



Université Sidi Mohamed Ben Abdellah

Faculté des Sciences Dhar El Mehraz – Fès

Master MLAIM
Probabilité et processus Markoviens

Analyse des Transitions en Langue Arabe

Encadré par :

Professeur . Hassan SATORI

Préparé par :

Ayat BOUHRIR
CNE:N13003366

Année Universitaire 2024-2025

I. Introduction

1.1 Contexte :

Le traitement automatique des langues (TAL) joue un rôle essentiel dans l'analyse et la modélisation des structures linguistiques complexes. La langue arabe, avec sa riche morphologie et son système unique de tachkils (diacritiques), représente un défi particulier dans ce domaine. La compréhension des transitions entre les lettres et les tachkils est cruciale pour diverses applications, notamment la génération automatique de textes vocalisés et la reconnaissance optique de caractères.

1.2 Objectifs du Projet :

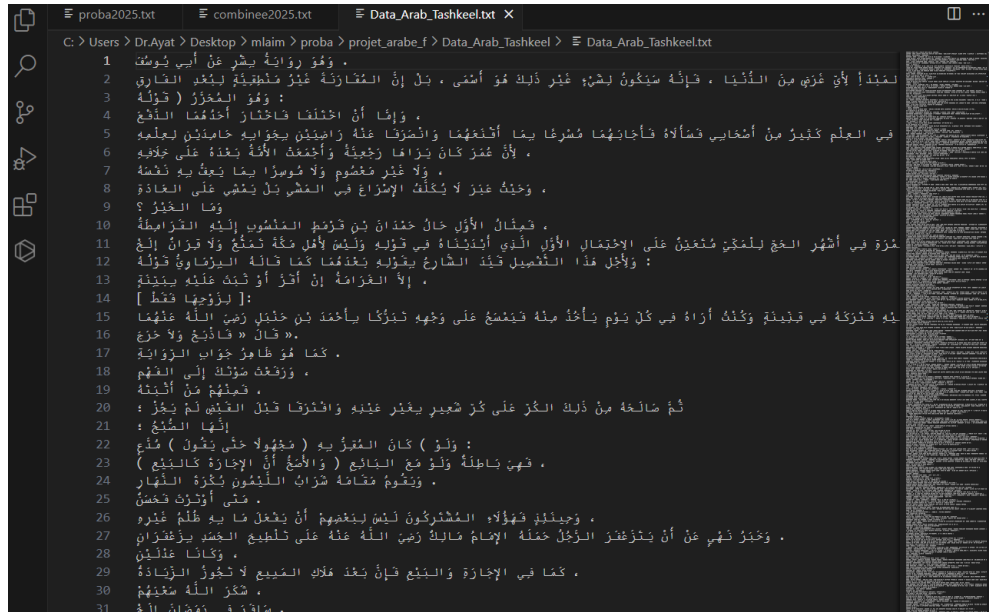
Ce projet vise à modéliser les relations entre les lettres et leurs tachkils dans des textes arabes. Les objectifs principaux incluent :

- Découper les phrases en mots et analyser chaque mot en paires successives de lettres et tachkils.
- Calculer les probabilités de transition pour chaque combinaison lettre-tachkil.
- Construire une matrice de transition des combinaisons possibles.
- Construire une matrice de stabilité
- Générer des fichiers texte et Excel pour visualiser et analyser les transitions et probabilités.

II. Méthodologie

2.1 Collecte de la Base de Données :

Un groupe de 13 personnes a été constitué pour collecter une base de données de 100 000 phrases en arabe, annotées avec les tachkils. Ces phrases proviennent de divers domaines afin d'assurer une couverture linguistique et stylistique étendue. La base de données a été stockée dans un fichier intitulé Data_Arab_Taskeel.txt, qui constitue la source principale pour l'analyse des transitions.



2.2 Découpage des Données :

Lors de cette étape, les phrases du fichier sont découpées en mots, puis chaque mot est divisé en lettres arabes et leurs tachkils associés. Les lettres sont extraites de l'alphabet arabe suivant : [أ, ب, ت, ث, ج, ح, خ, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق,]

[ك, ل, م, ن, ه, و, ي]. Chaque combinaison lettre-tachkil est ensuite enregistrée sous forme de paires successives.

Symbole	Nom	Position	Fonction	Exemple
-	فتحة	Au-dessus	Indique une voyelle courte (a)	كُتِبَ
ـُ	ضمة	Au-dessus	Indique une voyelle courte (u)	كُتِبَ
ـِ	كسرة	En dessous	Indique une voyelle courte (i)	كِتَابَ
◌	سكون	Au-dessus	Indique une absence de voyelle	يَكْتُبُ
ـّ	شدة	Au-dessus	Double la consonne (renforcement)	كُتِبْ
ـً	تنوين الفتح	Au-dessus	Indique un son nasal (an)	كِتَابًا
ـٍ	تنوين الضم	Au-dessus	Indique un son nasal (un)	كِتَابٌ
ـٍ	تنوين الكسر	En dessous	Indique un son nasal (in)	كِتَابِ

Mot: صحيحة

Lettres: ['ص', 'ح', 'ي', 'ج', 'ة']

Tachkil: ['FIN', ' ', ' ', ' ', ' ', ' ', 'DEBUT']

III. Calcul des Probabilités

3.1 Espace probabilisé :

Un espace probabilisé est une structure mathématique utilisée pour modéliser des expériences aléatoires. Il est défini par un triplet (Ω, \mathcal{F}, P) où :

- Ω (l'ensemble fondamental) : L'ensemble de tous les résultats possibles de l'expérience aléatoire. On l'appelle souvent l'espace des événements élémentaires.
- \mathcal{F} (la tribu) : Un ensemble de sous-ensembles de Ω , appelé une tribu, qui contient les événements (résultats mesurables). Elle doit respecter des propriétés spécifiques (fermeture par union, intersection et complémentaire).
- P (la probabilité) : Une fonction qui associe à chaque événement $A \in \mathcal{F}$ une valeur $P(A)$ telle que :
 - $P(A) \geq 0$ pour tout $A \in \mathcal{F}$,
 - $P(\Omega) = 1$
 - P est additive : $P(A \cup B) = P(A) + P(B)$ pour tout A, B disjoints.

Au niveau de notre modèle de travail on a :

- ✓ Ω : Les résultats élémentaires sont les paires successives de lettres et Tashkeel dans le texte analysé.
- ✓ F : L'ensemble des événements correspond aux transitions observées entre les paires (les sous-ensembles de l'espace d'états).
- ✓ P : La probabilité de transition est calculée comme :

$$P(X_n = j | x_m = i) = P_{ij}^{(m,n)}$$

3.2 Espace d'États :

L'espace d'états est l'ensemble de toutes les configurations possibles qu'un système peut prendre. Dans notre modèle, chaque état représente une paire de lettre et tachkil. L'espace d'états est donc défini par :

l'espace d'états est l'ensemble des paires successives composées de :

- Une lettre arabe (parmi les lettres de l'alphabet arabe).
- Un Tashkeel (parmi les signes diacritiques).

Chaque état représente une combinaison unique entre une lettre et, éventuellement, un Tashkeel, voici une représentation formelle de l'espace d'états :

$$S = \{(\text{lettre}, \text{tachkil}) \mid \text{lettre} \in \text{LETTRES}, \text{tachkil} \in \text{TACHKIL}\}.$$

3.3 Probabilité de Transition :

La probabilité de transition est un concept clé des chaînes de Markov. Elle mesure la probabilité qu'un système passe d'un état à un autre en une seule étape. Soit la chaîne de Markov et son espace d'états, la probabilité de transition entre deux états et est définie par :

Où :

est la probabilité de transition de l'état à l'état , est l'état du système à l'instant .

```

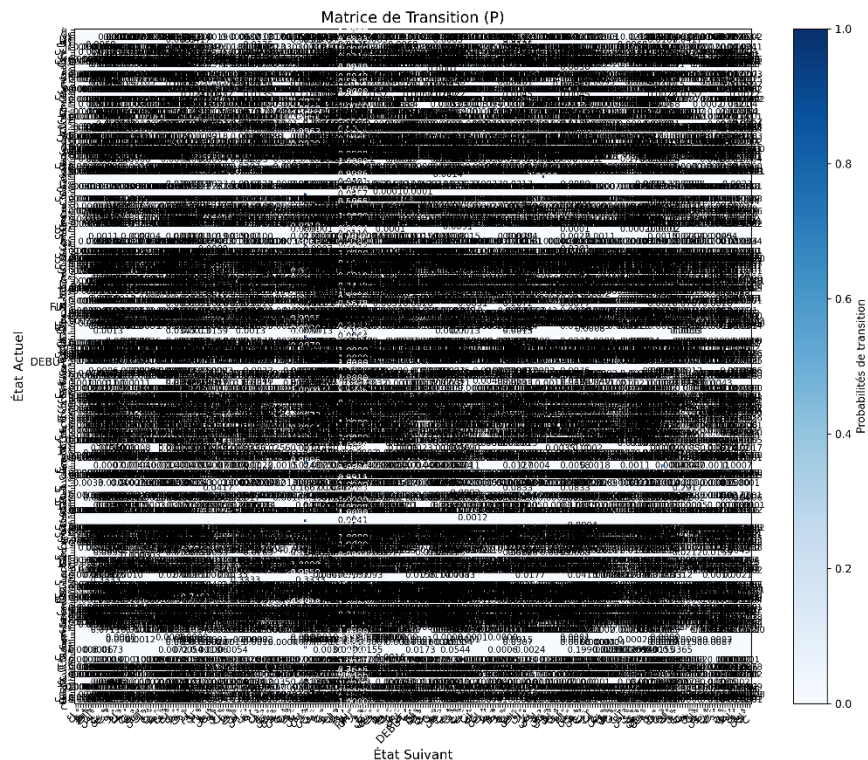
≡ proba2025.txt  ≡ combinee2025.txt  ≡ Data_Arab_Tashkeel.txt
C: > Users > Dr.Ayat > Desktop > mlaim > proba > projet_arabe_f > Data_Arab_Tashkeel > resultats > ≡ proba2025.txt
1 Probabilités de transition des paires successives:
73431 0.0000 :ج <- ا
73432 0.0000 :ج <- ا
73433 0.0000 :ب <- ا
73434 0.0000 :ب <- ا
73435 0.0000 :ع <- ا
73436 0.0000 :ج <- ا
73437 0.0000 :ف <- ا
73438 0.0000 :ا <- ا
73439 0.0000 :ن <- ا
73440 0.0000 :ج <- ا
73441 0.0000 :و <- ا
73442 0.0000 :ا <- ا
73443
73444 Somme totale des probabilités de transition: 270.0000
73445 Proba totale = 269.9999999999988 / 272 = 0.9926
73446

```

3.4 Matrice de transition :

The screenshot shows an Excel spreadsheet with a transition matrix. The matrix is a 29x29 grid where rows and columns are indexed by Arabic letters. The values represent the probability of transitioning from one letter to another. The diagonal elements are all 0, indicating no self-transitions. The off-diagonal elements are non-zero, representing the transition probabilities. The bottom row is labeled 'Matrice de Transition'.

3.5 Illustration des Probabilités de Transition dans une Matrice de Transition :



❖ Axes :

- L'axe vertical correspond aux états actuels du processus (labelé "État Actuel").
- L'axe horizontal correspond aux états suivants possibles (labelé "État Suivant").

❖ Valeurs dans la matrice :

- Chaque cellule de la matrice contient la probabilité de transition $P(i,j)$, où i est l'état actuel et j est l'état suivant.
- Ces valeurs sont des nombres entre 0 et 1, représentant la probabilité de passer de l'état i à j .

❖ Colorbar :

- Une échelle de couleurs est affichée à droite, indiquant les valeurs des probabilités. Les couleurs plus claires correspondent à des probabilités faibles (proche de 0), tandis que les couleurs plus foncées correspondent à des probabilités élevées (proche de 1).

❖ Interprétation

Cette matrice représente les dynamiques du système : elle montre comment il évolue d'un état à un autre.

- Les lignes et colonnes étiquetées (souvent des codes ou des noms d'états) indiquent l'ensemble des états du système.

- Une concentration de valeurs sombres dans une région de la matrice peut indiquer une forte probabilité de rester dans certains états ou de transitions préférentielles entre eux.

3.6 Modèle Associé à la Chaîne de Markov :

Une chaîne de Markov est un modèle mathématique décrivant un système évoluant d'un état à un autre, où chaque transition ne dépend que de l'état actuel, selon la propriété markovienne :

Dans notre modèle, une chaîne de Markov est associée aux transitions entre les paires de lettres et tachkils, représentant un système où chaque état correspond à une combinaison unique de lettre et tachkil

3.7 Matrice Stochastique

Une matrice stochastique est utilisée pour représenter les transitions entre les états dans une chaîne de Markov. Elle contient les probabilités de transition entre les états du système. Une matrice P est stochastique si elle satisfait les conditions suivantes :

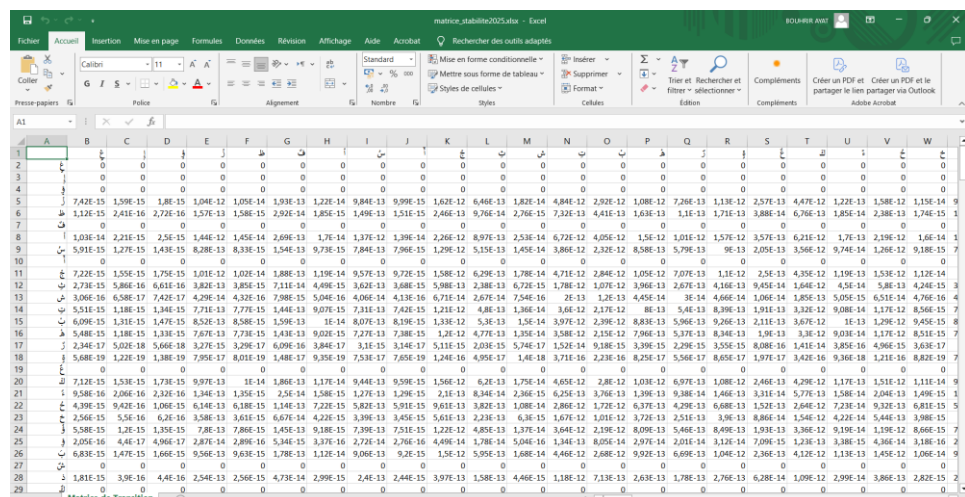
Les probabilités sont positives ou nulles,

La somme des probabilités dans chaque ligne est égale à 1.

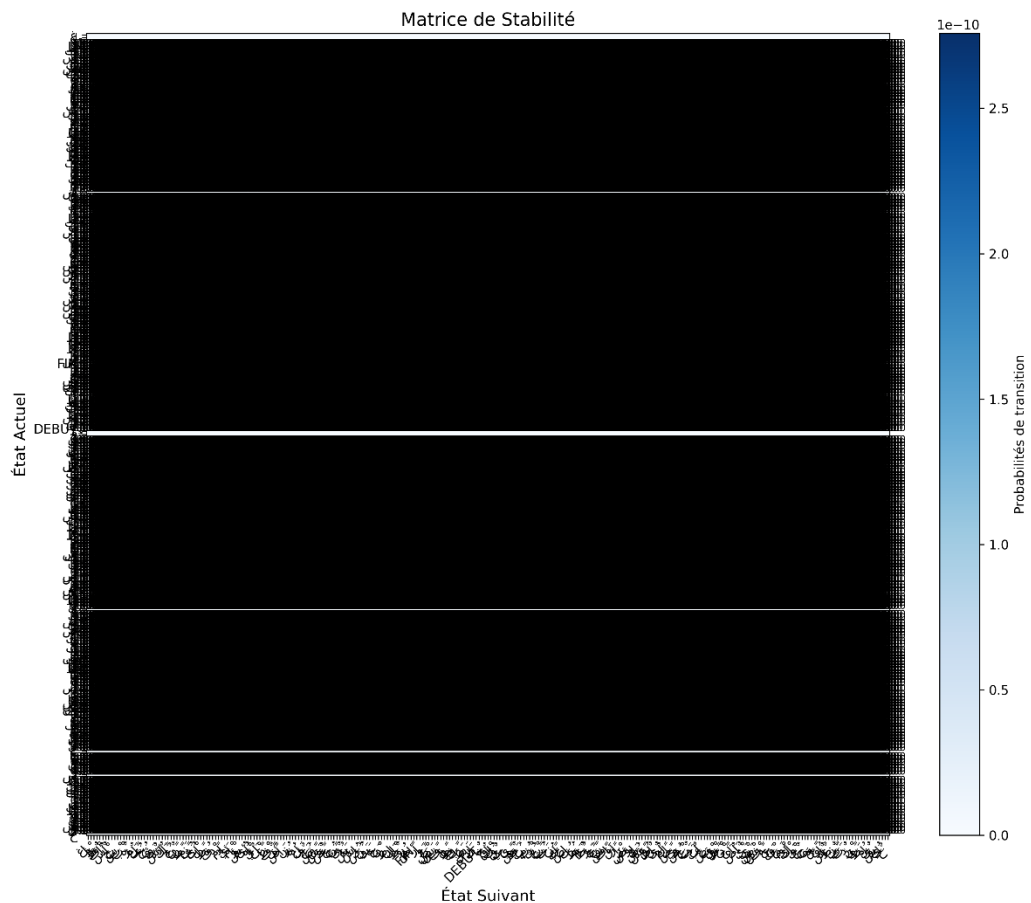
3.8 Matrice de Stabilité d'Ordre 2

La matrice de stabilité d'ordre 2 est une extension des matrices stochastiques, où les transitions entre états prennent en compte les deux états précédents successifs au lieu d'un seul. Dans une chaîne de Markov d'ordre 2, la probabilité de transition dépend des deux états précédents et, permettant de prédire :

Cette matrice permet d'étudier des transitions plus complexes, en tenant compte de l'historique plus large des états précédents. Elle est construite en calculant les probabilités pour toutes les combinaisons de trois états consécutifs.



3.9 Illustration des Probabilités de Transition dans une Matrice de Stabilité



❖ Axes :

- L'axe vertical indique les états actuels (les points de départ).
- L'axe horizontal montre les états suivants (les points d'arrivée possibles après une transition).

❖ Valeurs des probabilités :

- Chaque cellule de la matrice contient une valeur correspondant à la probabilité de transition d'un état actuel vers un état suivant.
- Ces probabilités sont très faibles, de l'ordre de 10^{-10} , ce qui reflète une situation où les transitions sont rares ou négligeables.

❖ Échelle de couleurs :

- ➔ La barre de couleurs à droite de l'image représente la magnitude des probabilités de transition. Les teintes plus foncées (proches du noir) correspondent à des probabilités proches de zéro, tandis que les teintes plus claires (bleues) correspondent à des valeurs légèrement plus élevées.

❖ Densité visuelle :

- L'image semble majoritairement sombre, ce qui indique que la plupart des transitions ont des probabilités très faibles ou nulles

IV. Le code

L'objectif du programme est de modéliser un texte en langue arabe en analysant les lettres et diacritiques (appelés « Tashkeel ») afin de construire une matrice de transition. Cette matrice est essentielle pour créer un modèle de chaîne de Markov, qui permet de modéliser les transitions entre différents états (lettres et diacritiques).

4.1 Bibliothèques utilisées :

os : Pour la gestion des fichiers et répertoires, comme la lecture des fichiers texte et la création de dossiers de sortie.

collections.defaultdict : Pour simplifier la gestion des dictionnaires en leur attribuant des valeurs par défaut.

openpyxl : Pour manipuler des fichiers Excel et enregistrer les matrices générées.

matplotlib.pyplot et numpy : Pour créer des représentations graphiques des matrices et manipuler les données sous forme de tableaux.

4.2 Fonctions principales :

4.2.1 Fonction analyse mot(mot) :

Objectif : Analyser un mot arabe pour extraire ses composants linguistiques (lettres, diacritiques, et combinaisons).

Détails du traitement :

Identifie les lettres arabes et les diacritiques (Tashkeel) dans le mot.

Forme des paires lettre + diacritique ou des lettres seules lorsque les diacritiques sont absents.

Retourne : Un dictionnaire avec trois clés :

lettres : Liste des lettres dans le mot.

tachkil : Liste des diacritiques associés.

paires : Liste des combinaisons lettre + diacritique (ou lettre seule).

4.2.2 Fonction calculer matrice transition(mots analyses) :

Objectif : Générer une matrice de transition représentant les probabilités de passage d'un état (paire) à un autre dans un texte.

Étapes du calcul :

Comptabiliser les transitions : Compte le nombre d'occurrences de chaque paire et des transitions entre elles.

Calculer les probabilités : Normalise les transitions pour obtenir une probabilité entre 0 et 1 pour chaque passage entre deux paires.

Retourne :

proba : La matrice des probabilités de transition.

etats : Les différents états (paires lettre + diacritique) présents dans le texte.

nombre_de_cas : Le nombre total de transitions analysées.

4.2.3 Fonction sauvegarder_combi_txt(mots analyses, chemin combin) :

Objectif : Sauvegarder les résultats de l'analyse des mots dans un fichier texte pour une consultation et une vérification ultérieures.

Contenu sauvegardé :

Les lettres extraites.

Les diacritiques associés.

Les transitions entre paires (sous forme de séquences "État actuel → État suivant").

Importance : Permet de documenter les données analysées et de s'assurer de leur cohérence.

4.2.4 Fonction sauvegarder_matrice_excel(matrice transition, etats, chemin excel) :

Objectif : Enregistrer la matrice de transition sous forme de tableau Excel.

Détails :

Les lignes représentent les états actuels (lettres ou paires).

Les colonnes représentent les états suivants, avec leurs probabilités de transition correspondantes.

Utilité : Offre une représentation structurée, facilitant l'analyse et l'interprétation des résultats.

4.2.5 Fonction sauvegarder_proba_txt(proba, nombre de cas, chemin proba) :

Objectif : Sauvegarder les probabilités de transition dans un fichier texte pour évaluer les résultats.

Fonctionnalités principales :

Documente les probabilités entre chaque paire et chaque transition.

Calcule et vérifie la cohérence des probabilités en s'assurant que la somme des probabilités est proche de 1.

Utilité : Assure une traçabilité complète des résultats pour validation ou réutilisation.

4.3 Conclusion :

Le programme présenté offre une solution robuste pour analyser les textes en langue arabe. Il identifie les transitions entre lettres et diacritiques, puis construit une matrice de transition utilisable dans des modèles de chaînes de Markov. De plus, il permet d'exporter et de visualiser les résultats dans différents formats (texte, Excel, graphique), ce qui facilite leur analyse et leur exploitation. Chaque fonction joue un rôle essentiel dans la chaîne de traitement, de l'analyse initiale des mots à la sauvegarde et à la visualisation des matrices.

V. Conclusion

Ce projet a permis d'explorer les particularités de la langue arabe, notamment ses lettres et ses diacritiques (tachkils), afin de comprendre et modéliser leurs relations. En partant d'une base de données diversifiée et en analysant les mots sous forme de combinaisons lettre-tachkil, nous avons construit des outils précieux comme des matrices de transition et de stabilité. Ces outils permettent de mieux appréhender les transitions entre les éléments d'un texte arabe, ouvrant ainsi la voie à des applications comme la vocalisation automatique ou la reconnaissance de caractères.

Le travail réalisé démontre l'importance d'une approche méthodique pour modéliser des langues complexes. Chaque étape, de l'analyse des mots à la visualisation des résultats, a contribué à fournir une vision claire et exploitable des dynamiques linguistiques.

Ce projet est avant tout un point de départ. Les résultats obtenus montrent un fort potentiel pour améliorer les technologies liées au traitement automatique de la langue arabe, mais aussi pour les appliquer à d'autres langues. En combinant expertise linguistique et outils techniques, nous pouvons relever les défis du TAL et continuer à développer des solutions utiles et innovantes.

VI. Bibliographie

1. *Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition."*
 - ➔ Introduction aux modèles de Markov cachés, essentiels pour la reconnaissance vocale et le traitement du langage.
2. *Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing.*
 - ➔ Une référence clé sur le traitement automatique des langues, couvrant les bases des modèles statistiques et linguistiques.
3. *Habash, N. Y. (2010). Introduction to Arabic Natural Language Processing.*
 - ➔ Livre dédié à la langue arabe, abordant ses spécificités, comme les diacritiques, et leurs défis dans le TAL.
4. *Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing.*
 - ➔ Ressource fondamentale pour comprendre les approches statistiques dans le traitement des langues.
5. *Corpus linguistiques arabes :*
 - ➔ Des sources ouvertes comme Leipzig et OpenSubtitles fournissent des bases de données textuelles en arabe utiles pour des analyses approfondies dans le site
<https://www.kaggle.com/datasets/hamzaabbad/tashkeela-processed-fully-diacritized-arabic-text/data>