Sur la constitution de corpus de deux langues à tradition orale (le berbère tamazight et l'arabe marocain) parlées à Orléans

On the Constitution of a Corpus of Two Languages of Oral Tradition (Berber Tamazight and Moroccan Arabic)

Samira Moukrim

Édition électronique

URL : <http://journals.openedition.org/praxematique/1186>

DOI : 10.4000/praxematique.1186

ISSN : 2111-5044

Éditeur

Presses universitaires de la Méditerranée

Édition imprimée

Date de publication : 1 janvier 2010

Pagination : 367-386

ISBN : 978-2-36781-012-6

ISSN : 0765-4944

Référence électronique

Samira Moukrim, « Sur la constitution de corpus de deux langues à tradition orale (le berbère tamazight et l'arabe marocain) parlées à Orléans », *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 21, mis en ligne le 01 janvier 2013, consulté le 24 septembre 2020. URL : <http://journals.openedition.org/praxematique/1186> ; DOI : <https://doi.org/10.4000/praxematique.1186>

Samira Moukrim
L.L.L. — Université d'Orléans

Sur la constitution de corpus de deux langues à tradition orale (le berbère tamazight et l'arabe marocain) parlées à Orléans

Introduction

Pour étudier l'expression du « présent » en français, berbère tamazight et arabe marocain, nous avons fait le choix de travailler sur des données orales authentiques. C'est la diversité d'emplois de la forme du « présent » en français (présent actuel, habituel, gnomique, de vérité générale, narratif, historique, etc.¹) qui a suscité notre intérêt et nous a poussée à explorer la question dans d'autres langues, notamment en arabe marocain et en berbère tamazight². Nous nous sommes interrogée sur l'existence et l'expression de ce phénomène dans deux langues

1. Le présent de l'indicatif en français continue à poser problème malgré les nombreuses mises au point dont il a fait l'objet. Le débat sur ses valeurs et ses emplois est loin d'être clos. Les linguistes proposent des explications tout aussi variées selon la théorie adoptée. Pour les uns (DAMOURETTE & PICHON (1911-1926/1970), MELLET (1980-2000), SERBAT (1980-1988), CHUQUET (1994)), le présent est une forme neutre, capable d'inscrire le procès dans n'importe quelle époque. Pour d'autres (BENVENISTE 1970/ 1974, GUILLAUME 1929, GOSSELIN (2005), HAILLET (2005)), le présent est une forme temporelle déictique, marque de coïncidence entre le moment d'énonciation et le moment du procès. Une troisième hypothèse est avancée par WILMET (1997 : 341) qui définit le présent comme étant une « forme verbale qui affirme la concomitance d'un procès au repère de l'actualité ». Avis partagé par BEAUZEE (1782-1786/1986), JAUBERT (2001), REVAZ (2006). Cf. MOUKRIM (2010 : 295-309) pour plus de détails sur la diversité des traitements des emplois du présent de l'indicatif en français.

2. L'étude de ce phénomène dans des langues différentes constitue un enjeu théorique pour la linguistique. Comme le précisent FUCHS & ROBERT (1997 : 1) : « il n'est sans doute pas exagéré de dire que la question centrale pour la linguistique est celle de la diversité des langues : c'est en effet à partir des langues — au pluriel — que la linguistique tente d'appréhender le langage — au singulier. » Dans ce travail, l'enjeu est de taille car les systèmes verbaux des langues étudiées sont très différents.

dont les systèmes verbaux reposent sur une opposition purement aspectuelle.

Pourquoi une énième étude sur le « présent de l'indicatif » en français ? Pour deux raisons : i) Explorer la *démarche comparative* ii) travailler sur des *données orales authentiques*. Nous supposons que la difficulté de rendre compte de la multiplicité d'emplois de la forme du « présent » en français pourrait être élucidée en examinant le fonctionnement de formes équivalentes dans d'autres langues. Nous supposons également que l'observation de la langue parlée pourrait apporter du nouveau sur la question, car la plupart des études fabriquent des exemples, ou empruntent des exemples à l'écrit¹. Ce travail consistait donc en deux volets i) la constitution du corpus et ii) l'observation des données.

La constitution du corpus n'est pas une tâche facile, surtout qu'il s'agit de corpus de deux voire trois langues différentes. L'hétérogénéité des corpus sur lesquels nous avons travaillé se manifeste tant au niveau de la *langue*, différente d'un corpus à l'autre, qu'au niveau de sa *codification* graphique et du degré de son *informatisation*. Nous nous sommes confrontée à des problèmes assez différents. Nous avons essayé de voir en quoi la différence des corpus implique une spécification des méthodes et outils, en particulier si les outils et instruments utilisés pour les corpus oraux de langues à grande diffusion (notamment celui d'ESLO) s'ajustent aux langues à tradition orale.

Dans ce papier, il sera question des différentes étapes de la constitution des corpus de l'arabe marocain (désormais AM) et du berbère tamazight (désormais BT), des choix opérés et de la démarche adoptée sur le terrain. Seront abordés également les aspects juridiques du corpus.

I. Contexte

Pour les trois langues à l'étude (AM, BT et français désormais FR), notre corpus a été constitué à Orléans. Cette unité de lieu (Orléans) se justifie d'abord par des raisons pratiques : nous travaillons dans le Laboratoire Ligérien de Linguistique (L.L.L. — université d'Orléans et Tours), qui pilote le projet Enquêtes Socio-Linguistiques à Orléans

1. Une étude basée sur des données empiriques orales permet non seulement d'observer les emplois de la forme du « présent » qui sont en rapport direct avec le moment de la parole mais aussi de déterminer les paramètres qui font que cette même forme énoncée dans l'actuel puisse renvoyer au non-actuel.

(ESLO¹), l'un des corpus les plus importants du français oral. C'est dans ce dernier que nous avons puisé les données du français.

Quant à celles de l'arabe marocain et du berbère tamazight², nous avons constitué le corpus auprès de locuteurs marocains arabophones et berbérophones résidant à Orléans. Notre objectif était de recueillir un échantillon authentique de ces deux langues pour les comparer avec le français parlé à Orléans. Nous avons voulu parallèlement à travers cette étude dessiner une image de l'arabe marocain et du berbère tamazight parlés actuellement en France, à Orléans. Ces deux langues³ qui, depuis la signature de la Charte européenne des langues régionales et minoritaires en 1999, figurent parmi les « langues de France non territoriales⁴ ».

Cette étude s'inscrit également dans le cadre du programme Langues en contact à Orléans (LCO), module d'ESLO, qui étudie la vie des langues en contact avec le français, d'où le choix des thématiques proposées (langues parlées/ en contact, représentation des langues, culture, traditions...).

2. Méthodologie

Notre corpus a été constitué en trois étapes⁵, qui représentent autant de moments de choix méthodologiques :

a Le travail préparatoire à la phase d'enregistrement ;

b L'enregistrement des données ;

1. Cf. site d'ESLO : www.univ-orleans.fr/eslo pour plus de détails sur le programme ESLO.

2. Il s'agit du berbère parlé par des témoins berbérophones originaires du Maroc central (dialecte *tamazight*) et de l'arabe marocain parlé par des témoins arabophones originaires de différentes villes du Maroc (Casablanca, Rabat, Berkane, Khémisat, Meknes, Khénifra...).

3. Pour plus de détails sur la situation de l'arabe dialectal et du berbère en France, cf. MOUKRIM (2010 : 42-55).

4. À signaler que la Charte des langues régionales et minoritaires, bien que signée le 7 mai 1999, n'a pas encore été ratifiée par la France. La Charte étant contraire à la Constitution, la France ne peut la ratifier sans engager une révision constitutionnelle qui rendrait les deux textes compatibles.

5. E. Delais-Roussarie (2003 : 91-92) a présenté les principales étapes de la constitution de corpus oraux en prenant appui sur la littérature traitant des problèmes de constitution des corpus écrits (cf. entre autres, HABERT *et al.* 1997 a et b).

c La mise en forme des données à des fins d'exploitation (transcription et annotation).

Pour la constitution du corpus, nous avons pris en compte un certain nombre de critères pertinents pour définir la sélection des données et décider des types d'enregistrements, au nombre desquels : i) les objectifs du travail, sa finalité ii) le type de données requis pour satisfaire ces objectifs iii) les modalités de collecte.

Le corpus constitué a pour fin une étude des formes verbales du « présent » en AM et en BT parlés à Orléans. Pour répondre aux exigences de cette recherche, le corpus doit permettre d'attester les formes verbales du présent dans leurs différents usages. Se pose alors le problème de savoir quelles sont les données qu'on peut considérer comme représentatives de ce phénomène ?

La représentativité du corpus étant difficilement envisageable¹, nous avons essayé de l'améliorer, comme le précise à juste titre Habert (2000) : « Améliorer la représentativité d'un corpus consiste à préciser la production et la réception de chacun de ses composants, en lien avec les motifs qui ont conduit à la création du corpus, mais aussi à pouvoir déterminer sur des bases objectivables les différents emplois du langage auquel on s'intéresse. »

Dans ce travail, nous avons essayé de diversifier les situations enregistrées ainsi que les catégories de locuteurs en différenciant sociologiquement les témoins par l'âge, le sexe, le niveau scolaire, la profession et les langues parlées. Nous avons tenu également à documenter les données ainsi que leur contexte de production.

Partant du fait que les données recueillies sont toujours l'effet d'un conditionnement consécutif au procédé d'enquête, les modes de recueil de notre corpus ont été choisis dans l'esprit de ce que souligne Maurer B. (1999) : « Le choix du mode de recueil des données constitue même une étape fondamentale dans la démarche de recherche puisque l'objet effectivement étudié en dépend étroitement. »

1. Pour A. METTOUCHI & A. LARCHERET-DUJOUR (2006), un corpus de référence pour les langues à petite diffusion peut « se définir comme une base contenant non pas tous les genres ou types de données ou tous les types de locuteurs, ni tous les points d'enquête pour une langue donnée, mais comme un ensemble structuré et mutualisable de textes rattachés à leur enregistrement sonore, accompagné par des informations linguistiques pertinentes pour la langue en question, ainsi que par des métadonnées riches » (cf. <http://crdo.risc.cnrs.fr/ecoles/elco/bilan-ELCO.pdf>).

Pour réunir le plus de données possible, nous avons eu recours à l'entretien semi-directif (face à face). Un guide d'entretien a été réalisé afin de faire parler les témoins, en ciblant les contextes propices à l'émergence des formes verbales au présent. Les questions que nous avons choisies portent d'une part, sur les langues utilisées par nos informateurs à Orléans, sur leur importance et sur ce qu'elles représentent pour eux ; d'autre part, sur la culture et les traditions transplantées du pays d'origine au pays d'accueil. L'entretien constitue une incitation à des réflexions sur la situation présente, passée et sur les perspectives de nos informateurs. Ce qui va nous permettre de cerner le fonctionnement de la forme du présent sans pour autant nous restreindre à une époque (présente, passée ou future) particulière.

Et pour varier les situations enregistrées, figurent dans le corpus d'autres genres de parole comme les communications téléphoniques, les recettes de cuisine, le récit et le récit de vie, la conversation et le commentaire de photos. Ces situations sont naturellement produites dans certains cas, mais sollicitées dans d'autres (recettes de cuisine, commentaire de photos).

En ce qui concerne les témoins, ils ont été différenciés en termes d'âge, de sexe, de niveau scolaire et de catégorie socio-professionnelle. Nous avons eu recours à 11 témoins locuteurs du berbère tamazight et 20 de l'arabe marocain.

Corpus du berbère tamazight (2008-2009)	
Nombre d'heures d'enregistrement	7- 8 h
<u>Situations de parole</u>	<ul style="list-style-type: none"> – Entretien face à face – Recettes de cuisine – Communications téléphoniques – Récit/ récit de vie – Conversation – Commentaire photos
<u>Catégories des locuteurs :</u>	
– Âge :	<ul style="list-style-type: none"> – 20-30 ans – 30-50 – 50 et plus
– Sexe :	<ul style="list-style-type: none"> – Hommes : 40 % – Femmes : 60 %

	Niveau d'études
– Niveau scolaire :	– non scolarisé
	– primaire/collège
	– bac
	– supérieur
	(professions)
– Profession/CSP :	– Secrétaire
	– Préparateur en pharmacie
	– Technicien
	– Commerçant/ Aide commerçant
	– Ouvrier
	– Couturière
	– Sans profession

Concernant les langues parlées par nos témoins, nous sommes partie de leurs déclarations en sorte qu'il est difficile, à partir de cette enquête, de mesurer le degré de compétence réel. Deux témoins peuvent être considérés comme monolingues, sept comme bilingues et dix-huit trilingues ou plus.

3. L'enregistrement

La qualité de la constitution du corpus est un facteur de la relation de confiance établie entre nous, en tant que personne (de la même communauté linguistique) et en tant que chercheur (qui peut, du point de vue des locuteurs, contribuer à la promotion de leur langue), et les enquêtés.

Avant de commencer l'enregistrement, nous essayons de nous familiariser¹ avec le témoin pour qu'il se sente parfaitement à l'aise et que la conversation enregistrée soit aussi naturelle que possible. En effet, nous avons pris systématiquement le temps de discuter avec le témoin pour avoir sa confiance. Tout cela a l'avantage d'atténuer « le paradoxe

1. Les problèmes de proximité/distance du chercheur aux situations enquêtées « ont été traités en termes de paradoxe d'observateur — selon lequel le phénomène enquêté se dissout dès qu'il est observé (tel le vernaculaire pour LABOV 1972) — aussi bien qu'en termes de violence symbolique entre l'enquêté et l'enquêteur (BOURDIEU, 1993). Ils ont aussi été traités en termes de réflexivité — par des chercheurs intégrant leur présence et celle du dispositif d'enquête dans l'analyse de l'objet enquêté (en anthropologie notamment, CLIFFORD & MARCUS, 1986, MONDADA, 1998) » (BAUDE 2006 : 52).

de l'observateur » et d'accéder au vernaculaire du locuteur. Comme le précise Bourdieu : « La proximité sociale et la familiarité assurent en effet deux des conditions principales d'une communication "non violente" » (Bourdieu 1993 : 1395).

Notre relation privilégiée avec les témoins vient du fait que nous partageons avec eux un certain nombre de traits : même origine (pays d'origine), même langue(s), même situation en France (appartenance à la population immigrée), etc. Nous sommes membre à part entière de la culture vernaculaire. Nous la connaissons de l'intérieur et nous la comprenons profondément. Cela permet, selon Labov (1978 : 10) de « réussir une percée plus profonde ».

Les enregistrements ont été recueillis à Orléans entre 2008 et 2009. Huit enregistrements ont été effectués à A.C.M. Formation¹ et le reste (38 enregistrements) chez les témoins, soit 46 enregistrements (audio) au total, d'une durée variant de 5 minutes (recettes de cuisine/communications téléphoniques) à 45 minutes².

4. Des données situées

Depuis le développement de la linguistique du corpus, la documentation de ce dernier est devenue fondamentale. Celle-ci consiste à fournir des renseignements sur la situation de collecte et le profil des témoins. Dans ce travail, il a été question de l'observation et la description des usages authentiques³ de la langue. D'où l'introduction du locuteur réel, d'une part, en tant que « voix » indissociable de la transcription

1. A.C.M. Formation (Accès aux clés de la modernité) est une association 1901. Son objet est d'intervenir, notamment auprès des publics les plus socialement défavorisés, au travers d'actions de formation et de pratiques culturelles.

2. Qu'il s'agisse du français extrait du corpus ESLO1, de l'arabe marocain ou du berbère tamazight, chaque corpus représente un peu plus de sept heures d'enregistrement, soit environ vingt-deux heures au total. Une différence d'importance est à signaler : le corpus du français, tiré de l'ESLO, date de 1968-1970, alors que ceux de l'AM et du BT ont été constitués en 2008-2009 à Orléans. Le corpus du français analysé présente un seul genre de parole, des entretiens en face à face.

3. Concernant la question de *données authentiques*, quelques restrictions sont à signaler : *i)* le corpus est lui-même un *construit* car il résulte d'une sélection : « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue. » (HABERT B. 2000) ; *ii)* le corpus n'est pas représentatif des différents usages de la langue.

(transcription alignée au son) et d'autre part, par la reconstitution de son profil sociologique. Ce qui peut « rendre à la linguistique la méthodologie d'une véritable science des données attestées et *situées*¹ » (Abouda et Baude 2006 : 9).

Nous avons constitué un corpus de « données situées » : il contient, en plus des données primaires (les enregistrements de la parole), une riche documentation sur ces données et sur leur contexte de production². Ce qui n'est pas sans importance pour l'analyse du phénomène étudié : la plupart de nos résultats émergent grâce à ces données situées. Nous avons tenu également à expliciter notre démarche, à documenter les conditions de constitution du corpus ainsi que nos choix théoriques et techniques. Toutes ces informations sont d'une grande importance pour rendre le corpus disponible³ et ré-exploitable par d'autres chercheurs.

5. La transcription du corpus

La collecte des données de l'arabe marocain et du berbère tamazight s'est effectuée sous forme d'enregistrements. Ces données sonores brutes ne peuvent pas être analysées sans un travail préalable de transcription et de segmentation. Transcrire des données sonores consiste à fournir une représentation symbolique du signal⁴ en respectant les conventions de la langue en question. Mais la transcription en elle-même n'est pas une tâche anodine. Le processus de transcription soulève un certain nombre d'enjeux théoriques, méthodologiques et pratiques.

Transcrire des langues à tradition essentiellement orale soulève des problèmes spécifiques, surtout si l'on entend produire un corpus partageable. En effet, en plus des difficultés que pose toute transcription de l'oral, d'autres interrogations, d'une autre nature, surgissent lors

1. C'est nous qui soulignons.

2. « La linguistique du corpus prend sens dans la réintroduction de la question de l'usage, elle amène à *situer*, c'est-à-dire à replacer les phénomènes observés et décrits dans un contexte. » (JACQUES 2005 : 29).

3. Pour chaque locuteur, il y a une fiche d'information récapitulant l'âge, le sexe, le niveau scolaire... complétée par des indications sur l'enregistrement (n°, type (situation de parole), participant(s), lieu, date et durée de l'enregistrement...), situation d'enregistrement...

4. E. DELAIS-ROUSSARIE (2003 : 98).

de la phase préparatoire à la transcription de ces deux langues « non latines » et pour lesquelles il existe une concurrence entre plusieurs alphabets et donc plusieurs traditions orthographiques, des problèmes liés à la fois au système graphique, au mode, aux conventions et aux outils de transcription à adopter...

Si le français dispose d'une tradition écrite ancienne et bien établie (standard), ce n'est le cas ni du berbère ni de l'arabe marocain. Le berbère est une langue à tradition orale, sans tradition écrite standardisée. L'arabe marocain est une langue à tradition orale, pour laquelle il existe certes un standard de référence, l'arabe classique, avec laquelle il a un lien de parenté, mais dont le système linguistique est très différent ¹.

5.1. Système(s) de notation

Plusieurs systèmes graphiques sont en usage pour écrire la langue berbère : graphie tifinaghe ², graphie arabe ³, graphie latine et graphie hébraïque ⁴. Il y a eu deux tentatives de normalisation du système graphique du berbère. La première est pilotée depuis 2001 par une institution publique au Maroc, l'Institut royal de la culture amazighe (IRCAM) qui a adopté un système graphique à base tifinaghe. La deuxième a commencé au début du xx^e siècle, en dehors de toute intervention institutionnelle ou étatique, par des universitaires et militants

1. En se basant sur le fait que les systèmes linguistiques de l'arabe marocain et de l'arabe classique présentent des différences sur les plans phonique, morphosyntaxique, voire lexical, et en tenant compte de l'absence de compréhension immédiate, les linguistes leur accordent le statut de deux langues différentes.

2. La majorité des berbérissants postule que les tifinaghe sont d'origine phénicienne. Ainsi pour CHAKER (1984), « l'alphabet est très certainement d'origine phénicienne — punique, comme la quasi totalité des systèmes existants ». Pour HACHID (2001), « Cette écriture est soit un emprunt à l'alphabet phénicien, soit une invention locale, ou encore un emprunt à un prototype fort ancien que l'on ne connaît pas encore. » (Cité dans BOUHJAR, 2003).

3. L'alphabet arabe a été utilisé en particulier chez les Chleuhs (sud du Maroc). Cette pratique est encore très vivante chez les écrivains contemporains chleuhs dont la majorité utilise l'alphabet arabe pour écrire leurs oeuvres (poésie, nouvelle, manuels grammaticaux...) : MOUSTAOU, AKHIAT... Et aussi chez CHAFIK.

4. Les communautés juives ont employé l'alphabet hébreu pour transcrire le berbère comme en témoigne une version de la Haggadah de Pessah (Cf. GALAND-PERNET & H. ZAFRANI, *une version berbère de la Haggâdâh de Pesah : texte de Tinrhir du Todrha* (Maroc), Paris, 2. vol., 1970 (Comptes rendus du GLECS, supplément I).

associatifs, avec des publications écrites en caractère latin, et se poursuit jusqu'à présent à travers des ateliers et journées d'études organisés par les chercheurs ¹.

Quant à l'arabe dialectal, il a été noté à l'aide de plusieurs systèmes graphiques. Avant la colonisation française, les musulmans ont utilisé la graphie arabe pour noter les parlers arabes maghrébins, et les Juifs du Maghreb, la graphie hébraïque ². Vers la fin du XIX^e et au début du XX^e siècle, la graphie latine a été introduite par des linguistes comme William Marçais, Marcel Cohen, Georges S. Colin, Jean Cantineau, etc. qui ont publié des textes en appui de leurs descriptions de l'arabe dialectal.

Le choix de la graphie latine pour transcrire les corpus de l'arabe marocain et du berbère tamazight suppose la réalisation des phénomènes spécifiques à ces deux langues (comme le phénomène de l'emphase, la labiovélarisation, la spirantisation, la tension, la gémination, etc.). Quel mode de transcription choisir ? Quelles conventions de transcription adopter et pourquoi ?

5.2. Mode de transcription du corpus

Choisir un mode de transcription implique des enjeux qui dépendent des objectifs de la recherche, mais également de la représentation ³ de la langue parlée, ou encore de la représentation que l'on veut en donner (Bilger, 2008 : 248). Et comme le souligne (Gadet 2008 : 37), « la transcription ne peut être regardée comme une opération banale, car on transcrit pour donner à voir quelque chose ».

Pour la transcription de l'arabe dialectal et du berbère, il existe plusieurs modes de notation : phonétique, phonologique, morpho-phonologique et orthographique (notation usuelle). C'est ce dernier mode de transcription (la notation usuelle) qui a été adopté pour la transcription du corpus. La *notation usuelle* n'est ni une transcription phonétique, ni une transcription phonologique stricte, mais une

1. Ces deux tentatives de normalisation sont présentées en détails dans MOUKRIM (2010 : 85-116).

2. Dès leur insertion dans la civilisation musulmane, les Juifs du Mashreq et du Maghreb adoptèrent la langue arabe dans la communication courante comme dans leurs productions littéraires. La transcription de la langue se fait alors en caractères hébreux d'où son appellation de « judéo-arabe » (Joseph TEDGHI (2002 : 415).

3. Représentation que le chercheur et le transcrip-teur se font de la langue.

notation d'inspiration phonologique qui prend en compte la structure morpho-syntaxique des énoncés.

Le choix de la notation usuelle s'explique par les raisons suivantes :

- le thème de recherche ne requiert pas le recours à une transcription phonétique (difficulté de décodage), ni à une transcription phonologique stricte.
- la volonté de fournir un corpus facilement lisible (décodable) par tout le monde et qui prend en compte également la structure morpho-syntaxique de l'énoncé.
- la possibilité d'accompagner la notation usuelle par une transcription phonétique et/ou phonologique pour une étude spécifique (phonétique, intonation...) : possibilité d'ajouter d'autres niveaux de transcription en fonction des objectifs de recherche et des outils de transcription utilisés.

Cependant, le choix d'une transcription « orthographique » (usuelle) pose problème en l'absence d'un standard stabilisé. Une transcription orthographique partageable suppose l'acceptation d'un standard ou à défaut la reprise des pratiques orthographiques les plus courantes au sein de la communauté scientifique.

5.3. Les conventions de transcription

Les corpus de l'arabe marocain et du berbère tamazight ont été transcrits sous TRANSCRIBER ¹, un logiciel d'aide à la transcription manuelle de fichiers audio qui permet de transcrire de nombreuses langues y compris non européennes.

Le choix des conventions de transcription n'est jamais neutre, ce qui explique une variété dans les pratiques (Bilger, 2008 : 34). Il dépend de différents facteurs : les objectifs de la recherche, la nature des données primaires (audio ou vidéo), la taille du corpus, la représentation que le chercheur se fait de la langue parlée ou la représentation qu'il veut en donner, la langue du corpus, etc.

Les conventions ne sont pas les mêmes pour une langue à tradition écrite ancienne et bien établie, comme le français, et pour le berbère ou pour l'arabe marocain (langues sans tradition orthographique solide). En effet, deux types de conventions de transcription ont été distingués : des conventions « spécifiques » à chaque langue ; et des conventions

1. Téléchargeable sur : <http://trans.sourceforge.net>.

« communes » à tout corpus oral quelle que soit la langue. Pour la transcription des corpus de l'AM et du BT¹, on s'est conformé aux propositions des ateliers de l'INALCO (1996-1998) en tenant compte également des propositions de L'IRCAM² concernant le traitement des problèmes spécifiques que posent les dialectes berbères marocains. Quant aux phénomènes liés à l'oralité, ont été retenues les conventions proposées par le L.L.L. pour le corpus du français de l'ESLO³.

En somme, travaillant dans une perspective de partage et de mutualisation des données et en l'absence d'une norme stabilisée pour chacune des deux langues, nous avons repris les pratiques majoritaires au sein de la communauté scientifique travaillant sur le berbère et l'arabe dialectal, ainsi que sur la langue parlée d'une manière générale.

6. Les aspects juridiques du corpus

La diffusion des corpus oraux suppose une attention accrue au cadre juridique dans lequel s'inscrivent la réalisation et l'exploitation des corpus. L'aspect juridique des corpus linguistiques est une question relativement nouvelle, à laquelle Baude (2006) a apporté des réponses consensuelles et intéressantes.

Les données recueillies dans ce travail se présentent sous forme d'enregistrements de locuteurs s'exprimant en berbère tamazight ou en arabe marocain (ou les deux) dans des situations de communication ordinaires. Se pose alors la question suivante : quel est le statut juridique de ce corpus ?

Les aspects juridiques des corpus oraux concernent trois grands domaines⁴ : i) le droit d'auteur et la propriété intellectuelle ; ii) les don-

1. Le même système de notation a été adopté pour la transcription des deux corpus car il n'y pas une grande différence entre le système de notation proposé pour l'écriture du berbère et celui utilisé pour la notation de l'arabe dialectal en caractères latins.

2. L'Institut royal de la culture amazighe est une institution publique marocaine dévolue à la promotion du berbère et à son insertion dans l'enseignement, les médias et la vie publique en général.

3. Les conventions de transcription que nous avons adoptées ne sont pas individuelles. Elles sont, comme le signale MONDADA (2008 : 87) « le produit historique de discussions et souvent d'accords établis au sein de groupes théoriquement homogènes ».

4. BAUDE (2008 : 24).

nées personnelles et le respect de la vie privée et iii) les responsabilités des « exploitants » et diffuseurs.

La définition du statut juridique d'un corpus « nécessite de décrire avec précision son contenu puis les conditions d'élaboration et d'exploitation de celui-ci » (*Ibid.* : 25). Cela permet de « différencier les composantes susceptibles d'être protégées par le droit d'auteur ainsi que les composantes contenant éventuellement des données personnelles » (*Ibid.* : 26).

Pour qu'un corpus (ou une composante de ce corpus) soit considéré comme protégé par le droit d'auteur il faut qu'il remplisse trois conditions : qu'il corresponde à l'exigence d'une activité créatrice, qu'il ait une forme définie et que cette forme soit originale. Ayant pour objectif d'étudier la langue (BT ou AM) des témoins, on a traité de questions qui relèvent de la vie courante (ordinaire) et qui ne présentent aucune originalité créative. Bien que notre corpus ne soit pas directement concerné par les protections du droit d'auteur, toutes les précautions ont été prises pour rendre les témoins non identifiables.

Quant aux données personnelles et le respect de la vie privée¹, leur présence dans un corpus implique une mise en conformité avec la loi informatique et liberté (licéité et loyauté, information préalable, obtention du consentement) ou une anonymisation irréversible de celle-ci (Baude 2008 : 27). L'anonymisation consiste d'abord à repérer les données permettant l'identification directe ou indirecte et celles qui pourraient porter préjudice (propos diffamatoires). Ces données primaires² (enregistrements) sont ensuite traitées au moyen d'opérations techniques (bippage, effacement, déformation, etc.).

Le recueil du consentement des personnes enregistrées est la meilleure solution éthique et juridique (Baude 2006, 2008). Toutefois, il faut établir un consentement « éclairé » qui démontre que le signataire est informé des finalités de la recherche et des conséquences qu'entraîne sa

1. Parmi les données personnelles qui permettraient, dans un corpus, d'identifier directement un témoin on compte les formes nominatives, données personnelles, profession, statut, titre, activités sociales, parenté, réseaux, référence à des lieux, référence à des caractéristiques de la personne, caractéristiques physiques, etc. Il en va de même si le témoin peut être identifié par les possibilités de recoupement d'informations (*Ibid.* : 28).

2. Les corpus oraux sont en général composés d'enregistrements audio ou vidéo (données primaires) et d'annotations de ces derniers (données secondaire) (BAUDE 2006 : 45-46, pour plus de détails).

participation au projet : « [...] sans informations préalables précises la demande d'autorisation n'a pas d'objet ni de sens » (Baude 2006 : 60). Avant de commencer l'enregistrement, le projet et les finalités de l'enquête sont clairement expliqués aux témoins. Une fois l'enregistrement terminé, ils sont invités à signer l'autorisation qui permettra de disposer des enregistrements et de leur transcription à des fins de recherche. La quasi-totalité des locuteurs ont préféré donner leur consentement oralement. Cela s'expliquerait par le fait que cette procédure de signature les inquiète, du fait qu'elle peut être liée à d'autres pratiques avec lesquelles elle pourrait être confondue comme la signature de chèques (Baude 2006 : 64). La seule personne qui a accepté de signer l'autorisation écrite a un niveau universitaire de troisième cycle. Par ailleurs, et malgré nos explications, certains de nos témoins trouvent contradictoire le fait d'écrire leur nom pour signer l'autorisation et « l'anonymisation ».

En écoutant les enregistrements, on a constaté que les témoins étaient tous très discrets sur tout ce qui pourrait les rendre identifiables, comme s'ils prenaient toutes les précautions pour ne pas être reconnaissables, en sorte que les données sont anonymisées à la « source ». Il n'est donc pas nécessaire d'anonymiser les données ni d'obtenir un consentement : « L'obligation d'obtenir un consentement préalable peut être levé si retrouver les personnes concernées s'avère difficile » (Baude 2006 : 109).

Les aspects juridiques des corpus oraux concernent également le domaine de la responsabilité de ceux qui auront à intervenir dans la « *vie du corpus* » : responsabilité des créateurs, responsabilité des exploitants, responsabilité des diffuseurs...

Collecteur, transcripteur et traducteur (les séquences contenant le phénomène linguistique étudié sont accompagnées d'une traduction), nous sommes aussi le premier exploitateur du corpus et on se doit de veiller à la qualité des données, au respect des finalités indiquées, au respect du principe de licéité et aux conditions de conservation (*Ibid.* : 121).

Conclusion

Bien que le corpus de l'arabe marocain et du berbère tamazight ne soit pas sociologiquement représentatif, la diversité qui a présidé à sa

constitution permet de le concevoir comme un réservoir dans lequel on peut rechercher des attestations concernant les distributions de faits de langue. Les échantillons retenus ne sauront être véritablement représentatifs de toute la communauté marocaine présente sur Orléans. Néanmoins, se voulant diversifiés, ils offrent également une catégorie particulière de locuteurs, rarement inclus dans les échantillons, des locuteurs en situation irrégulière¹ (des « sans papiers »).

Une fois le corpus constitué et transcrit, nous sommes passée à l'observation des données. Nos résultats proviennent, d'une part, de la comparaison de trois langues différentes (la diversité d'emplois de la forme du « présent » est appréhendée dans la perspective du fonctionnement du langage) et d'autre part, du fait que nous avons travaillé sur des données orales authentiques.

La démarche comparative a permis de dégager les propriétés spécifiques à chacune des langues étudiées. Et le fait de travailler sur des données orales réelles et situées a permis de mettre en lumière des fonctionnements linguistiques qui échappent à l'intuition. Bien que ce travail ne soit qu'une ébauche, la transformation des méthodes d'analyse à partir de nouvelles données va faire progresser la description grammaticale d'une manière générale.

Références bibliographiques

ABOUDA L. & BAUDE O.,

2005, « Du français fondamental aux ESLO », colloque international Français fondamental, *corpus oraux, contenus d'enseignement*, 50 ans de travaux et d'enjeux, SIHFLES — Laboratoire ICAR, Lyon, 8, 9 et 10 décembre 2005.

ABOUDA L. & BAUDE O.,

2006, « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO », in RASTIER F. & BALLABRIGA M. (dir.), *Corpus en Lettres et Sciences sociales — Des documents numériques à l'interprétation*, actes du XXVII^e colloque d'Albi, *Langages et signification*, publiés par DUTEIL-MOUGEL C. & FOULQUIÉ B.

1. Un de nos témoins est en situation irrégulière en France, c'est un « sans papiers » (il a été sollicité pour un récit de vie); deux l'ont été pendant cinq et six ans, ils viennent d'être régularisés.

- BAUDE O. (coord.),
2006, *Corpus oraux, Guide des bonnes pratiques*, C.N.R.S. éditions et PUO.
- BAUDE O.,
2008, « Le droit de la parole », in BILGER M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, PUP, 24-33.
- BENJELLOUN S., 2002, « Une double graphie, latine et arabe, pour enseigner l'arabe marocain », in CAUBET D., CHAKER S. & SIBILLE J. (éd.), *Codification des langues de France*, Paris, L'Harmattan, 331-340.
- BERGOUIGNIOUX G. (dir.),
1992, « Enquêtes, Corpus et Témoins », *Langue française* 93.
- BERGOUIGNIOUX G. et al.,
1992, « L'Étude socio-linguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus », *Langue française* 93, 74-93.
- BILGER M. & CAPPEAU P.,
2004, « L'oral ou la multiplication des styles », *Langage et société* 109, 13-30.
- BILGER M.,
2008, « Les enjeux des choix orthographiques » in BILGER M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, PUP, 248-257.
- BLANCHE-BENVENISTE C. & JEANJEAN C.,
1987, *Le français parlé : transcription et édition*, Paris, Didier-Érudition.
- BOUHJAR A., 2003, « Le système graphique tfinagh-Ircam », in *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'aménagement linguistique à Rabat, 8-9 décembre 2003, publication de l'Institut royal de la culture amazighe, série : *Colloques et séminaires* n° 3.
- BOUKOUS A., 2003, « La standardisation de l'amazighe : quelques prémisses », *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'aménagement linguistique à Rabat, 8-9 décembre 2003, Publication de l'Institut royal de la culture amazighe, série : *Colloques et séminaires* n° 3.
- BOURDIEU P. (dir.),
2003, *La misère du monde*, Paris, Seuil, coll. « Point ».
- CAPPEAU P., 2008, « Perception et reconstruction » in BILGER M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, PUP, 235-247.
- CAUBET D., 1999, « Arabe maghrébin : passage à l'écrit et institutions », in *Faits de Langues* 7, 13, 235-244.

- CAUBET D., 2002, « Arabe maghrébin, langue de France : entre deux graphies », in CAUBET D., CHAKER S. & SIBILLE J. (éd.), *Codification des langues de France*, Paris, L'Harmattan, 307-330.
- CERQUIGLINI B., 1999, *Les langues de la France*, rapport aux ministres de l'Éducation nationale et de la Culture et de la Communication. (en ligne : www.dglf.culture.gouv.fr/lang-reg/rapport_cerquiglini/langues-france.html).
- CHAKER S., 1996, « Propositions pour la notation usuelle à base latine du berbère » (Atelier du 24-25 juin 1996, INALCO-CRB. *Synthèse des travaux*).
- CHAKER S., ACHAB R., NAÏT-ZERRAD K., 1998, « Aménagement linguistique de la langue berbère : atelier organisé du 5 au 9 octobre 1998 », *Synthèse* préparée par CHAKER S., ACHAB R., NAÏT-ZERRAD K.
- CHAKER S., 2002, « Variation dialectale et codification graphique en berbère. Une notation usuelle pan-berbère est-elle possible ? », in CAUBET D., CHAKER S. & SIBILLE J. (éd.), *Codification des langues de France*, Paris, L'Harmattan, 341-354.
- GADET F., 2000, « Derrière les problèmes méthodologiques du recueil des données », in BILGER M. (dir.), *Linguistique sur corpus*, Perpignan, Presses universitaires de Perpignan.
- GADET F., 2008, « L'oreille et l'œil à l'écoute du social » in BILGER M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, Presses universitaires de Perpignan, 35-47.
- HABERT B., NAZARENKO A. & SALEM A., 1997, *Les linguistiques de corpus*, Paris, A. Colin.
- HABERT B., 2000, « Des corpus représentatifs : de quoi, pour quoi, comment ? », in BILGER M. (dir.), *Linguistique sur corpus*, Perpignan, Presses universitaires de Perpignan.
- HACHID M., 2000, *Les Berbères, aux origines de l'histoire*, Aix-en-Provence, Alger, Édisud-Inna yas.
- LABOV W., 1976, *Sociolinguistique*, Paris, Éd. de Minuit.
- LABOV W., 1978, *Le parler ordinaire*, Paris, Éd. de Minuit, coll. « Le sens commun ».
- MAURER B., 1999, « Quelles méthodes d'enquête sont effectivement employées aujourd'hui en sociolinguistique », in CALVET L.-J. & DUMONT P. (dir.), *L'enquête sociolinguistique*, Paris, L'Harmattan.

- MONDADA L., 2008, « La transcription dans la perspective de la linguistique interactionnelle » in BILGER M. (éd.), *Données orales : les enjeux de la transcription*, Perpignan, Presses universitaires de Perpignan, 78-109.
- MOUKRIM S., 2010, *Morphosyntaxe et sémantique du « présent » : une étude contrastive à partir de corpus oraux, arabe marocain, berbère tamazight et français (ESLO/LCO)*, thèse de doctorat, université d'Orléans.
- TEDGHI J., 2002, « Usage de la graphie hébraïque dans la transcription des parlers judéo-arabes modernes au Maghreb », in CAUBET D., CHAKER S. & SIBILLE J. (éd.), *Codification des langues de France*, Paris, L'Harmattan, 415-441.