Interoperability of audio corpora : the case of the French corpora

Olivier Baude[1], Michel Jacobson[2], Atanas Tchobanov[3], Richard Walter[4]

1. Laboratoire Coral
UFR Lettres, Langues et Sciences Humaines de l'Université d'Orléans, 10, rue de Tours, 45072 Orléans cedex 02
olivier.baude@univ-orleans.fr
2. Laboratoire Lacito
UMR 7107, 7 rue Guy Môquet, Bât. D, 94801 Villejuif
jacobson@idf.ext.jussieu.fr
3. Laboratoire MoDyCo
UMR 7114, Université Paris 10, 200, avenue de la République, 92000 Nanterre cedex
atanas.tchobanov@u-paris10.fr
4. Laboratoire MoDyCo
UMR 7114, Université Paris 10, 200, avenue de la République, 92000 Nanterre cedex
richard.walter@u-paris10.fr

Abstract

We present here the choices which were made within the framework of three oral corpora projects : Socio-linguistics studies on Orleans (ESLO), Phonology of the Contemporary French (PFC), The Archivage corpus of the LACITO lab. This comparative presentation of three corpora of audio linguistic resources comes from a analysis about the options the project have to operate to describe them for discovery purposes and to compare the contents. The aim is to illustrate the interest to think the interoperability and the methodology of codings and the metadata. Through this step, we want to simplify the technical creation of audio corpora and thus the constitution of linguistic resources, usable by enlarged academic and industrial communities.

Today, labs, institutions, and firms working on digitizing and diffusing audio corpora are much more numerous than in the past. This can be explained by, amongst other things, the maturity and low cost of digitization techniques. The difficulty of conserving analog audio has also led many institutions to digitize their data in order to preserve it. But preserving, sharing and exchanging data requires more know-how than to simply digitize or computerize it. One must create a corpus carefully, with considerable attention to the question of how to diffuse the data.

The multiplicity of audio corpora in linguistics has led to a confusing heterogeneity of codings, formats, and methods of cataloguing, referencing and diffusion. This diversity slows down the appropriatation of the data by communities of users. Software projects relying on critical volumes of data cannot be carried out without joining different resources together. Clearly, heterogeneous data is leading to exorbitant costs for such projects. Ultimately, this will endanger the plurality of knowledge, and the transfer towards other academic as well as industrial fields.

Interoperability is the key concept which makes it possible to create a convergence of techniques and formats, resulting in a diversity of the practices. It is important to establish a dialogue and to promote exchanges and transfers of experience between the various initiatives. The specific processing of each corpus will be thus improved. With a minimum of interworking, average users of the corpus will be able to work with it independently of the tools developed by the authors. Data could be added to already existing grid of codings.

We need some standardization in order to be able to satisfy the greatest number of users and in the perennial possible way. However, interoperability requirement should not be an additional hassle for corpora designers. One should not force the producers of resources brutally to change working method; it is rather a question ofsetting up import/export facilities in common formats. The aim is to provide the user with the necessary tools to have a diachronic and synchronic glance on big corpora.

We present here the choices which were made within the framework of three oral corpora projects. Both projects rely on linguistic resources, but historically, scientifically and technologically, they differ on many aspects.

Each project will be presented by the person in charge for the software development of the corpus. Convergence and divergence points will be stressed. It is indeed important to be able to establish comparative analyses between these various audio corpora, as well from the point of view of the contents as of the techniques and methodologies.

1. Corpora

1.1 Socio-linguistics studies on Orleans (ESLO)

The Socio-linguistics studies on Orleans (ESLO) is an investigation carried out to the end of the Sixties by British academics to didactic aims (teaching of French in the english public system of education). The investigation was into a sample representative of the urban "orléanaise" community, includes approximately 200 interviews, all referred, and more than 300 hours of sound recordings recordings (which hidden recordings, telephone conversations, public meetings, medico-teaching talks...). This corpus constitutes, by its width and its coherence, most important testimony on French spoken before 1980.

The first objective otf this project is to digitize the sound documents starting from the tape recordings and to propose of it an indexing and a first transcription in order

to put the data in an organization of storage and consultation.
In parallel, an exhaustive exploitation of a subset is committed. With this experience and the analys of the first results, a recent socio-linguistics investigation, called ESLO2, is in hand on the same geographical area. The objective is to evaluate, forty away years, the social dynamics of French (the uses of the language and the judgements on its employment). This project will provide diachronic linguistic resources for the same area. The diversity of the changes is reported to the social parameters, revealing the inequality of resistances or transformation of the language, but also the typology and the dynamics of the evolutions.

1.2 Phonology of the Contemporary French (PFC)

The Phonology of the Contemporary French project (PFC) aims at constituting a vast corpus of phonology of contemporary French, through the whole francophonie and according to precise geographical, social and linguistic criteria. The corpus is composed of recordings, annotations, socio-linguistic information and codings of certainphonological phenomena (schwa, prosody, etc).
With the help of some forty researchers and fifteen PhD students, PFC tries to document and describe the pronunciation of French in its diversity and on the basis of attested usage. The main general objectives of PFC are the following:

1. test phonological and phonetic models from a synchronic and diachronic point of view, giving pride of place to intra speaker and inter-speaker variation;
2. develop a close collaboration between phonologists, experimental phoneticians and specialists in NLP;
3. allow for the conservation of a representative part of French usage across the world;
4. allow the development of better pedagogical material on the basis of usage-based data.

Launched in 1999, the first phase of the project involved the large-scale gathering of data (surveys, digitalisation and transcription of the recordings) on the basis of a uniform methodology which permits a strict comparability of the results. Simultaneously, we devised several systems of annotation and coding for various phenomena (phonological inventories, schwa, liaison) and developed a first family of tools for the partial exploration of the corpus.
The ambition of PFC project is first of all to constitute our data as one of the reference corpora of spoken French. This requires completing the network of survey points, introducing a prosodic level in our coding system and making the corpus inter-operable. To this end, we must systematise our encoding norms as well as adapt and develop various tools for the treatment and the manipulation of the data. Secondly, we are ready to exploit our data on a large scale in order not only to provide better descriptions but also to engage in the current theoretical debates between various approaches such as stochastic models, principles and parameters, optimality theory or laboratory phonology.

1.3 The Archivage corpus of the LACITO

The Archivage corpus of the LACITO lab has been a programme of data safeguard and diffusion of languages with strong oral tradition and on reduced geographical areas. Data collection has been carried out for more than 40 years by researchers in linguistics, anthropology and ethnomusicology all over the world. The corpus is composed of recordings, annotations and transcriptions aligned to the audio.
The main and first goal of the « Archivage » project from the LACITO was to preserve the data harvested by the researchers from this laboratory. This project concerned only the primary data, those which was harvested on the field, i.e. the speech recordings, together with annotations of differents kinds like: transcriptions, translations, morpho-phonemic analysis etc. done during the fieldwork with the help of the speakers.
A digitalization policy has been defined in the laboratory to save some of the recordings (old analog tapes) which was degrading, to catalogue and to document them. The choices concerning the digitalization process where quite comparable of those chosen for the two other projects (ESLO and PFC), i.e. wav/pcm files, 44.1 Khz, 16 bits. On the other hand, for the encoding of the linguistic annotations, we have chosen to create a specific formal syntax especially build for this project but inspired by another one (the Text Encoding Initiative DTD). This syntax defines the objects and concepts the researchers used, i.e. the texts, the word lists or sentence lists, the corpora, the sentences ou breath group, the words, the morphemes, the transcriptions, the translations, etc. This syntax is implemented in an XML DTD and all the annotation's file we have done until now conforms to this syntax.
To disseminate these data, we have created a web architecture, build on the concepts defined by the Open Archive Initiative (OAI). The result of this work is one open archive, i.e. an archive harvestable with the use of the protocole OAI-MHP. This archive provide a free access to all the metadata which describes the resources of the archive in a Dublin-Core coding and in a OLAC (Open Langage Archives Community) coding. This archive at this day disseminate freely not only the metadata of some 150 documents (about 30 languages, most of them unwritten) but also the documents themselves (recordings and linguistic annotations).
This program growth so that a number of others laboratories sharing the same preoccupations came tu us adding some other users of our tools and infrastructure.

2. Convergence and divergence points

This work allows us to identify the actors and functions indispensables to define conservative organisation for sharing this kind of data. What this work has teach us was from different orders:

1. In technical terms this work allows us to understand that it is not possible to separate the preoccupations of safety from those of choising the formats and the encoding of the informations. The concepts of encoding without any loosing of information and the free access to the description of the encoding and to the format are primordiaux. That is for this reason that we do nont accept any proprietary formats or any formats with legal restrictions. Are ignored to all the formats and codec for audio compressions with loosing. In revanche all free, open and standardised

formats are welcome. The main problem today comes from the absence of any standard for linguistic pratices.

2. In organisational terms this work allows us to identify the two main missings wich are the porduction side (The digitisation of the old tapes represent today the only alternative for the preservation of the date. The struggle against the loosing of the colected data is a struggle against the time becose the amount of data to treat is so big and the 'moyens' we have usualy in our kind of laboratories so small that we progress with ant's foot step). The second side we miss is the long time preservation. The disitization is not an end and do not 'garantie' the preservation of the data. What it does is just facilitate the preservation in the mesure of thta duplicate the data can be done without any loosing. We can't think about preservation outside of an institution in which preservation is is its main goal. Only the organisation, the technology 'veille' on new supports, encoding, formats, etc. can avoid the obsolescence of the data.

2.1 Interoperability with a architecture of cataloguing

Sound recording standards being defined better, interworking between the three projects is already practically established. On the other hand, the descriptors of resources are not the same ones according to projects. The coding of the metadata varies because the names of the fields and their possible values were selected for the needs for the investigation. Needs will be indeed different for sociolinguists, ground linguists, dialectologists, phonologists, specialists in TAL or linguistic engineering. The linguistic analyses and thus their codings can also vary. Arranged orthographical transcriptions are the minimum, but linguists will certainly be happy with phonetic transcriptions, morpho-syntactic cuttings, lematisations, etc.

Because of this diversity, interworking between these corpora is currently quasi impossible. Facing this failure, we decided not to change methods and working tools. This is currently impossible, because part of the investigations are already finished or are under development. Instead we work on the definition of common export format and to adopt a common architecture of cataloguing. Exporting data in a standardized common format allows the exploitation with generic tools. The choice of XML in this context is quite natural. On the otherhand there is not standard for the structure of XML documents (schemas or DTD) for the annotation of speech. The chapter devoted to the transcription in the TEI is at the same time too poor and unsuited to the existing practices. And existing metadata standards (DCMI, OLAC, IMDI, MARC, etc.) do not cover completely the needs of oral corpora management.

2.2 Interoperability with a tank of data

The results of these projects (PFC, ESLO and Archivage) direct us towards the definition of an organization of storage and cataloguing of the data centered, around the concept of tank of data, in the definition given by the OAI, i.e. moissonable, with a strict separation of the data and metadata. This tank, in the course of construction, will accept only data and metadata in formats and codings, open and free of right, whose definition will also have to be stored in the tank.

This objective is dictated by the need to maintain the data in the medium and long term. Another key concept for the conservation and the mutualisation of the data is the separation of the logic structure and the typographical structure. The first represents the abstract or scientific of the physical structure; the second represents the usual forms of consultation for this type of data. This separation makes it possible to carry out calculations on the data with terms which the linguists can handle; at the same time, it makes it possible to make evolve easily the various representations of information (on paper, screen, multi-media, adaptation to a handicap, etc).

3. The CatCod Initiative

This comparative presentation of three corpora of audio linguistic resources aims to illustrate the interest to think up stream the interoperability and the methodology of codings and the metadata. Through this step, we want to simplify the technical creation of audio corpora and thus the constitution of linguistic resources, usable byenlarged academic and industrial communities.

The vocation of the CatCod initiative is to organize in France the community of the oral around a common practice of coding and cataloguing for the oral corpora. The purpose of the CatCod group which gathers participants of various laboratories and French universities is to describe the current practices. The result of this work should be to define a standard and its formalization in order to propose them to the Text Encoding Initiative (TEI) consortium.

References

Baude O., Jacobson M., Tchobanov A., Walter R. (2005), Interopérabilité des corpus sonores. In *Phonological Variation : The Case of French, Bulletin PFC 5* (http://www.projet-pfc.net).

Bray, T., Paoli, J. et Sperberg-McQueen, C. M. (Eds) (1998). Extensible Markup Language (XML) Version 1.0, In *Word Wide Web Consortium*.

Jacobson, M. (2004). Corpus oraux en linguistique de terrain. *Traitement automatique des langues*, 45/2, pp. 63-88.

Jacobson, M., Lowe, J. B. & Michailovsky, B. (2001). Linguistic documents synchronizing sound and text, *Speech Communication*, vol. 33, n° 1-2, pp. 79-96.

Sperberg-McQueen, C. M., & Burnard, L. (1994), *TEI Guidelines for Electronic Text Encoding and Interchange (P3),* Chicago and Oxford : ACH/ACL/ALLC Text Encoding Initiative.

Délégation générale à la langue française et aux langues de France, Ministère de la culture et de la communication (2005), *Guide des bonnes pratiques pour la constitution, l'exploitation, la diffusion et la conservation des corpus oraux*, Paris : CNRS éditions (2006).

The Open Archives InitiativeOAI: http://www.openarchives.org

The Dublin Core Metadata Initiative DCMI: http://dublincore.org

The Open Language Archives Community OLAC:
http://www.language-archives.org
The CatCod initiative:
http://icar.ens-lsh.fr/wiki/index.php
The Phonology of the Contemporary French (PFC) :
http://www.projet-pfc.net
The Archivage corpus of the LACITO lab:
http://lacito.vjf.cnrs.fr/archivage