Les corpus oraux entre science et patrimoine

L'expérience de *l'observatoire des pratiques linguistiques*

Les premières initiatives de constitution de "corpus oraux" (recueils ordonnés d'enregistrements à des fins scientifiques) datent à peine d'un siècle et les nouvelles technologies permettant le traitement informatique de données sonores sont à l'aube de leur développement. Il serait cependant erroné de réduire le peu de reconnaissance que l'on concède aux corpus oraux à un problème exclusivement technique (de diffusion). En effet, c'est bien plus du statut de la voix et de la langue orale[1] dans le monde social dont il s'agit.

Depuis quelques mois, le conseil scientifique de l'Observatoire des pratiques linguistiques de la Délégation Générale à la Langue Française et aux Langues de France[2] développe une initiative en faveur *de la constitution, l'exploitation et la diffusion des corpus oraux en France*. Cette initiative en est à ses prémices, il ne s'agit donc pas de proposer un bilan, mais d'apporter simplement quelques réflexions sur une action qui mêle la *diffusion* de recherches et *la mise en public* d'objets scientifiques.

1 *L'oral*, du champ scientifique à l'espace social: la disparition du *locuteur*

L'intérêt porté à "ce que parlent" les Français est lié historiquement aux marchés linguistiques dès la première enquête de l'Abbé Grégoire et son *Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française* (1794). Toutefois, la constitution du champ scientifique va complexifier cette approche. Ainsi, la dialectologie se restreindra à un objectif de description des patois et des dialectes avant que l'unification linguistique ne les fasse disparaître.

Par la suite, la méthodologie des Atlas linguistiques qui consiste à reproduire sur fonds de carte des variétés régionales en disparition (par simples isoglosses sans prendre en compte la notion de système linguistique seul accès à la langue et donc à l'identité du témoin)[3] exclura le locuteur et son espace social au profit d'une représentation géographique. Pourtant dès 1911, Ferdinand Brunot avait créé à la Sorbonne les Archives de la Parole, première initiative de publicisation d'enregistrements de variétés régionales et de *français parlé*, modifiant à la fois le champ scientifique et l'espace publique soumis à la norme unificatrice. Cette ambition ne survécut pas à la guerre, et l'on abandonna l'idée "*d'étude, d'archivage et d'analyse de parlers d'hommes et de femmes parlant comme à l'auberge ou à la fontaine*"[4] pour se consacrer principalement à la conservation d'une culture folklorique. En 1938 la création de la Phonothèque nationale permettra à Roger Dèvigne de développer les "croisières folkloriques" pour constituer une *Encyclopédie nationale des parlers, patois et vieux chants de France*. La phonothèque intégrée par décret en 1977 au sein d'un département spécialisé de la Bibliothèque nationale s'orientera vers une activité de recueil[5], et sera en 1995, totalement incorporée au département de l'audiovisuel de la BNF.

La linguistique de l'oral ne saurait se réduire à la dialectologie. Si cette dernière s'est construite en réaction à une norme unificatrice (le français), la linguistique de l'oral s'opposera à l'académisme normatif prépondérant à la linguistique (de l'écrit).

Cette contrainte normative trouve un renfort historique dans la réception des concepts fondamentaux de la linguistique moderne élaborés par Ferdinand de Saussure. La lecture des dichotomies fondatrices synchronie/diachronie, signifiant/signifié et langue/parole par les linguistes provoqueront le rejet de la description des variations en générale et des productions orales en particulier par une science vouée à la recherche d'invariants structurés en système. En France, seule la linguistique variationniste essaiera de refuser cette séparation entre invariants et variations, en proposant une méthodologie de l'enquête et du recueil de données *attestées* et *situées*. Cette approche, la plus rigoureuse et novatrice d'une sociolinguistique qui souhaite reconstruire la

[1] Plus exactement des *manifestations orales de la langue*.
[2] DGLFLF, direction du ministère de la Culture et de la Communication
[3] Laks 2003
[4] Callas 1992, Vecken 1984
[5] convention avec le CNRS en 1979 pour l'archivage des données des Atlas

linguistique moderne sera théorisée principalement par Encrevé[6], restera fortement dominée dans le champ, tout comme les initiatives françaises de publicisation de grands corpus de référence (le français fondamental, le corpus du Groupe Aixois de Recherche en Syntaxe et les corpus d'applications).
Ainsi, le français et surtout sa forme normative (l'écrit littéraire) apparaît comme le seul objet scientifique légitime. Cela sera aussi le seul objet légitime de l'espace social, reconnu comme capital culturel par excellence.

2 Frémissements et bouleversements du champ

Depuis quelques années la situation se modifie tant dans le champ scientifique que dans l'espace social.

2.1 Une politique linguistique fondée sur la diffusion scientifique

Le premier frémissement provient de l'élaboration de nouvelles politiques linguistiques. La DGLF dont *le rôle est de contribuer à la mise en œuvre de la politique linguistique* est transformée, en 2001, en DGLF***LF*** (*et aux Langues de France*), avec une volonté de développer deux axes. L'un est traditionnel (promotion du français comme langue et norme de l'état et de la république, loi Toubon de 1994, commission de terminologie), l'autre est novateur (prise en compte des variétés et des variations linguistiques dans leurs usages sociaux, réforme de l'orthographe de 1990, féminisation, charte européenne de protection des langues régionales et minoritaires, assises des langues de France,…). Autre indice, l'installation en 1999 de l'Observatoire des pratiques linguistiques dont la mission est *de recenser et de rendre disponible les savoirs issus de la recherche scientifique relatifs à la situation linguistique en France*.
L'Observatoire, rattaché à la Mission de langues de France est doté d'un conseil scientifique qui propose des actions de diffusion d'informations scientifiques auprès de deux cercles distincts : les acteurs des politiques linguistiques (direction des ministères de la culture, de l'éducation, les collectivités territoriales,…) et un public plus large (acteurs scientifiques, éducatifs, sociaux, culturels,…). L'observatoire valorise la recherche fondée sur les données attestées et situées par la diffusion de "synthèses vulgarisées" (en ligne et par un bulletin) et par l'aide financière à certains projets de recherche (appel d'offre).
Depuis peu, le ministère de la Culture connaît un autre bouleversement : la volonté de rendre accessible des masses de données reconnues comme patrimoine. Ainsi, pour l'oral, l'INA a commencé la numérisation de 60 ans de radios (et 50 ans de télé) soit plus de 700 000 heures d'archives sonores et a annoncé, en février 2004 le lancement, de *la première banque mondiale d'archives numérisées*.

2.2 L'ère de la mise en public des masses de données

La notion du traitement informatique de masse de données est également au cœur du champ scientifique avec l'émergence du Traitement Automatique du Langage et *des linguistiques de corpus*[7].
Les requêtes complexes sur un corpus très vaste (plusieurs centaines de millions de mots) apportent une méthodologie très puissante mobilisée par les linguistiques de l'écrit. Plus récemment, la numérisation de données sonores permet d'envisager un engouement similaire. Il ne s'agit donc plus d'enregistrements isolés de témoignages sonores mais de la constitution de bases de données de grande ampleur. Les outils se développent, qu'il s'agisse de numérisation, de transcription, d'annotation et de balisage, mais aussi de conservation et de diffusion, ce qui modifie la notion d'accès aux données. En effet, une tendance très forte se dessine autour de *l'interopérabilité*. La normalisation permet de concevoir les corpus comme des objets scientifiques

[6] Encrevé 1976, 1983, 1992.
[7] Habert Les linguistiques de corpus, 1998.

ré-exploitables (le fait le plus marquant est la création du métalangage XML[8] qui facilite l'échange des outils et des documents, et les initiatives internationales de normalisation comme la TEI[9]).
Cette interopérabilité est aussi à la pointe des réflexions sur la conservation et la mise à disposition des politiques culturelles et patrimoniales. La Commission Européenne a financé le projet Minerva (Réseau ministériel pour la valorisation des activités de numérisation) et le Groupe des représentants nationaux (GRN) des Etats membres de l'UE sur la numérisation du patrimoine culturel a adopté en 2001 les principes de "Lund" pour faciliter l'accès aux ressources numérisées. Cette même année la France créait le *comité de concertation pour les données en sciences humaines et sociales* auprès des ministres chargés de l'économie, de l'emploi, de l'éducation nationale et de la recherche, dont le rôle est de promouvoir la diffusion des données ayant un intérêt scientifique.

3 L'initiative de l'Observatoire

Le traitement automatique des corpus oraux apparaît alors comme l'occasion d'apporter la reconnaissance souhaitée à l'hétérogénéité des pratiques linguistiques et à la légitimité des corpus oraux comme objet scientifique et patrimonial. Le conseil scientifique de l'Observatoire ne pouvait être insensible à cet enjeu[10] et fut donc à l'initiative de la création d'un comité de pilotage *pour la constitution et l'exploitation des corpus oraux*. Outre les membres du conseil scientifique, celui-ci est composé de *conservateurs* dépendant du ministère de la culture (INA, Archives, BNF) de représentants des institutions scientifiques (CNRS), de linguistes avec le statut d'experts et de juristes spécialistes de la propriété intellectuelle et du droit de la science.
Le travail de réflexion de ce comité est en cours et je ne peux donc que mentionner l'ébauche de celle-ci qui, en tout état de cause, porte sur la valorisation des corpus oraux et la mise en public de cette légitimité scientifique et patrimoniale. Je me bornerai à présenter trois aspects de cette réflexion: les aspects juridiques, la normalisation et les outils techniques.

3.1 Les aspects juridiques.

Le comité de pilotage a créé un premier groupe de travail composé de linguistes, de conservateurs et de juristes, sur *les aspects juridiques*. Si ceux-ci forment le premier verrou évoqué par les chercheurs et les conservateurs c'est justement parce que les nouveaux outils de traitement et de diffusion de ces données modifient considérablement la donne. Ainsi la diffusion des enregistrements implique une procédure d'anonymisation, qui ne peut se réduire, à l'ère du croisement des bases de données, à une opération simpliste comme le bippage des noms propres. De même l'interopérabilité rend imprévisible *les finalités* et complique donc le recueil de consentement du témoin. Enfin la diffusion et l'utilisation industrielle posent la question de la propriété de la voix. Il faut donc reconnaître un *auteur*, un *propriétaire*, un *responsable*, aux données orales. Le comité de pilotage a ainsi décidé de rédiger un *guide des bonnes pratiques* dont l'objectif n'est pas de fournir des solutions clés en mains, mais de construire au sein de la communauté des chercheurs une éthique qui oblige à repenser les liens entre *données* et *donneur* et à reconnaître au locuteur un *corps socialement situé.*

[8] eXtensible Markup Language www.w3c.org
[9] Text Encoding Initiative
[10] Le président du conseil scientifique, P. Encrevé, à l'origine de la linguistique variationniste est le principal acteur des politiques linguistiques novatrices en France (membre du cabinet du 1er Ministre M. Rocard en charge de la langue, puis avec les mêmes fonctions au cabinet de C. Trautman Ministre de la culture) Le second membre B. Laks, issu lui aussi du champ de la linguistique variationniste et de la phonologie cognitive est directeur d'un grand laboratoire de recherche universitaire et coresponsable du projet actuel sur la phonologie du français contemporain (grand corpus oral).

3.2 Normalisation et métadonnées : pour des données situées

Cette reconnaissance passe aussi par les enjeux d'une *normalisation* dans un but d'interopérabilité. Les normes internationales de stockage et d'échange de données, en cours de constitution, définissent la structure même des objets scientifiques. Techniquement les standards actuels séparent la représentation physique et logique des documents (les données et les métadonnées). Tout document XML comporte l'identification des éléments possibles et leurs relations possibles (Définition de Type de Document) *et* les données identifiées selon cette DTD. C'est alors la notion même de données brutes qui est redéfinie. Ainsi la TEI rend obligatoire la constitution d'un header (en-tête) en début de corpus qui recense les informations sur le contexte de production des données. Un certain nombre d'informations sont donc fournies sur le locuteur avec la possibilité de reconnaître celui-ci comme être *socialement situé*. Tout l'enjeu réside ici : quels vont être les choix de standardisation? Et qui participent à ces choix? Actuellement ces normes sont soumises à deux forces, celle des utilisateurs (aucune recommandation ne peut se pérenniser sans l'assentiment des utilisateurs, c'est encore vrai dans le champ scientifique) et celles des producteurs (W3C Word Wide Web consortium, ISO–TC37 pour les ressources linguistiques,…).

3.3 Diffusion de l'oral et transcription

L'exemple de la transcription des enregistrements permet également d'expliciter les enjeux de l'interopérabilité. Le travail scientifique sur la langue parlée passe toujours par la *transcription* de la parole. Ce n'est pas le moindre des paradoxes pour des linguistes qui souhaitent travailler sur l'oral que d'être confronté au plus fort des effets de normalisation de l'écrit : l'orthographe. Pendant longtemps les transcripteurs devaient choisir entre l'orthographe normée, un alphabet phonétique, des conventions pour les formes orales et/ou l'ajout d'informations supplémentaires (prosodie,…). Ces choix fondamentalement liés aux enjeux théoriques de la recherche ont eu pour conséquence de rendre confidentiels les corpus transcris avec le plus de finesse (quand les marques du locuteur situé étaient présentes) au profit de corpus transformés pour les besoins des outils de traitement automatique de l'écrit et de la norme sociale.
Les logiciels[11] développés pour l'aide à la transcription sont fondés sur l'interopérabilité et ils risquent bien de modifier le champ scientifique et par delà même l'espace social.
Premièrement ceux-ci reposent sur l'alignement du signal reliant définitivement la *transcription* à la *voix* en un unique objet. Deuxièmement, la majorité de ces logiciels est prévue pour la multitranscription et proposent donc une granularité modulable et une représentation hétérogène (une première ligne de transcription peut contenir une transcription orthographique, une seconde phonétique, une troisième être aménagée pour une recherche spécifique, etc.).
C'est l'impact même de la représentation graphique et de sa norme qui est estompé et il devient bien plus difficile d'oublier que derrière les mots il y a une *voix* et un *locuteur*.

Ces trois aspects de la publicisation des corpus oraux offrent l'opportunité de replacer la *nature sociale* de la langue au centre de la linguistique et de rappeler par la même à la communauté scientifique comme à l'espace social qu'il n'y a pas de langue sans locuteur. Considérer l'hétérogénéité des usages de l'oral comme inhérent à la constitution de l'objet d'étude comme du patrimoine doit permettre de respecter les deux principes qui définissent, selon Labov, la *dette du chercheur* : la responsabilité sociale envers le terrain et le rendu scientifique à la communauté observée.

[11] Praat, Tasx, Anvil, etc.

Olivier Baude
MCF Université d'Orléans, **C**entre **O**rléanais de **R**echerche en **A**nthropologie et en **L**inguistique
Celith, EHESS, Paris.
DGLFLF

Les corpus oraux entre science et patrimoine.
L'expérience de *l'observatoire des pratiques linguistiques*

mots-clés: corpus oraux, interopérabilité, politique linguistique, publicisation, normalisation.

Bibliographie:

Baude O. 1999, *L'observation des pratiques linguistiques en France*, Culture et Recherche n°75, Ministère de la Culture et de la Communication.
Biber D. 1988, *Variation across speech and writing*, Cambridge University Press.
Bergounioux G. (dir) 1992, *Enquêtes, corpus et témoins*, Langue Française n°93, Larousse.
Blanche-Benveniste C. 2000, *Corpus de français parlé*, in Corpus méthodologie et applications linguistiques, édité par M. Bilger, Champion.
Calas M-F. 2002, *Le statut documentaire de la source orale*, in de la source à l'archive, actes des journées d'études, AFAS.
Encrevé P. 1976, *Présentation*, in Sociolinguistique, W. Labov, Minuit
Encrevé P. 2001, La langue de la République, in La République, Pouvoirs n°100, Seuil, Paris
Gibbon D, Moore R, Winski R., 1997*, Handbook of standards and ressources for spoken language ressources*, Mouton de Gruyter. New-York.
Habert B. 2000, *Des corpus représentatifs: de quoi, pourquoi, comment?* In Cahiers de l'Université de Perpignan, N° 31, Presses universitaires de Perpignan.
Laks B. 2003, *Les grandes enquêtes phonologiques en France*, in la prononciation du français dans sa variation, La tribune internationale des langues vivantes n°33.