accept to be recorded; the teacher who asks for authorisation from their pupils or students etc.). This delegation is part of the collaborations in the field between interviewers and interviewees; it can however be a source of misunderstandings and difficulties.

In the same way, authorisation can concern the very people who signed or people who depend on them (subordinates, children, students, etc.). In the latter case, it is important to take into account the fact that *authorisation* is not always the same as *acceptance*. In societies where individual right doesn't exist, the opinion and authorisation of the group as a whole or of some of its leaders (political or religious) are often essential.

WHAT DOES INFORMING MEAN?

The necessity to inform the recorded participants lies at the heart of informed consent. However, as soon as this necessity is questioned, questions arise. What does "informing" mean? What should people "be informed about"? Under what conditions can we say this information leads to the "informed" status of the surveyee.

The very notion of "information" can lead people to think of it as a simple transfer of messages and contents. It tends to obliterate the processes, contexts and contingencies which characterize this communicational activity through which a researcher explains the objective of their study to his partners in the field. As soon as we think of it in terms of activity types, informing surveyees raises a series of problems to solve.

– *Its appropriateness to the addressee*: explaining the research project and making sure it is understood and shared means it has to be adjusted to the skills, register and level of understanding of the addressee; this adjustment also concerns the survey context and methods, taking into account the relevance between what the partners see done in the field and the explanations given about it;
– Explaining the *objectives of the survey* should be done without going against it: this raises the issue of the balance to be found between the clarity of the survey and the effects it can induce on the participants' behaviour;
– Explaining the research project can be done at *different levels of generality* (from "it's a study about the way people speak" to "it's a survey about the frequency and the contexts of use of non-obligatory link in French").

The information given to surveyees not only includes the explanations about the scientific project but also precise information about, for example:

– the people *responsible for* the study and their institutional affiliation, as well as sponsors,
– a contact *address*,
– the *people who will have access to the data* and who will work on it,
– *the way* in which interviewees were chosen and the population they belong to,
– the way in which the data will be a*nonymised*,
– the fact that data will be transcribed according to *particular conventions* (possibility to give an example),
– the way in which data will be *archived* once the study is finished (conservation or destruction at the end of the survey, conservation by which guarantor, possible terms for reuse, distribution to other researchers),
– the *modes of access* to the information related to the project and concerning in particular the data/analyses referring to the person (possibility to have access to files and information about the person in particular),
– human rights, especially the right to *retract*,
– possible *risks* as well as positive spin-offs of the survey, whether moral or material.

Information modes can also differ from one addressee to the next according to their culture, in particular:

– The information can be given orally, either individually in *informal conversations* or collectively in *information meetings*…
– It can be given in writing (in a brochure, leaftlet…) or by email.

In the context of a written culture, it is recommended to leave a written trace; in the same way, it can be useful to refer to a website where the evolution of the project can be followed (possibly with different modes of access).

SCOPE OF THE AUTHORISATION REQUEST

Only once the information phase is completed can the request for authorisation to collect data take place. The question which arises is to know how to define the scope of this authorisation.

Indeed, authorisation concerns the following dimensions which can interact and overlap:

- the *actions* carried out by researchers within the framework of the project: recording, constructing the corpus (transcription, translation, annotation, etc.), archiving conditions (place of storage, planned duration for the conservation, answerable institutions...), analysis within the framework of the objectives announced, use of all or some of the data, disseminating the results of the analysis, the conservation/destruction of the data once the survey is completed;
- the *formats* and the conditions for recording: audio/video, with many cameras/microphones, at times which are or aren't known by the participants, for limited or long-lasting periods of time, every technical choice which affects the way in which the person will feature in the data can be explained or even negotiated;
- the conditions *of distribution* of the data and results: in whole or in part (short extracts whose maximum length can be decided beforehand), in text form exclusively (transcriptions) or in audio-visual form (in Powerpoint slideshows for example);
- the contexts of distribution of the data and results: research contexts (*workshops*, conferences, congresses), university teaching contexts, broader training and popularization contexts, field-related contexts (for example, an explicit request for authorisation to reuse data in a training course within the institution where the data was originally collected has to be made as it can turn out to be highly sensitive data);
- broad contexts of distribution: in the form of a CD or a website.

Explaining these contexts overlaps with explaining the activities in which the data will be used; in both cases, the stake is that of the people who will have access to the data within the framework of these activities. Distribution contexts where the researcher has some sort of control (through agreements, for example) can be distinguished from distribution contexts where no control is possible by definition (websites, for example).

It is possible to allow surveyees to add personal constraints; nevertheless, this possibility raises the question to know whether it is legal or not and to know how it can be interpreted. One of the major problems that arises when requesting authorisation – as it does when informing participants – has to do with the possible shift in the objectives of the survey which are not always completely determined at the start and can thus evolve throughout the work in the field and on the corpus. This is why it is important to express the objectives of the survey in a way which is general enough to allow possible evolutions of these objectives to be later added as research progresses. However, any change to the objectives entails a new authorisation request (cf. *infra*).

FORMS OF AUTHORISATION

The request for authorisation can take different forms, which also depend on the socio-cultural context in which the survey takes place: for example, requiring the signature of the surveyee only makes sense in writing and literacy cultures, in which this procedure is meaningful, doesn't frighten people and isn't related to other practices with which it could be confused (such as signing cheques).

The forms of the request can therefore be differenciated according to the medium on which they were recorded:

- written and signed request,
- oral request,
- it is possible and useful to plan for oral authorisations themselves to be recorded, in audio form or as a video, which ensures its traceability; This solution should be preferred when working in societies with an oral tradition, making sure that, according to need, the degree of formality required by the language practices of the community concerned and the choice of language are respected (for example, an individual recording with the speaker for a punctual authorisation or an authorisation recorded during a more formal meeting with the authorities).

In the case of a written request, it can take up different forms – in a prewritten text (form):

- A *compact* text which synthesises the different aspects of the request for authorisation and which requires a global agreement (or refusal).
- A text with boxes to tick and therefore giving *choices*: compared with the first one, this format has the advantage of materializing real choices for interviewees and therefore of giving them the possibility for partial refusal (for example, they can agree to an audio recording but refuse a video recording), or even for adding other constraints (for example, they can demand the video to be anonymised as well as the audio). The question which arises then is that of the formulation of alternatives so that they don't become redundant, too complicated or too long to process for the surveyee.

A problem can arise when a collective request is made, when groups are concerned (for example, in the case of meeting recordings): if too many alternatives are given to the participants to choose from, it is possible for the answers to lead to contradictory results where no common denominator emerges; In this way, requesting authorisation from groups can present problems and contraints which are not the same as for individuals.

For a more in-depth study of this point, refer to examples of authorisation requests in the appendices.

3.3.4 AFTER THE SURVEY: FEEDBACK AND DEBRIEFING

Emphasis is often laid on preparing the field, but it is also important to plan the departure and return to the field. This is important from a scientific point of view as well as from an ethico-judicial one: the return to the field can prove to be necessary at any time to check data, to complement the survey or to make contact with informants again. If leaving the field went badly, then returning there will be impossible. Moreover, presence on the field not only induces relationships of trust but also expectations the researcher is committed to in the long term: leaving the field simply by disappearing, after having been immersed, having often built up close relationships with the participants, requested their help and expected performance from their part can result in people feeling deceived. Once the knowledge, answers and corpora have been "extracted" from the participants, it is a question of knowing how to "give something back" to the people without whom the survey would have been impossible (see also the payment issues dealt with earlier). For example, it is now impossible to work in certain fields (in the case of endangered languages) without considering giving something back to the speaker or the community, or even to promise some sort of commitment from the researcher (involvement in educational or literacy projects etc.)[7].

It should be pointed out that giving "feedback", "debriefing" and disseminating the results of the experiment can already be done during fieldwork in the form of reports of partial results for example. The distinction between what is done "during" and "after" fieldwork can thus be put into perspective.

Many types of practices are possible to ensure that something is given back to the surveyed populations. Some of these will be listed here, ranging from the presentation of results, which is closest to the academic context, to the formulation of knowledge and skills closest to the field. Choosing a "field policy" probably lies in the assessment of the distance between the "feedback" and academia or the field:

- Presenting results at the end of the project: formulating results can be more or less popularized, closer or further from the preoccupations of the surveyees, the presentation of results can consist of examples of *transcription* and of analyses of transcriptions: participants react in a very different way (they are sometimes surprised, sometimes shocked) to the representation of their voice.
- *Empowerment* method (restitution): this consists in not simply considering "returning" the data as a way of giving "information" but also as a way of bringing knowledge and skills to the surveyees' community (Cameron *et al.*, 1991): in this way, not only can presenting analyses be considered but also allowing participants to continue to collect data and to analyse their own data for their own purposes, the

[7] Concept paper from UNESCO (2001), *Language vitality and Endangerment*: "Any research in endangered language communities must be reciprocal and collaborative. Reciprocity here entails researchers not only offering their services as a quid pro quo for what they receive from the speech community, but being more actively involved with the community in designing, implementing, and evaluating their research projects".

spin-offs of the analysis can be expressed in terms of diary entries, themes, participants' preoccupations, *requests for expert reports* often expressed by communities can also be answered to the best of the reasercher's ability (for example, workshops reflecting about the shift from oral to written language, or about the translation of official documents, involvement in bilingual educational programmes), the *knowledge generated from the survey* can also be made available to the community by reproducing it in forms other than traditional university writings (for example, in the form of exhibitions, or other cultural by-products), *training* based on the results and methods of the survey can be offered; more generally, passing on tools for analysis and skills which could be useful in the field can be considered.

– The issue of "returning" the data itself in the form of a corpus or archives can turn out to be delicate: it can be called for in certain cases (thus in the case of endangered languages, national heritage archives bequeathed to the community should be put together) but it should also be possible to avoid it in order to protect informants (in this way, in the case of surveys in companies or institutions, the data collected could be of interest to certain superiors but could be harmful to subordinates). Returning archives, when relevant, often raises a number of issues: the limited access granted to the people who can refer to them, taking into account the risks and advantages induced by them available in the field, and the modes and technological choices to access archives: if archives are formatted in a way that makes it easy for the population concerned to access them, the technology has to be adapted to the uses and abilities of the populations (it is pointless making a DVD if nobody has access to a DVD player, or making a website if nobody has access to a computer: the question raised here is that of managing the asymmetry between "academia" and the "field"), the assurance that people will have access to publications raises isues that are similar to those related to the access to data, although these are often easier to solve.

3.4 ANONYMISATION

Giving the possibility or the guarantee (which will later be put into perspective) to anonymise the collected data is an important element as regards the respect of people's privacy and the lawfulness of corpora built by researchers. Anonymising data is however neither a simple process nor a problem-free guarantee as it raises a host of technical, scientific and sociological issues.

Data anonymisation is an important gauarantee in terms of data lawfulness and use. In certain cases, when it truly guarantees that people concerned will not be identified and when data is not copyrighted, data can be lawfully used even if no prior request for authorisation has been made. One should be cautious in this case as many limitations and difficulties stand in the way of anonymisation (cf. below).

3.4.1 DEFINITION

Even though anonymisation is often referred to, the legal issue that is raised is that of the *impossibility to identify people*: the objective is to make it impossible to identify the people concerned from the collected data and from the different forms in which it is represented (for example, its transcription). Identification procedures have changed dramatically with the adavances of technology which now offers easy ways of storing and distributing data as well as powerful tools to process it (sorting, cross-referencing, crosstab queries etc.).

Considerations include:

- anything that makes it possible to *directly* identify someone: when referring to the speaker, or to a third party and to their private domain, or from the speaker's characteristic features such as their voice or physical appearance;
- anything that could cause prejudice to them;
- anything that can indirectly help identify the speaker concerned by cross-referencing information.

Procedures that remove these references to and characteristics of the speaker are called data "anonymisation" procedures.

3.4.2 DATA CONCERNED

Anonymisation isn't just about recordings or transcriptions, it's about a whole set of data contained in corpora which differs according to its media and formats – which anonymisation techniques depend on:

- primary video data,
- primary audio data,
- primary textual data: documents, either official or not, collected in the field,
- secondary data: transcriptions, field notes, metadata, analyses, ethnographic descriptions,
- visual secondary data: screenshots, voice representations (oscillograms, spectrograms etc.).

It should be noted that certain personal data isn't subjected to anonymisation: this is the case for people in the public eye performing actions of a public nature (for example, politicians on television) with full knowledge of the fact that their image is being broadcast and knowing that what they are saying is considered as a public speech. In this case, what they say – provided that it is considered as "original" – will be subjected to broadcasting constraints that govern copyright law with the exception of a set time period during which it will be considered as "current news". As soon as public speeches are no longer related to current news, they no longer come under this label[8].

[8] Cf. art. 122.5 in the law on the intellectual property code.

3.4.3 When should anonymisation take place?

Different times when anonymisation is carried out can be distinguished. Depending on the objectives and contexts of the survey, anonymisation can be considered to be needed either as *early* or as *late* as possible. The former ensures increased confidentiality of the people involved, the latter ensures maximum possibilities of analysis for the researcher. Time frames can also vary with data type:

- anonymising original primary data should be avoided as it could damage the data itself; however, non-anonymised data has to be kept in a safe place,
- data can/should/shouldn't (according to the principles followed) should be anonymised when it is deposited to be preserved; Institutions in charge of preserving data are thus answerable for it,
- it is possible to work (as part of a clearly defined research team which ensures that the data will not be distributed outside its boundaries) on data which hasn't been anonymised and to ensure however that any extract used in a paper or an oral presentation will be anonymised,
- anonymisation is always carried out on copies that will be distributed between researchers who are not part of the project and sometimes on copies that will be distributed between researchers involved in the project (as it is the case for large research consortiums or for research projects involving important team networks).

3.4.4 How should anonymisation be carried out?

Anonymisation methods concern both data media and formats and thus bring into play technological possibilities and constraints; they also concern symbolic forms and representations of people's identities and thus bring into play analysis issues.

Forms or elements concerned by anonymisation

As we will see, it is difficult – or even impossible – to establish a finite list of forms concerned by anonymisation. The main forms can nevertheless be highlighted:

- nominative forms (surname, first name, nickname, company acronym etc.),
- personal data (address, phone number, passport number, bank account number, age, place of birth etc.),
- profession, status, titles,
- social activities,
- blood relationships, networks,
- reference to places (toponyms, institutions, departments etc.),
- reference to a person's characteristic features (physical, cultural, medical etc.) which are unique or rare within the specified environment,
- physical characteristics: voice, face, bodily features,
- etc.

The "etc." that ends this list highlights the fact that any element, depending on the context of the recording and of its reception, can carry information about people's identities. Specifying the forms concerned by anonymisation implies the fact that researchers should have sociological and cultural skills which would put them in a

position to imagine the uses, knowledge and associations which could lead to identify someone from a specific form.

FORMS OF SUBSTITUTION

Once they have been identified, forms which can help identify people have to be transformed as part of the anonymisation procedure.

It should be pointed out that the most radical anonymisation method consists in purely and simply *deleting* data, even though other methods are often sought to ensure a better preservation of the data in the anonymisation process. It should be noted, however, that data can be partially deleted (it is possible to consider destroying extracts containing too many problematic and confidential elements for them to be used in their original form).

The anonymisation method that is usually adopted consists in *replacing* confidential elements with neutral forms. These vary according to the technical media concerned: a difference will be made here between text, audio and video.

- *Text*

The texts concerned are firstly the transcription and any extract from it used in research papers, exemplars, classes, conferences etc. Another type of text which needs to be anonymised is primary textual data (documents collected in the field). This data can take the form of texts or pictures (this is the case for letters, administrative documents or manuscripts kept as photocopied or digital files).

The principle of substitution consists in making the replaced portion of text visible and in giving general information about it (such as its duration).

- *Replacing text with a "blank":* this solution is the least informative and, above all, the least visible.
- *Replacing text with a hyperonym* or an abbreviation such as NN or NTOWN or NHOSPITAL for name, name of town, name of hospital etc. This solution can remain informative (the type of reference of the anonymised form is specified). It is useful in cases where using pseudonyms of substitution (cf. *infra* here) is impossible, difficult or not plausible. This solution requires specific conventions for hyperonyms to be developed, as the text they replace is of a different nature (this is why it is sometimes suggested to use capitals as long as they don't interfere with other uses of capitals specified in the transcription conventions).
- *Replacing text with a pseudonym*: this is the most commonly used solution, at least for people's names, as it allows the form of substitution to be well integrated in the course of discourse, it is not conspicuous, it is plausible and it maintains some of the indications included in the original form. However, it is only possible if the choice of pseudonyms has been well thought out and allows the following problems to be solved: pseudonyms should be chosen in the same paradigmatic field as that of the form it replaces (for instance, "Ahmed" could be replaced with "Moustapha" rather than with "Albert" in an attempt to maintain ethnic characteristics), in certain cases, especially if the recorded

interaction makes it relevant, potential connotations of the name should be maintained (for example, if it is a source of jokes or puns), as should be the number of syllables and certain phonetic and prosodic characteristics (if they are used in the interaction); pseudonyms have to be chosen in a way that makes it impossible to reconstitute the original name (in this way, choosing a pseudonym that starts with the same letters as the original name should be avoided, even if it makes it easier to memorize); pseudonyms should be chosen in a way that doesn't ridicule people (in this way, pseudonyms making direct reference to people's characteristic features should be avoided – for example "Mr Chubby"); street names, telephone numbers etc. can be replaced in the same way as people's names.

It should be noted that it is easier to choose pseudonyms for people than for town names (it is possible to imagine names for small towns, neighbourhoods or streets but it is a lot more difficult for big cities or capital cities); pseudonyms for names of institutional departments or services can sometimes be considered but not always (it makes no sense to replace "department of surgery" with "department of dermatology" when in a hospital). In cases where choosing a pseudonym is either difficult or not plausible, the solution consisting in using a hyperonym can be resorted to.

- *Audio*

– *replacing audio with silence*; the drawback of this solution is that the form of substitution can be mistaken for a pause,
– *replacing audio with a beep or another sound* which can't be mistaken for any other signal included in the recording,
– *replacing audio with a filtered or deformed version of the original signal;* this technique is mainly used in the media to make voices impossible to identify; when it is used by non-experts, its irreversible nature can raise problems (possibility to restore the original signal).

- *Picture*

The pictures concerned are the dynamic ones part of video recordings. But fixed pictures can also be considered, like photographs from documents or screenshots from transcriptions. Anonymising visual representations of audio streams (in spectrograms, for example) can be considered in cases where the pronunciation of a name or number could be identified.

– *deleting* this data can be considered by cutting it out when editing; in this case, it is advisable to note down the length of the deleted segment on the tape and to avoid giving the impression that the stream is continuous;
– *replacing data with a jammed signal:* using blurring, pixelating, edge detection or applying other types of filters (these types of image processing can be applied to the *entire picture or part* of it only); in the latter case, the technique is more complex when the part concerned is in motion;

– hiding people's eyes with a black strip.

3.4.5 Limits of anonymisation

Even though anonymisation is a fundamental procedure to ensure the legal distribution of data, one should remain cautious as to what can be promised and guaranteed to the surveyees in terms of data anonymisation.

Limitations are mostly of two very different natures: the first one deals with the contexts which increase or decrease the risks for people to be identified, while the second one deals with the constraints that anonymisation induces on the very objects of research.

Production and distribution contexts

Anonymisation is relativised by several factors coming into play either during data production – and depending on the specific events occuring while recording – or during data reception:

– Anonymisation is first carried out on a series of forms that are supposed to contain the main information making it possible to identify people. Nevertheless, any reference or form can potentially – depending on the context – lead to identifying people and very often in a way that originally goes unnoticed by the researcher. Thus, for example, a rare detail mentioned during the interaction (a person's rare pathology, an exceptional characteristic well known in the person's region etc.) can turn out to be revealing for some people (in certain cases, without the researcher even realising it).

– The recognizable nature of these details crucially depends on the context of reception and more specifically on the people who will refer to or peruse corpora. In this way, the members of the department of anaesthetics will easily recognize one of their colleagues based on typical expressions, specific areas of expertise or characteristic ways of expressing themselves or acting. On the other hand, the same details will go unnoticed among the staff of another hospital or even more so among students in linguistics. Yet, here again, the recognizable nature of these elements doesn't simply depend on the geographical or social distance of the context in which the data was recorded: people move in space and in social environments making it possible for a patient's son to identify his father in a university lecture on therapeutic consultations. The identifying potential of a detail thus depends on the data's context of reception.

– Depending on the cases, references to institutions or organizations can sometimes make it necessary to anonymise data: for example, references to a well-known shop should be anonymised if it refers to an employee's workplace, can be left in if it only refers to an element which is part of an itinerary, and should again be anonymised if it appears in defamatory statements.

– Other aspects are related to *cross-referencing* information coming from different sources (as could be the case for the link between anonymised data and metadata).

PRACTICES OF ANALYSIS

Anonymisation's limits can come from other types of considerations which are more related to researchers' practices of analysis.

The fundamental problem is raised by the possible contradiction between anonymisation and availability of details for the analysis (see Mondada, 2003 on the principle of availability). Indeed, recordings and transcriptions aim to make observable details available to be used in the analysis; on the contrary, anonymisation can make some of these details unavailable as they have been deleted or modified.

This can be the case, for example, when anonymising a name using a beep as it is said while overlapping another speech turn, which makes the analysis of the overlap impossible.

This can be the case when anonymising phone numbers in emergency calls, which makes the way the caller gives their phone number in a stressful and emotional situation unavailable and can thus have a crucial impact on this piece of information.

This can be the case when anonymising faces in a video, which makes it impossible to analyse looks.

In the same way, voice filtering (as it is done in the media) is not an option in most linguistic surveys which are based on the intrinsic qualities of the sound signal?

This is why researchers often stress the necessity and ask for the right to work from data which hasn't been anonymised, while guarateeing its safety and inaccessibility, to keep it as such and to carry out anonymisation procedures as late as possible and in a way which takes into account what is relevant for the analysis.

3.5 TRANSCRIPTION

Transcription is a practice which doesn't simply come down to a technical exercise of reproduction but rather involves numerous theoretical and interpretation stakes (Ochs as early as 1979). When moving from oral to graphic and visual data, many categorisation procedures are carried out related to either linguistic forms which are visually broken down into units (Blanche-Benveniste & Jeanjean, 1987; Mondada, 2000), or to the speakers' very own identities (Mondada, 2003). From the point of view of the recorded surveyees' image and identity protection, these effects have to be taken into account in order to avoid overinterpretation, stereotypical interpretation (Jefferson 1996) and stigmatisation of the speakers and the way they talk. Only these stakes of the transcription process will be covered here; in the following section, a completely different aspect will be dealt with: the issues of transcription standardisation and conventions.

3.5.1 ETHNOGRAPHIC DESCRIPTIONS

Transcription is often supplemented by a brief ethnographic description outlining the context in which the data was collected as well as the type of activities and the

participants' identities. This description, which includes elements from the corpus metadata, can have different repercussions on the way the corpus is read (or received in the case of an oral presentation):

- It can contain information allowing people to be identified, which goes against the principles of anonymisation.
- It can include information which influences the way the data is read or interpreted. When these indications refer to the fact the surveyee belongs to a certain category or to some other relevant dimension, they can give a particular image of the activity and of the speakers.
- In particular, the description can include references and can allow certain stereotypical interpretations to be inferred (or even to be used to add humour in order to win over an audience, which is quite common in oral presentations).

These remarks don't only concern data description, but also copus names which can sometimes include confidential information. In this way, even though they act as reminders, speakers' names should not be included in the name of the corpus.

3.5.2 IDENTIFICATION OF THE SPEAKERS

Transcription should include the results of the anonymisation process. Where annotation includes a code for turns, parts of the transcription can be allocated to different speakers identified in different ways. Using pseudonyms is common practice, but other alternatives exist even if they have various effects on the way the text that follows them can be interpreted. Whatever the choice that will be made, it raises the issue of the way the speaker will be dealt with. For example:

- A, B, C…: this is the least connoted option which nevertheless puts speakers in a particular order (as first, second, third etc.) by simply using the letters of the alphabet.
- P1, P2, P3... (for pupils): this solution makes the people in the same class homogeneous as they are referred to as part of a unique category. The same principle applies to S1, S2, S3 where S refers to the speaker: if linguists consider that all speakers are equal and that social actors are first and foremost interesting as speaking beings, from the point of view of the activity in progress, they participate in the first place as part of other categories, whether as interviewer/interviewee, father/son, doctor/patient, etc.
- M, F (for male and female): here again, this option gives preference to the sex/gender category over any other category, thus making this category overall relevant to understand the activities in progress.

These considerations prompt us to wonder what interpretive effects the choice of identifiers lead to. From this point of view, it is advisable to wonder what identifiers are relevant to the participants, especially in the case of analytical procedures which put emphasis on the participants' perspective (like conversation analysis). That's why the following options can be good alternatives:

– EVA, MAR, ROB, AND…: mention of the first three letters of participants' pseudonyms, be it first names or surnames, depending on the tone of the conversation,
– CAL/OPE for caller/operator, DOC/PAT for doctor/patient, or INTE/VIEWER for interviewee/interviewer when the institutional activity is governed by such category pairs. On these issues, one can refer to H. Sacks's considerations about categorising people and about the relevance of categories according to the activity and context in progress (someone who is a doctor in one particular context can be a father in another; the way he is identified thus depends on the activity in progress) (Sacks, 1972, 1992).

As the notions of privacy and intimacy don't have the same meaning in all societies, researchers will have to inquire about the speakers' positions as regards data anonymisation. In certain communities, not mentioning people's names is considered to be a lack of respect for the author of the story or for the people involved in it, whereas in other communities, mentioning them is a breach of privacy right. There are important differences on this point, for example, between certain fields in Africa (where speakers want to be mentioned) and fields like those in Amazonia, especially French Guyana.

3.5.3 STAKES

When transcribing, decisions are constantly made as regards the way speakers and their ways to express themselves are represented. Thus, analysis (and sometimes judgement) becomes instantly part of the transcription practice. A few of the stakes of the choices made in the transcription itself will now be discussed.

(ORTHO)GRAPHICAL STAKES

For over twenty years many discussions have taken place about the use of standard orthography, non-standard orthography and IPA in transcriptions (see 2.1.3). Phonetic transcriptions (IPA or others) can only be read by specialists in the case of short texts only. Thus, for larger copora, many European linguists have agreed to the use of standard orthography and, at the same time, have suggested to make it possible to import other notations when specific phenomena have to be observed in more detail.

On the contrary, in certain fields like phonetics, transcription using standard orthography can sometimes be irrelevant (for example, when transcribing logatoms, pseudowords, etc.).

However, the written representation of language often surprises speakers who can even dislike it a lot. They sometimes reject the representation of their language as it is shown in the transcription, disown the researcher and reject their work.

Representation of exolingual speech

Choosing to use IPA to transcribe certain passages or only those spoken by certain speakers rather than others allows greater accuracy in the representation of the details of their speech but can also cause uncontrolled asymetry effects.

Thus, resorting to IPA and non-standard orthography can have stigamisation and asymmetry effects for "non native" speakers when they are represented in a different way to that of "native" speakers (the latter using standard orthography, the former using special orthography highlighting not only their difference but also their "abnormality" and their "non normativity").

In a similar way, explicitly mentioning (as stated in the conventions) the speaker's language variety (differentiated through the use of fonts, styles and alphabets which are specific to the different languages used in a bilingual conversation, or to the learner's specific interlanguage in an exolingual conversation) implies carrying out a pre-categorisation process of this variety: however, this variety often happens to be an element negociated by participants and changes throughout the conversation (where certain forms are sometimes identified as "foreign" or "strange" and where, at other times, their difference is not at all taken into account.

The same questions arise when it comes to translating the transcription:

- translating certain speakers' words rather than others' can be considered as a value judgment;
- the way we translate, more or less litterally, can lead to a depleted version of the speaker's speech and to delete or, on the contrary, emphasize the difference;
- there are different formats for translations (as a line by line note following the original; or as an equivalent to the original form in order to maintain a quasi-litteral link to the original, in order to provide a grammatical gloss) which all give a different image of the other one's culture and language (Traverso, 2003).

It should be made clear that these are translations in the particular context of oral corpora. Translation is necessary to work on languages other than French but it is often only a tool for researchers and, in this case, it should not strive to be the reflection of the speaker's words. It should be supplemented by metalinguistic information which enables a better transcription of the nuances necessary for an in-depth analysis of the language to be carried out. Thus, if a bilingual corpus is to be made available, real translation work should be considered from a completely different angle to that of data collection in order to analyse the language.

Stakes of multimodal and detailed transcriptions

Choosing to transcribe verbal activities exclusively and to leave out other communicational indications, as it is currently the case in most transcriptions, can give an absurd image of certain speakers' behaviours. This can be the case for aphasic speakers or for children using alternative means of communication to the standard linguistic means. Not taking into account all of the resources used by these speakers means that only a partial image of them is given, which makes their behaviour pathological or abnormal.

In the same way, a transcription's various degrees of granularity (Jefferson, 1985) can be detrimental to the representation of unconventional behaviours (for example, an aphasic patient's vocalisation can be significant and require appropriate transcription; but it can also be reduced to a mere "sound" which becomes meaningless in a superficial transcription).

The more or less detailed or in-depth nature of a transcription doesn't only meet scientific requirements; it also meets ethical and judicial requirements which enable the speakers' image to be made more subtle and complex by distancing itself from the risk there is to caricature or stigmatise speakers through stereotypical behaviours.

4 ORAL CORPORA: NATIONAL HERITAGE OBJECTS?

A SOLUTION FOR THE PRESERVATION OF AND ACCESS TO ORAL CORPORA?

4.1 A REMINDER OF THE SITUATION

ORAL CORPORA DEVELOPED BY RESEARCHERS WITHIN INSTITUTIONS

The recording of oral copora has a history dating back to over a century to which the ability to capture the voice has given a new and singular dimension. As early as 1896, scholars, researchers (anthropologists, ethnomusicologists, linguists) recorded their collected data onto cylinders and then onto disks. Researchers were aware of the fact they were creating new collections to pass on to future generations and the productions recorded during "ethnographic missions" naturally found their place in institutes under the aegis of the state. The *Archives de la Parole*, a conservatory of the languages and dialects of France, were created within the University of Paris in 1911, the sound archives of the *Musée l'Homme* in 1932, the *Phonothèque Nationale* (National Sound Archive) in 1938 and it later became part of the Audiovisual Department of the *BnF* (French National Library) in 1977. Large ethnographic collections carried out by the *Musée National des Arts et Traditions Populaires*[9], as well as the *Centre d'Ethnologie de la France* (ethnologic centre of France), deal with, for example, Brittany in 1939 and the major pluridisciplinary survey on the *Aubrac* territory which, between 1964 and 1968, produced about four thousand phonograms and a dozen films. Linguists and then ethnologists were the ones who thought about the future of their recordings as a priority, including their use by other researchers. In the 1970s some sociologists like Daniel Berteaux[10] introduced "life stories" in their methods. This opened the way for pluridisciplinary research, among which "ethnotexts" were one approach experimented by Jean-Claude Bouvier and Philippe Joutard.

FRANCE IS NEVERTHELESS A COUNTRY WITH A WRITTEN TRADITION WHERE THE SPOKEN WORD DOESN'T HAVE ANY CULTURAL VALUE, AND EVEN LESS A NATIONAL HERITAGE STATUS.

Universities haven't therefore developed any specific critical methodolgy adapted to the specificities of oral speech. The absence of standard vocabulary to define the different forms of oral corpora is very telling as regards the absence of a real scientific and national heritage status of oral corpora. Each discipline uses its own terminology and gives it a particular meaning. Claude Martel[11] mentions the variety

---

[9] The *MNATP* became the *MCEM (Musée National des Civilisations de l'Europe et de la Méditerrannée)* in June 2005.

[10] Daniel Berteaux, « L'approche biographique. Sa validité méthodologique, ses potentialités » *Cahiers internationaux de sociologie*, 1980.

[11] Claude Martel « la recherche et les sources orales, les mots pour le dire » in: *Bulletin de liaison des adhérents de l'AFAS* 10, 1998.

of definitions known for terms such as life stories, testimonies, interviews depending on the field the person using them is involved in.

Historians were reluctant for a long time to consider oral testimonies as a reliable source and worthy of consideration. Philippe Joutard, one of the promoters of oral history, mentions France's isolation compared with European countries such as Great Britain, Italy, Spain, Argentina who saw, within their universities, a dynamic and expanding development in this field. This vitality is evidenced by many journals (see bibliography).

The outstanding survey carried out between 2001 and 2003 by Françoise Cribier and Elise Feller at the request of the Ministery for Research, has proved that in the last thirty years, French researchers, in all humanities with the exception of History, have recorded a tremendous amount of data. However, their recordings had no official recognition and no place to be stored and have thus been confined to laboratories. Above all, they have been neither described nor documented, and the interviewees' authorisations whenever they exist are, at best, to be used only by the researchers who carried them out.

Collections are often privately stored and owned as, most of the time, oral data carried out as part of official recording campaigns are an embarrassment to authorities. In this respect, the major venture carried out by the *DGRST* at the beginning of the 1960s in the Plozévet region, a Bigouden village, is altogether exemplary. This large-scale survey, which was carried out by the *Musée de l'Homme*, and lasted for nearly five years, mobilized historians, geographers, sociologists, economists, and ethnologists. Many of them were equipped with tape recorders. However, this survey didn't produce a pluridisciplinary work but a set of monographs and no-one cared about the recordings that were made, except for those made by ethnologist Donatien Laurent. He is one of the few researchers who not only documented his whole collection, but also deposited it in the Centre of Research and Celtic and Breton culture of Brest University. Today, his recordings are digitised and can be consulted in the University. The other recordings have been lost or, due to lack of financing, the tapes have been re-recorded.

THE DIGITAL NETWORK AGE: THE 1980S

Additonally, these collections without a scientific status raise, for some of them, unresolved judicial issues with regards to their preservation and their consultation.

Analog sound documents can only be consulted in real time. Their indexing isn't always enough for them to be consulted quickly. This task discourages most researchers.

In the 80s, digitisation techniques marked renewed interest in the spoken word, sensitive data with often unique content. Indeed digital recordings, indexed by researchers themselves at the time of recording, allow the sound document to be quickly skimmed through as can be done with written texts.

But if digital techniques have revolutionized access to oral corpora, as was the case for written texts and pictures, they have introduced another intellectual revolution, due to the perfect nature of the copies made, which is far more important for later use. *By making the notion of "original copy" irrelevant, they have obliterated the markers* which

up to this point stretched along the field of collections. As they are deposited by their builders within a national heritage institution, oral corpora become objects part of collections but it is then impossible to distinguish between the first recording considered as the "original" and the subsequent copies of an oral corpus.

As the recording medium doesn't allow the different elements to be identified, who is to select and freeze Time "T" of the version which will be evidence of the builder's research on entering an institution? What type of *metadata* will be integrated simultanuously into collections?

ORAL COLLECTIONS WITHOUT STATUS

Oral corpora are not mentioned in the Law on the Intellectual Property Code as protected works, unless they have an identified form and, as such, are protectable: *testimonies, interviews, discussions, radio broadcasts*.

Overall, oral collections, and the sound dimension in general, are not taken into account in the large-scale cultural venture launched in 1964 by André Malraux: *l'Inventaire général des monuments et richesses artistiques de la France*. None of the devices which lay the foundations of a national heritage are attributable to them. There is no classification or inscription and, as a consequence, no commission "specialised in national heritage" is concerned with them. Only UNESCO has taken initiatives in this way (see *UNESCO appendix*). More modestly, the *Mission du Patrimoine ethnologique* created in the 80s within the Ministry for culture and communication was to rank oral corpora among objects. This preoccupation has very quickly disappeared from the programmmes.

4.1.1 ORAL CORPORA COLLECTIONS

PRACTICES AND USES IN NATIONAL HERITAGE INSTITUTIONS

As oral corpora solely built by individuals or by institutions aren't considered as a particular category with regards to patrimony and the Law on the Intellectual Property Code, the law doesn't make allowances for a particular procedure to collect them and see to their preservation.

Universities didn't show any interest in this rich and profuse set which emerged from disciplinary fields that were too diversified. *Therefore no registration of copyright exists for oral corpora.*

Oral corpora can only be protected in a national heritage institution through *voluntary initiative* (gift or deposit) of the person who collected them or through *a decision from the institution* concerned with putting together oral collections about themes which are their own. National heritage institutions can therefore be, either at the same time or successively, builders of oral corpora and "curators" seeing to the preservation of oral corpora built by others. Institutions responsible for this type of collection carry out research on the preservation of sound documents.[12] They also make use of selective criteria to put collections together.

---

[12] Calas, M.-F. Fontaine, J-M. (1996) *La conservation des documents sonores*, Paris: CNRS-Editions.

- Generally speaking, the *collection coherence principle* governs the constitution of collections within national heritage institutions (archives, national heritage libraries, museums). An isolated recording is only significant for itself. The unique recording of the voice of a writer in a museum which is dedicated to them is anecdotal.
- This means that putting together a coherent collection results from *a rigourous selection policy* following the main lines defined by the institution (spoken collections for the *BnF*, collections on deportation for the National Archives) which are nevertheless broad and comprehensive enough for them to make up significant sources of reference for the future. In social history museums, inheritors of the ecomuseums defined in the 1970s by Georges-Henri Rivière, collecting oral surveys aims to make up for the absence of objects or their difficulty in showing the human dimension within a collection. In Fécamp, the recording of female workers in old fisheries is a testimony of a form of social organisation in the city in the first half of the twentieth century, which no object or piece of writing could reveal[13]. This is the same in the Museum of Tobacco Manufacturing in Morlaix, the Ecomuseum of the urban community of Creusot-Montceau-les-Mines (Saône-et-Loire).
- The collected data is not always considered as a collectable object or as a work of art. In the BnF, in the National Archives, cataloguing isn't determined by the medium of the collection. This is not the case in Museums. With the exception of the *Musée National des Civilisations de l'Europe et de la Méditerranée* (formerly known as *Musée national des Arts et Traditions populaires*) and of the *Musée Dauphinois* which very early on integrated Charles Joïsten's surveys on the Museum inventory in the same way as the objects, most museums, such as the Salagon Museum-conservatory, have integrated oral corpora into their library-type inventories. In the same way, the Saint-Quentin-en-Yvelines ecomuseum, has opted to include the interviews it carries out with political figures and inhabitants on a seperate register which lists survey collections. At the end of the 90s, a strong – even excessive – interest was shown in the quest for identity and the duty of rememberance. These oral archives are still not given the consideration they deserve.
- The collection of oral data can't simply be reduced to voice recordings. It only makes sense when the temporal, technical and scientific data about its development is made available. All of these elements of contextualisation (metadata), which are specific to the recorded corpus, together with the recorded corpus itself, make up an

[13] This series of interviews was carried out in collaboration between the Museum and the town's archives department and was released as a disk and booklet entitled *Femmes de marins, compagnes de pêche*, Fécamp, Musée des Terre-Neuvas, 2003.

inseparable whole, and without them, the recording wouldn't have any sense or temporality. It could then be interpreted in many different ways.

- As with any heritage object, an oral document, even though it is dated and identified, cannot, as many researchers believed for a very long time, be reduced to its producer's sole use. Oral surveys often go beyond the project they were carried out for. They can be used in the framework of other fields

> *"A new reading entails what was said to be looked at in a different way, because time has gone by and the questions that we once asked ourselves have shifted"*[14].

- It should be possible for different researchers to analysize them through time with their own grid. However, the audio document digitisation scheme put into place by the Minister for Culture and Communication at the end of 1999 highlighted the lack of information about these oral collections. Certain collections, considered as historical, couldn't arouse a great deal of interest due to the absence of certain documents necessary for their contextualisation. Moreover, none of the collections having answered the call for digitisation held the exploitation rights allowing them to oraganise public consultations, in particular via the internet.
- What type of protection do oral corpora in private and public institutions benefit from? Depositing collected data within an institution doesn't have any **evidentiary value*. Can the deposit date possibly show evidence of its precedence against a recording which would turn out to be a counterfeit? With the exception of the deposits which are inherently always removable, collected data enters the institution's collection (in the form of hard or digital copies) in a definitive and imprescriptible way. This transfer, as aforesaid, does not entail the cession of the exploitation rights. Institutions commit themselves in theory to assuring the physical durabililty of oral corpora and to organising their consultation while respecting the rights of those who have played a part in their construction, but transferees have to at least agree to transfering their consultation authorisations. Since the 80s, consulting collections from a distance has in a way "aroused" interest in oral corpora, and has enabled possibilities of access which were unimaginable before to be considered. The accessibility of analogical corpora raises the issue in terms of conservation and identification of the sources available prior to their digitisation. This problem also has a cost in terms of financial and human resources.

[14] Françoise Cribier & Elise Feller, *op. cit.*

4.1.2 THE BIBLIOTHEQUE NATIONALE DE FRANCE

A CONSERVATORY OF ORALITY

Following in the footsteps of the *Archives de la Parole* founded in 1911 by Ferdinand Brunot, of the *Musée de la Parole et du Geste* which replaced them in 1928, then of the *Phonothèque nationale* (national sound library) created in 1938, the Audiovisual Department of the *Bibliothèque nationale de France* aims at pursuing the actions initiated by these institutions. Today, over a century's worth of orality's history is therefore preserved and made available to the public.

But, at the same time, the Audiovisual Department follows an active policy for the development of its collections, in particular in the field of orality. Indeed, besides collecting legal deposits (see appendix *Bibliothèque nationale de France*), the *Bibliothèque nationale de France* has made it its mission to enrich its collections through purchases, donations, gifts, legacies, payments in kind etc. It is therefore the case of the Audiovisual Department which, as a complement to the legal deposit of sound, videographic, multimedia and computer documents it is in charge of, has defined the outlines of its collection enrichment policy with original sound recordings. At the end of this document, the reader will find a few of the collections which have recently been added by the Audiovisual Department and which are representative of the position of oral documents in its collections.

PRINCIPLES FOR THE ENRICHMENT OF THE AUDIOVISUAL DEPARTMENT'S COLLECTIONS

The Audiovisual Department considers documents as "unpublished" documents which are "original" and "unique", which haven't been distributed in large quantities and which aren't determined by a specific editorial form. With this in mind and faced with the undetermined boundaries of the field, the multiplicity of contents (linguistics, ethnology, oral history…), the multiplicity of possible sources (institutions, independent researchers…), the necessary complementarity with other institutions, but also the void to be filled as regards preservation, distribution and promotion, the Audiovisual Department has outlined a certain number of leading principles which will serve as guidelines for its enrichment policy in this field.

THE DOCUMENTARY AND NATIONAL HERITAGE CRITERION

The policy of the department rests first and foremost on the principle of selection. The fundamental criterion which leads to the acceptance or refusal of a donation of unpublished material is above all the documentary and/or national heritage interest of the proposed collections. This criterion can be considered as akin to "national memory". In other words, which unpublished recordings can be considered to be a matter of memory or national heritage? This criterion doesn't restrict the domain of documentary policy to the French "field", but gives priority to collections having a connection with France – either in terms of source (the collector, the institution…) or in terms of contents. The donation of Deben Bhattacharya's collection, an Indian ethnomusicologist who has made recordings throughout the world and who lived in Paris from 1954 to 2001, or the donation of Simha Arom's pygmy collection (Lacito-CNRS), are good examples of this.

In close conjunction with the documentary/national heritage interest criterion which sets precise limits, the Audiovisual Department gives particular attention to documents or collections which have no set place of conservation and/or consultation. This is the case, for example, of certain personal archives or of collections in escheat in certain laboratories due to a lack of appropriate structures.

ACCEPTABILITY OF A COLLECTION AND DOCUMENTARY PRINCIPLE

Once this selection criterion has been established as well as the documentary and national heritage criteria, conditions of acceptability are set out when considering a collection. It is firstly a matter of documentary conditions. In this way, to be acquired for free or purchased, unpublished sources should be documented or exploitable from a documentary point of view. It is either possible for document processing to be provided as metadata at the same time as the archive, or for all the information to be given to the *BnF* in one form or another to allow documentary processing to be carried out.

ACCEPTABILITY OF A COLLECTION AND LEGAL PRINCIPLE

Legal aspects make up another component of the conditions of acceptability. The person – a natural person or legal entity – who carries out a donation should notably make sure that:

- He/she is the owner of the physical media on which the recordings have been made and that these recordings are able to be given to the library;
- That he/she owns or vouches for copyrights on the works carried out and the neighbouring rights of the phonogram producer and possibly of the music players or singers.

When receiving physical media, the *BnF* needs to have the copyrights and neighbouring rights required to copy them and make them available to the public, as sound documents will be subjected to copying and public performances within the context of their preservation and consultation. However, the person – a natural person or legal entity – who makes the donation can't always legally give the authorisations for reproduction and communication.

The following rights should be transferred to the *BnF*:

- The right to reproduce the document, that is to say the possibility to transfer its contents onto appropriate digital media in order to preserve the signal;
- The performing rights. These are understood to be, at the very least, the possibility given to a public of researchers to consult the documents in room P (on the "research" floor of the library). Authorisation for communication on a case by case basis should be considered. In the same way, for certain documents a delay for reserved communication can be required for reasons other than those linked to copyright law (confidentiality of data related to privacy…).

SOME EXAMPLES AMONG THE LATEST UNPUBLISHED COLLECTIONS DONATED TO THE AUDIOVISUAL DEPARTMENT

(listed in order of their integration in collections):

- The regional linguistic altas collection (1979 and after);
- The Historical Research Centre collection, EHESS/CNRS (1979): oral history, life stories from the years 1970-1980;
- The Félix Quilici collection (1981): Corsican music with an oral tradition, 1959-1963;
- The Geneviève Massignon Collection (1985): ethno-linguistic data, Acadia, Western France, Corsica…, 1946-1963;
- The Nicole Revel Collection (1995): Palawan epics, the Philippines, in the 1980s;
- The Gilles Deleuze collection (1997): lectures, Paris VIII University, 1979-1984;
- The Deben Bhattacharya Collection (2003): ethnomusicological data from Asia and Europe…, 1954-2000;
- The Archiving programme, LACITO/CNRS (2005): rare languages, transcriptions, annotations, http://lacito.vjf.cnrs.fr/archivage/

4.1.3 FRENCH NATIONAL ARCHIVES (LES ARCHIVES DE FRANCE)

In the second book of the Code of Patrimony, archives are defined in article L 211-1 as follows:

"Archives are all documents, whatever their date, form or medium, whether produced or received by a natural person or legal entity, or by any private or public department or organization when practising their professional activity. The conservation of these documents is organized in the public interest, as much to meet the needs for management and justification of the rights of individuals and legal entities, whether public or private, as to ensure the historical documentation of research." Archives are made up of two categories: public archives, which originate in the activity of the state, local communities and public companies, and private archives (see appendix about *Archives: legislation*).

It is the development method and not the type of medium or the topic which determines which category the data belongs to. The recording of a regional council meeting is a public archive document whereas the recording of a political figure on the radio is a private archive document.

The consultation of sound collections depends on whether the archives are public or private. If the former follow clear regulations, the will of the person who deposits the documents sets the rules as far as private archives are concerned.

- *The Archives nationales (French National Archives)*

They are under the responsibility of the managing board of the *Archives de France* and they comprise five different centres.

- The *Centre Historique des Archives Nationales (CHAN)* (French National Archives Historical Centre) in Paris. It is within the 20th century section that a unit for oral archives was created in the 80s. This unit receives deposits, for example those made by the Foundation for the Memory of Deportation *(Fondation pour la mémoire des déportés)*, but it also produces testimonies to complement written archives "in saying what can't be written, in putting factual history on a human scale, and in filling existing gaps in history, if necessary, with narratives of concealed details"[15]. The same principle applies for video recordings in judicial archives (the Klaus Barbie trial, the Paul Touvier or the "blood conspiracy" cases) and the French Presidency's Archives, i.e. speeches and press conferences by Presidents Georges Pompidou and Valéry Giscard d'Estaing.
- Sources created by curators follow one of two approaches: autobiographical narratives are used to write the history of the elite, and thematic corpora make it possible to cross-reference several narratives dealing with the same fact or topic (for example, working as a Primary School teacher in the 50s).
- The *Centre des Archives contemporaines (CAC)* (Centre for Contemporary Archives) in Fontainebleau. For example, this is where the 400 hours worth of recordings carried out within the context of the programme initiated by the Committee for the History of Social Security *(Comité d'histoire de la Sécurité sociale)* chaired by Dominique Aron-Schnapper are deposited (refer to the "Status of archive collections" section...).
- The *Centre des Archives du Monde du Travail (CAMT)* (Centre for working life archives) in Roubaix which collects any type of archive in its field, including recordings.
- Of the other two centres, the Esperran centre only stores microfilms and the *outre-mer* archive centre mainly stores closed written collections.

- *The Services d'archives départementales (regional archives departments)*

These departments were decentralized long before others. They often collect copies of radio broadcasts, amateur films or documentaries and carry out oral survey programmes, either on their own or in collaboration with organisations and universities. They have a wide range of documents and the importance given to oral

---

[15] Agnès Callu, « Aux Archives nationales, une politique raisonnée en faveur des témoignages oraux » *Colonnes: archives d'architecture du XX^e siècle*, 20, décembre 2002.

collections depends on the topics covered and, above all, on the manager's motivation and interest.

The *services d'archives municipales* (municipal archives departments), involved in a memory *patrimonalisation* process, have often resorted to *emploi-jeune* contracts (5-year contracts for the under-thirties) for the collection of oral archives, thus employing young workers as "memory keepers" (as, for example, in Martigues and Lille).

4.1.4 THE PLACE OF ORAL CORPORA IN MUSEUMS

Museums are, in the wider meaning of the word, all permanent collections consisting of items whose preservation and dissemination are of public interest and organized with a view to the public's knowledge, education and pleasure.

Collections comprise any type of object and work with tangible materiality.

Oral recordings, by definition, are considered as immaterial by museums. However, ICOM (International Council of Museums), a non-governmental organization which sees to the development of all types of museums within UNESCO, opened the debate about the immaterial nature of the intangible heritage. The fact that western museums in general feel uneasy when it comes to integrating the sound, audiovisual and landscape dimensions into their collections perfectly brings to light this sort of contradiction between objects and orality for museums.

On the other hand, history museums, ecomuseums and social history museums have been using, sometimes for many years (as in the case of the *Musée dauphinois* in Grenoble), the recording of oral memory as one of the main elements of the cultural and scientific project the museum is centered on. Oral collections are entered on the museum's inventory register, just as any other collection, but this is far from being common practice and so many sound and video recordings are, at best, entered on the register for survey or documentary collections.

If oral corpora were considered, as in the *Musée dauphinois*, as works entered on the inventory register whose writing procedure is determined by legal texts, they would be inalienable and imprescriptible.

4.1.5 "ORAL CORPORA" AT THE INA

In allowing the public to consult legal deposits for radio and television broadcasts, the *Ina* gives people access to a wide variety of oral corpora comprising various testimonies, spoken words and speeches recorded with a view to being broadcast.

Researchers who use the consultation centre of the *Inathèque* put together different corpora, with their specific needs in mind, from the sources of radio and television broadcasts, and these corpora will then be used along the lines of one specific field: linguistics, sociology, history…

Corpora can be studied according to the discourse strategies at work in a particular type of broadcast (TV interviews, comments on the radio…), different types of discourse analysis (political, journalistic…), the creation of lexicons, sociolinguistic analyses (the dancer's speech, female workers' speech) etc.

Some of the broadacast collections which are stored at the *Ina* make up "closed corpora" of oral speech at once.

To mention but a few: the *"Archives du vingtième siècle"* produced by Jean-José Marchand, which is a collection of interviews of people working in the fields of literature and arts, the *"Conteurs"*, which is a collection developed by André Voisin and produced by the research department of the *ORTF* (Office for French Radio and Television Broadcasting), collections of personal stories, collections of regional stories (*Ceux de La Hague, Au cœur de l'Aubrac…*).

Besides, the *Ina* has been developing collections of national heritage recordings and testimonies.

These interviews, which vary in length (up to 15 hours long for some of them), can be accessed through an interactive consultation interface called "@propos" which makes it easy to browse the programme.

- Thus, the "*Musique Mémoires*" collection is based on an archiving campaign aiming to collect testimonies from composers, singers, conductors and personalities whose creations and actions have left their mark on the music scene in the last sixty years. These interviews, carried out by Bruno Serrou, offer to explore each artist's very own background: their origin, training, influences, the people they met, the way they practised their profession… The following people have already been interviewed: François Bayle, Claude Helffer, Betsy Jolas, Claude Ballif, Pierre Boulez, Marius Constant, Antoine Duhamel, Luis de Pablo, Yvonne Loriod, Michel Fano, Ivo Malec.
- "*Histoires d'historiens*" is a collection of contemporary historians' self-portraits; their lifestories thus told give a better understanding of their works. The following people have already been interviewed: Maurice Agulhon, Pierre Chaunu, Emmanuel Le Roy Ladurie, Claude Nicolet, Pierre Nora, Robert Paxton, Madeleine Rebérioux, René Rémond, Zeev Sternhell, Jean Tulard.
- "*Télé notre histoire*" is a collection of long interviews which offers television's memory as told by those whose personal path and professionnal practice shed light on this medium: writers, artists, producers, decision-makers, pioneers or more recent practitioners. The following people have already been interviewed: Igor Barrère, Marcel Bluwal, Yves Jaigu, Jacques Krier, Claude Santelli, Pierre Tchernia…
- Other interviews, which are not in line with a collection's principles, nevertheless offer testimonies from people who play a central part in contemporary cultural, scientific and artistic life. The following people have already been interviewed: Françoise Gilot, K.S. Karol, Claude Lévi-Strauss.

- "*Mémoires de la Shoah*" is a collection which is being developed and which comprises 110 3-hour-long interviews from people who witnessed the Shoah: deported people, orphans, "righteous".

All of these collections will eventually be accessible at the consultation centre of the *Inathèque*.

4.2 PRIVATE INITIATIVES

The recording of oral testimonies has been subjected to an outstanding development since 1972 (when the permanent committee for educational history was created) in the context of programmes put into place by the *Committees for Oral History* created by public institutions willing to promote the memory of their institutions.

Today, there are 67 committees and departments[16] part of an institution (the History Committee within the Ministry of Culture and Communication, the History Committee within the BnF).

The *AHICF*, *Association pour l'histoire des chemins de fer en France* (Association for the History of French Railways) is a case on its own. It offers its services to institutions whose history it intends to recount. The *AHICF* was created in 1987 and set itself two objectives: research and preservation of the national heritage. It promotes the preservation of sources but doesn't intend to carry it out itself. It offers free-choice services (historians) to help create memory in the industrial field.

Overall, these committees consider the recordings made as private archives protected by copyright laws. The devolution clause in the case of oral corpora, to the advantage of an archive department whenever organizations get disbanded, is common practice.

Among the active partners of an "oral archive" network, the following can be mentioned: the Associate centres of the BnF such as the *FAMDT, DASTUM*, the *MMSH Maison Méditer-ranéenne des Sciences de l'Homme* in Aix-en-Provence (cf. BnF, Associate Centres). These centres very rarely own the full rights on the corpora they preserve.

4.3 ACCESSING COLLECTIONS

There is no collective catalogue for oral corpora. Several initiatives have made it possible to identify the different departments within institutions or organisations involved in building or collecting oral corpora with a view to preserving them and making them accessible for consultation. However, these initiatives mostly describe corpora in a general way rather than give a detailed description of their contents which is due to the fact that corpus builders haven't described them at great length. A book[17] which was published this year arose from sorting through the answers to a

---

[16] Guide des Comités d'histoire et des services historiques. Paris, Comité pour l'Histoire économique et financière de la France

[17] Callu, A., Lemoine, H. (2004) *Patrimoine sonore et audiovisuel français: entre archive et témoignage: guide de recherche en sciences sociales*, Paris, Belin, 7 vol., 1 CD-Rom, 1 DVD-Rom.

large-scale survey about oral sources in social sciences preserved in France. It may become the starting point for the constitution of a collective source for orality if the electronic catalogue can be updated by the network of corpus builders.

The terms for the consultation of corpora are specified in the contract. There is, however, no standard contract.

Within institutions, the majority of oral recordings are considered to come under the law on the Intellectual Property Code. Generally speaking, witnesses have the right to examine the way their voice is used (law of 17 July 1970). Nobody has the right to fix, keep or disclose a natural person's words and image within a private setting without their consent. Article 9 in the Civil Code and article 226-1 in the Penal Code make it compulsory to obtain the person's written consent. Witnesses can be considered as authors whenever their words show originality thus making them liable to related moral rights. Using their recordings can be subjected to the obligation to give them payment as defined in their contract. The person collecting data should get permission for the data to be consulted on the widest scope possible.

> *Data accessibility raises questions regarding rights and ethics (respecting people's privacy, people's right to their own voice, life stories, sensitive testimonies, words that could become defamatory…). These corpora can therefore neither be consulted on site nor be made available on the internet due to the very nature of their content (life stories and testimonies implicating other people, interviews carried out in a psychiatric environment).*

Each case is therefore unique and recognizing people's rights requires a detailed and tricky analysis which should make it possible to answer the following questions: Who owns the rights? Will the owner transfer their rights? Under what terms? For what use? How long for? Will it take effect immediately or will it be deferred?

The collector-researcher, whose corpora recordings correspond to one step of their in-depth research, should have their rights as an author protected. In most cases, they are referred to as *collectors*. In order to grant copyrights to the interviewer, they would have to be able to bring to light the original nature of the points they make.

Consequently, institutions can only allow on-site consultation in their own premises, and any digitisation work they originate is often completed with no authorisation from the actual right holders (national digitisation programme).

Many problems still remain.

The issue of the paid collector who carries out their public duties and who is supposed to transfer their rights to the state highlights the problem of paid "authors'" rights, which comes up against unsolved financial issues in the public sector.

What should be made of the rights that some students, with no training in interviewing techniques, could claim when they are paid to ask a list of questions in the order of a pre-existent questionnaire?

The status of oral archive collections does matter.

The case of the first large-scale survey on the history of social security, carried out between 1973 and 1975 by Dominique Schnapper at the Committee for the History of Social Security's request, led to the recording of 200 witnesses who produced 400 hours worth of interviews and testimonies. By definition, these made up private archives. And yet, before the campaign started, it was decided that the whole survey would be classified as public archives and that, as such, it would be possible to consult it after a sixty-year delay. This decision had serious consequences. Philippe Joutard mentioned this case on several occasions as he considers it to be one of the possible reasons why the development of oral history in France is not very dynamic.

In the same way, Florence Descamps agrees with this analysis and condemns these innovative oral archives that were "frozen" from the start.

Researchers were reluctant to have their corpora institutionalised and thus kept them in their possession; they did not get any encouragement from institutions such as the CNRS and universities (except for the agreement signed by the CNRS and the BN in 1979 with a view to preserving linguistic atlases) which, until recently, had never taken any constructive initiative to preserve oral corpora which did not match any academic definition, whereas oral history developed profusely in Great Britain, its place of birth, as well as in Latin countries other than France.

4.3.1 WHAT TYPE OF NETWORK FOR TOMORROW?

A NETWORK FOR THE MANAGEMENT AND PROTECTION OF ORAL CORPORA COLLECTIONS RUN BY UNIVERSITIES AND RESEARCH INSTITUTIONS OR BY NATIONAL HERITAGE INSTITUTIONS?

Apart from national heritage institutions, universities and research organisms, following the example of many other European countries, might be able and willing to set up a wide humanities and social sciences network through which corpora made available to other researchers could be protected and made accessible to more.

In their Report[18], Françoise Cribier and her collaborator Elise Feller examined the situation and the existing networks involved in preserving and making available qualitative datasets in social sciences in six European countries. Two initiatives will be presented here as possible models for French researchers: Qualidata (Great Britain) and SIDOS (Switzerland).

Qualidata[19] was created in Great Britain in 1994. It is based in Colchester as part of the Sociology Department at the University of Essex. This initiative came within a university context which had been widely sensitized to the preservation of oral data, in particular through the study carried out by Paul Thomson at the ESRC's (Economic and Social Research Council) request. It could serve as an example. This service is highly selective when it comes to acquiring collections that were developed after 1995 (some of the selection criteria are: clearly defined themes, documented corpora, digitised high-quality sound documents whose legal characteristics are clearly determined).

This service also works from useful criteria with a view to a forthcoming secondary analysis. It is interesting to notice how involved it is in the training of researchers who are tomorrow's data developers.

It can thus be a way of having more control over research in specific fields and of avoiding doubles.

[18] *op.cit.*

[19] Qualidata, UK Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK. www.qualidata.ac.uk. See also Appendix 3 (Cribier, 2005).

SIDOS, a Swiss service for data information and archiving in social sciences created in 1992 by the Swiss Academy of Humanities and Social Sciences, is also a sort of agency involved in managing qualitative and quantitative datasets developed by researchers[20].

SIDOS considers data developers as authors and archiving as data and documentation editing.

Archiving aims towards sharing data with other researchers. It is a tool which makes scientific activity richer, provided that the data is then correctly distributed.

Setting up such networks would undeniably be interesting for research. We are not sure that the national heritage status and durability of these oral collections would be better ensured..

WHAT ORAL SOURCES FOR TOMORROW?

Since the beginning of digital recording, the issue of long-term durability has been a problem maily due to the quick obsolescence of standards and of system compatibility. However, the future coherence of collections is disrupted by data archiving methods. When depositing their collections, researchers carry out genuine editing work on their corpus and its documentation so as to always share douments that are understandable and coherent. This work should always be carried out by researchers. But when will they take time to do it? What image of their work will they be willing to give? Which form should they keep? What is the interest for tomorrow's researchers? There isn't a single answer.

Researchers who want to use oral corpora built by others need mediation, i.e. documentation which gives them information about the variables as well as the way the data was collected and the context of the project.

In the latter case, the researcher-developer isn't in the best position to describe their data knowing that it will be used by people who are not familiar with their specialist field. It is down to professionals in document processing, be it librarians, filing clerks or archivists, to describe corpora aimed at third parties *using standardised tools that anyone can understand.*

A description that is too precise, "*subsequent" or "retrospective" testimonies, "a posteriori life stories",* based on the notion of temporality, which would certainly be useful in the researcher's analysis, is not effective for the management of these collections within the institutions in charge of their preservation. These criteria are admittedly part of the objective description of an oral document, but it is not up to the institution to

[20] See the survey carried out by F. Cribier & E. Feller , *op. cit.* Appendix 3: 14-20.

classify them in overly limited categories, which is already a form of analysis and which restricts the prospective users' freedom by forcing a point of view on them.

In a nutshell, corpus developers are the only people who can document their oral corpora. It will be possible for them to be used by third parties only if the description is given by documentation professionals.

4.3.2 TOWARDS THE RECOGNITION OF A REAL STATUS FOR ORAL HERITAGE

The future of oral sources isn't solely a question of law. This aspect can be taken care of through pragmatic contractual solutions. This *Guide* has no other goal but to show it.

However, the real stake in the issue of oral sources is of a cultural and political nature. For them to be recognized, it is necessary to come up with strict selection criteria, without which no heritage worthy of the name can exist, and, at the same time, society has to be made aware of the fact that these scientifically-developed documents should be given *the status of heritage objects*.

Their integration in the system which governs heritage objects will then come naturally.

It should be noted that France is remarkably backward in the field of immaterial heritage.