Constitution d'un corpus d'arabe tunisien parlé à Orléans

Ben Ahmed Youssra
Laboratoire Ligérien de Linguistique (UMR 7270), 10 Rue de Tours – BP 46527, 45065, France
ben.ahmed.yossra@gmail.com

RESUME

La constitution de corpus des parlers arabes se heurte à l'absence de ressources et le manque d'outils pour le traitement de ces derniers. Comme conséquence directe de cette situation, nous avons rencontré de maints problèmes lors de la construction d'un corpus de l'une des variétés de ces parlers, i.e. l'arabe tunisien. Nous proposons dans cet article d'exposer les principaux choix méthodologiques et techniques pour lesquels nous avons opté afin de répondre aux contraintes auxquelles nous avons été confrontée.

ABSTRACT

Composition and exploitation of a Tunisian Arabic corpus - Orleans

The composition of Arabic corpus dialect faces the absence of resources and the lack of tools in order to treat them. As a direct consequence of this condition, we encountered many problems during the construction of a corpus of one of the varieties of these dialects, i.e. Tunisian Arabic.

We propose in this article to present the main methodological and technical choices for which we chose to respond to the restrictions we confronted.

MOTS-CLES : arabe tunisien, corpus oral, transcription, annotation.

KEYWORDS : tunisian arabic, corpus oral, transcription, annotation.

1 Introduction

Cette étude s'aventure dans un terrain à peu près vierge, celui de l'analyse des expressions du futur en arabe tunisien.

C'est la multitude d'emplois de la forme du «futur» en français (futur historique, futur de vérité générale, d'injonction, futur de bilan, etc.) qui a attiré notre attention et nous a amenée à explorer la question dans une autre langue, notamment en arabe tunisien. Il nous a semblé utile, dans ce cadre, de s'interroger sur l'existence du futur et son expression dans les systèmes verbaux de ces deux langues, qui se différencient fondamentalement sur le plan aspectuel.

Dans la plupart des études sur le futur, les linguistes ont eu recours soit à la fabrication d'exemples soit à l'emprunt d'exemples écrits ; des choix susceptibles d'empêcher de voir son fonctionnement réel.

Cette considération nous a poussée à proposer une étude contrastive de l'emploi du futur en nous basant sur des données orales authentiques. Notre choix s'est porté, pour la partie française, sur des données puisées dans les Enquêtes Socio-Linguistiques à Orléans (désormais ESLO). Quant à celle de l'arabe, nous avons constitué un corpus auprès des locuteurs tunisiens résidant à Orléans.

Dans ce papier, après une première partie où il sera notamment question des principaux choix opérés lors de la constitution du corpus de l'arabe tunisien (désormais AT), nous nous attarderons dans la deuxième partie sur les problèmes rencontrés lors de la phase de transcription. En troisième partie, nous aborderons les difficultés affrontées lors du processus d'annotation.

2 Méthodologie

Pour la constitution de notre corpus d'arabe tunisien[1], nous avons suivi en grande partie la démarche adoptée dans la construction du corpus ESLO. C'est sur ce dernier que nous nous sommes basée dans la sélection des données en français.

D'une taille conséquente (à ce jour, environ 7 millions de mots) et d'une diversité de genres (conférences universitaires, entretiens, repas en famille ou entre amis…), ce corpus offre un avantage essentiel dans le domaine de la temporalité où le pertinence des analyses est tributaire d'une bonne prise en compte de la situation de communication : il contient des données situées, enrichies par des métadonnées renseignant sur la situation de communication et précisant pour chaque locuteur son profil en termes d'âge, de sexe et de catégorie socio-professionnelle.

S'agissant d'une recherche contrastive, les choix présidant à la constitution du corpus de l'AT ont été en grande partie dictés par la recherche de la plus grande comparabilité possible avec le corpus ESLO.

A commencer par le lieu de l'enquête : même si Orléans constitue sans doute un « non-choix » (Abouda & Baude, 2009 : 133), il nous a paru prudent de respecter cette unité de lieu, d'autant que notre corpus de l'AT a pu ainsi intégrer, pour être partagé, la base de données constituée par le programme « Langues en Contact à Orléans » (LCO) en bénéficiant des moyens et de l'expérience accumulée dans ce cadre. Il va de soi que ce choix, qui a écarté l'autre possibilité un temps envisagée de mener une enquête comparable dans une ville tunisienne, présentait quelques inconvénients : il était particulièrement long et difficile de constituer un corpus suffisamment grand et équilibré.

Le corpus, que nous avons constitué entre 2013 et 2014, représente un volume de 17h, fractionné en 37 enregistrements, ce qui nous semble suffisant pour les investigations envisagées. Il parvient à

[1] Il s'agit d'une langue vernaculaire non standardisée, parlée par douze millions de personnes vivant principalement en Tunisie et par de nombreuses familles résidant à l'Europe, notamment en France. Il s'emploie principalement dans tous les usages informels, i.e. à la maison, dans la rue, au travail ou à l'école, et même dans les médias audiovisuels.

capter une certaine diversité de locuteurs (en fonction des variables d'âge, de sexe, de CSP, de régions, etc.), ce qui permet d' « améliorer [sa] représentativité »[2].

Corpus de l'arabe tunisien (2013-2014)	
Nombre d'heures d'enregistrements	17h
Situation de parole	Entretien face à face
Âge	- 20- 30 ans : 54% - 30- 50 : 35% - 50 et plus : 11%
Sexe	- Hommes : 65% - Femmes : 35%
Niveau scolaire	- primaire collège 15% - secondaire 15% - bac 24% - supérieur 46%
Profession	- Chef de cuisine - Animateur - Commerçant - Assistant d'éducation - Etudiant - Gérant - Femme de ménage…

TABLEAU 1 : Caractéristiques quantitatives du corpus

Respectant les procédures suivies par ESLO, nécessaires pour rendre le corpus disponible, nous avons procédé à une documentation précise de nos données[3].

En ce qui concerne le mode de recueil des données, nous avons privilégié l'entretien en face-à-face, « situation certes très formelle, mais qui avait l'avantage d'être (…) contrôlable » (Abouda & Baude, 2009 : 134).

Aussi, dans le même souci de comparabilité, nous avons réalisé un questionnaire basé sur les six thèmes retenus par ESLO (logement/Orléans, travail, loisirs, questions évaluatives sur Orléans, langue, recette), afin de faire parler les locuteurs, en ciblant les contextes propices à l'émergence des formes verbales au futur.

[2] Cf. Habert 2000.

[3] Pour chaque locuteur, nous avons réalisé 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…)

3 Transcription

Après la collecte des données, s'est posée la question de leur transcription. Cette étape a soulevé plusieurs interrogations depuis le système graphique jusqu'aux outils de transcription, en passant par le mode de transcription et les conventions adoptées.

3.1 Système de notation

Les travaux sur l'arabe tunisien, peu nombreux, hésitent entre les deux systèmes graphiques, i.e. latin et arabe. Le choix de l'un ou l'autre système est dicté par de nombreux paramètres, allant de la tradition du champ, jusqu'aux préférences idéologiques, en passant par la facilités techniques.

C'est précisément pour cette dernière raison, que nous avons opté pour une transcription avec une graphie latine, qui aura également l'avantage de fournir un corpus partageable et facilement lisible par les non-natifs.

3.2 Mode de transcription

Ce sont non seulement les objectifs de recherche qui définissent le choix d'un mode de transcription, mais aussi les spécificités de la langue parlée. En bref, ainsi le note Gadet (2008 : 37) « la transcription ne peut être regardée comme une opération banale, car on transcrit pour donner à voir quelque chose.»

La transcription de notre corpus a impliqué un choix entre plusieurs types de notation (phonétique, phonologique, morphologique et usuelle). Nous avons opté finalement pour une notation orthographique (usuelle) *d'inspiration phonologique* qui tient compte de l'aspect morphosyntaxique des énoncés. Ce choix a été motivé par les raisons suivantes :

- l'objet d'étude ne nécessite ni une notation phonétique, ni une notation phonologique stricte ;
- la simplicité de ce mode de notation permet un décodage rapide et facile par le lecteur en écartant les ambiguïtés et les hésitations, principalement au niveau syntaxique ;
- l'ajout possible des deux autres modes de transcription (phonétique ou/et phonologique) selon les besoins des chercheurs.

Nonobstant, l'absence d'un standard stabilisé a exigé la reprise des pratiques orthographiques les plus usitées au sein de la communauté scientifique.

3.3 Les conventions de transcription

En ce qui concerne l'outil de transcription, nous avons choisi TRANSCRIBER[4], un logiciel d'aide à la transcription manuelle de fichiers audio qui permet de transcrire de nombreuses langues y compris non européennes.

[4] Téléchargeable sur : http://www.ldc.upenn.edu/mirror.Transcriber/

Lors de la transcription, un problème d'encodage s'est manifesté par le fait que quelques caractères spéciaux ne s'affichent pas correctement. Il nous a paru dés lors indispensable d'opter pour l'encodage UTF-8.

De différents facteurs sont entrés en jeu pour la détermination des conventions de transcription, allant des finalités de la recherche, jusqu'à la taille du corpus, en passant par le type des données primaires (audio ou vidéo).

Les conventions varient selon la nature de la langue, i.e. écrite (comme le cas du français) ou orale (comme pour l'arabe tunisien). Elles se divisent en deux types : des conventions « *spécifiques* » à chaque langue ; et des conventions « *communes* » à tout corpus oral quelle que soit la langue.

Privé d'une tradition orthographique solide, nous avons choisi de transcrire l'arabe tunisien en nous basant sur les propositions de l'INALCO (1996-1998). En ce qui concerne les phénomènes associés à l'oralité, ont été maintenues les conventions avancées par le LLL pour le corpus du français de l'ESLO.

4 Annotation

L'annotation, étape incontournable en ce qu'elle permet de croiser approches qualitative et quantitative, consiste dans l'apport d'informations de nature différente. On parle dans ce sens d'une «valeur ajoutée» (Leech 1997) aux données brutes.

S'il existe pour l'arabe standard des étiqueteurs morphosyntaxiques, relativement importants (Arabic Part-of-speech Tagger[5], Sakher[6], Sebawai[7], Aramoph[8], etc.), la situation concernant l'arabe tunisien est nettement différente ; car il n'y a pas à notre connaissance de systèmes complets et disponibles pour l'étiquetage de celui-ci.

Etant donné que « l'arabe dialectal se distingue de l'arabe classique par une syntaxe simplifiée, un lexique plus riche en vocables étrangers et une phonologie altérée » (Boukadida, 2008 : 37), nous n'avons pas pu se servir de ces étiqueteurs. Comme conséquence de cette situation, nous avons été amenée, pour exploiter le corpus constitué, à baliser manuellement les occurrences de futur. Dans le fichier Transcriber, nous avons ainsi ajouté une balise sur chaque portion du texte exprimant un futur : (<Event desc="FUT" type="lexical" extent="begin"/> … <Event desc="FUT" type="lexical" extent="end"/>). Ce balisage dans le corps de transcription rend possible un retour facile vers les occurrences dans leur contexte et un accès rapide au signal sonore, nécessaire pour l'analyse de données.

---

[5] L'étiqueteur APT 'Arabic Part-of-speech Tagger' de Khoja (Khoja 2001) se présente comme une adaptation à l'arabe du système du British National Corpus (BNC) qui combine des techniques statistiques et des règles linguistiques pour déterminer tous les traits morphologiques d'une unité lexicale.

[6] Analyseur morphologique Sakher est un système développé par Chalabi (Chalabi, 2004). Il traite aussi bien l'arabe classique que l'arabe moderne, et il permet de déterminer la racine possible d'un mot en supprimant tous les affixes et suffixes, et en décrivant la structure morphologique de celui-ci.

[7] Analyseur morphologique Sebawai est un système développé par Darwish en 2003. Il permet de trouver les racines de mots.

[8] Aramorph est un analyseur distribué par le LDC (Linguistic Data Consortium) qui permet de segmenter un mot en trois séquences (préfixe racine post-fixe).

Les 2731 occurrences de futur ainsi identifiées ont par la suite été extraites grâce au logiciel d'analyse textométrique TXM[9], et exportées dans un tableau CSV, afin d'y être annotées. Chacune des occurrences du futur a ainsi été sous-spécifiée pour un certain nombre de traits morphosyntaxiques et sémantiques.

La dernière étape de ce processus a consisté à réinjecter sous TXM les occurrences et leurs annotations affinées dans l'objectif d'obtenir une analyse linguistique fine croisant approches qualitative et quantitative.

La réinjection sous TXM après annotation des occurrences identifiées permettra une analyse linguistique fine croisant approches qualitative et quantitative.

5 Conclusion

Dans cet article, il a été question de la constitution et du traitement de corpus de l'arabe tunisien parlé à Orléans.

La rareté des travaux sur l'arabe tunisien et le manque d'outils de traitement automatique de cette langue nous ont posé beaucoup de problèmes lors de la constitution de notre corpus, depuis le recueil de données jusqu'à l'annotation, en passant par la transcription. Afin de construire un corpus partageable et exploitable, il a été nécessaire d'opter pour quelques choix méthodologiques jugés les plus adéquats.

Bien qu'il ne soit pas assez représentatif, notre corpus peut être perçu comme un échantillon diversifié permettant d'observer plusieurs phénomènes qui découlent du fonctionnement du parler tunisien. Il propose une catégorie bien spécifique de locuteurs, habituellement non intégrés dans les corpus oraux, i.e. les locuteurs « sans papiers ». Néanmoins, cet échantillon ne représente pas toute la communauté tunisienne d'Orléans.

Après avoir transcrit et enrichi le corpus d'AT, il est devenu dés lors possible d'examiner nos données sur les plans quantitatif et qualitatif. Travailler sur des corpus oraux authentiques de deux langues différentes nous a permis d'observer la multiplicité d'emplois du futur et de nous approcher des propriétés de son fonctionnement à l'oral.

Bien que ce travail ne soit qu'une ébauche, la nature des données constituées peut apporter des informations concrètes sur la validité des modèles généraux de la description grammaticale.

[9] http://textometrie.ens-lyon.fr/

Références

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* F. Rastier, M. Ballabriga (dir.), *Corpus en Lettres et Sciences sociales — Des documents numériques à l'interprétation,* actes du XXVII colloque d'Albi, *Langages et signification,* publiés par C. Duteil-Mougel et B. Foulquié.

Abouda L. (2015). *Syntaxe et Sémantique en corpus. Du temps et de la modalité en français oral,* mémoire HDR, Université d'Orléans.

Baude O. (coord.) (2006). *Corpus oraux, Guide des bonnes pratiques.* CNRS éditions et P.U.O.

Baude O. (2008). Le droit de la parole, dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP.

Benjelloun S. (2002). Une double graphie, latine et arabe, pour enseigner l'arabe marocain, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, 331-340, L'Harmattan, Paris.

Bergounioux G. (dir.) (1992). Enquêtes, Corpus et Témoins, *Langue Française* 93.

Bergounioux G., et al. (1992). « L'Etude 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 dans Bilger, Mireille (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-Erudition.

Boukadida N. (2008). Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale).

Bourdieu P. (2003). (sous la direction de) *La misère du monde*, Paris, Seuil – Collection Point.

Caubet D. (1999). Arabe maghrébin : passage à l'écrit et institutions, In *Faits de Langues*, vol. 7, n° 13, 235-244.

Caubet D. (2002). Arabe maghrébin, langue de France : entre deux graphies, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, p.331-340, L'Harmattan, Paris, 2002.

Cerquiglini, B. (1999), *Les langues de la France*, rapport aux ministres de l'Éducation nationale et de la Culture et de la Communication.

Gadet F. (2000). Derrière les problèmes méthodologiques du recueil des données, dans M. Bilger (dir.), *Linguistique sur corpus*, Presses Universitaires de Perpignan.

Gadet F. (2008). L'oreille et l'œil à l'écoute du social, dans Bilger, Mireille (éd.). *Données orales: les enjeux de la transcription*. Perpignan. PUP, 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 ?, dans M. Bilger (dir.), *Linguistique sur corpus*, Presses universitaires de Perpignan.

Khoja S. (2001). Arabic part-of-speech tagger. In Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics, Carnegie Mellon University, Pittsburgh, 81–86.

Leech G. (1997). Introduction corpus annotation. In R. Garside, G. Leech, A. McEnery (Eds.), Corpus annotation: Linguistic information from computer text corpora. London, Longman, 1-18.

Maurer B. (1999). Quelles méthodes d'enquête sont effectivement employées aujourd'hui en sociolinguistique, dans L.-J. Calvet et P. Dumont (dir.), *L'enquête sociolinguistique*, L'Harmattan.

Mondada L. (2008). La transcription dans la perspective de la linguistique interactionnelle, dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription*. Perpignan. PUP, 78-109.