Interdisciplinarity and the sharing of oral data open new perspectives to field linguistics

Bernard Bel, Médéric Gasquet-Cyrus

Interdisciplinarity and the sharing of oral data open new perspectives to field linguistics

Bernard Bel & Médéric Gasquet-Cyrus
Laboratoire Parole et Langage (LPL)
Aix-en-Provence (CNRS & Aix-Marseille Univ): http://lpl-aix.fr

Abstract

Our laboratory is engaged in resource collection, analysis and theorizing (socio)linguistics with particular focus on links between experimental and field approaches. Projects dealing with endangered languages on the border area of Provençal and Francoprovençal (Valjouffrey and Valbonnais, Isère) and near St-Rémy-de-Provence are addressing this issue. Data collection builds on multitrack sound recording and full video coverage allowing accurate studies of interactions.

Dealing with large amounts of data beyond the initial focus of these projects has become a reality thanks to the availability of medium-term and long-term preservation at the Speech & Language Data Repository (SLDR, www.sldr.org). Resource pooling contributed to encouraging scholars and expert amateurs to hand over unpublished data for its preservation and non-commercial distribution. The enthusiastic public response empowered project informants who became fully-fledged members of our team, thereby setting new priorities on the research agenda: the design of a script for their revitalized language and an inventory of place names that delineate their living space.

Our projects

Work refered to in this presentation has been the focus of two projects sponsored by Fédération *Typologie et universaux* (TUL, FR 2559 of CNRS) and *Institut de linguistique française* (ILF, FR 2393 of CNRS) in a joint venture with Ministry of Culture (*Délégation générale à la langue française et aux langues de France*, DGLFLF). The initial call was labelled: *« Corpus en français et en langues de France (constitution, formatage, valorisation, base de données) »*.

Initial team members comprise a linguist, a research engineer, a PhD student in linguistics, a video assistant and a number of local participants hosting our visits and taking part in the collection of information. From the start, contacts in Valjouffrey became possible thanks to the commitment of a young musician and composer (Etienne Champollion) whose family is rooted in the valley.

Valjouffrey's dialect/*patois* is part of the French linguistic heritage. For centuries it was the spoken language of the Valjouffrey valley and it strongly shaped its local culture. It may be classified as a dialectal variant of Alpine Provençal (a language of the Oc family) with a notable influence of Francoprovençal, the language spoken in an area starting at La Mure. As per our knowledge, only four persons (3 men and 1 woman) have maintained (or revitalised) fluency in this language.

In a related project we are recording interactions between five native speakers of a regional variant of Provençal near St-Rémy-de-Provence. Our initial motivation was to compare practices of code-switching between this group and Valjouffrey speakers (Gasquet-Cyrus & Bel 2011).

Technical conditions for fieldwork

Our projects are less focused on the description of languages than on a extensive documentation of their usages, variations, history and cultural-anthropological background. We made a point to create technical conditions for collecting high-quality data in a least-intrusive capture of speech interactions. Each participant is given a head-worn microphone ensuring a stable optimal recording level, and their signals are written to separate tracks. We were lucky enough to acquire a cheap multitrack recording device with eight phantom-powered symmetrical microphone inputs. In addition, every session is covered by two video cameras, one of which provides a fixed large view and the other one is carried by a video operator.

We collected much more material than expected in the initial projects. The Valjouffrey 2010-2011 corpus currently offers 230 Gbytes of sound/video/pictures/texts.

Collecting a large corpus is motivated by: (1) its cultural heritage (patrimonial) value, given the age and very small number of speakers; (2) our hope that good-quality data will be enriched and reused by other teams and disciplines, including for the verification of speech production hypotheses tested in laboratory conditions; (3) our plan to extend this fieldwork to formal experimental approaches, such as setting up sessions for the AMPER project (*Atlas Multimédia Prosodique de l'Espace Roman*).[1]

In our view, a 'corpus' is the whole set of *primary data*, thereby meaning all data collected during a field or laboratory session. This includes sound, video, but also photographs, drawings, maps and written documents handed over by participants.

Long-term preservation and sharing of the corpus

Laboratoire parole et langage (LPL) is in charge of an archive submission site named *Speech & Language Data Repository* (SLDR, www.sldr.org). The aim of SLDR is to preserve data eligible for speech/language research and facilitate its non-commercial sharing. The device is constructed on an interoperable OAIS (*Open Archival Information System,* ISO 14721) currently involving two major computing centres (CINES and CC-IN2P3) in a project initiated by TGE-Adonis.[2]

SLDR has been designed for processing generic items (unrestricted tree-structures of documents) which may be defined (by their *descriptive metadata*) as *primary data* (corpus), *secondary data* (annotations), *tools* or *collections*. Items and documents may be shared in open access or reserved for downloading by identified research scholars after agreeing with a non-commercial licence in compliance with the French *Code du patrimoine* (Heritage Code).[3] SLDR also has provision for facilitating contacts between producers and users of resources (community-building in a Web 2.0 approach)[4] which is conducive to expanding the scope of associated projects.

Our *Code du patrimoine* (L211-4) states that public archives are *"documents produced by the activity of State, local governments, public institutions and other legal persons under*

---

[1] w3.u-grenoble3.fr/dialecto/AMPER/pub.htm

[2] www.sldr.org/doc/show/SldrPresentation-en.pdf

[3] www.sldr.org/wiki/CodePatrimoine

[4] Users' communities, see www.sldr.org/com

public or private law who are in charge of a public service, as part of their public service remit." Most language resources collected for their patrimonial value and/or usability by research scholars connected with public institutions are eligible for this qualification. This is the case of audio and video files collected for the two projects. Another statement (L213-1) stipulates that *"public archives shall be immediately in open access, unless subject to restrictions as per article L213-2."* Nonetheless, article L213-2 gives provision for 25 cases of derogations to this open-access obligation. These derogations have been encoded and tokenized to set the ground for a systematic management of access rights.[5]

Most files collected in the two projects qualify for AR048 derogation: "*50 years. Documents disclosure of which undermines the protection of privacy or for appreciation or value judgments about a person named or easily identifiable, or which reveal the behavior of a person under circumstances which might bring him/her injury.* (art. L213-2, I, 3)" Consequently, access may be restricted unless all participants have signed permissions. The minimum permission grants access to research scholars and students for data analysis. At a later stage recordings (or their anonymised versions) may be shared in open access.[6]

Empowering participants

Resource sharing via SLDR contributed to the popularity of projects with the effect of persuading amateurs or professionals to hand over unpublished resources (recordings, manuscripts or theses) for their preservation and non-commercial distribution.

Valjouffrey *patois* had so far been described in the unique and monumental work by Clément Girard: *Le patois de Valjouffrey, 1970*.[7] The author contacted us after discovering our project on the SLDR site and linked articles on another site dedicated to the valley and its culture. He quickly became a member of our team in full stand.

Clément also submitted to SLDR the 1970-1980s' tape recordings of his close relatives: altogether more than 10 hours including an accurate reading of his entire memoir by his mother Germaine. These recordings are in open-access (MP3 streaming) while high-resolution files (WAV format) may be downloaded by identified research scholars after accepting SLDR licence.[8]

Similarities with Valbonnais *patois* (a neighbouring valley) have been made evident owing to the remarkable thesis by Marcelle Péry who also joined our team as she became aware of early publications of the project: *Étude sur le patois de Valbonnais, 1943*.[9] Marcelle Péry took the initiative of organizing a *Journée Patois* each summer in

---

[5] This table has been translated to all navigation languages of SLDR. The English version is available from www.sldr.org/wiki/table_derogations_en

[6] The full process is exposed on www.sldr.org/wiki/accessRightsManagement_en

[7] www.sldr.org/sldr000006

[8] www.sldr.org/licence/en

[9] www.sldr.org/sldr000005

her Valbonnais home. These events have been video-documented by our team for open-access exposure on the website.[10]

The amplification phenomenon of data sharing, and the extremely friendly nature of relations between all participants (altogether 'scientists' and 'informants') empowered the speakers of Valjouffrey and Valbonnais patois. They became active members of the team, appropriating research topics that they felt most relevant: designing a script for their revitalized language and undertaking a detailed inventory of place names that delineate their living space (toponymy).

Much like other dialects/patois in the same area, Valjouffrey patois has never been a written language. Participants in the project (and notably fluent speakers of the language) expressed their wish to design a script in response to (*partly contradictory*) needs, among which: (1) keeping a memory of the language after the disappearance of all speakers; (2) passing on significant fragments of the language in the form of written documents; (3) stressing characteristic features (*cultural identity*) of the language; (4) facilitating its comprehension by untrained readers. The problematic elaboration of this script has been fully traced on audio/video recordings on the site. Its analysis gave way to an important reflexion on the meaning of, and motivations for standardisation (Thomas et al. 2011).

In Summer 2010, Julien Gaillard, our senior informant in Valjouffrey, had decided on his own to document the names of places and details relevant to mountain climbing in the valley. To this effect he sought assistance from Robert Jamos, an artist and mountain-climber, for the drawing of accurate views of the documented sites. This event oriented our research to the documentation of verbal interactions associated with the elaboration of this toponymy.

Conclusion

We may say that our fieldwork on Valjouffrey and neighbouring dialects is 'event-driven' rather than 'protocol-driven' (Bel 2011) as it tends to comply with an agenda decided by local participants. This type of research produces information packages of very diverse contents: hand-written dissertations have been scanned and archived for an open-access dissemination. Drawings and maps, lexica will also be preserved as 'linguistic resources' attached to the corpus.

This is an incentive for developing long-term preservation and sharing work environments able to handle generic information packages. Consequently, the identification and proper access to documents strongly depends on high-quality metadata and descriptive files (including annotations). At this stage it is crucial for us to participate in the elaboration of methodological guidelines that will facilitate interoperability between several data repositories without imposing technical restrictions that might be detrimental to the diversity of methodologies.[11]

References

Bel, B. (2011). Technology at the meeting point of hardware, software and 'mindware'. Proceedings of CLARIN/DARIAH conference *Supporting Digital*

[10] www.sldr.org/sldr000736

[11] In this direction, check the ORTOLANG project: www.sldr.org/wiki/ortolang

Humanities 2011: Answering the unaskable (November 17-18: Copenhagen, Denmark). www.lpl-aix.fr/article/4803

Gasquet-Cyrus M.; Bel, B. (2011). (Re)parler sa langue : l'alternance codique à la recherche de langues 'oubliées. Langues et cité, no. 19. 2011, p. 1-2. www.lpl-aix.fr/article/4869

Thomas, A.; Gasquet-Cyrus, M.; Bel, B. (2011). Revitalisation ou reniement de la langue locale ? L'élaboration problématique d'une graphie pour le « patois » de Valjouffrey. Actes, Colloque international AULF et LESCLaP-CEP : Standardisation et vitalité des langues de France (2011 octobre 13-14 : Amiens, France) (forthcoming) www.lpl-aix.fr/article/4712