Production des outils pour l'étude des langues peu dotées : le cas de l'arabe écrit contemporain

Ouafae Nahli
ouafae.nahli@ilc.cnr.it

Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli"

Bruit de fond ou valeur ajoutée?
Gérer le bruit lors des traitements informatiques des corpus linguistiques
28 avril 2023 Grenoble

Langue vs dialecte

- Une **langue** peut être considérée un dialecte qui :
  - → a un **statut officiel** reconnu
  - → a des **normes grammaticales** et des **dictionnaires** décrétées par des académies ou des autorités reconnues
    - →→ gagne le statut **écrit** ce qui contribue à la rendre plus **stable**

- Un **dialecte** généralement appartient à une zone géographique limitée:
  - → il ne possède ni normes ni dictionnaires et donc il n'a pas un statut officiel
  - → il reste à l'état **orale** et risque la **disparition** si le nombre de ses locuteurs diminue.

- Il n'existe aucune limite claire entre une langue e un dialecte: on parle d'un **continuum linguistique** où il existe un éventail de **dialectes intercompréhensibles**

  →Malgré les différences qui caractérisent leurs dialectes (appartenant à la même langue), les personnes se comprennent entre eux.

Dans plusieurs réalités linguistiques, une ou plusieurs langues officielles se superposent à plusieurs langues vernaculaires ou régionales.

→ Dans ce cas les personnes perfectionnent deux langues ou plus et sont capables de passer de l'une à l'autre en fonction des contextes où ils s'expriment: phénomène de **multilinguisme**

Arabe classique vs dialectes arabes

❖ La langue et les dialectes arabes constituent un cas extrême en raison de son **extension géographique**.

➢ La **langue arabe standard** est reconnue grâce au **statut religieux** (langue du Coran), elle a un statut social, intellectuel et littéraire : c'est la **langue écrite** par excellence dans tous les pays arabes.

➢ Les **dialectes arabes** sont relégués à un **statut orale,** cependant :

- ✓ avec l'avenue de Internet et des **réseaux sociaux**, les personnes communiquent grâce à la forme linguistique qu'il connaissent au mieux : leur **dialecte qui acquiert un statut écrit**
  →→ textes écrits arabes qui reflètent **la réalité linguistique des parlants natifs**: **l'arabe écrit contemporain**

❖ **L'intercompréhension** est possible, mais elle est loin d'être évidente entre deux dialectes éloignés géographiquement (l'arabe du Maroc et celui du Qatar par exemple).

CWALM
A lexical corpus-based model of Contemporary Written Arabic

Le projet propose une nouvelle approche qui permettrait d'analyser, selon la **tradition linguistique de corpus**, **l'arabe come réalité linguistique des locuteurs natifs**

→ étude de **l'arabe contemporain écrit** après l'analyse d'un corpus représentatif qui se définit selon un ensemble de critères externes et objectifs (comme le temps, le genre, le domaine)

→ compte tenu de la description objective de **l'arabe contemporain écrit**, il sera produite une ressource lexicale qui présentera l'arabe moderne standard et ses variétés dialectales comme **un ensemble complexe mais unifié**

CWALM
A lexical corpus-based model of Contemporary Written Arabic

1- Construction d'un corpus représentatif de l'arabe écrit contemporain

2- Annotation du corpus avec des informations lexicales, morphologiques, syntaxiques et sociolinguistiques

3- Construire un modèle lexicographique capable d'accueillir dans un même item lexical (« hyperlemme » ou « arco-lemme ») les lemmes en MSA et dans les différents dialectes

4- Construire un modèle permettant une connexion systématique entre les données du corpus, les items lexicaux et les informations lexicales

5- Construire un lexique structuré qui sera peuplé par les données extraites des annotations et de l'analyse du corpus. En plus, la signification des items lexicaux sera liée (quand c'est possible) au concept d'une ontologie (par exemple SUMO (Pease, A. ; 2006))

6 - Documentation nécessaire qui permet à d'autres chercheurs de poursuivre le développement

- **À Noter que l'arabe contemporain écrit** est une **langue peu dotée :** une langue parlée avec peu de ressources linguistiques. Il est difficile de faire :
  - la collecte et la construction de corpus
  - l'annotation et l'analyse des données
  - le développement des instruments et des outils

Recommandations et phases de construction de ressources pour les langues peu dotées

1- Préparation

- il est crucial de faire des recherches sociolinguistiques et statistiques sur la dynamique et les pratiques d'utilisation de la langue au sein de la communauté:
  - → les statistiques **d'alphabétisation** des utilisateurs parlants
  - → la disponibilité de **systèmes d'écritures**
  - → les pratiques **multilingues** et linguistiques mixtes dans la communauté

- Enquête et utilisation des **ressources de données existantes** comme **points de départ** (bien que petits ou pas de haute qualité)
- elles donnent un **aperçu préliminaire** sur les caractéristiques de la langue et des phénomènes linguistiques
- elles sont généralement accompagnées d'une description des **métadonnées** (par exemple, participants/communauté, contexte) qui peuvent aider à s'orienter dans la collecte

2- Collecte de données

- dans la plupart des cas, on utilise des ressources de langues riches pour construire des données de langue peu dotée, par exemple par biais de traductions, le résultat ne reflète pas la réalité linguistique de la communauté en étude
- Par contre, il faut **gérer la variation linguistique et ne pas ignorer les influences linguistiques** étrangères (par exemple, le code-switching; emprunts..)

→ **La collecte doit être faite du bas vers le haut (manière ascendante) auprès des membres de la communauté**

3- Nettoyage et normalisation des données

❖ **La disponibilité des données** des médias sociaux telles que Twitter et Facebook ont créé des besoins et des défis nouveaux :

→ **le contenu n'est pas standard**

→ il ne respecte pas **l'orthographie standard**

→ les écritures diverses (arabe ou Arabizi)

→ utilisation de **emoji** ☺ ☹

… ➔ **Existence de bruits dans le texte**

Pour **faciliter le travail automatique**, la variation linguistique des données porte à normaliser les données

❖ Les **processus de normalisation** supposent, que pour chaque langue, existe une norme souvent associée au **monolingue**

- **ignorer le multilinguisme** e les **aspects dynamiques** des données
- « normaliser » les données vers un dialecte standard ou vers une langue formelle associée à un registre formel (MSA)
  - **ignorer les communautés** qui n'utilisent pas le dialecte standard dans leurs communications quotidiennes.
  - **adapter toutes les données à un domaine précis**: dans le cas de l'arabe au domaine écrit, littéraire et religieux du MSA

Définition des bruits

Un **bruit est nuisible** lorsqu'il affecte le sens voulu du texte.

Un **bruit est utile** lorsqu'il sert pour un but important de l'étude et/ou pour améliorer les performances du système TAL

Définition des bruits

Caractéristiques intrinsèques de l'alphabet arabe

Il comporte vingt-huit lettres et s'écrit horizontalement de droite à gauche.

C'est un système d'écriture qui note seulement les consonnes.

→ Les voyelles peuvent être ajoutées sous forme de diacritiques.

→ Le redoublement de consonne et le symbole de la hamza sont aussi des formes diacritiques.

Un mot écrit peut généralement admettre plusieurs lectures suivant la répartition (ou l'absence) des diacritiques.

Racine	**كتب**									
	كتب	**ktb**			**كاتب**	**kAtb**		مكتب	mktb	
	كَتَبَ كُتِبَ كِتَّبَ كُتُب كَتْب	kataba kutiba kattaba kutub kat°b	il a écrit il a été écrit il fait écrire livres écriture		كَاتَبَ كَاتِب	kaAtaba kaAtib	Il a écrit à quelqu'un écrivain	مَكْتَب	maktab	bureau

Racine	**سيل**	**syl**			**س ءل**		
Verbe	سال	**sAl**	ا		**سأل**	**s'l**	أ
	سَالَ	saAla	il s'est renversé		سَأَلَ	sa'ala	il a demandé

3- Nettoyage et normalisation des données

Définition des bruits

- Les signes diacritiques ont une valeur:
  - flexionnelle: كَتَبَ kataba « il a écrit » / كُتِبَ kutiba « il a été écrit »
  - dérivationnelle: كَتَّبَ kattaba « il a fait écrire »
  - sémantique: سَالَ saAla « il s'est renversé » / سَأَلَ sa'ala « il a demandé »

جمُل	جمِل	جمَل
jamula	jamila	jamala

- Le problème se pose en raison du **manque de cohérence** dans l'utilisation des signes diacritiques : kataba - katb – ktab - ktba …
- Une pratique courante dans la gestion automatique des textes arabes consiste à normaliser le texte d'entrée en enlevant tous les signes diacritiques tels que les voyelles brèves, la gémination et le symbole hamza.
- L'analyse se fait systématiquement sur une forme non diacritique.
  - ✓ **La normalisation accélère le processus en résolvant la variabilité d'entrée mais**
  - ✓ **l'ambiguïté augmente**.

Définition des bruits

La substitution

❖ Écriture en Arabizi

Arabizi permet d'écrire un texte arabe en lettres latines. Pour le fait que le clavier latin ne couvrent pas toutes les consonnes arabes, on procède au **phénomène de la substitution** c'est-à-dire utiliser des chiffres arabes pour représenter des lettres qui n'ont pas d'équivalent dans l'alphabet latin. En raison de la nature informelle, la transcription peut être différente d'une région à une autre :

ش	š	ch/sh
ط	ṭ	t/T/6
ح	ḥ	7

Le verbe شَطَحَ **šaṭaḥa** «il a dansé» peut être écrit

chata7a	**chaTa7a**	**cha6a7a**
shata7a	**shaTa7a**	**sha6a7a**

Définition des bruits

→ **Commutation de code (Code-switching)** il est courant de basculer entre scripts différents pour transcrire la parole empruntée. On finit par écrire le même mot dans les deux scripts dans différentes instances du mème corpus.

→ **Variantes orthographiques** : سوريا / سورية

→ **Contraction des paroles :**

مَا	كَتَبْتُ	شَيْئًا
maA	katabtu	shay'an
ما كتبتش		ماكتبش
mA ktbtsh		**mA**ktbtsh

→ **Allongement des mots et répetition de la ponctuation**

قمرررررر qamarrrrr!!!!!! trèèèèèès belle!!!!!!

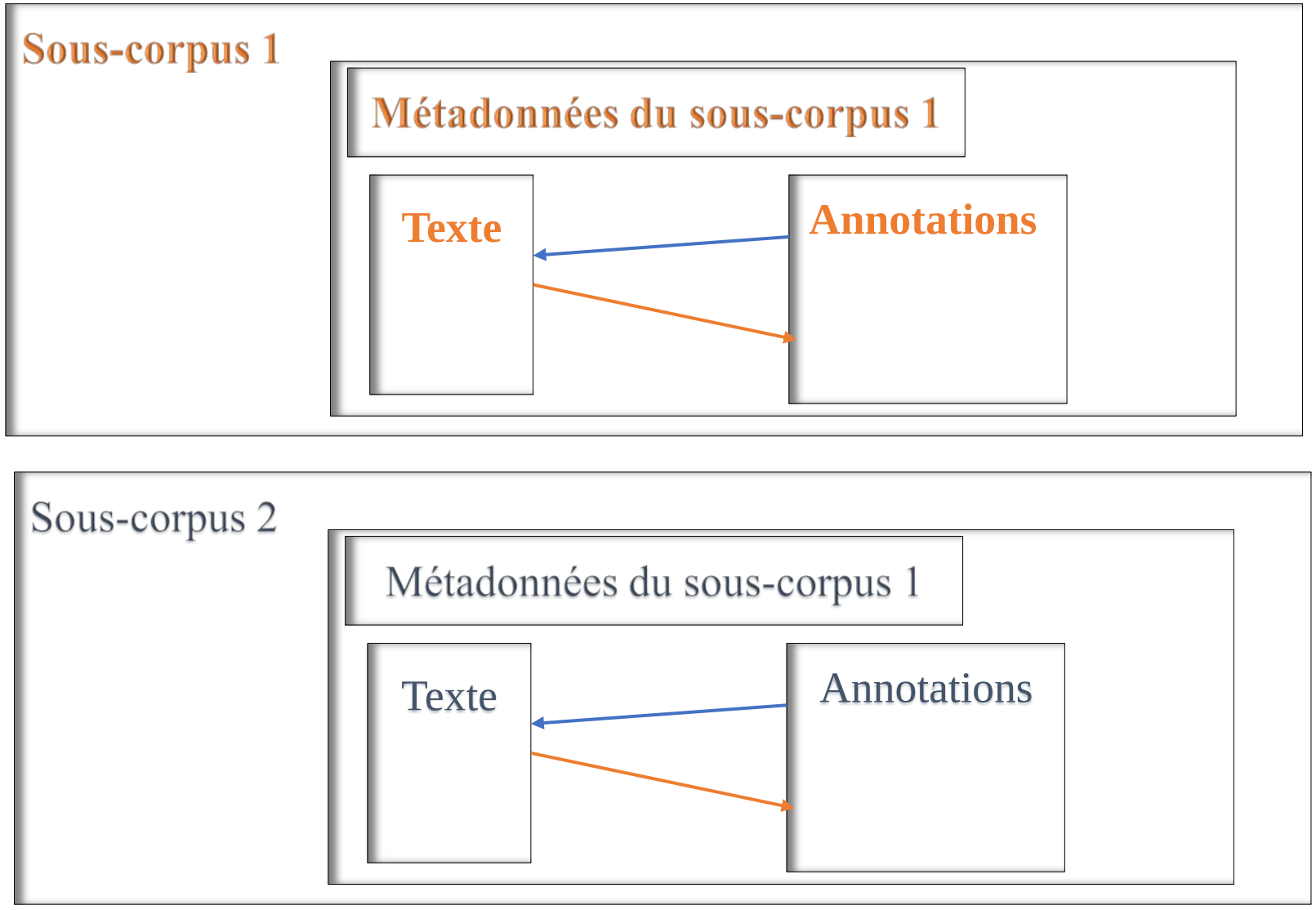
→ **Emoji** ☺☹

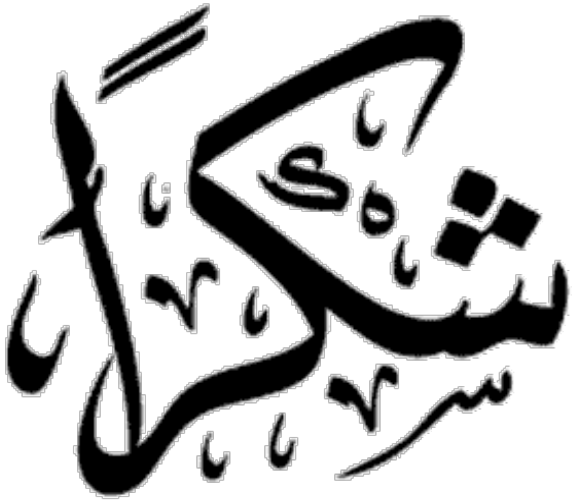
3- Nettoyage et normalisation des données

Selon le but de l'étude on peut décider quel type de normalisation peut-on appliquer et donc décider si un bruit est utile ou nuisible

3- Annotations

Métadonnées de l'ensemble du corpus (documentation)





Merci

Ouafae Nahli
Ouafae.nahli@ilc.cnr.it

References

Al Sharou, K., Li, Z., & Specia, L. (2021, September). **Towards a better understanding of noise in natural language processing**. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 53-62).

Choudhury, M., & Deshpande, A. (2021, May). **How Linguistically Fair Are Multilingual Pre-Trained Language Models?.** In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 14, pp. 12710-12718).

Doğruöz, A. S., & Sitaram, S. (2022). **Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights**. In Language Resources and Evaluation Conference (LREC) pages 92-97.

Dogruöz, A.S., Sitaram S., Bullock B. E., and Toribio A. J.. (2021). **A survey of code-switching: Linguistic and social perspectives for language technologies**. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1654–1666, Online. Association for Computational Linguistics.

HIRST, D. (2010). **Quand est-ce qu'un dialecte devient une langue ? La langue et l'être communiquant**. Hommage à Julio Murillo., Editions du CIPA. pp.179-190.

Michel, P., & Neubig, G. (2018). **MTNT: A testbed for machine translation of noisy text**. arXiv preprint arXiv:1809.00388.
Shoufan A. and Alameri S. (2015). **Natural Language Processing for Dialectical Arabic: A Survey**. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 36–48, Beijing, China. Association for Computational Linguistics.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & De Jong, F. (2016). **Computational sociolinguistics: A survey**. Computational linguistics, 42(3), 537-593.

Poudat, C., & Landragin, F. (2017). **Explorer un corpus textuel: Méthodes-pratiques-outils**. De Boeck Superieur.