

Langues et cité

Corpus de la parole

La France dispose d'une richesse linguistique fondée sur la diversité. À côté du français, langue nationale, les langues de France constituent un patrimoine culturel unique. Ce patrimoine est méconnu, et si des documents sonores existent pour la quasi-

totalité de ces langues, ils ne sont accessibles ni à l'ensemble de la communauté scientifique, ni au grand public. Plus grave encore, de nombreux documents sonores uniques, conservés sur des supports physiques en fin de vie, sont voués à disparaître à tout jamais dans un délai très bref. La numérisation offre non seulement la possibilité de sauver ces documents, mais aussi de les valoriser en les transformant en de véritables ressources linguistiques numériques. Ces corpus oraux, sous la forme de collections ordonnées d'enregistrements de productions linguistiques orales et multi-modales, prennent alors une valeur scientifique autant que patrimoniale.

Le développement des corpus oraux du français et des langues parlées en France est un enjeu de première importance pour la recherche et le développement de l'ingénierie linguistique, mais aussi pour l'enseignement de ces langues, pour la sauvegarde et la diffusion du patrimoine oral et la reconnaissance de la diversité linguistique.

Ce numéro présente un état de la recherche sur les corpus oraux et témoigne des nombreuses initiatives en cours dans ce domaine. C'est aussi l'occasion de présenter les actions du programme « corpus de la parole » mené par l'Observatoire des pratiques linguistiques de la DGLFLF et plus particulièrement la publication de l'ouvrage « Corpus oraux, guide des bonnes pratiques 2006 », ou encore les opérations de numérisation d'archives sonores dans le cadre du Plan de Numérisation du ministère de la culture et certains projets de recherche en cours.

Langues et cité

Grands corpus	p. 2
Statut patrimonial	p. 4
Des <i>Archives</i> au numérique	p. 5
Entrevue	p. 6
Projet phonologique	p. 8
Inventaire	p. 8
Enquête	p. 9
Projet CLAPI	p. 9
Projet ILF	p. 10
C-ORAL-ROM	p. 10
Projet LACITO	p. 11

Bulletin de l'observatoire des pratiques linguistiques



Depuis la fin des années 1960, de nombreux pays ont favorisé l'étude de leurs langues parlées, en multipliant les collections d'enregistrements, les transcriptions et les études de prononciation, de lexique, de grammaire, de discours, de sociolinguistique, de psycholinguistique, etc. L'impulsion a souvent été donnée par les académies chargées de veiller sur les langues nationales (par exemple, pour l'Europe, celles de Grande-Bretagne, d'Espagne, d'Italie, du Portugal, des Pays-Bas, d'Allemagne, des pays scandinaves). Des sommes importantes ont été consacrées à ces recherches, qui demandent de gros budgets pour l'organisation des banques de données, le développement des moyens électroniques adaptés et la formation des spécialistes.

Les grands corpus de langue parlée

Claire Blanche-Benveniste

Université de Provence

École Pratique des Hautes Etudes, Paris

Quel est l'intérêt de ces recherches ?
Où en est-on en France ?

Le premier intérêt de ces études est de permettre un grand progrès de la connaissance. Tout ce que nous savons de la relation entre langue parlée et langue écrite s'en trouve changé. Elles nous révèlent, en effet, que nous ne pouvons pas compter seulement sur notre intuition pour avoir une bonne représentation de la langue parlée : il faut pouvoir disposer de très nombreux exemples, avec toutes les caractéristiques possibles de prononciation, d'intonation, de vocabulaire et de grammaire, produits dans des situations très diversifiées et par des locuteurs très différents. Les faits statistiques sont ici primordiaux. Il importe de savoir si les phénomènes considérés sont très fréquents ou peu fréquents, et s'ils sont produits par tout le monde ou seulement par certaines personnes dans des circonstances déterminées. Du coup, la plupart des préjugés habituels sont dissipés : la langue parlée ne peut pas se ramener aux prises de parole familières, incomplètes et pleines de fautes que l'on cite souvent pour l'opposer à de bons exemples de langue écrite. Elle comprend aussi des prises de parole publiques, soignées, voire solennelles, avec quantité de « genres » différents, conversations, descriptions, récits, explications techniques,

argumentations, jeux de rôles, enregistrements de radio et télévision, etc. Il est possible d'aborder par là certains domaines de science cognitive, par exemple en observant comment les locuteurs s'adaptent aux différents genres, comment ils gardent en mémoire ce qu'ils viennent de prononcer et prévoient ce qu'ils vont dire ensuite (des mécanismes spécifiques apparaissent quand ils cherchent leurs mots), comment ils manipulent les répétitions, comment ils utilisent leurs voix et leurs gestes ou comment ils mènent les interactions avec autrui dans les conversations.

Les applications pratiques de ces recherches sont nombreuses. Il faut citer en premier lieu tout ce qui tient au Traitement Automatique du Langage (TAL), par exemple la reconnaissance et la synthèse de la parole, la consultation de données en langue parlée ou les dialogues entre hommes et machines. S'il existe actuellement des possibilités de demander oralement des renseignements à des machines, s'il existe des machines capables de lire des journaux et des livres pour les mal-voyants (une de ces machines fonctionne à l'Université de Caen), c'est grâce à ces recherches.

Les grandes collectes de langue parlée (les corpus oraux) fonctionnent comme des bases de données permettant de faire des comparaisons, ce qui se révèle

nécessaire dans de nombreux domaines. Les comparaisons entre les parlers de différentes régions sont nécessaires pour calculer les politiques linguistiques. Les anglophones disposent, à cet effet, d'une immense documentation sur les différentes sortes d'anglais parlées dans le monde. Une grande documentation existe aussi aujourd'hui sur les principales différences géographiques qui affectent la langue portugaise, selon qu'elle est parlée au Portugal, dans les îles, dans différentes régions du Brésil, au Mozambique, en Angola, en Guinée-Bissau, à Timor ou dans d'autres régions d'Asie. Les enseignants peuvent ainsi choisir les particularités qu'ils veulent conserver et celles qu'ils veulent écarter. Les grands corpus aident aussi à évaluer l'acquisition de la langue maternelle, en montrant ce qui est spécifique aux enfants de tel ou tel âge et ce qui se trouve aussi bien chez les enfants que chez les adultes. La comparaison est absolument indispensable pour tous les secteurs pathologiques, par exemple pour savoir si une prononciation défectueuse est significative ou non d'un type de « maladie du langage », ou dans quelle mesure certaines répétitions de lexique sont banales alors que d'autres signalent au contraire des troubles importants.

Les éditeurs du monde anglophone utilisent largement les résultats de ces

recherches pour diffuser des manuels d'enseignement de l'anglais comme langue étrangère (Collins, par exemple). Ils publient du matériel pédagogique qui tient compte de la fréquence des phénomènes grammaticaux et de leur répartition, en fournissant tous les exemples nécessaires. On dispose maintenant, pour plusieurs langues, de corpus dits alignés, qui permettent d'écouter une portion d'enregistrement sonore tout en lisant sur un écran la transcription écrite qui correspond, groupes de mots par groupes de mots. Ces corpus alignés fournissent des outils d'enseignement remarquables, qu'on peut utiliser seul ou avec le secours d'un moniteur.

Ces recherches demandent des investissements plus importants qu'on ne pourrait le croire quand on ne connaît pas le domaine. La partie technique d'enregistrements et d'équipements informatiques est chère, mais la participation de linguistes spécialisés l'est également. Transcrire des enregistrements de langue parlée est une opération délicate, qui exige une formation préalable, du temps et de la patience (il s'agit, par exemple, de circuler à travers des prononciations variées, de bien noter les répétitions et d'éviter de transcrire ce qu'on a cru entendre). Au début des années 2000, un des responsables du corpus de langue néerlandaise estimait qu'il fallait prévoir un euro par mot transcrit. Or les corpus de langue parlée actuels sont estimés utiles s'ils comptent au moins dix millions de mots (entre 800 et 1 000 heures d'enregistrement). Le calcul est simple : c'est un investissement lourd.

Où en sont actuellement les recherches sur le français parlé ?

En France, les premières enquêtes avaient commencé assez tôt, dans les années 1950, avec des collectes d'enregistrement destinées à l'enseignement du

français comme langue étrangère (Français fondamental, Corpus d'Orléans), de taille assez réduite. A partir des années 1980, des linguistes se sont intéressés à la description systématique de la langue parlée (équipe du GARS, à l'Université de Provence, projet de recherche sur la Phonétique du Français Contemporain, PFC, à l'Université de Toulouse, nombreuses recherches éparpillées, en France et hors de France). Jusqu'à présent, cependant, aucun projet national de grande envergure n'a été mené à bien. Les plus grands corpus comptent tout juste deux millions de mots et ils ne correspondent pas aux standards internationaux qui ont cours actuellement. Alors que la France a été pionnière dans la collecte de corpus de langue écrite (base FRANTEXT), elle est en retard pour la langue parlée.

La connaissance des relations entre langue parlée et langue écrite est encore souvent marquée par d'anciens préjugés (le poids de l'orthographe grammaticale du français y contribue pour beaucoup), ce qui a des conséquences importantes sur l'enseignement de la langue. En voici deux exemples. Que faire pour savoir comment les Français conjuguent effectivement leurs verbes lorsqu'ils parlent, dans différentes situations, en tenant compte de leurs formations différentes, compte non tenu des marques orthographiques ? Une réponse facile, fondée sur l'ignorance, serait de dire que la langue parlée n'a « pas de grammaire » et de s'en tenir là. Il serait pourtant fort utile, quand on enseigne à de jeunes enfants qui ont appris leur langue en écoutant parler les adultes, de savoir à quelle sorte de conjugaison des verbes ils ont été exposés et de distinguer ce qu'ils savent et ce qu'ils ne savent pas. Une étude rapide montre que, dans la conversation usuelle, les adultes ne conjuguent largement qu'une dizaine de verbes fréquents. Dans leur grande majorité, les autres sont utilisés à

l'infinitif, au participe passé et à la troisième personne du présent.

Deuxième exemple : il serait indispensable de savoir quelles fautes sont spécifiquement enfantines et quelles fautes sont produites par tout le monde, adultes et enfants : prononcer *quat'* pour *quatre*, disloquer les sujets comme dans *mon père, il est venu*, sont des particularités qui se manifestent chez quantité de locuteurs, depuis fort longtemps et qui n'ont rien à voir avec l'âge. Une documentation sérieuse permettra de voir que ces fautes, si fréquentes dans le langage de conversation, sont très rares dans les situations de parole surveillée, les prises de parole publiques, ou les discours professionnels standardisés.

La documentation fait également défaut pour l'étude des pathologies. Dans les hôpitaux où l'on soigne les troubles de langage (les nombreux accidents de moto en créent beaucoup chez de jeunes adultes), le personnel soignant est souvent amené à juger sans bases de comparaison avec d'autres productions orales qui passent pour « normales ». Comment savoir, dans ces conditions, quelles sortes de « phrases inachevées » sont à considérer comme banales et quelles autres sortes sont, au contraire, les indices d'un trouble particulier ? Comment savoir si un usage massif des verbes avoir et être, au détriment d'autres verbes, est l'indice d'une pathologie ? Il est difficile de répondre à des questions de ce type sans une bonne base de données de comparaison.

Les linguistes plaident souvent, au nom des connaissances fondamentales, pour que l'on développe de grands corpus de français parlé. Comme on l'a vu dans d'autres pays, ces corpus de langue parlée permettent aussi des applications pratiques qui répondent à de nombreuses demandes sociales ●

Références :

Revue Française de Linguistique Appliquée

- IV-1, juin 1999, *Grands corpus. Diversité des objectifs, variétés des approches*
- 1-2, décembre 1999, *Corpus, de leur constitution à leur exploitation*
- IV-2, décembre 1999, *L'oral spontané*

Claire BLANCHE-BENVENISTE, 2002, *Approches de la langue parlée en français*. Paris : Ophrys.

Claire BLANCHE-BENVENISTE, Christine ROUGET et Frédéric SABIO, 2000, *Choix de textes de français parlé : trente-six extraits*. Paris : Champion (Collection « Les français parlés, textes et études »).

Le statut patrimonial des enregistrements de paroles

Marie-France Calas

Les linguistes ont été parmi les premiers chercheurs à mettre en œuvre les possibilités techniques offertes au début du 20^e s. par l'invention de l'enregistrement sonore.

Ils ont vu dans cette prodigieuse invention un moyen efficace de faciliter et de rendre leurs collectes plus fiables. Mais par les techniques de l'enquête, par les caractéristiques de l'enregistrement numériques et l'implication des métadonnées, la création des corpus oraux présente de nombreuses similitudes avec la production d'enquêtes orales des ethnologues, des anthropologues, démographes, sociologues, historiens. La question du devenir des enregistrements ainsi créés s'est posée très vite, trouvant une solution originale, mais limitée dans le temps avec la création en 1911, au sein de l'Université, des *Archives de la Parole* par Ferdinand Brunot. Les enregistrements produits par ce service excèdent la production des enquêtes orales réalisées dans les missions devenues aujourd'hui historiques, en ouvrant la série des Voix célèbres. Les *Archives de la Parole* aujourd'hui intégrées au sein du département de l'Audiovisuel de la BNF légitiment une des orientations de cette institution autour de la parole.

Au-delà de la préservation matérielle des enregistrements contextualisés, se pose avec acuité la pérennité patrimoniale de ces corpus et de leur lecture par d'autres usagers de disciplines fort différentes. Dans ce domaine, notre pays a un grand retard. Les documents oraux ont été très longtemps ignorés par les institutions de conservation et par les textes juridiques sur le droit d'auteur et les droits voisins qui ne leur reconnaissent

pas de statut original. Si la B.N.F a intégré en 1977 les collections sonores de la Phonothèque nationale, cela n'a en rien conféré un statut officiel aux documents oraux, consultables en fonction des accords contractuels signés avec les ayants-droits. Ils sont exclus, comme tels, du dépôt légal, seule l'initiative volontariste, contractualisée ou non, peut les intégrer aux fonds sonores et audiovisuels.

Leur présence attestée dans les institutions de conservation montre souvent que les corpus oraux sont considérés, sur un sujet donné, comme des documents d'accompagnement parmi d'autres : l'ensemble des documents collectés dans le cadre de la grande enquête sur l'histoire de la sécurité sociale dirigée par Dominique Schnapper dans les années 70, est considérée comme des archives publiques, non consultables, sauf demande particulière, pendant soixante ans, comme des documents d'illustration de fonds écrits pour les Archives nationales, comme objets d'accompagnement dans les enquêtes orales acquises ou commandées par des musées d'ethnologie ou de société.

Pour l'INA, chargé de la collecte du dépôt légal de la radio et de la télévision, l'oralité est présente en permanence dans les émissions de radio ou de télévision dont la forme est protégée et la consultation très règlementée.

Produits avec rigueur et dans une perspective de préservation par des équipes de recherche, les corpus oraux doivent pouvoir bénéficier d'un statut patrimonial d'objet oral. Cela passe par la reconnaissance scientifique et culturelle de l'oral dans une

société qui l'a si longtemps méprisé.

Mais pour tenir compte de cette part essentielle de notre culture, trop souvent négligée par les institutions nationales, reconnue désormais officiellement par l'UNESCO, il convient de créer des outils spécifiques d'analyse et de trier, comme pour tout document, la masse des collectes à l'aune de critères d'évaluation exigeants et neutres. Le respect de règles techniques et déontologiques de production, reconnues et partagées par tous, constitue des éléments indispensables, à défaut d'être suffisants, pour définir en toute neutralité la part orale de notre patrimoine culturel.

Si la constitution de réseaux universitaires et de recherche offrent des moyens souples et efficaces pour gérer l'accès à ces corpus, les institutions nationales comme la BNF et dans une moindre mesure, les Archives nationales, l'INA, le réseau des musées d'ethnologie et de société doivent pouvoir assumer, de façon partagée, le rôle indispensable et lourd de la conservation pérenne des corpus oraux. L'interopérabilité entre les bases de données et la conscience partagée que l'oral est partie intégrante de notre culture devrait faciliter l'intégration des documents oraux aux collections patrimoniales

Cette évolution devra, dans tous les cas, intégrer une définition claire du statut patrimonial de l'objet oral ●

DES ARCHIVES DE LA PAROLE AU NUMÉRIQUE : les fonds sonores du département de l'Audiovisuel de la Bibliothèque nationale de France

Pascal Cordereix

Bibliothèque nationale de France,
Département de l'Audiovisuel,
Service des documents sonores
pascal.cordereix@bnf.fr

Avec plus d'un million de pièces, la collection d'enregistrements sonores du département de l'Audiovisuel de la Bibliothèque nationale de France est l'une des plus importantes au monde. L'oralité y tient une place importante puisque le fondement historique du département remonte aux *Archives de la Parole*, créées par Ferdinand Brunot en 1911. Depuis, parallèlement au dépôt légal des phonogrammes institué en 1938, l'enregistrement, la conservation, la diffusion auprès du public de la langue et de l'oralité, n'ont cessé d'être au cœur de l'action du département de l'Audiovisuel. On citera pour exemple les enquêtes des *Archives de la Parole* et du *Musée de la Parole et du Geste* entre 1911 et 1953, ou les fonds reçus par la Phonothèque nationale, comme celui des atlas linguistiques régionaux du CNRS, au début des années 1980, etc.

En complément, un certain nombre d'appareils de phonétique expérimentale (issus du laboratoire de l'abbé Rousselot et de l'Institut de phonétique de Paris) et plusieurs centaines de gramophones et de phonographes, sont conservés par le département de l'Audiovisuel.

Aujourd'hui, ce dernier a entamé un vaste plan de sauvegarde de ses collections en

les numérisant. Il s'agit ici non pas d'un simple transfert de support, de l'analogique au numérique, mais bien d'assurer la pérennisation à très long terme de cet archivage numérique, grâce à un stockage sur mémoire de masse informatique. Un autre volet de l'activité du département de l'Audiovisuel, la coopération au plan national et international, s'inscrit d'ailleurs de plus en plus dans cette perspective de l'archivage numérique, le département de l'Audiovisuel recevant aux fins de conservation et de communication des fonds numérisés par d'autres institutions (c'est le cas par exemple d'un certain nombre de fonds sonores relevant du plan de numérisation du ministère de la Culture et de la Communication).

Outre la conservation de cette mémoire de plus d'un siècle d'oralité, cet archivage numérique a également pour objectif d'en faciliter la consultation au delà de la BNF, en permettant progressivement, par exemple, d'en restituer une partie au public le plus large possible grâce à la diffusion en ligne sur internet. C'est ainsi qu'au printemps 2007, l'intégralité des enregistrements effectués par Ferdinand Brunot entre 1911 et 1914 devrait être consultable sur le site Web de la Bibliothèque nationale de France (<http://www.bnf.fr>) ●

Programme « Corpus de la parole »

La DGLFLF s'efforce de mettre en œuvre une action en faveur de la conservation, la numérisation, la mise à disposition, la diffusion et la valorisation des corpus oraux. Ce programme dirigé par le conseil scientifique de l'Observatoire des pratiques linguistiques a d'ores et déjà donné lieu à différentes actions en 2004-2006 :

- la création d'un groupe de travail comprenant des linguistes (CNRS et Université), des juristes, des informaticiens et des conservateurs (BNF, INA, Archives), pour réfléchir sur les questions théoriques et méthodologiques relatives à la numérisation et à l'exploitation des corpus oraux, a abouti à la rédaction d'un « Guide des bonnes pratiques », à la fois juridique, éthique et technique publié aux éditions du CNRS ;
- un inventaire des corpus oraux disponibles ;
- un soutien à différents projets de recherche en partenariat avec les fédérations des laboratoires de recherche en linguistique du CNRS (Institut de linguistique française, ILF-FR 2393, et Typologie et Universaux Linguistiques, TUL-FR 2559) pour la sauvegarde, la constitution et l'exploitation de corpus oraux ;
- la numérisation d'archives linguistiques sonores. Dans le cadre du plan de numérisation piloté par la MRT (Mission pour la recherche et la technologie) du ministère, la DGLFLF a présenté un programme consistant à numériser des fonds sonores du français et des langues parlées en France (numérisation des fonds fragiles dont les supports analogiques sont dans un état de détérioration, numérisation de fonds plus récents pour permettre leur intégration dans une base de données, indexation, catalogage et établissement de normes d'inter-opérabilité), à les valoriser par la création d'un site portail présentant les corpus de français et de langues de France, et à intégrer dans ce site une base de données regroupant une riche collection de corpus desdites langues. Cette base de données permettra une mise à disposition de ressources représentant la diversité des pratiques linguistiques en France ●

Entrevue avec Isabelle de Lamberterie

Isabelle de Lamberterie est directrice de recherche au CNRS, responsable de l'équipe « Normativité et société de l'information » du Centre d'études sur la coopération juridique internationale (CECOJI - UMR 62-24), membre du Comité d'éthique du CNRS et du Conseil supérieur de la recherche et de la technologie (CSRT), elle a lancé et accompagné les travaux de rédaction de l'ouvrage *Corpus oraux, Guide des bonnes pratiques, 2006*.

Langues et Cité : Quels sont les problèmes juridiques que posent la constitution et l'exploitation des corpus oraux ?

Isabelle de Lamberterie : Les questions juridiques se concentrent, principalement, autour de deux domaines : 1. les aménagements nécessaires pour assurer la protection de la vie privée (particulièrement quand le corpus traite de données sensibles et que les finalités de recherche justifient sa conservation ainsi qu'une possible ré-exploitation scientifique) ; 2. les questions de propriété intellectuelle lors de chacune des étapes que sont la constitution, l'exploitation, la diffusion et la conservation des corpus. Ces questions portent, d'une part, sur les « objets » de droit : les contenus, comme les résultats du travail de constitution du corpus peuvent-ils ou non faire l'objet d'une appropriation privative ou rentrent-ils dans le patrimoine commun ? D'autre part, il s'agit de cerner quels sont les titulaires de droits (locuteurs, chercheurs qui interviennent aux différents stades d'élaboration des corpus, institutions qui prennent l'initiative ou gèrent des étapes importantes de la vie du corpus...) ainsi que l'étendue et les limites des droits respectifs de chacun.

Ces questions juridiques sont, par ailleurs, étroitement imbriquées avec les questions de politique scientifique. Il s'agit d'organiser et d'aménager - en tenant compte des cadres juridiques existants - la mise en œuvre des choix stratégiques relatifs à la création, au partage ou à la conservation des objets scientifiques et patrimoniaux que sont les corpus oraux (contenus et contenant). Tout le problème est de déterminer une politique qui permette de reconnaître la responsabilité

scientifique de chacun, sa part de travail tout en favorisant la circulation et la conservation d'un patrimoine commun.

L&C : Le *Guide des bonnes pratiques* est le résultat d'un travail interdisciplinaire entre juristes, linguistes, conservateurs et informaticiens. Quelle est la place des juristes dans cette démarche ?

IdL : On ne peut pas plaquer du droit directement sur les pratiques scientifiques, il est très important qu'un travail conjoint permette aux intéressés de se réapproprier les textes juridiques. Le rapport au droit dans la communauté scientifique, comme dans la société en général, des chercheurs, est souvent limité à l'approche répressive. Et la peur d'une sanction juridique n'est pas forcément, suffisamment, incitative à respecter les règles de droit. La sanction peut même être perçue comme un risque dont on apprécie ou non la probabilité. Il est impératif de créer non seulement une sensibilisation, mais une véritable culture juridique : prendre en considération les enjeux juridiques, c'est apprendre à bien vivre ensemble et prévenir les conflits. Le droit n'est pas seulement répressif, c'est aussi, et surtout, la possibilité de prendre en considération les intérêts de chacun (et pas seulement sur le plan de la gestion des profits d'intérêts). C'est une approche préventive, qui permet une régulation par le biais d'accords et de conventions entre les différents acteurs. Pour cela, le juriste doit être pédagogue sans imposer une réponse univoque sur le permis et l'interdit. En accompagnant cette démarche de régulation juridique, le juriste montre sa capacité de participer à un travail interdisciplinaire.

La force de ce travail interdisciplinaire est que les différents acteurs ont accepté que leurs pratiques soient mises sous les feux de la rampe et littéralement décortiquées, puis ils se sont impliqués dans la lecture des cadres juridiques existants avant de poser directement des questions aux juristes et d'effectuer alors, avec eux, une lecture croisée des textes juridiques. Le *Guide* est le résultat de cette réappropriation, des lectures croisées, de la transposition et de la reprise par les linguistes de textes juridiques. Ce qui le distingue des autres travaux qui contiennent une série de règles à respecter, c'est le résultat de cette invitation faite aux lin-

guistes à comprendre la manière dont ils vont mettre en œuvre les textes, à se réapproprier la mise en œuvre de la législation existante.

L&C : Quelles ont été les différentes étapes de cette démarche ?

IdL : Cette régulation passe par une pédagogie du droit. La première étape nécessite une culture croisée : le juriste écoutant les questions du linguiste sans a priori et le linguiste acceptant de lire les textes juridiques pour reformuler ses questions. La deuxième étape consiste pour le linguiste à accepter d'explicitier d'une façon critique ses pratiques et de les mettre en perspective par rapport à la législation existante. Ce travail doit être accompagné par le juriste.

La troisième étape est l'aménagement par le linguiste de ses propres pratiques en fonction des deux étapes précédentes. Il tire alors des cadres juridiques les éléments nécessaires à l'évolution de ses

pratiques.

Enfin, la quatrième étape permet de faire émerger les points qui font difficulté dans l'état actuel du droit. L'ensemble de ces pratiques s'inscrivent dans le contexte plus large de la société de l'information, que ce soit pour le traitement, la conservation ou la diffusion des données. Comme les corpus sont des données numériques, on peut se demander si le cadre juridique est toujours adapté à l'environnement numérique et aux finalités de recherche.

L&C : Comment se situe la régulation juridique par rapport à une régulation éthique ?

IdL : Même si ces deux formes sont étroitement imbriquées, il convient de les distinguer. Le respect des règles de droit se caractérise par les possibilités de sanctions alors qu'il n'en est pas de même de la régulation éthique. Leurs rapports se manifestent dans la mesure où lorsqu'on

7
parle d'éthique, cela renvoie à une demande de régulation fondée sur des valeurs, ou quand le respect du droit existant se fonde sur des valeurs éthiques. Toutefois, ces distinctions ne sont pas aussi claires : ainsi quand le terme est utilisé dans les documents de l'Union européenne qui exigent la prise en compte d'une dimension éthique dans les programmes de la recherche, il est, souvent, question de respect des normes juridiques.

Au delà de la démarche éthique, la manière dont il est souhaité, dans ce *Guide*, que soient prises en compte de bonnes pratiques par les chercheurs concernés, contribue à créer la confiance et à crédibiliser la recherche. Cet effet induit participe aussi à une valorisation de la recherche dans la société et, par delà, de l'objet scientifique lui-même ●

Corpus Oraux 2006, Guide des bonnes pratiques. CNRS Éditions et PUO. Mai 2006.

Depuis une vingtaine d'années, les études sur les corpus de langues parlées ont complètement renouvelé les sciences du langage. Les toutes nouvelles technologies en matière de stockage, de diffusion, mais aussi d'exploitation des enregistrements sonores, couplées aux outils de traitement automatique du langage (transcriptions synchronisées sur le signal, annotations, etc.) ouvrent des perspectives prometteuses. Toutefois, cette situation ne va pas sans poser de nombreuses questions juridiques et éthiques, mais aussi techniques, méthodologiques et théoriques. Ce sont les réponses à ces questions que souhaitent présenter le *Guide des bonnes pratiques*.

Rédigé par un groupe de travail constitué de linguistes, juristes, informaticiens et conservateurs, cet ouvrage a pour vocation d'éclairer la démarche des chercheurs, de repérer les problèmes et les solutions juridiques et de favoriser l'émergence de pratiques communes pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux ●

Diffusion en librairie, 14 €.

Le projet Phonologie du français contempo- rain (PFC)

Le projet *Phonologie du français contemporain* (PFC) a débuté en 1999 et constituera à terme la plus grosse base de données orales portant sur le français contemporain et l'une des plus grosses bases toutes langues confondues. Il a démarré sur l'initiative conjointe de chercheurs du CNRS et d'universitaires français et étrangers ; il est dirigé par Jacques Durand (ERSS, CNRS – Université de Toulouse Le Mirail), Bernard Laks (MoDyCo, CNRS – Université Paris X Nanterre) et Chantal Lyche (Université d'Oslo).

Pour ces animateurs, le projet veut fournir une meilleure image du français parlé dans son unité et sa diversité, dans la réalité de ses usages attestés et dans sa diversité géographique, sociale et stylistique. En favorisant les échanges pluridisciplinaires entre les connaissances phonologiques et les outils du traitement automatique de la parole, il permettra la constitution de meilleurs matériaux pédagogiques pour la description du français. Enfin, objectif non négligeable, il assurera la conservation d'une partie importante du patrimoine linguistique du monde francophone et, ce, en contrepoint aux corpus déjà constitués.

Ce projet se concrétise par une base de données rassemblant des matériaux recueillis à partir d'un protocole d'enquête uniforme et en prenant appui sur des méthodes d'analyse et des outils développés en commun. L'ambition est d'offrir une vision globale et unitaire du français contemporain, en respectant la diversité des usages de la langue. Sur les enregistrements recueillis, est effectué un considérable travail de transcription et d'alignement du texte sur le signal.

PFC propose alors une structure de consultation des données recueillies et homogénéisées, via les protocoles inter-

net. Une base de données fortement structurée et relationnelle est ainsi accessible avec un simple navigateur. L'interface d'interrogation permet des requêtes larges et fines sur ces données, avec un croisement inédit entre les données documentaires textuelles et les données sonores numérisées. Par contre, l'accès à ces données respecte les usages recommandés par le *Guide des bonnes pratiques pour la constitution, l'exploitation, la diffusion et la conservation des corpus oraux*, avec anonymisation de certaines données et différents niveaux d'autorisations personnalisées.

L'objectif majeur est de construire un corpus favorisant différents niveaux d'approches, adapté donc à différents publics (étudiants, enseignants, chercheurs, ingénieurs). La variété des exploitations possibles est très grande par la mise à disposition d'une ressource à la masse critique importante et aux données standardisées et donc interoperables. L'enseignant ayant besoin d'un tutoriel comparatif de français oral pour des publics même jeunes comme l'ingénieur devant construire un système de reconnaissance vocale pourront se baser utilement sur cette ressource ●

<http://www.projet-pfc.net>

Inventaire des corpus oraux

Paul Cappeau et Magali Seijido

En même temps que l'entreprise du *Guide des bonnes pratiques* se développait et prenait forme, il est apparu nécessaire de compléter ce document par des informations sur les corpus oraux qui existaient aussi bien en France que dans d'autres pays européens. Cet inventaire permet d'avoir une meilleure visibilité des corpus déjà existants ou en cours de constitution et de mieux apprécier leurs caractéristiques (le type d'enregistrements, le support de conservation, la taille, l'état de la transcription, etc.). Il indique aussi si le corpus est consultable : sous quelle forme (son, texte, dans quelle proportion (extrait, totalité) et à quelles conditions (accès sur place,

consultable en ligne) ainsi que la personne ou l'équipe qui doit être contactée. Ces informations apportent un éclairage utile sur l'état des lieux et se prêtent à une double lecture : il existe un nombre finalement assez important de corpus ; toutefois, c'est plutôt une impression d'éclatement qui domine (les choix effectués par les diverses équipes ne sont pas semblables, les corpus recensés pour l'instant restent de taille modeste).

Cet inventaire pourrait faciliter les contacts et les échanges entre équipes, permettre d'identifier les manques les plus flagrants dans le domaine des données orales constituées et aider les futurs projets de constitution de grandes

banques de données à mieux cerner les forces disponibles et les besoins ●

Cet inventaire (partiel à cause des oublis et de l'évolution rapide dans un secteur en pleine activité) sera encore plus utile si les lecteurs aident à l'améliorer et à le compléter.

Toute information utile pourra être communiquée à :

Paul.Cappeau@univ-poitiers.fr

L'enquête ESLO (Enquête Socio-Linguistique à Orléans), conduite par des universitaires britanniques à des fins didactiques en 1968, comprend environ 200 interviews, toutes référencées, et plus de 300 heures de parole incluant des enregistrements cachés, des conversations téléphoniques, des réunions publiques, des entretiens médico-pédagogiques, etc. Ce corpus constitue, par son ampleur et sa cohérence, le plus important témoignage sur le français parlé avant 1980.

Le premier objectif est de numériser les documents sonores, puis d'en proposer une indexation et un premier balisage afin de mettre les données en ligne sur internet. Parallèlement, l'exploitation exhaustive d'un sous-ensem-

ble est engagée. Partant de l'expérience acquise, le CORAL (Centre Orléanais de Recherche en Anthropologie et Linguistique) en partenariat avec d'autres laboratoires (CELITH-MODYCO) a mis en chantier une nouvelle enquête dénommée ESLO2. L'objectif est d'évaluer, à une quarantaine d'années de distance, la dynamique sociale du français (des usages de la langue comme des jugements sur son emploi).

Cette façon de procéder présente l'avantage de préfigurer la référence attendue dans un domaine qui en est encore à se structurer et dans lequel se manifeste de manière récurrente une demande de définition pour un format standardisé de *collecte*, de *conservation*, de *traitement* et d'*analyse* :

- la *collecte* sur le terrain est première, non seulement dans ses aspects techniques, aujourd'hui bien maîtrisés, mais dans la définition du profil de l'échantillon représentatif et dans la problématisation des interactions entre les témoins et les enquêteurs ;
- la *conservation*, qui inclut la préservation des supports, l'indexation des contenus et l'accessibilité (c'est-à-dire la protection) des données, conditionne le partage des sources à des fins d'étude scientifique et d'expertise politique ;
- le *traitement*, en lien étroit avec le développement des matériels et des langages informatiques, suppose la maîtrise d'une chaîne d'opérations, depuis la conversion numérique des enregistrements jusqu'à une transcrip-

tion balisée et ouverte à l'ensemble des interrogations pertinentes ;

- l'*analyse* met les théories (et les logiciels) à l'épreuve des faits.

Avec la constitution et la comparaison de telles enquêtes, les politiques et les acteurs de la transmission linguistique ont à leur disposition un outil d'aide à la décision irremplaçable, qui permet d'appréhender, aussi objectivement que possible, le devenir du français parlé dans toutes ses dimensions. La définition d'un standard rigoureux et réaliste devrait orienter les descriptions du français parlé en France au service de la recherche, des applications et de l'expertise ●

LE PROJET CLAPI

(Corpus de Langue Parlée en Interaction) du laboratoire ICAR

Le laboratoire ICAR (ex-GRIC) (UMR 5191 du CNRS) mène à Lyon des recherches sur les interactions depuis une trentaine d'années et est reconnu internationalement dans ce domaine. Depuis quelques années, une banque de données de Corpus de Langue Parlée en Interaction, CLAPI, est développée dans le but d'assurer la sauvegarde de corpus anciens, à valeur patrimoniale et historique, et de stimuler la production et l'exploitation informatique de nouveaux corpus.

La base CLAPI compte en octobre 2005 :

- 600 h d'enregistrements audio et en partie vidéo, dont 350 h numérisées ;

- 100 h de transcriptions non alignées ;

- 25 h de transcriptions alignées avec le signal sonore et en format XML (150 000 mots) ;

- des corpus d'interactions dans des situations sociales très variées (conversation quotidienne, activités de travail, situations institutionnelles).

La base CLAPI ne se limite pas à gérer des corpus ; elle est avant tout fondée sur un savoir-faire développé par une équipe, qui concerne :

- le *terrain* : recueil de données en situation « naturelle » reposant sur une approche ethnographique ;
- l'*enregistrement des données* ;

- la *transcription* (convention ICOR) ;

- l'*identification des corpus et les métadonnées* : un ensemble fonctionnel de descripteurs adaptés aux corpus oraux (75 rubriques) a été mis au point ;

- les *dimensions juridique, déontologique et éthique* ;

- l'*hébergement sécurisé*.

En ce qui concerne l'accès aux corpus, CLAPI a stimulé les expériences sur la négociation de la diffusion des données interactionnelles, et sur les moyens humains qu'elle requiert, dans le respect du droit et de l'éthique. ICAR a pris l'option de rendre interrogeables en ligne librement, par les outils de la plateforme, des extraits de corpus

choisis par leurs auteurs (en février 2006, une vingtaine d'extraits de corpus, soit 4 h 30, dont 3 h avec signal). Enfin, la conception et développement d'outils de traitement et d'analyse des corpus, permet d'interroger les transcriptions au format xml, de reconnaître automatiquement les variantes graphiques générées par l'usage de « l'orthographe adaptée », de fournir les résultats des requêtes dans un concordancier, avec alignement entre la transcription et les bandes audio/vidéo ●

La plate-forme CLAPI est consultable à l'adresse suivante :

<http://clapi.univ-lyon2.fr>

Les projets soutenus par l'Institut de la langue française (ILF)

Christiane Marchello-Nizia
directrice de l'ILF

La mission spécifique de l'ILF est de favoriser, d'impulser et de développer la synergie entre laboratoires travaillant sur la langue française, et de mettre à la disposition des chercheurs un socle commun de ressources, qu'il s'agisse des grands corpus du français ou des outils logiciels destinés à leur exploitation et à leur analyse.

Actuellement, 21 projets fédératifs effectués en coordination entre plusieurs laboratoires sont en cours (et sept achevés déjà), qui réunissent près de 150 person-

nes. Des acquis considérables et des développements théoriques majeurs sont le résultat de cet effort fédératif.

Depuis 2000, un effort particulier a été dirigé vers les corpus oraux, la France ayant pris un retard notable dans ce domaine. La dotation propre de l'ILF (CNRS) et les contrats spécifiques conclus pour ce faire avec la DGLFLF ont permis de débloquer cette situation de façon remarquable. Tout en continuant de soutenir les corpus écrits, l'ILF a donc collaboré depuis quatre ans à promouvoir le développement de corpus oraux : car désormais si une langue n'offre pas de grand corpus disponible en ligne, elle court le risque d'être minorée.

Cinq projets soutenus par l'ILF et par une dotation spécifique de la DGLFLF sont d'ores et déjà capables de présenter des acquis considérables, sous forme de CD-ROM, de bases de données consultables en ligne, etc. Ce sont :

1. Le Projet PFC : « Phonologie du Français Contemporain » (voir p. 8) ;
2. Le Projet FPI : « Corpus de français parlé en interaction (FPI) : Recueil, numérisation, identification, exploitation », développé au sein des UMR 'ICAR'

(Lyon-2 et ENS-LSH) en relation avec d'autres laboratoires ; en relation avec ces projets, une Grammaire du français parlé en interaction est en cours d'élaboration sous la direction de Catherine Kerbrat-Orecchioni (UMR 'ICAR' Lyon).

3. Le Projet Gardette : « Numérisation du fonds sonore franco-provençal de l'Institut Pierre Gardette », développé également au sein de l'UMR 'ICAR' en liaison avec plusieurs autres laboratoires ;

4. Le Projet THESOC : « Thesaurus occitan », développé au sein de l'UMR de Nice en liaison avec l'UMR 'ERSS' (Toulouse) ;

5. Le Projet PRAX, qui consiste dans le développement d'outils logiciels dédiés au traitement des corpus oraux : « Plateforme de requêtes et d'annotations de corpus en XML », développé à l'UMR 'LPL' (Aix-Marseille).

D'autres projets sont également en cours ou en gestation dans d'autres unités, qui à leur tour seront soutenus, autant que faire se peut ●

Présentation de C-ORAL-ROM

Emanuela Cresti et Massimo Moneglia

C-ORAL-ROM, est une compilation, à des fins de comparaison, de quatre corpus de langue parlée spontanée, pour quatre langues romanes : français, italien, portugais et espagnol. Chaque corpus compte environ 300 000 mots transcrits. L'ensemble comporte 772 textes et représente un peu plus de 123 heures de parole.

Les enregistrements, pris dans des circonstances naturelles, dans des contextes différenciés, font que la base C-ORAL-ROM donne une représentation satisfaisante de ce qu'on peut entendre par « langage parlé spontané », à la fois sur le plan prosodique et syntaxique.

Chaque enregistrement, stocké dans un fichier *wav*, constitue une unité de corpus multimédia accompagnée par le logiciel d'analyse prosodique *WinPitchCorpus* (© Pitch France). *WinPitchCorpus* permet de

faire des alignements texte-son et son-texte, en même temps qu'une analyse acoustique avec tracé de fréquence fondamentale en temps réel, analyse spectrographique, synthèse après consultation de tous les paramètres prosodiques, etc.

Les corpus sont transcrits selon les normes orthographiques standard (format CHAT). Les principaux événements de dialogue y sont représentés : tours de parole, occurrences d'événements non-linguistiques et para-linguistiques, ruptures prosodiques. Une annotation spécifique divise la chaîne textuelle en *énoncés*, délimités par les ruptures prosodiques jugées pertinentes du point de vue de la perception. Des experts spécialisés ont été chargés de ces annotations. Les transcriptions sont alignées sur la source acoustique, énoncé par énoncé, et stockées à des niveaux distincts. Les bases

de données correspondantes, soit en gros 134 000 énoncés, peuvent ainsi être générées automatiquement.

C-ORAL-ROM est accessible sous deux versions. L'une, destinée aux laboratoires et aux industries de la langue, consiste en 9 DVD (fichiers non-compressés et non-cryptés), disponibles par un accord avec l'Agence européenne de distribution des ressources de langues (European Language Resources Distribution Agency). L'autre, destinée aux bibliothèques et aux usages personnels, est livrée sous un format compressé et crypté, en même temps que le livre : E. Cresti et M. Moneglia (eds.), 2005, *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam : Benjamins (Studies in Corpus Linguistics 15) ●

Le projet « Archivage » du LACITO

Michel Jacobson

Le LACITO (Laboratoire de langues et civilisations à tradition orale) est un laboratoire du CNRS dont les chercheurs (linguistes, anthropologues et ethnomusicologues) travaillent depuis plus d'une trentaine d'années à la description de langues pour la plupart sans écritures. De leurs enquêtes de terrain, ils ramènent des enregistrements audio, plus rarement vidéo, ainsi que des transcriptions, des traductions, etc. faites sur place avec l'aide de locuteurs. Ces enregistrements et analyses constituent les matériaux de base qui vont servir aux chercheurs pour poursuivre leurs recherches une fois revenus de leur mission.

Le chercheur durant son enquête sera amené à expliquer les buts de sa mission et tentera d'instaurer une « relation de confiance » entre lui et ses informateurs. Cette confiance est d'autant plus importante que les chercheurs sont parfois amenés à faire d'autres missions sur le même terrain. Elle peut être difficile à obtenir et facile à perdre, y compris par l'intervention ultérieure d'autres enquêteurs (missionnaires religieux, etc.) que les enquêtés risquent de classer dans une même catégorie.

Les enregistrements jusqu'à récemment, servaient principalement aux chercheurs qui les avaient récoltées. Des copies pouvaient en être faites pour des collègues, mais il n'existait ni catalogue ni organisation pour le stockage, la conservation et la copie. Quand un chercheur disparaissait, toutes ces données accumulées risquaient donc de disparaître avec lui. Afin de lutter contre ce risque, un programme appelé « Programme archivage » s'est mis en place au LACITO vers

la fin de années 90. C'est dans ce cadre que de nombreux enregistrements analogiques et notes de terrain ont été numérisés et catalogués.

Ce « Programme archivage » répond à deux buts principaux qui sont 1. la préservation et la pérennisation des données d'enquête et 2. leur valorisation / diffusion.

1. La préservation est assurée par la numérisation des sources. Celle-ci se fait en utilisant des formats et des codages ouverts et libres. Les enregistrements sont numérisés sans compression en qualité CD-Audio. Les notes de terrain sont structurées avec un langage de balisage de texte. Les transcriptions sont codées la plupart du temps avec l'alphabet phonétique international. L'ensemble de ces ressources (fichiers audios et fichiers d'annotations) sont référencées au sein d'un même catalogue. Les champs utilisés pour les décrire sont ceux qui sont préconisés par la communauté scientifique. Les ressources une fois préparées sont déposées dans un *entrepôt de données* où elles seront régulièrement recopiées sur des supports de sauvegarde. Enfin, un accord est en cours de négociation avec la BNF afin que cette institution prenne en charge l'aspect conservation à plus long terme.

2. La valorisation de ces ressources et leur diffusion sont assurées par l'intermédiaire de sites web. Celui du LACITO donne accès, à ce jour, à quelque 150 documents dans une trentaine de langues (principalement des langues de Nouvelle-Calédonie, du Népal et du

Caucase). Une interface de consultation du catalogue des ressources a été définie, qui permet d'effectuer des recherches multi-critères, mais il est possible aussi d'interroger ce même catalogue avec des moteurs de recherches spécialisés comme celui de la LinguistList ou plus génériques comme Google. Une interface de consultation a aussi été définie afin de consulter de manière synchronisée les documents d'enregistrement et leurs annotations. Les outils de consultation, comme ceux qui ont été développés pour la création et la diffusion de ces ressources, sont des *logiciels libres* ●

Parutions

Liselotte BIEDERMANN-PASQUES et Fabrice JEJCIC, *Les rectifications orthographiques de 1990 : analyses des pratiques réelles (Belgique, France, Québec, Suisse, 2002-2004)*. Coll. Les Cahiers de l'Observatoire des pratiques linguistiques, n° 1. Presses Universitaires d'Orléans, 2006, 154 p., préface de Pierre Encrevé.

L'Observatoire des pratiques linguistiques de la DGFLFLF inaugure avec ce n° 1, la série des *Cahiers de l'Observatoire* qu'il consacrera à divers "états des lieux" des pratiques langagières en France. Parmi les analyses proposées, on peut en retenir trois, à titre d'illustration : 1. Les rectifications proposées en 1990 ont été largement adoptées par les dictionnaires, au premier rang desquels celui de l'Académie française. 2. La connaissance des rectifications varie fortement d'un pays à l'autre : ce sont les Français qui les connaissent le moins (nul n'est prophète en son pays !),

mais beaucoup les pratiquent spontanément, en toute ignorance ! 3. Dans les différents pays étudiés, ce sont les propositions touchant l'accent circonflexe qui ont le moins été retenues, celles-là même qui avaient provoqué la plus grande émotion, car elles touchent directement à l'iconicité du signifiant graphique. Articles de L. Bidermann-Pasques et F. Jejcic, JP Simon, R. Muller, M. Lenoble-Pinson.

Les questions du bilingüisme à la Réunion. Les dossiers de l'ARC, vol. 8, Association réunionnaise communication et culture, Paris ; CD audio, 2 x 65 mn.

Ces deux CD audio rassemblent une série de communications enregistrées en 2001 et 2002 à Paris, Marseille et Caen, de H. Gerbeau, P. Fioux, MC Hazaël-Massieux, D., A. Gauvin, D. Caro-Delorme, O. Douville, G. Ramassamy.

Henri BOYER (dir.), *De l'école occitane à l'enseignement public : vécu et représentations sociolinguistiques. Une enquête auprès d'un groupe d'ex-« calandrons »*. L'Harmattan, Paris, 2006, 162 p.

L'enquête par entretiens dont rend compte cet ouvrage concerne un groupe de jeunes gens partageant un vécu scolaire et un apprentissage linguistique : ils ont été scolarisés dans l'une des premières Calandretas (écoles bilingues associatives et laïques), créée en 1979.

Anne ABEILLÉ et Danièle GODARD (dirs), *La syntaxe de la coordination*. *Langages*, n° 160, déc. 2005.

À retourner à

Délégation générale à la langue française et aux langues de France
Observatoire des pratiques linguistiques
6, rue des Pyramides
75001 Paris
ou par courriel :
olivier.baude@culture.gouv.fr

Si vous désirez recevoir **Langues et cité**, le bulletin de l'observatoire des pratiques linguistiques, merci de bien vouloir nous adresser les informations suivantes sur papier libre

Nom ou raison sociale :

Activité :

Adresse postale :

Adresse électronique :

Date :

Ce bulletin applique les rectifications de l'orthographe, proposées par le Conseil supérieur de la langue française (1990), et approuvées par l'Académie française et les instances francophones compétentes.

Langues et cité

Directeur de publication : Xavier North
Président du comité scientifique de l'observatoire : Pierre Encrevé
Rédacteur en chef : Olivier Baude
Secrétaire de rédaction : Jean Sibille
Coordination : Dominique Bard-Cavelier
Composition : Éva Stella-Moragues
Conception graphique : Doc Levin / Juliette Poirot
Impression : Graph 2000

Délégation générale à la langue française et aux langues de France
Observatoire des pratiques linguistiques
Ministère de la Culture et de la Communication
6, rue des Pyramides, 75001 Paris
téléphone : 01 40 15 36 91
télécopie : 01 40 15 36 76
courriel : olivier.baude@culture.gouv.fr
www.dgflf.culture.gouv.fr
ISSN : 1772-757X