



Master Recherche en Informatique et Télécommunications

Faculté des Sciences de Rabat

Université Mohammed V

Projet du module Traitement automatique des langues

Gestionnaire de corpus

Professeur :

- **Mme. Fadoua ATAA ALLAH**

Réalisé par :

- **Halima EL YAROUSSE**
- **Kawtar ET-TACHTI**

Année Universitaire 2020/2021

Tables de matières

TABLES DE MATIERES	2
TABLES DE FIGURES	ERREUR ! SIGNET NON DEFINI.
REMERCIEMENT	5
CHAPITRE 1 : LA BOITE A OUTILS DE GESTIONNAIRE DE CORPUS	6
1 INTRODUCTION	6
2 GENERALITES SUR LE GESTIONNAIRE DE CORPUS	6
2.1 CORPUS	7
2.1.1 Récupérez et explorez le corpus de textes	8
3 REALISATION DE LA BOITE A OUTILS	9
3.1 IMPLEMENTATION	9
3.1.1 Matériels	9
3.1.2 Logiciels	10
3.2 RESULTATS	10
3.2.1 Les composantes de l'interface	11
3.3 TEST D'EXECUTION	12
CONCLUSION	24
REFERENCES	25

Tables de figures

FIGURE 1 : LE CYCLE DE TRAITEMENT D'UN CORPUS DE TEXTE-----	8
FIGURE 2 : LES DIFFERENTS TYPES DE STRUCTURATION DU TEXTE-----	9
FIGURE 3 : L'INTERFACE GRAPHIQUE TKINTER DE GESTIONNAIRE DE CORPUS-----	11
FIGURE 4 : EXEMPLE DE SAISIE D'UN TEXTE PAR L'UTILISATEUR-----	12
FIGURE 5 : RESULTATS DE SEGMENTATION EN MOT DE TEXTE SAISI -----	13
FIGURE 6 : RESULTATS DE SEGMENTATION EN PHRASE DE TEXTE SAISI -----	13
FIGURE 7 : RESULTATS DES STATISTIQUES DE TEXTE SAISI -----	14
FIGURE 8 : CHOIX D'EMPLACEMENT POUR ENREGISTRER LE FICHIER XML DE TEXTE SAISI -----	15
FIGURE 9 : RESULTAT OBTENUS EN FICHIER XML DE TEXTE SAISI-----	15
FIGURE 10 : EXEMPLE DE CHARGER LE TEXTE الصاروخ الصيني -----	16
FIGURE 11 : AFFICHAGE DE TEXTE الصاروخ الصيني-----	17
FIGURE 12 : RESULTATS DE SEGMENTATION EN MOT DE TEXTE الصاروخ الصيني -----	17
FIGURE 13 : RESULTATS DE SEGMENTATION EN PHRASE DE TEXTE الصاروخ الصيني -----	18
FIGURE 14 : RESULTATS DES STATISTIQUES DE TEXTE الصاروخ الصيني -----	18
FIGURE 15 : CHOIX D'EMPLACEMENT POUR ENREGISTRER LE FICHIER XML DE TEXTE الصاروخ الصيني -----	18
FIGURE 16 : RESULTAT OBTENUS EN FICHIER XML DE TEXTE الصاروخ الصيني-----	19
FIGURE 17 : LE TABLEAU EXCEL-----	19
FIGURE 18 : LES QUATRE TEXTES POUR LE TEST DE CORPUS-----	19
FIGURE 19 : EMEUPLE DE CHARGER LE CORPUS -----	19
FIGURE 20 : EXEMPLE DE CHARGER LE CORPUS -----	21

FIGURE 21: RESULTATS DE SEGMENTATION EN MOT DE CORPUS-----	21
FIGURE 22 : RESULTATS DE SEGMENTATION EN PHRASE DE CORPUS-----	22
FIGURE 23 : RESULTATS DES STATISTIQUES DE CHAQUE TEXTE DE CORPUS-----	22
FIGURE 24 : RESULTATS DES STATISTIQUES DE CORPUS -----	23
FIGURE 25 : AFFICHAGE DE REMARQUE LEUR DE QUITTER L'INTERFACE-----	23

Remerciement

Nous tenons à remercier, le corps professoral et administratif de la faculté des Sciences de Rabat, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisé.

Mme. Fadoua ATAA ALLAH. *Nous vous remercions aussi d'avoir partagé avec nous votre passion pour l'enseignement. Nous avons grandement apprécié votre soutien, votre implication et votre expérience tout au long de votre séance de cours.*

Chapitre 1 : La boîte à outils de gestionnaire de corpus

1 Introduction

A travers ce chapitre, nous allons intéresser par la réalisation d'un outil qui assure la gestion des fichiers d'un corpus an assuré les fonctionnalités suivantes :

- segmenter le texte en phrases et en mots ;
- présenter le résultat de la segmentation dans un fichier XML en marquant les parties : le nom du fichier, l'auteur et la dates de rédaction extrait d'un fichier Excel, ainsi que la phrase et le mot tout en précisant leur emplacement ;
- présenter les statistiques du corpus en termes de fréquence de mots et de caractères, la longueur des fichiers, longueur moyenne et maximale des mots du texte.

Au cours de ce rapport, nous allons présenter, dans une première partie une idée générale concernant le gestionnaire de corpus. Dans la deuxième partie l'implémentation de notre application.

2 Généralités sur le gestionnaire de corpus

Un gestionnaire de corpus (navigateur de corpus ou système de requête de corpus) est un outil d'analyse de corpus multilingue, qui permet une recherche efficace dans les corpus. Un gestionnaire de corpus représente généralement un outil complexe qui permet d'effectuer des recherches de formes ou de séquences de langage. Il peut fournir des informations sur le contexte ou permettre à l'utilisateur de rechercher par attributs de position, tels que lemme, balise, etc. On les appelle concordances. D'autres fonctionnalités incluent la possibilité de rechercher des collocations, des statistiques de fréquence ainsi que des informations de métadonnées sur le texte traité. La signification plus étroite du gestionnaire de corpus se réfère uniquement au côté serveur ou au moteur de requête de corpus, alors que le côté client est simplement appelé interface utilisateur. Un gestionnaire de corpus peut être un logiciel installé sur un ordinateur personnel ou il peut être fourni en tant que service Web.

- ✓ Liste des gestionnaires de corpus :
 - **BNCweb** : une interface Web pour le British National Corpus
 - **CQPweb** : une interface Web pour l'étude d'une grande variété de corpus, y compris le Spoken

- **BNC2014 BYU-BNC** : un site Web qui permet des recherches dans les corpus nationaux britanniques et d'autres créés à l'Université
- **Brigham Young Coma** : une extension d'outils du système
- **EXMARaLDA** : pour travailler avec des corpus oraux sur un ordinateur
- **NoSketch Engine** : un système de gestion de corpus open-source gratuit combinant Manatee (back-end) et Bonito (interface web)
- **KonText** : une interface Web étendue et modifiée pour
- **NoSketch Engine (un remplacement Bonito) Sketch Engine** : logiciel de gestion et d'analyse de corpus de texte avec plus de 400 corpus en 80 langues
- **WordSmith Tools** : un progiciel destiné principalement aux linguistes

Remarque :

Dans notre projet le gestionnaire de corpus est un outil permet l'utilisateur de faire des fonctionnalités un peu différent comme il est mentionné dans l'introduction.

2.1 Corpus

Les corpus sont des outils indispensables et précieux en traitement automatique du langage naturel. Ils permettent en effet d'extraire un ensemble d'informations utiles pour des traitements statistiques.

Corpus est le pluriel de corpora ; Une collection de données linguistiques, parfois une compilation de textes écrits, ou de transcriptions d'enregistrement de discours. La raison principale d'un corpus est de vérifier une hypothèse sur le langage – par exemple : déterminer comment l'utilisation d'un son particulier, d'un mot ou d'une construction syntaxique varie. Les corpus linguistiques agissent avec les lois et les pratiques d'utilisation de corpora dans l'étude du langage. Un corpus informatique contient un ensemble vaste de textes traduisibles en langage-machine.

Ainsi, les corpus sont des

- collections de textes de taille importante
- constituées de textes authentiques
- rassemblées selon des critères spécifiques
- collectées sous format électronique

2.1.1 Récupérez et explorez le corpus de textes

La première étape du traitement des données texte est de récupérer le texte et le nettoyer afin de pouvoir l'utiliser ultérieurement dans vos algorithmes. C'est aussi bien de passer par une petite exploration afin de mieux le comprendre.

✚ **Le pré-traitement du texte comprend les étapes suivantes :**

1. Récupération du **corpus** par scraping ou en téléchargeant des fichiers textes, par exemple. Cela peut demander l'utilisation de regex afin de récupérer uniquement les parties qui vous intéressent.
2. La **tokenization**, qui désigne le découpage en mots des différents documents qui constituent votre corpus
3. La **normalisation** et la construction du dictionnaire qui permet de ne pas prendre en compte des détails importants au niveau local (ponctuation, majuscules, conjugaison, etc.)



Figure 1 : Le cycle de traitement d'un corpus de texte

Le corpus peut être organisé de plusieurs manières différentes :

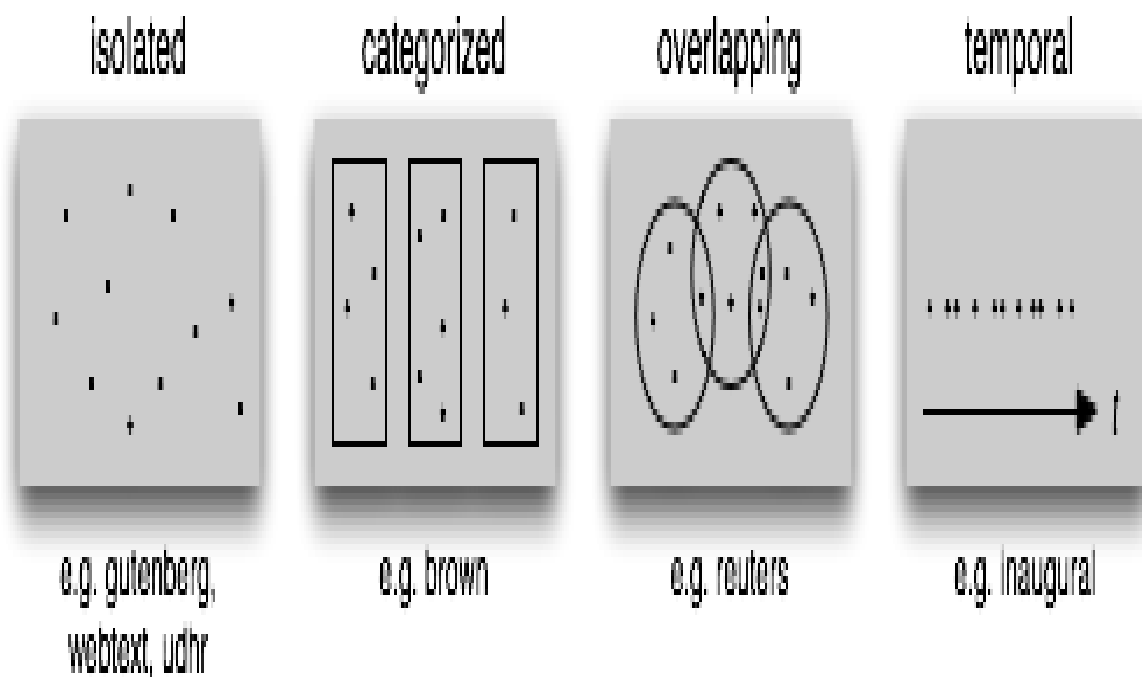


Figure 2 : Les différents types de structuration du texte

3 Réalisation de la boîte à outils

Dans ce chapitre nous allons présenter les différents matériels et logiciels pour réaliser notre application de gestionnaire de corpus, puis on démontrons les résultats d'exécution de notre interface graphique.

3.1 Implémentation

3.1.1 Matériels

✓ PC portable

On a utilisé un pc avec les propriétés suivantes :

Spécifications de l'appareil

Nom de l'appareil	DESKTOP-SA18GJN
Processeur	Intel(R) Core(TM) i5-5300U CPU @ 2.30GHz 2.30 GHz
Mémoire RAM installée	8,00 Go (7,88 Go utilisable)
ID de périphérique	F55BF809-1217-4F7C-B895-4390BD907CA8
ID de produit	00331-10000-00001-AA817
Type du système	Système d'exploitation 64 bits, processeur x64
Styilet et fonction tactile	La fonctionnalité d'entrée tactile ou avec un stylet n'est pas disponible sur cet écran

3.1.2 Logiciels

- ✓ Python

On utilisant la version 3.8.3

- ✓ Bibliothèque Tkinter

Le module Tkinter pour la conception de l'interface graphique

- ✓ Bibliothèque nltk

La bibliothèque nltk pour la segmentation en mot et en phrase

3.2 Résultats

Notre interface graphique permet à l'utilisateur de faire les fonctionnalités suivant :

- ✓ segmenter le texte en phrases et en mots ;
- ✓ présenter le résultat de la segmentation dans un fichier XML en marquant les parties : le nom du fichier, l'auteur et la dates de rédaction extrait d'un fichier Excel, ainsi que la phrase et le mot tout en précisant leur emplacement ;
- ✓ présenter les statistiques du corpus en termes de fréquence de mots et de caractères, la longueur des fichiers, longueur moyenne et maximale des mots du texte.

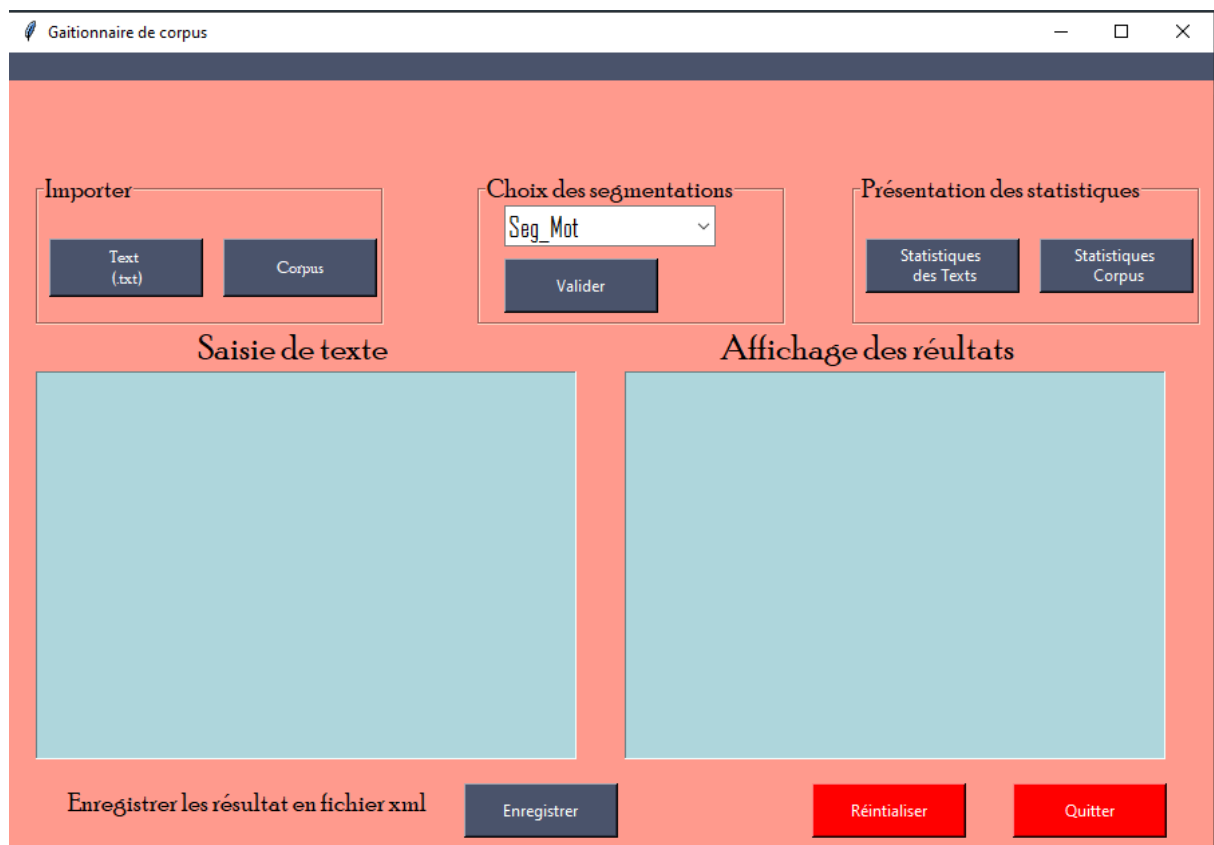


Figure 3 : L'interface graphique Tkinter de gestionnaire de corpus

3.2.1 Les composantes de l'interface

L'interface compose principalement par des boutons et des champs soit pour la saisie ou l'affichage des résultats.

Pour les boutons en trouve :

- ✓ deux boutons pour charger un texte et l'autre pour charger un corpus
- ✓ bouton pour la validation de choix de segmentation en mot ou en phrase
- ✓ deux boutons pour les statistiques de texte et l'autre le corpus
- ✓ bouton pour enregistrer les résultats de segmentation en fichier XML
- ✓ les derniers boutons un pour vider tous les champs et l'autre pour quitter l'interface

Pour les champs :

- ✓ Un champ pour le saisi de texte
- ✓ Un champ pour l'affichage des résultats

3.3 Test d'exécution

Puis ce que l'utilisateur peut utiliser à la fois de saisie un texte puis appliquer s les différents fonctionnalités ou peut charger un texte ou plutôt un corpus.

✓ Test de saisie

La figure ci-dessous présente l'exemple de saisie d'un texte par l'utilisateur.

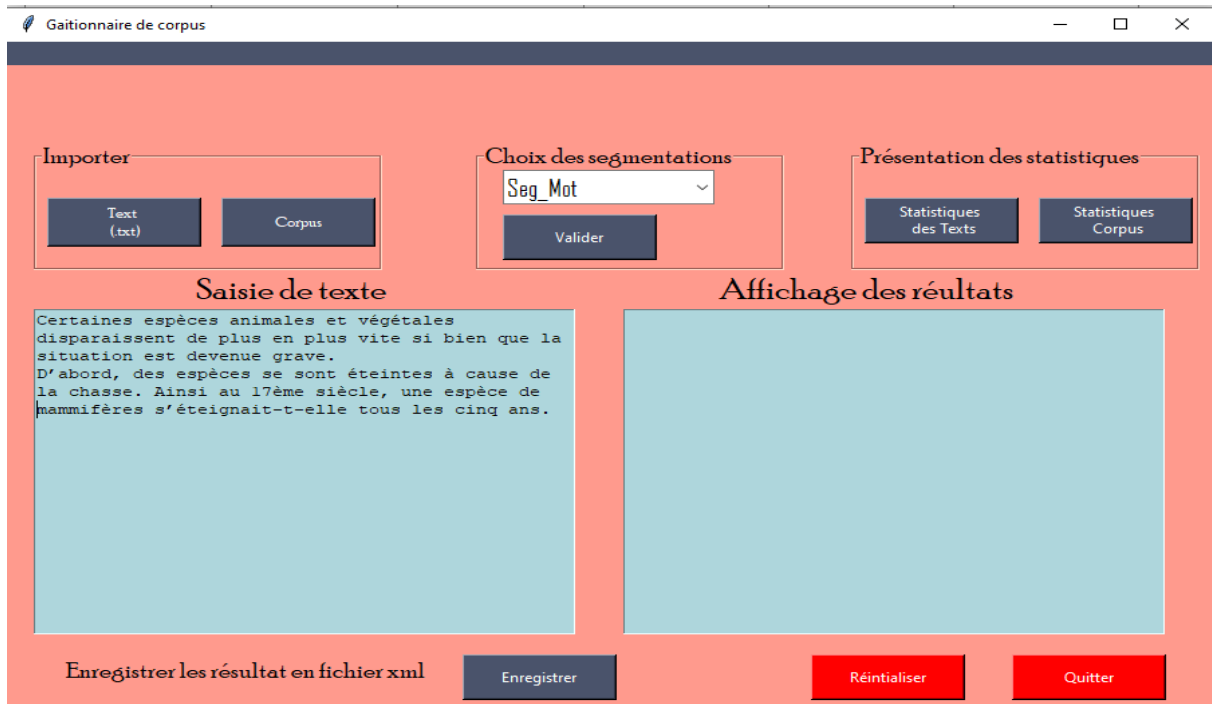


Figure 4 : Exemple de saisie d'un texte par l'utilisateur

La deuxième et la troisième figure affiche les résultats de segmentation en mot et en phrase respectivement.



Figure 5 : Résultats de segmentation en mot de texte saisi

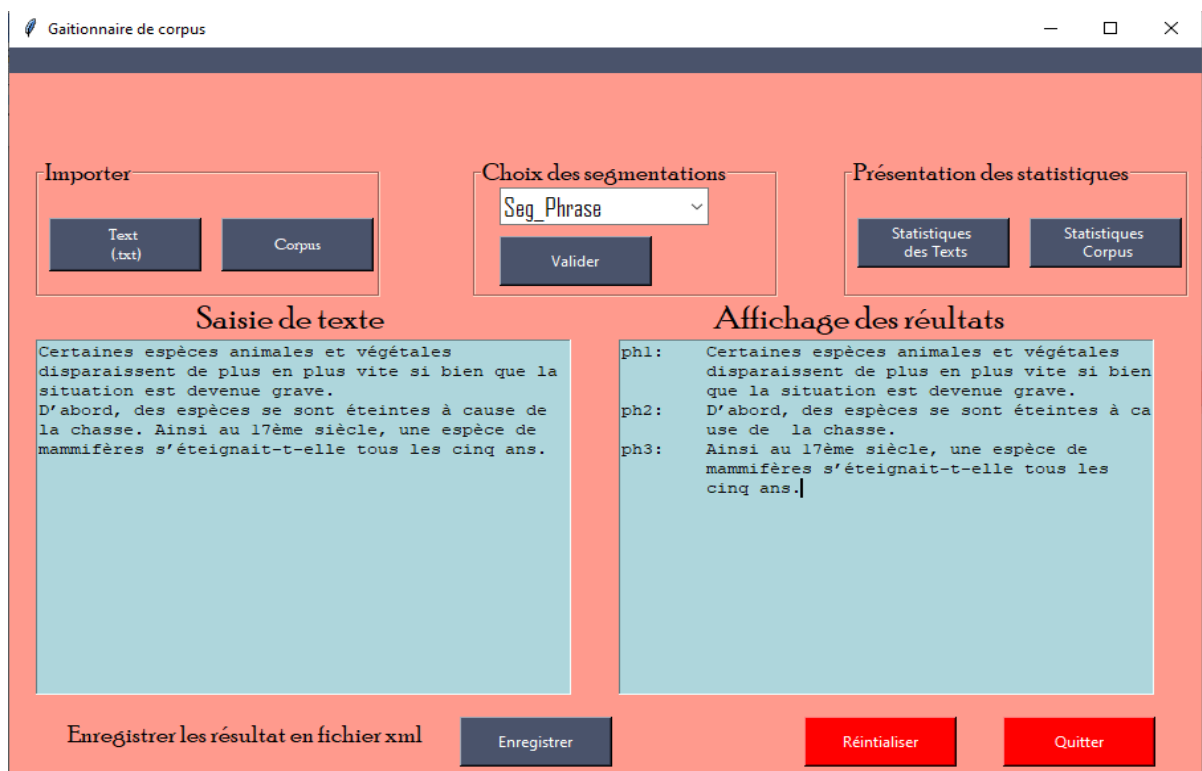


Figure 6 : Résultats de segmentation en phrase de texte saisi

La quatrième figure affiche les résultats des statistiques de texte saisi.

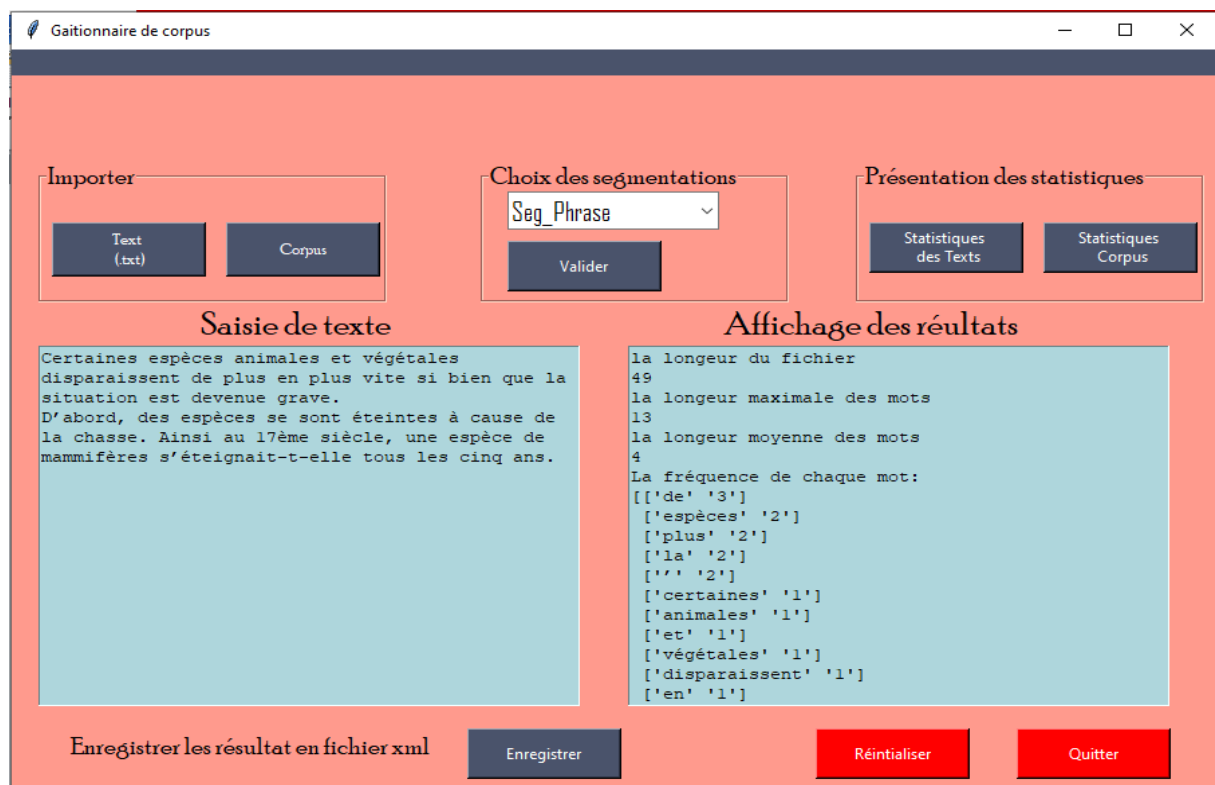


Figure 7 : Résultats des statistiques de texte saisi

Finally the two last figures present the choice of location to save the xml file and the other the result obtained in xml file.

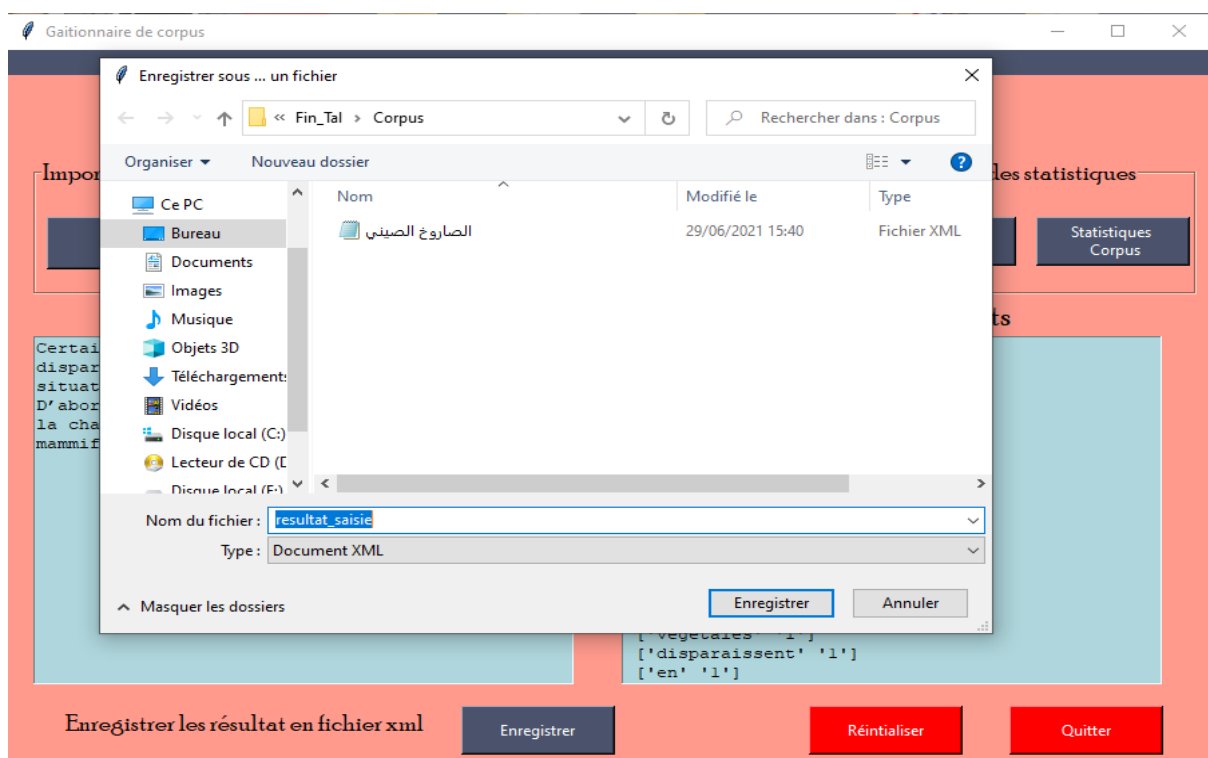
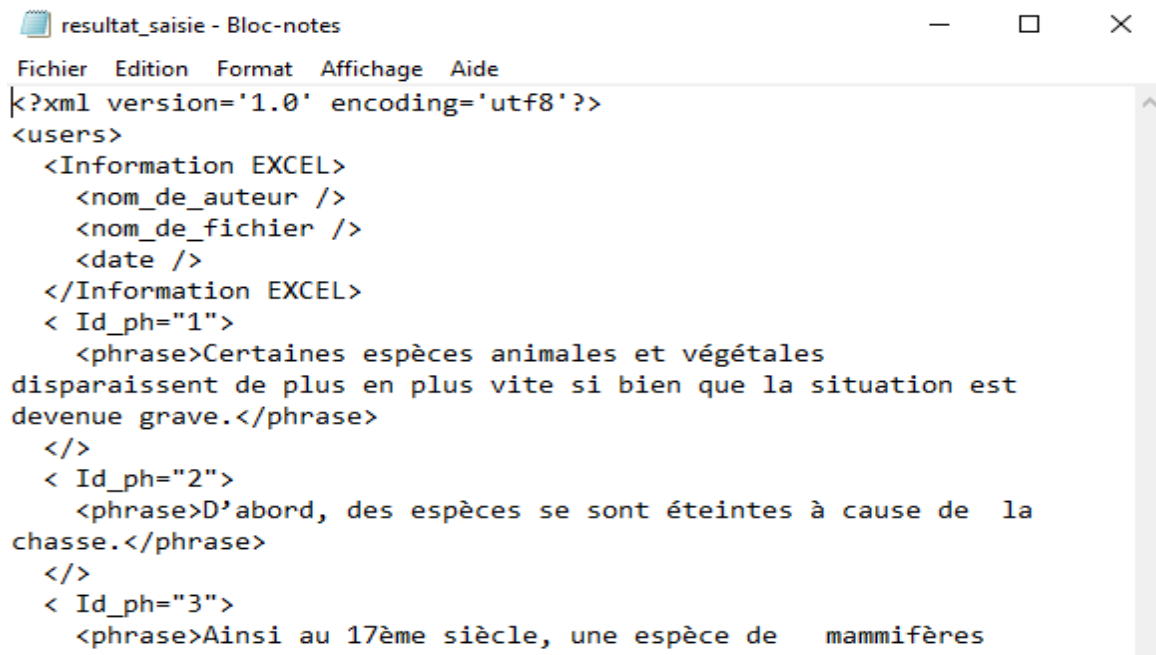


Figure 8 : Choix d'emplacement pour enregistrer le fichier xml de texte saisi



```
<?xml version='1.0' encoding='utf8'?>
<users>
  <Information EXCEL>
    <nom_de_auteur />
    <nom_de_fichier />
    <date />
  </Information EXCEL>
  < Id_ph="1">
    <phrase>Certaines espèces animales et végétales
disparaissent de plus en plus vite si bien que la situation est
devenue grave.</phrase>
  </>
  < Id_ph="2">
    <phrase>D'abord, des espèces se sont éteintes à cause de la
chasse.</phrase>
  </>
  < Id_ph="3">
    <phrase>Ainsi au 17ème siècle, une espèce de mammifères
```

Figure 9 : Résultat obtenus en fichier xml de texte saisi

Remarque :

Pour le texte saisi les informations extrait par le tableau Excel saurons vide et le l'utilisateur peut le remplir

✓ **Test de charger un texte**

Puis ce que on travaillons par deux types de langue diffèrent l'arabe et latin, on limitons le test d'exécution pour la langue arabe, on fait le même démarche que le test de texte saisi, sauf que cette partie en donnons la possibilité de charger un texte par l'utilisateur.

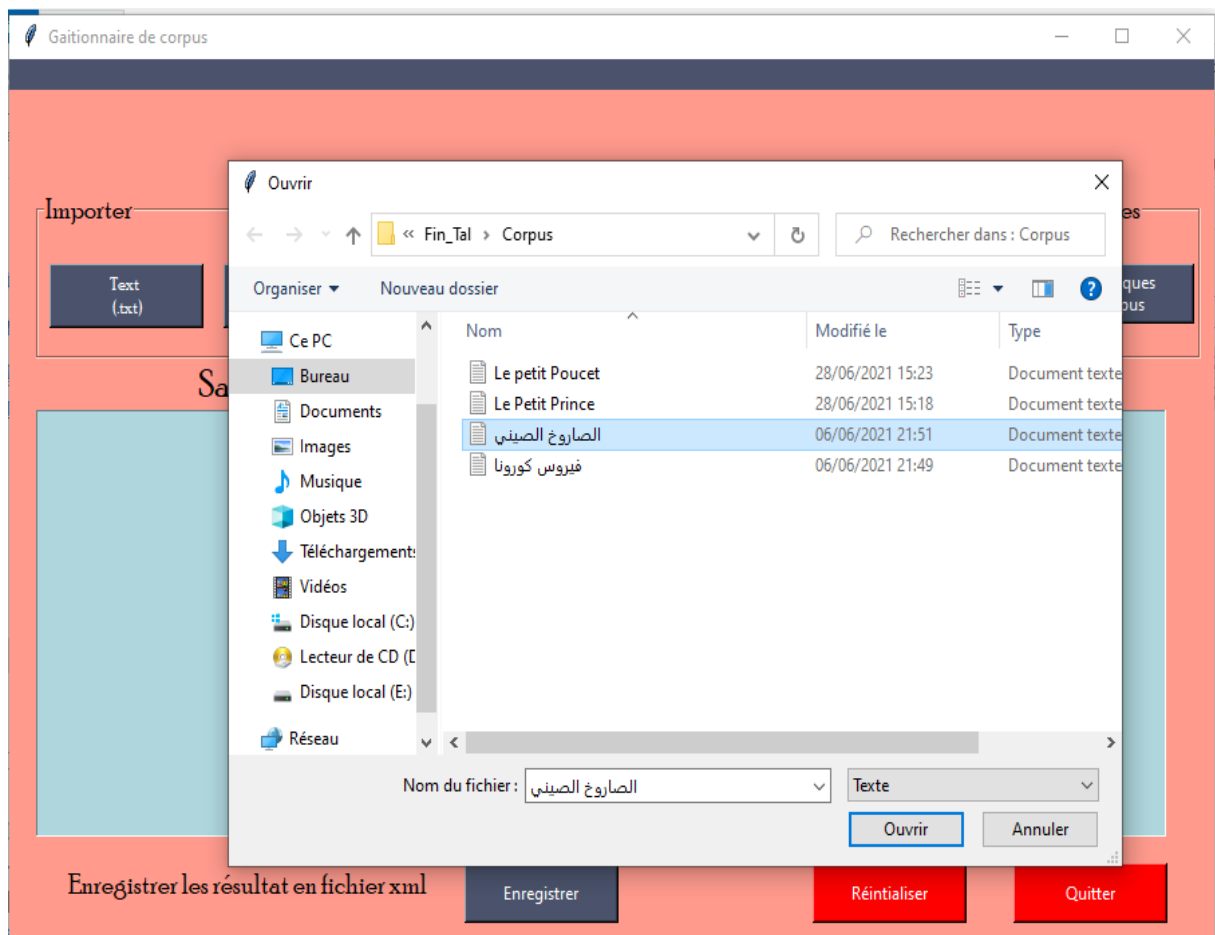


Figure 10 : Exemple de charger le texte الصاروخ الصيني



Figure 11 : Affichage de texte الصينى الصاروخ

Importer

Text (txt)

Corpus

Choix des segmentations

Seq Mot

Valider

Présentation des statistiques

Statistiques des Texts

Statistiques Corpus

Saisie de texte

تُخلق القرص بالمنافسة؛ لذا تبدو الصين أمام العالم و في حالتها مع الصاروخ "التائه" كشاب متحمس، يريد إثبات قوته للآخرين، رغم أنه لم يكن للصين أن تحظى بكل ه ذا الحضور لولا تجاربها، ومحاولاتها المستميتة لإثبات ن فسها، وأنها قادرة على أن تجعل نفسها في مصاف الدول المتقدمة .

جميع العالم بلا استثناء، كبرهم وصغيرهم، كان يتربح ما سيحدث للصاروخ الذي وُصف بـ"التائه"، رغم ظهور تصر يحات صينية تذكر أن هذا هو النظام المعمول به في حال تسليم حمولة فضائية، وأن جميع الصواريخ يحدث لها ما حدث وسيحدث مع هذا الصاروخ؛ إذ لا يوجد مفهوم تحكّم ف ي حطام الصاروخ، وإن رحلة الحطام محسوبة بعناية كما يحدث دائمًا، وسيحدث لاحقًا. ورغم ذلك إلا أن التصعيد وا لتربح من الجميع كان مستمرًا حتى لحظة سقوط بقية الحط ام!

Affichage des résultats

mot1: تُخلق

mot2: القرص

mot3: بالمنافسة؛

mot4: لذا

mot5: تبدو

mot6: الصين

mot7: أمام

mot8: العالم

mot9: وفي

mot10: حالتها

mot11: مع

mot12: الصاروخ

mot13: التائه

mot14: كشاب

mot15: متحمس،

mot16: يريد

mot17: إثبات

mot18: قوته

Enregistrer les résultats en fichier xml

Enregistrer

Réinitialiser

Quitter

Figure 12 : Résultats de segmentation en mot de texte الصينى الصاروخ

Importer

Text (txt)

Corpus

Choix des segmentations

Seq Phrase

Valider

Présentation des statistiques

Statistiques des Texts

Statistiques Corpus

Saisie de texte

تُخلق القرص بالمنافسة؛ لذا تبدو الصين أمام العالم و في حالتها مع الصاروخ "التائه" كشاب متحمس، يريد إثبات قوته للآخرين، رغم أنه لم يكن للصين أن تحظى بكل ه ذا الحضور لولا تجاربها، ومحاولاتها المستميتة لإثبات ن فسها، وأنها قادرة على أن تجعل نفسها في مصاف الدول المتقدمة .

جميع العالم بلا استثناء، كبرهم وصغيرهم، كا ن يتربح ما سيحدث للصاروخ الذي وُصف بـ"التائه"، رغم ظهور تصر يحات صينية تذكر أن هذا هو النظام المعمول ب ه في حال تسليم حمولة فضائية، وأن جميع الصواريخ يحد ث لها ما حدث وسيحدث مع هذا الصاروخ؛ إذ لا يوجد مفهوم م تحكّم في حطام الصاروخ، وإن رحلة الحطام محسوبة بعن اية كما يحدث دائمًا، وسيحدث لاحقًا. ورغم ذلك إلا أن التصعيد وا لتربح من الجميع كان مستمرًا حتى لحظة سقوط بقية الحط ام!

Affichage des résultats

ph1: تُخلق القرص بالمنافسة؛ لذا تبدو الصين أمام ،العالم وفي حالتها مع الصاروخ "التائه" كشاب متحمس يريد إثبات قوته للآخرين، رغم أنه لم يكن للصين أن تح ظى بكل هذا الحضور لولا تجاربها، ومحاولاتها المستميتة لإثبات نفسها، وأنها قادرة على أن تجعل نفسها في مصا ف الدول المتقدمة .

ph2: جميع العالم بلا استثناء، كبرهم وصغيرهم، كا ن يتربح ما سيحدث للصاروخ الذي وُصف بـ"التائه"، رغم ظهور تصريحات صينية تذكر أن هذا هو النظام المعمول ب ه في حال تسليم حمولة فضائية، وأن جميع الصواريخ يحد ث لها ما حدث وسيحدث مع هذا الصاروخ؛ إذ لا يوجد مفهوم م تحكّم في حطام الصاروخ، وإن رحلة الحطام محسوبة بعن اية كما يحدث دائمًا، وسيحدث لاحقًا. ورغم ذلك إلا أن التصعيد والتربح من الجميع ك

ph3: !ان مستمرًا حتى لحظة سقوط بقية الحطام

Enregistrer les résultats en fichier xml

Enregistrer

Réinitialiser

Quitter

Figure 13 : Résultats de segmentation en phrase de texte الصاروخ الصيني

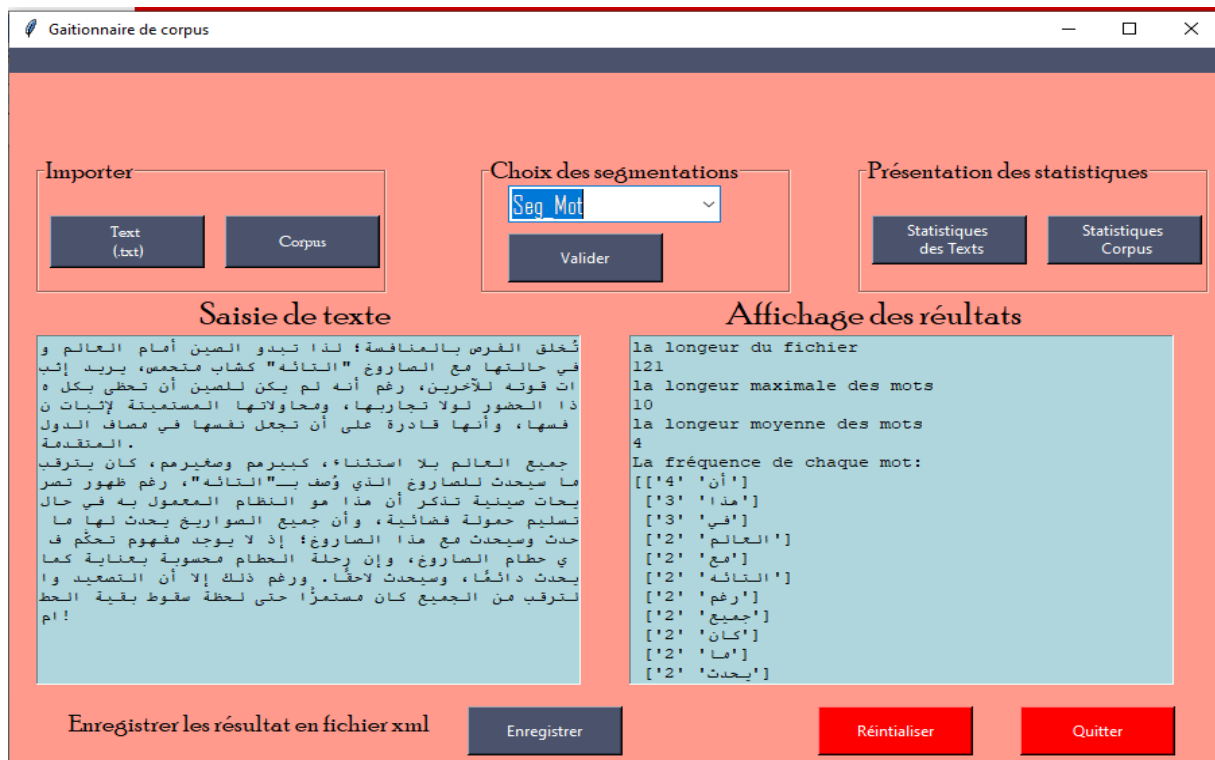


Figure 14 : Résultats des statistiques de texte الصاروخ الصيني

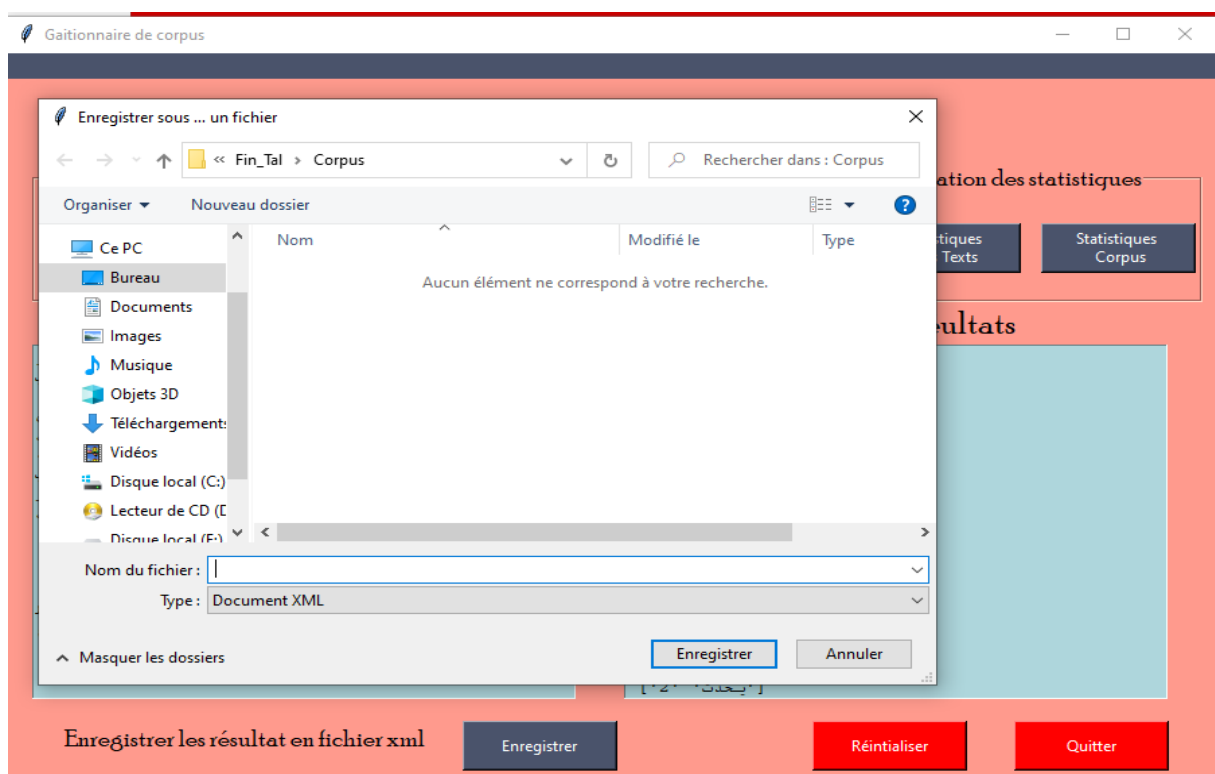


Figure 15 : Choix d'emplacement pour enregistrer le fichier xml de texte الصاروخ الصيني

```

<?xml version='1.0' encoding='utf8'?>
<users>
  <Information EXCEL>
    <nom_de_auteur>مها الجبر</nom_de_auteur>
    <nom_de_fichier>الصاروخ الصيني</nom_de_fichier>
    <date>2021.0</date>
  </Information EXCEL>
  < Id_mot="1">
    <mot>تُخلق</mot>
  </>
  < Id_mot="2">
    <mot>الفرص</mot>
  </>
  < Id_mot="3">
    <mot>بالمنافسة</mot>
  </>
  < Id_mot="4">
    <mot>لذا</mot>
  </>
  < Id_mot="5">
    <mot>تبدو</mot>
  </>
  < Id_mot="6">
    <mot>الصين</mot>
  </>
  < Id_mot="7">
    <mot>أمام</mot>
  </>
</users>

```

Figure 16 : Résultat obtenus en fichier xml de texte الصاروخ الصيني

Remarque :

Pour le cas de charger un texte les informations au-dessous de la balise « Information EXCEL » sont extrait par ce tableau Excel

	A	B	C	D	E
1	nom de l'auteur	Antoine de Saint-Exupéry	charles Perrault	وكالة ابوظبي	مها الجبر
2	nom du fichier	<i>Le Petit Prince</i>	<i>Le petit Poucet</i>	فيروس كورونا	الصاروخ الصيني
3	date de rédaction	1943	1697	2020	2021

Figure 17 : Le tableau Excel

✓ **Test de charger un corpus**

Pour le test de corpus l'utilisateur peut charger un corpus qui contient plusieurs textes de différentes langues. Dans notre cas le corpus contient 4 textes, deux en arabe et deux autres en français





 Le petit Poucet_Mot	29/06/2021 15:59	Fichier XML	5 Ko
 Le Petit Prince_Mot	29/06/2021 15:59	Fichier XML	3 Ko
 الصاروخ الصيني_Mot	29/06/2021 15:59	Fichier XML	7 Ko
 فيروس كورونا_Mot	29/06/2021 15:59	Fichier XML	7 Ko

Figure 18: Les quatre textes pour le test de corpus

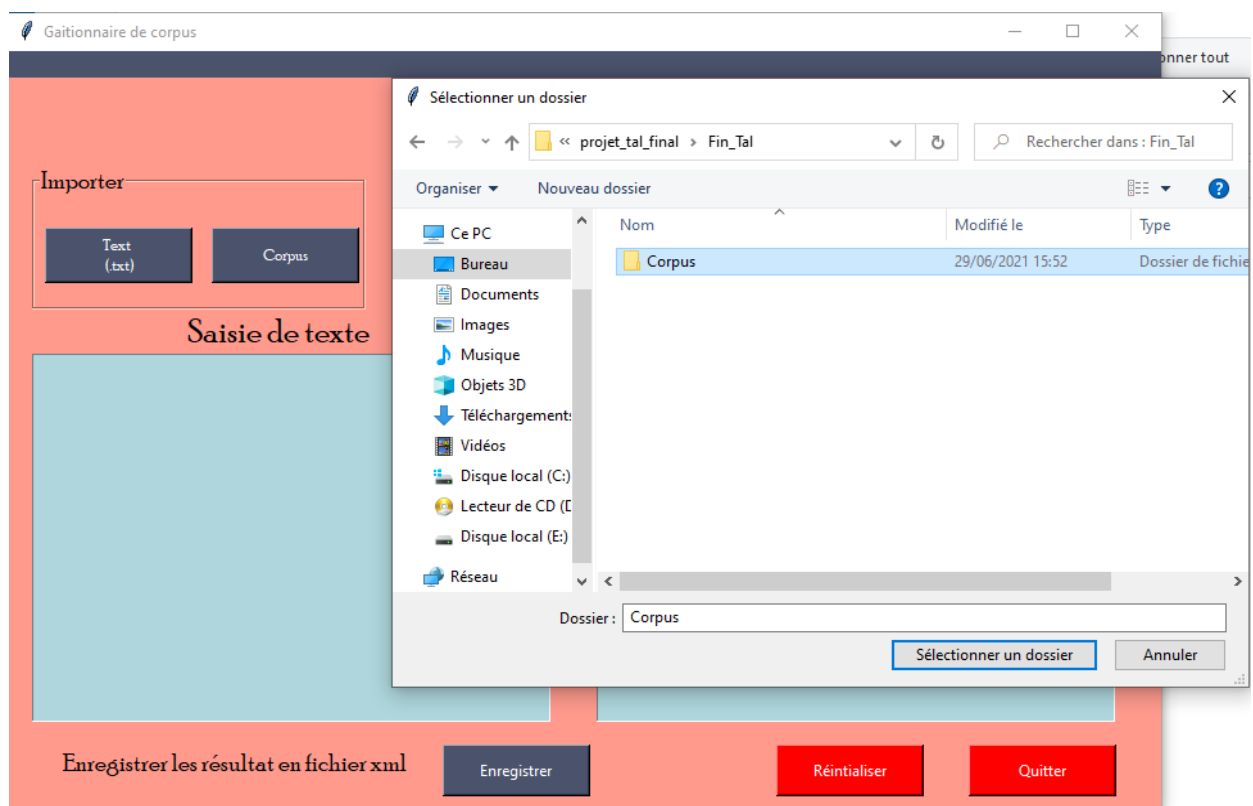


Figure 19 : Exemple de charger le corpus



Figure 20 : Exemple de charger le corpus

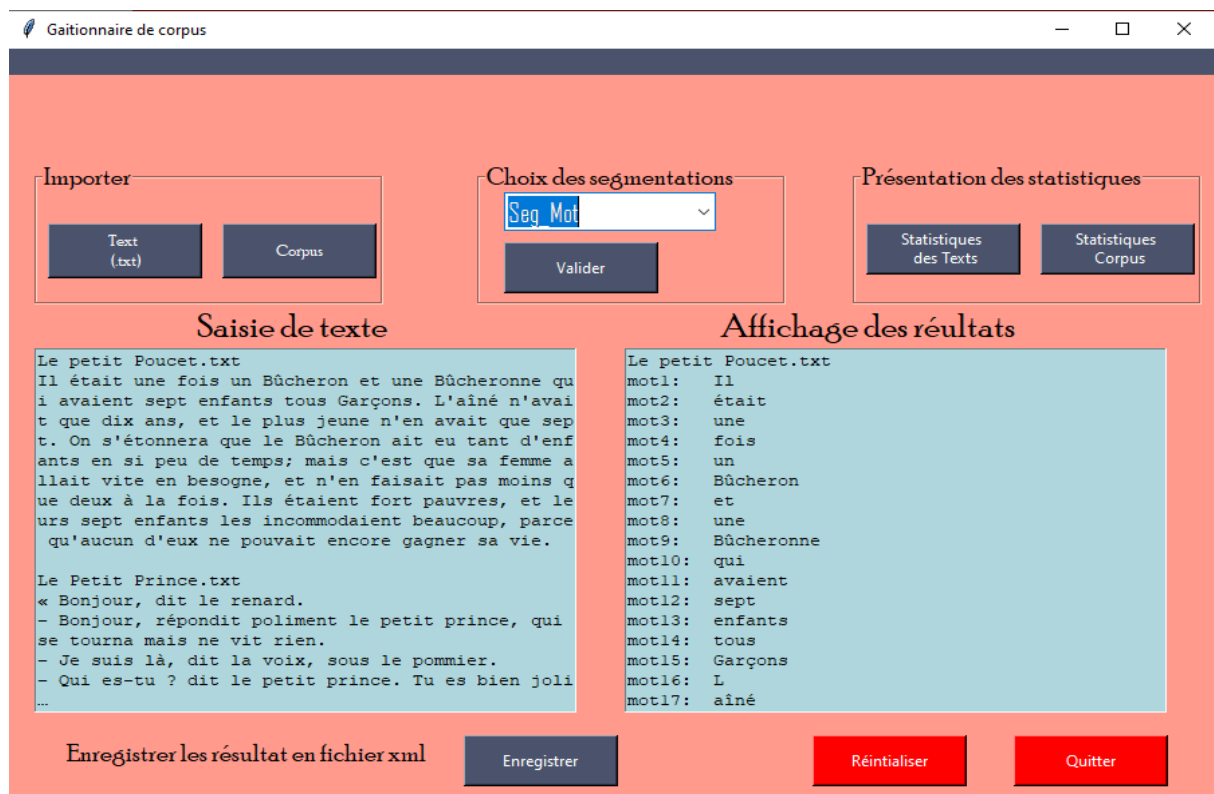


Figure 21: Résultats de segmentation en mot de corpus



Figure 22 : Résultats de segmentation en phrase de corpus

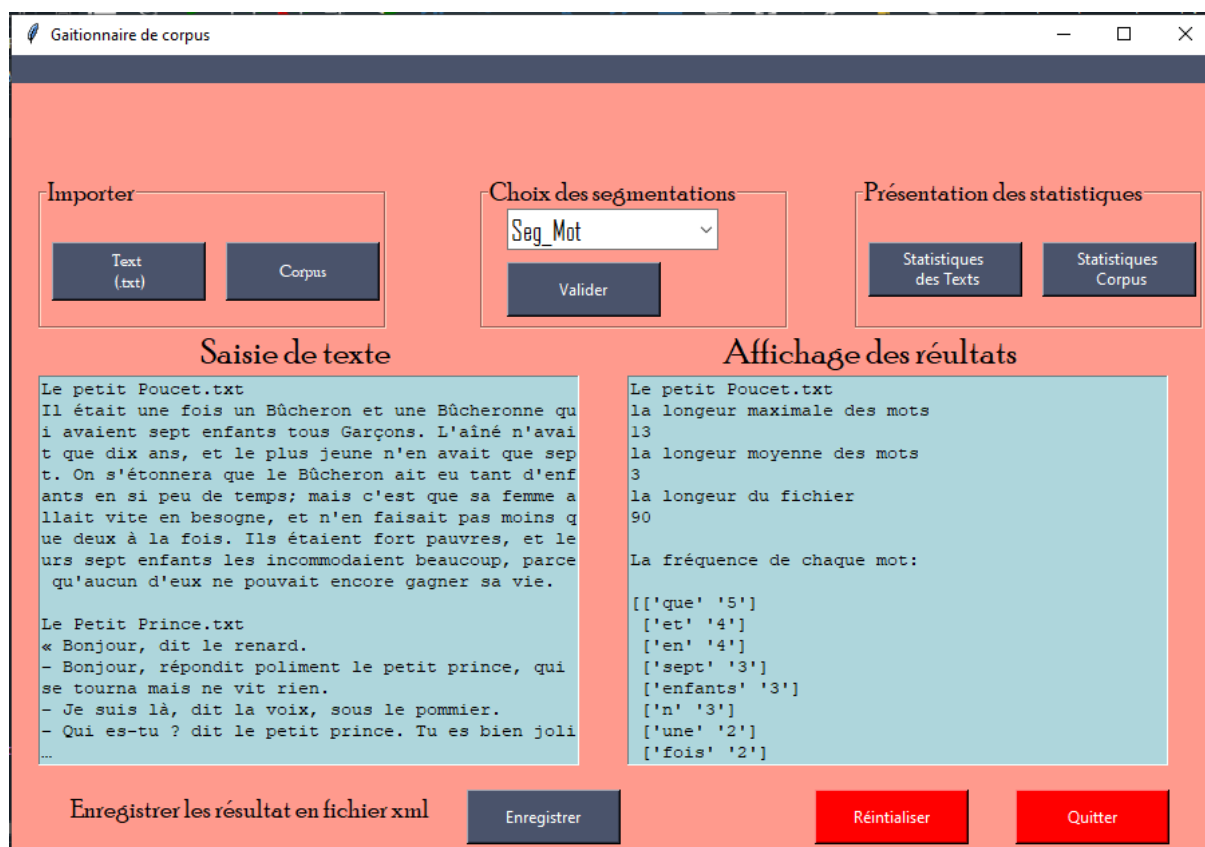


Figure 23 : Résultats des statistiques de chaque texte de corpus

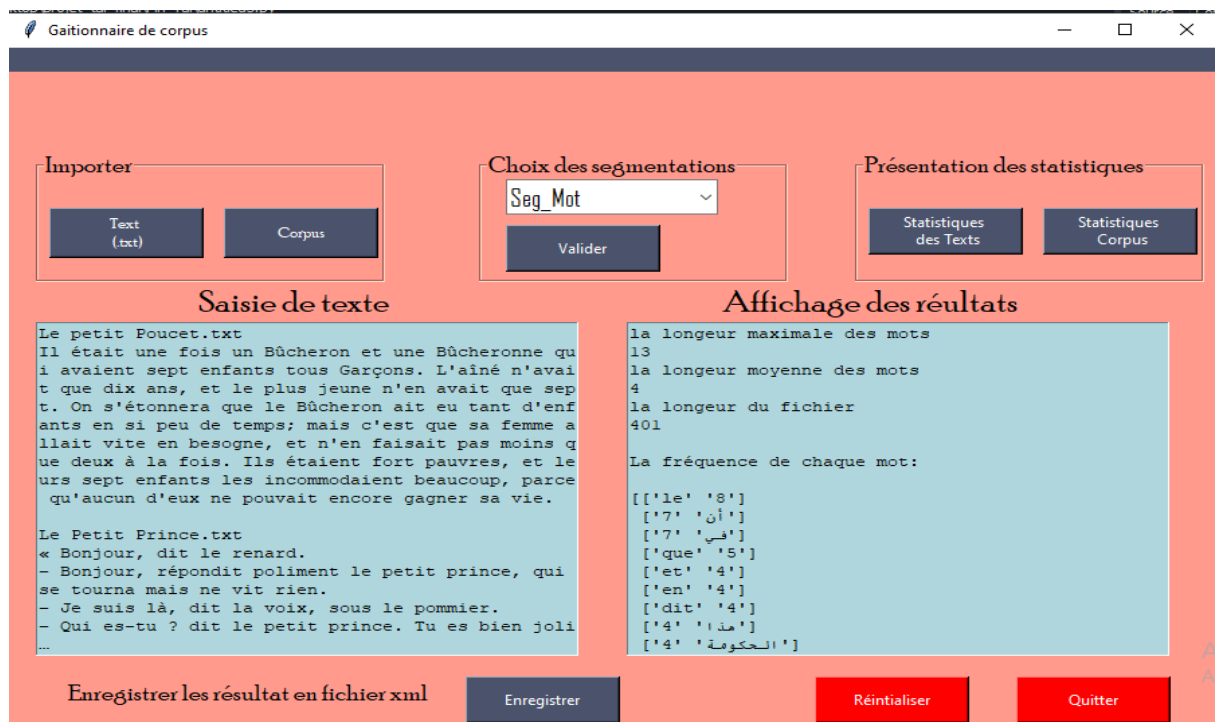


Figure 24 : Résultats des statistiques de corpus

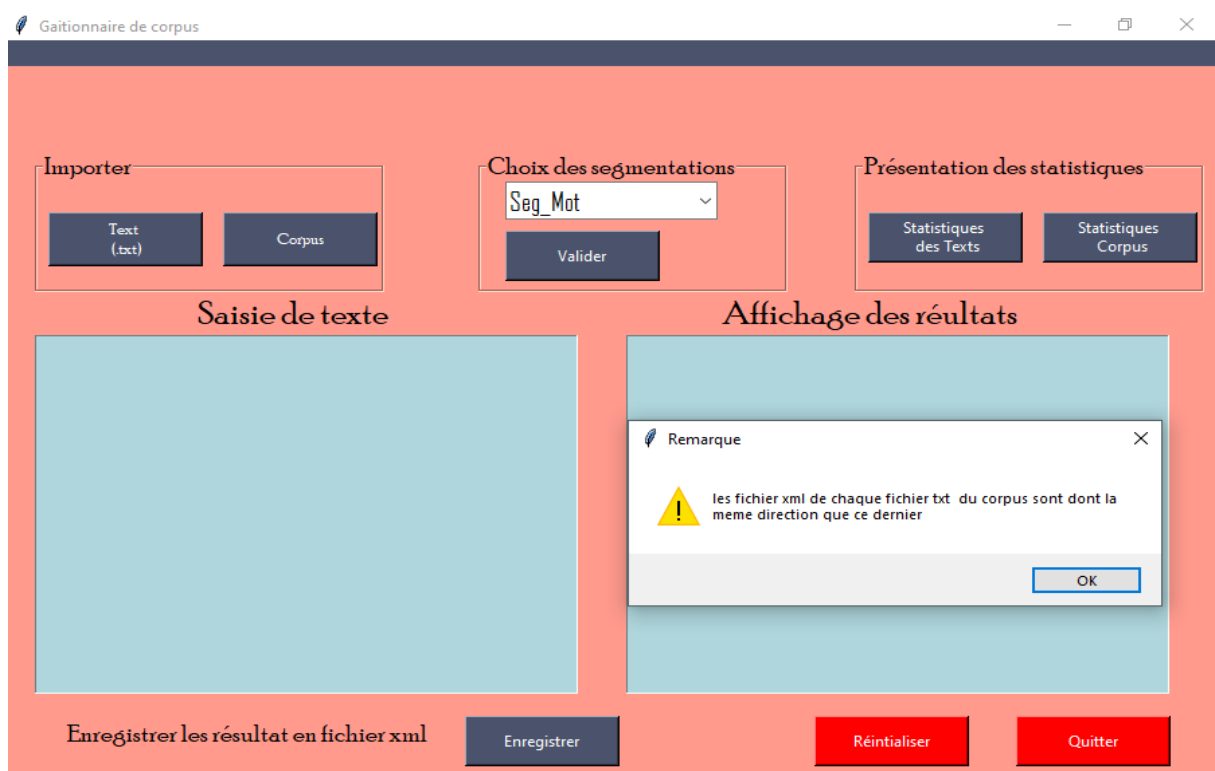


Figure 25 : Affichage de remarque lors de quitter l'interface

Remarque :

Dans le cas de corpus l'enregistrement des fichiers xml seront gérés automatiquement dans le corpus, comme il montre la figure ci-dessous.

Conclusion

Nous allons réaliser en premier temps chaque tâche indépendante, ce qui facilitera notre travail et d'autre part de bien développer nos connaissances en Traitement automatique des langues. Enfin on a rassemblé toutes les fonctionnalités dans une seule interface.

Le gestionnaire de corpus est un outil très important pour l'analyse et la recherche efficace dans les corpus, l'accent sera mis sur la segmentation des textes en mot et en phrase, et calculer leur statistique en termes des fréquences et de longueurs moyennes. Nous n'approfondirons pas sur tous les détails délicats de gestion de corpus ou les corpus sous-jacents de ces différents types. Dans la partie d'implémentation nous avons examiné la boîte à outils de notre gestionnaire de corpus.

Références

- Gestionnaire de corpus - https://fr.xcv.wiki/wiki/Corpus_manager
- <https://openclassrooms.com/fr/courses/4470541-analysez-vos-donnees-textuelles/4470548-recuperez-et-explorez-le-corpus-de-textes>