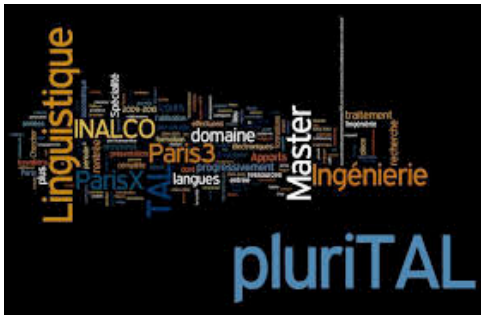


inalco
Institut national
des langues
et civilisations orientales

Rapport du projet

Translittération et traitement de corpus multilingue : anglais, chinois, hébreu, birman, grec

Sous la direction de Madame Ilaine WANG

CHEN Lian 陈恋

30 décembre 2024

Sommaire

Translittération et traitement de corpus multilingue : anglais, chinois, hébreu, birman, grec

CHEN Lian 陈恋

30 Décembre 2024

Table des matières

1 Introduction

Ce projet se concentre sur le traitement d'un corpus multilingue axé sur le changement climatique. L'objectif principal est d'analyser les caractères de différents systèmes d'écriture en utilisant des articles provenant de Wikipédia, rédigés en cinq langues distinctes : birman, hébreu, chinois, anglais et grec. Pour établir ce corpus, nous avons appliqué des critères spécifiques, notamment l'inclusion d'au moins deux langues inconnues et la nécessité de fichiers contenant au minimum 2000 caractères.

Le projet se découpe en plusieurs étapes. Tout d'abord, nous avons constitué le corpus en récupérant le contenu des articles choisis, puis nous avons mis en place une analyse détaillée des caractères, en nous concentrant sur la segmentation, l'analyse statistique et la relation entre graphèmes et phonèmes. Enfin, la translittération des systèmes d'écriture vers l'alphabet latin a été réalisée pour faciliter l'interprétation des résultats. Ces analyses visent à améliorer notre compréhension des variations linguistiques et des structures phonétiques à travers différentes langues, tout en fournissant des insights sur l'usage du français dans un contexte multilingue.

2 Constitution du corpus

Notre corpus se compose de cinq articles Wikipédia rédigés en cinq langues différentes, chacune utilisant un type de système d'écriture distinct :
Birman (système alphasyllabaire)
Hébreu (système consonantique, ou abjad)
Chinois (système logographique/syllabique)
Anglais (système alphabétique)
Grec (système alphabétique)

Pour constituer notre corpus, nous avons respecté plusieurs critères. Tout d'abord, parmi les cinq langues choisies, il était nécessaire qu'au moins deux nous soient totalement inconnues. De plus, tous les articles devaient porter sur le même thème, à savoir le changement climatique. Enfin, chaque fichier devait contenir au moins 2 000 caractères, espaces exclus.
Pour récupérer le texte de ces pages, nous avons utilisé le script Python du professeur Ilaine Wang, puis nous l'avons adapté nous-mêmes.

```
import requests
from bs4 import BeautifulSoup
import re
# page request
response = requests.get(
    url="https://en.wikipedia.org/wiki/Climate_change"
)
# Parsing page content
soup = BeautifulSoup(response.content, 'html.parser')
# Get all paragraphs, unordered list and ordered list
allParagraphs = soup.find(id="bodyContent").find_all(["p", "ul", "ol"])
text = ''
for item in allParagraphs:
    # Remove possible numeric marks from labels
    clnItem = re.sub(r'\[\d+\]', '', item.get_text())
    text += clnItem + '\n'  # 每个条目后加换行符
#print(text)
# 将The result is written to a text file
with open("wiki_changement_climatique_my.txt", "w", encoding='utf-8') as file:
    file.write(text)
```

Fig. 1 : Capture d'écran du script de scraping pour extraire du texte en ligne

Le texte de chaque langue est enregistré sous les noms suivants :

wiki$_{c}hangement_{c}limatique_{m}y.txt(pourlebirman)$

$wiki_{c}hangement_{c}limatique_{h}e.txt(pourl'hbreu)$

$wiki_{c}hangement_{c}limatique_{z}h - Hans.txt(pourlechinois)$

$wiki_{c}hangement_{c}limatique_{e}n.txt(pourl'anglais)$

$wiki_{c}hangement_{c}limatique_{e}l.txt(pourlegrec)$

Ensuite, nous utilisons la commande suivante :

grep -o '.' wiki$_{c}hangement_{c}limatique_{z}h - Hans.txt$

Cette commande nous permet d'extraire chaque caractère du fichier, et les résultats sont enregistrés dans les documents suivants :

wiki$_{t}heme_{b}ir3_{n}ew$

$wiki_{t}heme_{c}hi3_{n}ew$

$wiki_{t}heme_{e}n3_{n}ew$

$wiki_{t}heme_{g}rec3_{n}ew$

$wiki_theme_hebreux3_new$

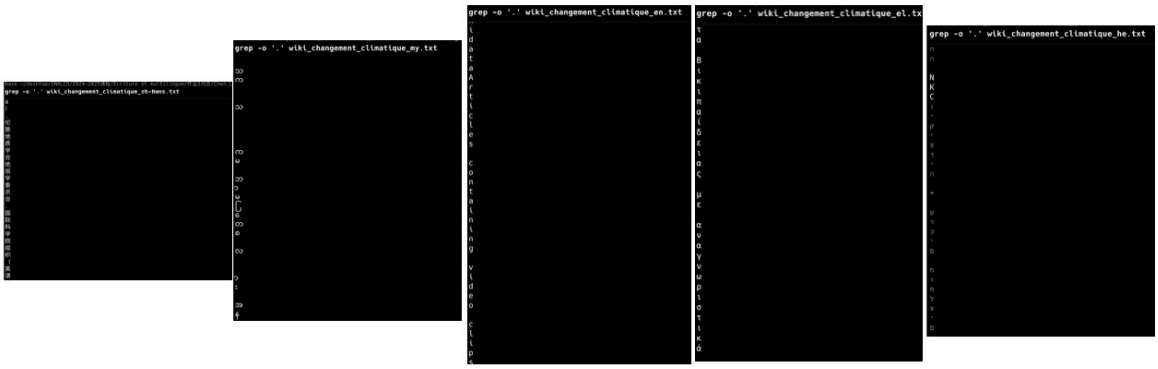


Fig. 2 : Capture d'écran des résultats de segmentation du texte dans ces cinq langues

Comme le montrent les images ci-dessus, en chinois, chaque caractère est séparé, y compris les signes de ponctuation, qui sont également considérés comme des caractères distincts. Il en va de même pour le birman, où chaque caractère est segmenté ligne par ligne. En revanche, en anglais, en grec et en hébreu, chaque lettre est traitée individuellement.

3 Analyse des caractères

3.1 Segmentation et analyse statistique

La première étape de cette partie consiste en la segmentation du fichier obtenu précédemment en caractères (non-mots compris). Pour ce faire, nous avons utilisé la commande (modifié selon pour chaque langue) :

grep -o '.' wiki$_c hangement_c limatique_m y.txt > wiki_t heme_b ir3_n ew$

Les documents sont nommés suivants :
wiki$_t heme_b ir3_n ew$
$wiki_t heme_c hi3_n ew$
$wiki_t heme_e n3_n ew$
$wiki_t heme_g rec3_n ew$
$wiki_t hewme_h ebreux3_n ew$

Pour pouvoir faire une analyse statistique des caractères, nous avons ensuite un tri :

1) **Première étape : trier la liste de caractères**

Dans cette étape, on utilise la commande sort pour organiser les caractères de manière ordonnée. La commande **sort wiki_theme_bir3_new.txt > sorted_wiki_theme_bir3_new.txt** permet de trier les caractères contenus dans le fichier **wiki_theme_bir3_new.txt** et d'enregistrer le résultat trié dans un nouveau fichier intitulé **sorted_wiki_theme_bir3_new.txt**. Comme le montrent les illustrations suivantes :

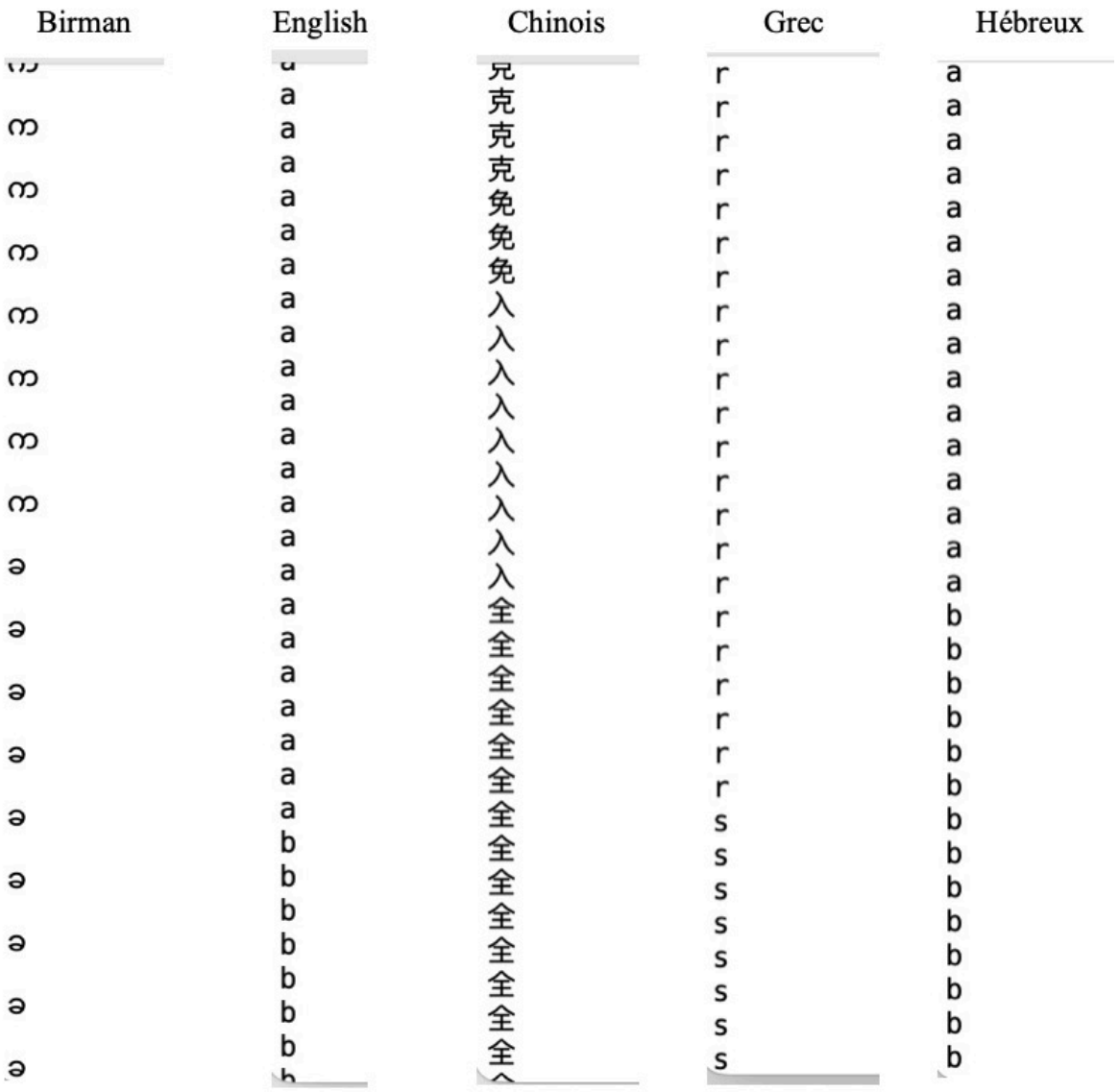
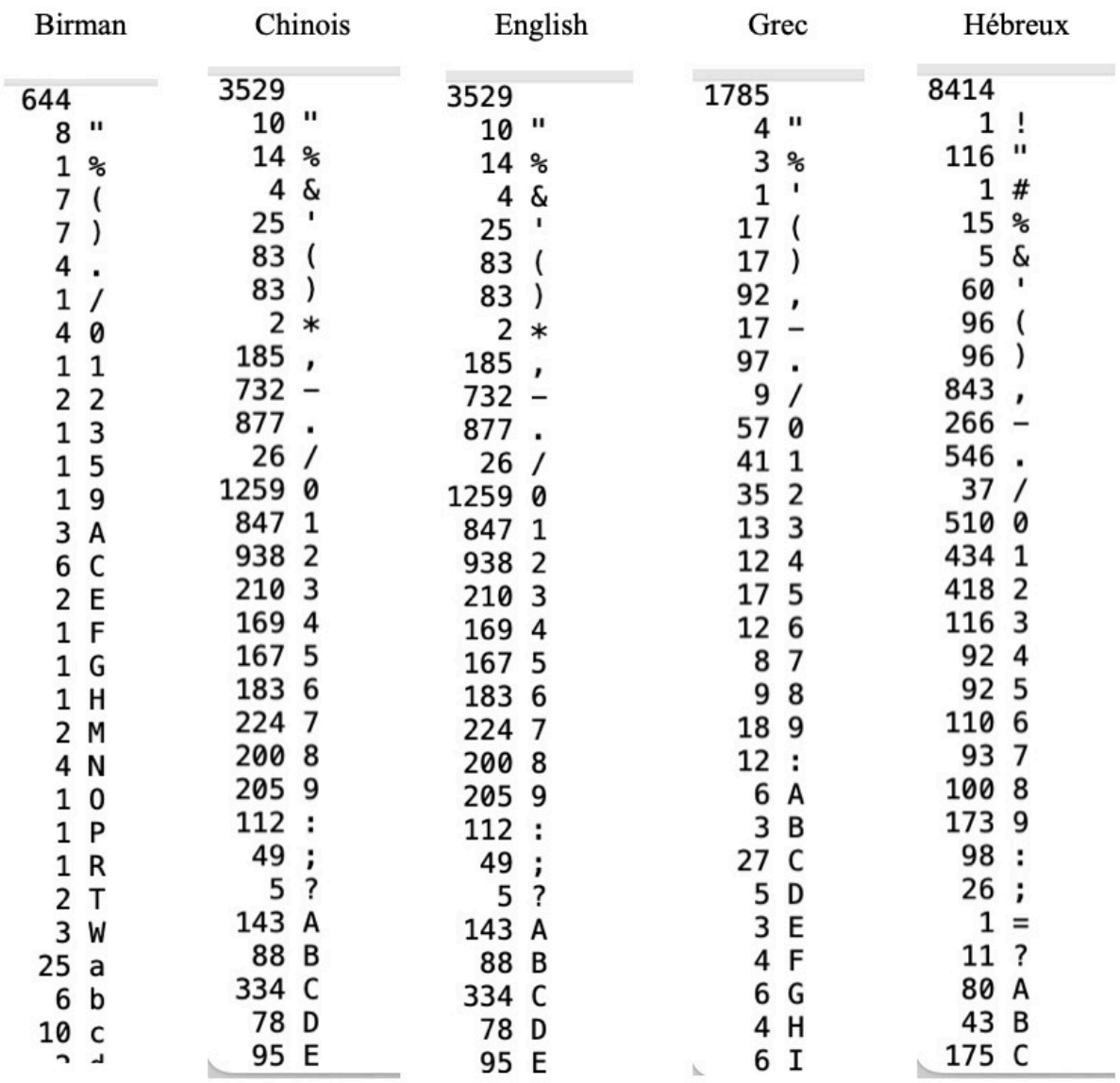


Fig. 3 : Captures d'écran de la liste des caractères pour chaque langue

2) **Deuxième étape : compter les caractères**

Après avoir trié la liste, nous utilisons la commande **uniq -c** pour compter l'occurrence de chaque caractère dans le fichier trié. La commande **uniq -c sorted_wiki_theme_bir3_new.txt > counted_sort** (où "bir3" est remplacé selon la langue) génère un fichier qui répertorie le nombre d'occurrences de chaque caractère, en l'enregistrant sous le nom de **counted_sorted_wiki_theme_bir3_new.txt**.

Voici les résultats :



Birman	Chinois	English	Grec	Hébreux
644	3529	3529	1785	8414
8 "	10 "	10 "	4 "	1 !
1 %	14 %	14 %	3 %	116 "
7 (	4 &	4 &	1 '	1 #
7)	25 '	25 '	17 (	15 %
4 .	83 (	83 (	17)	5 &
1 /	83)	83)	92 ,	60 '
4 0	2 *	2 *	17 -	96 (
1 1	185 ,	185 ,	97 .	96)
2 2	732 -	732 -	9 /	843 ,
1 3	877 .	877 .	57 0	266 -
1 5	26 /	26 /	41 1	546 .
1 9	1259 0	1259 0	35 2	37 /
3 A	847 1	847 1	13 3	510 0
6 C	938 2	938 2	12 4	434 1
2 E	210 3	210 3	17 5	418 2
1 F	169 4	169 4	12 6	116 3
1 G	167 5	167 5	8 7	92 4
1 H	183 6	183 6	9 8	92 5
2 M	224 7	224 7	18 9	110 6
4 N	200 8	200 8	12 :	93 7
1 O	205 9	205 9	6 A	100 8
1 P	112 :	112 :	3 B	173 9
1 R	49 ;	49 ;	27 C	98 :
2 T	5 ?	5 ?	5 D	26 ;
3 W	143 A	143 A	3 E	1 =
25 a	88 B	88 B	4 F	11 ?
6 b	334 C	334 C	6 G	80 A
10 c	78 D	78 D	4 H	43 B
? ?	95 E	95 E	6 I	175 C

Fig. 4 : Illustration des occurrences de chaque caractère dans ces cinq langues

3) **Troisième étape : trier par fréquence**

Dans cette dernière étape, on trie les caractères selon leur fréquence d'apparition. La commande **sort -rg counted_sorted_wiki_theme_bir3_new.txt > wiki_ecriture_bir3.txt** permet d'effectuer ce tri en ordre décroissant. Le résultat final est enregistré dans un fichier nommé **wiki_ecriture_bir3.txt**.

Birman	Chinois	English	Hébreux	Grec
1172 ်	142 [	28246	8414	1785
644	138 温	17096 e	י 3870	1000 α
625 ာ	137 球	11626 a	ו 3794	778 τ
611 ိ	137 ^	10315 i	ה 2816	703 ι
449 ု	135 容	10242 t	מ 2199	690 ο
444 c	128 変	10129 n	ת 2164	668 ε
427 ေ	126 地	9524 o	ל 2060	547 ν
413 ွ	124 P	9181 r	ר 1893	487 η
383 ဖ	120 度	7501 s	א 1523	473 ρ
374 ာ	118 大	6047 l	ב 1465	461 σ
357 ာ	112 :	5975 .	ש 1363	442 ς
292 ်	111 国	5479 c	נ 1085	414 κ
283 ာ	106 候	4579 h	ע 1048	384 μ
277 ပ	105 全	4346 d	ם 1028	383 π
259 ်	105 中	3970 m	ת 919	328 λ
250 ု	99 是	3700 2	פ 859	271 υ
227 ဝ	99 R	3548 0	843 ,	224 ί
	95 E	3381 u	840 e	208 ό
	92 人	3272 1	ק 816	204 ά
	92 F	3170 g	ד 805	198 έ
	90 T	2641 p	כ 647	166 γ
	88 B	2422 ,	615 a	165 θ
	84 G	2367 C	ט 586	158 ή
	83)	2324 f	586 i	143 ω
	83 (	2053 b	570 n	142 δ
	78 D	1910 y	546 .	97 .
	77 学	1591 v	ג 527	92 ,
		1456 S	510 0	90 χ
		1256 P	494 t	79 ύ
		1236 A	ס 484	73 ώ

Fig. 5 : Illustration du classement des caractères selon leur fréquence d'apparition dans ces cinq langues.

Nous analysons ces corpus en répondant les questions suivantes :

a) **La segmentation en caractères pour ces langues est-elle aisée ?**

Nous remarquons qu'en grec et en anglais, chaque lettre de chaque mot est séparée et les mots sont délimités par des espaces, parfois accompagnés de signes de ponctuation (comme les virgules et les points). En revanche, en chinois, les signes de ponctuation et les caractères sont également distincts. Comme déjà souligné, chaque caractère chinois est clairement séparé, y compris les signes de ponctuation, qui sont considérés comme des caractères autonomes. La segmentation des caractères chinois est relativement simple grâce à la nature du système d'écriture, où chaque caractère représente généralement un mot ou un concept unique. Ainsi, bien qu'il n'y ait pas d'espaces entre les mots, la segmentation reste facile tant que l'on dispose des caractères appropriés pour l'analyse.

En revanche, la situation semble moins claire pour le birman. En tant que personne n'ayant pas de connaissances en birman, je ne peux pas garantir que la segmentation soit correcte. Il m'est difficile de comprendre quand il faut insérer des espaces, et je ne suis pas sûr si un caractère est séparé ou non. D'après mes recherches sur Wikipédia (voir`https://fr.wikipedia.org/wiki/%C3%89criture_birmane`, bien que chaque caractère birman soit généralement une unité autonome, certains caractères composés peuvent représenter des combinaisons de sons, ce qui complique la segmentation.

Par ailleurs, la commande **grep -o '.'** a été utilisée pour extraire chaque caractère comme une ligne distincte, rendant ainsi cette tâche relativement simple. Cependant, cette méthode ne prend pas en compte les nuances grammaticales et syntaxiques propres à la langue birmane, qui peuvent poser des défis supplémentaires lors de l'analyse.

Dans l'ensemble, bien que la segmentation en caractères soit généralement simple dans ces cinq langues, des complexités peuvent surgir, en particulier avec le birman. Cela peut rendre la lecture du texte moins fluide et moins agréable.

b) **Avez-vous rencontré des problèmes ? Si oui, les avez-vous résolus et comment ?**

Des problèmes peuvent se manifester en raison de la diversité des caractères spéciaux et des glyphes présents dans les écritures birmane et chinoise. Par exemple, pour birman, certains caractères peuvent posséder des variantes de forme ou des diacritiques qui modifient leur prononciation. Pour surmonter ces défis, il est crucial de bien gérer les encodages en utilisant des outils appropriés afin d'assurer un traitement correct des données, notamment en vérifiant que les fichiers textes sont codés en UTF-8 un encodage largement utilisé pour le traitement de caractères multilingues.). Nous avons employé le script suivant :

```
# Lire le fichier
with open('/Users/lianchen/Desktop/INALCO/2024-2025课程/Ecriture et multilingue/作业1月交/Chen_Lian/
wiki_theme_bir3_new', 'r', encoding='utf-8') as file:
    content = file.read()
# Écrire dans le fichier
with open('/Users/lianchen/Desktop/INALCO/2024-2025课程/Ecriture et multilingue/作业1月交/Chen_Lian/
wiki_theme_bir3UTF_new.txt', 'w', encoding='utf-8') as file:
    file.write(content)
```

Fig. 6 : Script pour encoder des fichiers en UTF-8

Les problèmes potentiels peuvent inclure des caractères non reconnus ou mal encodés, surtout si le texte contient des caractères moins courants ou des symboles spécifiques. Pour résoudre de tels problèmes, il serait utile de s'assurer que le fichier est encodé avec UTF-8, d'utiliser des bibliothèques adaptées au traitement de caractères chinois, et de vérifier les résultats avec un affichage correct des caractères.

c) **Sans compter l'espace, quels sont les 10 caractères les plus fréquents et les moins fréquents pour les 5 langues que vous avez choisies ? Dans quelle mesure ces statistiques sommaires vous informent-elles (ou pas) sur ces langues ou ces écritures ?**

	Birman	Chinois	Grec	Anglais	Hébreux
Caractère le plus fréquents	1172 ◌ွ 644 625 ◌ာ 611 ◌း 449 ◌့ 444 ဂ 427 ဖ◌ 413 ◌ြ 383 ဖ 374 ◌ာ	498 的 240 英 229 语 208 气 194 于 192 存 182 档 160 年 160 化 150 原	1000 α 778 τ 703 ι 690 o 668 ε 547 ν 487 η 473 ρ 461 σ 442 ς	17096 e 11626 a 10315 i 10242 t 10129 n 9524 o 9181 r 7501 s 6047 l 5975 .	387 ׳ 3794 ו 2816 ה 2199 מ 2164 ת 2060 ל 1893 ר 1523 א 1465 ב 1363 ש 1085 נ 1048 ע
Caractères les moints fréquents	1 O 1 H 1 G 1 F 1 9 1 5 1 3 1 1 1 / 1 %	1 乏 1 乌 1 举 1 」 1 「 1 ’ 1 ‘ 1 ú 1 é 1 |	1 — 1 Z 1 Ή 1 ± 1 z 1 _ 1 V 1 U 1 K 1 '	2 ± 2 | 2 > 2 = 1 ● 1 → 1 1 ć 1 ã 1 <	1 , 1 _ 1 ~ 1 é 1 á 1 _ 1 Q 1 = 1 # 1 !

Fig. 7 : les 10 caractères les plus fréquents et les moins fréquents pour les 5 langues

Pour les moins fréquents, cela peut inclure des caractères comme les chiffres et les ponctuations.

d) **Les signes de ponctuation sont-ils les mêmes qu’ en français ?**

Dans notre analyse, il semble que les signes de ponctuation en anglais et en grec soient similaires à ceux du français. Concernant l’hébreu, cette langue utilise également des signes de ponctuation similaires, tels que les virgules et les points, mais leur emploi et leur position peuvent varier. Dans les textes hébreux, certains signes de ponctuation peuvent apparaître à des emplacements inversés par rapport à leur usage en français, ce qui est dû à la direction de l’écriture.

Par ailleurs, les signes de ponctuation utilisés en chinois diffèrent de ceux employés en français. Par exemple, la virgule (，) et le point (。) sont utilisés de manière distincte en termes de position et de fréquence par rapport à l'usage français. De plus, les guillemets (« » ou “ ”) présentent des différences tant dans leur forme que dans leur utilisation.

3.2 Relation graphème/phonème

Le travail effectué à l' étape 1 nous a aidé à identifier les caractères les plus fréquents et les moins fréquents dans un système d' écriture, basé sur un article Wikipédia. Cependant, cela ne montre pas comment ces caractères sont utilisés et leur valeur phonémique dans la langue. Nous avons donc tester le plateforme Wikicolor `https://wikicolor.alem-app.fr`. Cet outil utilise un code couleur pour mettre en évidence les graphèmes d' un mot et leur correspondance phonétique. Cela aide à visualiser comment les lettres se prononcent dans différents mots.

a) **Testez et commentez le travail de l' outil WikiColor**

Nous avons choisi la langue française et chinoise pour testet.
Nous avons d'abord tester cinq mots "Wikicolorisés" : WikiColor montre la relation entre l' orthographe et la pronunciation. Par exemple :

La plateforme Wikicolor utilise un code couleur pour représenter visuellement la correspondance entre l' orthographe des mots et leur prononciation phonétique, en mettant en évidence les graphèmes. Les graphèmes sont les unités de son ou de graphe qui constituent un mot, et leur couleur permet de mieux comprendre les variations de prononciation. Voici un aperçu de la relation couleur-graphèmes, ainsi que des exemples que vous avez testés :
Exemples de mots wikicolorisés :

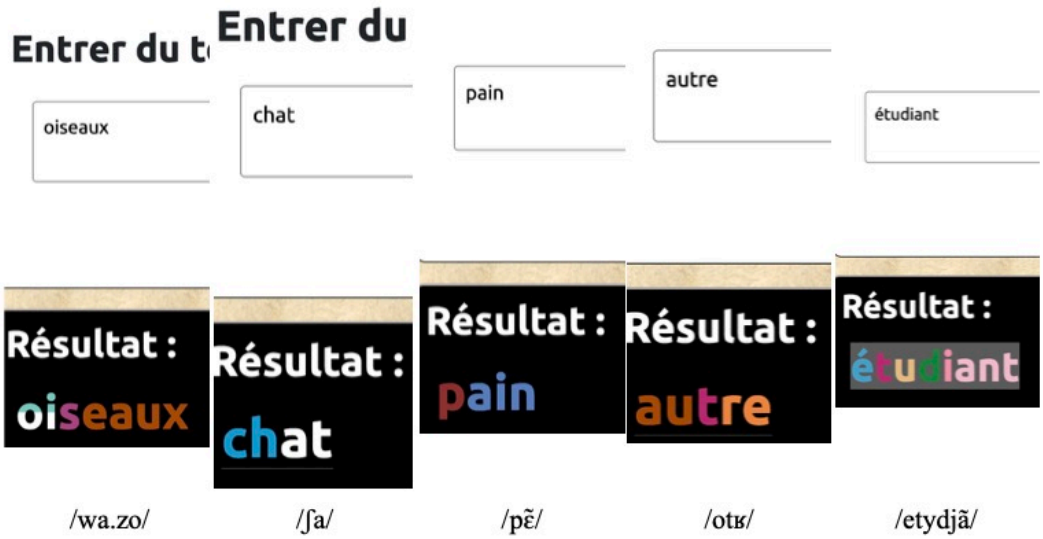


Fig. 8 : Exemples de mots wikicolorisés

Wikicolor facilite la compréhension de la phonétique en associant chaque graphème à une couleur spécifique, permettant ainsi aux utilisateurs de visualiser comment les lettres et les sons interagissent dans le français. Ce système est particulièrement utile pour l'apprentissage de la langue et l'enseignement de la prononciation correcte des mots.

Ensuite, nous avons testé le mandarin.

Nous avons d'abord sélectionné le chinois parmi les trois langues disponibles en haut à droite : chinois, anglais et français. Ensuite, nous avons testé des phrases simples telles que 你好 (Nǐ hǎo) pour «bonjour», 老师 (lǎoshī) pour «enseignant», et 今天 (jīntiān) pour «aujourd'hui». Nous avons constaté que les caractères chinois sont colorés, cependant, nous n'avons pas réussi à identifier de lien entre ces couleurs et le rythme des mots, car aucune explication n'était fournie.

D'après les résultats, nous avons noté que les caractères 老 (lǎo, «vieux») et 好 (hǎo, «bon») contiennent chacun trois sons, ce qui explique pourquoi ils sont associés à trois couleurs différentes : bleu, blanc et vert. En revanche, pour 今天 (jīntiān, «aujourd'hui»), bien qu'il se compose également de deux premiers sons, les couleurs qui lui sont attribuées semblent différentes.

Nous avons donc réessayé d' analyser leur pinyin et avons remarqué que tous les éléments étaient affichés en blanc. Cela indique qu'il y a une marge d'amélioration pour le système Wikicolor pour le chinois, afin d'assurer une meilleure correspondance entre les couleurs et les

phonèmes.

Fig. 9 : Exemples de mots en chinois wikicolorisés

L’interface de Wikicolor est très confortable et claire en un coup d’œil. WikiColor est un outil intéressant qui vise à illustrer la relation entre les graphèmes (les unités écrites, comme les lettres) et les phonèmes (les unités de son, comme les sons) dans différentes langues.

Fig. 10 : Exemples de pinyin chinois wikicolorisés

WikiColor utilise des couleurs pour représenter les différentes parties d'un mot, ce qui permet une visualisation immédiate des relations entre l'écriture et la prononciation. Cela peut être particulièrement utile pour les apprenants de langues, car cela facilite la compréhension des correspondances entre les lettres et les sons. En intégrant la transcription phonétique (API), WikiColor aide les utilisateurs à saisir non seulement comment un mot s'écrit, mais aussi comment il se prononce, ce qui est essentiel pour maîtriser correctement une langue, surtout celles avec des nuances de prononciation comme le français, l'anglais ou le mandarin. Cet outil peut potentiellement être utilisé pour diverses langues（français, anglais, chinois), ce qui le rend polyvalent.

Néanmoins, Pour des langues comme le mandarin, où les tonalités jouent un rôle crucial dans la signification, WikiColor peut rencontrer des défis lorsqu'il s'agit de représenter ces nuances. On a vu que dans les tests, l'outil a eu des difficultés à traiter correctement les tonalités, ce qui pourrait fausser l'apprentissage. Comme nous l'avons mentionné, des erreurs peuvent survenir, telles que des éléments affichés en rouge ou des symboles incorrects comme un "x". Ces problèmes techniques peuvent nuire à l'expérience utilisateur et nécessiter une attention continue pour garantir la fonctionnalité de l'outil. Aussi, La compréhension des graphèmes et phonèmes dépend aussi du contexte dans lequel ces mots sont utilisés. L'outil doit donc être accompagné d'explications sur les variations possibles en fonction du contexte pour être totalement efficace.

b) **Un tel outil serait-il utile pour la langue inconnue que vous avez choisie pour le système alphabétique ?**

Un tel outil serait indéniablement utile pour l'apprentissage de la langue grecque, notamment en raison de ses particularités orthographiques et phonétiques. Par exemple, en grec, chaque lettre correspond à un son spécifique, mais certaines lettres peuvent avoir des prononciations différentes selon leur position dans un mot. La lettre «γ» (gamma) est généralement prononcée comme un «g» dur, mais lorsqu'elle est suivie des lettres «ε» ou «ι», elle se prononce comme un «y» doux (ex : «γένος» /□□en.os/) (voir https://fr.wikipedia.org/wiki/Alphabet_grec). Un outil qui utilise des couleurs pour représenter ces variations de prononciation permettrait aux apprenants de visualiser facilement ces différences et d'associer chaque graphème à son phonème approprié.

Les accents jouent un rôle crucial dans la prononciation et la signification des mots. Par exemple, le mot «πού» (pou, où) a un accent tonique, tandis que «που» (pu, que), sans accent, peut changer de signification selon le contexte. Un outil qui met en évidence les accents à l'aide

de couleurs pourrait aider les apprenants à comprendre l'importance de la prosodie et à prononcer correctement les mots en fonction de leur accentuation.

Le grec contient également de nombreuses diphtongues, telles que « αι » (ai) et « ει » (ei), qui peuvent être difficiles à maîtriser. En associant des couleurs spécifiques à ces diphtongues, un outil comme WikiColor pourrait aider les apprenants à les reconnaître et à les prononcer correctement. Par exemple, dans le mot « καί » (kai, et), le son « αι » pourrait être mis en évidence d'une manière qui aide l'apprenant à visualiser la transition entre les sons.

Ainsi, un outil comme WikiColor serait extrêmement bénéfique pour l'apprentissage du grec, car il offrirait une approche visuelle et interactive qui facilite la compréhension des relations entre l'écriture et la prononciation, tout en prenant en compte les spécificités phonétiques et orthographiques de la langue.

4 Translittération vers l' alphabet latin

Selon l'Organisation internationale de normalisation (ISO), la translittération est un « procédé qui consiste à représenter les caractères d'un système d'écriture alphabétique ou syllabique à l'aide des caractères d'un alphabet de conversion ». La dernière phase de ce projet vise à étudier la question de la translittération pour l'un des systèmes d'écriture que nous avons sélectionnés, parmi lesquels figurent : le système alphabétique inconnu, le système consonantique et l'alpha-syllabaire. Nous avons opté pour la langue grecque et, par conséquent, Nous utiliserons le script suivant pour le transmettre et l'enregistrer sous le nom : **wiki_ecriture_grec3_translit.txt**.

Fig. 11 : Script pour la translittération de la langue grecque.

Voici les résultats :

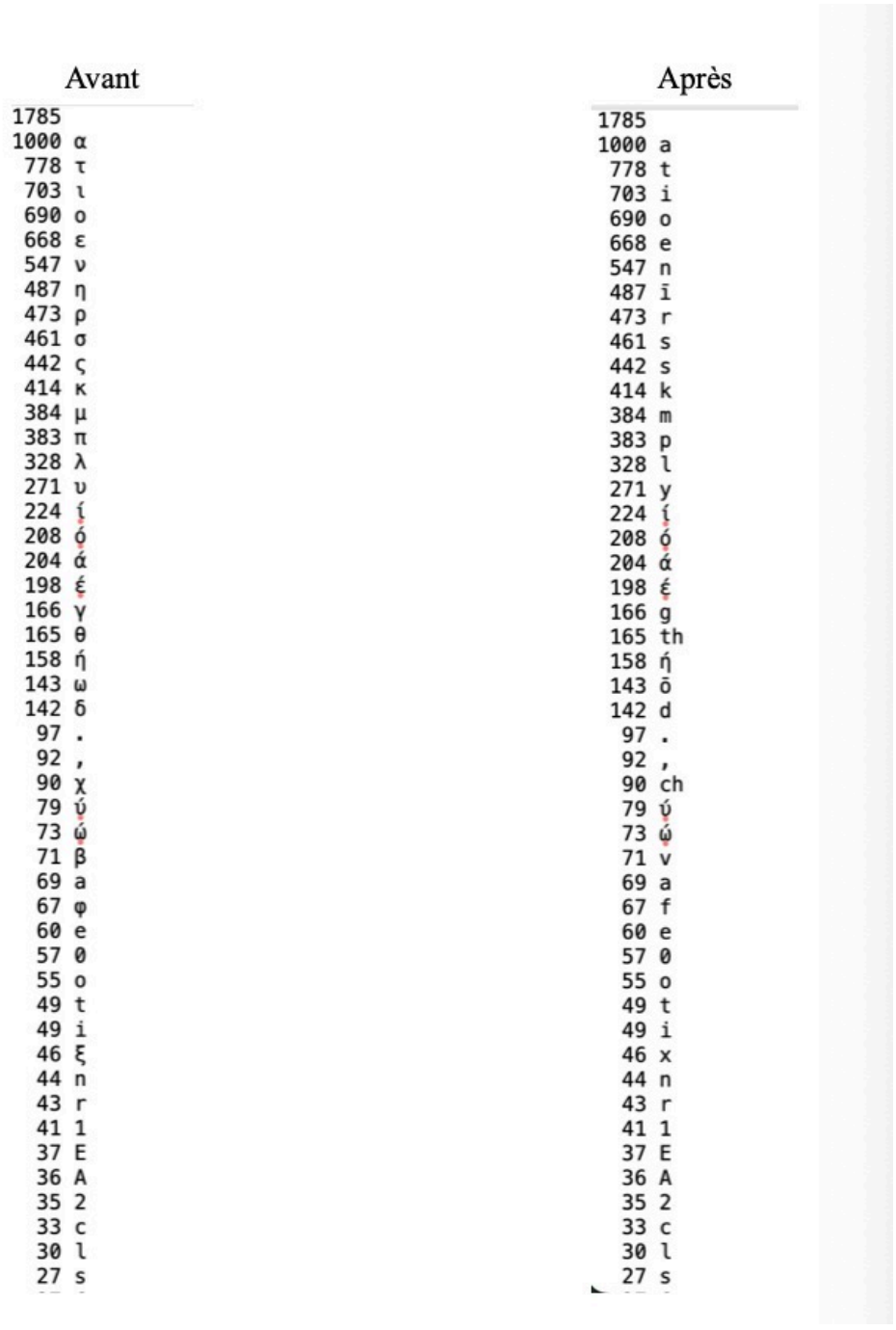


Fig. 12 : Résultats de la translittération de la langue grecque

5 Conclusion

Ce projet de traitement de corpus multilingue a permis d'explorer en profondeur les spécificités des systèmes d'écriture de cinq langues distinctes, à savoir le birman, l'hébreu, le chinois, l'anglais et le grec, à travers le prisme du changement climatique. En mobilisant des outils informatiques pour constituer un corpus, segmenter les textes, et analyser les caractères, nous avons pu recueillir des données significatives et effectuer une analyse comparative des

structures linguistiques.

L’ analyse des caractères a révélé non seulement la fréquence d’ utilisation de certains graphèmes, mais également des différences notables dans l’ usage de la ponctuation et des règles de segmentation entre ces langues. Cela souligne la richesse et la diversité linguistique du corpus étudié. Les défis rencontrés, notamment liés à l’encodage et à la gestion des caractères spécifiques, ont également mis en lumière l’importance d’utiliser des outils adaptés pour de telles analyses.

De plus, l’utilisation de l’outil WikiColor a montré comment la visualisation des relations entre orthographe et prononciation peut enrichir l’apprentissage des langues. Cependant, il reste des obstacles à surmonter, notamment dans le traitement des langues tonales comme le mandarin, ainsi que l’ amélioration continue de l’ outil pour maximiser son efficacité.

La phase de translittération vers l’ alphabet latin a également été une étape cruciale, facilitant ainsi la comparaison et l’interprétation des résultats de l’analyse. À travers ce projet, nous avons non seulement acquis une meilleure compréhension des dynamiques linguistiques, mais nous avons également mis en évidence l’importance de l’approche multilingue dans la recherche linguistique.

Références

[1] *Wikicolor*. https://wikicolor.alem-app.fr/fr.

[2] *Apprentissage des langues et multimodalité (ALeM)*. https://alem.hypotheses.org/category/tutoriels/tutoriels-wikicolor.

[3] *Alphabetic grec*. https://fr.wikipedia.org/wiki/Alphabet_grec.

[4] Keyman. *écriture birmane*. rhttps://fr.wikipedia.org/wiki/%C3%89criture_birmane.