Transcription de corpus oraux d'arabe parlé en interaction. Convention AraPI et annexes.

Lina Choueiri, Loubna Dimachki, Catherine Pinon, Véronique Traverso

Transcription de corpus oraux d'arabe parlé en interaction : La convention AraPI

Auteurs :
Véronique Traverso, Catherine Pinon, Loubna Dimashki, Lina Choueiri

Présentation du projet AraPI

- *L'équipe (groupe AraPI – Arabe Parlé en Interaction) :*
  - Ifpo (Véronique Traverso, Catherine Pinon)
  - Université libanaise (Loubna Dimashki, Moustafa al-Hajj)
  - AUB (Lina Choueiri)
  - CNRS - laboratoire ICAR (Joseph Dichy, Carole Etienne)
  - CNRS - laboratoire SEDYL (Stefano Manfredi)
  - CNRS - laboratoire LLL (Layal Kanaan).

- *Les financements :*
  - Ifpo
  - Labex ASLAN
  - Laboratoire ICAR (CNRS)
  - Université libanaise

- *Les objetcifs :*
  - établir une convention de transcription pour les corpus d'arabe parlé en interaction
  - alimenter une base de données (corpus transcrits et alignés)

- *Les publications :*
  - Traverso, Pinon, Dimachki et Choueiri : « Corpus d'arabe parlé (1) : Quels corpus d'arabe parlé en libre accès ? », *les carnets de l'Ifpo, la recherche en train de se faire à l'Institut français du Proche-Orient* [En ligne] https://ifpo.hypotheses.org/8865
  - Traverso, Pinon, Dimachki et Choueiri : « Corpus d'arabe parlé (2) : contraintes et problèmes liés à la réalisation des corpus d'arabe parlé en interaction », *les carnets de l'Ifpo, la recherche en train de se faire à l'Institut français du Proche-Orient* [En ligne] https://ifpo.hypotheses.org/9039
  - Traverso, Pinon, Dimachki et Choueiri : « Corpus d'arabe parlé (3) : Choix pour la notation des sons et des phénomènes dans la réalisation de corpus d'arabe parlé », à paraître dans *les carnets de l'Ifpo, la recherche en train de se faire à l'Institut français du Proche-Orient* [En ligne] https://ifpo.hypotheses.org/9305
  - Pinon, Traverso, Dimachki et Choueiri : « Corpus d'arabe parlé (4) : La convention de transcription ARAPI pour l'arabe parlé en interaction », à paraître dans *les carnets de l'Ifpo, la recherche en train de se faire à l'Institut français du Proche-Orient* [En ligne].

ARABE PARLE EN INTERACTION (ARAPI)
CONVENTIONS

INTRODUCTION

Les transcriptions de corpus d'interaction en arabe parlé pour des analyses de linguistique interactionnelle ou pragmatique en sont encore à leurs balbutiements. Chaque chercheur est confronté aux mêmes difficultés, et forge ses propres solutions en fonction des caractéristiques de ses données et de ses objectifs. Les choix effectués, eux, sont très variés, s'étageant autour de quatre possibilités majeures : une transcription phonétique, qui cherche à rendre compte de façon précise des spécificités sonores des productions ; une transcription phonologique, qui ne retient que les variantes distinctives de la variété concernée ; une translittération, c'est-à-dire une représentation fidèle, lettre par lettre, de l'orthographe arabe en alphabet latin ; et enfin une transcription en caractères arabes.

La convention ARAPI a opté pour une transcription en quatre tiers pour rendre les corpus transcrits exploitables par des chercheurs de différents domaines, éventuellement non arabophones, y compris en recourant aux outils informatiques sur script latin ou arabe. Les quatre tiers concernent 1) une transcription interactionnelle ; 2) une transcription morpho-phonologique assortie d'une glose ; 3) une transcription en caractères arabes ; 4) une traduction en français. Chacune des tiers est présentée en détail dans le document qui suit.

La convention est établie à partir de corpus syro-libanais, mais elle est conçue pour être étendue à d'autres dialectes arabes, moyennant son adaptation à leurs spécificités. De même, la convention peut être utilisée par les chercheurs de différents horizons, quelle que soit leur spécialité : chaque chercheur transcrit selon ses besoins, en sélectionnant la ou les tiers qui sont nécessaires à sa recherche, voire en ajoutant des éléments conventionnels qu'il doit expliciter, lorsqu'un phénomène qu'il a besoin de représenter n'est pas prévu dans la convention.

0. PRESENTATION GENERALE DE LA CONVENTION

Ce document détaille les conventions établies pour la transcription de sources produites en arabe parlé en interaction. Il est augmenté de plusieurs annexes. La première détaille la convention de notation pour les phénomènes oraux. La deuxième propose la liste de tous les éléments linguistiques conventionnels sous forme de tableaux. Cette liste est amenée à évoluer avec l'ajout de nouvelles transcriptions au corpus. La troisième annexe présente les règles générales pour la transcription des dialectes, inspirées de CODA (Conventional Orthography for Dialectal Arabic, Habash et al. 2012[1]). La quatrième récapitule les gloses morphosyntaxiques utilisées dans la tier 2b. Et enfin, la cinquième est un tableau synoptique qui permet d'appréhender rapidement les différences notables entre les tiers.

0. 0. Forme de la transcription

La police utilisée pour les tiers 1 et 2a est Courier New, 10 pt, en noir. Pour la tier 2b, 3 et 4, on utilise Times New Roman 11. On insère un saut de ligne entre chaque tier [2].

Les tiers commencent par l'abréviation du pseudonyme du locuteur (généralement en trois lettres, ex. AMI), suivies d'une tabulation, puis du texte prononcé par le locuteur. Toute parole transcrite doit être rapportée au locuteur qui l'a produite (voir le détail des mises en page dans l'annexe 1).

Exemple :

AMI →	`transcription interactionnelle (tier 1)`
`transcription morpho-phonologique (tier 2a)`
glose morpho-syntaxique (tier 2b)
كتابة عربية (tier 3)
traduction dans la langue du chercheur (tier 4)

0. 1. Une transcription en plusieurs tiers

Pour l'établissement d'un corpus complet, la transcription est réalisée en quatre tiers.

Tier 1 : transcription interactionnelle
La tier 1 est une transcription adaptée aux analyses pragmatiques, discursives et interactionnelles. Elle fonctionne en lien avec les données primaires (enregistrements), et elle indépendante de la tier 2. La transcription choisie est phonético-phonologique ; elle conserve le découpage en mots et ne prétend pas rendre compte de toutes les variantes de prononciation. Elle applique également certaines conventions de segmentation des morphèmes et indique les emprunts aux langues étrangères.

[1] Habash, Nizar & Diab, Mona & Rambow, Owen. (2012). Conventional Orthography for Dialectal Arabic. Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul.
[2] Ces informations sont pertinentes pour un travail sous word. Elles n'ont plus lieu d'être pour des transcriptions effectuées à l'aide de logiciels de transcription, qui disposent d'interfaces à multiples tiers (comme ELAN ou Praat).

Tier 2 dédoublée : transcription morpho-syntaxique
La tier 2 correspond à une représentation phonologique d'un standard du dialecte concerné qui intègre un certain nombre de choix conventionnels (tier 2a). Elle est assortie d'une glose morpho-syntaxique suivant la référence LeipzigGlossingRules (tier 2b). Voir l'annexe 4 pour la liste des étiquettes morpho-syntaxiques et des exemples en arabe libanais.

Tier 3 : transcription en caractères arabes
La tier 3 est une transcription en caractères arabes s'inspirant de la convention CODA (Conventional Orthography for Dialectal Arabic) [3]. Voir dans l'annexe 3 les règles générales pour la transcription des dialectes en caractères arabes.

Tier 4 : traduction
Enfin la tier 4 propose une traduction. Elle se veut fidèle au sens davantage qu'à la structure morpho-syntaxique de l'énoncé. Certains éléments difficilement traduisibles font l'objet de notations conventionnelles.

0. 2. Les phénomènes interactionnels

La notation des phénomènes interactionnels ne concerne que la tier 1. Celle-ci comporte tous les éléments relatifs à la production de la langue parlée en interaction : chevauchements de parole, pauses, enchaînements immédiats, notation sommaire de la prosodie.
Pour ces phénomènes, la convention AraPI reprend la convention ICOR, qui s'inspire des conventions de Jefferson (2004) généralement utilisées en linguistique interactionnelle, avec de légères modifications. Ces conventions sont présentées dans l'annexe 1, Conventions de transcription pour les phénomènes oraux et interactionnels.
Une version étendue des conventions interactionnelles est consultable sur le site CORINTE, http://icar.univ-lyon2.fr/projets/corinte/.

0. 3. L'alphabet de transcription

La convention ARAPI utilise l'alphabet phonétique international (API) avec quelques modifications :

- les voyelles longues sont notées avec un trait suscrit (ā, ū, ī, etc.) plutôt qu'avec les deux points (a:, u:, i:, etc.) déjà employés pour la transcription interactionnelle ;
- les consonnes emphatiques sont marquées avec des points sous les consonnes non-emphatiques correspondantes (ṣ, ḍ, ṭ, ẓ). La transcription de l'interdentale emphatique devrait être le ð avec un point souscrit. Nous la noterons ð.

Dans les tableaux ci-dessous, pour des raisons pratiques, les sons de la langue sont présentés selon l'ordre de l'alphabet arabe.

[3] Habash, Diab et Rambow, "Unified Guidelins and Resources for Arabic Dialect Orthography" [En ligne] http://www.lrec-conf.org/proceedings/lrec2018/pdf/395.pdf
Habash, Diab et Rambow, "Conventional Orthography for Dialectal Arabic" [En ligne] http://www.lrec-conf.org/proceedings/lrec2012/pdf/579_Paper.pdf

0. 3. 1. Consonnes

Arabe	API	Exemple	Arabe	API	Exemple
ب	b, p	baba, paspor	ط	ṭ	ṭawla
ت	t	tīn	ظ	ð, ḍ, ẓ	ðohor, ḍəhr, būẓa
ث	θ, t, s	θaqāfe, talʒ, masalan	ع	ʕ	ʕənd
ج	ʒ, dʒ, g	ʒamāl, dʒamāl, gamāl	غ	ɣ	ɣēm
ح	ħ	ħabīb	ف	f, v	ʔalf, narvaz
خ	x	ʔəxt	ق	q, g, ʔ	ħuqūq, grīb, ʔarīb
د	d	dār	ك	k, g	ktāb, garāʒ, sangari
ذ	ð, d, z	hāða, hāda, ʔəstēz	ل	l, ɻ	lēl, aɻɻāh
ر	r	rās	م	m	mabrūk
ز	z	zēt	ن	n	naʒāħ
س	s	səne	ه	h	hənnen
ش	ʃ	ʃams	و	w	walad
ص	ṣ	ṣēf	ي	j	jōm
ض	ḍ, ẓ	ḍamīr, maẓbūṭ	ء	ʔ	suʔāl

Des caractères arabes particuliers peuvent être utilisés pour la transcription en arabe des mots étrangers courants contenant les sons [p] ou [v].

Arabe	API	Exemple
پ	p	پويات (des pots)
ڤ	v	ساڤا (ça va)
گ	g	سنگري (plombier)

0. 3. 2. Voyelles

API	Exemple	API	Exemple
a	abjaḍ	ā	ktāb
ɛ	bɛrke	ē	lēʃ, bēt, bēred
e	ṭajjeb	ī	dīk
ə	sətt, təffēħa	ō	ʃōb
i	kīfik	ū	sūʔ
o	bjektob	ã	ã... ok (exclamation)
u	suʔāl	ɔ̃	bɔ̃ʒūr
y	stydjo	ø	ʔø (hésitation)

Le son [e] en finale de mot, lorsqu'il ne s'agit pas d'une *imāla* mais d'un terme étranger, peut être transcrit يه en arabe. Exemple : kafe / كافيه (café).

1. LA TIER 1 : UNE TRANSCRIPTION INTERACTIONNELLE

La tier 1 est une transcription phonético-phonologique en caractères latins (alphabet phonétique international adapté), dans laquelle les phénomènes oraux et interactionnels sont notés.

1. 1. Phénomènes interactionnels

Tous les éléments relatifs à la production de la langue parlée en interaction sont notés dans cette tier selon la convention détaillée en annexe 1. Les segments inaudibles sont rendus par xx. En cas de troncation, on transcrit phonétiquement ce que l'on entend. Les transcriptions incertaines sont notées entre parenthèses.

1. 2. Transcription des sons et représentation de la réalité sonore (granularité)

Les sons sont notés tels qu'ils sont perçus :
- le *qaf* est noté (q, ʔ ou g), selon sa réalisation.
- les interdentales (θ, ð), sont notées comme elles sont prononcées (s, d, t, z, θ, ð).
- les emphatiques (ṣ, ḍ, ṭ, ẓ) ne sont notées que si la consonne est réellement vélarisée. Sinon, on notera leur réalisation (éventuellement : s, d, t, z).
- le [l] emphatique dans *aḷḷāh* est noté avec un point.
- la *hamza* initiale n'est pas transcrite ; la *hamza* médiane et finale n'est transcrite que si on la prononce.
- l'*imāla* est notée par le son [e].

La transcription de la tier 1 ne prétend pas atteindre une granularité très fine pour la représentation des sons prononcés, notamment les voyelles (c'est pourquoi on parle de représentation phonético-phonologique plus que réellement phonétique). Néanmoins, il n'y a pas d'essai de standardisation des sons transcrits ; le transcripteur note simplement ce qu'il perçoit.

1. 3. Notation de la longueur

1. 3. 1. Longueur phonologique et longueur expressive

On distingue la longueur phonologique (distinctive) notée par un trait sur la voyelle (ā, ī, ē, ō, ū) de la longueur expressive notée par les deux points après la voyelle (ba:). La longueur expressive n'apparaît que dans la tier 1.

1. 3. 2. Les voyelles longues finales de la graphie arabe

Les voyelles longues en finale de mot dans la graphie arabe ne sont pas transcrites avec une voyelle longue, mais conformément à leur réalisation. Cela concerne notamment :
- les pronoms personnels
- le *alif maqṣura* ou le *alif ṭawīla* à la fin des verbes, des pronoms, des noms et des particules
- la préposition *fi*
- certains pronoms suffixes.

1. 3 3. Les alif-s suscrits de l'orthographe arabe

On note la voyelle longue uniquement si elle est prononcée comme telle. Cela concerne certains démonstratifs, la particule *lākin* et le mot *allāh*. Le nom Allah est transcrit comme il est prononcé.
Ex. *aḷḷa / aḷḷah / aḷḷā / aḷḷāh*

1. 4. Les déterminants

1. 4. 1. L'article

L'article est noté comme il est prononcé et sans séparateur.
Ex. *lbēt*

1. 4. 2. Les démonstratifs

Le démonstratif réduit est noté [ha], accolé à l'article sans séparateur.
Ex. *halbēt*

Les démonstratifs complets sont notés tels qu'ils sont réalisés.
Ex. *hajda / haza ; hajde / hajdi / hazihi ; hadōl / hadōle / hajdōle ; hdāk / hadīke ; hāj* (liste non exhaustive).

1. 5. Les pronoms

1. 5. 1. Les pronoms personnels

Les pronoms sont transcrits selon leur réalisation. Les pronoms suffixes sont transcrits tels qu'ils se prononcent et sont accolés au verbe, au nom ou à la préposition.
Les pronoms compléments introduits par la particule [ijjā] sont notés conformément à leur réalisation, collés à la particule, cette dernière étant séparée de ce qui précède par une espace. Exemple de transcription des pronoms (non exhaustif) :

Pronoms isolés		**Pronoms possessifs**		**Pronoms compléments**	
ana,ani enta ente huwwe	hijje neħna,ʔēħna ento henne,hennen	ktēbe ktēbek,ktēbak ktēbek,ktēbik ktēbo	ktēba ktēbna ktēbkon ktēbon	jāni,jāne jāk,ijjāk, jēk jāki,ijjāki, jēke jē, ijjē, jāho	jē,ijjē, jāha, jēha jāna, jēna jekon jon,jāhon

Ex. *bejton, ʔāletlo, fīki, beddik jēha*

1. 5. 2. Le pronom relatif

Les relatifs sont notés au plus proche de la perception et sont séparés du mot suivant par une espace.
Ex. *jilli / əlli / li / l*

1. 5. 3. Les pronoms interrogatifs

Les interrogatifs sont notés selon leur prononciation.
Ex. *kīf, wejn, wēn, fēn, lē, lēʃ, ʃū*

1. 5. 4. Les pronoms démonstratifs

Les pronoms démonstratifs sont notés tels qu'ils sont réalisés.
Ex. *hēj*, ; *hajda ; hajde / hajdi ; hadōl / hadōle / hajdōle ; haydēk / hadīke* (liste non exhaustive).

1. 6. Les particules et prépositions

Les particules sont notées selon leur prononciation. Elles sont séparées du nom par une espace, même si elles sont accolées dans la graphie arabe. Voir la liste des prépositions dans l'annexe 2.
Ex. *bi lbēt*

1. 7. Les chiffres

Les chiffres sont transcrits phonétiquement.

1. 8. Autres éléments conventionnels

Certains éléments qui sont transcrits phonétiquement en tier 1 feront l'objet de notations conventionnelles dans les autres tiers. C'est le cas du lexique dialectal courant sans équivalent en arabe standard.

1. 8. 1. Les expressions figées

Les éléments à traduction variée (comme *jaʕni*) sont transcrits en tier 1 mais font l'objet d'une notation conventionnelle dans les autres tiers. On trouvera en annexe 2 une liste des notations conventionnelles.

1. 8. 2. Les mots ou syntagmes cités en langue étrangère

Les productions en langue étrangère sont délimitées par des dièses afin de permettre leur repérage rapide et systématique. Ils sont notés, sans mention de langue, dans l'orthographe de la langue concernée pour les segments (courts ou longs) dont la prononciation respecte le standard de la langue concernée, ou en transcription phonétique si l'on considère que cela apporte des informations pertinentes ou si la langue n'est pas identifiable.

On utilise, en cas de besoin, l'orthographe adaptée décrite dans la convention ICOR, qui fait intervenir l'antiquote ` pour les élisions non standard (ex. je l` fais). Voir Annexe 1.

1. 8. 3. Les mots étrangers entrés dans les usages

Les mots étrangers entrés dans les usages sont transcrits phonétiquement, sans dièses.
Ex. *b̥āj*

1. 8. 4. Les emprunts arabisés

Les emprunts arabisés sont transcrits phonétiquement et notés entre dièses.

1. 8. 5. Les régulateurs, marques d'hésitation, acquiescement, négation

La notation des régulateurs, de l'hésitation, de l'acquiescement ou de la négation fait l'objet de conventions. Voir Annexe 2.

1. 8. 6. Les noms propres

Qu'ils soient remplacés par des pseudonymes ou non, les noms propres sont transcrits dans leur orthographe courante si elle est connue ou transcrits phonétiquement selon les cas et notés entre guillemets hauts, avec majuscule initiale. Il peut s'agir d'une personne, d'un lieu, d'une institution, etc.

1. 8. 7. La mention de noms anonymisés

Les noms propres anonymisés qui sont cités dans les interactions apparaissent entre guillemets hauts, en entier ou sous leur forme abrégée selon les cas. Cela peut concerner une personne, un lieu, une institution, etc. S'il s'agit du pseudonyme de l'un des participants à l'interaction, on le conserve sans guillemets.

2. LA TIER 2 : UNE TRANSCRIPTION MORPHO-PHONOLOGIQUE

La tier 2a est une transcription morpho-phonologique d'un standard du dialecte concerné. Elle est notée en caractères latins (alphabet phonétique international adapté) et intègre de nombreuses notations conventionnelles. Elle est assortie, en tier 2b, d'une glose inspirée des Leipzig Glossing Rules (voir l'annexe 4 et le texte intégral [En ligne] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf). Elle comporte de nombreuses notations conventionnelles détaillées en annexe 2. Les choix de notations conventionnelles respectent le plus souvent possible la convention de notation des dialectes arabes CODA et se réfèrent à l'ouvrage *Parlons arabe libanais* de Fida Bizri (L'Harmattan, 2010).

2. 1. Phénomènes interactionnels

Les phénomènes interactionnels ne sont pas notés, à l'exception des pauses notées par (.) quelle que soit leur durée et des chevauchements marqués par [.
Les segments inaudibles sont rendus par xx.
En cas de troncation, on ajoute un tiret à la fin du mot inachevé.
Les transcriptions incertaines sont notées entre parenthèses.

2. 2. Transcription des sons

La tier 2 est un équivalent en caractères latins de la tier 3, elle-même établie en suivant la convention de notation des dialectes arabes CODA (voir l'annexe 3 et la présentation de la tier 3 ci-dessous). Cette tier repose sur de nombreuses notations conventionnelles, puisqu'elle correspond, dès que possible, à une translittération du *ductus* de l'arabe standard, tout en préservant de manière conventionnelle les caractéristiques du dialecte dans lequel l'interaction a lieu.
Les consonnes sont transcrites conformément au *ductus* standard et les voyelles sont transcrites phonologiquement. De ce fait, dès que le mot dialectal a son équivalent en arabe standard, on le translittère selon l'orthographe classique, et lorsque ce n'est pas le cas, on le transcrit phonologiquement selon une notation conventionnelle renvoyant à un un standard dialectal. Les notations conventionnelles concernent notamment les pronoms personnels, les particules, les interrogatifs, les démonstratifs, les désinences verbales, etc. (voir l'annexe 2).

Voici les règles générales de transcription des sons :
- Le *qaf* est noté q.
- Les interdentales seront notées θ ou ð.
- Les emphatiques ṣ, ḍ, ṭ, ẓ.
- La *hamza* est notée ʔ en médiane et finale de mot uniquement, et non notée à l'initiale.
- Les sons vocaliques sont réduits aux sons [a] [u] [i] [e] [o] pour les voyelles brèves, à [ā] [ū] [ī] [ē] [ō] pour les longues.
- L'*imāla* est transcrite par le son [e].
- Les diphtongues sont rétablies.
- Les voyelles élidées sont rétablies.
- Les voyelles épenthétiques sont supprimées.

2. 3. Notation de la longueur

2. 3. 1. Longueur phonologique et longueur expressive

Seule la longueur phonologique est notée en tier 2.

2. 3. 2. Les voyelles longues finales de la graphie arabe

Les voyelles longues finales de l'orthographe arabe standard, même si elles ne sont pas prononcées, sont notées dans la tier 2 qui privilégie une translittération de l'orthographe arabe standard.
Ex. *fī ; ilā ; ʕalā ; maqhā*

2. 3. 3. Les alif*-s suscrits de l'orthographe arabe*

On note systématiquement une voyelle longue si celle-ci correspond à un *alif* suscrit dans l'orthographe standard. Cela concerne certains démonstratifs, la particule *lākin* et le mot *allāh*.
Ex. *lākin, hāðā, allāh*

2. 4. Les déterminants

2. 4. 1. L'article

L'article est segmenté au moyen d'un tiret. En début d'énoncé, il est systématiquement noté [el-]. En cas de liaison, il est toujours noté [-l-] même en cas d'assimilation.
Ex. *el-bēt ; bi-l-bēt ; bi-l-rās*

2. 4. 2. Les démonstratifs

Le démonstratif *ha* est noté avec séparateur (ha-l-).
Ex. *ha-l-kitāb ; bi-ha-l-bajt*

Les démonstratifs complets font l'objet d'une notation conventionnelle.
Ex. *hāðā ; hāðihi ; haðūl ; haðāk ; haðīk*

2. 5. Les pronoms

2. 5. 1. Les pronoms personnels

Les pronoms suffixes sont segmentés par un tiret.
Les pronoms isolés et suffixes apparaissent sous une forme dialectale conventionnelle.
Les pronoms compléments introduits par la particule [ijjā] seront notés de manière conventionnelle sous la forme *jē*-pronom conventionnel.

Pronoms isolés		**Pronoms possessifs**		**Pronoms compléments**	
anā	hija	kitāb**-ī**	kitāb**-hā**	jē**-nī**	jē**-hā**
enta	neħnā	kitāb**-ka**	kitāb**-nā**	jē**-ka**	jē**-nā**
enti	entō	kitāb**-ki**	kitāb**-kon**	jē**-ki**	jē**-kon**
huwa	hennen	kitāb**-hu**	kitāb**-hon**	jē**-hu**	jē**-hon**

2. 5. 2. Le pronom relatif

Le relatif est systématiquement noté *jalli*, séparé du mot suivant par une espace, quelle que soit sa réalisation.

2. 5. 3. Les pronoms interrogatifs

Les interrogatifs font l'objet d'une transcription conventionnelle, quelle que soit leur réalisation. Voir Annexe 2.

2. 5. 4. Les pronoms démonstratifs

Les pronoms démonstratifs font l'objet d'une notation conventionnelle.
Ex. *hāðā ; hāðī ; haðūl ; haðāk ; haðīk*

2. 6. Les particules et prépositions

Un tiret sépare les prépositions monolitères du mot qui suit. Elles sont notées conventionnellement (*la-, bi-, wa-, ta-*) ; les autres prépositions sont séparées du nom par une simple espace. Voir la liste des prépositions dans l'annexe 2.

2. 7. Les chiffres

Les chiffres sont transcrits conformément à leur équivalent en arabe standard.
Ex. arbaʔat ʕaʃar

2. 8. La morphologie verbale

Le radical du verbe est isolé de l'ensemble des affixes par un tiret. On segmente aussi les morphèmes de modes/temps et personnes qui font l'objet d'une notation conventionnelle (voir annexe 2).
Ex. *b-ja-ktub-ū*

Les préfixes de conjugaison comme *ʕam* et *raħ* sont séparés du verbe par une espace.
Ex. *ʕam ja-ktub ; raħ ta-ktub*

Le morphème *ħa-* est segmenté par un tiret du préfixe de conjugaison.
Ex. *ħa-ja-ktub*

En cas d'élision du morphème de conjugaison en tier 1, on rétablit en tier 2 le morphème manquant.
Ex. *ʕam fakkir* (tier 1) = *ʕam u-fakkir* (tier 2a)

De même, pour l'inaccompli à valeur modale (sans le préverbe b-) employé après différentes particules comme حتى، لازم، بدّ، في, etc., la *hamza* de la première personne du singulier est rétablie dans la tier 2 (badd-ī aħʒuz).

La voyelle du radical et celle des personnes pour les formes dérivées sont notées conformément à la voyelle présente dans le verbe correspondant en standard (voir annexe 2).
Ex. *b-ja-ʕmal ; b-ju-balliʃ*

Pour le détail des conventions concernant la transcription des voyelles de conjugaison et des verbes conjugués à l'accompli, à l'inaccompli et à l'impératif, voir l'annexe 2.

2. 9. La morphologie nominale

Les marques du genre et du nombre des noms sont segmentées par un tiret.
Ex. *ʕāmil ; ʕāml-e ; ʕāml-īn*

2. 10. Autres éléments conventionnels

Certains éléments qui sont transcrits phonétiquement en tier 1 font l'objet de notations conventionnelles dans les autres tiers. Voir Annexe 2.

2. 10. 1. Le lexique dialectal

Les productions typiquement dialectales font l'objet de notation conventionnelle.

2. 10. 2. Les expressions figées

Pour les éléments très récurrents à l'oral à traduction variée (comme les particules discursives, par exemple *jaʕni*) ou les expressions figées d'inspiration coranique ou non, on utilise une notation conventionnelle composée de quatre lettres capitales (voir annexe 2).
Ex : *b-ʃūf-ki* INCH

2. 10. 3. Les mots ou syntagmes en langue étrangère

Toutes les productions en langue étrangère sont notées entre dièses afin de permettre leur repérage rapide et systématique. Elles sont notées sans mention de la langue, dans l'orthographe de la langue concernée, que les segments soient courts ou longs, ou en transcription phonétique si l'on ignore la langue en question.
Ex. ENNO Nicolas #hair# badd-hu #bonnet#

2. 10. 4. Les mots étrangers entrés dans les usages

Les mots étrangers entrés dans les usages, dont on fournit une liste non exhaustive en annexe 2, seront notés entre dièses, dans l'orthographe de la langue et sans mention de la langue.
Ex. #bonjour# ; #sorry# ; #merci#

2. 10. 5. Les emprunts arabisés

Les emprunts arabisés, dont on fournit une liste non exhaustive en annexe 2, seront notés entre dièses en transcription et sans mention de la langue.

Ex. #pojēt# ; #taperwarāt#

2. 10. 6. Les régulateurs, marques d'hésitation, acquiescement, négation

La notation des régulateurs, de l'hésitation, de l'acquiescement ou de la négation font l'objet de conventions, présentées en annexe 2.

2. 10. 7. Les noms propres

Qu'ils soient remplacés par des pseudonymes ou non, les noms propres sont transcrits dans leur orthographe courante si elle est connue et notés entre guillemets hauts, avec majuscule initiale. Il peut s'agir d'une personne, d'un lieu, d'une institution, etc.

2. 10. 8. La mention de noms anonymisés

Les noms propres anonymisés qui sont cités dans les interactions apparaissent entre guillemets hauts, en entier (avec une majuscule) ou sous leur forme abrégée selon les cas. Cela peut concerner une personne, un lieu, une institution, etc. S'il s'agit du pseudonyme de l'un des participants à l'interaction, on le conserve sans guillemets.

3. LA TIER 3 : UNE TRANSCRIPTION EN GRAPHIE ARABE

La tier 3 propose une transcription en caractères arabes. Elle reprend la convention CODA pour la transcription des dialectes, avec quelques aménagements (voir l'annexe 3). Selon cette convention, on se réfère, chaque fois que cela est possible, au *ductus* de l'arabe standard. En même temps, on conserve les traits spécifiques du dialecte utilisé, en faisant émerger un standard dialectal par le biais de différentes notations conventionnelles.

3. 1. Phénomènes interactionnels

Les phénomènes interactionnels ne sont pas notés, à l'exception des pauses notées par (.) quelle que soit leur durée, et des chevauchements marqués par]. Les segments inaudibles sont rendus par xx. En cas de troncation, on ajoute un tiret à la fin du mot inachevé. Les transcriptions incertaines sont notées entre parenthèses. Si le nom des intervenants est cité dans un énoncé, il figure en caractères latins en trois lettres majuscules, selon le code donné par le transcripteur.

3. 2. Transcription des sons

La graphie standard est rétablie dès que cela est possible : on réintègre toutes les consonnes selon l'orthographe classique.
- Le *qaf* est noté ق, quelle que soit sa réalisation.
- Les interdentales sont notées ث et ذ, quelle que soit leur réalisation.
- Les emphatiques sont notées ص, ض, ط, ظ.
- La *hamza* initiale n'est notée que si elle correspond à une *hamza* stable.
- Les géminations morphologiques sont systématiquement notées par une *chadda*.
- La transcription n'est pas vocalisée (ni voyelles brèves, ni *alif* suscrit).
- Aucun signe de lecture n'est ajouté pour l'article (*hamza*, *waṣla, sukūn* et *chadda*).
- Le *tanwīn* [an] est noté.

Pour les mots qui n'ont pas de graphie standard et certains termes dialectaux courants, on se réfère à une notation conventionnelle, présentée dans l'annexe 2.

3. 3. Les voyelles longues finales de la graphie arabe

Les voyelles longues finales figurent conformément au *ductus* classique. Cela concerne notamment :
- les pronoms personnels dans leur notation conventionnelle (أنا)
- le *alif maqṣura* ou le *alif ṭawīla* à la fin des verbes, des pronoms, des noms et des particules (علٰى)
- certains pronoms suffixes (ها).

3. 4. Les déterminants

3. 4. 1. L'article

L'article est toujours noté الـ, sans aucun signe de lecture, conformément à la graphie standard.

3. 4. 2. Les démonstratifs

Le démonstratif *ha-* est noté accolé à l'article هالـ.
Les autres démonstratifs font l'objet d'une notation conventionnelle, quelle que soit leur réalisation. Voir l'annexe 2.

3. 5. Les pronoms

3. 5. 1. Les pronoms personnels

Les pronoms sont notés de manière conventionnelle.

Pronoms compléments		Pronoms possessifs		Pronoms isolés	
إياها	إياني	كتابها	كتابي	هي	أنا
إيانا	إياك	كتابنا	كتابك	نحنا	إنت
إياكن	إياكي	كتابكن	كتابكي	إنتوا	إنتي
إياهن	إياه	كتابهن	كتابه	هنن	هو

3. 5. 2. Le pronom relatif

Le relatif sera systématiquement noté اللي.

3. 5. 3. Les pronoms interrogatifs

Les interrogatifs font l'objet d'une transcription conventionnelle, quelle que soit leur réalisation :

كيف، وين، فين، مين، ليه، ليش، إيش، شو، قديش، قديه، إيمتى، إمتى...

Voir l'annexe 2 pour une liste complète.

3. 5. 4. Les pronoms démonstratifs

Les pronoms démonstratifs font l'objet d'une notation conventionnelle, quelle que soit leur réalisation. Voir l'annexe 2.

هذا، هذي، هذول، هذاك، هذيك

3. 6. Particules et prépositions

Les particules et prépositions monolitères sont attachées conformément à l'orthographe arabe.
Les particules tronquées ne sont pas rétablies et sont traitées comme des monolitères.
Ex. : عالبيت ne sera pas noté على البيت

3. 7. Les chiffres

Les chiffres sont transcrits conformément à leur équivalent en arabe standard.
Ex. أربعة عشر

3. 8. La morphologie verbale

La transcription des verbes conserve les marqueurs modaux du dialecte (*b-*, *ʕam*). En cas d'élision du morphème de conjugaison en tier 1 (*ʕam fakkir*), le morphème manquant est rétabli (عم أفكّر). De même, pour l'inaccompli à valeur modale (sans le préverbe b-) employé après différentes particules comme حتى ،لازم ،بدّ ،في , etc., la *hamza* de la première personne du singulier est rétablie dans la tier 3 (بدّي أحجز).

Les tableaux complets de conjugaison à l'accompli, à l'inaccompli et à l'impératif figurent dans l'annexe 2.

3. 9. Autres éléments conventionnels

Des conventions orthographiques sont introduites pour le vocabulaire dialectal ainsi que pour les emprunts (Voir annexe 2).

3. 9. 1. Les expressions figées

Les éléments à traduction variée ou les expressions figées d'inspiration coranique ou non, sont écrits en toutes lettres au plus près de l'orthographe standard, selon une notation conventionnelle qu'on s'attachera à respecter pour faciliter les recherches automatiques (Annexe 2).

3. 9. 2. Les mots ou syntagmes en langue étrangère

Les productions en langue étrangère sont notées avec les caractères et dans l'orthographe de la langue concernée.
Ex. #bonnet# بدّه #hair# إنّه نقولاس

3. 9. 3. Les mots étrangers entrés dans les usages

Les mots étrangers entrés dans les usages seront transcrits en caractères arabes, entre dièses, sans menton de la langue.
Ex. #بنجور# ، #سوري# ، #مرسي# ، #فا سا#

3. 9. 4. Les emprunts arabisés

Les emprunts arabisés seront transcrits en caractères arabes, entre dièses, sans menton de la langue.
Ex. #تيبيروارات# ، #پويات#

3. 9. 5. Les régulateurs, marques d'hésitation, acquiescement, négation

La notation des régulateurs, de l'hésitation, de l'acquiescement ou de la négation font aussi l'objet de conventions. Voir Annexe 2.

3. 9. 6. Les noms propres

Qu'ils soient remplacés par des pseudonymes ou non, les noms propres sont transcrits dans leur orthographe courante si elle est connue et notés entre guillemets hauts, en caractères arabes. Il peut s'agir d'une personne, d'un lieu, d'une institution, etc.

3. 9. 7. La mention de noms anonymisés

Les noms propres anonymisés cités apparaissent en entier ou sous leur forme abrégée selon les cas, entre guillemets hauts et en caractères arabes. Cela peut concerner une personne, un lieu, une institution, etc. S'il s'agit du pseudo de l'un des participants à l'interaction, on le conserve en caractères latins et sans guillemets.

4. LA TIER 4 : UNE TRADUCTION

La tier de traduction offre un accès direct au sens. Les choix de traduction doivent être guidés par deux principes : l'intelligibilité et la conformité à l'original. On ne fera donc pas du mot-à-mot et on veillera à ce que le rendu soit compréhensible en français, tout en s'assurant de rester au plus proche de la structure arabe originale. Par ailleurs, aucun signe de ponctuation n'est ajouté.

4. 1. Phénomènes interactionnels

Les phénomènes interactionnels ne sont pas indiqués, à l'exception des pauses notées par (.) quelle que soit leur durée, et des chevauchements marqués par [. Les segments inaudibles sont rendus par xx. En cas de troncation, on ajoute un tiret à la fin du mot inachevé. Les éléments incertains dans la transcription sont traduits entre parenthèses. Comme dans les autres tiers, si le nom des intervenants est cité dans un énoncé, il figure en caractères latins en trois lettres majuscules, selon le code donné par le transcripteur.

4. 2. Vouvoiement et ajustements de traduction

Les énoncés où le vouvoiement serait d'usage en français sont traduits en utilisant "vous", plutôt que "tu".
Les ajouts aidant à comprendre le sens de la traduction seront mentionnés entre {}.

4. 3. Expressions figées

Les éléments à traduction variée (comme *jaʕni*) ou les expressions figées d'inspiration coranique ou non, sont notés conformément à la notation conventionnelle établie utilisée en tier 2, composée de quatre lettres capitales (voir la liste en annexe 2).
Ex : LAYK je passe te voir demain INCH

4. 4. Mots ou syntagmes en langue étrangère

Les mots prononcés dans une langue autre que l'arabe sont traduits en français, mais figurent entre # pour marquer l'emprunt, la langue utilisée étant indiquée entre double parenthèses à l'intérieur des dièses en se référant au code ISO. De même, les réalisations en arabe classique (CLA) ou standard moderne (MSA) figurent entre dièses avec mention de la langue entre double parenthèses.
Ex : #merci ((FR))#

4. 5. Régulateurs, marques d'hésitation, acquiescement, négation

La notation des régulateurs, de l'hésitation, de l'acquiescement ou de la négation fait l'objet de conventions (voir l'annexe 2).

4. 6. Les noms propres

Qu'ils soient remplacés par des pseudonymes ou non, les noms propres sont notés dans leur orthographe courante si elle est connue, sans guillemets mais avec majuscules initiales.

5. LISTE DES DOCUMENTS ANNEXES

Annexe 1 : Convention de transcription pour les phénomènes oraux et interactionnels

Annexe 2 : Liste des notations conventionnelles

Annexe 3 : Liste des gloses morpho-syntaxiques

Annexe 4 : Règles générales de transcription des dialectes en caractères arabes

Annexe 5 : Tableau synoptique de la convention AraPI

6. BIBLIOGRAPHIE

BICKEL, B., COMRIE, B., HASPELMATH, M. (2015). *Leipzig Glossing Rules* [En ligne] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

BIZRI, F. (2010). *Parlons l'arabe libanais*, Paris, L'Harmattan.

BLANC, H. (1960). "Stylistic Variations in Spoken Arabic: A Sample of Interdialectal Educated Conversation", in C. A. FERGUSON (ed.), *Contributions to Arabic Linguistics*, Cambridge, Massachusset, Harvard Middle Eastern Monographs, p. 81-161.

CAUBET, D. (1999). « Arabe maghrébin : passage à l'écrit et institutions », *Faits de Langue* 13, p. 235-244.

DIAB, M., HABASH, N. (2014). *Natural Language Processing of Arabic and its Dialects* [En ligne] http://emnlp2014.org/tutorials/4_notes.pdf.

DICHY, J. (1994). « La pluriglossie de l'arabe », *Bulletin d'études orientales* 46, Damas, p. 19-42.

HABASH, N., SOUDI, A., BUCKWALTER, T. (2007). "On Arabic Transliteration", in SOUDI A., BOSCH A., NEUMANN G. (eds). *Arabic Computational Morphology. Text, Speech and Language Technology*, 38. Springer, Dordrecht.

HABASH, N., DIAB, M., RAMBOW, O. (2012). *Conventional orthography for dialectal Arabic* [En ligne] http://www.lrec-conf.org/proceedings/lrec2012/pdf/579_Paper.pdf.

HABASH, F., et al. (2018). *Unified Guidelines and Resources for Arabic Dialect Orthography* [En ligne] http://www.lrec-conf.org/proceedings/lrec2018/pdf/395.pdf.

ICOR (2013). *Convention ICOR* [En ligne] http://icar.cnrs.fr/projets/corinte/documents/2013_Conv_ICOR_250313.pdf

IZRE'EL, S., METTOUCHI, A. (2015). "Representation of speech in Corpafroas. Transcriptional strategies and prosodic units", in METTOUCHI A., VANHOVE M., CAUBET D. (dir.), *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages*, John Benjamins Publishing Company, p. 14-41.

JEFFERSON, G. (2004). "Glossary of transcript symbols with an introduction", in LERNER G. (dir.), *Conversation Analysis: Studies from the First Generation*, Amsterdam, John Benjamins, p. 13-31.

MAAMOURI, M, GRAFF, D., JIN, H., CIERI, C. (2004a). *Dialectal Arabic Orthography-based Transcription & CTS Levantine Arabic Collection*. EARS PI Meeting and RT-04 Workshop, IBM Executive Conference Center, Palisades, NY, USA; November 7-11, 2004 [En ligne] http:// www.sainc.com/richtrans2004/

MAAMOURI, M, GRAFF, D., JIN, H., CIERI, C. (2004b). "Dialectal Arabic Telephone Speech Corpus: Principles, Tool design and Transcription Convention", in Proceedings of the NEMLAR Arabic Language Resources and Tools Conference Sept. 22-23, 2004. Cairo, Egypt, p. 55-60 [En ligne]

https://www.researchgate.net/publication/228876943_Dialectal_arabic_telephone_speech_corpus_Principles_tool_design_and_transcription_conventions

METTOUCHI, A., VANHOVE, M., CAUBET, D. (dir.) (2015). *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages*. John Benjamins Publishing Company.

METTOUCHI, A., CHANARD, C. (2010). "From Fieldwork to Annotated Corpora: the CorpAfroAs Project", *Faits de Langue-Les Cahiers* 2, p. 255-265.

TRAVERSO, V. (2006). *Des échanges ordinaires à Damas*. Lyon/Damas: PUL/Presses de l'IFPO.

VAILLANT P., LEGLISE I. (2014). « À la croisée des langues. Annotation et fouille de corpus plurilingues », *Revue des nouvelles technologies de l'information*, Cépaduès-Editions, RNTI-SHS-2, p. 81-100.

ZAGHOUANI W. (2014). "Critical Survey of the Freely Available Arabic Corpora", Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools [En ligne] https://arxiv.org/ftp/arxiv/papers/1702/1702.07835.pdf

ZAWAYDEH Bushra et al. "Guidelines for Transcribing Arabic Dialects" [En ligne] https://catalog.ldc.upenn.edu/docs/LDC2005S08/BBN-Babylon-transcription-guidelines.pdf

ZRIBI I. , GRAJA, M., ELLOUZE, M., JAOUA, M., HADRICH BELGUITH, L., (2013). "Orthographic Transcription for Spoken Tunisian Arabic", CICLing 1, p. 153-163.

ZRIBI I., BOUJELBANE R., MASMOUDI A., ELLOUZE M., HADRICH BELGUITH L., HABASH N. (2014). "A Conventional Orthography for Tunisian Arabic", LREC 2014, p. 2355-2361.

ZRIBI I., ELLOUZE M., HADRICH BELGUITH L., BLACHE P. (2015). "Spoken Tunisian Arabic Corpus "STAC": Transcription and Annotation", *Research in Computing Science* 90, p. 123-135.

Annexe 1 : convention de transcription pour les phénomènes oraux et interactionnels

Phénomène	Conventions	Exemples
Identité du participant		
Participant identifié	Toute parole est rapportée au locuteur qui l'a produite, qui est indiqué en début de ligne généralement par son identifiant, qui est le diminutif de son pseudonyme (utilisé pour des raisons éthiques et juridiques). L'identifiant est composé d'un, de deux ou de trois caractères (en majuscules). Dans les documents word, il est suivi d'une tabulation. Attention: c'est le seul endroit dans une transcription où l'usage de la tabulation est admis.	COR hāj
Doutes sur l'identité du participant	Lorsque le locuteur n'est pas identifiable, on utilise des X à la place du pseudo, en spécifiant F(emme) ou H(omme) si possible et nécessaire.	XXX hāj XXF bāj XXH salēm
Tour		
Notation du tour	La notation du tour est toujours rapportée à un identifiant. Un tour peut s'étendre sur plusieurs lignes. Le paragraphe d'un tour est aligné à gauche.	COR u baʕdēn ħanballeʃ duɣri duɣri #direct# waɬɬa
Enchaînement immédiat	= indique deux tours de parole qui s'enchaînent immédiatement sans micro-pause entre eux (*latching*).	COR kīfek/= SAR =kīfak anta
Chevauchement	Les crochets [et] encadrent le chevauchement dans chacun des tours concernés.	SAR hāj lī[na] COR [ahlē:n]
Silences et pauses	Les pauses d'une durée inférieure à 0,2 secondes sont notées (.) (micro-pauses) Au-dessus, elles sont chronométrées à l'aide d'un logiciel au 10ième de seconde près.	(.) (0.7)
Production de la parole		
Relation phonie-graphie	Pour les productions en arabe, on utilise une transcription phonético-phonologique (voir la convention ARAPI). Pour les productions dans d'autre langues, on utilise une transcription orthographique	SAM ahlēn ʃu xbārik SAR monsieur
Elision non standard	En cas de besoin, pour les langues transcrites en orthographe, on peut recourir à l'élision non standard, notée par une antiquote ` (suivie d'un espace en fin de mot ; sans espace à l'intérieur du mot). Cette notation distinguant l'élision standard de l'élision non standard est importante pour le fonctionnement des outils de recherche automatique dans les données.	SAR c'est d`jà ça  COR j'ai compris (.) j` recommmence
Segments inaudibles	Les segments incompréhensibles sont représentés au moyen d'une série de x, chacun valant une syllabe. Si le nombre de syllabes ne peut être perçu, on note (inaud.) Si le transcripteur hésite entre plusieurs notations, les deux sont notées entre parenthèses, séparées par des points virgules.	COR xxx COR (inaud.) SAR (bʔūl; btəʔūl)
Allongement	Le son allongé est noté par : (sans espace avant). Les : sont répétés en fonction de la durée perçue de l'allongement.	COR wē:nek/ SAR hō:::ne ma ʃeftini

Troncation	L'abandon d'un mot en cours de production est noté par - (sans espace avant, un espace après, ce qui le distingue du trait d'union).	COR ajm- ajmta beddek teʒi
Aspiration	L'aspiration est notée par .h (h précédé d'un point sans espace).	COR .h: eh:: lima la
Expiration, soupir	L'expiration est notée par h	COR h:: laʔa akīd lā
Prosodie		
Montée et chute intonative	Les montées et chutes intonatives sont notées par « / » et « \ » sans espace avant. Les fortes montées et chutes sont notées par « // » et « \\ .	SAM xallaṣte// LEI lā\ lissa
Saillance perceptuelle	Les segments perçus comme saillants sont soulignés.	SAM waḷḷa ma baʕref
Production vocale		
Production vocale	Les caractéristiques vocales et du mode de production sont décrites entre doubles parenthèses et précèdent la transcription, l'ensemble est compris entre chevrons :	COR <((en riant)) ma fī:ni>
Actions et événements. Cela concerne tous les phénomènes décrits et non transcrits, c'est-à-dire qui ne correspondent pas à des paroles produites par les participants.		
Action ou événement attribuable	Description entre doubles parenthèses dans le "pavé" du tour, après le pseudo du locuteur en début de ligne	CAR ((tape un numéro de téléphone))
Action ou événement	Même notation entre doubles parenthèses, mais sans identifiant dans la colonne des pseudos.	((le tonnerre retentit))

Annexe 2 : Liste des notations conventionnelles

0. Présentation des 4 tiers

La convention AraPI comprend 4 tiers :

Tier 1 : *transcription phonético-phonologique (API modifié)*
Tier 2a : *translittération morpho-phonologique (API modifié)*
Tier 2b : *glose morpho-syntaxique*
Tier 3 : *translittération arabe (avec caractères modifiés)*
Tier 4 : *traduction*

Dans la **tier 1**, on transcrit phonétiquement sans atteindre un degré de granularité très fin. Dans les tableaux qui suivent, les exemples donnés pour la tier 1 ne représentent pas toutes les réalisations possibles.
La **tier 3** est une notation conventionnelle du dialecte en caractères arabes : les traits morphologiques spécifiques au dialecte sont conservés, mais le *ductus* des mots se rapproche dès que possible de l'arabe littéral standard. De nombreux éléments courants, tels que les pronoms, les interrogatifs, etc. font l'objet de notations conventionnelles. La **tier 2a** est la translittération de la tier 3 en API.
La **tier 2b** est une glose (cf. annexe 4 pour une liste des gloses et des exemples).
La **tier 4** est une traduction dans la langue du chercheur, qui rend le sens de l'énoncé tout en s'attachant à rester au plus près de l'original.

1. Transcription des sons

1. 1. Consonnes

Les consonnes sont présentées dans l'ordre alphabétique de l'arabe.

Tier 3	Tier 1	Tier 2a	Tier 3	Tier 1	Tier 2a
ء	ʔ	ʔ	ط	ṭ	ṭ
ب	b, p	b	ظ	ð̣, ḍ, ẓ	ð̣
ت	t	t	ع	ʕ	ʕ
ث	θ, t, s	θ	غ	ɣ	ɣ
پ	p	p	ف	f, v	f
ج	ʒ, dʒ, g	ʒ	ڤ	v	v
ح	ħ	ħ	ق	q, g, ʔ	q
خ	x	x	ك	k, g	k
د	d	d	گ	g	g
ذ	ð, d, z	ð	ل	l, ɫ	l
ر	r	r	م	m	m
ز	z	z	ن	n	n
س	s	s	ه	h	h
ش	ʃ	ʃ	و	w	w
ص	ṣ	ṣ	ي	j	j
ض	ḍ, ẓ	ẓ			

1. 2. Voyelles brèves

Il n'y a aucune vocalisation en tier 3. Pour les termes étrangers se terminant par le son [e], on peut le transcrire يه en arabe.

Tier 1	Tier 2a
a, ə	a
ɛ, e, ə	e
i, ə	i
o, ə	o
u, y	u

1. 3. Voyelles longues

Tier 3	Tier 1	Tier 2a
ا	ā	ā
ا-ي	ē	ē
ي	ī	ī
و	ō	ū
و	ū	ū
ʔø	ʔø	ʔø (hésitation)
	ã	hã (marqueur exclamatif)
		bɔ̃ʒūr

API	Exemple	API	Exemple
a	abjaḍ	ā	ktāb
ɛ	bɛrke	ē	lēʃ, bēt, bēred
e	ṭajjeb	ī	dīk
ə	sətt, təffēħa	ō	ʃōb
i	kīfik	ū	sūʔ
o	bjektob	ã	ã... ok (exclamation)
u	suʔāl	ɔ̃	bɔ̃ʒūr
y	stydjo	ø	ʔø (hésitation)

1. 4. Exemples

Tier 1	Tier 2a	Tier 3
zɣīr	ṣaɣīr	صغير
rfīʔa	rafīqa	رفيقة
tlēt	θalāθ	ثلاث
masalan	maθalan	مثلا
iza	iðā	إذا
bi ẓẓabəṭ	bi-l-ḍabṭ	بالضبط
mazbūṭ	maḍbūṭ	مضبوط
ẓāhra	ðāhira	ظاهرة
ḍall	ðalla	ظلّ
aḷḷā, aḷḷāh	allāh	الله
bēt	bajt	بيت
jōm	jawm	يوم
ktēb	kitāb	كتاب
sijēse	sijāse	سياسة
baḥər	baḥr	بحر
azraʔ	azraq	أزرق
rās	raʔs	رأس
samā	samāʔ	سماء

2. Longueur phonétique et expressive

Tier 1	Tier 2a	Tier 3
bʕī::d	baʕīd	بعيد
lē:ʃ	lēʃ	ليش
akī::d	akīd	أكيد

3. Voyelles longues en fin de mot

Tier 1	Tier 2a	Tier 3
ana	anā	أنا
ʕala	ʕalā	على
farʒīne	farʒī-nī	فرجيني
fi	fī	في

4. Transcription du *alif* suscrit

Tier 1	Tier 2a	Tier 3
aḷḷā, aḷḷāh	allāh	الله
lākin, lakin	lākin	لكن
hāza, haza	hāðā	هذا
halbēt	hāðā l-bajt	هذا البيت

5. Transcription de la *hamza*

Tier 1	Tier 2a	Tier 3
bīr	biʔr	بئر
rās	raʔs	رأس
raʔīs	raʔīs	رئيس
samā	samāʔ	سماء
tlēt ijjēm	θalāθat ajjām	ثلاثة أيّام
ħkīne	iħkī-nī	احكيني
esmi	ism-ī	اسمي
ʃū əsmak	ʃū ism-ak	شو اسمك
xajje	ax-ī	أخي

6. Les déterminants

6. 1. L'article

Tier 1	Tier 2a	Tier 3
à l'initiale		
elləʕbe	el-luʕba	االعبة
eddār	el-dār	الدار
eʃʃaməs	el-ʃams	الشمس
elbejt	el-bajt	البيت
en liaison		
bə lləʕbe	bi-l-luʕba	باللعبة
bi ddār	bi-l-dār	بالدار
taḥt eʃʃaməs	taḥt l-ʃams	تحت الشمس
zēt ellōn	ðāt l-lawn	ذات اللون
ʃakəl bēt elʒirān	ʃakl bajt l-ʒirān	شكل بيت الجيران
tannurt elbenet	tannurat l-bint	تنّورة البنت

6. 2. Les démonstratifs

Tier 1	Tier 2a	Tier 3

ha	ha	هـ
hajda, haza	hāðā	هذا
hajde, hajdi, hazihi	hāðī	هذي
hadōl, hadōle	haðūl	هذول
hdāk	haðāk	هذاك
hadīk, hadīke	haðīk	هذيك

Tier 1	Tier 2a	Tier 3
halktēb bi halbēt	ha-l-kitāb bi-ha-l-bajt	هالكتاب بهالبيت

7. Les pronoms

7. 1. Les pronoms personnels

7. 1. 1. Les pronoms isolés sujets

Tier 1	Tier 2a	Tier 3
ana, ani	anā	أنا
enta	enta	إنت
ente	enti	إنتي
huwwe	huwa	هو
hijje	hija	هي
neħna, ʔēħna	naħnā	نحنا
ento	entō	إنتوا
henne, hennen	hennen	هنن

7. 1. 2. Les pronoms suffixes possessifs

Tier 1	Tier 2a	Tier 3
ktēbe	kitāb-**ī**	كتابي
ktēbek / ktēbak	kitāb-**ka**	كتابك
ktēbek / ktēbik	kitāb-**ki**	كتابكي
ktēbo	kitāb-**hu**	كتابه
ktēba	kitāb-**hā**	كتابها
ktēbna	kitāb-**nā**	كتابنا
ktēbkon	kitāb-**kon**	كتابكن
ktēbon	kitāb-**hon**	كتابهن

7. 1. 3. Les pronoms suffixes compléments

Tier 1	Tier 2a	Tier 3
jāni, jāne	jē-**nī**	إياني
jāk, ijjāk, jēk	jē-**ka**	إياك
jāki, ijjāki	jē-**ki**	إياكي
jē, ijjē, jāho	jē-**hu**	إياه
jē, ijjē, jāha	jē-**hā**	إياها
jāna	jē-**nā**	إيانا
jekon	jē-**kon**	إياكن
jon / jāhon	jē-**hon**	إياهن

7. 1. 4. Exemples

Tier 1	Tier 2a	Tier 3
bētna, bajtna	bajt-nā	بيتنا
katbətha, katabəta	katab-at-hā	كتبتها
ʕaṭīne, ʕṭīne	aʕṭi-nī, aʕṭī-nī	اعطني / اعطيني
ʔāletle	qāl-at l-ī	قالت لي
biddi, bəddi	badd-ī	بدّي
bəddak, biddak, baddak	badd-ka	بدّك
bəddik, biddik, baddik	badd-ki	بدّكي

bəddo	badd-hu	بدّه
bədda	badd-hā	بدّها
badna	badd-nā	بدّنا
bədkon	badd-kon	بدّكن
bəddon	badd-hon	بدّهن

7. 2. Le pronom relatif

Tier 1	Tier 2a	Tier 3
jilli, elli, li, l	jalli	اللي

7. 3. Les pronoms interrogatifs

Tier 1	Tier 2a	Tier 3
kīf	kīf	كيف
wejn, wēn	wīn	وين
fejn, fēn	fīn	فين
mīn	mīn	مين
lē	lē	ليه
lēʃ	lēʃ	ليش
ēʃ	ēʃ	إيش
ʃū	ʃū	شو
ʔaddēʃ	qaddēʃ	قديش
ʔadde	qaddē	قديه
ejmata	ejmatā	إيمتى
emta	emtā	إمتى
ej / ejja	aj / aja	أي / أية

7. 4. Les pronoms démonstratifs

Tier 1	Tier 2a	Tier 3
hajda, haza	hāðā	هذا
hajde, hajdi, hazihi	hāðī	هذي
hadōl, hadōle	haðūl	هذول
hdāk	haðāk	هذاك
hadīk, hadīke	haðīk	هذيك

8. Particules et prépositions

Toutes les particules et prépositions sont séparées par une espace en tier 1. En tier 2a, les particules monolitères sont segmentées par un tiret. En tier 3, on respecte les règles d'orthographe de l'arabe standard.

Tier 1	Tier 2a	Tier 3
Particules monolitères		
bi ṣṣēf	bi-l-ṣajf	بالصيف
bi lbēt	bi-l-bajt	بالبيت
bi bēt elʒirān	bi-bajt l-ʒirān	ببيت الجيران
bi lxzēne	bi-l-xizāne	بالخزانة
ta jektob	ta-ja-ktub	تيكتب
la jkūn	li-ja-kūn	ليكون
la lmadām	li-l-madām	للمدام
ʕa tṭawle	ʕa-l-ṭāwile	عالطاولة
Autres particules		
mn əlbidāje	min l-bidāje	من البداية
foʔ ərrās	fawq l-raʔs	فوق الرأس

9. Les chiffres

Les chiffres sont transcrits phonétiquement en tier 1, mais notés conformément à leur équivalent en arabe standard dans les tiers 2a et 3.

Tier 1	Tier 2a	Tier 3
ṣefer	ṣifr	صفر
wāħad	wāħid / aħad / iħdā	واحد / أحد / إحدى
tnēn	iθnajn / iθnatajn	إثنين / إثنتين
tlēte	θalāθ / θalāθa	ثلاث / ثلاثة
arbʕa	arbaʕ / arbaʕa	أربع / أربعة
xamse	xams / xamsa	خمس / خمسة
sette	sitt / sitta	ستّ / ستّة
sabʕa	sabʕ / sabʕa	سبع / سبعة
tmēne	θamānin / θamāniya	ثمانٍ / ثمانية
tesʕa	tisʕ / tisʕa	تسع / تسعة
ʕaʃra	ʕaʃar / ʕaʃra	عشر / عشرة
ħdaʕʃ	aħad ʕaʃar / iħdā ʕaʃra	أحد عشر / إحدى عشرة
tnaʕʃ	iθnaj ʕaʃar / iθnataj ʕaʃra	إثني عشر / إثنتي عشر
tlattaʕʃ	θalāθat ʕaʃar	ثلاثة عشر
ʕarbaʕtaʕʃ	arbaʕat ʕaʃar	أربعة عشر
xamstaʕʃ	xamsat ʕaʃar	خمسة عشر
settaʕʃ	sittat ʕaʃar	ستّة عشر
sabaʕtaʕʃ	sabʕat ʕaʃar	سبعة عشر
tmantaʕʃ	θamānijat ʕaʃar	ثمانية عشر
tesaʕtaʕʃ	tisʕat ʕaʃar	تسعة عشر
ʕeʃrīn	ʕiʃrīn	عشرين
tlētīn	θalāθīn	ثلاثين
arbʕīn	arbaʕīn	أربعين
xamsīn	xamsīn	خمسين
settīn	sittīn	ستّين
sabʕīn	sabʕīn	سبعين
tmēnīn	θamānīn	ثمانين
tesʕīn	tisʕīn	تسعين
mit	miʔa	مئة
eləf	alf	ألف

10. Morphologie verbale

Les désinences verbales ne sont segmentées par un tiret que dans la tier 2 où le radical du verbe est isolé de l'ensemble des affixes (morphèmes de modes/temps et personnes) qui font l'objet d'une notation conventionnelle.

Attention : la convention s'écarte de la proposition de CODA dans la notation de la conjugaison à la 1ère personne du pluriel (on note m-n- et pas b-n-).

En cas d'élision du morphème de conjugaison en tier 1 (ʿam fakkir), on rétablit en tier 2 et 3 le morphème manquant (ʿam u-fakkir). La voyelle du radical et des personnes (pour les formes dérivées) est transcrite phonétiquement en tier 1 mais devient conventionnelle en tier 2.

Voici les tableaux de conjugaison pour le verbe *katab*, suivis de différents exemples pour d'autres verbes.

10.1. Inaccompli

Tier 1	Tier 2a	Tier 3
bektob	b-**a**-ktub	بأكتب
btektob	b-**ta**-ktub	بتكتب
btektobe	b-**ta**-ktub-**ī**	بتكتبي
bjektob	b-**ja**-ktub	بيكتب

btektob	b-**ta**-ktub	بتكتب
mnektob	m-**na**-ktub	منكتب
btektobo	b-**ta**-ktub-**ū**	بتكتبوا
bjektobo	b-**ja**-ktub-**ū**	بيكتبوا

10.2. Inaccompli modal

Pour l'inaccompli à valeur modale (sans le préverbe b-) employé après différentes particules comme حتى ,لازم ,بدّ ,في , etc., la *hamza* de la première personne du singulier est rétablie dans les tiers 2 et 3.

Tier 1	Tier 2a	Tier 3
ektob	**a**-ktub	أكتب
tektob	**ta**-ktub	تكتب
tektobe	**ta**-ktub-**ī**	تكتبي
jektob	**ja**-ktub	يكتب
tektob	**ta**-ktub	تكتب
nektob	**na**-ktub	نكتب
tektobo	**ta**-ktub-**ū**	تكتبوا
jektobo	**ja**-ktub-**ū**	يكتبوا

10.3. Accompli

Tier 1	Tier 2a	Tier 3
katabet, katabt	katab-**t**	كتبت
katabet, katabt	katab-**t**	كتبت
katabte	katab-**tī**	كتبتي
katab	katab	كتب
katbet	katab-**at**	كتبت
katabna	katab-**nā**	كتبنا
katabto	katab-**tū**	كتبتوا
katabo	katab-**ū**	كتبوا

10.4. Impératif

Tier 1	Tier 2a	Tier 3
ktōb	u-ktub	اكتب
ktebe	u-ktub-ī	اكتبي
ktebo	u-ktub-ū	اكتبوا

10. 5. Formes augmentées

	Tier 1	Tier 2a	Tier 3
II	ħammam	ħammam, ju-ħammim	حمّم ، يحمّم
	ʔattal	qattal, ju-qattil	قتّل ، يقتّل
III	ʕāmal	ʕāmal, ju-ʕāmil	عامل ، يعامل
	kētab	kētab, ju-kātib	كاتب ، يكاتب

IV	aẓhar	aẓhar, ju-ẓhir	أظهر ، يظهر
V	tħammam	tħammam, ja-ta-ħammam	تهمّم ، يتهمّم
	tkassar	tkassar, ja-ta-kassar	كسّر ، يتكسّر
VI	tʕāmal	tʕāmal, ja-ta-ʕāmal	تعامل ،يتعامل
	tkētaf	tkētaf, ja-ta-kātaf	كاتف ، يتكاتف
VII	nkasar	nkasar, ja-n-kasir	انكسر ، ينكسر
	nʔaṭaʕ	nqaṭaʕ, ja-n-qaṭiʕ	نقطع ، ينقطع
VIII	ʃtaɣal	ʃtaɣal, ja-ʃtaɣil	اشتغل ، يشتغل
	ʕtamad	ʕtamad, ja-ʕtamid	اعتمد ،يعتمد
X	staʔzan	staʔzan, ja-sta-ʔzin	ستأزن ، يستأزن
	staɣfar	staɣfar, ja-sta-ɣfir	ستغفر ، يستغفر

10. 6. Exemples

Tier 1	Tier 2a	Tier 3
minʔassemon	m-nu-qassim-on	منقسّمهن
jkūn / jikūn	ja-kūn	يكون
ʕam jektob	ʕam ja-ktub	عم يكتب
raħ tektob	raħ ta-ktub	رح تكتب
ħa jektob	ħa-ja-ktub	حيكتب
ʕam fakkər	ʕam ufakkir	عم أفكّر
bteʃrabe	b-ta-ʃrab-ī	بتشربي
jaʕmol	ja-ʕmal	يعمل
semʕet	samiʕ-t	سمعت
balleʃ	ballaʃ	بلّش
biballeʃ	b-ju-balliʃ	بيبلّش
betħebb	b-tu-ħibb	بتحبّ
mənrūħ	m-na-rūħ	منروح
fallet	fall-at	فلّت
nsīna	nasī-nā	نسينا
ħajaʕtūne	ħa-ja-ʕt-ū-nī	حيعطوني
xallīni nām	xallī-nī anām	خلّيني أنام
ʔāletlon	qāl-at la-hon	قالت لهن
ʔaltella	qul-t la-hā	قلت لها
biddi ħʒuz	badd-ī aħʒuz	بدّي أحجز

11. Morphologie nominale

Dans la tier 1, les marques de genre et nombre des participes et des pluriels externes ne sont pas segmentées. Dans la tier 2a, elles sont segmentées. Conventionnellement, le pluriel externe masculin est noté -īn.

Tier 1	Tier 2a	Tier 3
ʕāmil	ʕāmil	عامل
ʕāmle	ʕāmil-e	عاملة
ʕāmlīn	ʕāmil-īn	عاملين
mʕallimīn	muʕallim-īn	معلمين
mʕallimāt	muʕallim-āt	معلمات

12. Autres éléments conventionnels

12. 1. Les expressions figées (liste ouverte)

Voici la liste des expressions qui sont codées en tier 2 et 4 et qui font l'objet d'une notation conventionnelle en tier 3, par ordre alphabétique arabe.

Tier 1	Tier 2a et 4	Tier 3
akīd	AKID	أكيد
in ʃā ʔaɬɬāh, nʃaɬɬa	INCH	إن شاء الله
enno	ENNO	إنّه
bass, bəss	BASS	بس
hallāʔ	HALQ	هلاق
lħamdəɬɬa	HDLA	الحمد لله
aɬɬa ʔirħamo	ARHM	الله يرحمـx
lakēn	LKAN	لكان
lajke	LAYK	ليك
maʃaɬɬa	MCHA	ما شاء الله
waɬɬā	WALA	والله
u hēke	WHEK	وهاك
jaɬɬā	YALA	ياالله
jaʕne, jaʕni	YANI	يعني

Tier 1	Tier 2a	Tier 3	Tier 4
lħamdəɬɬa mēʃe lħāl kīfik ənte	HDLA māʃī l-ħāl kīf-kī enti	الحمد لله ماشي الحال كفيك إنتي	HDLA ça va et toi
lajka:/ kənna badna hēk naʕmol #réunion#	LAYK kun-nā badd-nā hāk na-ʕmal #réunion#	ليك كنا بدنا هاك نعمل #réunion#	LAYK on voulait comme ça faire une #réunion ((FR))#

12. 2. Les mots ou syntagmes en langue étrangère

Tier 1	Tier 2a	Tier 3	Tier 4
bɔ̃ʒūr ṣabāja sori ajja #facture# la ljōm #dernier délai#	#bonjour# ṣabāja #sorry# ajja #facture# li-l-jawm #dernier délai#	#بنجور# صبايا #سوري# أيّة #facture# لليوم #dernier délai#	#bonjour ((FR))# les filles #désolée ((GB))# quelle #facture ((FR))# est pour aujourd'hui #dernier délai ((FR))#
enno Nicolas #hair# baddo #bonnet#	ENNO Nicolas #hair# badd-hu #bonnet#	إنّه نقولا #hair# بده #bonnet#	ENNO Nicolas a besoin d'un #bonnet ((FR))# pour ses #cheveux ((GB))#

12. 3. Les mots étrangers entrés dans les usages (liste ouverte)

Voici des exemples de mots d'origine étrangère entrés dans les usages : ils figurent en tier 2a dans l'orthographe de la langue. C'est une liste ouverte qui s'enrichira au fur et à mesure des transcriptions de nouveaux corpus.

Tier 1	Tier 2a	Tier 3	Tier 4
bɔnʒūr / bɔ̃ʒūr	#bonjour#	#بنجور#	#bonjour ((FR))#
sori / soɣi	#sorry#	#سوري#	#désolé (e) ((GB))#
mersi	#merci#	#مرسي#	#merci ((FR))#
sa va	#ça va #	#سا ڤا#	#ça va ((FR))#
hāj	#hi#	#هآي#	#salut ((GB))#

ḅāj	#bye#	#باي#	#au revoir ((GB))#
oke	#okay#	#أوكي#	#okay ((GB))#

12. 4. Les emprunts arabisés (liste ouverte)

Voici des exemples d'emprunts arabisés : ils figurent en tier 1 et 2a en transcription phonétique. C'est une liste ouverte qui sera enrichie au fur et à mesure des transcriptions de nouveaux corpus.

Tier 1	Tier 2a	Tier 3	Tier 4
#pojēt#	#pojāt#	#پويات#	#des pots ((FR))#
#taperwarēt#	#taperwarāt#	#تيبروارات#	#des tupperwares ((GB))#

12. 5. Régulateurs, marques d'hésitation, acquiescement, négation

Sens	Tier 1	Tier 2a	Tier 3	Tier 4
clic = non	tsk	tsk	tsk	tsk
morphème de négation	hə?ə	hə?ə	hə?ə	hə?ə
oui	?ē	?ē	إيه	oui
accusé réception minimal	?m	?m	?m	ʔm
accusé réception double	mhm	mhm	mhm	mhm
tag	ha	ha	ha	ha
morphème d'exclamation	a::::h, ã	āh	آه	ah
hésitation	?ø	?ø	ʔø	euh
hésitation (mot tronqué)	k- ktīr	k- kaθīr	كـ كثير	beau- beaucoup

12. 6. Les noms propres

Les noms propres (anonymisés ou non) figurent entre guillemets hauts dans les trois premières tiers. Il peut s'agir d'une personne, d'un lieu, d'une institution, etc.

Tier 1	Tier 2a	Tier 3	Tier 4
mitəl "Woody Allen"	miθl "Woody Allen"	مثل "وودي آلن"	comme Woody Allen
aflām "Jusef ʃahin" el?adīme	aflām "Jūsuf ʃāhīn" el-qadīma	أفلام "يوسف شاهين" القديمة	les vieux films de Youssef Chahine

12. 7. La mention de noms anonymisés

Dans toutes les tiers de la transcription, les noms propres anonymisés cités apparaissent sous leur forme intégrale ou abrégée selon les cas. Cela peut concerner une personne, un lieu, une institution, etc. S'il s'agit du pseudo de l'un des participants à l'interaction, on le conserve sans guillemets.

Tier 1	Tier 2a	Tier 3	Tier 4
marħaba SOF (0.5) kīfik	marħaba SOF (.) kīf-ik	مرحبا SOF (.) كيفك	Salut SOF (.) ça va
marħaba ja "ʒorʒ"	marħaba jā "Georges"	مرحبا يا "جورج"	Salut Georges

13. Le lexique dialectal (liste ouverte)

Dans ce tableau, amené à être enrichi, on a relevé les mots dialectaux et expressions très courants ou sans équivalents en arabe standard. La tier 1 présente une liste des variantes alors que les tiers

2a et 3 proposent la notation conventionnelle correspondante. Les termes figurent dans l'ordre alphabétique arabe de la tier 3.

Tier 1	Tier 2a	Tier 3	Tier 4
alo, a:lo	alo	ألو	Allô
oke, okaj	oke	#أوكي#	okay
ənno	ʔnno	انه	que
jā… jā	jā… jā	أو	ou
wala	wa-lā	أو	ou
ūḍa	ūḍā	أوضى	chambre, pièce
awwal ma	awwal mā	اول ما	dès que
ajwa	ajwā	أيوا	oui, bien sûr
ḅāj	bāj	#باي#	salutation de clôture
beħkīki, mneħke	ba-ħkī-ik, m-na-ħkī	بحكيكي, منحكي	je te parle, on se parle (salutation de clôture, à bientôt)
bʃūfek nʃālla	ba-ʃūf-ik INCH	بشوفك إن شاء الله	je te vois INCH (salutation de clôture, à bientôt)
baddak / biddak badda	badd-ka badd-hā	بدّك بدّها	tu veux elle veut
beddi ʔəzəʕʒek	badd-ī u-zʕiʒ-ak	بدي ازعجك	je vais te/vous déranger
baddik ʃī	badd-ik ʃī	بدك شي	tu veux quelque chose ? (pré-salutation de clôture)
barra	barra	برّا	à l'extérieur
barrat…	barrat…	برّاة...	à l'extérieur de…
barḍo	barḍo	برضو	encore, aussi
bass	bass	بس	quand, seulement, juste
b ʔaleb	bi-qalb	بقلب	dans, à l'intérieur
bala	bilā	بلا	sans
balke, barke	balke	بلكي	peut-être
tabaʕ tabʕet tabʕūl	tabaʕ tabʕet tabʕūl	تبع تبعة تبعول	à, de
ʒuwwa	ʒuwwa	جوّا	à l'intérieur, dedans
ʒuwwat…	ʒuwwat…	جوّاة...	à l'intérieur de…
ħadd	ħadd	حد	à côté de
lħamdəlla: meʃe lħāl	HDLA māʃī l-ħāl	الحمد لله ماشي الحال	HDLA ça va
lħamdəlla tamēm	HDLA tamām	الحمد لله تمام	HDLA très bien, super
dəɣri	duɣrī	دغري	directement
rāħ, raħ	raħ	رح	(particule du futur)
zalame	zalame	زلمة	gars
zejj	zajj	زيّ	comme
salām	salām	سلام	(salutation d'ouverture)
ʃwaj, ʃwajj, ʃwejj	ʃwaj	شوي	un peu
ʃwajet…	ʃwajat-	شوية...	un peu de…
ʃī	ʃī	شي	(quelque) chose
ṭajjeb	ṭajjib	طيب	bon, bien

ʕaʃān, ʕalaʃān	ʕaʃān	عشان	pour
fəm, təm	fam	فم	bouche
karmēl	karmāl	كرمال	pour
kamēn, kamēna	kamān	كمان	aussi
lā	lā	لا	non
laʔ, laʔa	lāʔa	لأ	certainement pas
lessa, essa	lissa	لسّة	pas encore
mbārəħ	mbāriħ	مبارح	
mbala, əmbala	mbalā	مبلا	non, au contraire
madām, madāmtak	madām, madāmat-ka	مدام	madame
madri	madrī	مدري	je ne sais pas
məʃ, miʃ, muʃ	mʃ	مش	(négation)
meʃtēʔetlek	muʃtāqa la-ki	مشتاقة لك	tu me manques
miʃān, minʃān	miʃān	مشان	pour
mēʃe lħāl	māʃī l-ħāl	ماشي الحال	ça va
mnīħ, mlīħ	mnīħ	منيح	bien
mū	mū	مو	(négation)
nijjāl…	nijjāl-	نيال	tant mieux pour…
halla?	hallaq	هلاق	maintenant
hōn, hōne	hunā	هنا	là
honīk, hnēke	hunāk	هناك	là-bas
hēk	hāk	هيك	comme ça
hāj	hāj	هاي	hi (salutation d'ouverture)
jalla	jallāh	يللى	allez
jalla ḅāj	jallāh bāj	يللى باي	allez salut

Annexe 3 : Liste indicative des gloses morpho-syntaxiques

1	première personne
2	deuxième personne
3	troisième personne
ACC	accusatif
ACP	accompli
ACT	actif (vs. passif)
ADJ	adjectif
ADV	adverb(ial)
AFX	affixe
AGR	accord, agreement
ANN	annexion
ART	article
AUX	auxilaire
CHI	nombre cardinal ou ordinal
COLL	collectif
COMP	complémentiseur (ex. ENNO, que complétif)
COP	copule (pronominal copula)
DAT	datif
DEF	défini
DEM	démonstratif
DIM	diminutif
DU	duel
ELAT	élatif
F	féminin
FUT	futur
INA	inaccompli
IFX	infixe
IMP	impératif
INDF	indéfini
INTENS	intensifieur
INTER	interrogatif
LOC	locatif
M	masculin
N	nom
NEG	négation
NPR	nom propre
NOM	nominatif
PART	particule
PAST	passé
PRE	préposition
PAS	passif
PFX	préfixe
PL	pluriel
POSS	possessif
PRO	pronom personnel
PRV	préverbe
PRES	présent
PTA	participe actif
PTP	participe passif
REL	relatif
SFX	suffixe
SG	singulier
SUPER	superlatif
V	verbe
VOC	vocatif

Annexe 4 : Transcrire les dialectes arabes en linguistique. Comparaison des conventions existantes

Document réalisé par Catherine PINON

La transcription des dialectes arabes en caractères arabes est problématique, car il n'existe aucune orthographe officielle et les différences phonétiques et morpho-phonologiques peuvent être très importantes d'un dialecte à l'autre. Proposer une convention de transcription unique adaptable à tous les dialectes faciliterait les recherches et leur diffusion. Plusieurs propositions ont été faites dans ce sens. Nous avons choisi ici d'en présenter trois, sur lesquelles reposent les choix de transcription en effectués dans la convention AraPI : les conventions CODA, AMADAT et BBN. Nous mentionnerons les règles propres à AraPI quand elles s'écartent des propositions présentées ici.

La réflexion qui suit sur la transcription des dialectes en caractères arabes ne vise pas l'établissement d'une orthographe standardisée. Elle propose une comparaison de systèmes existants, analysant ce qui les rapproche et ce qui les différencie. Il faut garder en tête que ces conventions de transcription ont pour but de proposer une référence bien exhaustive et bien explicitée. Son objectif majeur est de permettre le partage des données entre des chercheurs de différentes disciplines.

Ce document vise, dans le même esprit que les *Leipzig Glossing Rules*, à déduire des différentes propositions faites des règles communes, afin d'uniformiser les pratiques de transcription des dialectes. Après l'énoncé de règles générales, on prendra l'exemple du dialecte syro-libanais pour illustrer la méthode de transcription à suivre.

Table des matières

1. Les travaux existants

Les linguistes travaillant sur l'arabe sont tôt ou tard confrontés à la question de la transcription de leurs sources. Chaque chercheur fait ses propres choix en fonction de ses objectifs scientifiques, mais aussi de contraintes techniques qui peuvent peser sur ses recherches. Il existe plusieurs conventions qui visent à unifier les systèmes de transcription utilisés par les linguistes arabisants.

1. 1. La convention CODA

La convention CODA (Conventional Orthography for Dialectal Arabic) a été établie par Nizar Habash, Mona Diab et Owen Rambow du Center for Computational Learning Systems de l'Université Columbia à New York, en 2011.

> références en ligne :
Habash, Diab et Rambow, "Unified Guidelins and Resources for Arabic Dialect Orthography" [En ligne] http://www.lrec-conf.org/proceedings/lrec2018/pdf/395.pdf

Habash, Diab et Rambow, "Conventional Orthography for Dialectal Arabic" [En ligne] http://www.lrec-conf.org/proceedings/lrec2012/pdf/579_Paper.pdf

1. 2. La transcription AMADAT

Il s'agit d'une transcription dialectale à base orthographique (Dialectal Arabic Orthography-based transcription) mise au point par Mohamed Maamouri, David Graff, Hubert Jin, Christopher Cieri et Tim Buckwalter du Linguistic Data Consortium de l'Université de Pennsylvanie, pour l'étude des conversations téléphoniques (Conversational telephone speech = CTS) en arabe levantin.

> références en ligne : Maamouri, Graff, Jin, Cieri et Buckwalter, "Dialectal Arabic Orthography-based Transcription & CTS Levantine Arabic Collection" [En ligne] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/ears-rt04-dialectal-arabic-transcription.pdf

1. 3. Les directives BBN

Les directives pour transcrire les dialectes arabes (Guidelines for Transcribing Arabic Dialects) ont été proposées par Bushra Zawaydeh, Dave Stallard et John Makhoul de BBN Technologies en 2002-2003.

> références en ligne : Zawaydeh, Stallard et Makhoul : "Guidelines for Transcribing Arabic Dialects" [En ligne] https://catalog.ldc.upenn.edu/docs/LDC2005S08/BBN-Babylon-transcription-guidelines.pdf

2. Les règles de transcription communes

D'une manière générale, les transcriptions se veulent à la fois panarabiques et spécifiques. Elles visent la facilité d'emploi et la lisibilité. À partir d'un examen détaillé de ces conventions, nous avons distingué les éléments communs et les règles spécifiques qui s'en dégagent. Nous en proposons un résumé sous la forme d'un ensemble de règles.

Règle 1 : un mot = une seule et unique transcription

Chaque mot a une unique forme orthographique en transcription, même si sa réalisation phonétique varie.

Par conséquent, on peut établir pour chaque mot une transcription conventionnelle et une liste des variantes.

Règle 2 : une orthographe ayant comme référence le MSA

Dès que cela est possible, les choix de transcription sont conformes à l'orthographe de l'arabe littéral (Modern Standard Arabic ou MSA), qui est assez stable sur l'ensemble du monde arabe et constitue une référence partagée par tous les arabophones scolarisés. Par exemple, la cliticisation des particules monolitères, le recours à la *šadda* pour la gémination phonologique, l'écriture de l'article ou l'emploi du *tā' marbūṭa* et du *alif maqṣūra* sont conformes à l'orthographe standard.

Par conséquent, on cherche à établir pour chaque mot un lien avec le MSA.

- Si le mot a une forme directement apparentée en MSA, utiliser l'orthographe MSA (ce qui peut donner une forme qui n'existe pas en MSA, cf. exemple 4).

Exemples :

(1) Libanais
/kti:ʀ/ ou /kati:ʀ/
كثير (non pas كتير) — *l'interdentale* [θ] *réalisée* [t] *est rétablie*

(2) Libanais
/daʔən/
ذقن (non pas دقن ou دأن) — *le* [ð] et le [ɋ] *réalisés* [d] *et* [ʔ] *sont rétablis*

(3) Libanais
/zɣīr/
صغير (non pas زغير) — *le* [ṣ] *réalisé* [z] *est rétabli*

(4) Libanais
/hadōl/
هذول (non pas هدول) — *le* [ð] *des démonstratifs standard est rétabli, même si cette forme n'existe pas en MSA.*

- Dans le cas où la racine commune au dialecte et au MSA a une consonne ajoutée ou élidée, on rétablit aussi la racine MSA.

Exemples : (5) Libanais
/u-noṣ/
ونصف (non pas ونص)

- On réintègre les voyelles longues de schèmes particuliers même si elles ne sont pas réalisées et on note les tanwīn-an اً*.*

Exemples : (6) Libanais
/ṭabūr/
طابور (non pas طبور) *le* [ā] *réalisé* [a] *est rétabli*

(7) Libanais
/matalan/
مثلاً (non pas مثلن) *le* [ɵ] *réalisé* [t] *est rétabli et* [an] *est noté* [اً]

- Si on utilise des textes déjà transcrits en dialecte, on corrige les erreurs typographiques (métathèses, omission ou ajout d'espaces). Par exemple, on rétablit la bonne orthographe du yā' / alif maqṣūra *dans les textes égyptiens.*

- Dans le cas où le lien entre la forme dialectale et le MSA n'existe pas, on livre une liste des transcriptions conventionnelles.

Exemples : (8) Libanais
/doɣri/
دغري (notation conventionnelle)

Règle 3 : une morpho-syntaxe qui préserve les spécificités dialectales

Bien que s'inspirant pour la transcription des sons du *ductus* classique, les spécificités morpho-syntaxiques des dialectes sont maintenues.

Par conséquent, on conserve les marques de temps et d'aspect dans la conjugaison, les terminaisons des pluriels externes, les pré-verbes dialectaux (par exemple, on ne remplace pas le marqueur de futur ḥa- *par* sa-*) et on conserve les spécificités lexicales dialectales.*

Exemple : (9) Libanais
/b-ya-ktub/
بيكتب (non pas يكتب)

(10) Tunisien
/bdī-t/
بديت (non pas بدأت)

(11) Libanais
/raħ balliʃ/
رح أبلّش (non pas سأبدأ)

(12) Libanais
/fēto ləmʕalmīn/
فاتوا المعلّمين (non pas دخل المعلّمون)

Règle 4 : adaptation, mais cohérence et pertinence

Les chercheurs adaptent les conventions existantes en fonction des objectifs de leur recherche. Ils doivent alors expliciter leurs choix. Dans tous les cas, les conventions établies doivent être suivies strictement et les exceptions listées.

3. Les règles de transcription spécifiques

En fonction des objectifs de la recherche, certains choix doivent être faits concernant des notations particulières.

Règle 5 : la transcription des mots étrangers

Les systèmes CODA, AMADAT et BBN s'accordent sur le fait de transcrire les mots étrangers en écriture arabe sans recourir à des caractères supplémentaires comme چ, پ ou ڤ . On utilise la transcription arabe usuelle si elle existe. Dans le cas contraire, on peut modifier la valeur de certaines lettres pour ne pas recourir à des caractères supplémentaires :
- on utilise le *bāʾ* (ب) pour transcrire [p]
- on utilise le *tāʾ+šīn* (تش) pour transcrire [tch]
- on utilise le *fāʾ* (ف) pour transcrire [v]
- on utilise le *jīm* (ج) ou le *kāf* (ك) pour transcrire [g]

Règle 5a (optionnelle) : préséance des formes levantines

En cas de concurrence entre plusieurs transcriptions connues, CODA préfère à la transcription égyptienne la transcription levantine : on écrit (كراج), que l'on transcrive de l'égyptien ou du libanais.

Règle 5b (optionnelle) : adaptation au standard dialectal

En cas de concurrence entre plusieurs transcriptions connues, AMADAT propose de se conformer avec les usages locaux : on écrit (كراج) si l'on transcrit du libanais et (جراج) si l'on transcrit de l'égyptien.

Règle 5c (optionnelle) : caractères arabes supplémentaires

AraPI recourt, dans le cas des emprunts arabisés, à un certain nombre de caractères supplémentaires, tels que گ, پ ou ڤ, en indiquant à l'aide de l'alphabet phonétique international leur valeur.

Règle 5d (optionnelle) : transcription en caractères latins

AraPI propose de transcrire les segments en langue étrangère en caractères latins, dans leur orthographe d'origine.

Règle 6 : code-switching et diglossic code-switching

Le marquage du passage d'une langue à une autre n'apparaît pas toujours dans les différentes conventions.

Règle 6a (optionnelle) : uniformisation (dialectalisation du MSA)

Dans CODA, si un terme est prononcé en MSA, on le transcrit dans le dialecte. Par exemple, si le mot « homme » est prononcé [raʒul] comme en MSA dans un discours en égyptien, il sera transcrit راجل [rāgəl] conformément à la spécificité lexicale de ce mot en égyptien.

Règle 6b (optionnelle) : marquage du code-switching

Dans AraPI, les mots ou syntagmes en langue étrangère ou dans une variété d'arabe différente figurent encadrés de #.

Règle 7 : noms propres et abréviations

La transcription des noms propres ou des abréviations peuvent poser problèmes. Dans l'ensemble des conventions, on suggère d'utiliser la transcription consensuelle quand elle existe. D'autres propositions sont parfois faites concernant spécifiquement les noms propres.

Règle 7a (optionnelle) : marquage des noms propres et abréviations

BBN propose de lier par un tiret bas les noms propres et abréviations composées, par exemple :

برج_البراجنة

Règle 7b (optionnelle) : marquage des noms propres et abréviations

AraPI propose d'entourer la transcription des noms propres par des guillemets. Exemple :

"ماركس"

4. Questions courantes

Les éléments linguistiques les plus courants ne font pas toujours l'objet de règles détaillées. Voici point par point les propositions partagées par les conventions (quand elles les abordent), avec la mention des divergences quand le consensus n'est pas établi.

Les pronoms isolés et suffixes
Chaque dialecte a son écriture conventionnelle des pronoms. La 2ème personne du féminin singulier est toujours marquée par un *yāʾ* : كي – إنتي
Pour les pronoms commençant par un *hāʾ*, on le note toujours même s'il n'est pas prononcé.

Les verbes
- on garde l'allongement vocalique des redoublés et des hamzés de 3ème radicale (بديت)
- on note la 2ème personne du féminin singulier par un *yāʾ* (بديتي)
- on note la 3ème personne du masculin pluriel par وا (le *alif* n'est pas noté uniquement s'il y a un suffixe après : بيكتبوا - بيكتبوها)

Les préverbes sont attachés au verbe s'ils sont monolitères, sinon ils sont séparés par une espace, conformément à l'écriture arabe.
حأكتب – تيروح – عم بتكتب – رح بيكتبوا

La variante négative شي du négatif ش est toujours orthographiée ش à la fin du verbe.

Variantes : Pour CODA, le paradigme de conjugaison du libanais est :
بأكتب – بتكتب – بتكتبي – بيكتب – بتكتب – بنكتب – بتكتبوا – بيكتبوا
Pour BBN, la 1ère personne du singulier s'écrit sans la *hamza* : باكتب . CODA recommande de ne jamais transcrire منكتب à la 1ère personne du pluriel, même si c'est prononcé ainsi. Pour AraPI à l'inverse, on note منكتب .

Les prépositions
Les particules monolitères sont attachées : والكتاب – بالبيت
Les autres particules ne doivent pas être cliticisées : ما de négation, ainsi que لـ + suffixe.
Par exemple, on écrira ما قلت لهاش (pas ماقلتلهاش).

Les démonstratifs
Les formes brèves comme *ha-* sont attachées dans CODA et AraPI, séparées par une espace dans BBN :

هالبيت - بهالبيت – عهالبيت vs ها البيت

Notation des chiffres et nombres en toutes lettres, jours de la semaine
Pour CODA et AraPI, les chiffres et les nombres écrits en toutes lettres sont conformes à l'écriture standard, quelle que soit la réalisation. Il en va de même pour les jours de la semaine. Pour BBN les nombres sont écrits en transcription dialectale (طنعشر…).

Les mots courants (*chose, eau, viens !*)
Pour CODA, on suit les règles du MSA dans les cas suivants :

Convention	Variantes non retenues
شيء	شي - إشي
ماء	ماي – مي – مایة
تعال – تعالي - تعالوا	تع – تعي - تعوا

AraPI propose une liste ouverte de mots dialectaux et expressions très courantes.

5. Bibliographie complémentaire

La convention CODA appliquée à différents dialectes :

- HABASH, DIAB et RAMBOW, *Conventional Orthography for Dialectal Arabic (CODA): Principles and Guidelines - Egyptian Arabic* [En ligne]
https://academiccommons.columbia.edu/doi/10.7916/D83X8562

- HABASH, JARRAR, ALRIMAWI, AKRA, ZALMOUT, BARTOLOTTI et ARAR, *Palestinian Arabic Conventional Orthography Guidelines - Technical Report* [En ligne]
http://www.jarrar.info/publications/HR15.pdf

- SAADANE et HABASH, *A Conventional Orthography for Algerian Arabic* [En ligne]
http://www.aclweb.org/anthology/W15-3208

- TURKI, ADEL, DAOUDA et REGRAGUI, *A Conventional Orthography for Maghrebi Arabic* [En ligne]
https://www.researchgate.net/publication/311589181_A_Conventional_Orthography_for_Maghrebi_Arabic

- ZRIBI, BOUJELBANE, MASMOUDI, ELLOUZE, BELGUITH et HABASH, *A Conventional Orthography for Tunisian Arabic* [En ligne]

Annexe 5 : Tableau synoptique de la convention AraPI

Toutes les notations conventionnelles figurent en annexe 1 (phénomènes interactionnels) et annexe 2 (liste des notations conventionnelles).

<table>
<tr><th>Éléments conventionnels</th><th>Tier 1</th><th>Tier 2a</th><th>Tier 3</th><th>Tier 4</th></tr>
<tr><td>Transcription (type)</td><td>interactionnelle
phonético-phonologique</td><td colspan="2">morpho-syntaxique
(référant arabe standard et standard dialectal)</td><td>traduction dans la langue du chercheur</td></tr>
<tr><td>Alphabet</td><td colspan="2">API modifié</td><td>arabe (modifié) / latin</td><td>latin</td></tr>
<tr><td>Segmentation</td><td>en mots</td><td>conventionnelle
morphémique par tiret (-)</td><td>en mots selon les règles de la graphie arabe</td><td>-</td></tr>
<tr><td>Interaction</td><td>ICOR modifiée</td><td colspan="3">aucune notation sauf pauses, enchaînements et chevauchements</td></tr>
<tr><td>Pronoms, déterminants, particules, prépositions</td><td>transcription phonétique</td><td colspan="2">notation conventionnelle
(référant arabe standard et standard dialectal)</td><td>-</td></tr>
<tr><td>Chiffres</td><td>transcription phonétique</td><td colspan="2">notation conventionnelle
(référant arabe standard et standard dialectal)</td><td>-</td></tr>
<tr><td>Morphologie nominale et verbale</td><td>transcription phonétique</td><td colspan="2">notation conventionnelle
(référant arabe standard et standard dialectal)</td><td>-</td></tr>
<tr><td>Expressions figées</td><td>transcription phonétique</td><td>codage 4 majuscules</td><td>notation conventionnelle</td><td>codage 4 majuscules</td></tr>
<tr><td>Mots ou syntagmes en langue étrangère</td><td colspan="2">entre dièses, dans l’orthographe de la langue</td><td>entre dièses, transcription dans l’orthographe de la langue que les segments soient courts ou longs</td><td rowspan="3">entre dièses, traduits, mention de la langue par le code ISO entre doubles parenthèses</td></tr>
<tr><td>Mots étrangers entrés dans les usages</td><td>transcription phonétique</td><td>entre dièses, dans l’orthographe de la langue</td><td rowspan="2">entre dièses, transcription en caractères arabes</td></tr>
<tr><td>Emprunts arabisés</td><td colspan="2">entre dièses, transcription phonétique</td></tr>
<tr><td>Régulateurs</td><td colspan="4">transcription conventionnelle</td></tr>
<tr><td>Noms propres</td><td>entre guillemets hauts orthographe officielle ou transcription phonétique avec majuscules</td><td>entre guillemets hauts orthographe officielle ou transcription standard avec majuscules</td><td>entre guillemets hauts orthographe officielle ou transcription en caractères arabes</td><td>orthographe officielle avec majuscules</td></tr>
<tr><td>Éléments dialectaux courants</td><td>transcription phonétique</td><td>notation conventionnelle en API modifié</td><td>notation conventionnelle en caractères arabes</td><td>-</td></tr>
</table>