Sur les enjeux méthodologiques de la construction d'un corpus d'arabe tunisien

Fatma Ben Barka
LLL (UMR 7270), 10 Rue de Tours – BP 46527, 45065, France
fatma.messaoudi@univ-orleans.fr

RESUME

La constitution de corpus est une tâche délicate, surtout quand il s'agit d'une langue pour laquelle il manque de ressources et d'outils qui en facilitent le traitement. Telle est la situation de la plupart des parlers arabes, qui sera exposée ici à travers l'exemple de la construction d'un corpus d'arabe tunisien. Dans cet article, nous exposerons les principaux choix méthodologiques et technologiques que nous avons opérés pour répondre aux contraintes auxquelles nous avons dû faire face lors de la constitution de notre corpus.

ABSTRACT

On the methodological issues of the construction of a Tunisian Arabic corpus

The constitution of corpus is a delicate task, especially when it is a language for which it lacks resources and tools that facilitate the processing. This is the situation of most Arabic dialects, which will be exposed here through the example of the construction of a Tunisian Arabic corpus. In this article, we will present the main methodological and technological choices that we have made to respond to the constraints that we had to face during the constitution of our corpus.

RÉSUMÉ EN LANGUE NATIONALE

حول التحديات المنهجية لتكوين مدوّنة شفاهيّة للهجة العربية التونسية

إنّ إنشاء مدوّنة هو عمل دقيق خاصّة عندما يتعلق الأمر بلغة تفتقر إلى مصادر وأدوات تسهّل معالجتها. ذلك هو شأن أغلب اللهجات العربية المحليّة الذي سنعرضه في هذا البحث معتمدين على نموذج إعداد مدوّنة شفاهيّة للهجة العربية التونسيّة. في هذا المقال سنقدم أهم الخيارات المنهجية والتقنية التي قمنا بها تماشيا مع الصعوبات التي كان علينا مواجهتها عند إعدادنا لمدوّنة بحثنا

MOTS-CLES : corpus oral – arabe tunisien – méthodologie – transcription
KEYWORDS : oral corpus – tunisian arabic – methodology – transcription

1 Contexte

Dans les nombreuses études traitant de la question de l'emploi du subjonctif en français de points de vue syntaxique et sémantique (Nordahl, 1969, Nolke, 1985, Soutet, 2000...), les linguistes ont eu massivement recours soit à la fabrication d'exemples soit à l'emprunt d'exemples écrits, appartenant généralement aux genres littéraire ou journalistique. Ce choix, qui n'est pas dénué de circularité - ne faisant que retrouver dans les données les prescriptions fournies par la grammaire standard -, a vraisemblablement emprisonné l'étude du subjonctif dans un cadre normatif, susceptible d'empêcher de voir son fonctionnement réel, et de répondre aux questions théoriques fondamentales qui se posent dans le domaine de la flexion verbale. De par ce fait, le débat sur ses contextes d'emploi et ses valeurs sémantiques est loin d'être clos.

Cette considération nous a poussée à proposer un réexamen de l'emploi de ce mode verbal en français en nous basant sur des données orales et authentiques et a suscité notre intérêt pour nous

lancer dans la recherche et l'analyse d'éventuels équivalents du subjonctif dans une autre langue, notamment en arabe tunisien[1].

Pour les données en français, nous avons opté pour les Enquêtes Sociolinguistiques à Orléans (ESLO). En ce qui concerne celles de l'arabe, nous nous sommes reposée tout d'abord sur le corpus de notre collègue Yossra Ben Ahmed, d'une durée totale de 17h de paroles enregistrées en entretien face à face.

Nonobstant, dans un souci d'accumulation, nous avons mené une deuxième collecte des données auprès des locuteurs tunisiens, dans l'objectif de parvenir à un corpus plus vaste.

La constitution et l'exploitation d'un corpus échantillonné et diversifié de 13 heures d'enregistrements de l'arabe tunisien, effectué dans le cadre de notre étude doctorale (en cours), a exigé la mise en œuvre d'un ensemble de procédures, allant du recueil de données jusqu'à la phase de transcription.

Dans cet article, après avoir exposé les différents choix méthodologiques et technologiques opérés lors de l'élaboration du corpus de l'arabe tunisien (désormais AT), nous reviendrons sur les problèmes rencontrés lors de sa transcription.

2 Démarche

Pour recueillir un échantillon authentique de l'arabe tunisien, nous nous sommes reposée fondamentalement sur la méthodologie d'ESLO.

Grâce à sa taille importante (actuellement autour de 7 millions de mots) et à ses genres interactionnels assez diversifiés (entretiens, repas, conférences universitaires…), ce corpus offre la possibilité d'entamer des recherches sur des données orales situées et enrichies par des métadonnées informant sur la situation de parole et précisant le profil de chaque locuteur (son âge, son sexe, sa profession et sa catégorie socioprofessionnelle).

S'agissant d'une recherche comparative, les choix présidant à la construction du corpus de l'AT ont été guidés par un souci de comparabilité avec le corpus ESLO.

Pour le mode de collecte des données, nous avons favorisé l'entretien en face-à-face, « situation certes très formelle, mais qui avait l'avantage d'être (…) contrôlable » (Abouda & Baude, 2009 : 134). Cependant, pour maintenir un certain équilibre au sein de notre corpus, nous avons intégré, à hauteur de 20%, deux autres genres interactionnels de « contrôle », i.e. les repas et les cours universitaires.

Le guide d'entretien a été puisé dans les principaux thèmes retenus par ESLO (logement, travail, loisirs, langues, Orléans), tout en rajoutant d'autres questions (par exemple sur la révolution tunisienne) susceptibles de faire parler les locuteurs tunisiens, en visant les contextes propices à l'apparition des formes verbales subjonctives.

Suivant la démarche d'ESLO, indispensable pour rendre le corpus disponible et interopérable, nous avons réalisé une documentation détaillée de nos données[2] et de leurs contextes de

[1] http://www.axl.cefan.ulaval.ca/afrique/tunisie.htm

[2] Chaque locuteur dispose d'une fiche d'informations précisant son âge, son sexe, son niveau scolaire sa profession…

production.

Soucieuse de respecter l'unité de lieu, nous avons commencé notre enquête à Orléans. Mais, afin de répondre aux contraintes sociologiques, nous étions obligée d'effectuer des enregistrements en Tunisie.

Le corpus, que nous avons élaboré entre 2016 et 2017 (à Orléans et en Tunisie) est d'une durée globale de 13h, fractionnée en 19 enregistrements, un volume jugé suffisant pour mener notre étude contrastive sur le subjonctif.

Corpus de l'AT (2016-2017)	
Lieux d'enquête	Orléans – Tunisie
Nombre d'heures	13
Nombre de mots	108705
Nombre de locuteurs	19
Situations de communications	entretien face à face – repas – cours universitaires
Catégories des locuteurs	Âge : 15-35 ans / 35-55 ans / plus de 55 ans Sexe : Hommes : 40% / Femmes : 60% Catégories socioprofessionnelles : • Ouvriers • Employés • Cadres, professions intellectuelles supérieures

TABLE 1 – Le corpus en question

3 Transcription

Une fois le corpus construit et afin de faciliter son traitement, un travail préalable de transcription a été nécessaire. En bref, ainsi le note Blanche-Benveniste (2000 : 24) : « on ne peut pas étudier l'oral par l'oral, en se fiant à la mémoire qu'on en garde. On ne peut pas, sans le secours de la représentation visuelle, parcourir l'oral en tous sens et en comparer les morceaux. »

La phase d'annotation de nos données brutes a soulevé plusieurs interrogations : quels système et mode de notation choisir, quelles conventions adopter et sur quel outil transcrire ?

complétée par des renseignements sur l'enregistrement (n°, genre interactionnel, témoin(s), lieu, date et durée de l'enregistrement…)

3.1 Système de notation

L'arabe tunisien est une langue à tradition orale. Il peut être considéré comme une variante de l'arabe classique, mais ces deux langues se distinguent par deux systèmes spécifiques aux niveaux phonologique, morphologique, syntaxique et lexical.

En général, les parlers arabes maghrébins ont été notés à l'aide de deux systèmes graphiques, i.e. arabe et latin selon la tradition du champ, les préférences idéologiques et les enjeux pratiques.

Afin de garantir une facilité technique, nous avons opté finalement pour la graphie latine, un choix qui nous paraît le moins discutable à l'heure actuelle. De par ce fait, il nous permet non seulement d'écarter les contraintes de l'écriture arabe (de droite à gauche), mais aussi de constituer un corpus de référence de l'arabe tunisien, partageable et lisible par les non-natifs.

3.2 Mode de notation

La sélection d'un mode de notation se fait selon les finalités de recherche et les degrés de représentativité de la langue parlée.

Pour réaliser une transcription alignée avec le son, nous avons eu la possibilité de choisir entre plusieurs types de notation (phonétique, orthographique, phonologique et morphologique). En définitive, notre choix s'est reposé sur une notation orthographique (usuelle) d'*inspiration phonologique* où les dimensions morphosyntaxiques des énoncés sont préservées.

La sélection de ce mode de transcription se justifie par les raisons suivantes :

- le sujet de notre recherche n'exige ni une notation phonétique, ni une notation phonologique stricte ;
- la nature de ce mode permet non seulement d'éliminer les ambigüités (notamment syntaxiques), mais aussi de fournir un décodage facile par tout type de lecteurs ;
- la possibilité d'ajouter d'autres niveaux de transcription (phonétique ou/et phonologique), suivant les attentes et les objectifs des chercheurs ;
- la facilité d'une lecture rapide et la simplification de l'exploitation des logiciels de concordance.

Néanmoins, en l'absence d'un standard stabilisé, nous avons dû tenir compte des pratiques orthographiques les plus usuelles au sein de la communauté scientifique.

3.3 Conventions de transcription

Les conventions de transcription varient d'une langue à une autre. En effet, si le français est une langue à tradition écrite standardisée, l'arabe tunisien est langue privée de toute tradition orthographique.

Autre point, ces conventions se partagent en deux:

- des conventions « *spécifiques* » à chaque langue ;
- des conventions « *communes* » à tout corpus oral quelle que soit la langue.

Le choix entre les différents types de convention se détermine donc par de nombreuses

considérations, allant de la taille du corpus jusqu'aux objectifs de la recherche, en passant par la nature des données (écrites, orales).

Pour la transcription de notre corpus de l'AT, nous nous sommes basée fondamentalement sur les recommandations de l'INALCO (1996-1998). Quant aux phénomènes de l'oralité, nous avons adopté les conventions proposées par le LLL dans le cadre du projet ESLO.

3.4 Outil de transcription

Notre corpus a été transcrit sous TRANSCRIBER[3], un logiciel d'aide à l'annotation des données audio permettant de transcrire plusieurs langues (européennes et/ou non européennes).

Ce logiciel, qui a été développé par Claude Barras et Edouard Geoffroy de la Direction Générale de l'Armement (DGA), facilite la visualisation, la manipulation et le traitement des sources sonores.

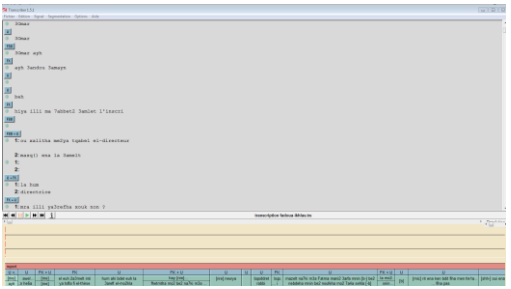

FIGURE 1 – Interface de Transcriber

Nonobstant, lors de la transcription, nous avons rencontré un problème d'affichage de ces caractères spécifiques :

- (ج) ja = ž (chuintante sonore)
- (ش) sh = š (chuintante sourde)
- (خ) kha = x (vélaire sourde)
- (ح) ha = ḥ (pharyngale sourde)
- (ق) qa = q (uvulaire sonore)
- (ع) aa = Ɛ (pharyngale sonore)
- (غ) gha = ġ (vélaire sonore)

Pour résoudre ce problème, il a fallu choisir l'encodage UTF-8.

[3] Ce logiciel est téléchargeable sur : http://trans.sourceforge.net.

4 Conclusion

Dans ce papier, nous avons présenté les différentes étapes de la construction et de l'exploitation d'un corpus de l'arabe tunisien.

Comme nous avons pu le montrer, la constitution de ce corpus s'est heurtée à la rareté des travaux sur le parler tunisien. Nous avons été donc confrontée à plusieurs points problématiques pour lesquels nous avons dû opérer des choix méthodologiques et technologiques afin de rendre le corpus disponible et interopérable.

Dans la volonté d'atteindre un certain degré de représentativité du parler tunisien, nous avons varié autant que possible les situations de communications et les profils sociologiques des locuteurs. Le résultat était un corpus assez diversifié et échantillonné.

Après l'avoir transcrit, il devient dès lors possible de l'annoter afin d'enrichir nos données syntaxiquement et sémantiquement.

L'objectif de notre travail était de constituer un corpus de référence d'arabe tunisien qui peut faire l'objet de prochaines recherches sur cette langue peu dotée.

Références

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

ABOUDA L., BAUDE O. (2005). Du Français Fondamental aux ESLO. In Actes du Colloque international Français fondamental, corpus oraux, contenus d'enseignement, pages 131-146.

BENJELLOUN S. (2002). Une double graphie, latine et arabe, pour enseigner l'arabe marocain.

In Codification des langues de France, pages 331-340.

BERGOUGNIOUX G. (dir.) (1992). Enquêtes, Corpus et Témoins. In Langue française 93, pages 3-22.

BILGER M., CAPPEAU P. (2004). L'oral ou la multiplication des styles. In Langage et Société 109, pages 13- 30.

BLANCHE-BENVENISTE C., JEANJEAN C. (1987). Le français parlé : transcription et édition. Paris : Didier-Erudition.

BLANCHE-BENVENISTE C. (2000). « Transcription de l'oral et morphologie », In Romania Una et diversa, Philologische Studien für Theodor Berchem (Gille M. et Kiesler R. Eds). Tübingen : Gunter Narr, pages 61-74.

BOUKADIDA N. (2008). Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale). Université Rennes 2; Université de Tunis

CAUBET D. (1999). Arabe maghrébin : passage à l'écrit et institutions. In Faits de Langues 13, pages 235- 244.

CAUBET D. (2002). Arabe maghrébin, langue de France : entre deux graphies. In Codification des langues de France, pages 331-340.

GADET F. (2008). L’oreille et l’œil à l’écoute du social. In Données orales: les enjeux de la transcription, pages 35-47.

LEECH G. (1997). Introduction corpus annotation. In Corpus annotation: Linguistic information from computer text corpora, pages 1-18.