Constitution d'un corpus oral de l'arabe tunisien : une ressource essentielle pour l'étiquetage morphosyntaxique

Yossra Ben Ahmed, Flora Badin, Linda Hriba

Constitution d’un corpus oral de l’arabe tunisien: une ressource essentielle pour l’étiquetage morphosyntaxique

Yossra Ben Ahmed[1,2] Flora Badin[1] Linda Hriba[1,2]

(1) Laboratoire Ligérien de Linguistique, 10, rue de Tours 45065 Orléans Cedex 2
(2) Université d’Orléans, 10, rue de Tours 45065 Orléans Cedex 2
ben.ahmed.yossra@gmail.com, flora.badin@univ-orleans.fr, linda.hriba@yahoo.fr

Résumé

La constitution d’un corpus oral d’arabe tunisien pour l’analyse des expressions du futur a soulevé plusieurs problématiques : collecte des données, transcription et annotation. Après avoir exposé les enjeux théoriques et méthodologiques de chaque phase de traitement nous montrerons en quoi notre corpus servira de ressource pour la création d’un étiqueteur morphosyntaxique de l’arabe tunisien translittéré. Disposer et rendre accessible de tels corpus et outil faciliteront les recherches sur cette langue peu dotée et ouvriront de nouvelles perspectives de traitement.

Abstract

The constitution of an oral corpus of Tunisian Arabic : an essential resource for part-of-speech tagging

The constitution of an oral corpus of Tunisian Arabic for the analysis of the expressions of future raised several problems : data collection, transcription and annotation. After presenting the theoretical and methodological stakes for each step of the process, we will show how our corpus can be used as a resource for the creation of a morphosyntactic tagger of transliterated Tunisian Arabic. Giving access to such corpora and tools will facilitate the research on a poorly documented language and will open new perspectives of language processing.

Résumé en langue nationale

لقد أدّى بناء عينة شفوية من اللغة العربية التونسية، بغية تحليل تعبيرات زمن المستقبل،إلى إثارة العديد من الإشكاليات، من اهمها :جمع البيانات، نسخها و توسيمها. بعد عرض القضايا النظرية لكل مرحلة من مراحل المعالجة، سنوضح كيف أن هذه العينة الشفوية سوف تعمل من جهة كمصدر مهم لإنشاء أداة تمييز بنيوية نحوية. ومن جهة أخرى،لإتاحة مادة من اللهجة التونسية مما من شأنه ان يسهل البحث في هذه اللغة غير المتطرق إليها كثيرا و فتح آفاق جديدة من المعالجات العلمية

Mots-clés : arabe tunisien, corpus oraux, transcription, annotation, étiquetage morphosyntaxique.

Keywords : Tunisian Arabic, oral corpora, transcription, annotation, part-of-speech tagging.

1. Introduction

La constitution des corpus oraux est indispensable si l'on veut éviter les pièges de l'intuition et confronter les modèles théoriques aux observables. Notre corpus d'arabe tunisien (désormais AT), constitué par l'un des auteurs de cette proposition dans le cadre d'une recherche doctorale (Ben Ahmed, en cours) sur l'expression du futur en français et en arabe tunisien parlés, ne consiste cependant pas en un simple enregistrement de données orales. Il s'agit de la construction d'une véritable base de données transcrites et annotées de façon à assurer son exploitation par la communauté scientifique. Nous nous proposons dans cette communication d'exposer dans un premier temps les principaux choix opérés lors de la constitution, de la transcription et de l'annotation du corpus d'AT (Ahmed, 2017) puis nous présenterons les premiers éléments méthodologiques qui conduiront à terme au développement d'un étiqueteur morphosyntaxique.

2. La constitution du corpus

2.1. Situation des corpus oraux en arabe tunisien

Nous recensons à l'heure actuelle peu de corpus d'enregistrements de données orales pour l'arabe tunisien :

- Le corpus Tunisian Dialect Corpus Interlocutor "TuDiCol", (Graja *et al.,* 2010) est constitué de conversations collectées dans des situations agents et clients enregistrées dans des gares tunisiennes. Le corpus contient 127 dialogues combinant 893 discours représentant 3403 mots.
- La même situation d'enregistrement a été utilisée pour la constitution du corpus Tunisian Arabic Railway Interaction Corpus "TARIC" (Masmoudi *et al.,* 2014). Il s'agit d'un corpus de conversations entre passagers et agents de la SNCFT composé de 20h d'enregistrements audio, 4662 dialogues, 18657 d'énoncés et 71684 de mots. L'objectif consistait à demander en dialecte tunisien des informations (horaires, prix, réservation des billets etc. . .) sur les services de chemin de fer dans une gare ferroviaire.
- Enfin dans le cadre d'un travail de recherche de doctorat, (Boujelbane *et al.,* 2015) a construit une ressource pour le traitement automatique du dialecte tunisien parlé dans les médias. Le corpus est constitué de 1h42 d'enregistrements (12 207 mots) issus de journaux télévisés et de 3h40 d'émissions de débat (25 757 mots).

Dans le cadre du corpus AT, que nous décrivons dans la section suivante, nous avons privilégié les entretiens en face à face qui offre l'avantage de disposer pour chaque locuteur de données sociologiques (âge, catégorie socioprofessionnelle, niveau d'études...).

2.2. Méthodologie de constitution du corpus de l'AT

Après avoir longtemps privilégié les exemples fabriqués ou attestés mais essentiellement écrits et littéraires (pour des raisons à la fois épistémologiques et pratiques), le domaine de la linguistique bénéficie depuis une décennie (si on fait abstraction de quelques précurseurs) de l'avènement de corpus oraux. Pour l'analyse des expressions du futur en arabe tunisien, nous avons constitué un corpus oral situé. La prise en compte de ces données orales y est vue comme une occasion d'étudier une langue peu dotée et de mettre en lumière des fonctionnements linguistiques qui échappent à l'intuition. Notre corpus a été constitué, entre 2013 et 2014, auprès de locuteurs tunisiens natifs résidant à Orléans (France). Il représente un volume de 17h, fractionné en 37 enregistrements (141 250 mots), ce qui nous semble suffisant pour les investigations envisagées. Pour améliorer la représentativité de notre corpus, nous avons diversifié les catégories de locuteurs en différenciant sociologiquement les témoins par l'âge, le sexe, le niveau scolaire, la profession et les langues parlées. Depuis le développement de la linguistique du corpus (Habert, 2000), 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, nous avons procédé à une documentation précise de nos données afin d'avoir un corpus attesté et situé :

> « 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).

En ce qui concerne le mode de recueil des données, nous avons favorisé l'entretien en face-à-face, « situation certes très formelle, mais qui avait l'avantage d'être (...) contrôlable » (Abouda et Baude, 2005). Un guide d'entretien a été réalisé afin de faire parler les locuteurs. Les questions que nous avons choisies portent sur six thèmes (logement/Orléans, travail, loisirs, questions évaluatives sur Orléans, langue, recette).

3. La transcription

3.1. Le mode de transcription

Les données sonores collectées ne peuvent être analysées et traitées sans un travail préalable de transcription. Cette étape a soulevé plusieurs interrogations : Quels systèmes graphiques privilégier? Quelles conventions adopter? Quels outils pour transcrire?

Pour la transcription de notre corpus, deux systèmes graphiques (latin et arabe) s'offrent à nous. De nombreux paramètres dictent le choix du système : tradition du champ, préférences idéologiques et facilité technique. Nous avons ainsi opté pour une

transcription avec une graphie latine dont l'avantage est de fournir un corpus partageable et facilement lisible par les non-natifs.

Le choix d'un mode de transcription implique des enjeux qui dépendent des objectifs de la recherche, de la représentation de la langue parlée, ou encore de la représentation que l'on veut en donner (Bilger, 2008). « La transcription ne peut être regardée comme une opération banale, car on transcrit pour donner à voir quelque chose.» (Gadet, 2008).
Il existe plusieurs types de notation : phonétique, phonologique, morphologique et orthographique (usuelle). Pour notre corpus, nous avons opté pour une transcription orthographique d'inspiration phonologique avec la prise en compte de l'aspect morphosyntaxique des énoncés. Ce choix a été motivé par les raisons suivantes :

- l'objet d'étude, qui dans notre cas, ne nécessite ni une notation phonétique, ni une notation phonologique stricte,
- la simplicité de ce mode de notation qui 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.

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

3.2. Les conventions de transcription

Notre corpus a été transcrit sous Transcriber (Barras *et al.,* 2001), un logiciel d'aide à la transcription manuelle de fichiers audio qui permet de transcrire de nombreuses langues y compris non européennes avec l'utilisation du codage de caractères UTF-8. Pour le choix des conventions, qui n'est jamais neutre, différents facteurs ont dû être pris en compte : les finalités de la recherche, la taille du corpus et le type des données primaires (audio ou vidéo). Ces conventions diffèrent entre une langue écrite et bien établie (i.e. le français) et les langues sans traditions orthographiques solides (i.e. l'arabe tunisien). Nous distinguons les conventions « spécifiques » à chaque langue et les conventions « communes » à tout corpus oral quelle que soit la langue. Pour les premières, le recours aux propositions retenues par l'INALCO (1996-1998), s'est imposé. En ce qui concerne le deuxième type de conventions, portant sur les phénomènes liés à l'oralité, les propositions du LLL pour le corpus ESLO[1] ont été privilégiées.

4. L'annotation du corpus

4.1. L'annotation manuelle des expressions du futur

L'annotation constitue incontestablement une « valeur ajoutée » (Leech, 1997) grâce aux informations de nature différente qu'elle permet d'ajouter aux données. A l'heure actuelle, il n'existe pas de systèmes complets et disponibles pour l'étiquetage de l'arabe tunisien, situation différente pour l'arabe standard pour lequel des étiqueteurs morphosyntaxiques relativement robustes ont été développés (Arabic Part-of-speech Tagger, Sakher, Sebawai, Aramoph, etc.).
Des différences syntaxiques, lexicales et phonologiques entre l'arabe tunisien et l'arabe classique (Boukadida, 2008) nous ont conduit à ne pas recourir à ces étiqueteurs pour l'annotation de notre corpus. Nous avons ainsi été contraintes d'annoter manuellement les occurrences du futur. Chaque portion de texte exprimant le futur a été balisée directement dans l'outil Transcriber sous la forme d'évènements (Figure 1)

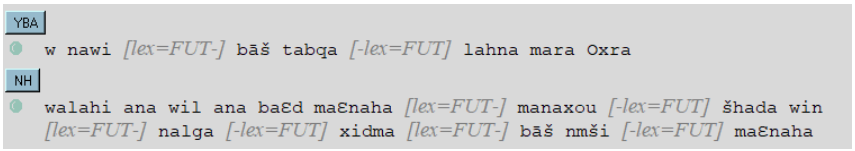


Figure 1 : Annotation du futur sous transcriber

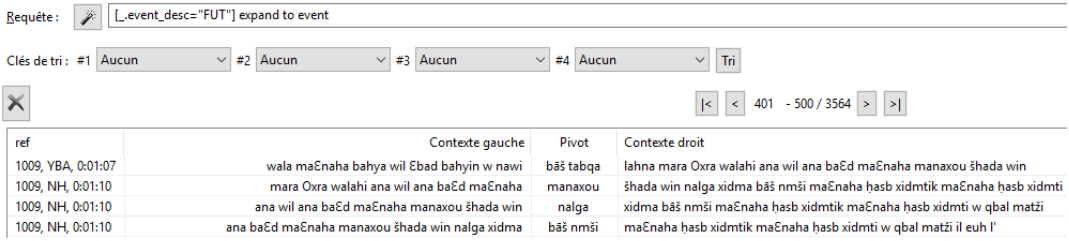
1

Le balisage dans le corps de la transcription rend de ce fait aisé l'identification des occurrences dans leur contexte en assurant un accès au signal sonore, nécessaire pour l'analyse des données. La transformation des données via un script python (Figure 2) pour l'exploitation dans le logiciel d'analyse textométrique TXM (Heiden *et al.*, 2010) a été nécessaire.

```
<Turn speaker="spk1" mode="spontaneous" fidelity="high" startTime="67.75" endTime="70.098">
<Sync time="67.75"/>
w nawi
<Event desc="FUT">
 bāš tabqa
</Event>
 lahna mara Oxra
</Turn>
<Turn speaker="spk2" mode="spontaneous" fidelity="high" startTime="70.098" endTime="75.419">
<Sync time="70.098"/>
walahi ana wil ana baƐd maƐnaha
<Event desc="FUT">
 manaxou
</Event>
 šhada win
<Event desc="FUT">
 nalga
</Event>
 xidma
<Event desc="FUT">
 bāš nmši
</Event>
 maƐnaha
</Turn>
```

Figure 2 : Document XML d'exploitation

L'utilisation de ce logiciel permet de plus la recherche couplée avec les métadonnées locuteurs recueillies lors de l'entretien.
Les 3564 occurrences du futur sont alors recherchées via le système de requête de l'outil. Le résultat se présente sous forme de concordancier (Figure 3).



ref	Contexte gauche	Pivot	Contexte droit
1009, YBA, 0:01:07	wala maƐnaha bahya wil Ɛbad bahyin w nawi	bāš tabqa	lahna mara Oxra walahi ana wil ana baƐd maƐnaha manaxou šhada win
1009, NH, 0:01:10	mara Oxra walahi ana wil ana baƐd maƐnaha	manaxou	šhada win nalga xidma bāš nmši maƐnaha ḥasb xidmtik maƐnaha ḥasb xidmti
1009, NH, 0:01:10	ana wil ana baƐd maƐnaha manaxou šhada win	nalga	xidma bāš nmši maƐnaha ḥasb xidmtik maƐnaha ḥasb xidmti w qbal matži
1009, NH, 0:01:10	ana baƐd maƐnaha manaxou šhada win nalga xidma	bāš nmši	maƐnaha ḥasb xidmtik maƐnaha ḥasb xidmti w qbal matži il euh l'

Figure 3 : Concordancier du logiciel TXM

L'export de ce dernier via la macro BuildWordPropTable a permis la sous-spécification des occurrences du futur avec des traits morphosyntaxiques et sémantiques. Enfin, la macro InjectWord- PropTable offre la possibilité de requêtage de cet enrichissement dans l'outil en vue d'une analyse linguistique fine.
Cette annotation manuelle a suscité l'envie de créer un étiqueteur morphosyntaxique de l'arabe tunisien. Dans la partie suivante, nous expliquerons notre démarche pour aboutir à cette ressource encore non disponible pour une telle langue.

4.1.1. Vers un étiquetage morphosyntaxique

L'étiquetage morphosyntaxique est une plus-value pour les analyses futures d'un corpus et est essentiel pour faciliter le traitement des données par des applications de Traitement Automatique des Langues (TAL). Pour le traitement automatique de l'arabe classique, il existe quelques analyseurs dont :

- APT 'Arabic Part-of-speech Tagger' (Khoja, 2001) qui se présente comme une adaptation à l'arabe du système du British National Corpus (BNC). L'outil combine des techniques statistiques et des règles linguistiques pour déterminer tous les traits morphologiques d'une unité lexicale,
- Sakher (Chalabi, 2004) qui traite aussi bien l'arabe classique que l'arabe moderne. Il permet de déterminer la racine possible d'un mot en supprimant tous les affixes et suffixes et d'en décrire la structure morphologique,

- Aramorph, analyseur distribué par le LDC (Linguistic Data Consortium), qui segmente un mot en trois séquences (préfixe | racine | post-fixe).

Cependant, ces outils ne peuvent être appliqués pour le traitement de nos données, notre objet d'études étant l'arabe tunisien oral. Or, les méthodes d'analyse utilisées peuvent être semblables pour la création de notre ressource. Notre travail s'effectuera en trois phases :

1. mise en place d'un système de segmentation lexicale de l'arabe tunisien, spécifiquement de la graphie latine,
2. lemmatisation en vue de créer un dictionnaire,
3. proposition d'un jeu d'étiquettes (verbe, nom, préposition, conjonction, pronom, article, interjection, adverbe, ...) pour l'annotation de chaque unité lexicale en partie du discours.

Notre jeu de données sera constitué de 2500 mots. Pour la segmentation nous nous baserons sur une liste déjà existante de clitiques, préfixes, suffixes et racines (Mars et al., 2016). Car, contrairement au français, chaque syllabe peut être une unité lexicale. Aussi, nous annoterons et segmenterons manuellement notre jeu de données en vue de servir de corpus d'apprentissage au logiciel Treetagger (Schmid, 1995). L'utilisation d'un lexique adapté aux fonctionnalités du logiciel sera incontournable. Une fois toutes ces ressources produites, l'entrainement de Treetagger pourra être amorcé. Dès les premiers résultats obtenus nous pourrons envisager les tests sur un nouveau jeu de données afin d'améliorer notre système.

5. Conclusion

La constitution d'un corpus d'arabe tunisien est une ressource indéniable pour la communauté scientifique. Aussi comme nous l'avons souligné elles nécessitent plusieurs phases de traitement qui soulèvent chacunes des problématiques liées entre autres au manque de cadre théorique et d'outils pour l'AT. Face à l'accroissement des données et pour pallier le manque de ressources nous proposons ainsi la mise à disposition d'un corpus transcrit et annoté de l'AT qui permettra à terme le développement d'un étiqueteur morphosyntaxique.

6. Références de bibliographie

ABOUDA, L. et BAUDE, O. (2005). Du français fondamental aux eslo. *In Grand corpus de français parlé, Bilan historique et perspectives de recherche,* volume 33, pages 131-146.

AHMED, Y. B. (2017). Constitution d'un corpus d'arabe tunisien parlé à orléans. *Actes des 9èmes Journées Internationales de la Linguistique de corpus,* page 173.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (2001). Transcriber : development and use of

a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5-22.

BILGER, M. (2008). *Données orales : les enjeux de la transcription.* Presses Univ. de Perpignan.

BOUJELBANE, R., ELLOUZE, M., BÉCHET, F. et BELGUITH, L. (2015). De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens. *Revue TAL*, pages rahma-boujelbane.

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

CHALABI, A. (2004). Sakhr arabic lexicon. *In NEMLAR international conference onArabic language resources and tools*, pages 21-24.

GADET, F. (2008). L'oreille et l'oeil à l'écoute du social.

GRAJA, M., JAOUA, M. et HADRICH BELGUITH, L. (2010). Lexical study of a spoken dialogue corpus in tunisian dialect. *In The international arab conference on information technology (acit), benghazi-libya.* Citeseer.

HABERT, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment. *Cahiers de l'Université de Perpignan*, 31:11-58.

HEIDEN, S., MAGUÉ, J.-P et PINCEMIN, B. (2010). Txm : Une plateforme logicielle open-source pour la textométrie-conception et développement. *In 10th International Conference on the Statistical Analysis ofTextual Data-JADT2010*, volume 2, pages 1021-1032. Edizioni Universitarie di Lettere Economia Diritto.

KHOJA, S. (2001). Apt : Arabic part-of-speech tagger. *In Proceedings ofthe Student Workshop at NAACL*, pages 20-25.

LEECH, G. (1997). Abrief users' guide to the grammatical tagging of the british national corpus. *UCREL, Lancaster University.*

MARS, M., ZRIGUI, M., BELGACEM, M. et ZOUAGHI, A. (2016). A semantic analyzer for the comprehension of the spontaneous arabic speech. *arXiv preprint arXiv :1610.02493.*

MASMOUDI, A., KHMEKHEM, M. E., ESTEVE, Y., BELGUITH, L. H. et HABASH, N. (2014). A corpus and phonetic dictionary for tunisian arabic speech recognition. *In LREC*, pages 306-310.

SCHMID, H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universitat Stuttgart*, 43:28.