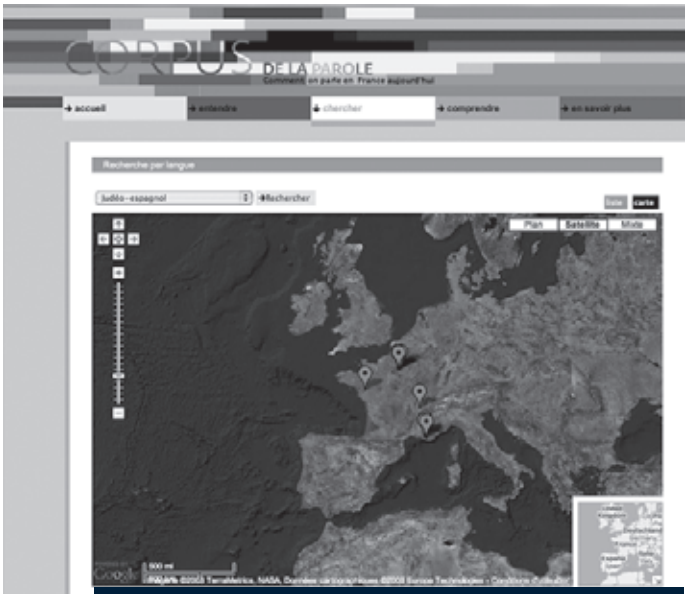
Les corpus de la parole :
patrimoine immatériel et langues de France

Olivier Baude et Michel Alessio
MCC / Délégation générale à la langue française et aux langues de France



http://corpusdelaparole.culture.fr

Verba volant, scripta manent. Dans un pays si fortement marqué par la tradition écrite et les usages littéraires, longtemps la langue parlée n'a pas été perçue dans toute son importance. C'est l'auteur de la grande *Histoire de la langue française,* Ferdinand Brunot, qui, le premier, s'est préoccupé d'enregistrer et de conserver les traces sonores de faits de langue, en créant les fameuses *Archives de la parole* en 1911. Ces enregistrements sur rouleaux de cire, pieusement conservés, forment le fonds premier du département de l'audiovisuel à la Bibliothèque nationale de France. Jusque-là, la parole vivante apparaissait curieusement – et paradoxalement – comme une forme subalterne, dérivée, et pour tout dire dégradée de la langue écrite. Effet de culture : telle était la représentation dominante qu'on se faisait du langage en France. Ç'a été le travail de la linguistique du XX^e^ siècle que de rétablir l'ordre des choses en se fondant notamment sur la description de données orales constituées en collections et ordonnées par des critères scientifiques : les corpus oraux.

Ce travail ne se fait pas d'un trait. À la suite des *Archives de la parole,* sont créées en 1932 la phonothèque du musée de l'Homme et en 1938 la Phonothèque nationale. C'est toutefois le musée national des Arts et Traditions populaires qui possédait le plus de témoignages oraux. Mais ceux-ci étaient exclusivement réalisés par des ethnologues, et destinés à leurs recherches. L'oral n'était pas encore collecté pour lui-même ; il n'était que le support d'études en sciences humaines et sociales. D'ailleurs, lorsqu'André Malraux lance l'Inventaire général des monuments et richesses artistiques de la France en 1964, l'oral n'y figure nulle part.

À l'ère du numérique, la sauvegarde de l'oral prend véritablement son essor. L'informatique permet de faciliter la classification des enregistrements et leur accès. Mais cela ne résout pas tous les problèmes. Un enregistrement isolé ne présente guère d'intérêt en soi. Il n'y a patrimoine de l'oral que lorsque plusieurs documents sont regroupés autour d'un thème. Enfin, la validation scientifique et le traitement des données enregistrées marquent la porte d'entrée du domaine du patrimoine. Celui-ci en effet ne s'étend pas aux données sonores brutes qu'un particulier a pu collecter à des fins personnelles, lors d'une conversation ou d'un entretien. Le « sceau de la science » doit garantir que les corpus oraux ont été correctement composés, c'est-à-dire indexés, transcrits, éventuellement traduits, balisés, annotés, catalogués.

En France, les premières grandes enquêtes sur le français ont été effectuées dans les années 1950 à des fins didactiques. Or, plus de cinquante ans après, nous ne disposons toujours pas d'un véritable corpus de référence qui permette toutes sortes de recherches (descriptions, analyses, applications) et on mesure le retard pris par notre pays, y compris vis-à-vis d'autres zones francophones, en comparant ces résultats avec les données engrangées ailleurs.

Le ministère de la Culture et de la Communication / Délégation générale à la langue française et aux langues de France (DGLFLF), en partenariat avec les chercheurs des universités et du CNRS, a entrepris de combler ce retard, dans une perspective de valorisation de la diversité. La France dispose en effet d'une grande richesse de langues. À côté du français, langue nationale, langue commune, présente sur les cinq continents, les langues de France constituent un patrimoine culturel unique : il y a sur le territoire de la République des langues romanes, des langues germaniques, le breton, langue celtique, le basque, qui n'est pas une langue indo-européenne, des créoles, des langues amérindiennes, des langues polynésiennes, des langues austronésiennes, etc. Plus de 75 langues sont reconnues comme « langues de France », c'est-à-dire parlées par des citoyens français en France depuis assez longtemps pour faire partie du patrimoine culturel national, et qui par ailleurs ne sont langue officielle d'aucun État.

Ce patrimoine est trop souvent méconnu, et si des archives sonores existent désormais pour la quasi-totalité de ces langues, la richesse qu'elles représentent n'était jusqu'ici accessible ni à l'ensemble de la communauté scientifique ni au grand public. Plus grave encore, de nombreux documents sonores conservés sur des supports physiques à bout d'usage (comme les enregistrements sur bandes magnétiques) sont voués à disparaître dans des délais très courts. Or, il s'agit souvent des derniers ou des seuls documents dont nous disposions sur des langues de France – comme pour certaines langues de Guyane ou de Nouvelle-Calédonie –, mais aussi sur le français. Ainsi, au ministère de la Culture, la DGLFLF a numérisé les seuls enregistrements de français constitués par des linguistes dans les années 1970.

Aujourd'hui, avec le progrès des nouvelles technologies, la numérisation offre non seulement la possibilité de sauver ce patrimoine mais aussi l'occasion de le valoriser en transformant les documents originaux en de véritables ressources linguistiques numé-

riques. En créant, en partenariat avec le CNRS, le programme *Corpus de la parole,* le ministère de la Culture et de la Communication/DGLFLF s'est engagée depuis 2004 dans une triple démarche.
La première étape de ce programme était d'ordre méthodologique ; il s'agissait de définir les conditions dans lesquelles les productions verbales devaient être recueillies à des fins d'études et de recherches, et c'est ainsi qu'a été entreprise l'édition de l'ouvrage *Corpus oraux, Guide des bonnes pratiques* (CNRS-Editions et PUO, 2006), consacré à la constitution, la conservation et l'exploitation des corpus oraux. Ce guide s'inscrit à l'exact croisement d'une démarche scientifique et d'une politique culturelle ; il constitue aujourd'hui une proposition de charte pour tous les chercheurs, auxquels il fournit les instruments, y compris les instruments de prescription d'ordre juridique, qui permettent de constituer ces données brutes en objets de savoir.
La seconde étape a consisté à lancer un vaste chantier de numérisation dans le cadre du plan national de numérisation du ministère de la Culture (DDAI/MRT). Ce plan a déjà permis de sauvegarder des centaines d'heures d'enregistrement.
La troisième étape a consisté à rendre les données accessibles à tous, d'abord à la collectivité des chercheurs, mais aussi, au-delà des chercheurs, au grand public, et c'est désormais possible avec le site *Corpus de la parole* (**http://corpusdelaparole.culture.fr**), dont la première version est en ligne depuis le début de l'année 2008. Ce site donne accès à un catalogue collectif de corpus oraux en français et en langues de France, sous la forme de fonds sonores transcrits et numérisés.
La sauvegarde et l'exploitation de ces enregistrements en français et en langues de France sont un enjeu de première importance. C'est un enjeu pour la recherche, pour le développement de l'ingénierie linguistique et pour l'enseignement, mais aussi pour le développement d'une politique culturelle qui reconnaisse les faits de langue comme éléments du patrimoine immatériel dans toute sa variété.