

Du son, du texte, des métadonnées.

L'évolution de la banque de données textuelles orales VALIBEL (1989-2006)

Anne DISTER, Michel FRANCARD, Philippe HAMBYE et Anne Catherine SIMON

Centre de recherche VALIBEL, UCLouvain

{prénom.nom}@uclouvain.be

Résumé

La banque de données textuelles orales VALIBEL a été constituée en 1989, dans un contexte particulier : la création d'un Centre de recherche sur la variation en français, qui se voulait un observatoire des pratiques et des représentations des locuteurs belges francophones. Dans cet article, nous retraçons les différentes phases de l'évolution de cette banque de données (la création ; le développement; la révolution tranquille apportée par de nouveaux outils informatiques). Ensuite, nous analysons la manière dont nous avons répondu, par la création d'outils appropriés, aux défis qui se sont posés : comment gérer la représentativité et l'équilibre d'un corpus ouvert ? Comment gérer la variation présente dans les transcriptions et les fichiers d'annotation ? Comment le corpus évolue-t-il en fonction des types d'exploitations (recherches) dont il fait l'objet. Dans une dernière partie, nous expliquons notre politique de diffusion.

1. Historique

1.1. Création de la banque de données textuelles VALIBEL (1988) : contexte et objectifs

La création, fin 1988, de la banque de données VALIBEL s'inscrit dans un contexte marqué par deux courants distincts dans les recherches sur ce qu'on appelait alors le « français parlé » : l'un, inspiré par les travaux des chercheurs du Groupe aixois de recherche en syntaxe (GARS) (Cl. Blanche-Benveniste, C. Jeanjean, etc.); l'autre, dans la ligne des travaux de sociolinguistes québécois soucieux de rendre compte de la variation du français, principalement à l'oral (notamment H. Cedergren, D. Deshaies, S. Poplack, D. Sankoff, P. Thibault, D. Vincent).

Du premier courant, l'équipe VALIBEL retiendra moins la priorité accordée aux phénomènes morphologiques et syntaxiques que l'attention minutieuse portée aux différentes opérations de transcription du matériau sonore et aux problèmes que cette « édition » du corpus soulève, bien au-delà des seules contraintes techniques : illusions auditives, lisibilité, « authenticité »,

etc. (Blanche-Benveniste et Jeanjean 1987). L'apport majeur du second courant sera d'intégrer, dès le départ, la récolte et la transcription de corpus oraux dans une démarche variationniste, attentive à prendre en compte les données sociolinguistiques et situationnelles.

La fin des années 1980 connaît, en Belgique francophone, une émancipation progressive des recherches linguistiques vis-à-vis de la tradition jusqu'alors très prégnante des études normatives (cf. la Belgique « terre de grammairiens »). Le purisme des « chasses aux belgicisms » cède la place à un intérêt pour la description des usages attestés, dont le célèbre *Bon usage* de M. Grevisse se fait lui-même de plus en plus l'écho, au fil des éditions successives.

À la même époque, le développement de supports informatiques conviviaux, accessibles aux chercheurs sans grand investissement préalable, va rendre possible la gestion de corpus textuels d'envergure, sans commune mesure avec les échantillons d'oral disponibles jusqu'alors.

Dans ce contexte, la nécessité de disposer de matériaux fiables pour documenter la variation linguistique du français s'impose avec une particulière urgence pour la Belgique francophone : les données textuelles orales disponibles y sont rares et celles qui existent ont été recueillies sans beaucoup de précautions méthodologiques, en référence à des études très ciblées en dehors desquelles le matériau récolté est quasi inexploitable.

La création du Centre de recherche VALIBEL (acronyme pour *Variétés Linguistiques du français en Belgique*) et, dans la foulée, d'une banque de données textuelles orales informatisée vise donc prioritairement à **documenter la variation linguistique en Belgique francophone** (Wallonie et Bruxelles), sans restriction de domaine – tant le lexique que la morphosyntaxe ou la phonétique pourront être « illustrés » – ou de frontière temporelle : il s'agira, non d'un corpus clos, mais d'une banque de données destinée à être enrichie régulièrement, un *observatoire* du français en Belgique francophone.

1.2. Développement (1989-2002)

Dès le départ, ce projet ambitieux a dû compter avec des limites matérielles qui en ont orienté le développement ultérieur. Disposant de ressources financières limitées, le centre VALIBEL a progressivement alimenté sa banque de données grâce aux apports de chercheurs « juniors »

(mémorants, doctorants) qui intégraient l'exploitation de corpus oraux dans leurs travaux. Cette modalité rendait indispensable l'établissement de protocoles précis pour chacune des étapes menant de l'enregistrement des données orales à leur encodage informatique, chacun des chercheurs adoptant ce « cahier des charges » assez exigeant mais indispensable pour préserver la cohérence et l'homogénéité de présentation des ressources textuelles orales dans la banque de données.

L'importance accordée à la fiabilité des données, à l'identification précise de leur contexte de production et à la mise en évidence de la richesse des matériaux recueillis sera une constante de la première phase du développement de la banque de données textuelles orales VALIBEL. Une attention particulière sera accordée à la mise au point des **conventions de transcription**¹ qui, à ce moment, ont pour finalité de mettre à la disposition d'un large public des textes oraux aisément lisibles² et disponibles sur support informatique.

Une dimension capitale n'est donc pas prise en compte, faute de moyens techniques adéquats : le son. Toutefois, dès le départ, le Centre VALIBEL conservera et gèrera les originaux des enregistrements sonores dans une sonothèque, de manière à ce que phonéticiens et phonologues puissent trouver des réponses aux questions identifiées à la « lecture » des corpus oraux transcrits. Dans le même ordre d'idée, les informations sur les locuteurs, les circonstances d'enregistrement, le contexte d'énonciation, etc. seront récoltées systématiquement, indexées et mises à la disposition des chercheurs désireux de les corrélérer à des faits linguistiques repérés dans les corpus.

À la différence d'un (méga) corpus clos, dont les principes de constitution sont régis par les objectifs que se fixe un chercheur, une banque de données textuelles ouverte est enrichie par des contributeurs qui, dans le cas de VALIBEL, avaient chacun des thématiques de recherche partiellement différentes.

¹ La première version de ces conventions sera mise au point en collaboration avec L. Péronnet, de l'université de Moncton (Acadie), de manière à rendre ces principes de transcription compatibles avec les spécificités linguistiques d'autres aires francophones (Francard et Péronnet 1989). Depuis lors, ces conventions ont été adoptées, intégralement ou légèrement adaptées, par plusieurs équipes de la francophonie « périphérique » (Acadie, Louisiane, La Réunion, etc.).

² Ce souci de lisibilité, notamment pour des non-spécialistes de l'oral, explique des choix comme le recours à une orthographe conventionnelle, la volonté d'éviter des surcharges (notations en API par exemple), sans toutefois que l'oralité des corpus soit totalement gommée.

Cette pluralité des thématiques de recherche (et donc des corpus encodés) aura des conséquences en ce qui concerne l'exploitation des données textuelles orales. À l'exception – notable – d'une recherche en lexicologie différentielle³, la banque VALIBEL n'a pas encore donné lieu à une exploitation de l'ensemble des données disponibles pour des travaux en (morpho)-syntaxe, en phonétique-phonologie ou en analyse du discours, domaines qui ont fait l'objet jusqu'à présent d'études menées sur des échantillons limités de corpus. Cette lacune est en passe d'être comblée, grâce à l'appui de techniques qui rendent ce type d'exploitations plus aisé (voir infra). Par contre, l'étude des attitudes et des représentations des locuteurs francophones de Wallonie et de Bruxelles, une des thématiques de recherche majeures du Centre VALIBEL dans les années 1990, s'est appuyée sur un grand nombre de corpus constitués autour de cette problématique.

Quant à la volonté de limiter les corpus encodés par VALIBEL aux seules productions de francophones belges, dans le souci de combler rapidement une lacune dans cette aire francophone (voir plus haut), elle n'a nullement empêché une concertation régulière avec d'autres équipes, surtout nord-américaines (Acadie, Québec, Louisiane). De ces échanges s'est très tôt dégagée la conviction d'une nécessaire collaboration entre les différentes aires francophones pour mieux identifier, dans une logique différentielle, les spécificités avérées et ce que les francophones ont réellement en partage. Des réseaux de chercheurs comme celui qui donnera naissance à la Banque de données lexicales panfrancophone (BDLP, voir <http://www.tlfq.ulaval.ca/bdlp/>) ou, plus tard, celui constitué autour du projet Phonologie du français contemporain (voir www.projet-pfc.net ; Durand et Lyche 2003) ont ainsi permis aux ressources récoltées par le Centre VALIBEL d'être intégrées dans une approche plus globale, à l'échelle de la francophonie.

La banque de données textuelles orales VALIBEL prendra rapidement son essor et deviendra, au terme d'un « premier lustre » (Francard 1995) la plus importante banque de données textuelles orales de ce type dans la francophonie. Elle atteint, en 1996, trois millions de mots,

³ Précisons qu'il s'agit d'une recherche d'exemples « authentiques » dans le cadre du *Dictionnaire du Français en Belgique*, et non d'un traitement quantitatif du lexique.

au départ de plus de 300 heures d'enregistrements réunissant quelque 450 informateurs⁴. En 2002, les quatre millions de mots sont atteints⁵.

À l'issue de cette première phase, centrée sur les questions techniques et méthodologiques ainsi que sur l'urgence d'atteindre une masse critique de données textuelles, le rythme d'accroissement de l'encodage des corpus va se ralentir, pour faire place à d'autres priorités, impliquant des progrès dans le traitement automatique des corpus (étiquetage morphosyntaxique, accessibilité des données, etc., voir Francard *et al.* 2002).

1.3. La révolution tranquille (2002-2006)

Divers facteurs ont amené le centre VALIBEL à repenser ses pratiques de constitution de corpus, à modifier certains principes de transcription ou d'annotation, et à se doter d'outils répondant à des besoins d'exploitation et de diffusion inédits.

Dès 2002, l'équipe s'est agrandie et le souci s'est fait jour de rendre les données existantes plus accessibles en interne. Parallèlement, les demandes extérieures de consultation des corpus ont augmenté, tandis que la consultation en ligne devenait un standard (alors qu'auparavant les chercheurs se déplaçaient pour venir faire des recherches sur les corpus existants⁶). Pour répondre à cette double demande, nous avons développé une base de données dynamique en ligne pour la gestion, l'archivage et la consultation des **métadonnées** (voir Gilles *et al.* 2006; voir ci-dessous)⁷.

Un second facteur décisif a été le développement du logiciel Praat⁸ et sa popularisation, chez les phonéticiens d'abord puis chez les linguistes travaillant sur l'oral. Praat, ainsi que d'autres

⁴ À titre de comparaison, le corpus du GARS atteint à la même date « environ deux millions de mots » (Blanche-Benveniste 1996 : 27) ; la composante orale du célèbre *British National Corpus* (BNC) comprend, sur un total de 100 millions de mots, 10 millions de mots transcrits de l'oral (Crowdy 1995).

⁵ En mai de cette même année, l'équipe VALIBEL est invitée à une journée ATALA sur le thème de « la constitution et l'exploitation de corpus de français parlé » (journée organisée par Cl. Blanche-Benveniste et J. Véronis le 25 mai 2002). On peut supposer que cette invitation reposait sur l'existence d'une des plus importantes bases de données textuelles orales en français, en termes quantitatifs.

⁶ À la fois parce que la diffusion via le réseau n'était pas aussi performante et sécurisée qu'actuellement, mais aussi parce que cela assurait une forme de contrôle sur l'usage des corpus partagés avec des chercheurs externes.

⁷ Cette base de données dynamique en ligne est comparable à celle développée dans le cadre du projet CLAPI et qui contient les descripteurs des corpus.

⁸ La première version du logiciel d'analyse phonétique date de 1993. Voir Boersma et Weenink (2006).

logiciels⁹, permet de créer et de manipuler des fichiers de type textgrid, qui contiennent une transcription orthographique ou une annotation quelconque alignée avec un fichier sonore. Après la révolution apportée aux études sur le français parlé par l'utilisation du magnétophone portatif, à la fin des années 1950, l'apparition de logiciels comme Praat constitue selon nous une seconde révolution dans les études sur le langage oral.

À VALIBEL, le début des années 2000 a coïncidé avec une numérisation systématique de tous les enregistrements (transcrits ou non) de la banque de données, afin de faciliter leur archivage et leur consultation. Parallèlement, nous entreprenons l'alignement systématique des corpus qui avaient été transcrits en format texte uniquement. Il nous semble aujourd'hui qu'il est indispensable de transcrire chaque nouveau corpus en **synchronisant le texte et le son**. Certes, cette nouvelle pratique peut présenter certains inconvénients, principalement dus aux raisons suivantes :

- pour les transcrip-teurs habitués aux traitements de texte de type Word, la fenêtre de Praat sous laquelle s'effectue la transcription n'est pas un modèle d'ergonomie, même si elle présente l'avantage de combiner l'écoute du son et la saisie du texte ; une pratique régulière du logiciel permet d'effectuer la plupart des opérations via des raccourcis clavier et rend ainsi la transcription au moins aussi rapide que dans un traitement de texte courant ;
- utiliser Praat force à utiliser un logiciel supplémentaire dans la chaîne des outils mobilisés lors de la constitution d'un corpus, voire plusieurs logiciels ou scripts supplémentaires si l'on veut ensuite convertir un textgrid vers un autre format de fichier texte ;
- l'utilisateur perd une partie des fonctionnalités habituelles du traitement de texte (formats de polices ; fonctions d'édition, de recherche, etc.) ;
- certains chercheurs, qui transcrivent des entrevues en vue d'en produire une analyse de contenu, ont l'impression que l'effort supplémentaire requis par la transcription synchronisée n'est pas justifié par leurs objectifs de recherche.

Pour autant, la transcription synchronisée présente des avantages qui compensent les inconvénients évoqués précédemment :

⁹ On peut citer WinPitch ou Transcriber qui permettent également de constituer des fichiers contenant à la fois une transcription ou une annotation et des repères temporels indexant ce contenu textuel à des segments temporels d'un fichier son.

- l'alignement du son et de la transcription redonne une première place à la matérialité sonore ; même si certaines études n'en font pas un usage prioritaire, le retour au support original se révèle parfois utile et la transcription peut être utilisée pour naviguer rapidement dans un enregistrement, dans lequel il sera possible de retrouver les détails de la production verbale ;
- l'absence de recours à des artifices tels que l'italique rend les transcriptions orthographiques transportables vers d'autres applications informatiques ;
- grâce au système des tires multiples de Praat, la transcription orthographique peut être séparée d'autres types de notations qui seront inscrites dans des tires parallèles et superposées à la première (transcription phonétique ou phonémique, codage de divers phénomènes, commentaires métalinguistiques, mise en évidence de phénomènes particuliers, etc.)

En conclusion, on soulignera que l'alignement du texte sur le son dans Praat a eu une conséquence rétroactive sur certaines conventions de transcription, puisque l'accès simultané au son a diminué la nécessité de recourir à des notations de type phonétique. La dissociation plus nette entre une transcription de base et une ou plusieurs annotations spécifiques présente l'avantage de rendre la transcription de base en orthographe conventionnelle réutilisable comme point de départ pour différents codages particuliers réalisés par différents chercheurs, ce qui est difficile quand on amalgame à la transcription orthographique des notations spécifiques.

En ce qui concerne les annotations spécifiques, nous enrichissons progressivement une partie de nos corpus à deux niveaux d'analyse :

- un alignement phonétique et une transcription prosodique : stylisation de la F0 à l'aide de prosogramme (Mertens 2004), réalisée à partir d'un alignement phonétique et syllabique semi-automatique (logiciel EasyAlign, voir Goldman 2007) ;
- un étiquetage morphosyntaxique (Dister 2007).

Les évolutions qui précèdent se sont imposées parce qu'elles procèdent d'une incontestable logique d'enrichissement ou d'amélioration qualitative de la banque de données. D'autres changements supposent des choix qui peuvent être discutés, car ils présentent des avantages mais aussi des inconvénients. Dans la suite de cet article, nous discutons précisément des

gains et des pertes liés à trois aspects des principes de gestion de la banque de données VALIBEL :

- la question de la représentativité des corpus constitués et la possibilité de constituer des « corpus omnibus », c'est-à-dire exploitables à des fins très différentes (voir point 2) ;
- la variation inhérente à la pratique même de la transcription, même lorsqu'elle respecte l'orthographe conventionnelle (voir point 3) ;
- la diffusion des corpus et l'équilibre à trouver entre la diffusion payante et la mise à disposition gratuite, en passant par la mutualisation (échange) des données entre équipes (voir point 4).

2. Gérer un corpus ouvert

2.1. La question de la représentativité

Si elle doit assumer pleinement le rôle d'observatoire du français parlé en Belgique, la banque de données VALIBEL doit-elle prétendre constituer un « corpus de référence » pour le français parlé en Belgique, censé en donner une image aussi complète et fiable que possible ? Elle ne pourrait effectivement remplir cette fonction sans répondre à la question de la représentativité du corpus et sans définir précisément son contenu et les exploitations qu'il rend possibles (Habert 2000 et 2005).

Faire de la banque VALIBEL un « corpus de référence » supposerait en effet de garantir à la fois une représentativité quantitative, en rassemblant une masse de données suffisante, et une représentativité en termes de catégories pertinentes, en veillant à la diversité et à l'équilibre des enregistrements (types d'interactions et profils sociolinguistiques des informateurs, notamment).

Or, étant alimentée au gré des projets de recherche, la base de données s'est rapidement trouvée confrontée à une tension entre ces deux types de représentativité : l'exigence de massification implique d'intégrer tous les enregistrements liés aux projets de l'équipe, ce qui crée un déséquilibre potentiel dans la mesure où certains projets phare contribuent bien plus que d'autres à la production d'un certain type de données. La même concurrence entre ces deux objectifs (quantité et diversité) se joue lorsqu'il s'agit de diffuser les données et de susciter une visibilité à l'extérieur : la valeur d'un corpus se mesure en effet tantôt à partir du

nombre de mots, tantôt en fonction de la disponibilité de données rares ou dont le recueil est couteux (par exemple données d'interactions spontanées recueillies *in vivo*).

Face à ce problème, nous avons décidé de ne pas nous engager dans une voie, mais d'offrir à la communauté scientifique la possibilité d'avoir accès tant à une masse importante mais disparate d'enregistrements de français parlé en Belgique qu'à des sous-ensembles de données plus équilibrés, plus homogènes ou plus spécifiques, répondant à des besoins de recherche précis, grâce au système d'exploitation [moca] (v. ci-dessous).

Ce choix nous permet de contourner le problème de la représentativité dont les solutions sont le plus souvent insatisfaisantes : tout corpus de la langue L, ou d'une variété V de la langue L, qui se veut représentatif, se fonde sur une définition particulière, et donc discutable, de cette langue ou de cette variété, de leurs frontières, du champ de leur variation, etc. C'est pourquoi, en constituant un "réservoir" de données toujours ouvert, la banque de données VALIBEL ne prétend pas *représenter* le français parlé en Belgique, mais vise plutôt, comme nous l'avons indiqué plus haut, à documenter la diversité des pratiques linguistiques observables sur ce territoire.

Depuis l'introduction des premiers corpus, nous effectuons cependant un suivi régulier de l'équilibre de la banque de données afin de lui assurer une diversité aussi grande que possible du point de vue des types de données. Nous tentons de procéder par ajustements successifs selon que tel ou tel indicateur révèle un déséquilibre flagrant dans l'économie générale des corpus encodés : à nouveau, l'importance de certains projets de recherche a pour conséquence que certains types d'interactions (les entretiens entre un informateur et un enquêteur par exemple) ou que certains profils de locuteurs (provenant de telle région, de telle catégorie sociale ou de telle tranche d'âge) sont surreprésentés par rapport à d'autres. Face à ce constat, et en suivant par ailleurs l'évolution des méthodes de recherche en sociolinguistique et en analyse de discours, nous favorisons l'intégration à la banque de données d'enregistrements d'interactions spontanées. Dans le même ordre d'idées, il nous a paru important d'inclure prochainement des enregistrements de locuteurs issus de l'immigration, ceux-ci étant jusqu'ici absents dans la banque de données, alors que l'on sait combien le français parlé aujourd'hui se modifie à travers le phénomène migratoire.

2.2. L'exploitation de données hétérogènes

On peut se demander dès lors si une telle hétérogénéité des données permet encore au chercheur qui les exploite de cerner l'objet sur lequel il travaille. Pour qu'une analyse puisse être menée sur des données aussi variées, il est nécessaire que leur mise en forme respecte certaines conventions et que leur comparabilité soit contrôlée, c'est-à-dire que les données en question soient caractérisées de façon précise, de sorte que le chercheur connaisse les conditions générales de leur recueil (quels informateurs ? quels enquêteurs ? quelles conditions d'interaction ? etc.) et puisse en tenir compte dans l'analyse.

À cet égard, la force de la banque de données VALIBEL est sans doute d'avoir dès l'origine imposé un strict cahier des charges (décrit ci-dessus) à tous les chercheurs recueillant un corpus. Cela implique que les données recueillies selon d'autres protocoles que le nôtre (comme c'est le cas pour les enquêtes menées dans le cadre du projet PFC) soient adaptées aux exigences de la banque de données avant d'y être intégrées. Si la quantité et la précision des méta-données recueillies peut varier sur l'ensemble des corpus VALIBEL¹⁰, nous disposons toujours d'informations minimales permettant de caractériser le corpus en question.

Cependant, il ne suffit pas de posséder une telle information, encore faut-il avoir les moyens de l'exploiter. Une des étapes majeures du développement récent de la banque de données VALIBEL a consisté à se doter d'un outil informatique pouvant traiter les méta-données qui avaient été depuis longtemps encodées sur support informatique, de manière à pouvoir les consulter de façon ergonomique et à pouvoir sélectionner les sous-ensembles de données pertinents pour telle ou telle analyse.

Pour répondre à ce besoin, nous avons développé en collaboration avec d'autres équipes de recherche le système [moca] (Multimedia Oral Corpus Administration)¹¹. Ce système comprend :

¹⁰ Lors de leurs enregistrements, les chercheurs de l'équipe VALIBEL prennent soin de compléter des fiches qui concernent les informateurs et la situation d'enregistrement. Les fiches d'identification des locuteurs sont très précises et certains types d'enquête ne permettent pas de les compléter entièrement, que ce soit parce que les informateurs ne sont pas en mesure de donner les informations demandées (lorsqu'elles concernent le niveau d'études de leur parents par exemple) ou parce que la relation construite avec l'informateur ne permet pas au chercheur de prendre la posture de l'interviewer nécessaire pour remplir les fiches d'identification.

¹¹ Voir <http://cental.fltr.ucl.ac.be:9080/moca> [moca] est un logiciel libre développé par VALIBEL en partenariat avec le CENTAL (Centre de traitement automatique du langage, UCLouvain) et les universités de Freiburg (Allemagne) et du Luxembourg.

- une base de données dynamique qui contient toutes les méta-données sur les enregistrements, sur les locuteurs et sur les corpus ;
- les données primaires (enregistrements sonores, téléchargeables en entier ou consultables par fragments) ;
- des données secondaires (fichiers contenant des transcriptions orthographiques ou des annotations), accompagnées elles-mêmes d'une description (auteur, date, type, symboles utilisés, etc.).

Les méta-données peuvent être aisément consultées pour chaque enregistrement ou groupe d'enregistrements. Un moteur de recherche permet de constituer des sous-corpus parmi les données existantes, en fonction de critères sur les locuteurs (par ex. des locuteurs âgés), sur les enregistrements (trilogues, interactions scolaires, informelles, etc.) ou sur les annotations disponibles (transcription phonétique).

À l'aide de [moca], les chercheurs disposent ainsi d'une série indéterminée de *corpus virtuels* répondant à des besoins très variés. Dans ce cas, ce qui fait l'intérêt du corpus exploité, ce n'est donc pas tant le nombre de mots que la possibilité de sélectionner les données les plus susceptibles de valider l'analyse linguistique ou de l'appliquer à des objets nouveaux et précis. Le principe des corpus virtuels autorise la logique d'ouverture de la banque de données, puisqu'il sera par la suite possible d'éliminer de l'analyse tout matériau qui paraîtrait trop hétérogène.

En résumé, le système [moca] permet de rendre exploitables pour d'autres recherches les corpus constitués préalablement, en tentant de donner au chercheur qui y accède des informations explicites qui sont en règle générale connues uniquement de l'enquêteur ayant constitué chaque corpus. C'est également grâce à [moca] que nous pouvons facilement avoir une radioscopie de nos données et connaître les caractéristiques de la banque de données à un moment défini.

3. Gérer la variation des transcriptions et des annotations

Nous l'avons dit, dès sa création, le centre de recherche VALIBEL a réfléchi à ses pratiques de transcription et a établi des conventions de transcription explicites (cf. Francard et Péronnet 1989, Dister et Simon, à paraître). Néanmoins, force est de constater que malgré la

publication d'un guide à l'usage des transcrip-teurs, on ne peut évacuer un problème central dans la constitution des données secondaires : celui de la variation.

Le premier facteur qui engendre de la variation dans les transcriptions est lié à des difficultés d'écoute (Blanche-Benveniste et Jeanjean 1987). Celles-ci sont diverses, conditionnées par la qualité des enregistrements, le nombre des locuteurs qui interviennent dans la conversation, la variété de langue de ceux-ci, etc.

Outre ces problèmes « techniques », c'est de la pratique même de transcription que naît la variation. En effet, c'est une évidence maintenant de dire que la transcription relève d'une construction : transcrire n'est pas une activité neutre et transparente faite par un quelconque copiste (Bilger 2000). Au contraire, transcrire est une activité sélective et interprétative, qui engage une théorie et oblige à poser des choix (voir l'article fondateur de Ochs 1979 ; Edwards 1993).

Au centre de recherche VALIBEL, nous transcrivons orthographiquement (mais sans ponctuation), refusant tout trucage qui viserait à rendre compte de la phonétique par une adaptation de la graphie standard (Blanche-Benveniste et Jeanjean 1987). Ces transcriptions laissent inévitablement de côté une série de phénomènes jugés intéressants par certains chercheurs, mais que nous avons décidé de ne pas intégrer dans la version textuelle « de base » (voir ci-dessus).

Mais même lorsque l'on a posé les grands principes, la pratique des différents transcrip-teurs peut varier : on citera par exemple l'usage non uniformisé des indications qui relèvent du non-verbal (toux, rires, etc.) dont on voit que les transcrip-teurs font un usage parfois très différent. Un autre facteur de variation important dans nos transcriptions concerne les indications de la pause. En effet, si nos transcriptions ne sont pas ponctuées, elles contiennent néanmoins des indications sur les pauses, réparties en trois catégories selon leur durée : pause brève, pause longue et silence. La pause ayant une réalité physique (un blanc dans un continuum de signal sonore), on pourrait s'attendre à ce que cette réalité soit objectivable, et donc facilement

décelable, et cela de manière identique par les transcrip-teurs. Or, on constate qu'il n'en est rien et les tests de perception effectués par certains chercheurs concordent sur ce point ¹².

La pause est jugée intuitivement par le transcrip-teur, et non mesurée à l'aide d'appareils d'analyse acoustique. Il s'agit donc dans nos corpus d'une marque subjective¹³, laissée à l'appréciation du transcrip-teur. Celui-ci prend en compte plusieurs facteurs dont le principal est sans doute la vitesse d'articulation¹⁴. La pause est donc notée relativement à la vitesse d'articulation et, ainsi, pour une même durée de blanc dans le continuum sonore, selon que le locuteur parle plus ou moins vite, une pause d'une durée pourtant équivalente sera ou non marquée par le transcrip-teur selon le locuteur auquel elle est attachée. Lors de la révision d'une transcription par un second chercheur, c'est assurément sur l'appréciation des pauses que les jugements divergent le plus.

En fait, quand l'on travaille sur des transcriptions de l'oral, il faut être conscient que l'on travaille sur des données construites : deux transcrip-teurs, formés dans la même équipe et suivant les mêmes conventions, ne fourniront jamais deux transcriptions identiques.

Certaines variations concernent la graphie, quand le recours aux ouvrages de référence ne permet pas toujours de trancher. Ainsi, dans les transcriptions de la banque de données VALIBEL, nous trouvons les trois variantes graphiques pour le préfixe *hyper* :

graphie avec soudure : *hypercorrectisme* [accFJ1r], *hypertension* [norPM1]

graphie avec trait d'union : *hyper-spécialisation* [iljPF1r], *hyper-riche* [ilpMJ1]

suite de deux mots séparés par un blanc : *hyper accentué* [accCP1r], *hyper joli* [famVV1]

Ces exemples, où la variation peut sembler anecdotique, posent néanmoins certains problèmes, notamment lorsque la masse de données textuelles s'accroît et qu'elle doit être consultée à l'aide de logiciels informatiques¹⁵.

¹² Voir Candéa (2000 : 112 et sv.) pour un bref résumé d'autres tests de perceptions sur les pauses et les phénomènes dits d'hésitation.

¹³ Dans le sens où elles sont laissées à l'appréciation du transcrip-teur ; nous ne prétendons pas que toutes les pauses sont subjectives au sens de Duez (1991) et de Candéa (2000) (= non présentes dans le signal sonore).

¹⁴ Lane et Grosjean (1973) ont montré « qu'un sujet qui articule rapidement ne présente pas automatiquement un temps de pause réduit et vice versa » (cité par Grosjean et Deschamps 1975 : 156).

¹⁵ Notons que nos conventions de transcription, revues en 2004, sont totalement compatibles avec une utilisation informatique des données.

C'est typiquement le cas lors du traitement automatique de texte. En effet, d'un point de vue informatique, on a un seul mot graphique lorsqu'il y a soudure, et on comptabilise deux mots graphiques quand un blanc ou un trait d'union sépare les composants. Pour l'analyse lexicale avec le logiciel de traitement de corpus Unitex (Paumier 2006) que nous utilisons, la variation graphique que nous illustrons avec le préfixe *hyper* n'est pas mineure. En effet, Unitex utilise des dictionnaires électroniques à large couverture (Courtois et Silberztein 1990). Pour être reconnu et analysé par le système, un mot doit donc être répertorié au préalable dans les dictionnaires. Dans le cas de la composition, on peut reconnaître des formes non répertoriées comme des formes composées, si chacun des termes de la composition est lui-même répertorié dans les dictionnaires (*hyper* comme préfixe et *spécialisation* comme nom, par exemple). Dans le cas de la soudure, cette reconnaissance des composés n'est pas possible à l'heure actuelle, et les mots sont considérés comme des mots inconnus par le système s'ils ne figurent pas en tant que tels dans le dictionnaire.

On a évoqué jusqu'ici la variation entre différentes transcriptions dans un même corpus. La variation peut aussi toucher deux versions d'une même transcription, ce qui montre que la variation est inévitable, même dans les transcriptions réalisées avec le plus grand soin. Pourtant, le chercheur a besoin de travailler sur des objets stables.

En effet, le fait que les transcriptions évoluent, en fonction des corrections qui y sont apportées, rend les résultats d'une recherche faite sur une version *alpha* impossibles à reproduire sur une version *bêta*, alors que ces deux versions correspondent aux mêmes données primaires. On objectera que pour que des résultats soient reproductibles, ils doivent partir des mêmes données. Mais qui a envie de repartir de données dont on sait qu'elles sont moins bonnes que d'autres dont on dispose par ailleurs ?

Cette évolution des textes a sans doute moins de répercussions quand on traite de grandes masses de données que des énoncés plus petits. Ainsi, on sait que de nombreuses « erreurs » dans les transcriptions concernent la notation des « disfluences » propres à l'oral (répétitions, amorces, etc. ; voir Cappeau 1997 et Pallaud 2002). Les corrections qui concernent les disfluences ont sans doute moins d'incidences sur les résultats d'une étude statistique faite sur un million de mots que lorsque l'analyse en profondeur d'un bref énoncé prend pour base de

sa démonstration une séquence disfluente qui sera corrigée dans une version ultérieure du texte.

De plus, dans le cadre d'un centre comme VALIBEL, où plusieurs chercheurs travaillent sur les mêmes données, se pose la question cruciale de la gestion des textes qui évoluent au cours du temps. Actuellement, nous menons une réflexion en fonction des outils informatiques qui permettent de gérer ces mises à jour, qui produisent des transcriptions toujours meilleures, mais jamais définitives.

4. Politique de diffusion

La « diffusion » liée à notre base de données concerne évidemment les données elles-mêmes (enregistrements sonores, transcriptions orthographiques, alignement phonétique, etc.) mais aussi les logiciels que nous développons pour leur gestion et leur exploitation.

Concernant les logiciels, notre politique est très claire : nous souhaitons développer des outils en partenariat avec d'autres équipes¹⁶ et rendre ces outils entièrement disponibles à la communauté scientifique, dans la logique de l'open source. Cette politique présente selon nous le double avantage de favoriser l'amélioration des outils¹⁷ et de favoriser leur utilisation par le plus grand nombre d'équipes. À nouveau, un modèle du genre est Praat – dont on ne doit pas vanter l'efficacité ni la large diffusion.

Concernant les données elles-mêmes, la réponse est plus nuancée :

- pour un corpus fraîchement recueilli, nous souhaitons laisser au chercheur responsable de la constitution la primeur des premières publications ;
- pour un corpus dont certains résultats d'analyse ont été publiés (comme c'est le cas pour les analyses sur l'insécurité linguistique en Belgique francophones réalisées à partir des interviews d'étudiants, de journalistes, de cadres et de politiciens), nous cédon les données à des chercheurs qui souhaitent traiter d'autres aspects. Notre politique est de lier chaque diffusion à un objectif de recherche précis et à une

¹⁶ C'est le cas actuellement, via des collaborations avec le Cental (UCL), les universités de Freiburg, de Genève, et de Saint-Denis de La Réunion (le projet VALIRUN dirigé par G. Ledegen).

¹⁷ Les utilisateurs proposent de nouvelles fonctionnalités qui sont intégrées si elles répondent aux exigences de qualité.

durée d'exploitation, et non de diffuser les données par simple téléchargement à partir de notre site¹⁸.

En fait, à l'heure où l'on prône l'échange et la mutualisation, il paraît aberrant que chaque centre de recherche garde jalousement les corpus qu'il a constitués. Si cette réaction peut s'expliquer par le coût exorbitant de constitution de corpus oraux (en termes d'argent, de temps, d'investissement humain)¹⁹, il nous semble pourtant que la meilleure manière de garder nos données en vie est de permettre qu'elles soient réutilisées par d'autres équipes, avec des objectifs de recherches différents.

5. Conclusion

La réflexion sur la constitution des corpus oraux s'est développée au point de constituer parfois un objet d'étude en soi ; pour autant, elle n'a de sens que si elle permet la réalisation de recherches et si elle contribue effectivement au progrès de la connaissance. Or, il ne suffit pas pour cela d'accumuler des données, mais il faut que ces données répondent aux besoins des chercheurs, besoins qui dépendent de leurs objets de recherche, de leurs méthodes et de leurs buts.

En jetant ainsi un regard rétrospectif sur l'évolution de la banque de données VALIBEL, on constate à quel point la constitution de corpus suppose une série de choix qui engagent les chercheurs au-delà des questions de la qualité scientifique et technique des données produites. Les principes décrits ci-dessus se fondent en effet sur des conceptions particulières du langage, de ce que sont des données pertinentes, de ce qui est souhaitable en termes de partage de données et de collaboration entre chercheurs. Ainsi, s'il est indispensable à nos yeux de disposer des métadonnées qui accompagnent les enregistrements de notre banque de données, c'est parce qu'une analyse linguistique qui négligerait complètement le contexte social et interactionnel des pratiques langagières nous paraît inconcevable. Si la possibilité d'un recours direct au matériau sonore nous semble primordiale, c'est parce que la plupart des recherches que nous menons reposent sur une analyse multidimensionnelle des interactions, que la transcription seule ne peut qu'imparfaitement rencontrer. Enfin, si la possibilité de

¹⁸ Nous signons ainsi un contrat avec les utilisateurs des données que nous diffusons, ce qui évite qu'« elles se perdent dans la nature ».

¹⁹ Si les corpus devaient se vendre au prix coûtant, ils seraient inabordables pour les équipes scientifiques rattachées aux universités. De plus, se pose la question éthique de la vente de « biens culturels » constitués grâce à des fonds publics.

partager des mêmes données entre plusieurs chercheurs fait partie des visées qui fondent la gestion de notre banque de données, c'est parce que la possibilité de croiser des analyses complémentaires sur un même corpus fait partie de nos objectifs de recherche à long terme.

Bien d'autres options auraient pu être privilégiées, en réponse à d'autres besoins (comme par exemple celui de disposer d'un enregistrement vidéo des interactions, ou comme celui d'augmenter davantage le nombre de mots total de la banque de données) et en fonction d'autres logiques, ou d'une autre histoire. Derrière les choix de développement d'une banque de données telle que VALIBEL, s'inscrit en effet toute l'évolution d'une équipe de recherche et de sa position dans le champ de la (socio)linguistique.

Références bibliographiques

BILGER Mireille (2000). « Petite typologie des conventions de transcription de l'oral. Quelques aspects pratiques et théoriques », *Linguistique sur corpus. Études et réflexions*. (Mireille Bilger coord.), *Cahiers de l'Université de Perpignan*, 31, Presses universitaires de Perpignan, pp. 77-92.

BLANCHE-BENVENISTE Claire, JEANJEAN Colette. (1987). *Le français parlé. Transcription et édition*. Paris : Didier Érudition.

BLANCHE-BENVENISTE Claire. (1996). « De l'utilité d'un corpus linguistique », *Revue française de linguistique appliquée* 1 (2), p. 25-42.

BOERSMA Paul, WEENINK David (2006). Praat : doing phonetics by computer (Version 4.5.08) [Computer program]. Retrieved December 20, 2006, from <http://www.praat.org/>

CANDEA Maria (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané*, Thèse non publiée.

CAPPEAU PAUL (1997). « Données erronées : quelles erreurs commettent les transcrip-teurs ? », *Recherches sur le français parlé*, 14, Université de Provence, pp. 117-126.

CLAPI (2007). « Corpus de Langues Parlées en Interaction. Banque de données et plateforme logicielle », page consultée le 8 janvier 2007, <http://clapi.univ-lyon2.fr>

COURTOIS Blandine, SILBERZTEIN Max (Éds) (1990). *Dictionnaires électroniques du français, Langue française* 87, Paris, Larousse.

CROWDY Steve (1995). « The BNC spoken corpus », *Spoken English on computer. Transcription, mark-up and application* (G. Leech, G. Myers, T. Jenny Eds), New York, Longman, pp. 224-234.

DELIC (2004). « Présentation du Corpus de référence du français parlé », *Recherches sur le français parlé* 18, pp. 11-42.

DISTER Anne (2007). *De la transcription à l'étiquetage morphosyntaxique de corpus de parole. Le cas de la banque de données VALIBEL*, Thèse de doctorat, Université de Louvain.

DISTER Anne, SIMON Anne Catherine (à paraître) « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », *Arena Romanistica* 1.

DUEZ Danielle. (1991). *La pause dans la parole de l'homme politique*, Paris, Éditions du CNRS.

DURAND Jacques, LYCHE Chantal (2003). « Le projet *Phonologie du français contemporain* (PFC) et sa méthodologie », in E. Delais & J. Durand (eds), *Corpus et variation en*

phonologie du français : méthodes et analyses. Toulouse : Presses Universitaires du Mirail, pp. 213-278.

EDWARDS Jane A. (1993). « Principles and Contrasting Systems of Discourse Transcription », *Talking Data. Transcription and in Coding Discourse Research* (J.A. Edwards and M.D. Lampert Eds), Hillsdale, Lawrence Erlbaum Associates, pp. 3-31.

FRANCARD Michel, PÉRONNET Louise (1989). « La transcription de corpus oraux dans une perspective comparative. La démarche du projet PLURAL ». *Actes du colloque La description des langues naturelles en vue d'applications informatiques* (Québec, université Laval, 7-9 décembre 1988) (Conrad OUELLON éd.), Centre international de recherche sur le bilinguisme (publication K-10), Québec, p. 295-307.

FRANCARD Michel (1995). « L'oral, un bon investissement ? La banque de données VALIBEL: bilan d'un premier lustre ». Dans *Présence francophone* 46, p. 9-34.

FRANCARD Michel, GERON Geneviève, WILMET Régine (2002). « La banque de données VALIBEL : des ressources textuelles orales pour l'étude du français en Wallonie et à Bruxelles ». Dans Pusch, Claus D. & Raible, Wolfgang (éd.), *Romanistische Korpuslinguistik – Korpora und gesprochene Sprache / Romance Corpus Linguistics – Corpora and Spoken Language* (= ScriptOra; 126). Tübingen : Gunter Narr, p. 71-80.

GILLES Peter, KEVERS Laurent, SIMON Anne Catherine (2006). « [moca], un système de gestion et d'annotation de données orales, communication », communication présentée à la 3e rencontre fribourgeoise de la linguistique sur corpus appliquée aux langues romanes, Freiburg-im-Breisgau, 14-17 septembre 2006.

GOLDMAN Jean-Philippe (2007). *EasyAlign, script d'alignement phonétique semi-automatique sous Praat, version du 10 septembre 2007*, <http://latlcui.unige.ch/phonetique/easyalign>

GROSJEAN François, DESCHAMPS Alain. (1975). « Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation », *Phonetica* 31, pp. 144-184.

HABERT Benoît (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? », *Linguistique sur corpus. Études et réflexions*. (Mireille Bilger coord.), *Cahiers de l'Université de Perpignan*, 31. Presses universitaires de Perpignan, pp. 11-58.

HABERT Benoît (2005). « Face à la disette dans la profusion », *SCOLIA* 19, pp. 41-61.

LANE Harlan, GROSJEAN François (1973). « Perception of reading rate by listeners and speakers », *Journal of Experimental Psychology* 97 (2), pp. 141-147.

MERTENS Piet (2004). « Le prosogramme : une transcription semi-automatique de la prosodie », *Cahiers de l'Institut de Linguistique de Louvain* 30 (1-3), pp. 7-25

OCHS Elinor (1979). « Transcription as theory », *Developmental pragmatics* (E. Ochs et B. B. Schieffelin Éds), New York, San Francisco, London, Academic Press, pp. 43-72

PALLAUD Berthille (2002). « Erreurs d'écoute dans la transcription de données orales », *Revue Parole* 22-23-24, pp. 267-294.

PAUMIER Sébastien (2006). *Unitex 1.2. Manuel d'utilisation*, <http://www-igm.univ-mlv.fr/~unitex/manuel.html>