Yacoub Ghérissi

AntConc

de Laurence Anthony

Édition augmentée et mise à jour pour la version 4.2.0

Guide d'utilisation en français

gh.yac.2013@gmail.com

2023

« *Le mariage entre la linguistique appliquée, surtout dans les domaines de l'enseignement des langues de spécialité et la lexicographie, et l'informatique a conduit à une discipline nouvelle, la linguistique des corpus.* »

Geoffrey Williams

Introduction

La linguistique de corpus est une nouvelle branche de la linguistique qui exploite les technologies informatiques sous forme de logiciels pour effectuer une meilleure étude de la langue en pratiquant des recherches sur de grands corpus constitués de textes numérisés. Avec l'amélioration des processus de récupération et les capacités de stockage des données de plus en plus grandes, les chercheurs de toutes les disciplines peuvent désormais profiter également d'un bon nombre de logiciels de traitement très performants et souvent offerts gratuitement au téléchargement sur Internet[1].

Cette linguistique facilite l'analyse de divers phénomènes linguistiques dans leurs contextes réels en profitant de la puissance des moteurs de recherche et participe également, entre autres, à la création de différents dictionnaires, de grammaires etc.

Ces logiciels sont créés par des informaticiens qui maîtrisent les langages informatiques, mais pour les linguistes et les littéraires qui, très souvent, ignorent ces langages basés sur les mathématiques, l'utilisation est parfois rébarbative et déconcertante.

Afin d'aider les chercheurs en sciences du langage qui, dans leurs recherches, manipulent des textes de tous genres (textes littéraires, discours politiques, articles de journaux, etc.) à tirer un meilleur profit des bases de données, nous avons choisi de présenter l'un de ces outils, le logiciel ***AntConc***.

C'est un logiciel concordancier, développé par le Professeur Laurence Anthony[2] . Il est téléchargeable gratuitement sur Internet[3] et tourne sous Windows, MacOS X et GNU/Linux. Il ne nécessite aucune installation et se lance par un simple double-clic sur le fichier exécutable (.exe). La version 2019, ***AntConc3.5.8w*** (w pour Windows), fonctionne correctement avec toutes les versions de Windows.[4]

Pour optimiser la recherche avec ***AntConc***, il est important de connaître les options de configuration que ce logiciel offre à l'utilisateur. Dans cette initiation, au premier chapitre, nous présentons dans les détails les configurations des réglages et les outils proposés. Nous avons également donné les traductions de tous les termes anglais que le chercheur non anglophone rencontrera au cours de la lecture.

Dans le deuxième chapitre, nous montrons comment faire, avec ***AntConc***, une recherche morphologique simple et facile. Le programme analyse statistiquement les textes numérisés et enregistrés sur le disque dur de la machine qui lui sont proposés. Il ne possède pas de dictionnaire intégré. Il génère principalement des concordances selon les formes alphanumériques saisies. Le résultat dépend donc forcément des

[1] Citons à titre d'exemple ***TXM, Tropes, Cordial, Lexico5, DTMVic, NooJ, Unitex, Hyperbas, IRaMuTeQ***.
[2] Center for English Language Education in Science and Engineering, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo. *Ant* et *Conc* sont respectivement la première syllabe du nom du concepteur et la première syllabe du mot *concordancier*.
[3] http://www.laurenceanthony.net/software/antconc/
[4] De nouvelles versions sont disponibles depuis. Voir Chapitre Quatrième pour les nouveautés et les améliorations.

formes saisies, et la moindre faute de frappe dans le texte numérisé ou dans la requête peut générer un résultat nul ou erroné.

Nous verrons, avec le troisième chapitre, qu'il est possible de dépasser le stade de « l'association entre un corpus brut, c'est-à-dire réduit au seul texte, sans annotations linguistiques, et une interface de visualisation de type concordancier »[5], et ce grâce au logiciel ***TreeTagger*** qui permet de catégoriser les mots des textes.

Nous avons ajouté un quatrième chapitre pour présenter les nouveautés de la dernière version (4.2.0) du logiciel en comparaison avec ses anciennes versions. Un tableau récapitulatif résumera les modifications apportées.

Enfin, pour une utilisation plus poussée des résultats obtenus par ***AntConc***, nous avons consacré un cinquième et dernier chapitre à l'enregistrement, puis au transfert des résultats obtenus avec ***AntConc*** vers le tableur ***Excel*** de ***Microsoft*** pour une analyse plus poussée. En effet, ces résultats, provenant de ***AntConc*** ou de tout autre logiciel lexicométrique, ne sont que données statistiques brutes que le linguiste doit croiser et interroger pour réaliser son analyse qualitative et aboutir à des conclusions pertinentes et rigoureuses. Quelques manipulations, avec l'outil ***Tableau croisé dynamique***, sont présentées et expliquées en détail. Le but est de montrer les avantages offerts par le différents filtrages et recoupements qui peuvent aider le chercheur en suggérant des pistes de recherche.

Pour plus de clarté, nous avons opté pour des captures d'écran qui permettent de visualiser et de suivre les étapes des manipulations. Ces captures comportent le plus souvent des numéros qui correspondent aux étapes de la manipulation décrite juste avant. Tout ce qui est présenté, nous l'avons testé à plusieurs reprises. Certaines manipulations nécessitant l'utilisation de corpus de textes ou de listes de mots ainsi que le programme de catégorisation ***TreeTagger***, ou le nouveau ***TagAnt***, nous invitons le lecteur qui veut utiliser ***AntConc***, à penser à créer, à l'avance, un répertoire avec des sous-répertoires pour recevoir les textes numérisés, les listes à préparer et les résultats, pour se donner l'occasion de vérifier par lui-même les réalisations présentées.

Le concepteur d'***AntConc*** a également développé une vingtaine de programmes[6] qui gravitent autour du concordancier, pour préparer, convertir, partitionner, encoder des fichiers, entre autres.

[5] Cécile Fabre, Didier Bourigault. « Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants ». Journal of French Language Studies, Cambridge University Press (CUP), 2008, 18 (1), pp.87-102.

[6] Téléchargeables gratuitement à l'adresse https://www.laurenceanthony.net/software.html.

Préparation

Avant de commencer à manipuler le logiciel, il est recommandé de créer à l'avance un répertoire – Appelez-le ***AntConc*** ou un nom au choix. Créez dans ce répertoire des sous-répertoires et donnez-leur des noms explicites, pour recueillir les textes au format .txt[7], qu'ils soient scannés et océrisés[8] ou téléchargés sur Internet, les listes et les applications à utiliser lors de certaines manipulations (***Stop List, French Lemma List***, etc.)[9] et les résultats des requêtes.

[7] En fouinant sur Internet, on peut trouver des textes du domaine public, par exemple à partir du Projet Gutenberg (http://www.gutenberg.org/) ou des corpus de presse sur le site du Centre National de Ressources Textuelles et Lexicales – Cnrtl (https://www.cnrtl.fr/corpus/), et bien d'autres encore. Voir aussi notre rapide présentation d'*AntConc* dans Jacques François, Yacoub Ghérissi. « Pour une linguistique orientée outils : la polysémie du verbe compter et les genres textuels ». 2012. ⟨https://hal.archives-ouvertes.fr/hal-01811292⟩.

[8] Du sigle OCR (optical character recognition), reconnaissance optique des caractères.

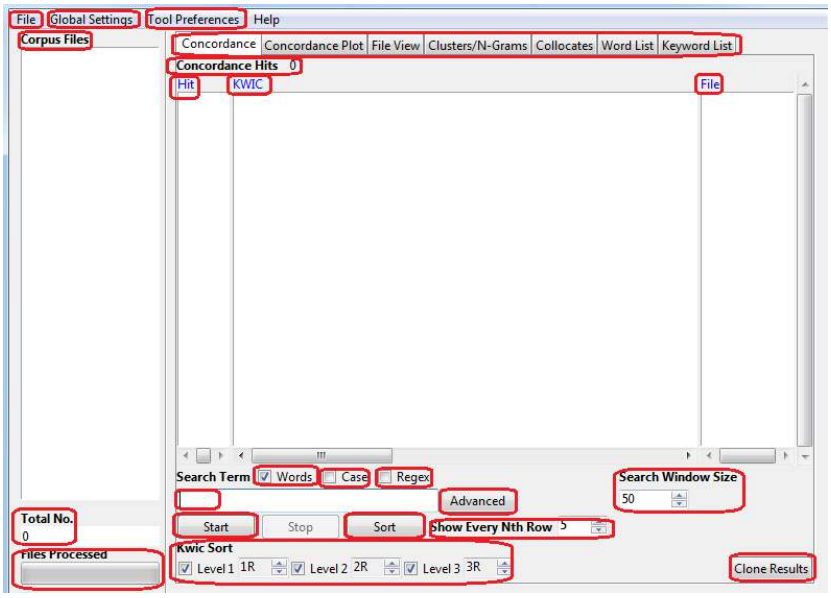
[9] Voir les références dans les § correspondants.

Chapitre Premier
La fenêtre principale

1.1. La fenêtre principale

Dans la version ***AntConc w3.5.8***, qui nous intéresse présentement, la fenêtre principale du concordancier présente 3 menus et 7 outils paramétrables selon plusieurs options.

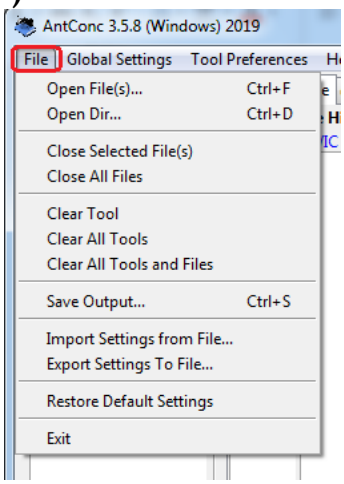
Nous allons d'abord expliquer brièvement les contenus des fenêtres de dialogue des menus et des outils en proposant leurs traductions en français. Ensuite nous aborderons les détails des outils et de leur manipulation.



En voici un premier aperçu, juste pour expliquer rapidement les menus et les fonctions des boutons qui seront détaillés après.

File	Fichier
Global Settings	Paramètres généraux
Tool Preferences	Préférences de réglage / Paramètres spécifiques
Corpus files	Liste des fichiers proposes au traitement
Concordance … Keyword List	Les 7 outils de recherche
Concordance Hits	Nombre d'énoncés trouvés
Hit	Numérotation des énoncés
KWIC	Mots-clés en context (*Key Word in Contexte*)
File	Noms des fichiers
Words	Recherche par mot graphique
Case	Recherche sensible à la casse
Regex	Recherche avec des expressions régulières
Search Term	Zone de saisie des requêtes
Advanced	Recherche avancée
Search Window Size	Compteur des caractères des cooccurrents
Total N°	Total des fichiers proposés
Start	Validation de la recherche
Sort	Filtrage – classement
Show every Nth Raw	Afficher chaque nième ligne
Files Processed	Progression de la recherche
Kwic Sort	Filtrage des résultats
Clone Results	Clonage des résultats

1.1.1. File (le menu Fichier)

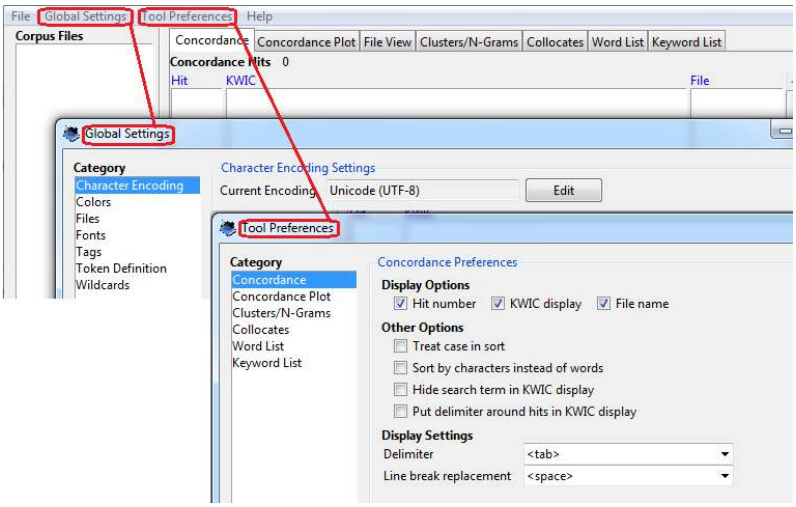


Pour le menu ***File***, voici les traductions des sous-menus en anglais, mais qui ne sont pas très difficiles à comprendre pour un francophone habitué aux logiciels.

Open File(s)File(s)	Ouvrir un ou plusieurs fichiers
Open Dir…ectory	Ouvrir tous les fichiers d'un répertoire
Close Selected File(s)	Fermer le ou les fichiers sélectionnés
Close All Files	Fermer tous les fichiers
Clear Tool	Effacer le dernier réglage
Clear All Tools	Effacer tous les réglages
Clear All Tool and Files	Fermer les réglages et les fichiers
Save Output	Enregistrer les résultats
Import Settings from File	Importer les réglages depuis un fichier
Export Settings To file	Exporter les réglages vers un fichier
Restore Default Settings	Restaurer les réglages
Exit	quitter

1.1.2. Les réglages préalables

AntConc offre deux types de configuration pour les réglages préliminaires (***Global Settings***) et les outils, ou paramètres spécifiques (***Tool Préférences***).



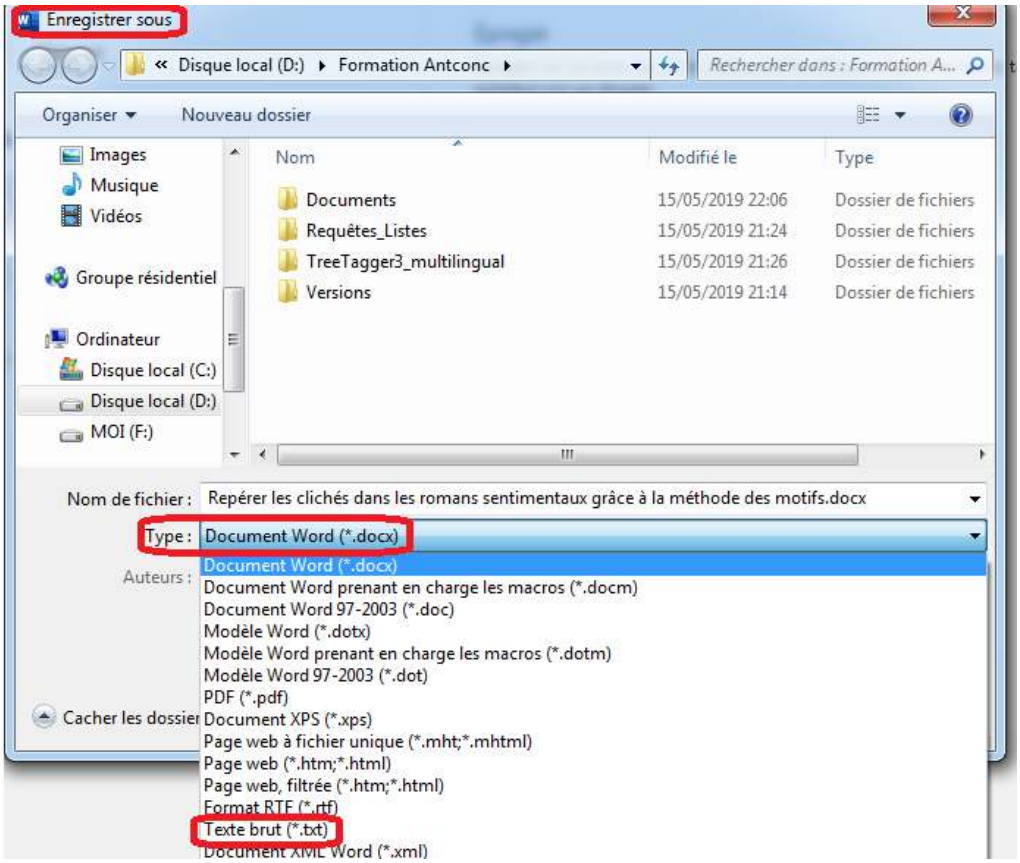
1.1.2.1. Global Settings (paramètres généraux)

1.1.2.1.1. File (type de fichier)

AntConc fonctionne, par défaut, sur des fichiers au format « Texte brut », reconnaissables à leur extension (.txt). Les fichiers de type document (*.doc) comportent un en-tête et diverses informations sur la mise en forme, l'auteur, les dates de création et de toutes les versions du fichier ainsi que des statistiques, ce qui le rend plus volumineux qu'un texte brut.

Si le chercheur ne dispose que de fichiers ***Word***, il est facile d'en faire des copies en les convertissant un à un en documents du type Texte Brut[10]. Un logiciel gratuit, comme ***Corpus Text Processor***[11] est capable de convertir plusieurs fichiers à la fois, à partir de et vers les principaux formats utilisés par les périphériques de lecture.

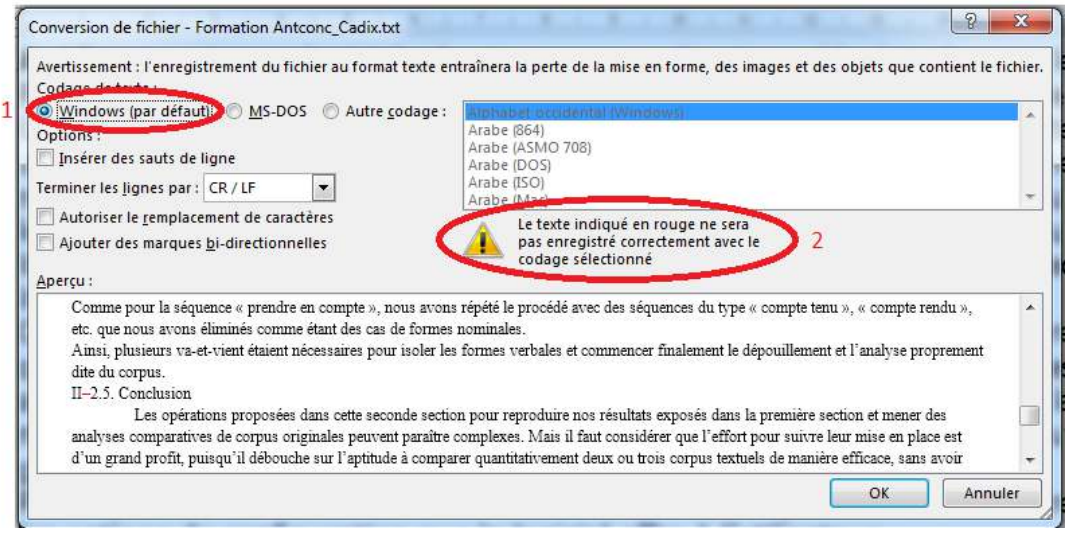
Pour convertir rapidement un fichier ***Word*** (.doc ou .docx) en fichier (.txt), il faut enregistrer le fichier source (Word) sous le type texte brut (*.txt) en gardant le même nom de fichier. Un nouveau fichier est créé dans le même répertoire, avec l'extension .txt.



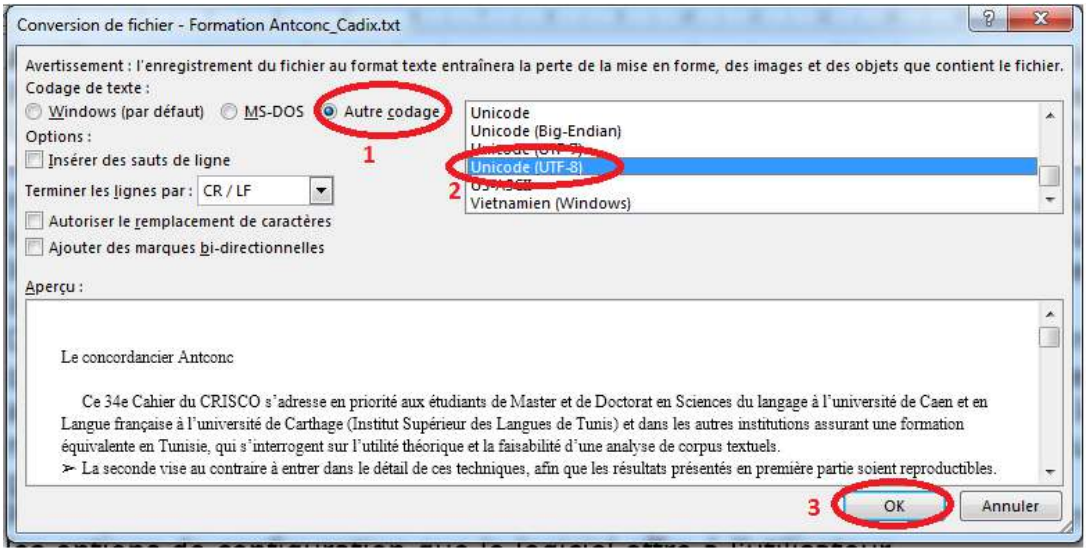
[10] Pour convertir des fichiers ***Word*** ou ***PDF*** en ***.txt***, le site de Lawrance Anthony propose de télécharger AntConverter , et pour obtenir des fichiers dans le bon encodage, il recommande EncodeAnt. Voir note 6.

[11] https://github.com/writecrow/ciabatta/wiki/1.-Tools:-Corpus-Text-Processor#installation

Lors de la conversion, Word signale (1), parfois, avec un triangle jaune (2), l'existence de caractères (symboles ou signes de langues non romanes) qu'il ne prend pas en charge.



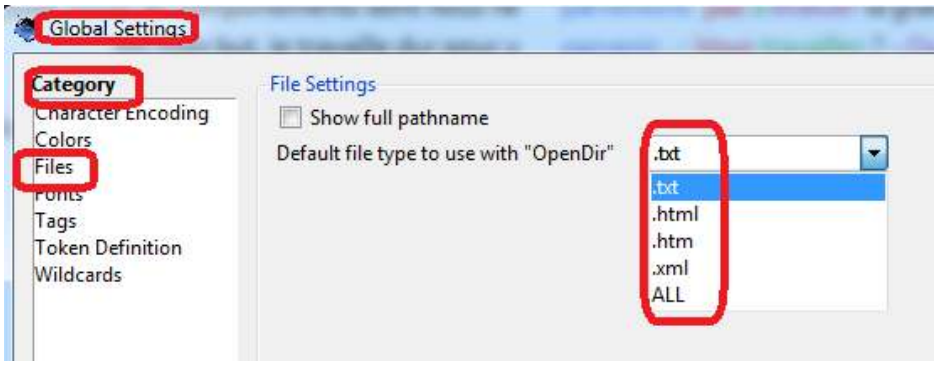
De préférence, avec des caractères refusés ou non, il est recommandé de changer le codage (1) et de choisir dans la liste déroulante (2) le codage Unicode (UTF 8) et valider avec le bouton OK (3).



Une fois ***AntConc*** lancé et le(s) fichier(s) à traiter choisi(s), ***AntConc*** propose une option concernant le type de fichier à traiter.

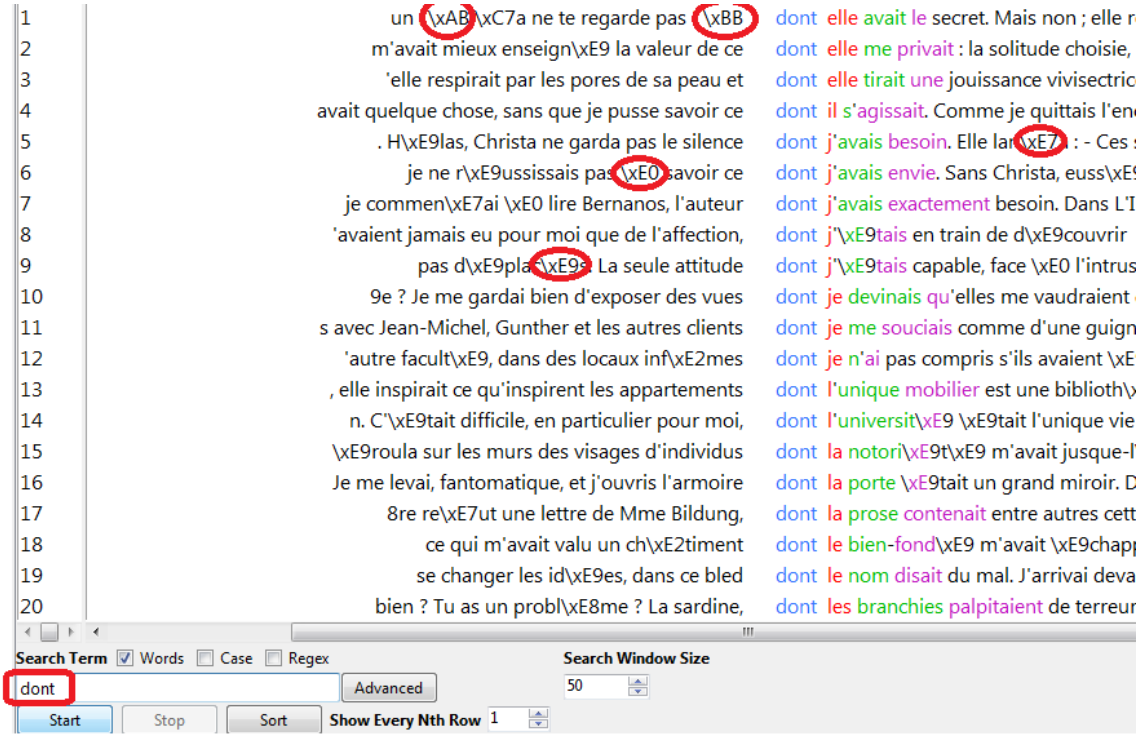
En outre ***AntConc*** peut être réglé pour accepter d'autres formats de fichiers : .html, .htm[12], .xml[13]. Le réglage est accessible avec le menu Global Setting > Category > Files.

[12] Le HyperText Markup Language, généralement abrégé HTML, est le langage de balisage conçu pour représenter les pages web. C'est un langage permettant d'écrire de l'hypertexte, d'où son nom.
[13] XML est un langage de balisage générique qui permet de structurer des données afin qu'elles soient lisibles aussi bien par les humains que par des programmes de toutes sortes

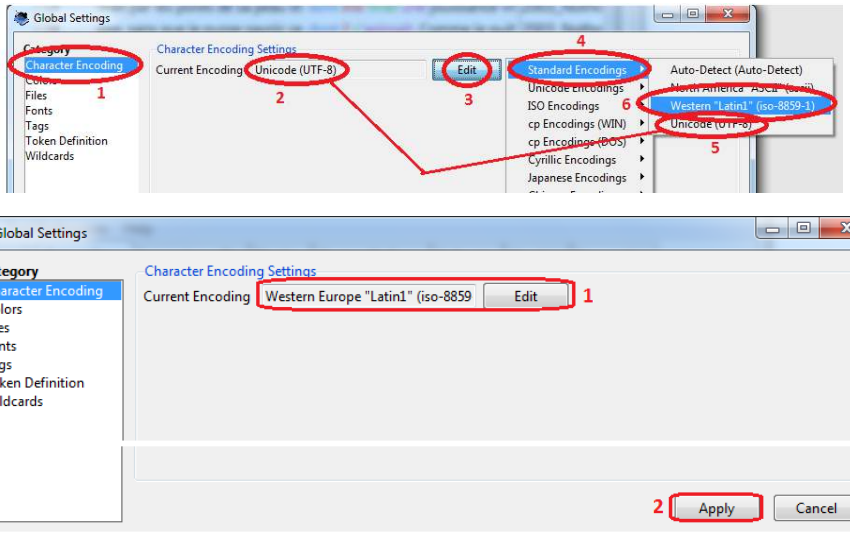


1.1.2.1.2. Character Encoding (encodage du fichier)

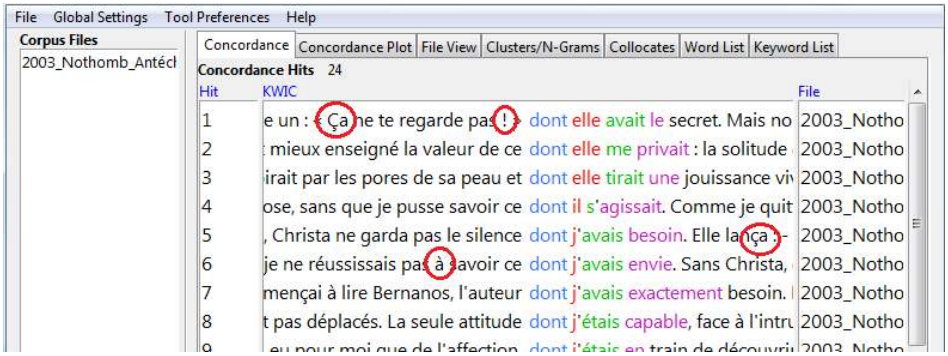
Pour vérifier que l'encodage est le bon, il faut, de se lancer dans la recherche, tester le codage en lançant une requête avec un mot simple qui ne doit pas comporter de lettres accentuées (é, è, ç, à, ê, etc.), par exemple **dont**.



Si, à la place des lettres accentuées, le résultat affiché comporte des symboles de type **\x** suivis d'une lettre majuscule et d'un chiffre ou de tout autre symbole comme **\xE9**, c'est que le codage du texte est erroné. Le menu ***Global Settings > Character Encoding*** (1) **>** ***Edit*** (3) **>** ***Standard Encodings*** (4) permet de changer le codage. Très souvent, il suffit de choisir ***Western "Latin1" (iso-8859-1)*** (5).



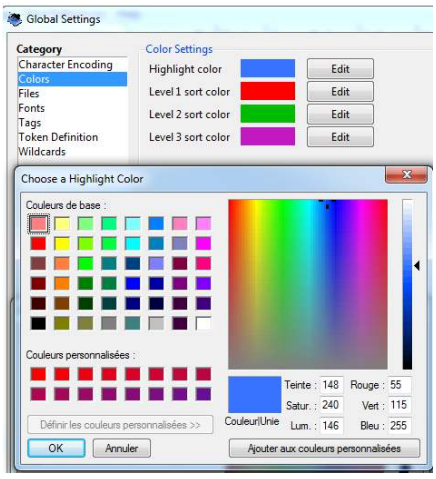
En validant, et en recommençant la requête-test, si le problème est réglé …



on peut passer à la deuxième étape, celle des réglages du programme avec le menu ***Tool Preferences***.

1.1.2.1.3. Colors (couleurs)

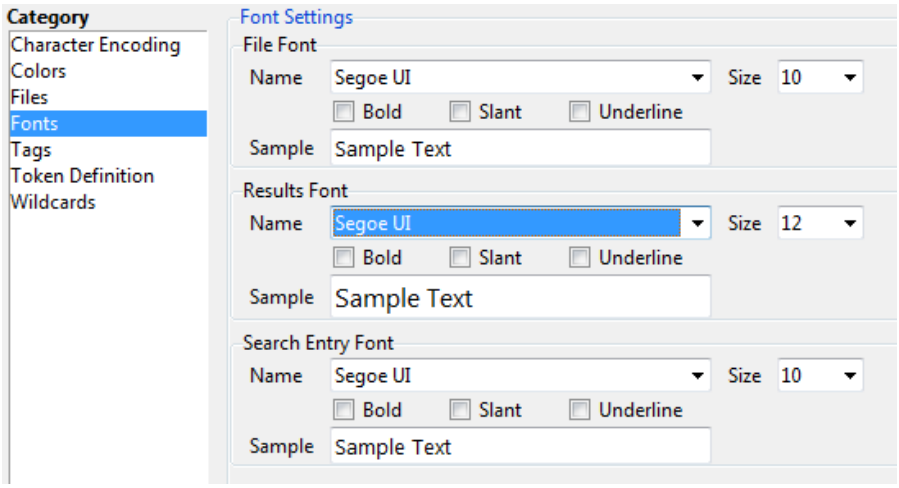
AntConc propose de changer les couleurs des niveaux de filtrage[14] pour plus de visibilité.



[14] Voir plus bas le filtrage du résultat (§ 2.1.3.).

1.1.2.1.4. Font (polices de caractères)

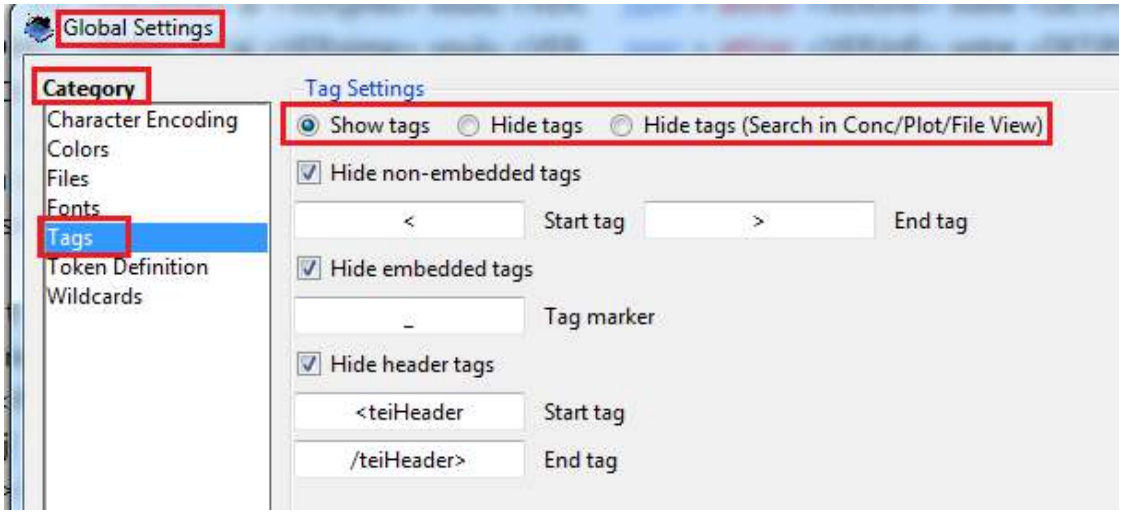
Le paramétrage concerne les polices de caractères à utiliser. Ces options ne sont pas vraiment très utiles pour la linguistique de corpus. Mais de préférence, le choix doit porter sur une police de caractères installée sur la machine utilisée.



1.1.2.1.5. Tags (balises)

En informatique, on utilise, par exemple en HTML, dans les codes sources, des tags. Ce sont des balises reconnaissables aux chevrons <…>, qu'on assigne aux mots ou suites de mots pour décrire leurs caractéristiques. En linguistique de corpus, on place entre chevrons des données sémantiques, syntaxiques, lexicales, etc., pour pouvoir opérer des regroupements faciles des données présentant les mêmes informations.

L'option ***Tag Settings*** permet de choisir, comme on le verra quand on abordera la question de la catégorisation, de cacher les balises (***Hide tags***) ou de les montrer (***Show tags***).



1.1.2.1.6. Token Definitions (jokers)

Pour définir le jeu des caractères, ***AntConc*** propose l'option, par défaut, de ne considérer que les lettres ***Letter Token Calasses*** (1), et/ou d'inclure dans la recherche les chiffres en cochant ***Number Token Classes*** (2) et les ponctuations ***Punctuation Token Classes*** (3).

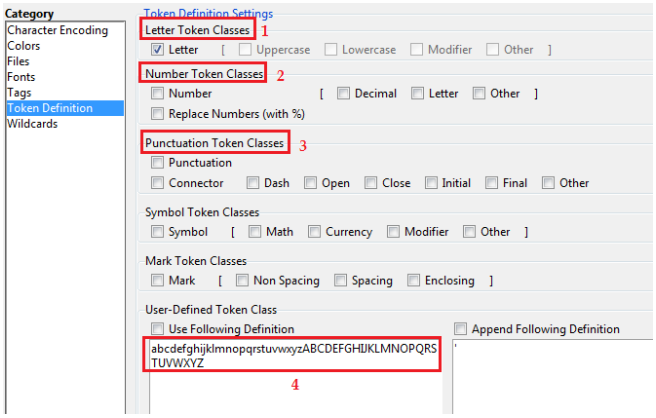
Par défaut donc, ***AntConc***, avec l'option ***Letter***… (lettre), indique qu'il considère comme lettres, au sens le plus large, les lettres minuscules (de a à z) et les lettres majuscules (de A à Z). Il est possible de définir sa propre définition des jokers ou d'ajouter des caractères aux classes standard. Pour les langues non latines, un réglage des éléments graphiques non-séparateurs de mots permet avec ***Use Following Definition*** (4) d'ajouter, par exemple, les graphèmes spécifiques de l'alphabet Tamazight pour que le logiciel reconnaisse les textes écrits dans cette langue.[15]

Le manuel d'utilisation qui accompagne le programme signale que le logiciel « est entièrement compatible Unicode, ce qui signifie qu'il peut gérer les données dans n'importe quelle langue, y compris toutes les langues européennes, l'arabe et les langues asiatiques ».

Pour des recherches, par exemple, sur les textes des Sms, qui utilisent des chiffres en remplacement de certaines syllabes[16], il est recommandé de cocher, au préalable, l'option ***Number Token Classes*** (2) pour faire apparaître les mots qui comportent des chiffres.

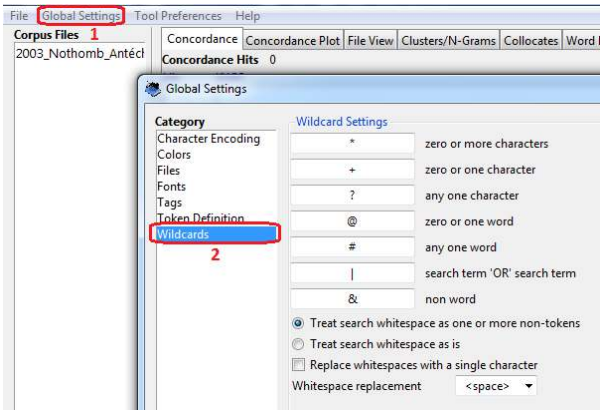
[15] Tigziri, N. (2016). « Analyse textuelle à l'aide du concordancier ANTCONC d'une oeuvre de Belaïd Ait-Ali ».

[16] n8 pour « nuit », 7 pour « cette », etc. Les Tunisiens, par exemple, quand ils utilisent le clavier latin, recourent beaucoup aux chiffres 3, 5, 7 et 9 pour transcrire certains phonèmes spécifiques à l'arabe. Ces chiffres imitent approximativement la calligraphie des lettres correspondantes.



1.1.2.1.7. Wildcards (métacaractères)

Dans le menu ***Global Settings***, et le sous-menu ***Wildcards***, ***AntConc*** propose 7 opérateurs logiques qui facilitent la recherche en utilisant des opérateurs logiques à la place des mots, des jokers en quelque sorte.



Ces opérateurs servent à lancer des recherches avec des expressions régulières (***regex***). Ce sont des métacaractères qui ont, chacun, comme le montre le tableau suivant, une signification particulière.

Voici leur traduction.

*	Le mot saisi seul ou suivi d'un ou de plusieurs caractères
+	Le mot saisi seul ou suivi d'un seul caractère (avec possibilité d'augmenter le nombre)
?	Le mot saisi obligatoirement suivi d'un seul caractère (avec possibilité d'augmenter le nombre)
@	Le mot saisi seul ou suivi d'un seul mot (avec possibilité d'augmenter le nombre)
#	Le mot saisi suivi obligatoirement d'un seul mot (avec possibilité d'augmenter le nombre)
|	L'un ou l'autre des mots saisis
&	Le mot en fin de phrase

1.1.2.2. Tool Preferences (réglage des outils)

La fenêtre principale donne accès à sept outils : ***Concordance***, ***Concordance Plot***, ***File View***, ***Clusters/N-grams***, ***Collocates***, ***Word List*** et ***Keyword List***. Ils sont directement activables à l'aide des raccourcis F1 à F7.

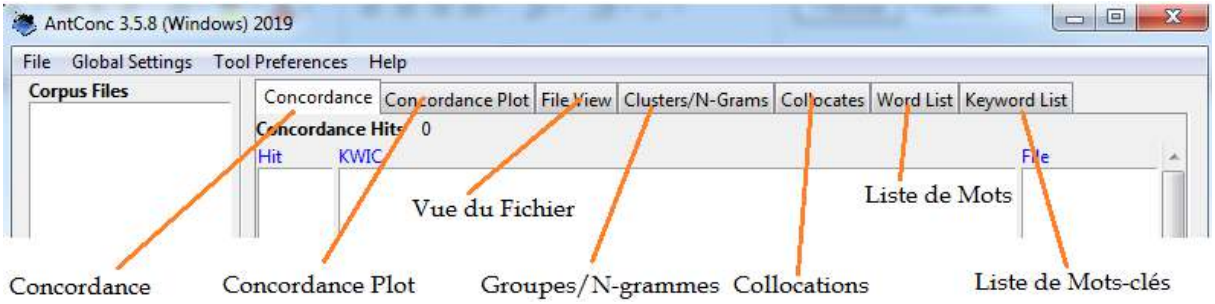
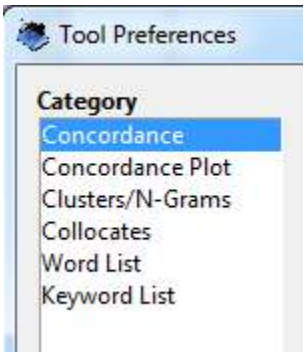


Figure x : Les onglets correspondant aux outils

A chaque changement dans les réglages, aussi bien avec ***Global Settings*** ou ***Tool Preferences***, il faut valider en cliquant sur le bouton ***Apply*** en bas à droite de la fenêtre.

Le menu ***Tool Preferences*** présente les options qu'il faut configurer avant de lancer chaque outil.

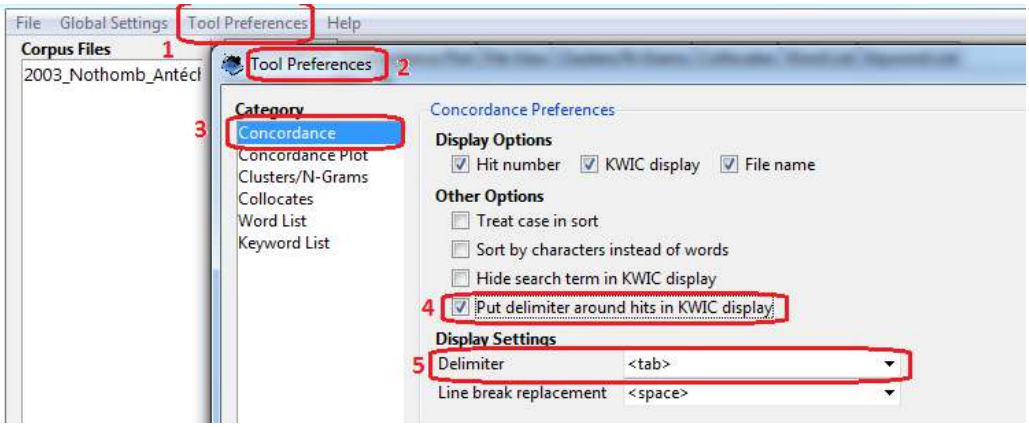


Chapitre Deuxième
Le fonctionnement des outils

2.1. Concordance

La première opération sert à préparer les résultats pour une meilleure utilisation ultérieure, notamment avec le tableur ***Excel*** de ***Microsoft Office***.

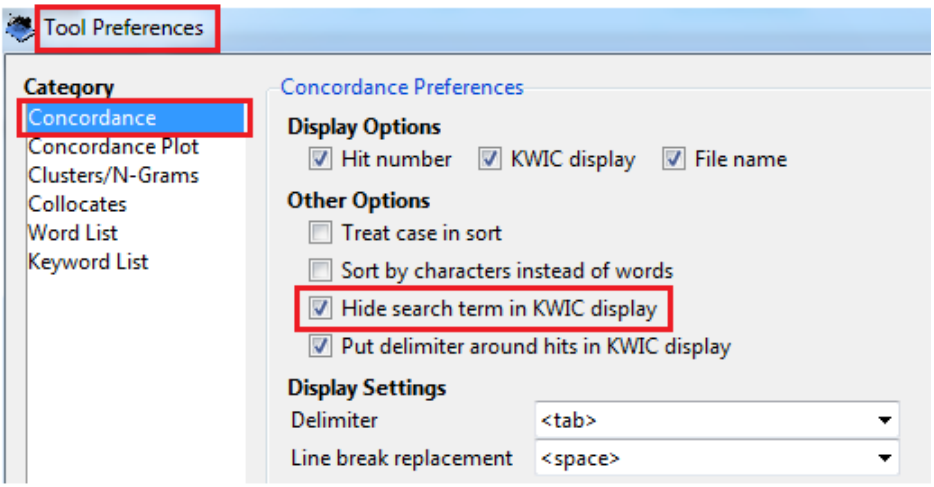
Pour ce faire, dans la commande ***Tool Preferences*** (1-2) > ***Concordance*** (3), on coche la case ***Put delimiter around hits in KWIC display***[17] (4) et on garde l'option ***Delimter*** <tab> (5) afin d'anticiper sur les résultats à obtenir pour que les termes de la requête soient isolés dans des colonnes différentes.



Après le lancement d'une requête, un espace est visible à gauche du mot recherché, ce qui signifie que la configuration avec tabulations est bien fonctionnelle. Cet espace correspondra, plus tard dans ***Excel***, à une colonne vide à gauche de la colonne du mot-clé.

[17] KWIC est l'acronyme de l'anglais ***Key word in context*** et signifie "mot-clé en contexte". Pour ***Delimiter***, il s'agit des possibilités pour choisir d'autres séparateurs dans le menu déroulant qui propose (,), (:) et (;) en plus de <tab> pour ***tabulation*** c.-à-d. la possibilité de trouver le mot pivot (ou mot-clé) dans une colonne, précédé du contexte gauche et suivi du contexte droit. Pour manipuler les résultats dans Excel, l'option <tab> est la plus appropriée.

D'autres options sont proposées. Il est possible de cacher les mots-clés d'une concordance et de ne garder que les contextes gauche et droit. Cela sert, entre autres, à créer des exercices à trous pour le cours de langue.



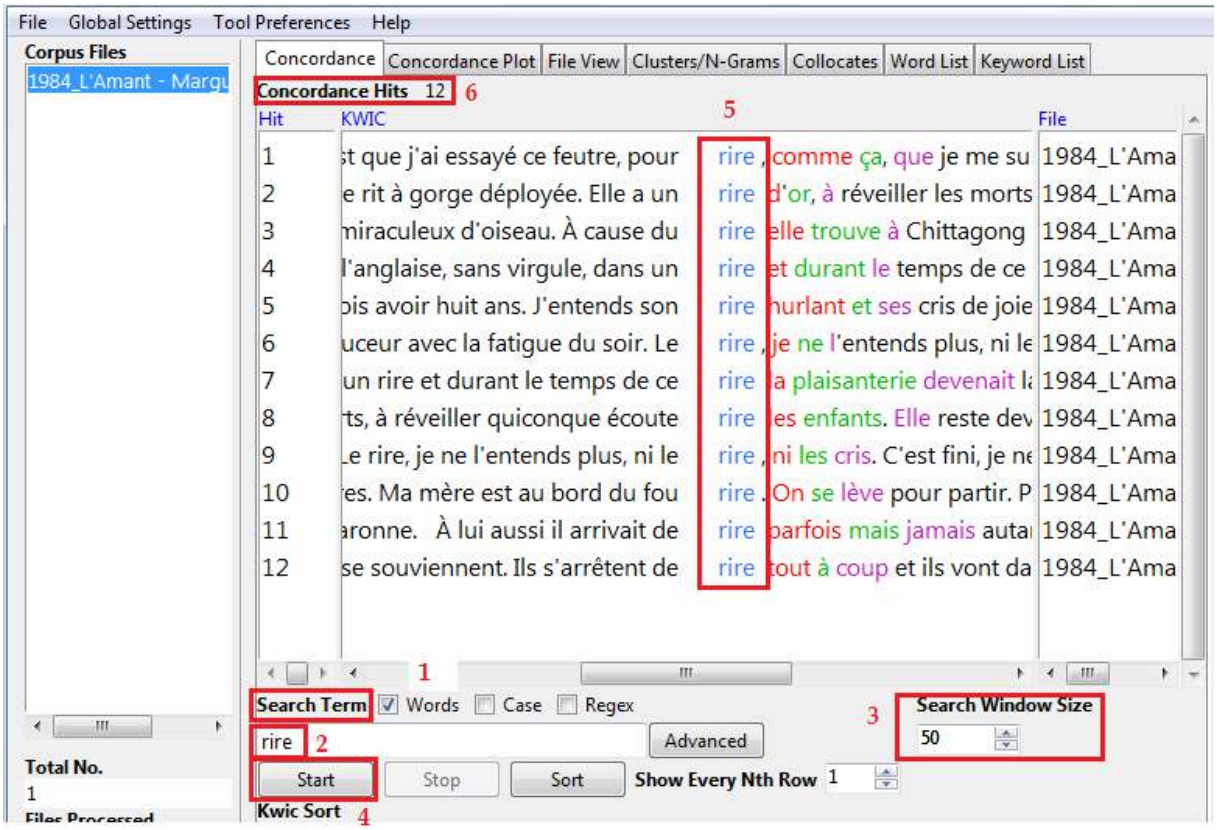
Le résultat de la requête n'affiche pas les mots-clés.

Pour obtenir un fichier de concordance avec un seul mot ou un groupe de mots[18] à rechercher, ***AntConc*** propose deux possibilités.

2.1.1. La recherche (Search Term)

La première possibilité est simple. Il faut laisser la case ***Word*** (1), cochée par défaut, et saisir sous ***Search Term*** dans la fenêtre de saisie (2), un mot ou une suite de mots puis lancer la recherche (ici le verbe ***rire***). Avec ***Search Window Size*** (3), ***AntConc*** donne la possibilité d'élargir la taille du contexte autour du mot-clé en réglant la longueur des contextes en nombre de caractères avant et après le mot-clé, de 0 à 1000 par pas de 5.

Enfin on valide la recherche en cliquant sur le bouton ***Start*** (4).



Le résultat est affiché et le nombre d'occurrences trouvées est indiqué au-dessus de la fenêtre des résultats, ***Concordance Hits*** (6).

Cependant, une requête avec la graphie d'un verbe à l'infinitif ne génère que la concordance du verbe à l'infinitif. Étant un concordancier morphologique, ***AntConc*** ne cherche et ne trouve donc que ce qui est saisi, c.-à-d. seulement la forme graphique tapée (2 et 5) et génère un contexte comme demandé avec ***Search Window Size*** (3) et filtré selon ***Kwik Sort*** (4).

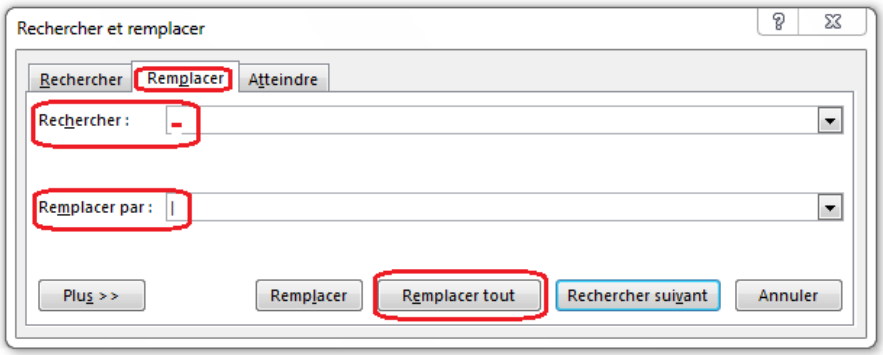
[18] Antconc distingue les mots simples et les groupes de mots (structures libres ou figées) tels qu'on les saisit dans un texte normal. Si on saisit deux ou trois mots séparés par une espace, le programme ne cherche pas chaque mot de la requête isolé des autres, mais la suite de mots, considérée comme un seul mot.

Pour gérer des requêtes de longueurs variables, comme rechercher, par exemple, le paradigme de conjugaison d'un ou de plusieurs verbes (exemple, la capture suivante avec le verbe ***rire***), il suffit de copier, à partir du fichier *French Lemma List* qu'on peut télécharger gratuitement sur le site du développeur Laurence Anthony[19], les paragraphes correspondants en prenant soin d'ajouter le participe présent et éventuellement les participes passés pour les verbes dont les participes sont variables.

La recherche peut concerner des listes de mots, comme les noms ou les adjectifs de couleur, les mots relatifs à un domaine donné, etc. Il faut alors saisir manuellement ou copier à partir d'un fichier créé à l'avance, une série de mots séparés par un slash droit (|)[20], obtenu avec la combinaison des touches AltGr+6, pour signaler au programme qu'on cherche l'un ou l'autre des mots saisis. La liste ne doit pas comporter d'espace sauf entre les mots d'un même groupe de mots.

Pour automatiser l'opération, on peut saisir dans ***Word*** la liste avec des espaces puis remplacer toutes les espaces[21] par des slashs droits, comme dans l'exemple suivant pour le paradigme de conjugaison du verbe ***rire***.

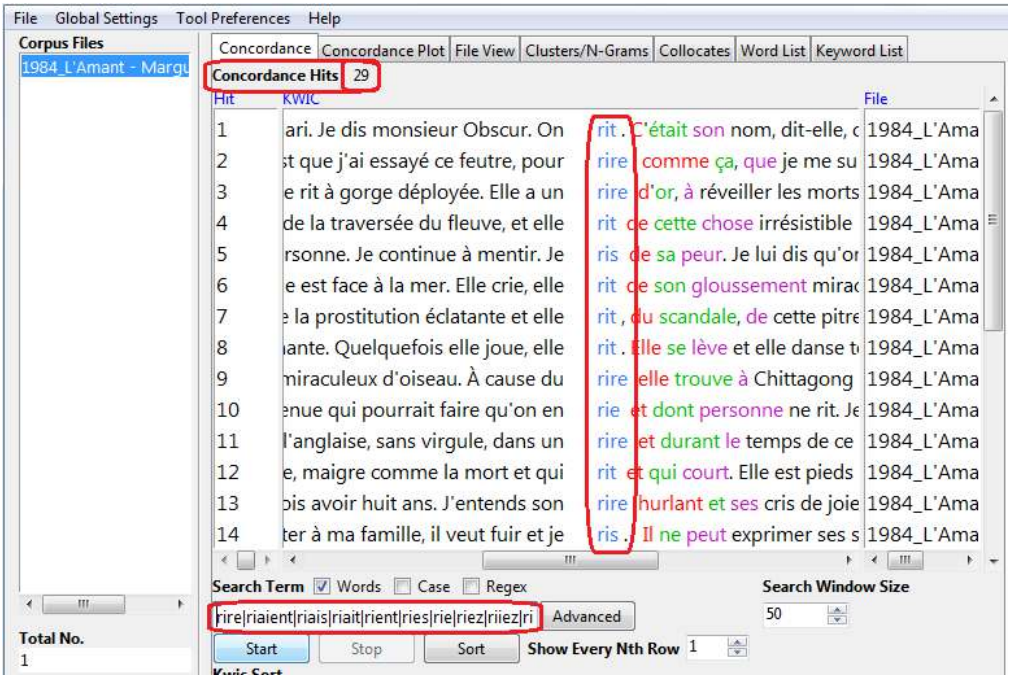
rire riaient riais riait rient ries rie riez riiez riions rîmes rions riraient rirait rirai riras rira rirent
rirez ririez ririons rirons riront rissent risses risse rissiez rissions ris rîtes rit rît rirais riant



On obtient la série suivante qu'on copie et qu'on colle dans la zone de saisie de ***Search Term*** avec la combinaison Ctrl+V.

rire | riaient | riais | riait | rient | rient | ries | rie | riez | riiez | riions | rîmes | rions | riraient | rirait | rirai | riras | rira | rirent | rirez | ririez | ririons | rirons | riront | rissent | risses | risse | rissiez | rissions | ris | rîtes | rit | rît | rirais | riant

[19] http://www.laurenceanthony.net/software/antconc/, à la rubrique **Lemma lists**.
[20] Voir, plus loin, les Jokers, au § 3.1.1.3.
[21] Le mot ***espace*** est masculin quand il signifie le lieu et féminin pour signifier la séparation des mots avec la barre du clavier, c'est un terme de typographie.

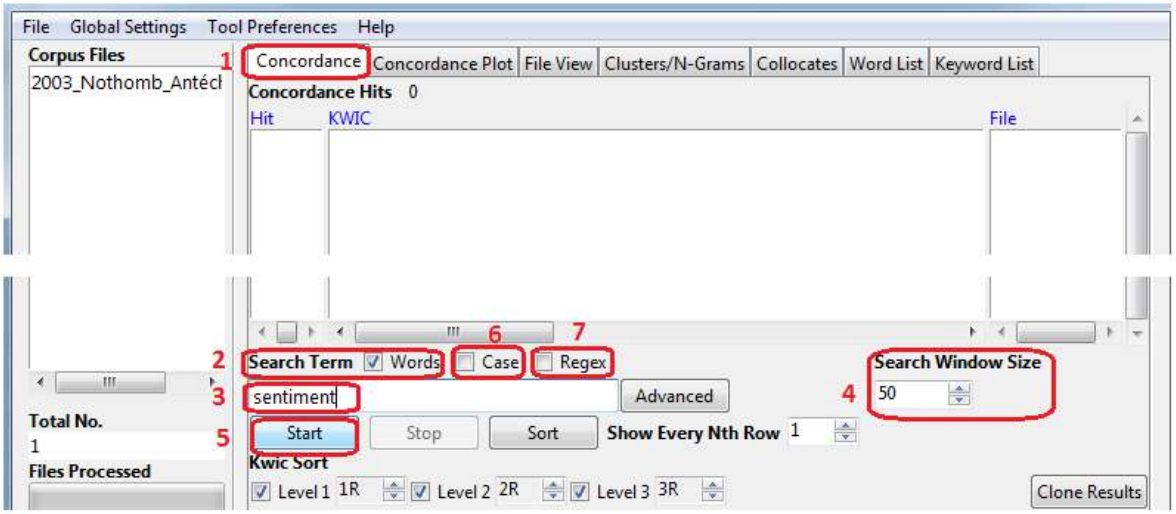


Pour une thématique donnée, on recherche des listes de mots dans les dictionnaires et/ou sur Internet pour constituer des requêtes et les préparer pour d'éventuelles recherches.

En cliquant sur une occurrence du mot-clé, l'onglet ***File View*** est immédiatement activé pour afficher l'occurrence dans son contexte élargi[22].

2.1.1.1. Les options de la concordance

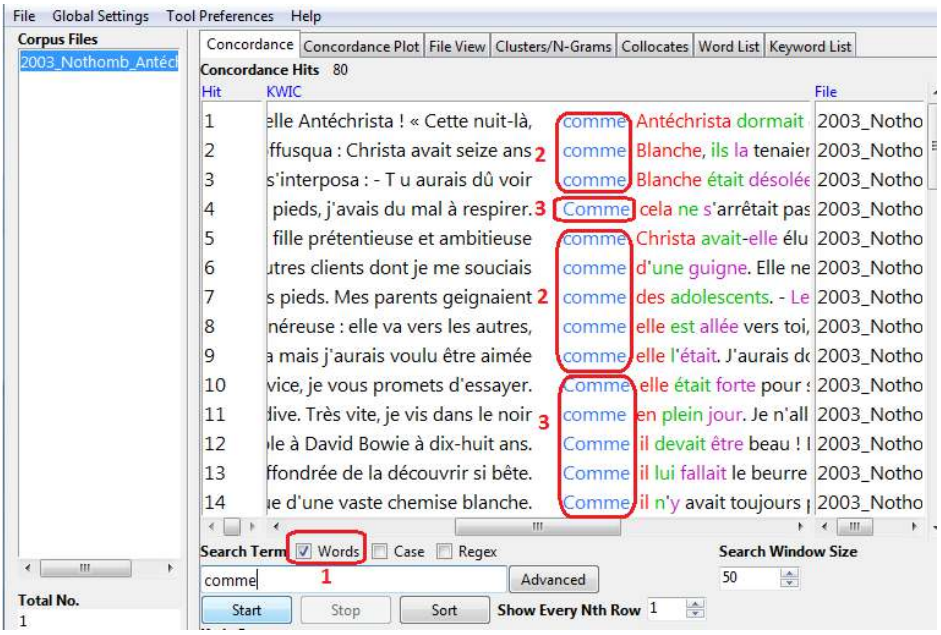
AntConc propose 3 options de recherche, ***Word*** (2), ***Case*** (6) et ***Regex*** (7).



[22] Voir § 2.3.

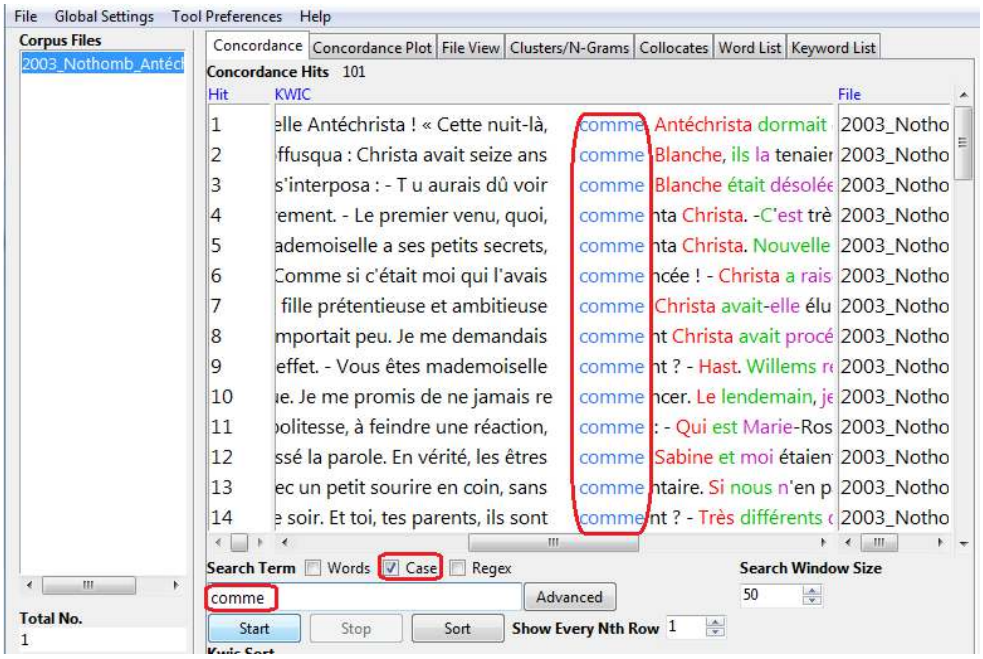
2.1.1.1.1. Recherche simple

Avec l'option ***Word*** (2), cochée par défaut et seule cochée, la recherche est effectuée sans prise en compte de la casse des caractères. Tous les cas (majuscules et minuscules) sont relevés.



2.1.1.1.2. Requête avec spécification de la casse

Une requête avec spécification de la casse (la case ***Case*** cochée) donnera seulement les mots à initiale en minuscule si on saisit la première lettre du mot en minuscule.



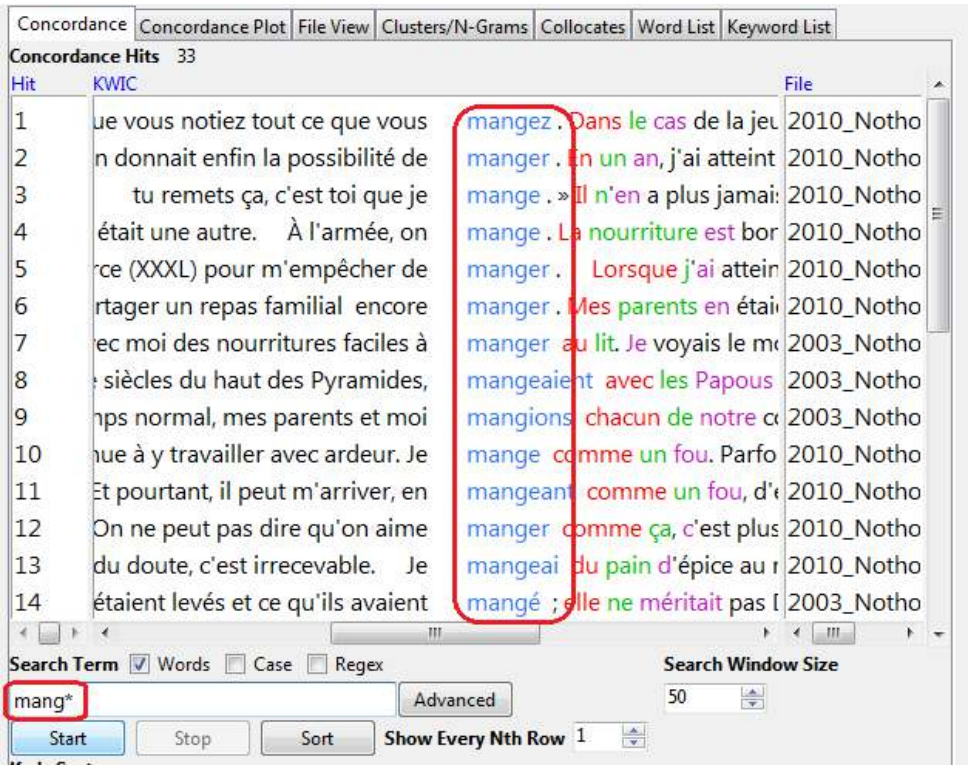
Si on saisit un mot en majuscule, le résultat ne concernera que les cas en majuscule. La recherche des majuscules permettrait de trouver les entités nommées (les Noms propres), mais également les débuts des phrases.

2.1.1.1.3. La recherche avec des expressions régulières

La recherche utilise les métacaractères signalés dans le § 1.2.1.7. ***Wildcards***.

1. L'astérisque (*)

L'astérisque permet de rechercher les caractères saisis ainsi que toutes les formes qui commencent par ces lettres, considérées comme un préfixe. Si l'astérisque est placé au début des caractères saisis, ces derniers seront considérés comme un suffixe. Cette requête est utile pour trouver les verbes à base unique comme ***chanter*** (chant*), ***manger*** (mang*), ou les mots se terminant par ***ation*** (*ation) ou ***ment*** (*ment).



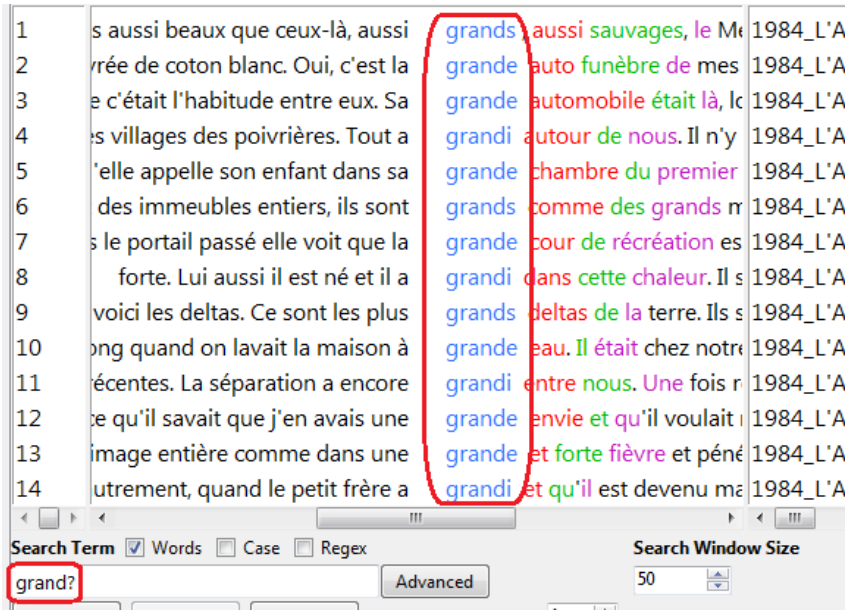
2. Le slash droit (|)

Le slash droit (appelé également ***pipe*** ou ***barre verticale***) est l'opérateur logique OU inclusif, qui recherche des exemples comprenant au moins l'un des termes saisis. Il a été question de cet opérateur plus haut[23].

Cette option peut être combinée avec ***Case***.

3. Le métacaractère ?

Quand le mot saisi est suivi de (**?**), le programme recherche le mot saisi suivi **obligatoirement** d'une seule lettre



[23] § 3.1.1.

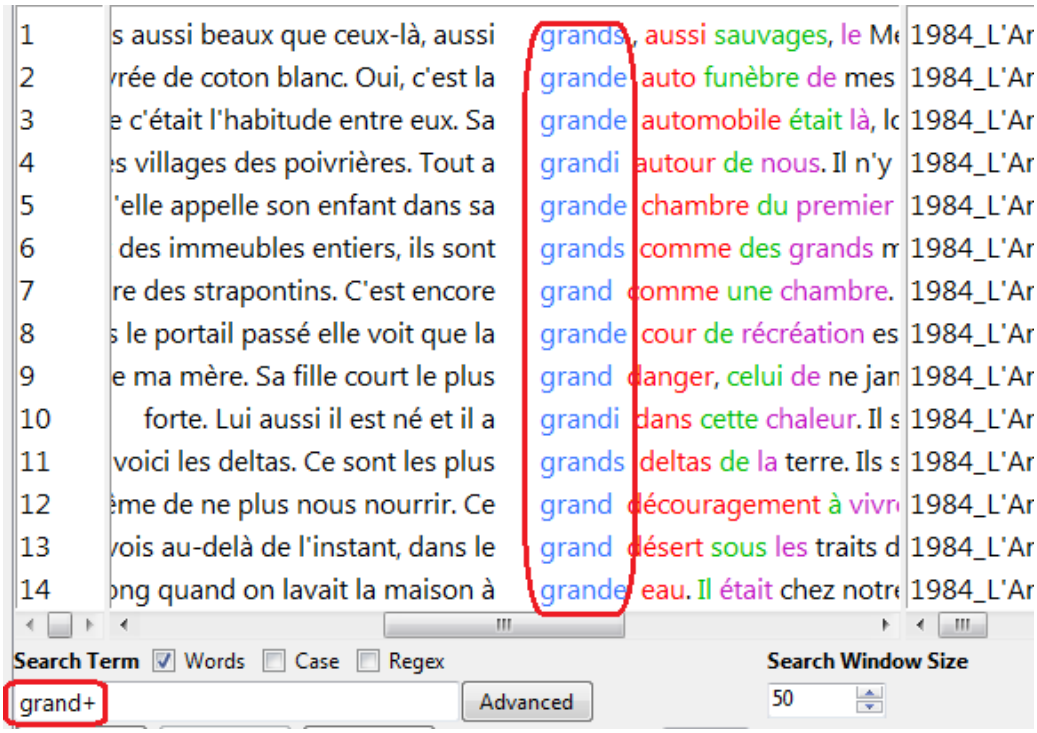
grand ? ciblera ***grandi, grande*** et ***grands.***

Cet opérateur peut rechercher une lettre à l'intérieur d'un mot comme dans les mots croisés.

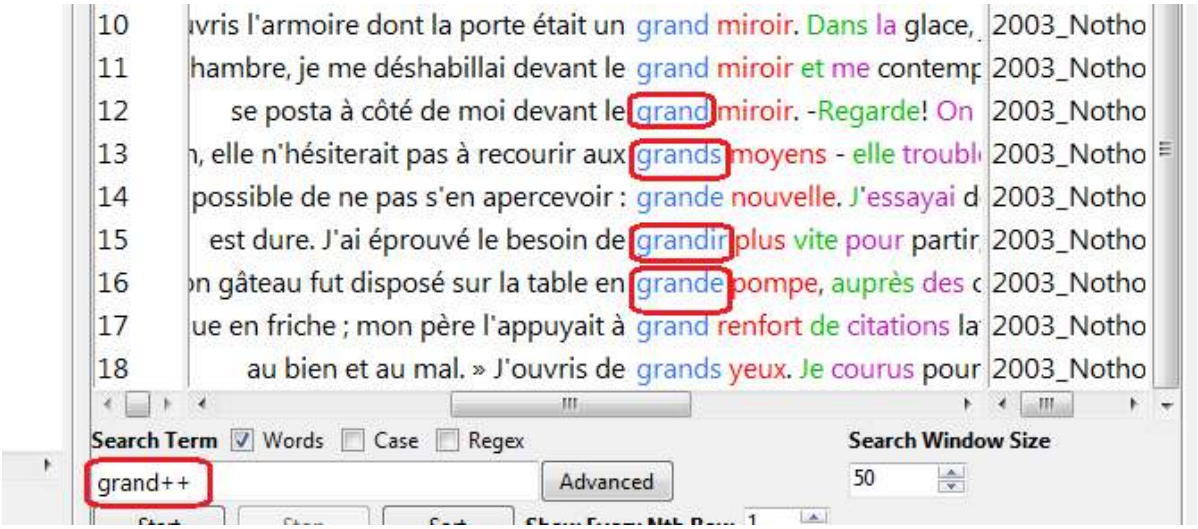
4. Le métacaractère +

Si le mot saisi est suivi de **(+)**, le programme recherche le mot saisi seul ou suivi d'une seule lettre. On peut utiliser ce caractère autant de fois que nécessite la variabilité d'un mot, par exemple.

grand+ donnera ***grand, grande*** et ***grands***



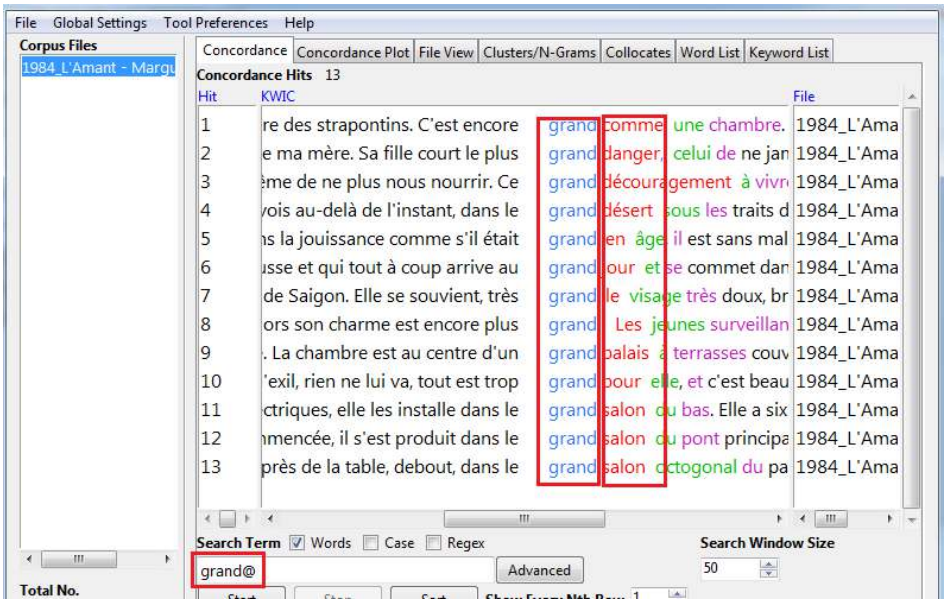
Avec (++), les formes recherchées sont le mot entier saisi **ainsi que** la même forme suivie d'une ou de deux lettres



grand++ devrait donner ***grand, grande, grands*** et ***grandir, grandis, grandit, grandie*** et ***grandît.***

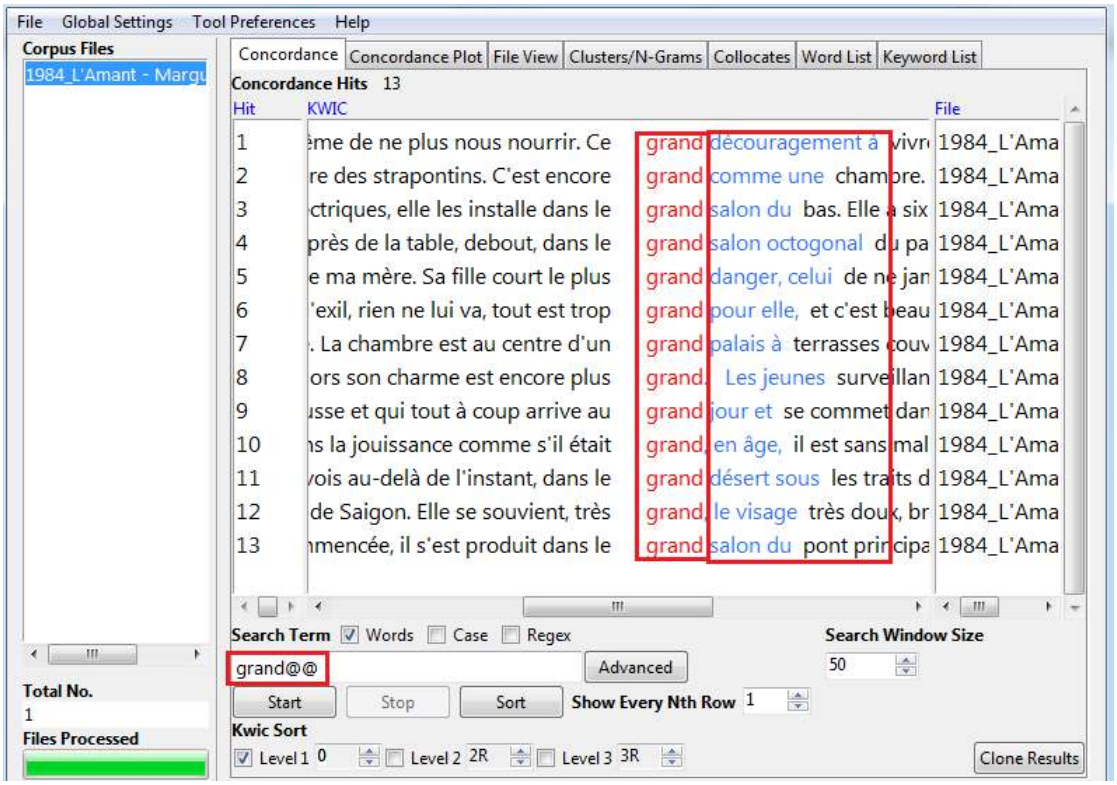
5. Le métacaractère (@)

Avec le dièse, une requête avec ***grand@*** donne ***grand*** suivi obligatoirement d'un seul mot.

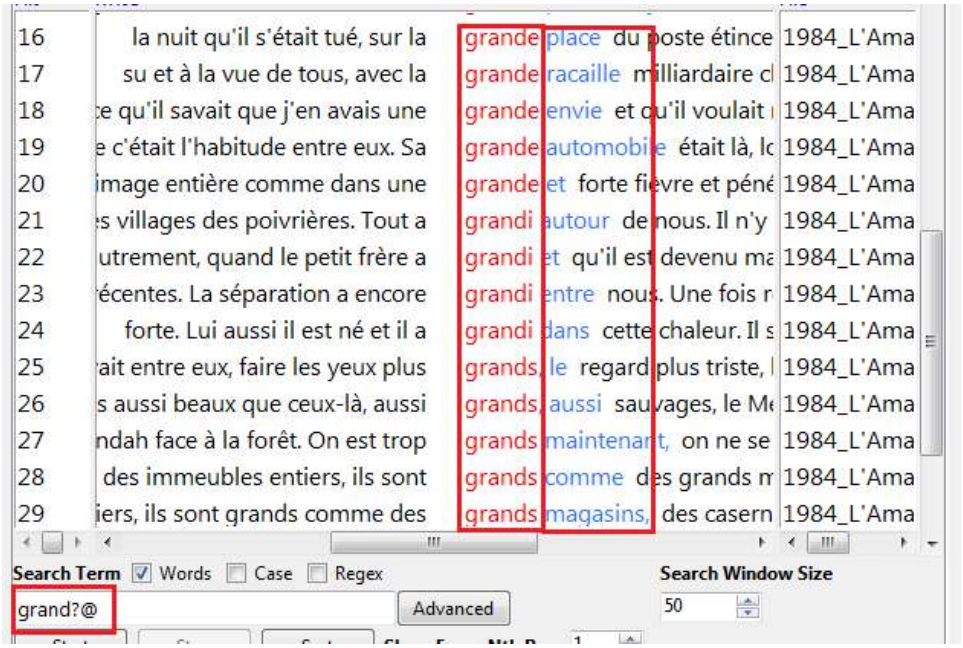


Les Jokers sont cumulables avec possibilité de répétition du même.

grand@@ donne ***grand*** suivi obligatoirement de deux mots.



grand ?@ donne ***grand*** suivi obligatoirement aussi bien d'une seule lettre (e, i ou s) que d'un mot (en bleu)

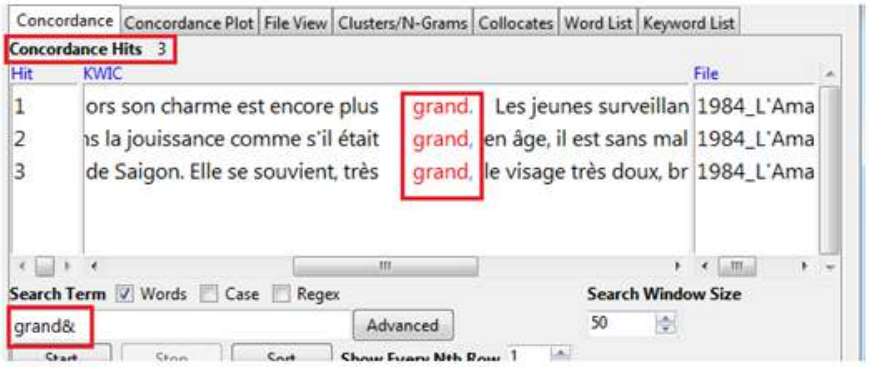


6. Le métacaractère (#)

L'expression régulière avec le dièse donne des exemples avec le mot saisi suivi d'une ponctuation ou non et d'un mot à sa droite.

7. Le métacaractère (&)

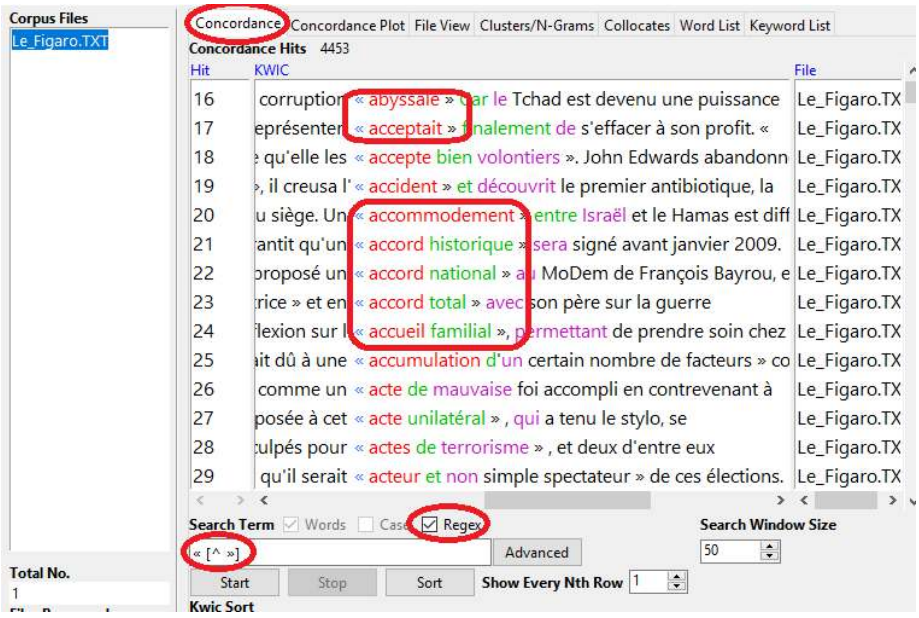
Ce signe typographique appelé ***perluète*** ou ***esperluette*** permet de retrouver les occurrences du mot saisi quand il est à la fin d'une structure, autrement dit suivi d'une ponctuation (, . ; : ? !)



Tous les opérateurs qu'on vient de voir peuvent être cumulés et/ou répétés.

8. Le métacaractère (^)[24]

Il est possible également de repérer toutes les séquences guillemetées en utilisant l'expression régulière **« [^ »]**. Les guillemets saisis doivent être du même type que celles du texte[25]. Il ne faut pas oublier de choisir l'option ***regex*** au niveau de ***Search Term***. Ainsi sont extraits les différents énoncés rapportés au discours direct tout comme les mots ou séquences que l'auteur du texte voulait signaler comme des emprunts, des néologismes ou employés métaphoriquement, etc.



[24] Le signe ^ est obtenu par la touche qu'il partage avec le tréma. Sur les claviers AZERTY, cette touche est appelée *touche morte* qui ne produit aucun résultat lorsqu'elle est enfoncée une seule fois puis relâchée. Dans ***Word*** elle modifie le comportement de la prochaine touche qui sera enfoncée, en ajoutant un accent circonflexe sur les voyelles : â, ê, î, ô, û et ŷ. Pour obtenir le signe ^ tout seul, il faut appuyer deux fois sur la touche, ce qui donne ^^, puis effacer le second.

[25] Les guillemets français « … » sont différents des guillemets anglais "...".

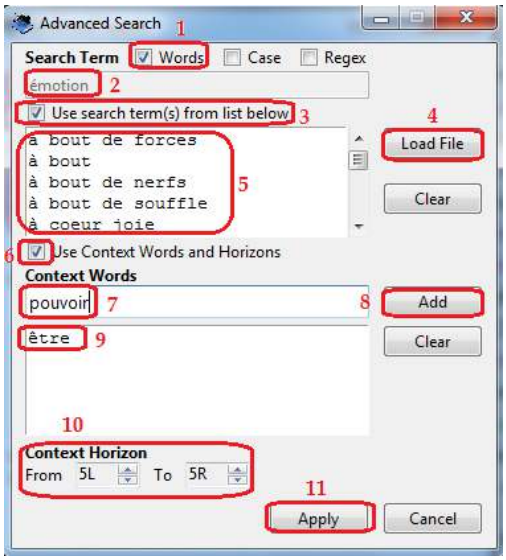
9. Le métacaractère (\)

Pour extraire les ponctuations (les points d'interrogation ou d'exclamation, par exemple), le requête doit comporter le symbole \ appelé anti-slash. Pour l'obtenir, il faut maintenir la touche **alt gr** enfoncée et appuyer sur la touche **8** qui sert d'habitude à taper le soulignement bas. La requête qui permet de retrouver les questions d'un texte est \? (3). Mais avant de cliquer sur ***Start*** (4), il faut cocher la case ***Regex*** (2) car dans la recherche simple, ***AntConc*** ne reconnaît que les lettres de l'alphabet, pour les ponctuations, il faut recourir aux expressions régulières[26].

2.1.2. La recherche avancée (Advanced)

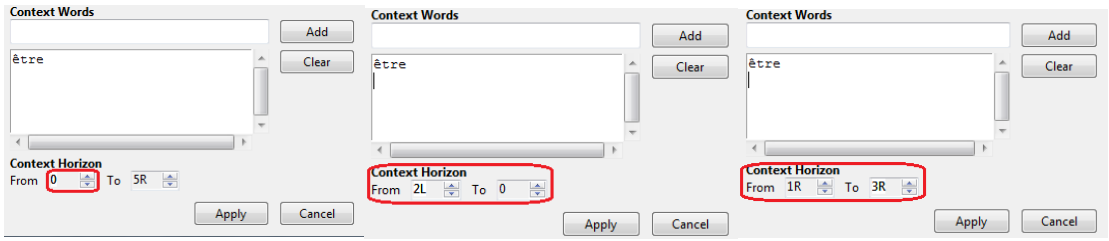
La seconde possibilité de recherche est réalisée avec le bouton ***Advanced*** (recherche avancée). Au préalable, un fichier avec l'extension .txt (texte brut) et comportant la liste des mots à rechercher placés chacun sur une ligne, doit être créé. En cliquant sur le bouton ***Advanced*** on obtient la fenêtre de dialogue suivante.

[26] ***Antconc*** est programmé entièrement avec le langage Perl 5.8. Pour les réglages, voir supra, § 1.2.1.6.

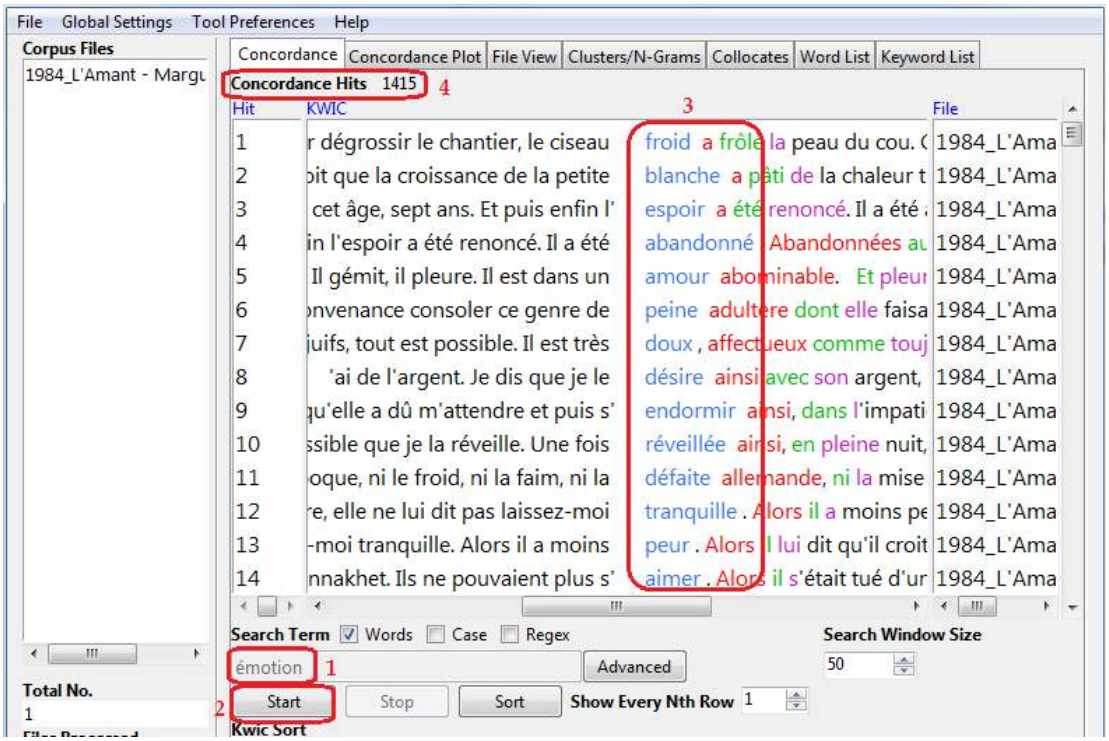


On laisse la case ***Word*** cochée (1), on nomme l'opération (ici ***émotion***) (2), on coche la case ***Use search term(s) from list below*** (3) et on demande l'ouverture du fichier qui contient la liste des mots qu'on veut rechercher ***Load File*** (4). Quand le début de la liste apparaît (5), on lance la recherche avec ***Apply*** (11).

Dans la même fenêtre, ***AntConc*** propose un second choix, ***Context Words***. Il s'agit de faire une recherche de mots de la liste ***Search Term*** déjà téléchargée, mais cette fois accompagnés d'un contexte (environnement phrastique) de longueur variable à souhait. Dans la fenêtre de dialogue ***Advanced***, on coche la case ***Use Context Words and Horizons*** (6) et on saisit un à un les mots qu'on voudrait trouver dans l'environnement des mots de la liste proposée et à chaque saisie on clique sur le bouton ***Add*** (8), pour valider le mot. Immédiatement, le mot est placé dans la fenêtre en dessous (9). On répète l'opération d'ajout autant de fois que de mots voulus. Enfin, on choisit la distance, mesurée en mots graphiques, qui séparerait le mot recherché (5) du mot qu'on voudrait qu'il se trouve dans son environnement (10). Un compteur permet de choisir cette distance et le côté droit (symbolisé par L "left") et/ou gauche (R pour "right") par rapport au(x) mot(s) de la liste proposée dans ***Use search term(s) from list below***. Dans l'exemple de la capture précédente, **From 5L to 5R** veut dire qu'on cherche les mots de la liste « émotion (dans l'exemple) » (2) accompagnés des mots ***être*** et ***pouvoir*** qui pourraient se trouver, au choix, à 5, 4, 3, 2, 1 mots à droite et également à 1, 2, 3, 4 ou 5 mots à gauche. Il est possible de ne choisir qu'un seul côté, droit ou gauche en donnant pour L et/ou R l'indice 0.



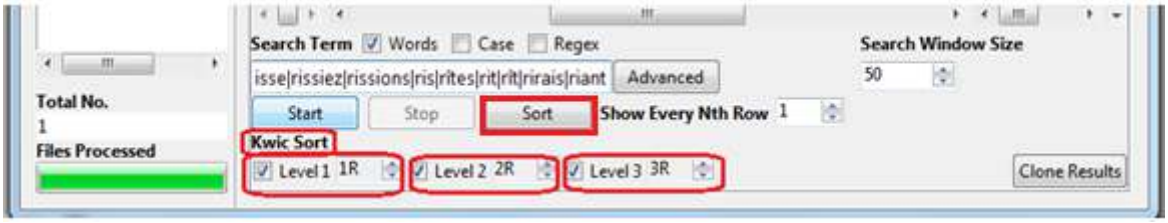
Selon le volume du fichier et la longueur de la liste, le logiciel peut mettre quelques minutes avant d'afficher les résultats.



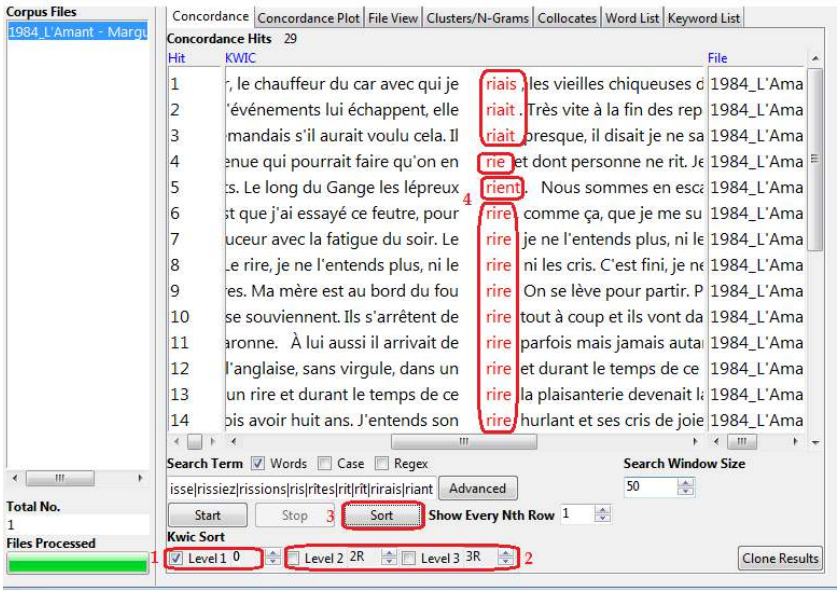
2.1.3. Le filtrage du résultat

AntConc permet de gagner du temps en préparant la concordance en amont avant de passer au traitement avec le tableur Excel. Pour ce faire, en bas de la fenêtre principale, le menu ***Kwic sort*** propose de classer aussi bien les mots-clés que leurs cooccurrents selon 1, 2 ou 3 niveaux (***Level 1, Level 2 et Level 3***).

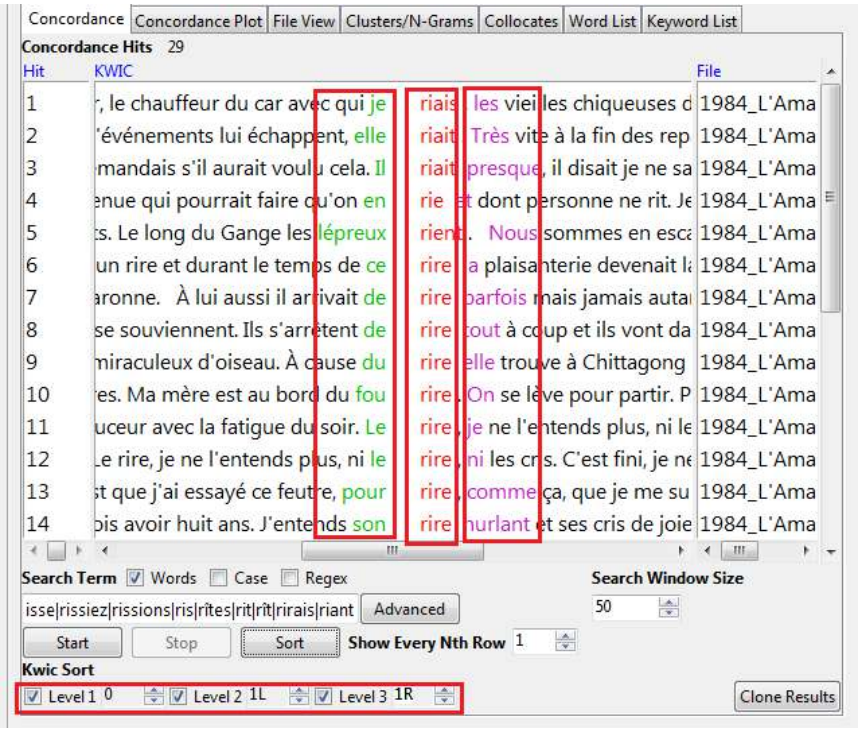
Par défaut, les 3 niveaux sont validés avec les cases cochées. On peut choisir de cocher 1 ou 2 ou les 3 ensemble, au choix et selon l'idée qu'on se fait du résultat escompté. Une fois le choix des options est fait, on doit cliquer sur le bouton du filtrage ***Sort*** et non ***Start***.



Comme pour l'environnement avec ***Advanced*** (L pour gauche et R pour droit), le réglage se fait en choisissant, par exemple, 0 pour ***Level*** 1 (1) puis en décochant ***Level*** 2 et ***Level*** 3 (2), et enfin en validant avec le bouton ***Sort*** (3), on obtient la concordance avec les mots-clés classés par ordre alphabétique (4) : successivement ***riais***, ***rie***, ***rient***, ***rire***, ***ris*** et ***rit***.



On peut aussi utiliser les 3 niveaux en les cochant tous les trois. Dans l'exemple suivant, avec Level 1 en ***0*** (ordre alphabétique), Level 2 en ***L1*** et Level 3 en ***R1***, on obtient un classement qui indique le mot-clé (en rouge) précédé du premier mot employé juste à sa gauche (***L1*** en vert) et le premier mot employé immédiatement à sa droite (***R1*** en violet).



Le filtrage est donc proposé pour observer les environnements des mots-clés recherchés et découvrir, pour un verbe par exemple, les prépositions avec lesquelles il est employé ou ses différentes structures. On découvrira, pour le verbe ***penser***, des emplois intransitifs, d'autres avec la préposition **à**, d'autres encore suivis d'une proposition subordonnée introduite par la conjonction **que**. ***Penser*** peut être suivi

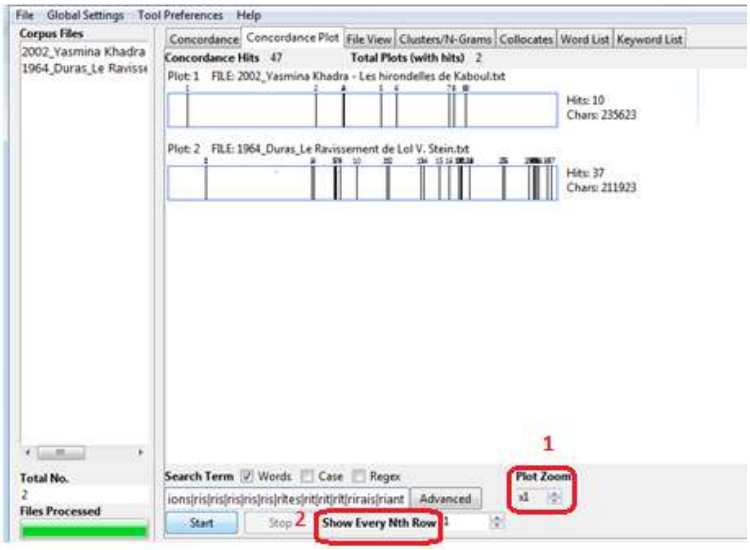
également d'un verbe à l'infinitif et même d'un nom. Il fonctionne aussi comme verbe de parole dans des phrases incises.

Nous ne **pensons** jamais **à tirer** les rois.
« Elle est intégrée », **pensai-je**.
… la plus grande objectivité **pensa** : « Qu'est-ce qui m'empêche… »
qu'est-ce que les Japonais **pensent du** pape
Je n'**en pensais** pas **un mot**.
Les organisateurs avaient dû **penser qu'**elle minerait le moral des détenus
il **pensait à** Plumpy.
Elle **pensait** m'**avoir « rivé mon clou »**.
Comme elle, je ne cesse de **penser à** vous.
Les bien-pensants **pensaient** tout haut **leur bien**

Signalons enfin qu'avec les 6 autres outils, tout clic sur une occurrence d'un résultat ouvre instantanément le volet ***Concordance*** pour afficher la concordance de l'occurrence en question. Pour ***Concordance*** uniquement, si on clique sur une occurrence l'affichage se positionne automatiquement sur la même occurrence dans le volet ***File View*** en affichant une portion très large du texte source.

2.2. Concordance Plot

L'onglet ***Concordance Plot*** permet de visualiser le résultat de la concordance et les statistiques selon une vue longitudinale originale. Chaque texte analysé est représenté par un rectangle bleu qui correspond à la longueur du texte entier et où toutes les occurrences sont tracées sous forme de barres comparables à celles d'un code à barres indiquant approximativement la position des mots recherchés. On peut ainsi découvrir la répartition des emplois d'un mot ou d'une série de mots sur la longueur du texte.
Toutes les barres sont des hyperliens qui permettent d'afficher, sur simple clic, le volet ***Concordance*** et la concordance de l'occurrence en question.



Avec la requête de l'exemple ci-dessus du paradigme du verbe ***rire***, nous découvrons, par exemple, qu'on rit plus dans le roman de Marguerite Duras ***Le ravissement de Lol***

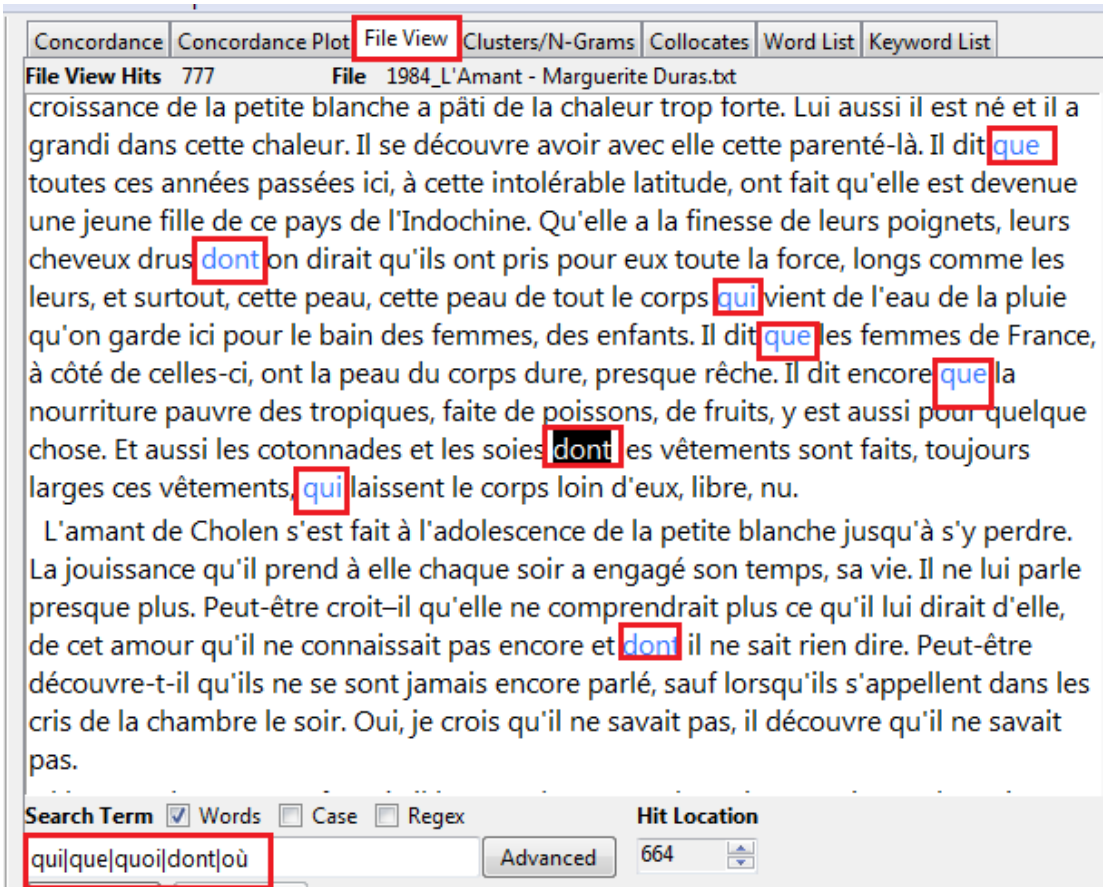
V. Stein (37 occurrences) que dans le roman de Yasmina Khadra ***L'outrage fait à Sarah Ikker*** (10 occurrences), bien que le premier compte un peu moins de caractères que le second (211923/235623). On peut ainsi déterminer approximativement les endroits (début, milieu ou fin) où un mot ou une série de mots apparaissent dans un texte et procéder à des comparaisons si la requête porte sur plusieurs fichiers. Avec ***Plot Zoom*** (1) un agrandissement jusqu'à 10 fois est possible afin de séparer les occurrences qui se confondent.

Enfin l'option ***Show every Nth Raw*** (2), permet d'afficher les résultats à partir chaque nième ligne. Avec le niveau 1, toutes les occurrences apparaissent. Plus le niveau monte, moins de résultats sont affichés pour ne garder que les plus significatifs.

2.3. Vue du Fichier

Avec les résultats obtenus par tous les outils, après un passage par le volet ***Concordance***, un simple clic permet de visualiser un large contexte à partir du texte dans la fenêtre ***File View*** qui s'ouvre instantanément.

Il est possible, ainsi, de retrouver un extrait qui présente, par exemple, plusieurs occurrences des mots recherchés, ici la liste simplifiée des pronoms relatifs saisis dans la zone ***Search Term***.



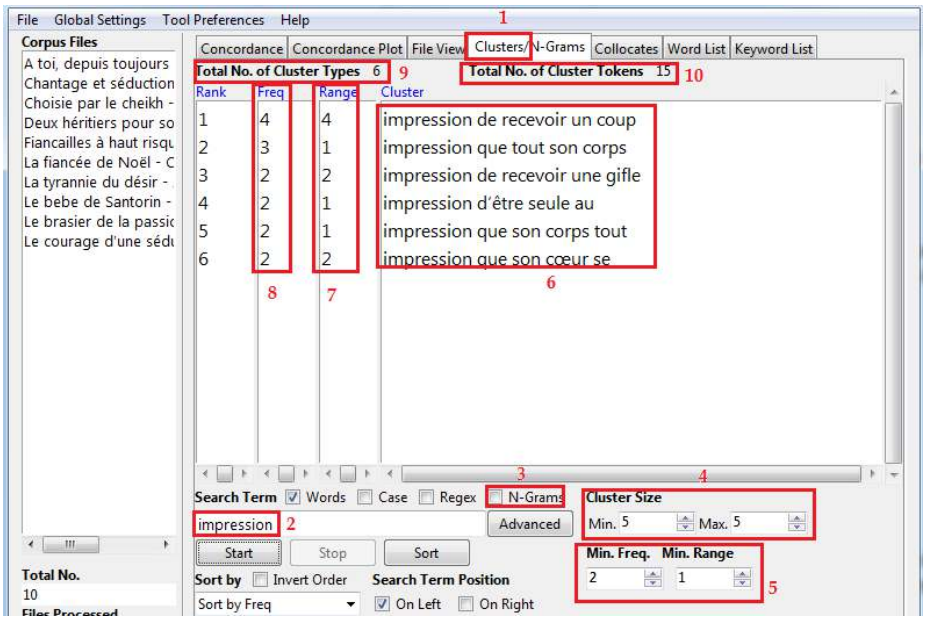
2.4. L'outil Clusters/N-Grams

Cet onglet correspond en fait à deux programmes qui recherchent les segments répétés, mais de deux manières légèrement différentes, le premier avec une requête spécifique et le second sans mot-clé.

2.4.1. Clusters (séquences)

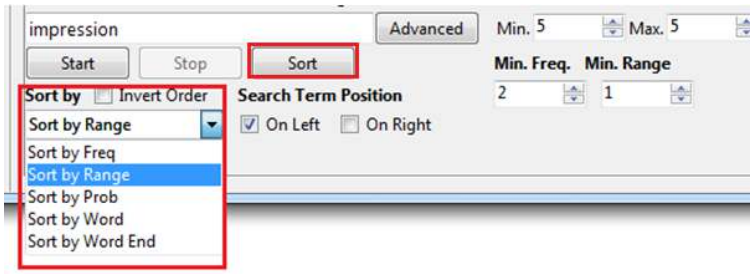
Cet outil produit une liste de groupes de mots contigus qui contiennent la requête tapée dans la zone de saisie. Il s'agit de retrouver les cooccurrents d'un mot ou d'une suite de mots qui pourraient constituer des clichés ou des habitudes de formules chez un ou plusieurs auteurs. Pour pouvoir saisir une requête ***Clusters***, il ne faut pas cocher la case ***N-Grams*** (3). Il faut, par contre, régler, avec l'option ***Clusters Size*** (4), la longueur en mots des suites à rechercher, y compris le mot saisi, avec la possibilité de choisir un nombre minimum et un nombre maximum. On peut également régler un seuil minimal pour les fréquences (***Min.Freq***.) et pour les textes (***Min.Range***) (5). Pour ce genre de recherches, le ***Min.Freq***. doit être réglé sur au moins 2 pour donner des résultats intéressants car le but de la manipulation est la comparaison entre les textes.

La capture suivante montre que le mot-clé recherché est ***impression*** (2). Le réglage ***Clusters Size*** (4) indique que la séquence doit comporter 5 mots (le mot-clé accompagné de 4 cooccurrents à sa droite (4+1=5). Deux options supplémentaires (5) servent à régler le minimum de fréquence (***Min.Freq***.), ici 2, et pour tous les textes, au moins 1 texte (***Min.Range*** 1). Avec un seuil ***Min.Range*** de 2, la recherche n'a produit que les suites répétées au moins deux fois dans au moins deux textes différents. Après validation des options, dans la colonne ***Range*** (7), les nombre des textes dans lesquels l'occurrence de la colonne ***Cluster*** (6) apparaissent. La colonne ***Freq*** (8) donne le nombre des occurrences par ordre décroissant. La fenêtre précise également qu'il a été trouvé 6 différents ***types*** (9) sur un total de 15, celui des ***Tokens*** (10).

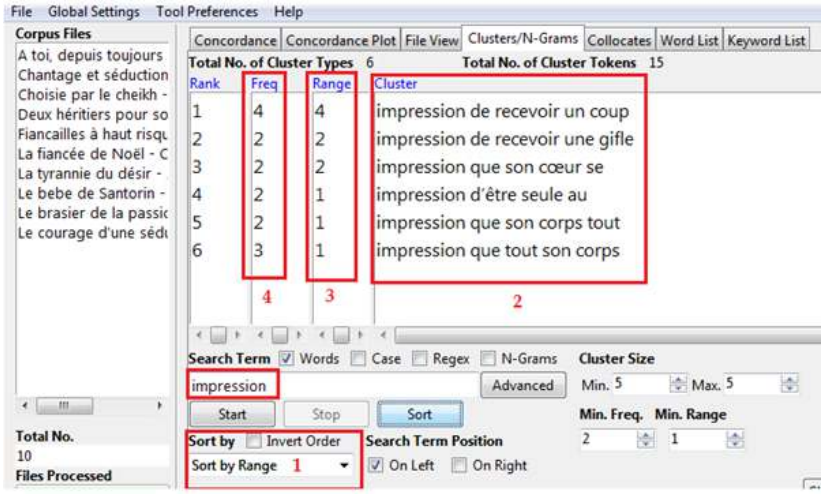


Quand le résultat est affiché, il est déjà filtré par fréquence. D'autres filtrages sont proposés en plus de celui par fréquence. Un filtrage par ***Range*** classera les résultats du

plus grand nombre de textes au plus petit selon le minimum de fréquence voulu. Le filtrage par ***Prob*** (probabilité) classera les résultats par la probabilité du premier mot du groupe précédant les mots restants. Les deux derniers filtrages par ***Word*** et par ***Word End*** classent respectivement l'index par le début ou la fin des mots. Pour valider les options de filtrage, il ne faut pas oublier qu'il faut cliquer sur le bouton ***Sort***.

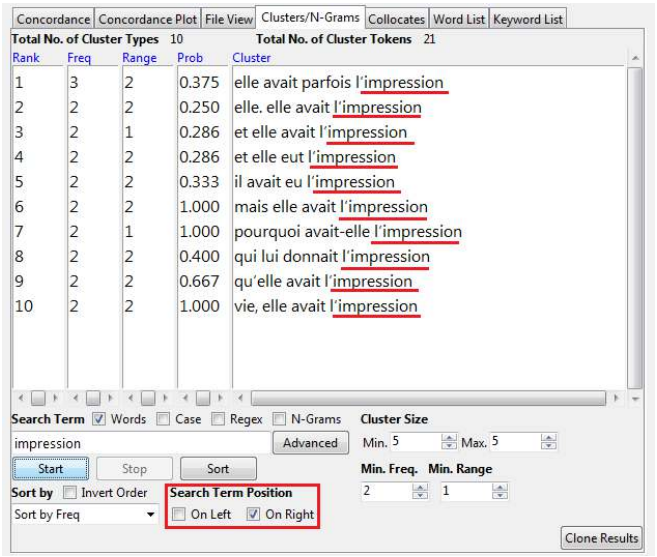


Ce classement peut, par exemple, être refait avec un filtrage par textes ***Sort by > Sort by Range***, par exemple, ce qui donne l'ordre des occurrences (2) par nombre de textes (3) du plus grand au plus petit sans considération de la fréquence (4).



Il faudra tester un à un les filtrages pour trouver des idées et des pistes de recherches intéressantes.

Enfin, il est possible de demander à inverser l'ordre des mots pour placer le mot saisi à la fin du groupe. ***On Right*** signifie qu'on demande que le mot-clé soit le dernier du groupe de mots.



2.4.2. N-Grams (suite de mots)

Un n-gram est « une sous-séquence de n éléments construite à partir d'une séquence donnée » (Wikipédia). En plus clair, il s'agit d'une séquence de taille n, par exemple une suite de 2, 3, 4 lettres ou mots ou plus qui se trouve dans une séquence de taille plus grande que n, un texte par exemple[27].

Contrairement à l'option ***Clusters*** qui cherche les cooccurrents d'un mot-clé spécifique, l'option ***N-Grams*** procède à une recherche à l'aveugle en recherchant les suites de mots, seulement selon la longueur en mots indiquée. Quand on active l'option ***N-Grams*** (2) de la rubrique ***Search Term***, l'outil produit toutes les séquences ayant les longueurs minimale et maximale spécifiées (3) – on ne tape donc pas de mot-clé particulier dans la zone de saisie sous ***Search Term***. Elle permet, dans le cas d'un corpus de plusieurs textes de retrouver des similitudes qui pourraient être considérées comme des clichés spécifiques à un genre particulier de textes[28], à un auteur ou à une époque.

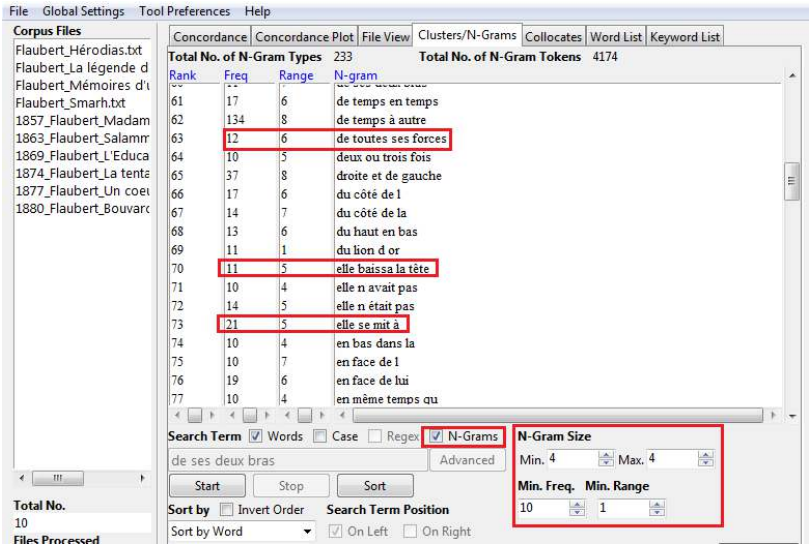
Chez un auteur comme Gustave Flaubert, certaines séquences de 4 mots comme ***de toutes ses forces, elle baissa la tête, elle se mit à***, apparaissent et nous donnent une certaine idée des personnages et des habitudes langagières du romancier.

Voici quelques exemples significatifs de ce que cette recherche produit comme résultat.

---

[27] Les utilisateurs des nouveaux smartphones connaissent l'option *Dictionnaire* qui effectue de la prédiction de texte lorsqu'on commence à taper les mots d'un sms : selon un modèle probabiliste, l'appareil construit des mots par n-gram, c'est-à-dire en proposant la combinaison suivante la plus probable, en fonction des lettres saisies. C'est également cette technique qui est utilisée par un moteur de recherche (Google, par exemple) lorsqu'il suggère de terminer la requête que l'internaute a commencé à taper.

[28] Dominique Legallois présente un didacticiel très clair sur Antconc à l'adresse : https://ecampus.unicaen.fr /pluginfile.php /345933/ mod_resource/content /6/ co/ UOH.html



Dominique Legallois[29] propose de classer ces segments répétés en trois catégories

Types de segments répétés :

- segments complets : la sécurité sociale, salle à manger, il y a longtemps, Monsieur Michel, tout de suite
- segments ouverts : il se mit à, je pense que, il est impossible de
- segments hétérogènes : et je le, bruit de la, de faire de

Seuls les deux premiers sont pratiquement exploitables.

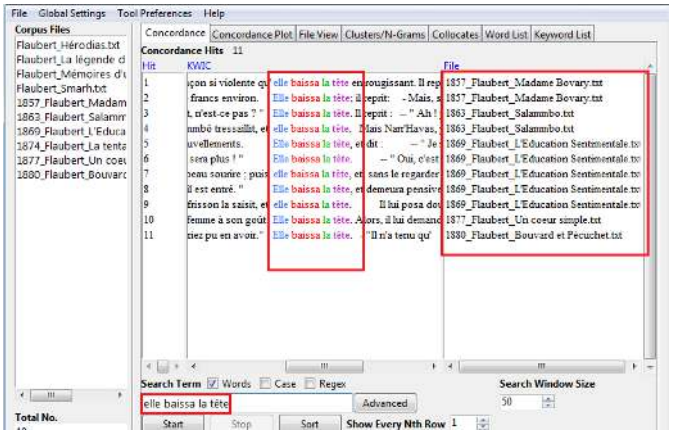
Le tableau suivant donne les segments répétés qui présentent une certaine signification. Bien évidemment le programme génère sans distinction des suites sans sens comme ***en face de l'*** (au rang 75, dans la capture d'écran précédente) et des suites significatives qui sont les seules à avoir un sens exploitable dans les analyses.

Fréquence	Nbr de textes	N-Grams 4
38	5	il se mit à
26	6	puis tout à coup
21	5	elle se mit à
21	5	et se mit à
18	6	sur le bord de
17	4	un quart d heure
16	6	et tout à coup
16	6	le long des murs
15	5	à la lueur des
14	4	pour la première fois
12	6	de toutes ses forces
12	4	quand tout à coup
12	5	sur le seuil de
12	6	à ras du sol
11	7	de ses deux bras
11	5	elle baissa la tête
11	4	et peu à peu
11	4	peu à peu se
10	4	d une voix basse
10	5	deux ou trois fois

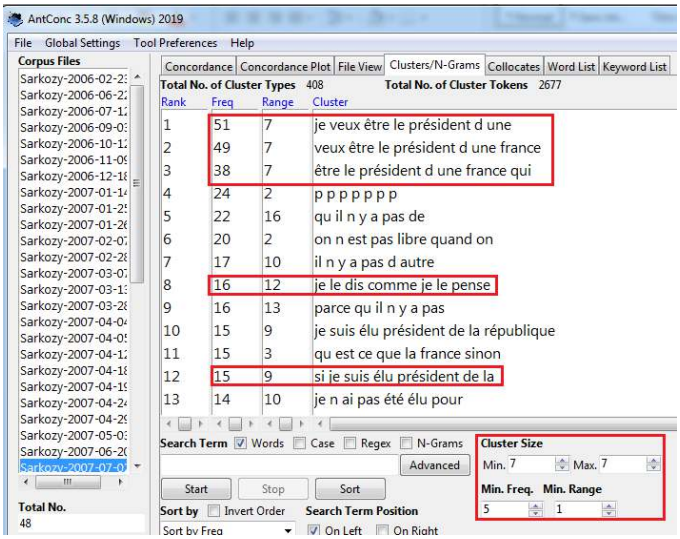
[29] ***Idem***.

10	3	peu à peu il
10	3	sur la grande route

Pour l'exemple ***elle baissa la tête***, on a obtenu 11 occurrences et on découvre ainsi que les personnages féminins de Flaubert dans cinq romans ***Madame Bovary, Salammbo, L'éducation sentimentale, un cœur simple*** et ***Bouvard et Pécuchet*** présentent, comme le signalent certains psychologues du comportement, un manque de confiance en eux-mêmes ou une certaine timidité.



Dans un corpus de discours politiques, par contre, les fréquences des séquences significatives sont très importantes, notamment pendant une campagne électorale. Des formules[30] comme *je le dis comme je le pense, je veux être le président* sont ressassées. Cette récurrence donne son propre style à un homme politique ou à une campagne électorale.



En cliquant sur une occurrence de la liste ***Cluster*** (ici l'exemple 12 avec l'option ***Cluster Size*** 7 : *si je suis élu président de la*), on affiche sa concordance dans l'onglet ***Concordance*** qui est activé automatiquement et où on découvre dans l'environnement de la séquence les promesses du candidat N. Sarkozy pour les Présidentielles françaises de

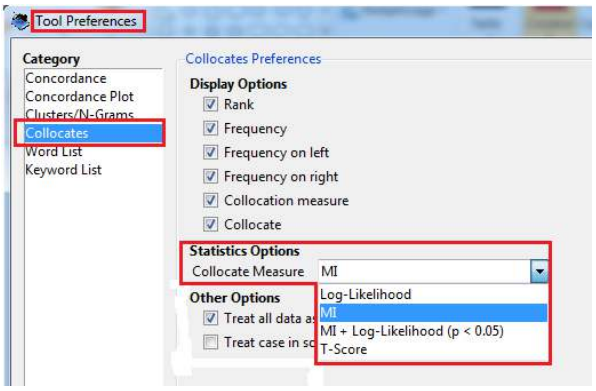
[30] Alice Krieg-Planque, 2009, *La notion de "formule" en analyse du discours : Cadre théorique et méthodologique*, Presses universitaires de Franche-Comté. Coll. Annales littéraires de l'université de Franche-Comté.

2007. Dans la colonne de droite sont indiqués les références des discours qui contiennent les énoncés correspondants.

2.5. Collocates

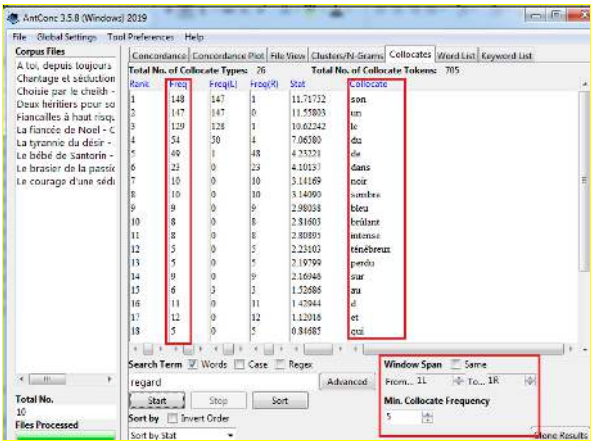
L'outil ***Collocates*** permet de rechercher les collocations d'un terme recherché, autrement dit les groupes de mots comportant un terme pivot donné afin de localiser dans les textes des d'expression idiomatique causée par une cooccurrence systématique, c.-à-d. des structures libres ou figées permises par les règles de combinaisons possibles, grammaticales[31] ou lexicales, dans une langue donnée. C'est le cas notamment des clichés.

L'option ***Statistics Options*** du menu ***Tool Préférences > Collocates >*** permet de régler la recherche en fonction du résultat escompté. Les deux options les plus pertinentes sont ***MI*** et ***T-Score***.

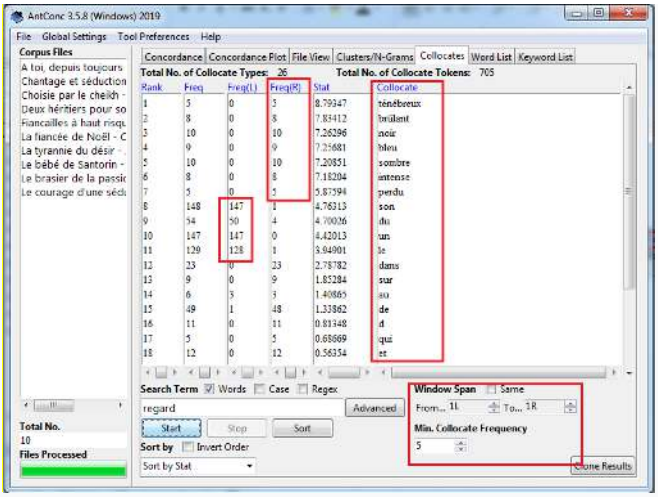


L'option ***T-Score*** classe les fréquences des cooccurrents simplement par ordre décroissant.

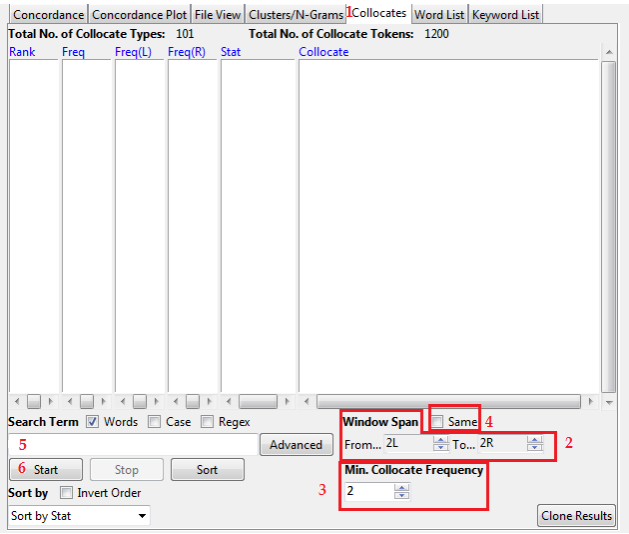
[31] Appelées ***colligations*** dans la linguistique contextualiste britannique.



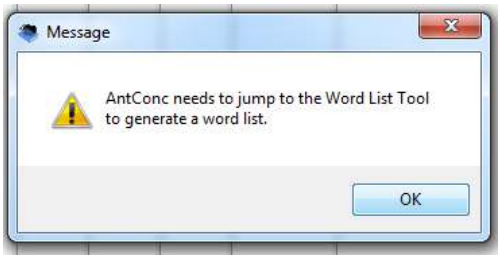
Les statistiques avec ***MI*** (***Information Mutuelle***) néglige totalement la fréquence et privilégie plutôt les mots lexicaux et relègue en fin de liste les mots grammaticaux dont le contenu informationnel est très faible.



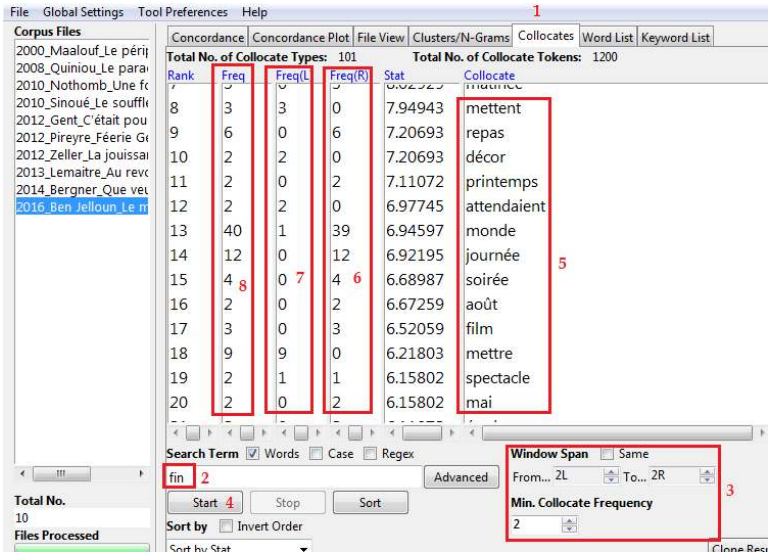
Si, par exemple, le chercheur a seulement besoin d'un même intervalle pour voir quels mots apparaissent directement à droite du terme recherché, il coche la case ***Same*** (identique) pour conserver la même taille d'envergure minimale et maximale.



Au lancement de la recherche, le programme demande au préalable l'indexation[32] des mots du texte pour pouvoir trouver les collocations. Il suffit de l'autoriser à le faire en cliquant sur le bouton OK.



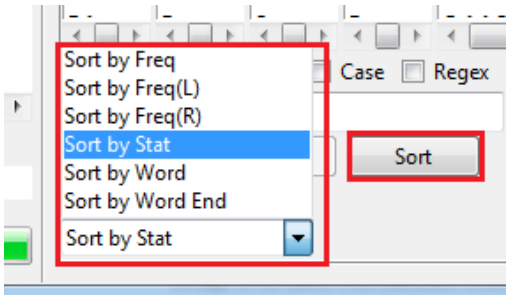
Dans l'exemple suivant, le programme est appelé à rechercher les cooccurrents de la forme ***fin*** (2). Il doit trouver les mots qui se trouvent dans un écart de 2 mots à gauche (***From* 2L**) et de 2 mots à droite (***To* 2R**) de la requête ***Window Span*** (3)[33] et ne retenir que ceux qui sont répétés au moins 2 fois ***Min.Collocate frequency*** (3). Quand on lance la recherche, le résultat offre les données suivantes. Dans la fenêtre (5) sont listés les mots qui peuvent être classés de différentes manières.



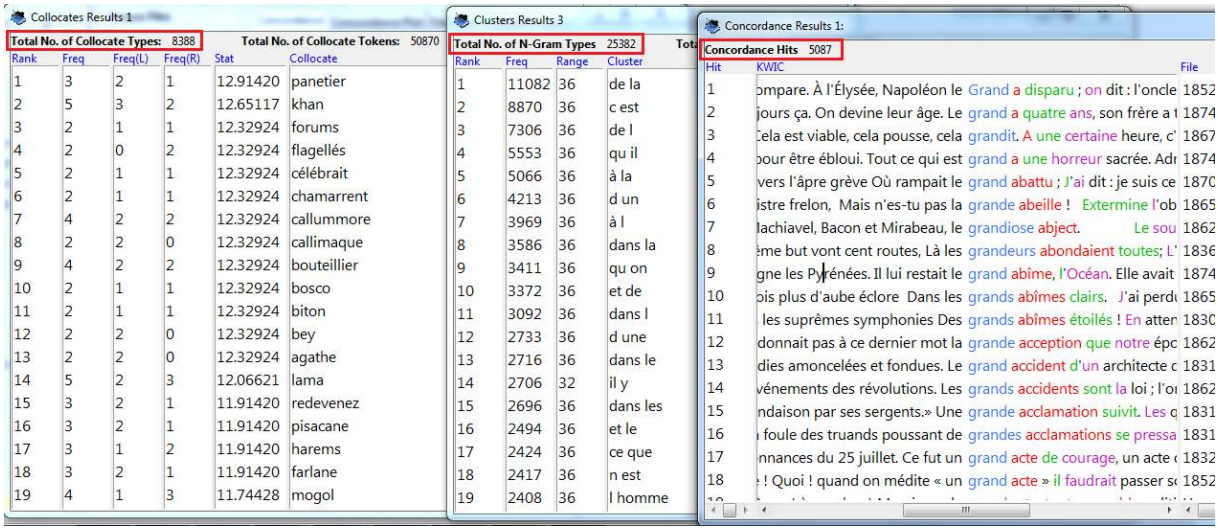
Comme pour tous les résultats, il est en effet possible de filtrer différemment les occurrences : par fréquence générale (***Sort by Freq***)(8), par fréquence des cooccurrents de gauche (***Sort by Freq***(L)) (7), de droite (***Sort by Freq***(R)) (6), par Statistique (***Sort by Stat***), par début (***Sort by Word***) ou fin des mots (***Sort by Word End***)

[32] L'AFNOR (1993) proposée la définition suivante : « l'indexation est le processus destiné à représenter par les éléments d'un langage documentaire ou naturel des données, résultat de l'analyse du contenu d'un document ou d'une question ».

[33] La recherche de cooccurrences ne tient pas compte des limites de phrase ou de paragraphe. Seule compte la distance en mots demandée.

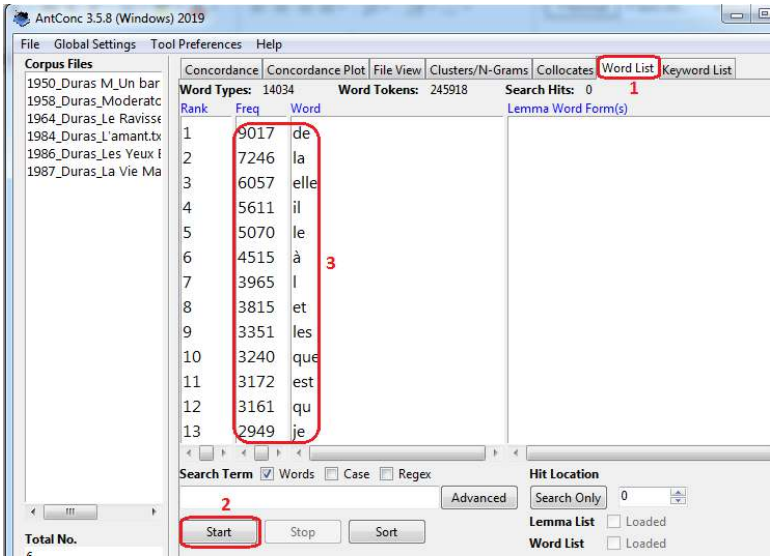


Il est enfin possible de cloner le résultat pour d'éventuelles comparaisons avec d'autres.



2.6. Word List

L'onglet ***Word List*** (pour *index lexical*) permet d'obtenir l'index et les fréquences des mots simples employés selon leur forme graphique dans le(s) texte(s) proposé(s). La liste obtenue classe les mots d'abord du plus au moins fréquent.

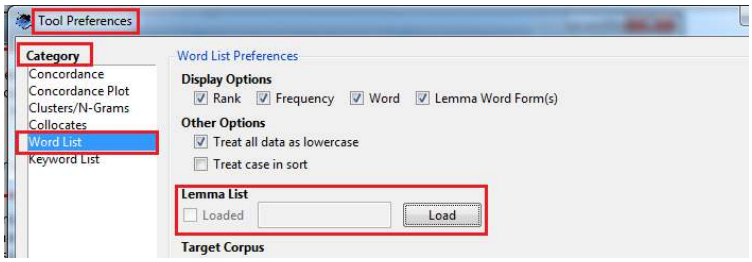


2.6.1. La recherche lemmatisée

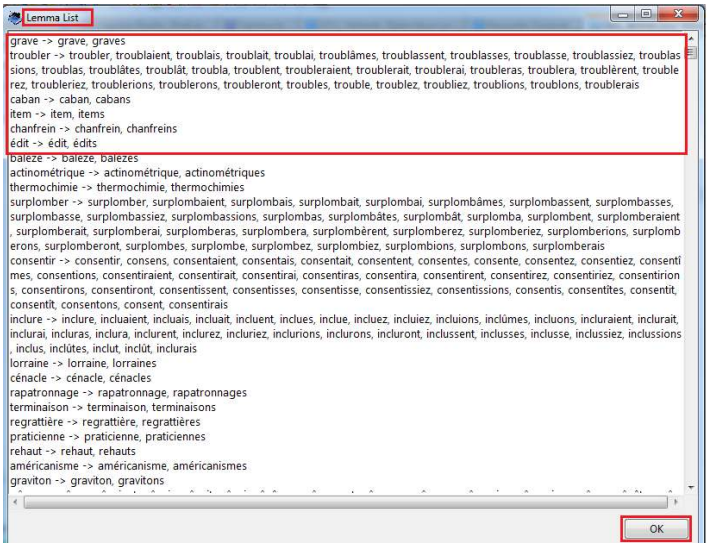
L'outil ***Word List*** offre la possibilité d'affiner davantage la recherche en proposant de regrouper les mots en les classant par lemmes (*lemma* en anglais)[34].

Le site du développeur d'***AntConc*** propose de télécherger des listes de lemmes pour quelques langues dont le français.[35]

Pour cette fonctionnalité, on appelle le fichier téléchargé des lemmes, à l'aide du bouton ***Load*** (charger).



L'affichage du fichier peut prendre quelques secondes. La fenêtre de dialogue doit montrer des lemmes suivis du signe -> et des flexions correspondantes séparées par des virgules.

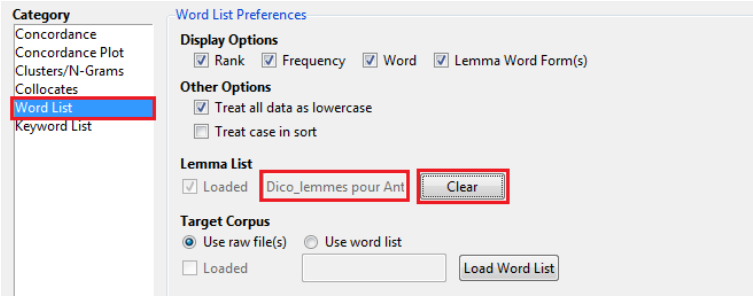


Après validation par ***OK***, on voit apparaître, comme dans la figure suivante, le nom du fichier des lemmes. Le bouton ***Load*** se transforme en ***Clear*** qui permet d'effacer la liste pour d'éventuelles corrections. On lance enfin la recherche en cliquant sur ***Apply***.

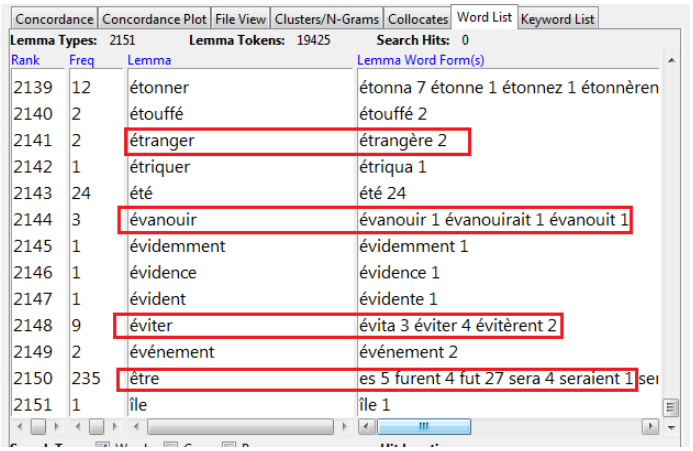
---

[34] Le lemme d'un mot est la graphie qu'un dictionnaire donne pour une entrée donnée. Pour un verbe, ce sera l'infinitif, pour un nom ou un adjectif, la forme du masculin singulier par opposition aux flexions qui sont les différentes formes fléchies d'un même mot. Les formes fléchies correspondent aux formes "conjuguées" d'un verbe ou "variables" d'un mot nom ou d'un adjectif.

[35] https://www.laurenceanthony.net/software/antconc/ . Voir dans le sous-titre *Listes de lemmes* la liste pour le français, créée par Benoît Sagot.



Le résultat s'affiche comme dans la capture suivante. Dans la colonne ***Lemma***, les lemmes sont classés par ordre alphabétique avec, à gauche, la colonne des fréquences et, à droite, les flexions correspondantes suivies chacune du nombre d'occurrences. Ce classement peut être changé avec ***Sort by*** comme expliqué plus haut.



Il faudra cependant faire attention à l'homographie qui n'est pas traitée par un concordancier morphologique comme ***AntConc***. Les résultats sont très souvent truffés de cas parasites (ou bruits) avec des occurrences qui n'ont aucun rapport avec les requêtes de la recherche entamée. Par exemple, la forme ***joue*** correspond aussi bien au verbe ***jouer*** conjugué aux présents de l'indicatif et du subjonctif avec la première et la troisième personne du singulier, qu'au substantif ***joue*** pour la partie d'un visage ou d'un navire, la forme ***rivière*** peut correspondre à un *cours d'eau* ou à un *collier de perles*. Aucun logiciel, à l'heure actuelle, même catégoriseur, n'est capable de faire la différence. Il faudra attendre de passer à ***Excel*** pour éliminer manuellement ces « parasites ».

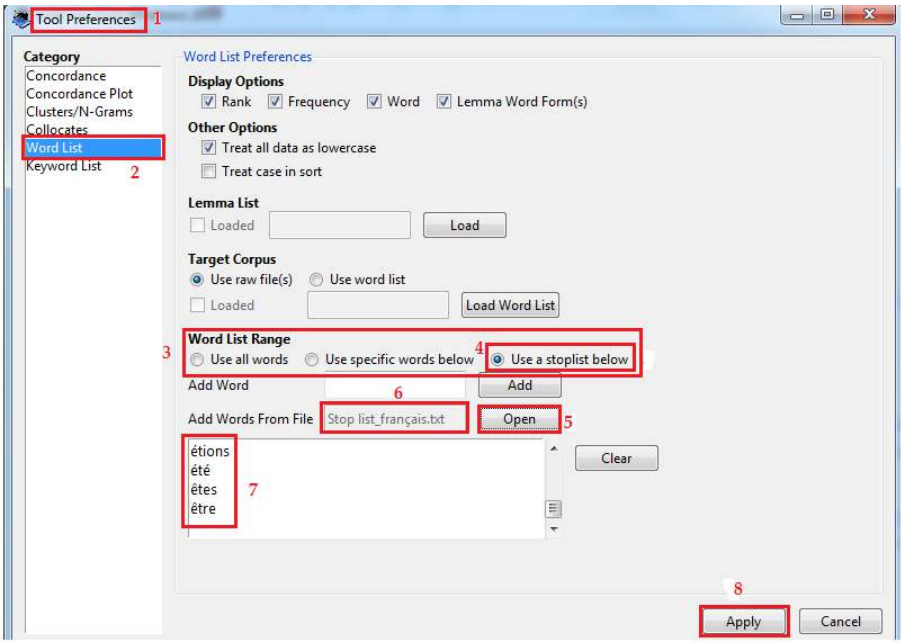
2.6.2. Stop List

La plupart des textes d'une langue présentent une très grande fréquence de mots grammaticaux[36]. En français, la préposition ***de*** vient toujours en tête du classement des mots de n'importe quel type de texte.

[36] Dans le lexique, on distingue les mots lexicaux (ou mots vides dont la liste, spécifique à une langue, est fermée) des mots grammaticaux (ou mots pleins). Dans les années 1930, G.K. Zipf, de l'université de Harvard, a démontré qu'en classant les mots d'un ou de plusieurs textes par fréquence décroissante, on observe que la fréquence d'utilisation d'un mot (le nombre de fois où on l'a trouvé dans les textes) est

Pour ne chercher que dans les mots sémantiquement pleins (nom, verbe, adjectif et adverbe) et empêcher le programme de produire l'index des mots grammaticaux, on utilise une ***Stop List*** (ou Stop Word), c.-à-d. une liste qui comporte les mots grammaticaux, le plus souvent vides de sens. Une simple recherche sur Internet fournit ce genre de listes pour beaucoup de langues[37].

Dans le menu ***Tool Preferences > Category > Word List*** (1-2-3), on choisit, dans la rubrique ***Word List Rang*** (3), l'option ***Use a stoplist below*** (4) et on demande au programme d'ouvrir le fichier qui contient la liste des mots à proscrire de la recherche en cliquant sur le bouton ***Open*** (5). Quand le nom du fichier ***Stop list_français.txt*** (6) et la liste apparaissent (7), on valide (8) et le tour est joué. La fenêtre de dialogue disparaît et on retrouve la fenêtre principale.



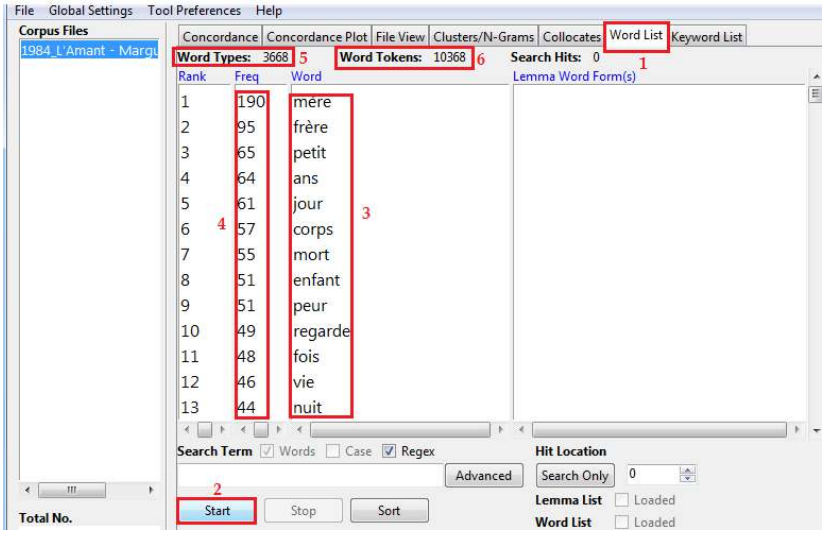
Dans la fenêtre principale, sans saisir de mot dans ***Search Term***, on lance la recherche avec l'onglet ***Word List*** (1). Le résultat s'affiche dans la fenêtre centrale où ne sont retenus que les mots sémantiquement pleins dans la colonne ***Words*** (3).

AntConc propose un filtrage selon la fréquence des mots dans la colonne ***Freq*** (4). Des statistiques s'ajoutent au-dessus de la liste, ***Word types*** (5) (nombre de formes trouvées correspondant à la colonne ***Word***) et ***Word Tokens*** (6) (nombre total des occurrences dans le texte correspondant au total de la colonne ***Freq***).

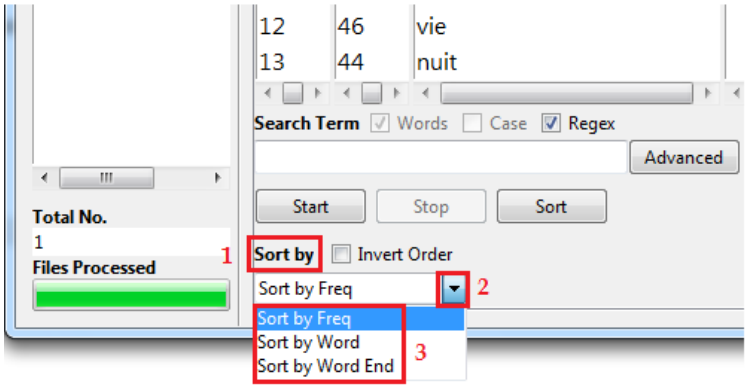
---

inversement proportionnelle à son rang (sa place dans le classement). Cette loi, appelée la loi de Zipf, stipule que la fréquence du second mot le plus fréquent correspond à la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, à son tiers, etc. Pour plus de détail sur cette question, voir la loi de Zipf. G.-Th. Guilbaud, « Zipf et les fréquences », in ***Mots***, n°1, octobre 1980. pp. 97-126.

[37] Pour le français, voir https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.txt. Une fois téléchargé, nommez le fichier, par exemple ***Stop list_français.txt*** et placez-le dans le dossier ***AntConc*** que vous avez créez au départ pour recevoir vos recherches.

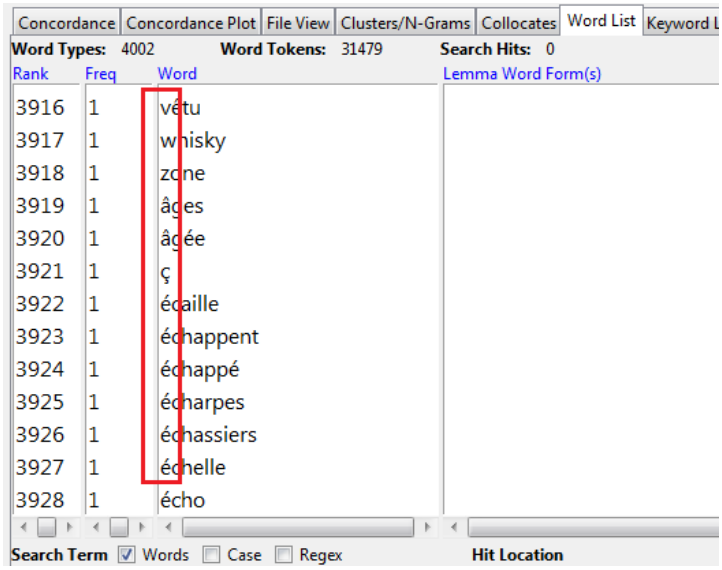


Les mots classés par fréquence peuvent être filtrés avec ***Sort by*** (1) de deux autres façons qu'on découvre en ouvrant (2) la liste déroulante (3). Une fois le choix fait, on clique sur le bouton ***Sort***.

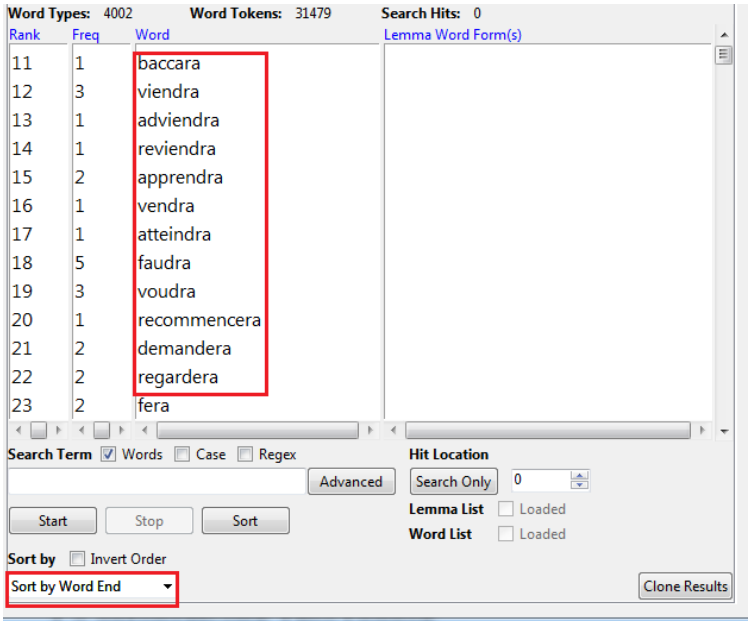


Dans l'ordre pseudo-alphabétique (terme non officiel), les caractères sont classés selon leur code informatique. Les formes commençant par une lettre accentuée sont donc reléguées en fin de liste. Il n'est pas possible de filtrer dans l'ordre alphabétique usuel.[38]

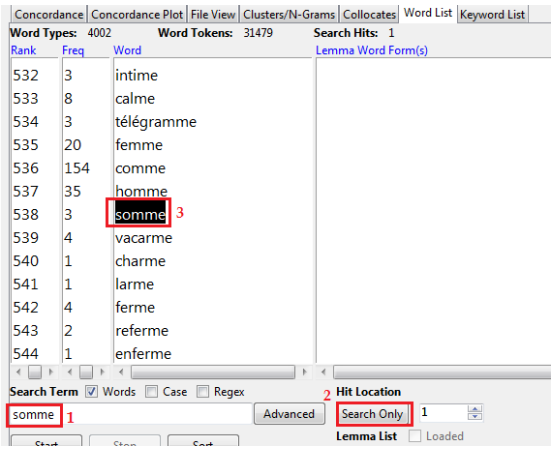
[38] Le tri alphabétique suit l'ordre lexicographique des dictionnaires : a<à<â<ä<A<À<Â<Ä ou e<é<è<ê<ë<E<É<È<Ê<Ë



Le dernier filtrage ***Word End*** propose un classement en considérant les caractères des mots de droite à gauche. Ce procédé permet, entre autres, de retrouver les mots qui riment ensemble.



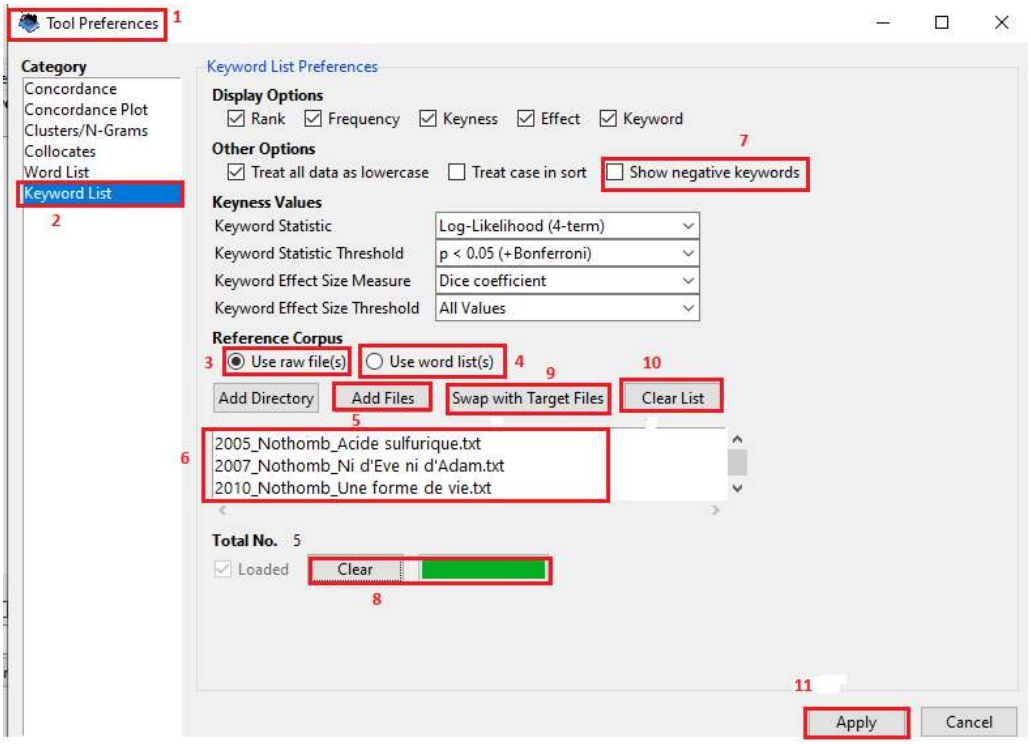
Il est possible de faire des recherches dans l'index. Le mot tapé dans la zone de saisie (1) est appelé avec le bouton ***Search Only*** (2). Il apparaît surligné en noir (3).



2.6.3. Keyword List (liste de mots-clés)

Cet outil sert à faire des comparaisons entre les mots-clés d'un corpus (un ou plusieurs textes d'un seul auteur, par exemple) proposé au programme et ceux d'un autre corpus de référence ou d'une liste préétablie. Le but est de trouver la liste des mots-clés les plus fréquents et donc spécifiques d'un auteur, d'un genre, etc., et ceux qui sont absents ou sous-employés par rapport aux mots du corpus de référence. On peut ainsi comparer deux auteurs entre eux ou le vocabulaire d'un même auteur dans deux textes différents, deux genres littéraires, deux types de discours, etc.

Pour réaliser la comparaison, il faut ouvrir l'onglet ***Tool Preferences*** (1) et choisir la catégorie ***Keyword List*** (2). Un premier réglage avec ***Other options > Show negative keywords*** (7), permet, au choix, de demander au programme de donner aussi bien les mots spécifiques du corpus source que ceux se trouvant dans le corpus ou la liste de référence, mais absents du corpus source. Dans ce cas la case doit être cochée (7).



La méthode de génération de mots-clés (une mesure statistique) sert à calculer la spécificité des mots du fichier cible.

Un mot qui est *positivement* clé apparaît plus souvent qu'on ne l'attendrait par hasard en comparaison avec le corpus de référence.

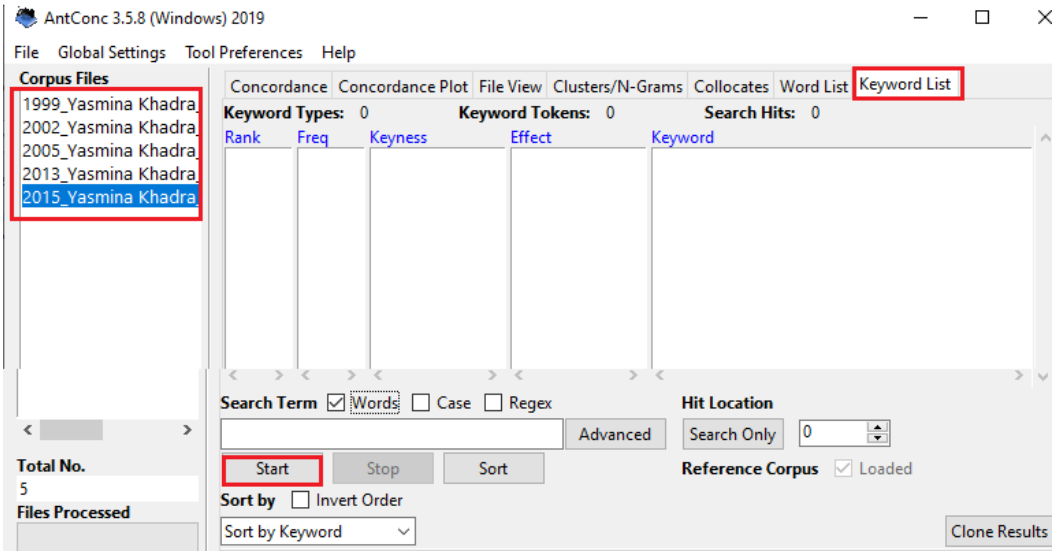
Un mot qui est *négativement* clé apparaît moins souvent que ce à quoi on pourrait s'attendre par hasard en comparaison avec le corpus de référence.

Le paramètre par défaut de ***Log-Liklihood*** (vraisemblance[39]) est donc recommandé.

Pour charger le corpus de référence, on choisit ***Use raw file(s)*** (3) pour le cas de fichier(s) ou ***Use word list(s)*** (4) si on prévoit une ou plusieurs listes qui doivent être déjà préparées dans des fichiers. Avec le bouton ***Add Files*** (5), on renseigne le programme sur l'emplacement des fichiers à utiliser.

La liste des textes apparaît dans la fenêtre (6) ainsi que leur nombre (***Total No.***). On valide avec le bouton ***Load*** (8) qui va prendre en charge le corpus de référence, à partir d'un répertoire créé à l'avance, et se transformer immédiatement en ***Clear*** (8), pour permettre, éventuellement, de rectifier le choix des fichiers ou des listes ***Clear list*** (10). On valide enfin le réglage avec ***Apply*** (11)
Au retour automatique à la fenêtre principale, il reste à choisir le ou les fichiers à comparer avec les fichiers de référence.

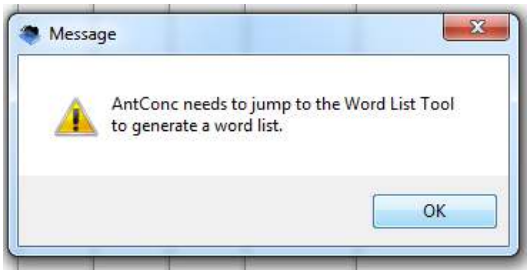
Quand les noms des fichiers choisis sont visibles sous ***Corpus Files***, on passe à l'onglet ***Keyword List*** et on valide avec le bouton ***Start***, sans rien saisir à ***Search Term***.



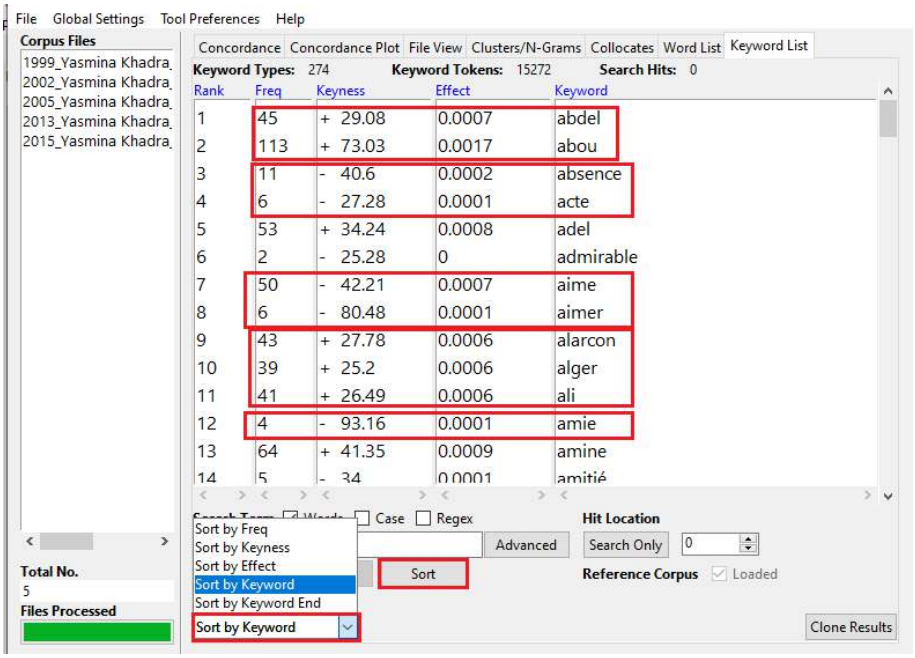
Pour optimiser la recherche, il est préférable de faire fonctionner au préalable l'option ***Stop List*** au menu de ***Word List*** et empêcher la prise en compte des mots grammaticaux[40] et de ne retenir que les mots lexicaux.

[39] La vraisemblance est une fonction qui associe à chaque paramètre la probabilité (ou densité de probabilité) d'observation dans l'échantillon donné.
[40] Voir plus haut **3.7.2.**

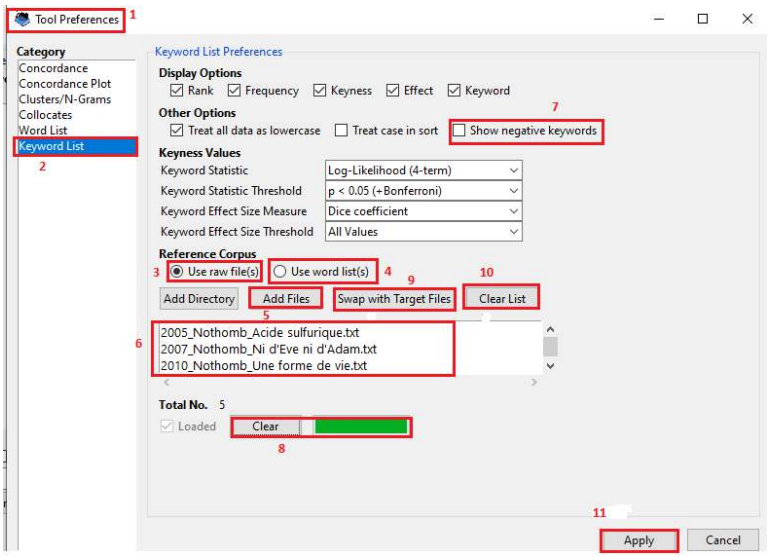
Comme avec ***Word List***, au lancement de la recherche, le programme demande au préalable l'indexation des mots du texte pour pouvoir trouver les collocations. Il suffit de l'autoriser à le faire en cliquant sur le bouton OK.



Dans l'exemple suivant, le programme est appelé à comparer cinq romans de l'écrivain algérien Yasmina Khadra avec le vocabulaire de Amélie Nothomb. Avec le filtrage par ***Keyword*** (mots-clés) et ***Sort***, on remarque que des mots comme les noms propres ***abdel***, ***abou***, ***adel***, ***alarcon***, ***alger***, ***ali*** et le participe passé ***allé*** sont marqués positivement, alors que les mots ***absence***, ***acte***, ***aime*** et ***aimer*** sont comptés comme négatifs, c.-à-d. qu'ils sont sous-employés chez Khadra par rapport à leurs emplois chez Nothomb. Par exemple, ***absence*** est noté -40,6 avec 11 occurrences (pour 32 occ. chez Nothomb) et ***acte*** -27,28 avec 6 occurrences (pour 20 occ. chez Nothomb).



Pour passer d'un corpus à l'autre, autrement dit pour renverser l'ordre des corpus en changeant le sens de la comparaison où le corpus de référence devient le corpus de base et vice-versa, on revient à ***Tool Preferences*** et on clique sur le bouton ***Swap with Target Files*** (9). Immédiatement les deux corpus échangent leur place et l'opération peut être refaite dès le début en vérifiant que l'échange s'est bien opéré (6) et en validant enfin les choix (8 et 11). Avant de lancer ***Keyword List***, il faut refaire la recherche avec ***Word List*** pour charger l'index du nouveau corpus de référence nécessaire à la comparaison. Sinon il n'y aura pas de résultat car l'indexation du corpus de référence n'a pas été réalisée.


Tool Preferences
1
Category
Concordance
Concordance Plot
Clusters/N-Grams
Collocates
Word List
Keyword List
2
Keyword List Preferences
Display Options
Rank
Frequency
Keyness
Effect
Keyword
Other Options
Treat all data as lowercase
Treat case in sort
Show negative keywords
7
Keyness Values
Keyword Statistic
Log-Likelihood (4-term)
Keyword Statistic Threshold
p < 0.05 (+ Bonferroni)
Keyword Effect Size Measure
Dice coefficient
Keyword Effect Size Threshold
All Values
Reference Corpus
3
Use raw file(s)
Use word list(s)
4
9
10
Add Directory
Add Files
Swap with Target Files
Clear List
5
6
2005_Nothomb_Acide sulfurique.txt
2007_Nothomb_Ni d'Eve ni d'Adam.txt
2010_Nothomb_Une forme de vie.txt
Total No. 5
Loaded
Clear
8
11
Apply
Cancel

Chapitre Troisième
La catégorisation

Comme expliqué à propos des tags (§1.2.1.5.), la catégorisation est le procédé qui consiste à ajouter automatiquement des balises, le plus souvent entre chevrons <…> ou des tabulations, pour renseigner les programmes de recherches en linguistique de corpus – les concordanciers notamment - sur les caractéristiques syntaxiques, sémantiques, lexicales, etc. Le but est pouvoir retrouver et regrouper facilement des structures qui combinent ou non des mots-clés.

Il existe deux possibilités de catégorisation (étiquetage) d'un texte, une en ligne et l'autre hors ligne.

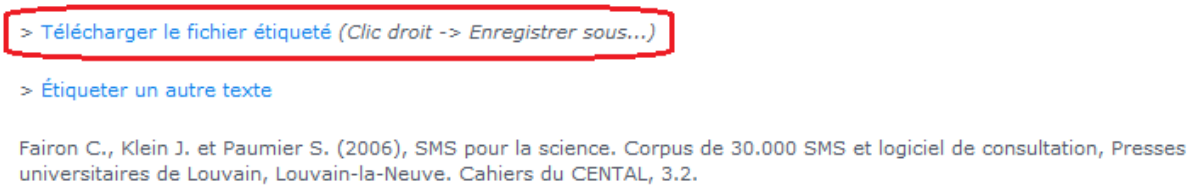
3.1. TreeTagger en ligne

La première possibilité est proposée par le Centre de Traitement Automatique du Langage sur les plates-formes technologiques de L'UCLouvain[41]. Le site permet d'étiqueter, grâce au logiciel de catégorisation ***TreeTagger***, des textes à soumettre au logiciel, uniquement en ligne. La procédure est très simple : il suffit de transférer[42], depuis l'emplacement du fichier sur la machine, un texte en français de type *.txt.

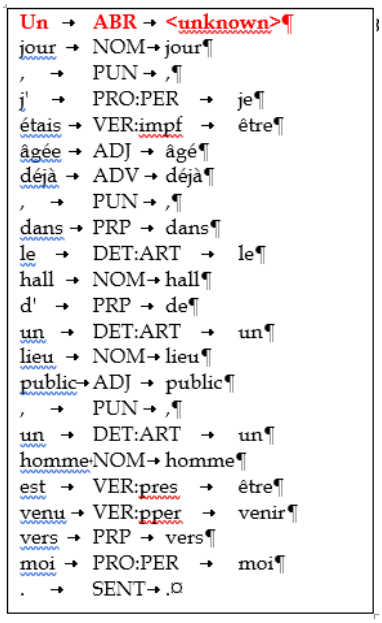
Au préalable, une vérification est d'importance qui concerne l'apostrophe. D'après Wikipédia[43], « en informatique, l'apostrophe droite ('), "***apostrophe-quote***" en anglais, correspond au caractère numéroté 30 par Unicode et ASCII : la numérotation hexadécimale donne donc U+0027. Par contre, l'apostrophe de forme incurvée (’), "***right single quotation mark***" ou "***single comma quotation mark***" en anglais, correspond au caractère Unicode U+2019 ». En français, l'apostrophe ne doit pas être droite, mais incurvée. Cependant les deux types donnent des résultats différents. Sans entrer dans les détails, signalons que l'apostrophe française, l'apostrophe incurvée, n'est pas considérée comme délimiteur et ne sépare donc pas, dans les cas d'élision, les lettres **c** (pronom démonstratif), **l** (pronom personnel ou article défini)**, m, t, s, j** (pronoms personnels), **d** (préposition) et **n** (adverbe de négation) suivies de l'apostrophe du mot à initiale vocalique qui suit. Au lieu d'avoir deux mots pour une suite comme **l'enfant** (***l'*** et ***enfant***), le résultat affichera un seul mot (**l'enfant**). Il faudra donc vérifier la forme de l'apostrophe dans le texte à soumettre et si les apostrophes sont incurvées, procéder au remplacement par des apostrophes droites avant de lancer la requête avec ***TreeTagger***.

[41] http://cental.fltr.ucl.ac.be/treetagger/index.html.
[42] En anglais, upload.
[43] https://fr.wikipedia.org/wiki/Discussion:Apostrophe_(typographie).

Une fois le traitement terminé, il est proposé de télécharger le résultat.



Pour comprendre la procédure, il faut savoir que lors de l'opération de catégorisation, ***TreeTagger*** procède à un retour de chariot (retour à la ligne) après chaque mot du texte. Ensuite il fait suivre les mots par leur catégorie grammaticale et leur lemme, séparés par des taquets de tabulation[44]. Ainsi les mots gardent leur ordre d'apparition dans le texte. Le résultat obtenu se présente comme suit mais sous forme de liste.



[44] Dans Word, les taquets de tabulation (la touche à deux flèches inversées) permettent de réaliser des alignements en colonne de textes sans avoir besoin de recourir aux tableaux, mais juste en positionnant sur une même ligne différentes données. Ils ne sont visibles que si on appuie sur la combinaison des touches Ctrl+Maj+8 ou sur ¶ dans le menu Paragraphe. Ils ne sont pas imprimables.

AntConc, lui, en donnant l'index des mots classés par fréquence ou par ordre alphabétique, défait (malheureusement) la linéarité du texte et empêche ainsi la recherche catégorisée des structures du type DET+ADJ+NOM, etc. L'avantage de la procédure avec ***TreeTagger*** est de pouvoir convertir la liste tabulée pour retrouver aisément le fil du texte à traiter.

Dans l'exemple précédent, excepté l'erreur qui assigne, pour le premier mot de la première phrase, à l'article indéfini **Un**, le symbole ABR pour ***abréviation*** et qu'on peut corriger facilement en « **Un → DET:ART → un** ». D'autres erreurs peuvent apparaître, mais le traitement est bon dans son ensemble.

Pour préparer le fichier au traitement avec ***AntConc***, une étape intermédiaire permet d'encadrer les symboles des catégories par des chevrons. Il suffit de passer par ***Excel*** et d'insérer une colonne à gauche de la colonne des catégories et de la remplir par le chevron ouvrant (<) et une autre à droite à remplir par le chevron fermant (>)[45]. L'utilité des chevrons réside dans le fait que certains programmes informatiques comme ***AntConc*** permettent des recherches en tenant compte ou non des informations contenues dans ces balises. Autrement dit les ajouts lors de la catégorisation gonflent le texte et le rendent illisibles, mais, comme nous l'avons montré plus haut, permettent de lancer des recherches sur le contenu même de ces balises.

Cette opération donne le résultat suivant.

	A	B	C	D	E	F
1	Amélie	<	NAM	>	Amélie	
2	Nothomb	<	NAM	>	<unknown>	
3	Une	<	NAM	>	<unknown>	
4	forme	<	VER:pres	>	former	
5	de	<	PRP	>	de	
6	vie	<	NOM	>	vie	
7	Roman	<	NOM	>	roman	
8	©	<	VER:pper	>	<unknown>	
9	Éditions	<	NOM	>	édition	
10	Albin	<	NAM	>	<unknown>	
11	Michel	<	NAM	>	Michel	
12	,	<	PUN	>	,	
13	2010	<	NUM	>	@card@	
14	9,7822E+12	<	NUM	>	@card@	
15	Chapitre	<	ABR	>	<unknown>	
16	1	<	NUM	>	@card@	
17	Ce	<	ABR	>	<unknown>	
18	matin	<	NOM	>	matin	
19	#NOM?	<	ADV	>	là	
20	,	<	PUN	>	,	

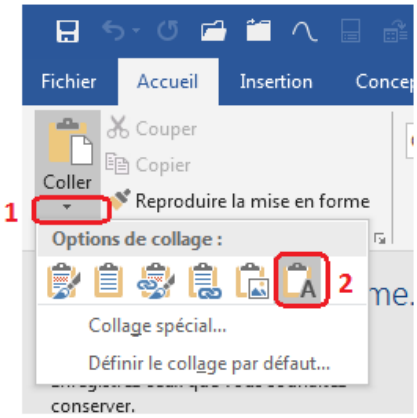
Pour recopier rapidement ce qui est saisi dans la première cellule dans la totalité de la colonne (le tableau contient 31552 lignes), on double-clique sur le petit carré noir[46] qui se trouve à l'angle droit en bas de la cellule B1 avec la souris qui prend alors la forme d'une petite croix noire.

[45] Il est recommandé de garder la colonne des lemmes si jamais on veut plus tard procéder à une recherche lemmatisée.

[46] Appelé « poignée de remplissage ou de recopie ».

	A	B	C	D	E
1	Amélie	<	NAM	>	Amélie
2	Nothomb		NAM	>	<unknown>
3	Une		NAM	>	<unknown>
4	forme		VER:pres	>	former
5	de		PRP	>	de
6	vie		NOM	>	vie
7	Roman		NOM	>	roman

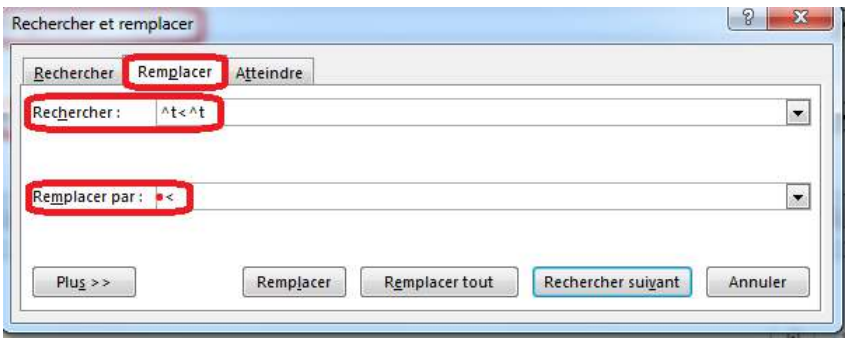
Ensuite on copie les 4 premières colonnes du tableau (A-B-C-D) dans ***Word*** en évitant de cliquer directement sur l'icône **Coller** mais en ouvrant le menu de ***Coller*** (1) et en choisissant de copier seulement le texte (2), surtout si le tableau est très grand. Pour l'exemple d'un texte comme ***L'Amant*** de Marguerite Duras, le fichier téléchargé depuis ***TreeTagger*** compte 1157 lignes.



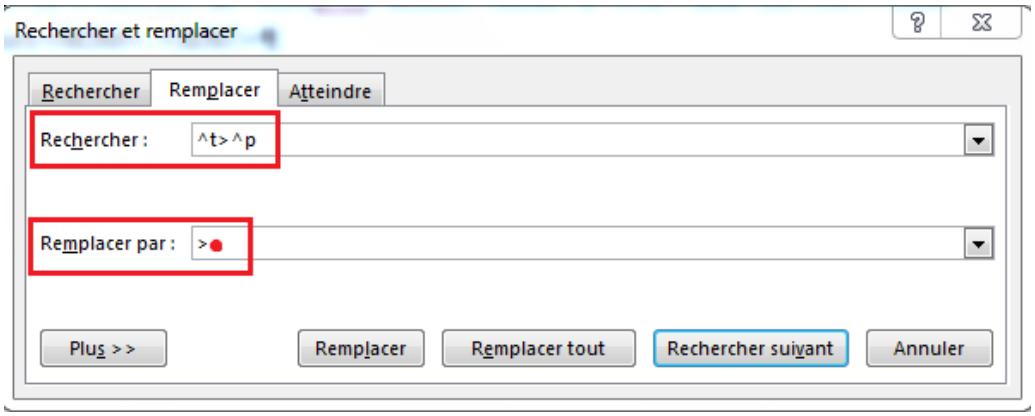
On obtient un résultat comme le suivant :

si → < → KON→> → si¶
tu → < → PRO:PER → > → tu¶
écris→< → VER:pres → > → écrire¶
chaque → < → PRO:IND → > → chaque¶
jour→< → NOM→> → jour¶
de → < → PRP→> → de¶
ta → < → DET:POS → > → ton¶
vie → < → NOM→> → vie¶
comme → < → KON→> → comme¶
une→< → DET:ART → > → un¶
possédée → < → VER:pper → > → posséder¶

On procède à un premier remplacement des tabulations (^t) qui précèdent les chevrons ouvrants ainsi que les chevrons ouvrants eux-mêmes (<) et les tabulations qui suivent par une espace (.) suivie du chevron seulement (le point rouge symbolise ici l'espace).



Un second remplacement remplacera la tabulation (^t), le chevron fermant (>) et le retour de chariot (¶) par une espace afin de transformer la liste en un texte suivi.



Cela donnera, pour une phrase du roman comme

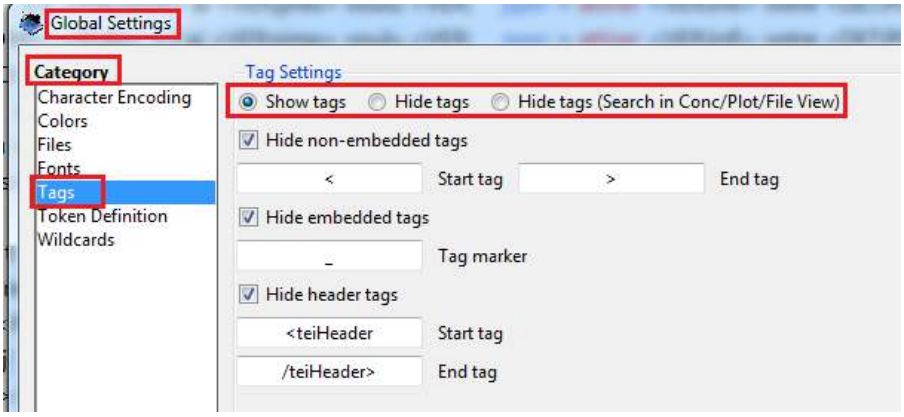
> ***Si tu écris chaque jour de ta vie comme une possédée, c'est parce que tu as besoin d'une issue de secours,***

la suite suivante.

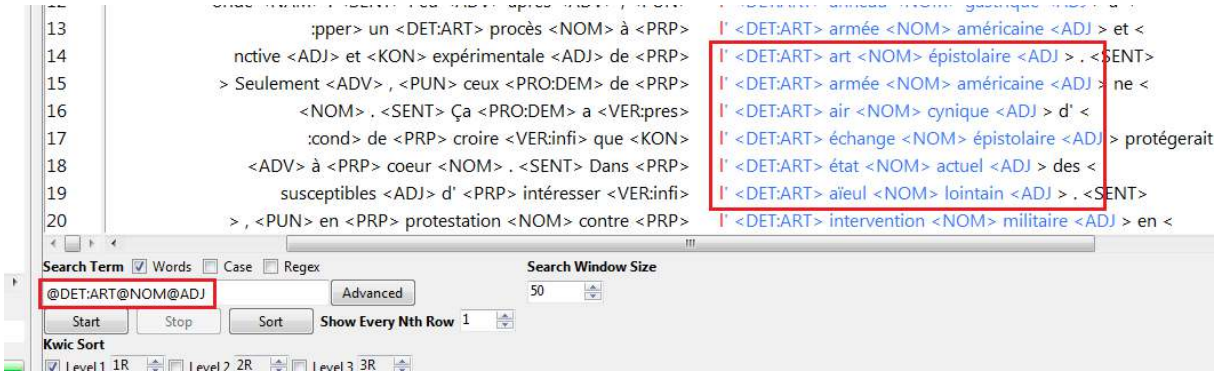
si <KON> tu <PRO:PER> écris <VER:pres> chaque <PRO:IND> jour <NOM> de <PRP> ta <DET:POS> vie <NOM> comme <KON> une <DET:ART> possédée <VER:pper> , <PUN> c' <PRO:DEM> est <VER:pres> parce <KON> que <KON> tu <PRO:PER> as <VER:pres> besoin <NOM> d' <PRP> une <DET:ART> issue <NOM> de <PRP> secours <NOM> . <SENT>

On aura compris que les chevrons seront considérés par ***AntConc*** comme des Tags.

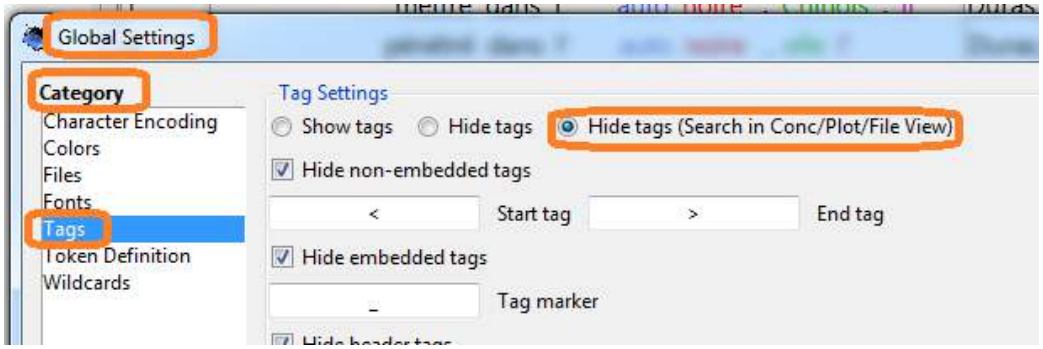
Avant de lancer une requête dans ***AntConc***, on peut passer par le menu ***Global Settings > Category > Tags*** pour découvrir que le programme propose de montrer (***Show tags***) ou de cacher les balises (***Hide tags***).



Avec les tags visibles, une requête comme @DET:ART@NOM@ADJ@ donnera toutes les occurrences de la structure **Déterminant+Nom+adjectif**.



Pour lancer des requêtes sans considération des tags et procéder à des recherches simples comme avec l'outil ***Concordance***, on choisit au préalable l'option ***Hide tags*** qui permet de cacher ce qui est placé entre chevrons (***Global Settings > Category > Tags > Hide tags***). Ainsi le programme fonctionnera comme si les balises n'existaient pas.



On peut simplifier la procédure en n'utilisant pas les chevrons, et en gardant le fichier tagué par **TreeTagger** tel quel et même sous forme de liste, mais on pert la possibilité de cacher les tags qui resteront toujours visibles.

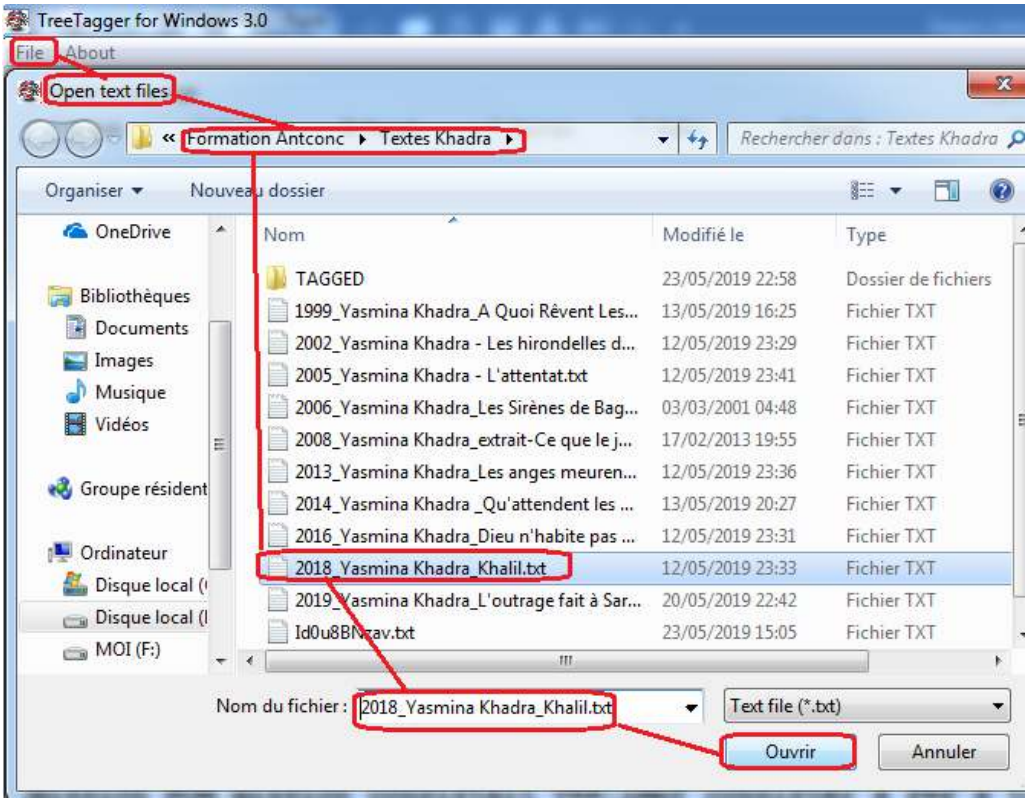
3.2. TreeTagger3_multilingual[47]

La seconde possibilité de catégorisation est réalisable avec le programme homonyme ***TreeTagger*** en ouvrant le fichier **TreeTagger3.exe**. Il est exécutable par double-clic et ne nécessite aucune installation.

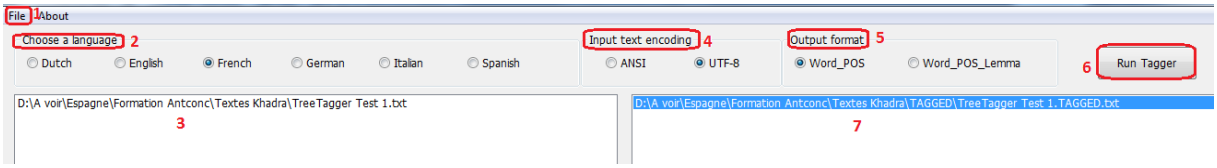
A l'ouverture du programme ***TreeTagger***, on obtient la fenêtre suivante.

[47] A télécharger en suivant la procédure présentée dans le didacticiel de Dominique Legallois à l'adresse https://ecampus.unicaen.fr/pluginfile.php/345933/mod_resource/content/6/co/Module_OATcore__projet_UoH__2.html.

On renseigne avec ***File*** (1) le programme sur le ou les textes à catégoriser.



Une fois le nom du fichier inséré et son nom affiché dans la fenêtre gauche (3), on règle le programme en choisissant la langue du texte (2) : 6 langues différentes sont proposées, ***Dutch*** (néerlandais), ***English*** (anglais), ***French*** (français), ***German*** (allemand), ***Italian*** (italien) et ***Spanish*** (espagnol).



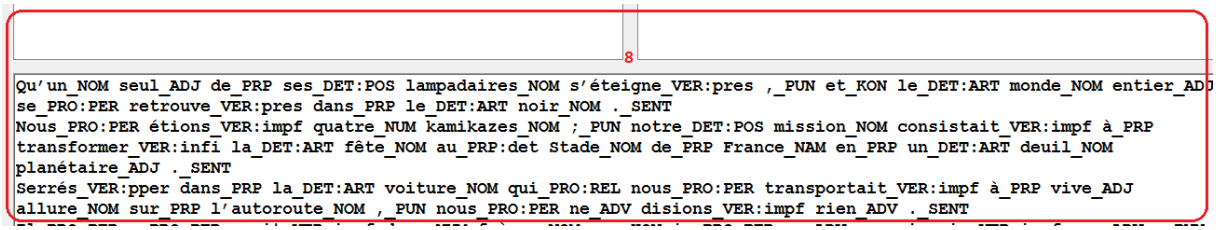
Il faut également paramétrer le codage du texte ANSI ou UTF-8 (4) : deux options qu'il faut tester si on n'est pas sûr de l'encodage du fichier.

On choisit, ensuite, pour le résultat souhaité (5) entre deux options : ***Word_POS***[48] qui, en respectant la linéarité du texte, donnera après chaque mot seulement sa catégorie grammaticale, ou ***Word_POS_Lemma***[49] pour obtenir après chaque mot d'abord sa catégorie grammaticale puis son lemme.

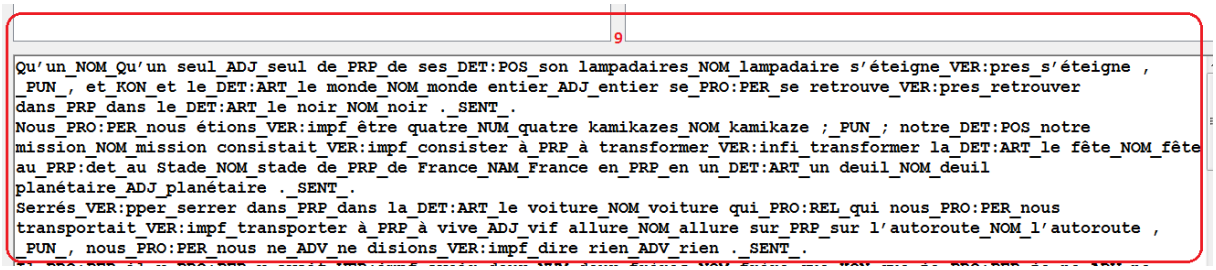
Enfin on valide avec ***Run TreeTagger*** (6). Le nom du texte tagué apparaît (7) dans la fenêtre droite.

[48] POS pour Part of speech (partie du discours).
[49] Lemma est le lemme du mot qu'un dictionnaire donne pour une entrée donnée. Pour un verbe, ce sera l'infinitif, pour un nom ou un adjectif, la forme masculin singulier.

Si le choix a été fait pour ***Word_POS***, quand on double-clique sur le nom du fichier de la fenêtre droite (7), le résultat apparaît dans la fenêtre inférieure (8). Chaque mot est suivi de sa catégorie. Il ne faut pas espérer un résultat impeccable, certaines erreurs peuvent survenir comme pour **Qu'un** considéré comme **NOM**[50]. Mais le résultat est assez satisfaisant dans son ensemble.



Si le choix a été fait pour ***Word_POS_Lemma***, le texte comportera les mots suivis de leur catégorie et de leur lemme. Pour ***ses***, on obtient **ses_DET:POS_son**, pour ***étions***, **étions_VER :impf_être**, etc.



Le ou les fichiers obtenus sont enregistrés automatiquement dans le répertoire du fichier choisi, dans un sous-répertoire nommé TAGGED, avec l'extension **.txt** et consultables comme n'importe quel autre fichier.

Voici le tableau des correspondances des Tags pour le français[51].

Tags	sens
ABR	abbréviation
ADJ	adjectif
ADV	adverbe
DET:ART	article
DET:DEM	Déterminant démonstratif
DET:POS	Déterminant possessif (ma, ta, ...)
INT	interjection
KON	conjonction
NAM	Nom propre
NOM	nom
NUM	numéral
PRO	pronom
PRO:DEM	pronom démonstratif
PRO:IND	pronom indéfini
PRO:PER	pronom personnel

[50] Voir, plus haut, ce qui est dit à propos de l'apostrophe §3.1.
[51] ***French TreeTagger Part-of-Speech Tags,*** Achim Stein, April 2003.

Tags	sens
PRO:POS	pronom possessif (mien, tien, ...)
PRO:REL	pronom relatif
PRP	préposition
PRP:det	préposition plus article (au,du,aux,des)
PUN	ponctuation
PUN:cit	ponctuation de citation
SENT	Point final d'une phrase
SYM	symbole
VER:cond	Verbe au conditionnel
VER:futu	Verbe au futur
VER:impe	Verbe à l'impératif
VER:impf	Verbe à l'imparfait
VER:infi	Verbe à l'infinitif
VER:pper	Verbe au participe passé
VER:ppre	verb au participe présent
VER:pres	Verbe au présent de l'indicatif
VER:simp	Verbe au passé simple
VER:subi	Verbe au subjonctif imparfait
VER:subp	Verbe au subjonctif présent

3.2.1. Recherches dans un fichier tagué

Pour comprendre le principe de ce type de recherche, il faut s'habituer aux suites de mots qui constituent les séquences taguées. L'arobase @ symbolise une suite de lettres entre deux espaces, quelle qu'elle soit, autrement dit un mot.

La phrase suivante

L'accolade que m'avait donnée Lyès était plus longue, mais moins appuyée que d'habitude

donne avec ***TreeTagger*** et ***Word_POS_Lemma***

L'_DET:ART_le accolade_NOM_accolade que_PRO:REL_que m'_PRO:PER_me avait_VER:impf_avoir donnée_VER:pper_donner Lyès_NOM_Lyès était_VER:impf_être plus_ADV_plus longue_ADJ_long ,_PUN_, mais_KON_mais moins_ADV_moins appuyée_VER:pper_appuyer que_KON_que d'_PRP_de habitude_NOM_habitude ._SENT_.

La voici sous forme de tableau avec chaque mot sur une ligne.

Mot	Catégorie	Lemme
L'	DET:ART	le
accolade	NOM	accolade
que	PRO:REL	que
m'	PRO:PER	me
avait	VER:impf	avoir
donnée	VER:pper	donner
Lyès	NOM	Lyès
était	VER:impf	être
plus	ADV	plus

Mot	Catégorie	Lemme
longue	ADJ	long
,	PUN	,
mais	KON	mais
moins	ADV	moins
appuyée	VER:pper	appuyer
que	KON	que
d'	PRP	de
habitude	NOM	habitude
.	SENT	.

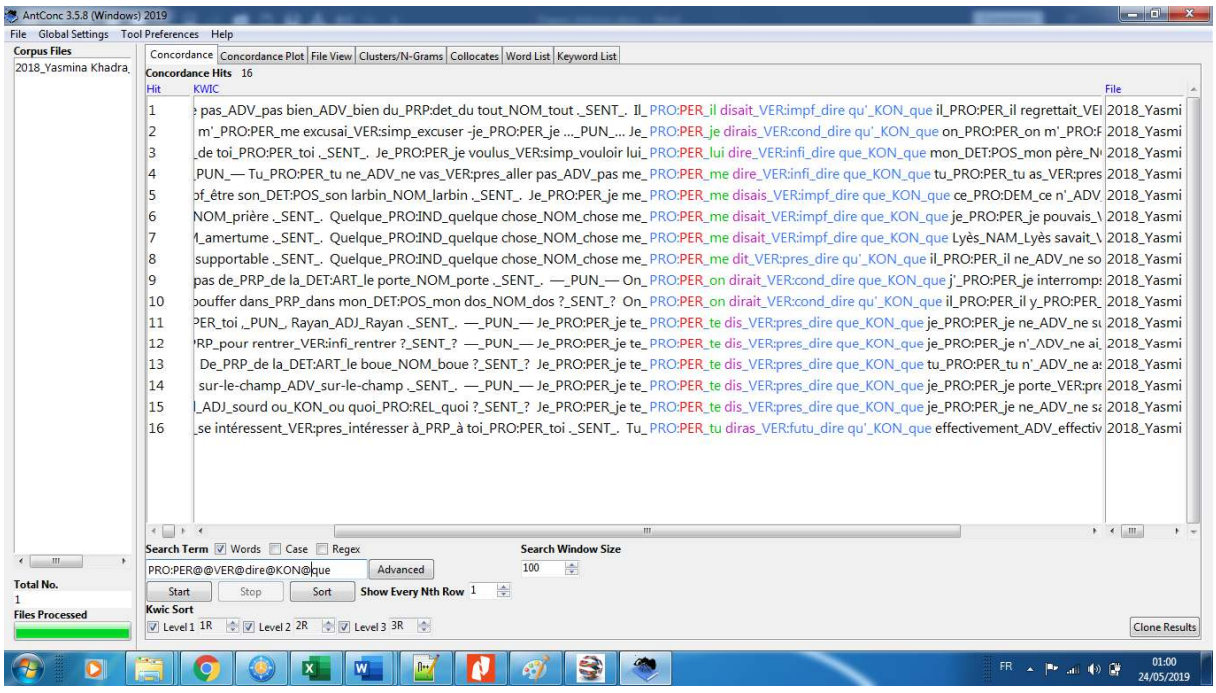
3.2.2. Quelques exemples de requêtes

Voici une série d'exemples de requêtes catégorisées.

- Pour repérer, dans le corpus, l'emploi du verbe ***dire*** précédé d'un pronom personnel et suivi d'une subordonnée conjonctive, on lance la requête suivante :

PRO:PER@@VER@dire@KON@que[52]

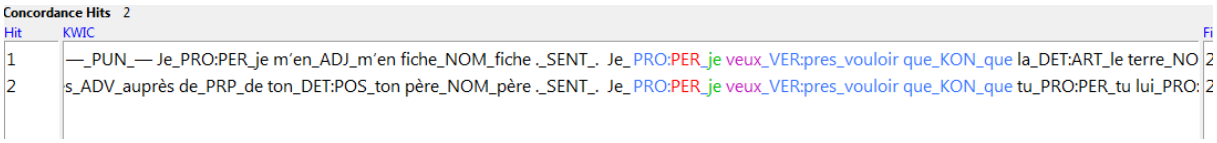
où PRO:PER est la catégorie <pronom personnel>, les deux arobases @@ signifient pour le premier le lemme du pronom (par exemple ***il***) et pour le second la forme du verbe ***dire*** tel qu'elle est dans le texte source (conjugué ou non), VER est la catégorie <verbe> ; l'arobase suivante représente le répertoire de conjugaison du verbe (par exemple ***impf*** pour imparfait). La forme ***dire*** est le lemme du verbe ***dire***. La quatrième arobase est pour la conjonction ***que*** (qui peut avoir la forme ***qu'***). KON est la catégorie <conjonction> et la dernière arobase pour séparer le lemme en ***que***.



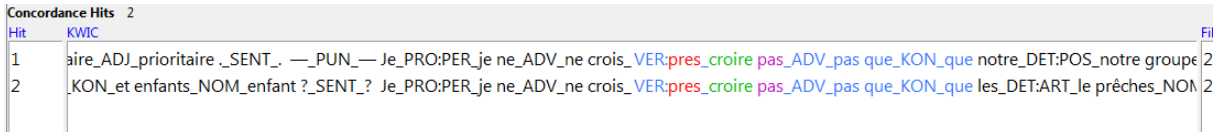
- Pour trouver l'emploi de la construction ***je + vouloir + que***, la requête est **:**

PRO@@PER@@je@@VER@@PRES@@vouloir@@KON@@que

[52] Attention, s'assurer qu'il n'y a pas d'espace avant et après les deux points (:).

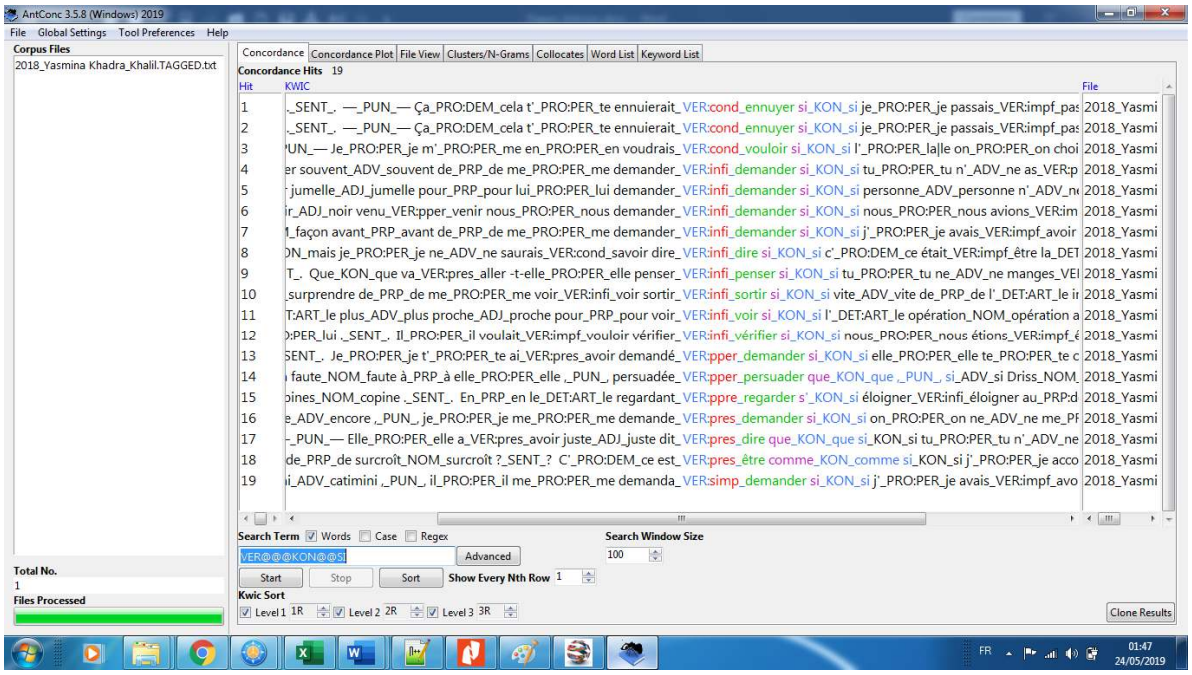


- Pour trouver les occurrences de la construction ***croire que*** à la forme négative, on lance la recherche : VER@@croire@@pas@@KON@@que. Nous avons ajouté l'adverbe de négation « pas ».



- Pour trouver les verbes suivis de la conjonction ***si.***

VER@@@KON@@SI



- Pour trouver une suite comme ***le plus grand,*** avec un déterminant suivi de l'adverbe ***plus*** et d'un adjectif

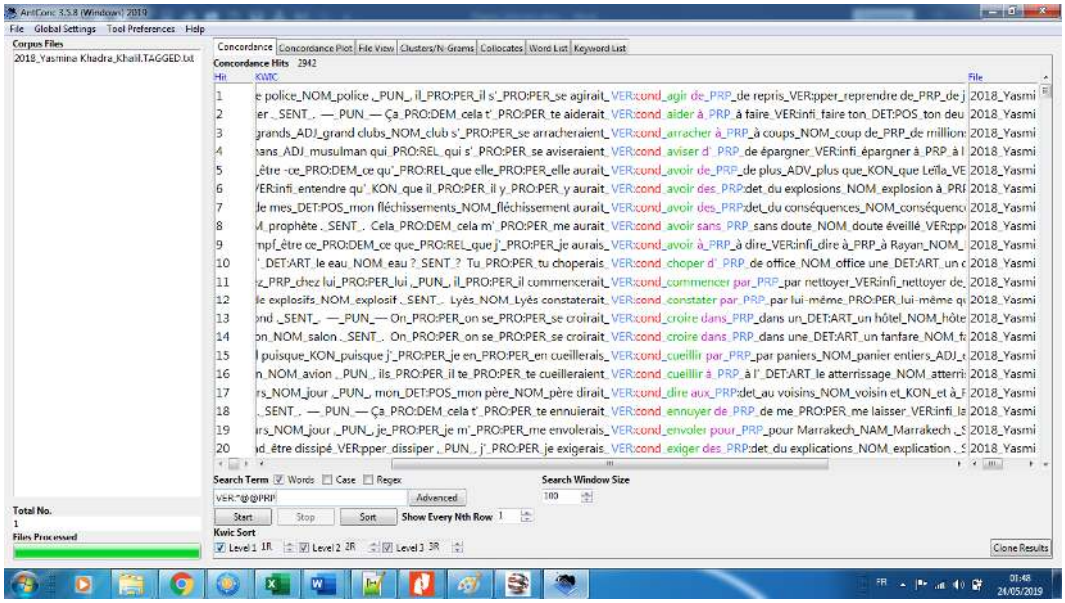
DET:ART@@@plus@ADJ@

- Pour trouver la structure : Déterminant + N + une préposition + N, comme dans ***la boite à pharmacie***, on demande

DET:ART@@NOM@@PRP@@NOM@

- Pour trouver les verbes suivis d'une préposition, on demande

VER:*@@PRP



- Pour trouver tous les verbes d'un texte, on demande

VER :*@

- Pour trouver tous les verbes d'un texte suivis de la conjonction QUE, on demande

VER :@@qu*KON que

- Pour trouver tous les verbes à l'impératif, on demande

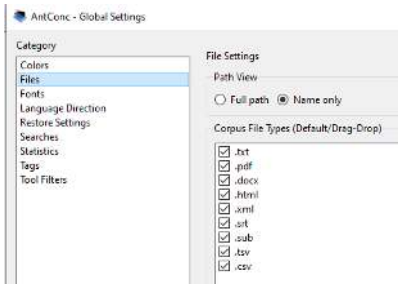
VER :impe

Chapitre Quatrième
La nouvelle version 4.2.0

Dans sa nouvelle version 4.2.0, téléchargeable toujours gratuitement[53], ***AntConc*** exige désormais l'installation du logiciel. Plusieurs changements et modifications ont été apportés.

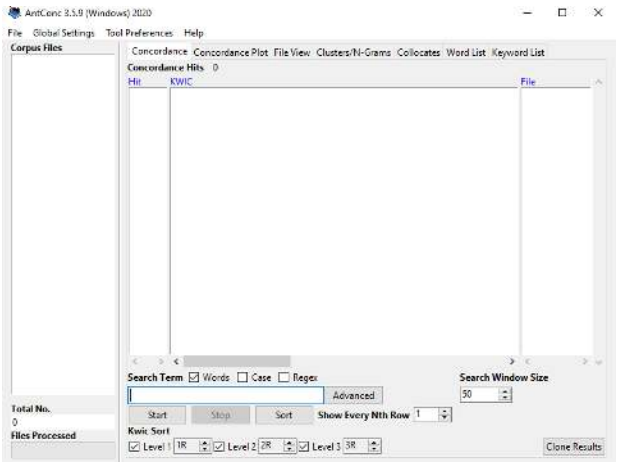
Le logiciel est désormais puissant et permet la création de bases de données. Il est développé en *Python*[54] et *Qt*[55] en utilisant le compilateur *PyInstaller*[56] afin de générer des exécutables pour les différents systèmes d'exploitation. Il utilise *SQLite*[57] comme base de données sous-jacente.

Dans cette nouvelle version, plusieurs types de fichiers sont acceptés, dont ***Word*** de ***Microsoft Office*** :



4.1. Les nouveautés de la fenêtre principale

L'interface est légèrement différente par rapport aux anciennes versions. La plupart des outils sont présents avec, pour certains, un changement de nom ou d'emplacement.



AntConc version 3.5.9w

[53] https://www.laurenceanthony.net/software/antconc/

[54] *Python* est un langage de programmation interprété, multiparadigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.

[55] *Qt* est un framework d'interface graphique et d'application multi-plateforme.

[56] *PyInstaller* est un outil pour regrouper les fichiers Python et toutes ses dépendances dans un seul exécutable ou répertoire.

[57] *SQLite* est un gestionnaire de base de données *SQL* utilisable localement ou sur un site Web. N'importe quel logiciel peut l'utiliser pour stocker des données plutôt que dans des fichiers texte, et donc effectuer des requêtes d'accès aux données.

Dans la nouvelle version, la fenêtre principale présente plusieurs modifications et ajouts.

Search Windows Size, qui permettait de choisir le nombre de mots à afficher de part et d'autre du terme recherché, devient ***Context Size*** (Taille du contexte).

Le bouton ***Sort*** disparaît et est confondu avec ***Start***.

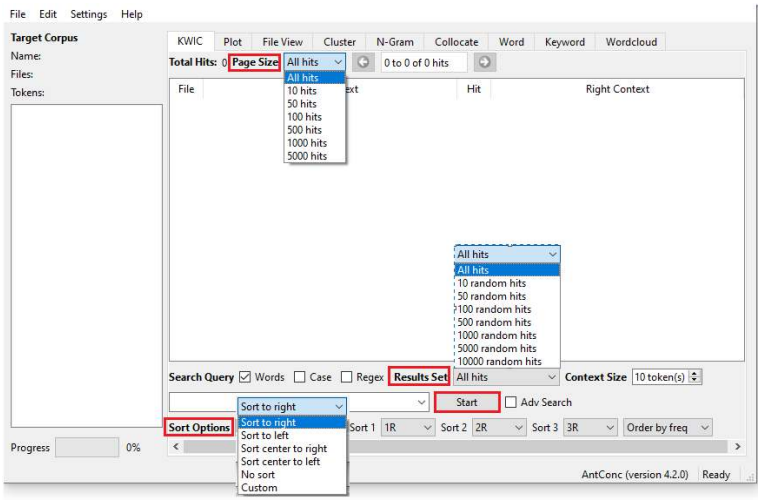
Avec l'option ***Page Size***, il est désormais possible de choisir d'afficher tous les résultats d'une requête (***All hits***) ou uniquement les n-premiers (10, 50, 100, 1000 ou 5000) si le nombre des concordances est très grand.

Par contre avec la nouvelle option, ***Result Set***, en remplacement de ***Show Every Nth Row*** qui permet d'afficher un nombre restreint parmi les concordances trouvées en divisant la totalité des résultats par le diviseur choisi, le programme, désormais, génère aléatoirement un échantillon parmi toutes les concordances selon le nombre choisi par le chercheur (***10, 50, 100, etc. random hits***). Le résultat change à chaque redémarage de la requête.

Les Options de filtrage (***Sort Options***) permettent de programmer, avant de lancer une requête, l'organisation et/ou la réorganisation des concordances si le résultat n'est pas satisfaisant. Chacune des options proposées se combine avec l'ordre souhaité : par fréquence ***Order by freq***, ou par ordre alphabétique ***Order by value***.

Avec ***Sort to left***, par exemple, le programme classe les concordances selon l'ordre choisi grâce aux quatre possiblités des listes déroulantes : ***Sort 1***, ***Sort 2***, ***Sort 3*** et ***Order by…***

Le chercheur a donc la possibilité de générer ou de réorganiser rapidement les résultats en fonction de ce qu'il cherche à trouver dans le corpus avec les différentes combinaisons offertes par ces options qu'il faut tester une à une pour se familiariser avec les différentes possibiltés.

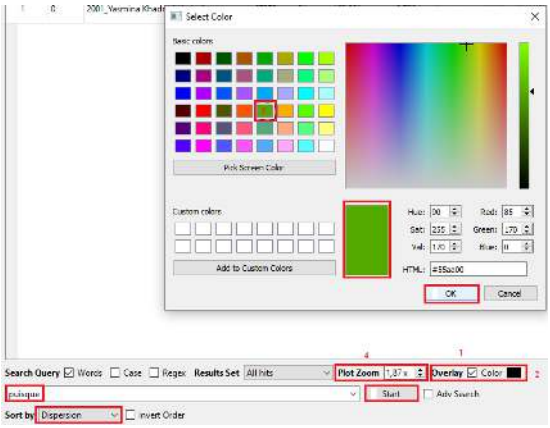


AntConc nouvelle version 4.2.0

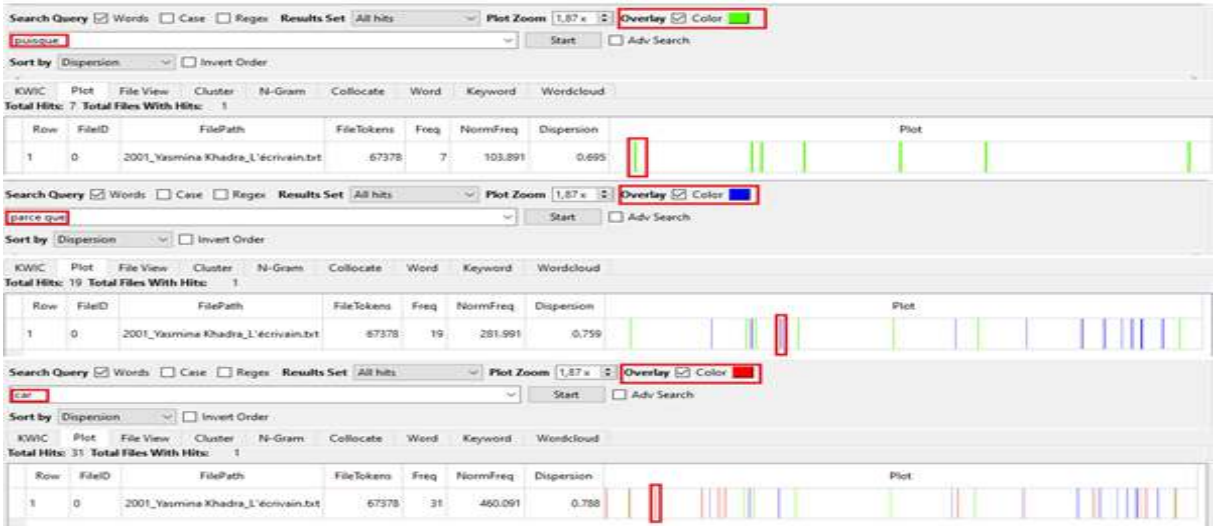
Les deux anciens menus ***Global Settings*** et ***Tool Preferences*** sont réunis dans un seul (***Settings***) avec les deux mêmes menus mais en sous-menus, et un nouveau, ***Edit***, est ajouté pour la sélection, la copie et l'enregistrement rapide des résultats, avec ***Select All*** et ***Copy***.

AntConc propose neuf outils accessibles, soit en cliquant sur l'onglet correspondant dans la fenêtre d'outils, soit en utilisant CTRL+TAB pour passer d'un outil à l'autre, ou encore en utilisant la combinaison de touches CTRL + numéro d'outil (par exemple, CTRL +1 pour ***KWIC***, CTRL +2 pour ***Plot***, etc.) pour sélectionner un outil spécifique.

Pour le menu ***Plot***[58], La nouveauté permet de superposer plusieurs résultats en cochant, au préalable, la case ***Overlay*** (1) et en attribuant à chaque requête une couleur spécifique en cliquant sur la case ***Color*** (2-3). Cela permet, dans les limites du possible, de voir comment différentes requêtes de recherche sont liées et/ou se chevauchent. ***Plot Zoom*** (4) permet de contrôler la taille du tracé et le degré de détail à afficher.



Voici un simple exemple qui montre la distribution de trois expressions de la cause, puisque (en vert), parce que (en bleu) et car (en rouge).



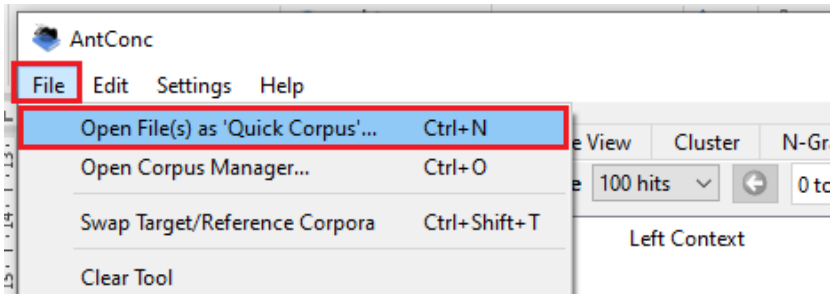
[58] Voir § 2.2.

Un nouvel outil, ***Wordcloud***, est ajouté, mais juste dans un but d'esthétique. Il permet une visualisation des résultats générés par les outils ***KWIC***, ***Fichier***, ***Cluster***, *N*-***Gram***, ***Collocation***, ***Word***, sous la forme d'un nuage de mots. Son utilité est très limitée.

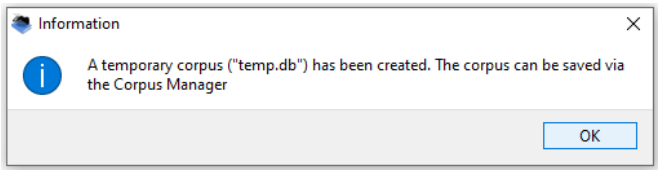
4.2. Le menu Fichier

4.2.1. Ouverture rapide

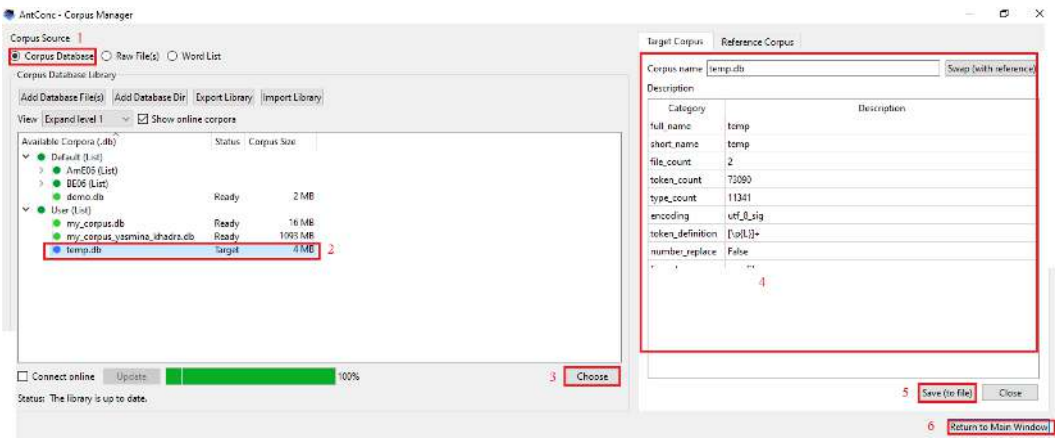
Pour ouvrir rapidement un ou plusieurs fichiers, le programme propose ***Open File(s) as 'Quick Corpus'*** pour une seule et rapide session.



Une fois le corpus proposé – composé d'un seul ou de plusieurs fichiers – est accepté, le logiciel signale qu'un corpus temporaire, nommé ***temp.db***, a été créé et qu'il est possible de l'enregistrer sous un autre nom pour une utilisation ultérieure, si l'utilisateur le décide.

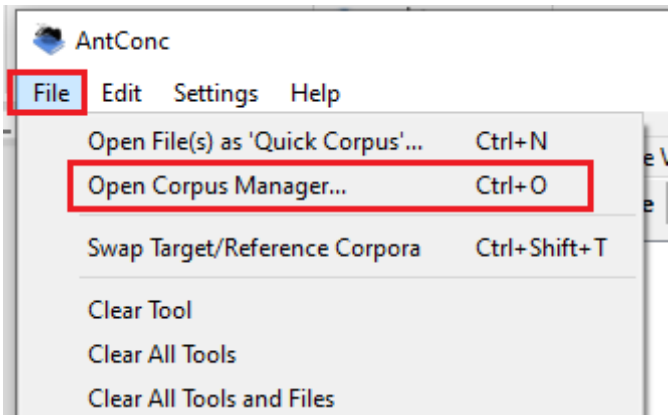


Le fichier reste à la disposition du chercheur et est visible dans le menu ***File > Open Corpus Manager >Corpus Database*** (1) qui affiche les corpus déjà créés. Pour vérifier la disponibilité du corpus ***temp.db*** et l'activer, le bouton ***Choose*** (3) donne différentes informations sur le corpus (4). Il est possible de lui donner un nom explicite et de l'enregistrer pour de nouvelles recherches avec le bouton ***Save (to file)*** en bas à droite de la fenêtre. Une fois l'enregistrement effectué, avec le bouton ***Return to Main Window***, on revient à la fenêtre principale pour procéder aux recherches.

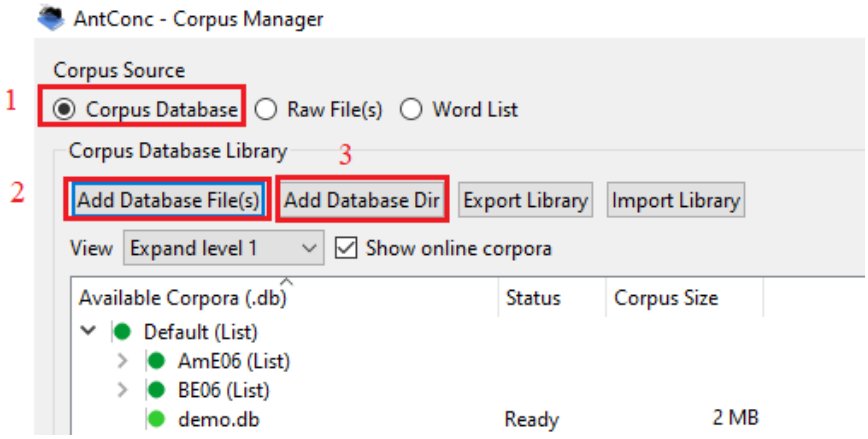


4.2.2. Ouverture du gestionnaire de corpus

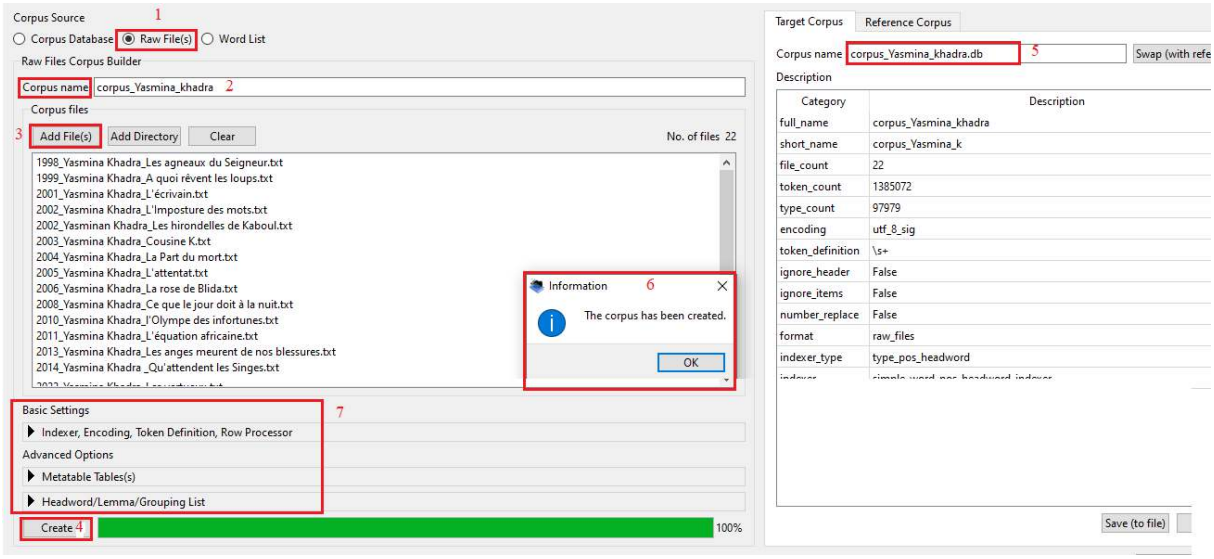
Si l'utilisateur compte réutiliser un corpus dans plusieurs sessions du programme, il est conseillé d'ouvrir le ou les fichiers avec l'option ***Open Corpus Manager*** du menu ***File***. Pour bien gérer et en profondeur sa base de données, le programme offre au chercheur plusieurs possibiltés qu'on va passer en revue.



L'option ***Corpus Database*** (1) étant sélectionné par défaut, le programme affiche les bases de données déjà existantes dans sa mémoire et permet d'y ajouter des fichiers (2) ou même tout un sous-dossier (3).

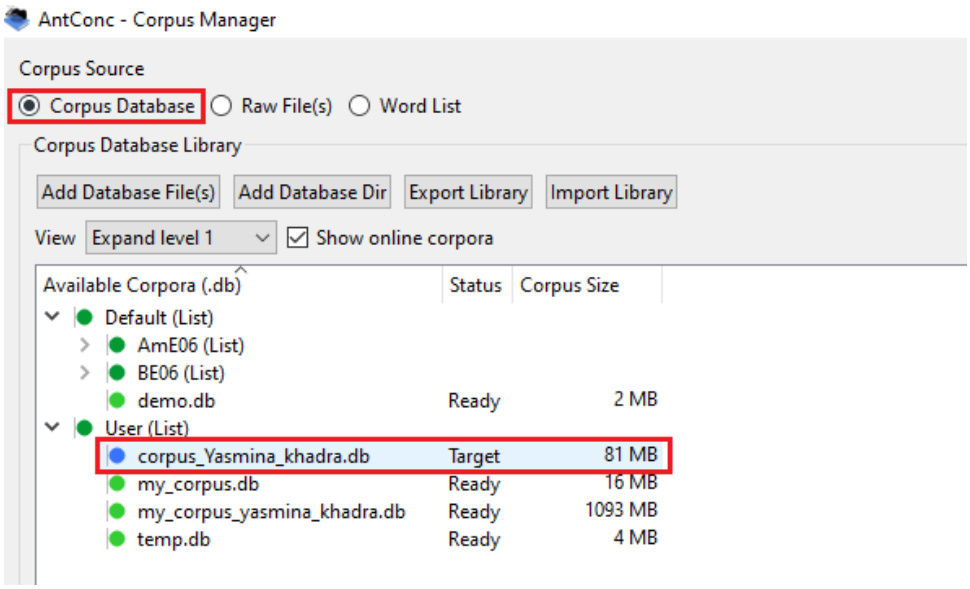


Dans l'exemple suivant, une base de données qui contient 22 fichiers correpondant aux 22 romans de l'écrivain algérien Yasmin Khadra va être créée sous le nom ***corpus_Yasmina_Khadra***. En cochant maintenant l'option ***Raw File(s)*** (1) puis en donnant un nom à la base (2), après un click sur le bouton ***Add File(s)*** (3), le programme demande l'adresse des fichiers à ajouter et leur ouverture. Pour finaliser l'opération, un click sur le bouton ***Create*** (4) démarre l'opération. Enfin une fenêtre signale la réussite de la création du corpus (6) et déjà le nom de ce corpus apparaît (5) dans le menu ***Target Corpus***.



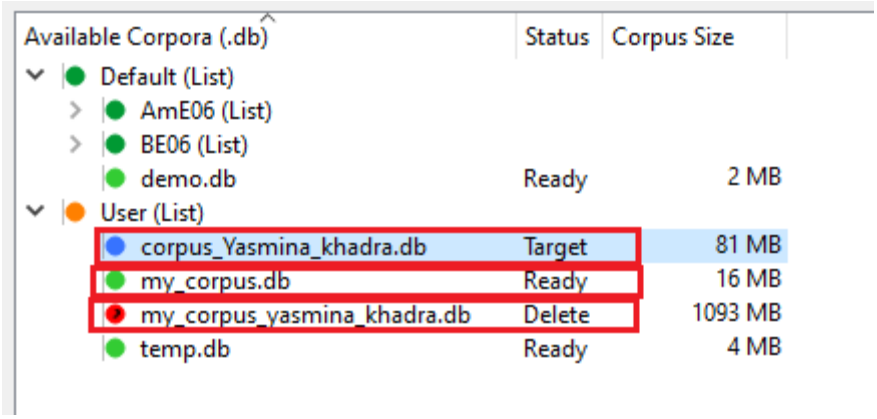
Pour une première utilisation, on peut se contenter de la configuration basique (7). Plus tard, des tests avec les options proposées permettront d'améliorer la prise en main du logiciel et de ses nouveautés.

De retour à ***Corpus Databse***, on vérifie que le corpus ***corpus_Yasmina_Khadra*** a bien été créé.

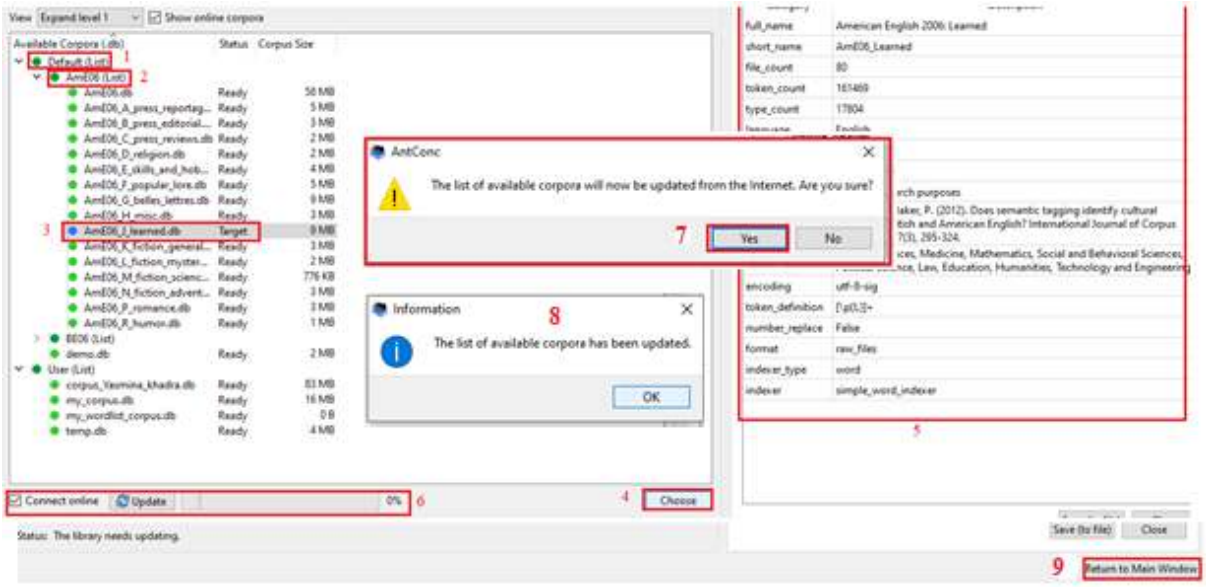


Le statut du corpus a une signification selon la couleur de la pastille située à gauche du nom du corpus et qui change avec un ou deux clics successifs dessus :

- La pastille bleue (***Target***) signifie que la base est ciblée pour toute modification ;
- La pastille bleue (***Ready***) signifie que la base est prête à l'utilisation ;
- La pastille bleue (***Delete***) signifie que la base est séléctionnée pour être effacée.

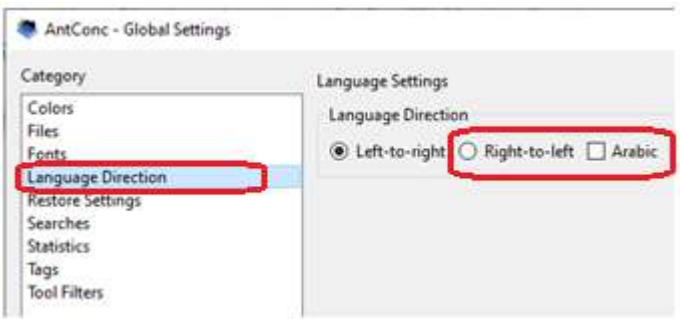


Il reste à savoir qu'il est possible de profiter d'une opération qui concerne des bases de données en langue anglaise déjà programmées qu'il faut sélectionner (1-2-3-4-5) et télécharger sur internet (6). Une fois l'opération acceptée (7) et la confirmation de la mise à jour de la base de données validée (8), le retour à la fenêtre principale est assuré via le bouton ***Return to Main Window*** (9).

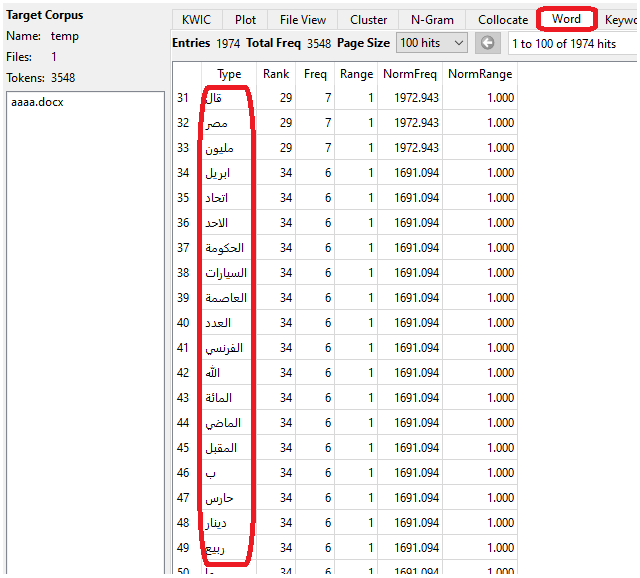


4.2.3. Le traitement de la langue arabe

AntConc intègre également la langue arabe avec la reconnaissance du sens de l'écriture de droite à gauche et qu'il faut configurer avant de lancer la recherche (***Setting > Global Settings > Language Direction > Arabic***)

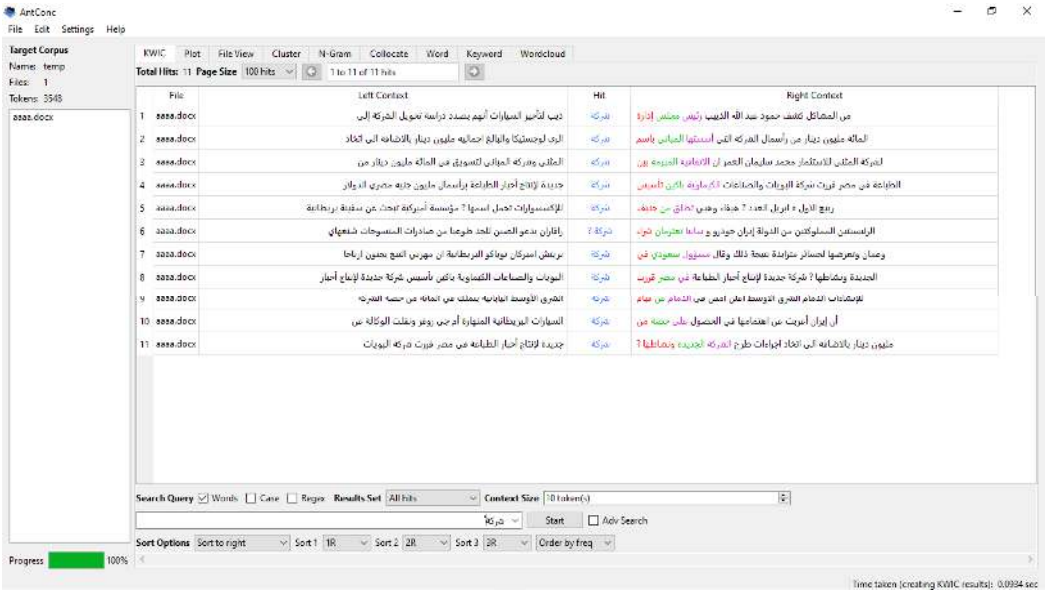


Voici un exemple du résultat de l'outil ***Word*** qui indexe les mots du texte.



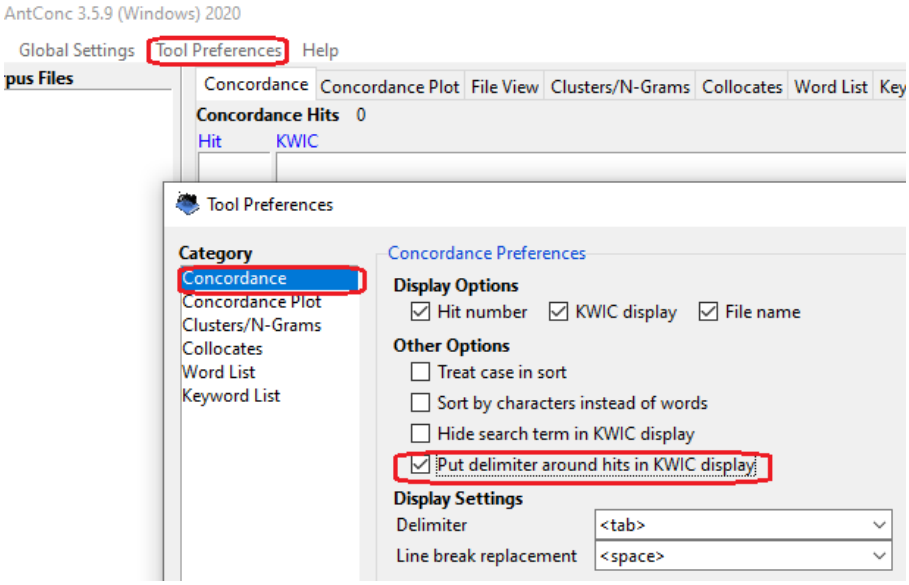
	Type	Rank	Freq	Range	NormFreq	NormRange
31	قال	29	7	1	1972.943	1.000
32	مصر	29	7	1	1972.943	1.000
33	مليون	29	7	1	1972.943	1.000
34	ابريل	34	6	1	1691.094	1.000
35	اتحاد	34	6	1	1691.094	1.000
36	الاحد	34	6	1	1691.094	1.000
37	الحكومة	34	6	1	1691.094	1.000
38	السيارات	34	6	1	1691.094	1.000
39	العاصمة	34	6	1	1691.094	1.000
40	العدد	34	6	1	1691.094	1.000
41	الفرنسي	34	6	1	1691.094	1.000
42	الله	34	6	1	1691.094	1.000
43	المائة	34	6	1	1691.094	1.000
44	الماضي	34	6	1	1691.094	1.000
45	المقبل	34	6	1	1691.094	1.000
46	ب	34	6	1	1691.094	1.000
47	حارس	34	6	1	1691.094	1.000
48	دينار	34	6	1	1691.094	1.000
49	ربيع	34	6	1	1691.094	1.000

et ce que ça donne avec l'outil ***KWIC***.

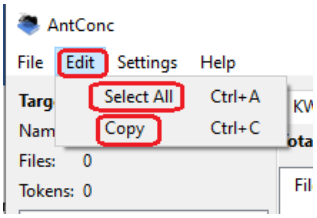


4.2.4. L'enregistramant des résultats

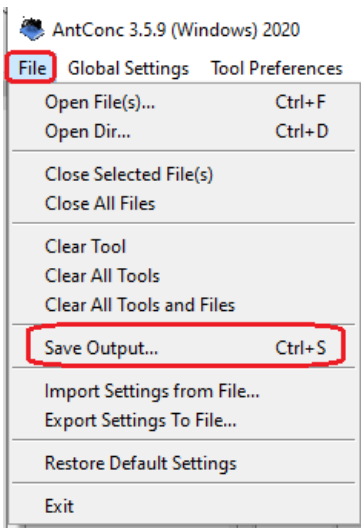
Une autre nouveauté concerne le rapatriement des résultats. Il n'est plus obligatoire, comme dans les anciennes versions, de passer par l'option ***Put delimiter around hits in KWIC display*** du menu ***Tool Preferences*** pour obtenir les tabulations qui séparent les contextes gauche et droit de la requête.

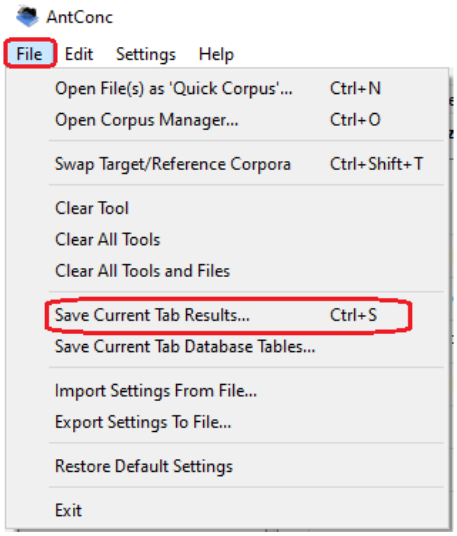


Les transferts ne passent plus obligatoirement par la création d'un fichier ***.txt***, mais les résultats sont selectionnés et copiés (***Edit > Select All > Copy***) directement dans le presse-papier pour être collés dans n'importe quel éditeur de texte.



Pour l'enregistrement classique des résultats dans des fichiers ***.txt***, comme dans les anciennes versions avec ***Save Output…***, le menu ***File*** propose une nouvelle opération (***File > Save Current Tab Results***).



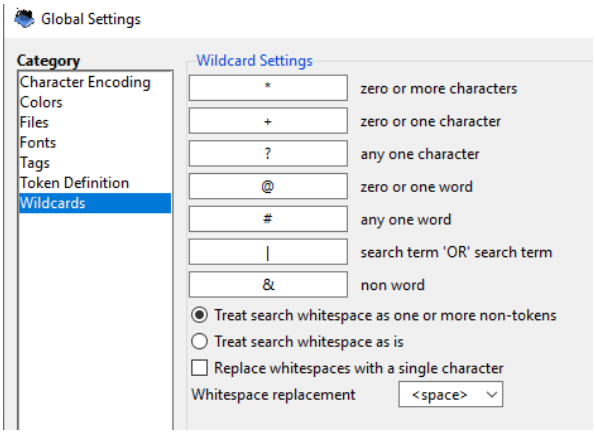


De plus, pour la langue arabe, lors du transfert des résultats des concordances obtenus vers ***Excel***, l'ordre des colonnes est correct, contrairement aux anciennes versions où il fallait inverser l'ordre des colonnes relatives aux contextes gauches et aux contextes droits pour obtenir des séquences lisibles et exploitables.

	A	B	C	D
1	aaaa.doc	ذيب لتأجير السيارات أنهم بصدد دراسة تحويل	شركة	من المشاكل كشف حمود عبد الله الذيبب رئيس مجلس إدارة
2	aaaa.doc	الري لوجستيكا والبالغ اجماليه مليون دينار بالاضافة	شركة	المائة مليون دينار من رأسمال الشركة الي أسستها المباني باسم
3	aaaa.doc	المثنى وشركة المباني لتسويق في المائة مليون دينار من	شركة	لشركة المثنى للاستثمار محمد سليمان العمر ان الاتفاقية المبرمة بين
4	aaaa.doc	جديدة لإنتاج أحبار الطباعة برأسمال مليون جنيه	شركة	الطباعة في مصر قررت شركة البويات والصناعات الكيماوية باكين تأسيس
5	aaaa.doc	للإكسسوارات تحمل اسمها ? مؤسسة أميركية تبحث	شركة	ربيع الاول هـ ابريل العدد ? هيفاء وهبي تطلق من جنيف
6	aaaa.doc	رافاران يدعو الصين للحد طوعيا من صادرات	شركة	الرئيسيتين المملوكتين من الدولة إيران خودرو و سايبا تعتزمان شراء
7	aaaa.doc	بريتش اميركان توباكو البريطانية ان مهربي التبغ يجنون	شركة	وعمان وتعرضها لخسائر متزايدة نتيجة ذلك وقال مسؤول سعودي في
8	aaaa.doc	البويات والصناعات الكيماوية باكين تأسيس شركة	شركة	الجديدة ونشاطها ? شركة جديدة لإنتاج أحبار الطباعة في مصر قررت
9	aaaa.doc	الشرق الأوسط اليابانية بتملك في المائة من حصة	شركة	للإنشاءات الدمام الشرق الاوسط اعلن امس في الدمام عن قيام
10	aaaa.doc	السيارات البريطانية المنهارة أم جي روفر ونقلت الوكالة	شركة	أن إيران أعربت عن اهتمامها في الحصول على حصة من
11	aaaa.doc	جديدة لإنتاج أحبار الطباعة في مصر قررت شركة	شركة	مليون دينار بالاضافة الى اتخاذ اجراءات طرح الشركة الجديدة ونشاطها ?
12				

4.2.5. Les requêtes avec les métacaractères

Pour les recherches avec les métacarctères, des modifications substantielles ont été apportées aux opérateurs logiques. La syntaxe de base suit à peu près le ***Common Elementary Query Language*** (CEQL)[59].



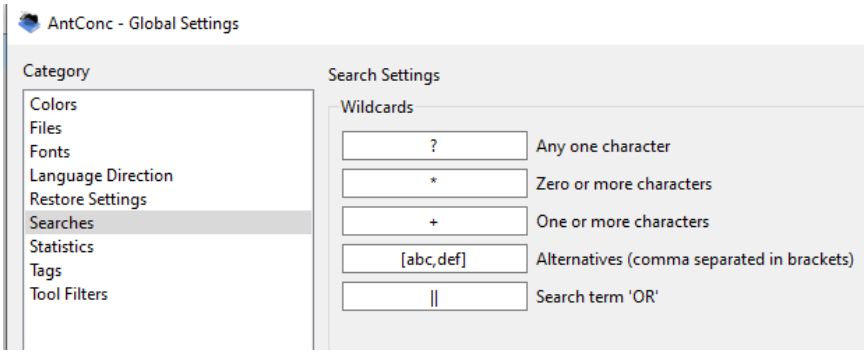
Les recherches d'une suite de deux mots (ou plus) avec un mot saisi et un autre aléatoire et séparés par une ponctuation (@), ou séparés ou non par une ponctuation (#) sont obtenus avec l'astérisque (*) avec possibilité d'augmenter le nombre de mots aléatoires en séparant les astérisques par une espace. Sans espace, l'astérique fonctionne comme dans les anciennes versions[60].

Les recherches d'un mot ou d'une suite de mots suivis obligatoirement d'une ponctuation (&) n'existe plus tout comme « zéro ou un seul caractère » (+).

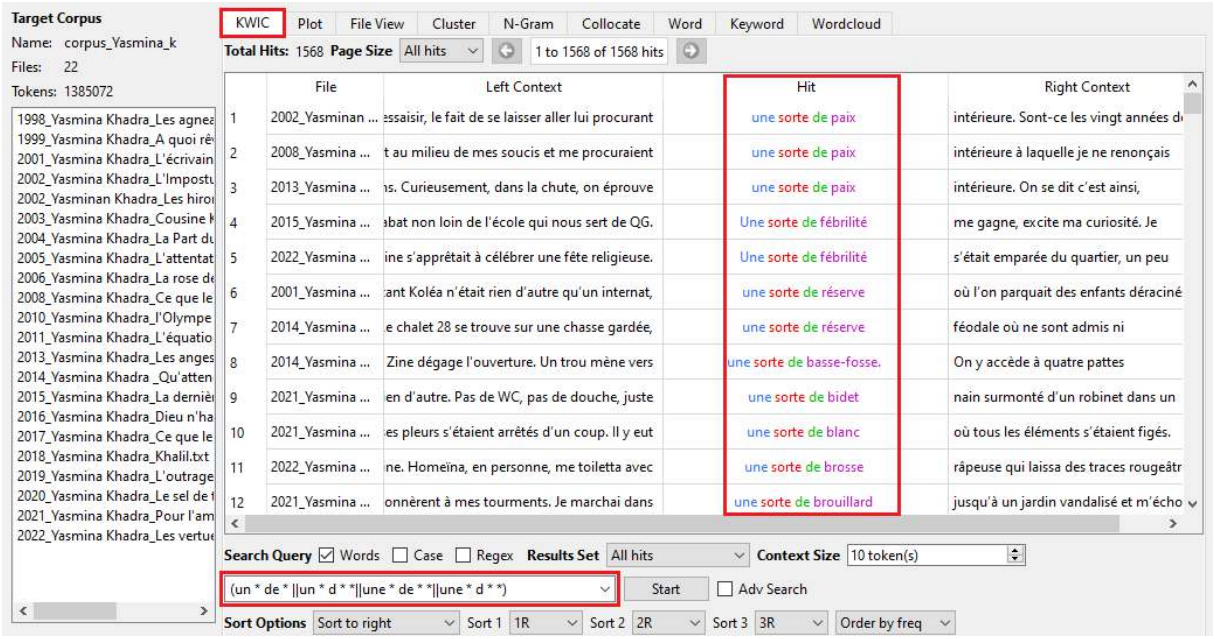
Le slash droit (|) est désormais doublé (||).

Une recherche nouvelle est ajoutée ([abc,def]) pour trouver ce qui, entre cochets, est séparé par une virgule.

[59] https://cwb.sourceforge.io/ceql.php.
[60] Voir § 2.1.1.3.1.



Avec une requête comme [***un * de * *||un * d * *||une * de * *||une * d * ****] génèrera les structures [déterminant indéfini singulier + N_1 + la préposition DE + N_2] du type ***une sorte de paix***, ***une dizaine de jours***, ***une fraction de seconde***, etc.



Il est toutefois préférable de garder une ancienne version, par exemple ***AntConc 3.5.9***, et de l'utiliser en vue de bénéficier des métacaractères éliminés.

Avec les anciennes versions jusqu'à ***AntConc v3.5.9***, la lemmatisation permet de créer la liste des seuls lemmes d'un texte[61].

4.2.6. La catégorisation

À partir d'***AntConc 4.0.1***, pour procéder à des requêtes catégorisées, il est recommandé, avant de charger les fichiers dans ***AntConc***, de les baliser et de les lemmatiser avec les différentes parties du discours[62], ***Part-Of-Speech*** (POS), par

[61] Voir § 2.7.1.
[62] Ne pas oublier la question de l'apostrophe, voir § 3.1.

exemple, en utilisant l'application ***TagAnt***[63]. Ensuite, les formes lemmatisées peuvent être affichées en choisissant les options du menu ***Word***.

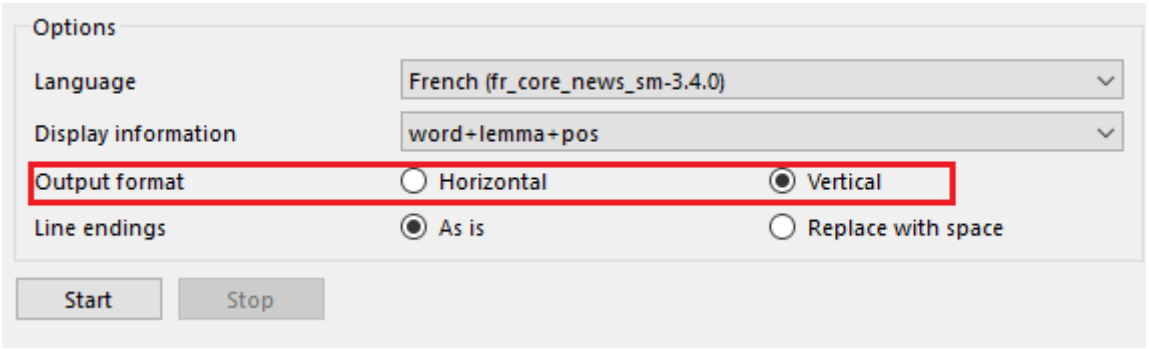
Après le choix des fichiers avec ***Input File(s)*** et ***Open File(s)*** (1 et 2), on procède aux réglages de la langue avec l'option ***Language*** (3). Un large choix est possible entre le chinois, l'anglais, le français, l'allemand, l'italien, le japonais et l'espagnol. Pour le fraçais, le meilleur choix est ***french (fr_core_nems_sm-3.4.0)***. Les fichiers tagués sont enregistrés automatiquement dans un sous-dossier nommé ***tagged*** et placé dans le même dossier que les fichiers soumis au programme.

Il faut également choisir entre les différentes combinaisons offertes pour l'affichage des informations avec ***Display information*** (4).

word	mot seul
pos	seulement la catégorie grammaticale du mot sans le mot
pos_tag	pas de différence avec le précédent
lemma	Seulement le lemme, sans le mot
word + pos	le mot suivi de sa catégorie grammaticale
word + pos_tag	pas de différence avec le précédent
word + lemma	le mot suivi de son lemme
word + pos_tag + lemma	le mot suivi de sa catégorie grammaticale puis de son lemme
word + lemma + pos	le mot suivi de son lemme pui de sa catégorie grammaticale

Pour les autres configurations, il est préfarable de garder celles d'origine.

En cas de doute sur le résultat, il est possible de refaire la catégorisation en choisissant l'option de sortie ***Output Format*** et choisir la sortie en colonnes (***Vertical***).

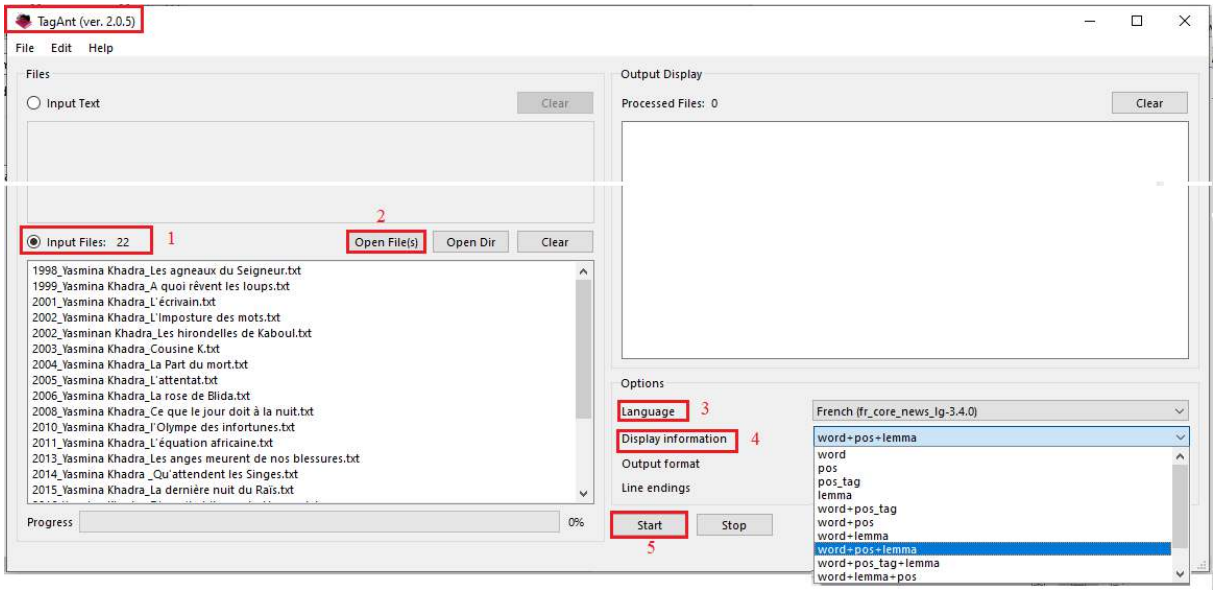


Le résultat sera sous forme de tableau facilement transférable dans ***Excel*** où il est possible, en profitant des grandes capacités de filtrage du tableur, d'automatiser les corrections et puis de convertir le résultat corrigé dans ***Word*** et retrouver l'ordre linéaire du texte.

Mot	Lemme	Catégorie
Seule	seul	ADJ
une	un	DET
bande	bande	NOUN
de	de	ADP

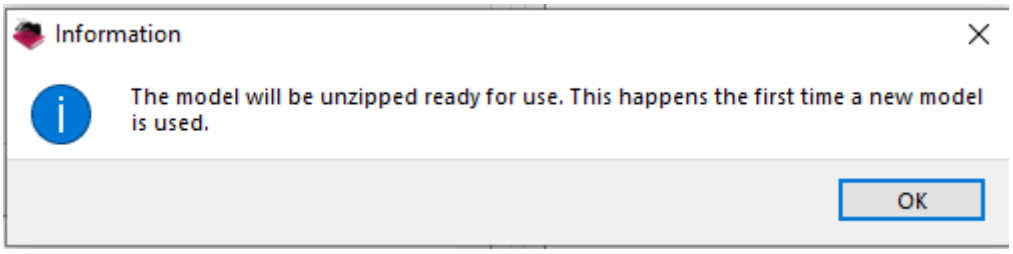
[63] Téléchargeable gratuitement sur lr site du concepteur https://www.laurenceanthony.net/software/tagant/

galopins	galopin	NOUN
continue	continue	VERB
d’	d’	ADV
écumer	écumer	VERB
les	le	DET
recoins	recoin	NOUN
,	,	PUNCT
aussi	aussi	ADV
ardente	ardent	ADJ
qu’	qu’	SCONJ
un	un	DET
essaim	essaim	NOUN
de	de	ADP
frelons	frelon	NOUN
.	.	PUNCT



En pressant le bouton ***Start***, une fenêtre de dialogue signale que le programme doit procéder à la compression de la base, opération obligatoire seulement à la première manipulation du corpus.

Une fois l’opération lancée, il faut attendre la fin de la progression de catégorisation des fichiers qui se fera texte après texte.

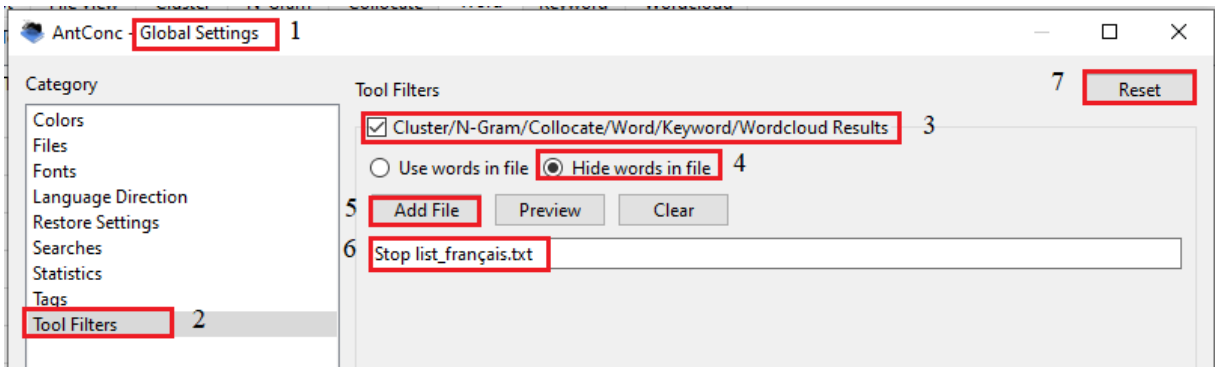


4.2.7. L'option *Tool Filters*

Dans le menu Setting>Category>Tool Filters, deux options sont proposées.

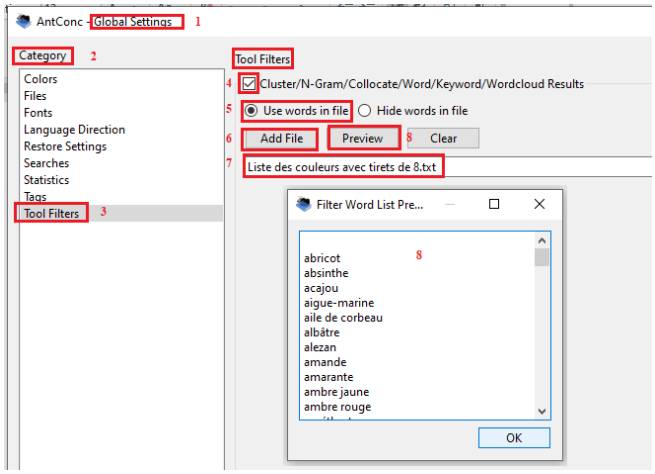
4.2.7.1. Hide words in file

L'option ***Stop List*** des anciennes versions est désormais placée dans le menu ***Global Settings*** (1) > ***Tool Filters*** (2) sous le nouveau nom de ***Hide words in file***. Il faut cocher la case ***Cluster/N-Gram/...*** (3) et l'option ***Hide words in file*** (4) et en renseignant le programme sur le fichier contenant les mots à ne pas considérer pendant la recherche (5)[64] et vérifier que le nom du fichier en question apparaît (6).



4.2.7.2. Use words in file

L'option ***Use wordw in file*** (5) fonctionne comme ***Advenced Search*** de l'onglet ***KWIC***. En choissant ***Add File*** (6), le nom du fichier contenant les mots dont on voudrait obtenir les statistiques apparaît (7). Une fois le contenu de ce fichier vérifié (8) et la validation faite, l'outil ***Word*** fournira les résultats.



4.2.8. Les modifications dans les outils

Une nouveauté dans le menu *N-**Gram***, qui a désormais son propre onglet séparé de ***Clusters***, consiste dans la possibilité de rechercher des suites de mots et leurs

[64] Voir § 2.7.2.

fréquences.[65] Avec une suite de quatre mots, le logiciel affiche, par exemple des structures comme *je n'ai pas, je n'avais pas, je n'étais pas*, etc., avec la fréquence de chaque suite.

	Type	Rank	Freq	Range
1	je n ai pas	1	34	1
2	je n avais pas	2	16	1
3	je n étais pas	3	11	1
4	je n arrivais pas	4	6	1
5	je n arrive pas	4	6	1
6	je n aimais pas	6	4	1
7	je n eus pas	7	3	1
8	je n osais pas	8	2	1
9	je n ose pas	8	2	1
10	je n éprouvais pas	8	2	1
11	je n aime pas	11	1	1
12	je n aimerais pas	11	1	1
13	je n allais pas	11	1	1
14	je n appréciai pas	11	1	1
15	je n arrêtais pas	11	1	1
16	je n aurais pas	11	1	1
17	je n entendais pas	11	1	1
18	je n entendis pas	11	1	1
19	je n entretiens pas	11	1	1
20	je n habite pas	11	1	1

Avec ***Open Slots***, on peut rechercher une suite avec un nombre donné, par exemple quatre (4), avec des emplacements ouverts, c'est-à-dire que le nombre de créneaux dans le ***n-gram*** peut prendre plusieurs valeurs. Le résultat ramassera toute la série visible dans l'image précédente en deux groupes, si le créneau est 1.

KWIC Plot File View Cluster N-Gram Colloc

N-Gram Types 220359 **N-Gram Tokens** 245704 **Page Size** All

	Type	Rank	Freq	Range
1	je ne + pas	1	132	1
2	je n + pas	2	98	1
3	ce n + pas	3	74	1

Si on choisit le créneau 2, les occurrences sont alors ramenées à un seul groupe.

KWIC Plot File View Cluster N-Gram

N-Gram Types 80076 **N-Gram Tokens** 122852 **Page Siz**

	Type	Rank	Freq	Range	
1	je + + pas	1	232	1	0.017
2	de + + de	2	222	1	0.626
3	ne + + pas	3	214	1	0.103

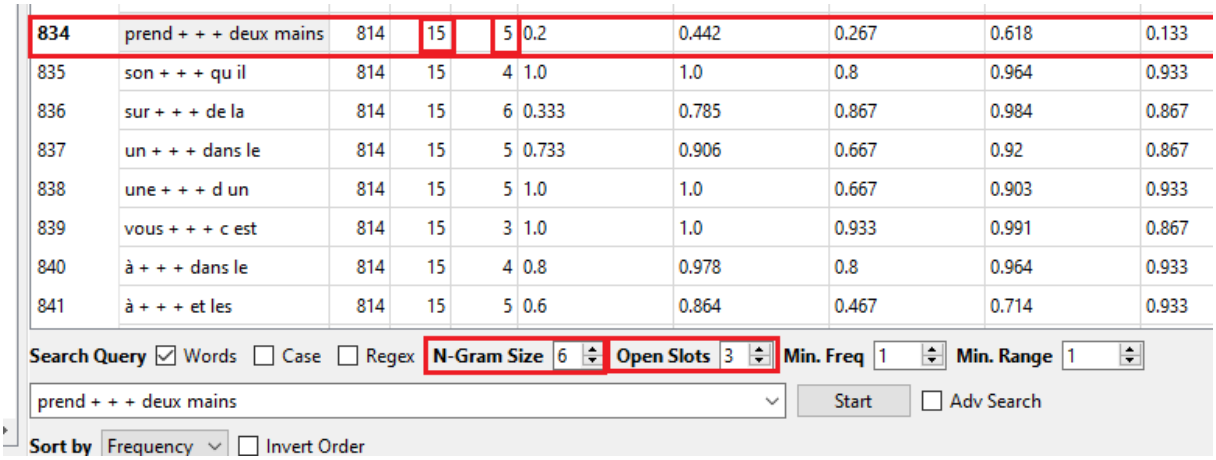
Le but de cette manipulation est de procéder à une recherche de combinaisons de mots et d'obtenir un nombre limité de classes sans trop de détails. Cela donne à voir également les cas des structures corrélatives, ou les cas de l'insertion d'un adverbe, par

[65] Voir § 4.2.2. pour l'ouverture des fichiers avec ***Open Corpus Manager***.

exemple, entre un auxiliaire et le participe passé d'un verbe conjugué à un temps composé.

On peut découvrir aussi comment un écrivain - ou tout locuteur - répète les mêmes séries de mots en variant selement un ou deux mots.

Dans l'exemple suivant la requête demande les suites de six mots (***N-Gram Size***) avec un créneau de 3 (***Open Slots***), ce qui donne pour cinq œuvres une fréquence de 15 occurrences pour la structure ***prend + + + deux mains***.



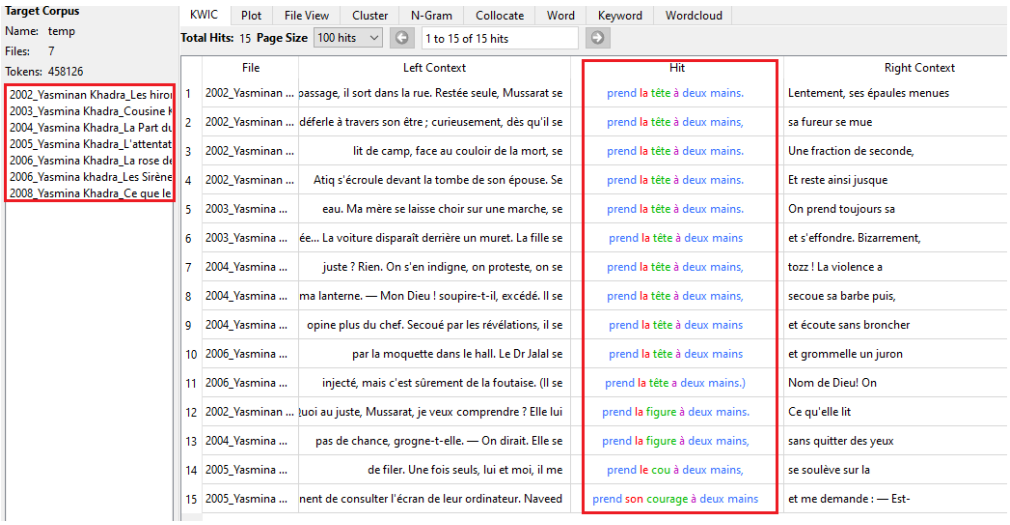
834	prend + + + deux mains	814	15	5	0.2	0.442	0.267	0.618	0.133
835	son + + + qu il	814	15	4	1.0	1.0	0.8	0.964	0.933
836	sur + + + de la	814	15	6	0.333	0.785	0.867	0.984	0.867
837	un + + + dans le	814	15	5	0.733	0.906	0.667	0.92	0.867
838	une + + + d un	814	15	5	1.0	1.0	0.667	0.903	0.933
839	vous + + + c est	814	15	3	1.0	1.0	0.933	0.991	0.867
840	à + + + dans le	814	15	4	0.8	0.978	0.8	0.964	0.933
841	à + + + et les	814	15	5	0.6	0.864	0.467	0.714	0.933

Search Query ☑ Words ☐ Case ☐ Regex N-Gram Size 6 Open Slots 3 Min. Freq 1 Min. Range 1

prend + + + deux mains Start ☐ Adv Search

Sort by Frequency ☐ Invert Order

Un double-clic ouvre l'onglet (***KWIC***) et affiche les concordances où on découvre les structures ***prend (la tête/la figure/le cou/son courage) à deux mains***.



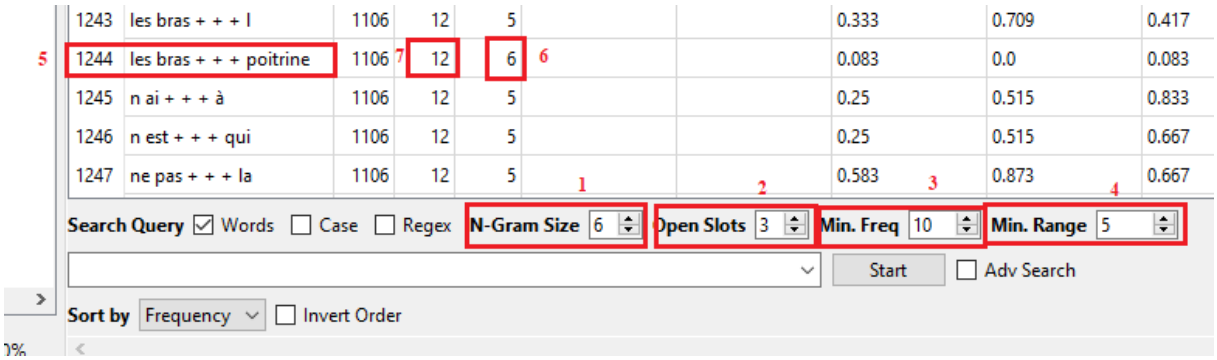
Target Corpus
Name: temp
Files: 7
Tokens: 458126
2002_Yasminan Khadra_Les hirou
2003_Yasmina Khadra_Cousine K
2004_Yasmina Khadra_La Part du
2005_Yasmina Khadra_L'attentat
2006_Yasmina Khadra_La rose de
2006_Yasmina khadra_Les Sirène
2008_Yasmina Khadra_Ce que le

KWIC | Plot | File View | Cluster | N-Gram | Collocate | Word | Keyword | Wordcloud

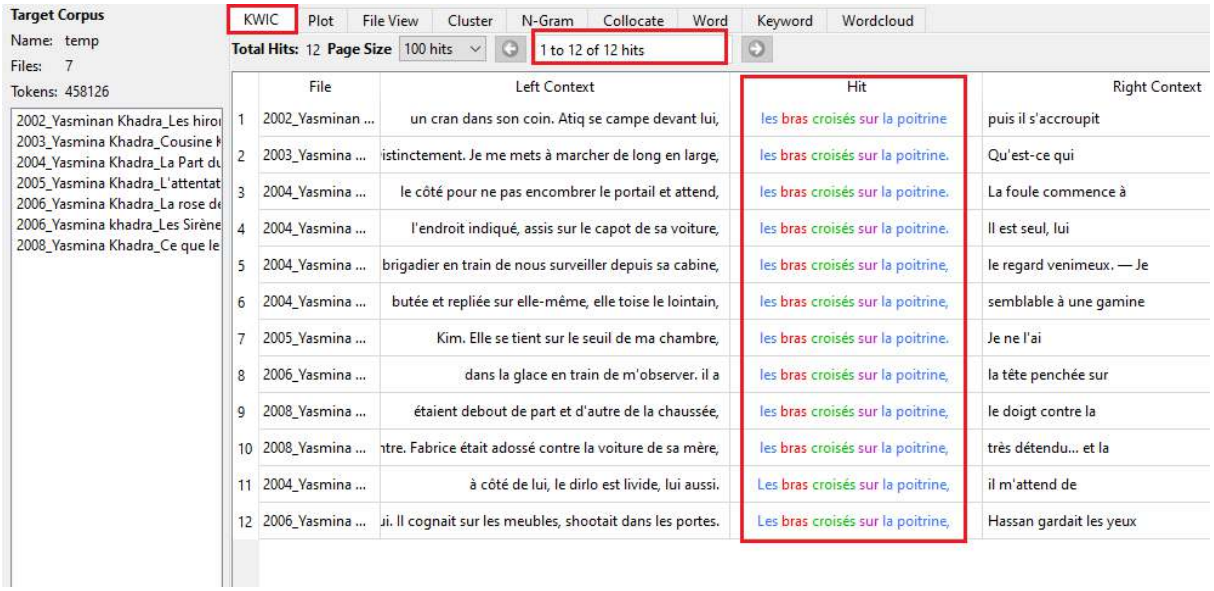
Total Hits: 15 Page Size 100 hits 1 to 15 of 15 hits

	File	Left Context	Hit	Right Context
1	2002_Yasminan ...	passage, il sort dans la rue. Restée seule, Mussarat se	prend la tête à deux mains.	Lentement, ses épaules menues
2	2002_Yasminan ...	déferle à travers son être ; curieusement, dès qu'il se	prend la tête à deux mains,	sa fureur se mue
3	2002_Yasminan ...	lit de camp, face au couloir de la mort, se	prend la tête à deux mains.	Une fraction de seconde,
4	2002_Yasminan ...	Atiq s'écroule devant la tombe de son épouse. Se	prend la tête à deux mains.	Et reste ainsi jusque
5	2003_Yasmina ...	eau. Ma mère se laisse choir sur une marche, se	prend la tête à deux mains.	On prend toujours sa
6	2003_Yasmina ...	ée... La voiture disparaît derrière un muret. La fille se	prend la tête à deux mains	et s'effondre. Bizarrement,
7	2004_Yasmina ...	juste ? Rien. On s'en indigne, on proteste, on se	prend la tête à deux mains,	tozz ! La violence a
8	2004_Yasmina ...	ma lanterne. — Mon Dieu ! soupire-t-il, excédé. Il se	prend la tête à deux mains,	secoue sa barbe puis,
9	2004_Yasmina ...	opine plus du chef. Secoué par les révélations, il se	prend la tête à deux mains	et écoute sans broncher
10	2006_Yasmina ...	par la moquette dans le hall. Le Dr Jalal se	prend la tête à deux mains	et grommelle un juron
11	2006_Yasmina ...	injecté, mais c'est sûrement de la foutaise. (Il se	prend la tête a deux mains.)	Nom de Dieu! On
12	2002_Yasminan ...	uoi au juste, Mussarat, je veux comprendre ? Elle lui	prend la figure à deux mains.	Ce qu'elle lit
13	2004_Yasmina ...	pas de chance, grogne-t-elle. — On dirait. Elle se	prend la figure à deux mains,	sans quitter des yeux
14	2005_Yasmina ...	de filer. Une fois seuls, lui et moi, il me	prend le cou à deux mains,	se soulève sur la
15	2005_Yasmina ...	nent de consulter l'écran de leur ordinateur. Naveed	prend son courage à deux mains	et me demande : — Est-

Pour accélérer la recherche, on peut règler la fréquence minimale et augmenter le nombre de textes où la structure est employée. Ici, avec une requête de six mots (1) et un crénau de 3 (2), pour une fréquence minimale de 10 (3) et un minimum de textes de 5 (4), le résultat donne ***les bras + + + poitrine*** (5), utilisé dans six œuvres de Yasmina Khadra sur les sept examinées (6) avec une fréquence de 12 (7).



Le résultat donne ***les bras croisés sur la poitrine***.



4.2.9. Tableau récapitulatif des changements

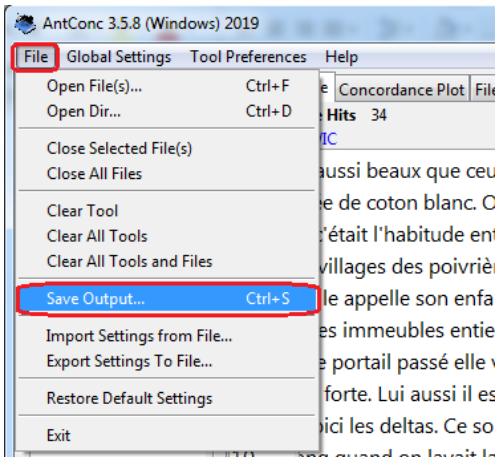
Voici un tableau qui résume les modifications les plus courantes apportées aux anciennes versions. Pour certains outils, l'emplacement a été changé. Pour d'autres, des améliorations ont augmenté l'efficacité pour une plus grande précision de la recherche. D'autres ont disparu ou ont été séparés ou regroupés. Enfin des nouveautés comme la création des bases de données et l'ajout de l'arabe qui faisait défaut à cause du sens de l'écriture, etc.

Version 3.5.9	**Version 4.2.0**	**Observation**
Clone Results	Ø	disparition
Concordance Plot	*Plot* avec *Overlay* et *Color*	développement
Context Size (Taille du contexte)	*Search Windows Size*	**Changement de nom**
File **>*Open File(s)/Open Directory***	***File***>*Open File(s) as 'Quick Corpus'* et *Corpus Database*	nouveauté
Global Settings et *Tool Preferences*	*Settings*	**déplacement**
Global Settings> *File* **>*File Settings***>txt, html, htm, xml	*Global Settings*> *Global Settings* **>***Files*>*Corpus File Types* : txt, pdf, docx, html, srt, sub, tsv, csv	**développement**
N-Gram et Clusters	*Onglets séparés*	**disjonction**

Version 3.5.9	Version 4.2.0	Observation
Plot Zoom	***Plot Zoom + Overlay + color***	**développement**
Put delimiter around hits in KWIC display du menu *Tool Preferences*	Ø	**disparition**
Save output	*Edit,* avec *Select All* et *Copy*	nouveauté
Save Output	*File* > *Save Current Tab Results*	**Nouveau nom**
Show Every Nth Row	*Result Set*	nouveauté
Sort	*Start*	**regroupement**
Sort	Ø	disparition
Stop List	*Hide words in file*	Changement de nom
*Global Setting**s*** *> Wildcards*	*Searches*	Changement de nom
*Wildcards @ et * *	*Wildcards * *	**nouveauté**
Ø	***KWIK>****Page Size*	nouveauté
Ø	*Wordcloud*	nouveauté
Ø	***File>****Corpus Database*	nouveauté
Ø	*Setting* > *Global Settings* > *Language Direction* > *Arabic*	nouveauté
Ø	*Open Slots*	nouveauté

Chapitre Cinquième
Enregistrement des résultats et manipulations avec Excel

Tous les résultats qui apparaissent dans la fenêtre principale, à part le texte dans l'onglet F***ile View***, peuvent être rapatriés et enregistrés sur le disque dur de l'ordinateur dans un fichier de type .txt. ***AntConc*** fonctionnant dans la mémoire vive de l'ordinateur ne conserve pas les opérations après fermeture de la session. Il propose par défaut d'enregistrer le fichier sur le ***Bureau*** sous le nom antconc_***results.txt***. Il suffit de le renommer et de le stocker dans un autre endroit. Le raccourci (Ctrl+S) facilite ce transfert.



Le contenu du fichier sauvegardé a la forme suivante

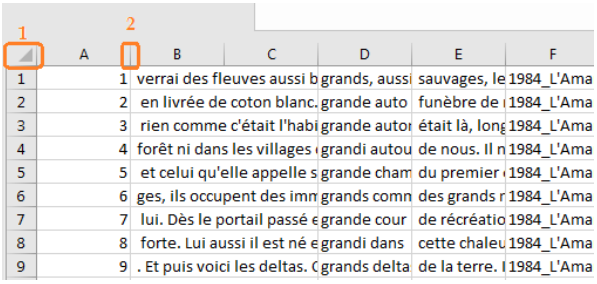
1 verrai des fleuves aussi beaux que ceux-là, aussi grands, aussi sauvages, le Mékong et ses bras qui descendent 1984_L'Amant - Marguerite Duras.txt
2 en livrée de coton blanc. Oui, c'est la grande auto funèbre de mes livres. C'est la Morris 1984_L'Amant - Marguerite Duras.txt
3 rien comme c'était l'habitude entre eux. Sa grande automobile était là, longue et noire, avec, à l' 1984_L'Amant - Marguerite Duras.txt
4 forêt ni dans les villages des poivrières. Tout a grandi autour de nous. Il n'y a plus d' 1984_L'Amant - Marguerite Duras.txt
5 et celui qu'elle appelle son enfant dans sa grande chambre du premier étage, celle où elle mettait des 1984_L'Amant - Marguerite Duras.txt
6 ges, ils occupent des immeubles entiers, ils sont grands comme des grands magasins, des casernes, ils sont ouver 1984_L'Amant - Marguerite Duras.txt

Chaque occurrence comporte un numéro d'ordre + contexte gauche + une espace + le mot-clé + le contexte droit + le nom du fichier du texte.

Avant d'aller plus loin, il faut donc vérifier le codage du texte et l'existence d'une espace avant et après le mot-clé.

5.1. Le Tableur *Excel*[66]

En sélectionnant la totalité du fichier (Ctrl+A) et en le copiant (Ctrl+C) puis en le collant dans une feuille vierge d'*Excel* de *Microsoft Office* dans la cellule A1 (Ctrl+V), on obtient la forme suivante :

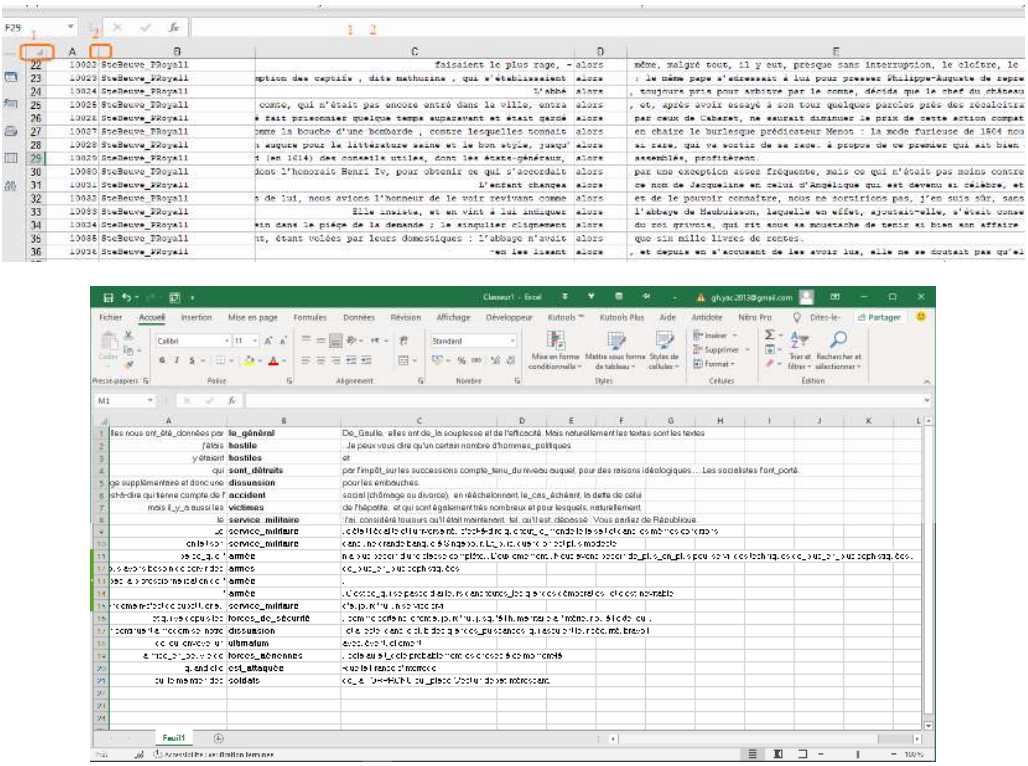


	A	B	C	D	E	F
1	1	verrai des fleuves aussi b	grands, aussi	sauvages, le	1984_L'Ama	
2	2	en livrée de coton blanc.	grande auto	funèbre de	1984_L'Ama	
3	3	rien comme c'était l'habi	grande autor	était là, long	1984_L'Ama	
4	4	forêt ni dans les villages	grandi autou	de nous. Il n	1984_L'Ama	
5	5	et celui qu'elle appelle s	grande cham	du premier	1984_L'Ama	
6	6	ges, ils occupent des imn	grands comn	des grands r	1984_L'Ama	
7	7	lui. Dès le portail passé e	grande cour	de récréatio	1984_L'Ama	
8	8	forte. Lui aussi il est né e	grandi dans	cette chaleu	1984_L'Ama	
9	9	. Et puis voici les deltas. (	grands delta	de la terre. I	1984_L'Ama	

En effet, le transfert réalisé vers ***Excel*** garde la même largeur des colonnes et les contenus des cellules n'apparaissent pas dans leurs totalités. Avec les deux manipulations suivantes, on arrive rapidement à régler le problème :

1. Simple clic sur la case du triangle (1) ;
2. Double-clic rapide sur le trait qui sépare les colonnes A et B (2) ;

Ce double clic permet d'élargir les colonnes pour montrer la totalité de leur contenu. On obtient la présentation suivante :



[66] Nous utilisons la version *Office 2019*.

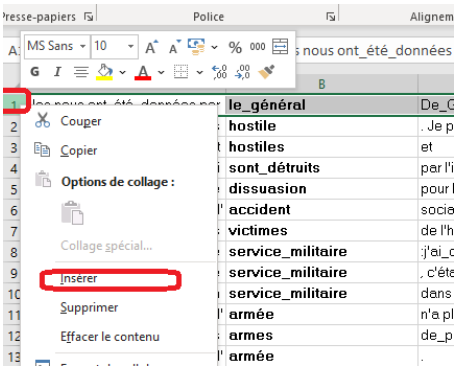
Il faut maintenant enregistrer le fichier et nommer la ***feuille1*** avec un nom explicite. Et répéter toute la procédure s'il y a d'autres transferts.

5.2. Le Tableau Croisé Dynamique d'Excel

Pour manipuler des données, ***Excel*** offre deux possiblités, une manuelle, l'autre automatique.

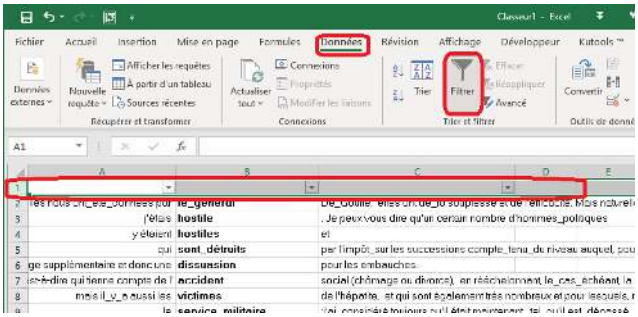
5.2.1. La procédure manuelle

Avant de commencer l'analyse, il faut insérer une nouvelle ligne vide au dessus de la première ligne du tableau en sélectionnant cette dernière avec le bouton droit de la souris et en choisissant ***Insérer***.

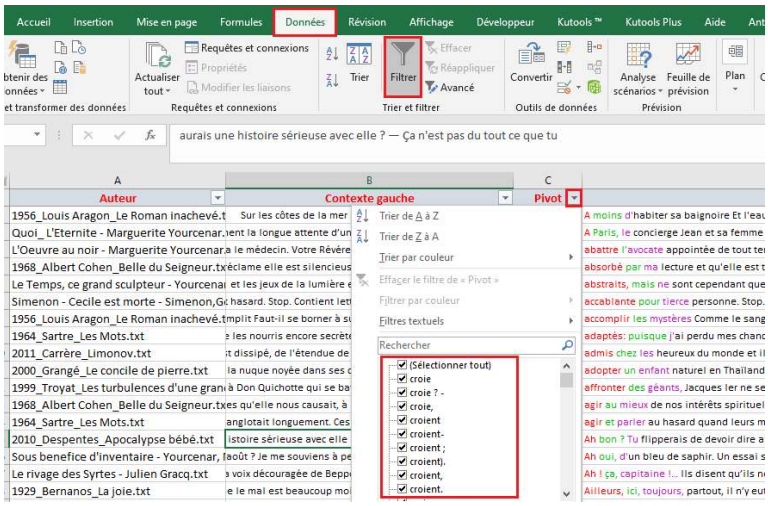


Ce qui donne la fenêtre suivante où apparaît une ligne vide au début du tableau. On profite pour donner des titres aux colonnes.

Dans le menu ***Données***, on choisit ***Filtrer*** pour obtenir les outils qui permettent de filtrer les résultats, à savoir les petits triangles noirs dirigés vers le bas. En cliquant sur l'un d'eux, ceci crée automatiquement une liste déroulante avec les mots ou les suites de mots de la colonne classés par ordre alphabétique sans répétition pour les mêmes formes.



Les possibilités de filtrage sont visibles dans une fenêtre qui donne à voir sous forme de cases cochées par défaut les différents contenus des cellules de la colonne en question. Après avoir décoché la case ***(Sélectionner tout)***, et que toutes les cases sont automatiquement décochées, on choisit de filtrer les données du tableau selon le contenu d'une ou de plusieurs cellules au choix dont on coche les cases.

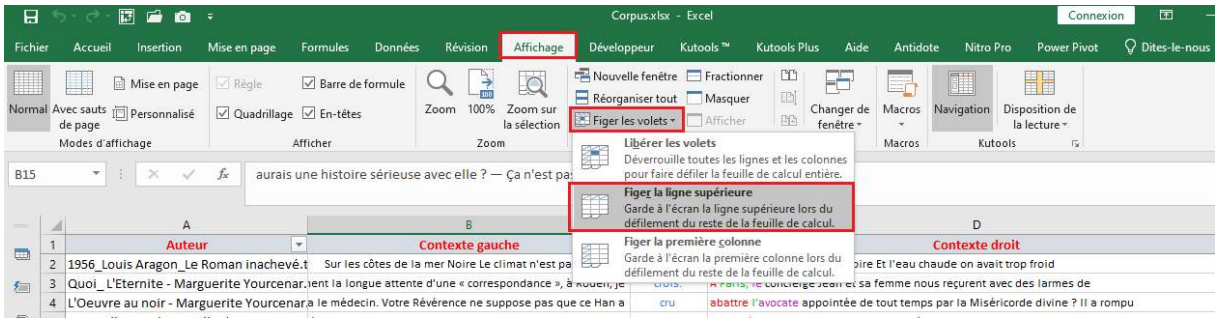


Le petit triangle noir est maintenant plus petit et est accompagné du symbole du filtrage (un entonnoir). Un autre filtrage peut être effectué sur le résultat obtenu par le premier filtre mais sur une autre colonne.

A la fin de l'opération, on n'oublie pas de restituer l'ensemble des données du tableau en cochant la case ***(Sélectionner tout)***, décochée lors du dernier filtrage effectué.

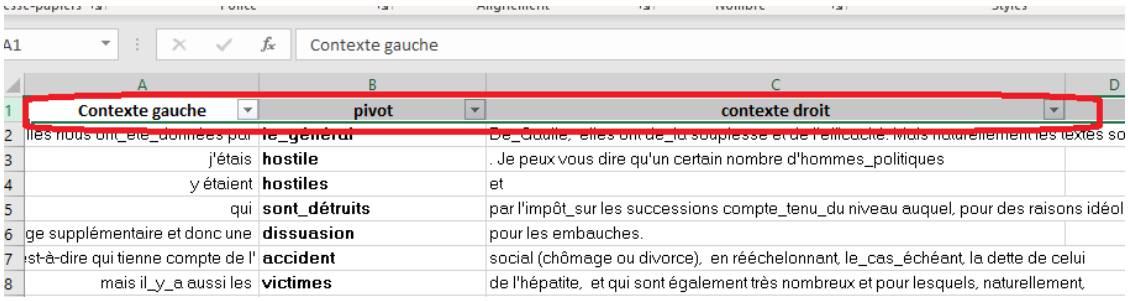
5.2.2. Figement des volets

Si le nombre des lignes du tableau dépasse le cadre de la fenêtre de départ, c-à-d à partir de 25 lignes et plus[67], il est recommandé de figer la ligne des titres (la ligne supérieure) pour qu'elle reste visible lors du défilement du tableau vers le bas.



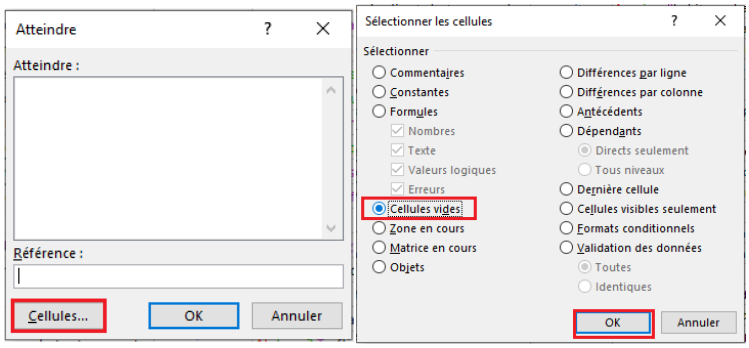
[67] ***Excel*** permet de gérer sur une seule et même feuille 1 048 576 lignes et 16 384 colonnes, mais il ne faut pas exagérer.

Si les nombre des colonnes excède lui aussi les limites de la fenêtre, il est possible de figer la première colonne si elle comporte des informations et même de figer tous les volets. Désormais, en se déplaçant dans le tableau de haut en bas ou de gauche à droite, ce qui a été figé reste visible. Les mêmes opérations permettent de libérer le tableau.

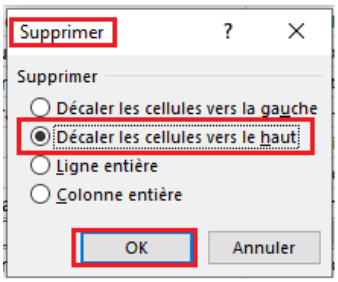


Un simple clic dans n'importe quelle cellule du tableau indique à ***Excel*** qu'on se propose de travailler sur la totalité du tableau.

Attention : si le tableau comporte une ou plusieurs lignes totalement vides, ***Excel*** ne tient pas compte des lignes non vides situées juste après : il ne filtre donc qu'une partie des données. Pour éliminer ces lignes vides, on utilse la touche du clavier **f5**[68], on clique dans un premier temps sur le bouton ***Cellules*** puis dans la seconde fenêtre qui apparaît on choisit l'option ***Cellules vides*** et on valide avec ***OK***. Les cellules vides sont sélectionnées, il suffit de les supprimer.



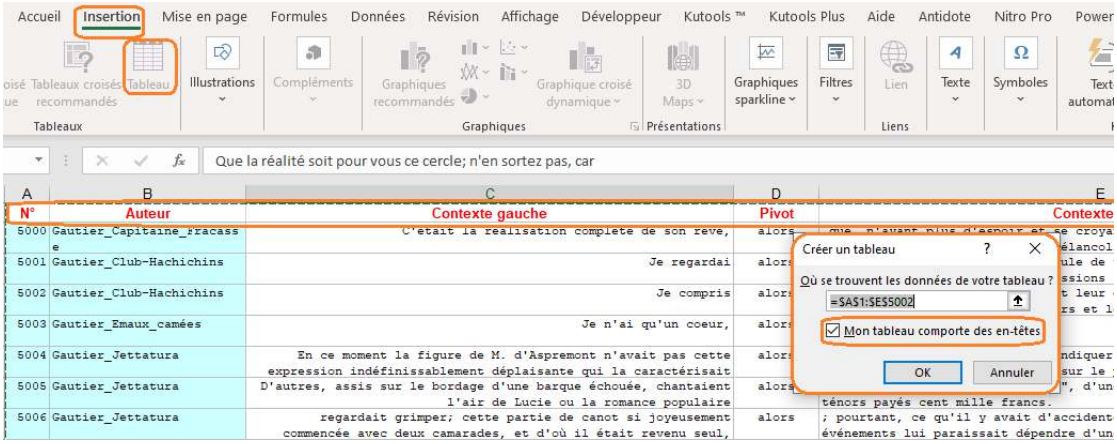
On clique avec le bouton droit de la souris sur une cellule d'une des lignes vides et on choisit la commande ***Supprimer***. Puis on demande au programme de décaler les cellules vers le haut avec l'option correspondante et on valide finalement avec ***OK.***



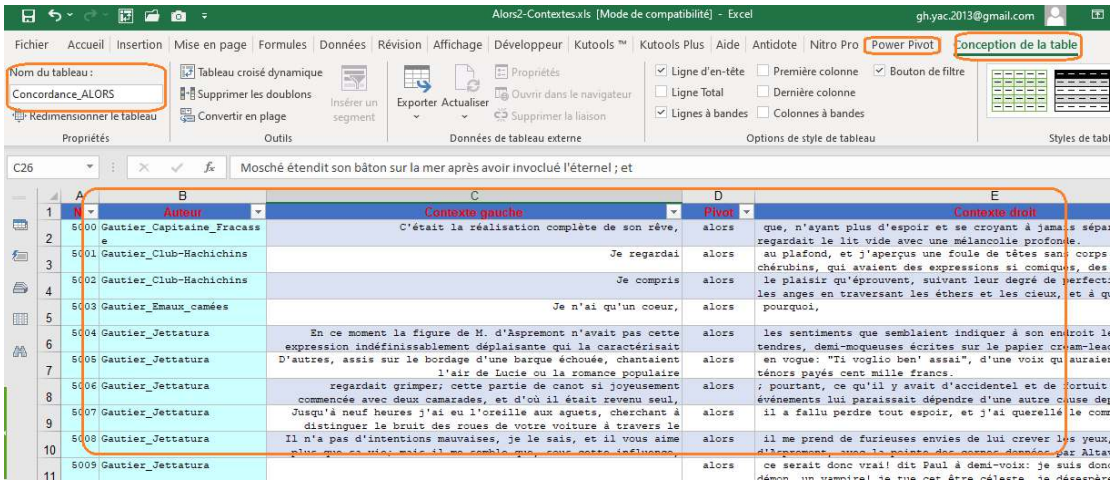
[68] Avec certains claviers, on utilise la combinaison fn + F5.

5.2.3. La procédure automatique (Tableau Croisé Dynamique)

Avant de procéder à une quelconque manipulation sur le tableau obtenu, il est fortement recommandé de signaler au logiciel qu'il s'agit d'un tableau[69]. Pour ce faire, après un simple clic dans n'importe quelle cellule du tableau des données, on va dans le menu ***Insertion***, où le logiciel propose, dans la rubrique ***Tableaux***, trois icônes : l'icône ***Tableau croisé dynamique***, ***Tableaux croisés recommandés*** et ***Tableau***. Il faut commencer par cliquer sur l'icône ***Tableau*** pour que le logiciel traite les données de la manière la plus adéquate[70]. Un cadre en pointillés signale le pourtour des cellules prises en charge et une fenêtre de dialogue apparaît pour demander si le tableau a des titres de colonnes. Enfin on valide avec ***OK***. Dans les versions antérieures d'***Excel***, la procédure est un peu différente et plus longue mais assez claire.



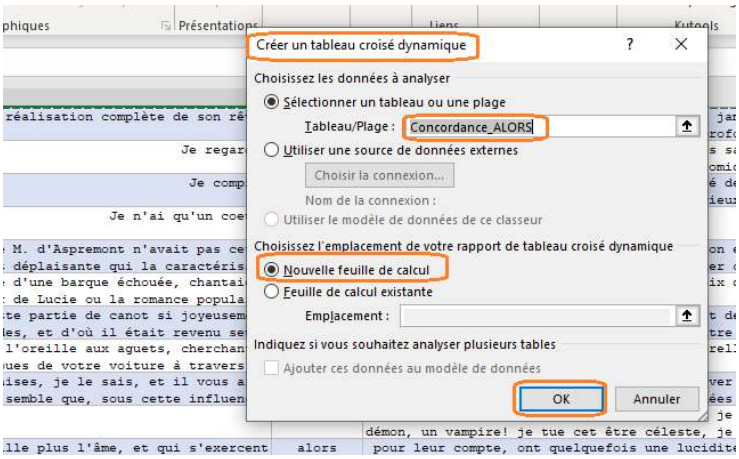
Le tableau change d'allure et deux nouveaux menus, ***Power Pivot*** et ***Conception de la table*** sont créés donnant à l'utilisateur la possibilité de changer les couleurs proposées. Il est également recommandé de donner au tableau un nom dans le haut à gauche de la fenêtre. La ligne des titres des colonnes est munie d'un système de filtrage grâce aux petits triangles noirs à droite de chaque titre. Le tableau est désormais prêt.



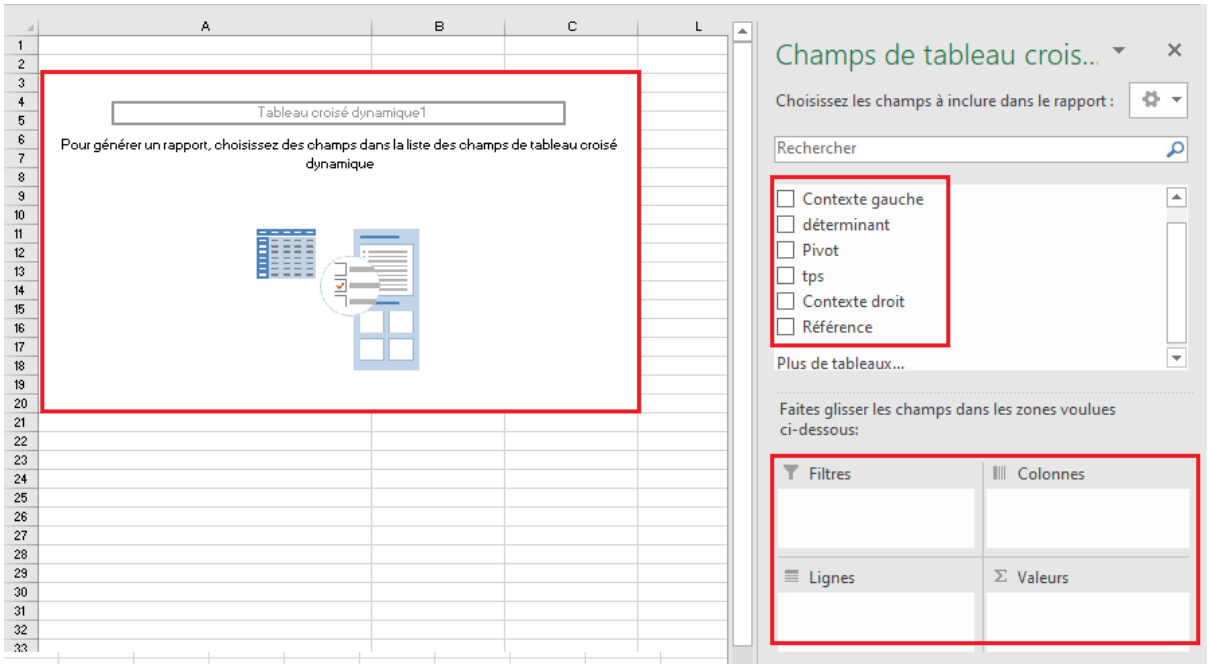
[69] Eh oui, ***Excel*** ne sait pas encore qu'il s'agit d'un tableau !

[70] Si on respecte cette consigne, toutes les modifications, ajouts ou suppressions sont systématiquement pris en charge.

De retour au menu ***Insertion***, un clic sur l'icône ***Tableau croisé dynamique*** fait apparaitre une fenêtre de dialogue qui signale qu'un Tableau croisé dynamique (désormais **TCD**) va être créé et rappelle le nom donné au tableau. Une option est proposée pour indiquer au logiciel s'il crée le TCD dans une nouvelle feuille de calcul vierge ou dans une autre feuille déjà préparée à cet effet. Le premier choix est le plus facile pour un débutant.



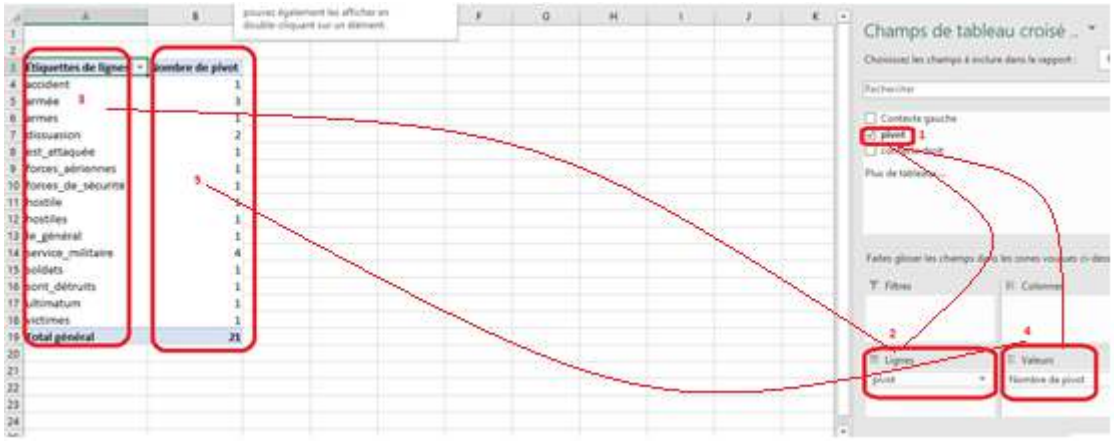
Une fois le choix validé, une nouvelle feuille est créée dans le même classeur et une fenêtre présente, à droite, sous ***Champs de tableau croisé***, sous forme de cases à cocher, les titres des colonnes du tableau créé dans ***Excel*** à la première étape et quatre zones ou rubriques : ***Filtres***, ***Colonnes***, ***Lignes*** et ***Valeurs***. A gauche vont se placer les résultats des croisements à effectuer.



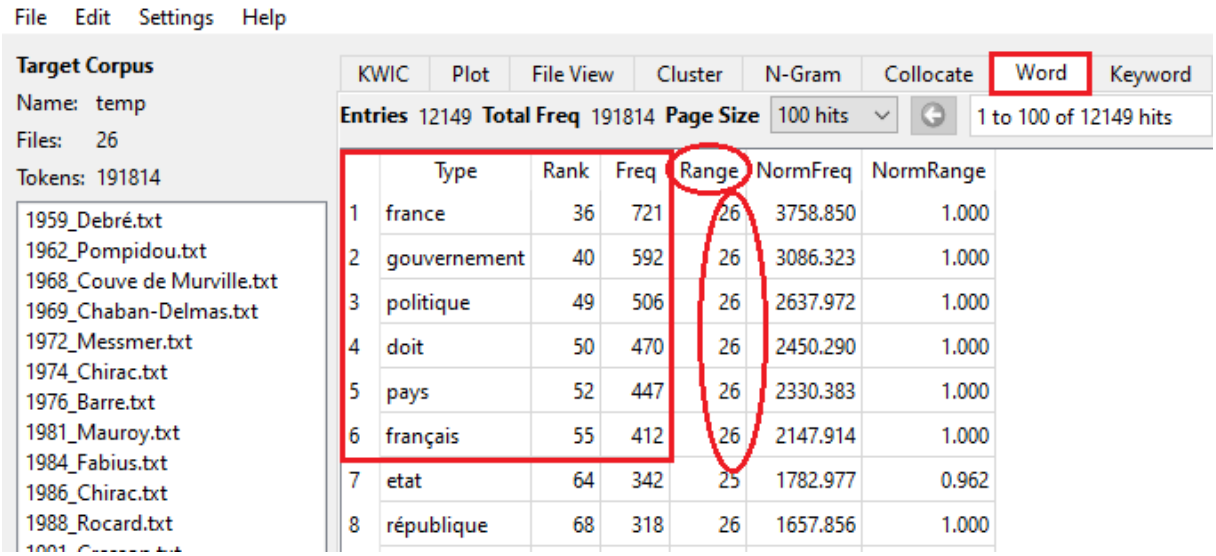
En cliquant dans la case correspondant à n'importe quelle colonne du tableau de la base de données dans notre exemple (1), ***Excel*** place immédiatement le nom de la case dans la rubrique ***Lignes*** (2) et affiche à gauche le contenu de la colonne du tableau filtré alphabétiquement (3).

Avec la souris on clique de nouveau sur la même case et sans lâcher (cliquer-déplacer) on déplace la souris vers la rubrique ***Valeurs*** (4), les statistiques des mots de la colonne apparaissent en face des mots (5).

Puisqu'il s'agit d'un **Tableau *croisé*** (et) ***dynamique***, il est facile de changer le ou les choix en décochant les cases cochées et de refaire un autre choix si les résultats qui s'affichent ne sont pas satisfaisants ou ne sont pas exploitables selon la perspective de la recherche en cours.



Dans l'exemple pratique suivant, l'analyse consiste à comparer, chez les 26 premiers ministres français de 1959 à 2022, lors de leurs discours de politique générale, les emplois des tirois de conjugaison avec le mot « gouvernement » quand il est sujet de la phrase. Le choix a porté sur le mot *gouvernement* car il est, selon l'index général des vingt-six discours obtenu avec l'option ***Hide words in file*** de l'outil ***Word*** d'***AntConc v4.2.0***[71], le deuxième mot plein après *France*, et avant *politique*, *pays* et *français*, tous communs à la totalité des fichiers (***Range*** 26).



On demande, d'abord, de considérer les locuteurs comme base du filtrage en cliquant sur la case du champ *Locuteur* (1) qui contient les noms des locuteurs, la mention

[71] Voir § 2.7.

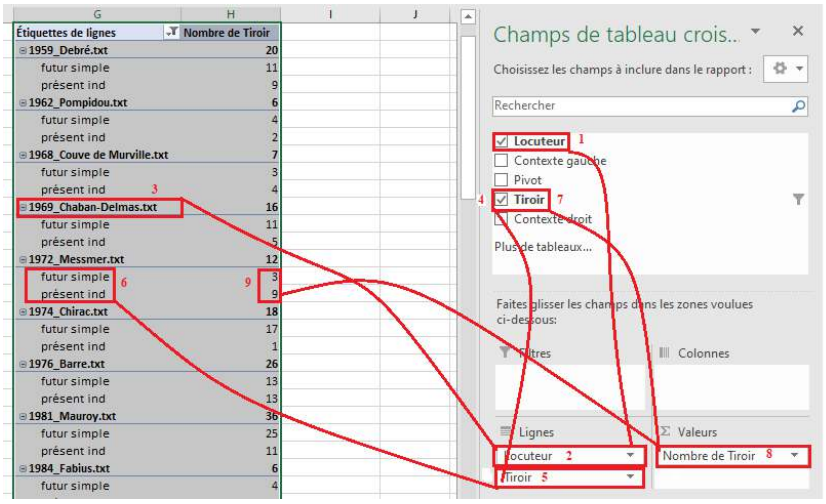
Locuteur (2) est automatiquement placée dans la rubrique *Ligne* et immédiatement la liste des différents locuteurs apparaît à gauche de l'écran et qui correspond aux noms des premiers ministres (3, *Etiquettes de lignes*).

Dans un deuxième temps, on voudrait connaître pour chaque locuteur les deux tiroirs de conjugaison les plus utilisés avec les verbes dont le sujet est le nom *gouvernement*. On répète l'étape précédente en cliquant sur la case ***tiroirs*** (4), ce qui place la mention ***tiroirs*** dans la rubrique ***Lignes*** (5) et automatiquement le tableau de gauche (6) affiche les tiroirs utilisés par chaque locuteur. Pour les occurrences où le nom *gouvernement* n'est pas sujet d'un verbe, le tableau affiche *vide*.[72]

Étiquettes de lignes	Nombre de Tiroir
1959_Debré.txt	21
futur simple	11
passé composé	1
présent ind	9
(vide)	0
1962_Pompidou.txt	9
conditionnel pr	1
futur simple	4
passé composé	2
présent ind	2
(vide)	0

Pour connaître maintenant les statistiques des emplois des tiroirs de conjugaison, on sélectionne de nouveau la case *Tiroir* (7)[73] et, avec un copier-glisser, on la déplace vers la rubrique *Valeurs* où s'affiche maintenant *Nombre de Tiroirs* (8). Le tableau s'actualise avec les données chiffrées (9).

La capture suivante récapitule l'ensemble des opérations effectuées.



[72] Pour supprimer le résultat des cellules vides, suivre la procédure expliquée dans « Comment masquer les lignes vides dans le tableau croisé dynamique dans Excel ? » à l'adresse internet https://fr.extendoffice.com/documents/excel/2361-excel-pivot-table-hide-blank-rows.html#a1

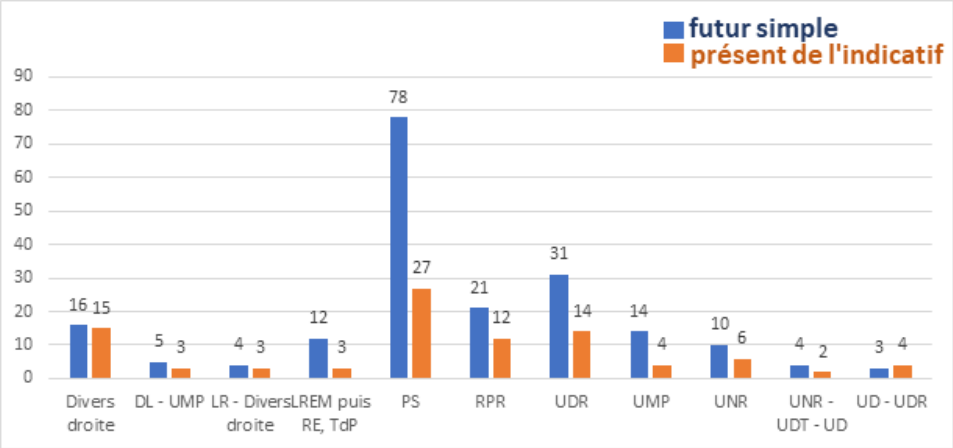
[73] Le chois de la case *Pivot* donnera le même résultat.

Nous avons ajouté aux données les obédiences politiques des premiers minstres. Les graphiques suivants donnent à voir la répartition des tiroirs de conjugaison selon les partis politiques des locuteurs.

1ers ministres	Président	Politique	futur	présent
1959_Debré	de Gaulle	UNR	10	6
1962_Pompidou	de Gaulle	UNR	4	2
1968_Couve de Murville	de Gaulle	UNR - UDT - UD	3	4
1969_Chaban-Delmas	Pompidou	UDR	11	5
1972_Messmer	Pompidou	UDR	3	8
1974_Chirac	Giscard d'Estaing	UDR	17	1
1976_Barre	Giscard d'Estaing	Divers droite	13	12
1981_Mauroy	Mitterand	PS	25	10
1984_Fabius	Mitterand	PS	3	1
1986_Chirac	Mitterand	RPR	11	7
1988_Rocard	Mitterand	PS	4	**1**
1991_Cresson	Mitterand	PS	1	1
1992_Beregovoy	Mitterand	PS	4	3
1993_Balladur	Mitterand	RPR	1	1
1995_Juppé	Chirac	RPR	9	4
1997_Jospin	Chirac	PS	15	8
2002_Raffarin	Chirac	DL - UMP	5	3
2005_Villepin	Chirac	UMP	4	
2007_Fillon	Sarkozy	UMP	8	3
2010_Fillon	Sarkozy	UMP	2	1
2012_Ayrault	Holland	PS	13	1
2014_Vals	Holland	PS	3	1
2016_Caseneuve	Holland	PS	10	1
2017_Philippe	Macron	LR - Divers droite	4	3
2020_Castex	Macron	Divers droite	3	3
2022_Borne	Macron	LREM puis RE, TdP	12	3

Les résultats provisoires indiquent que ce sont les membres du Parti socialiste qui utilisent le tiroir du *futur simple* (78 occurrences)[74] avec le sujet *gouvernement* plus que le double de l'Union des démocrates pour la Cinquième République (UDR) resque tous les autres réunis (109 occurrences). Pour le *présent de l'indicatif*, ce sont également les socialistes qui sont en tête du classement.

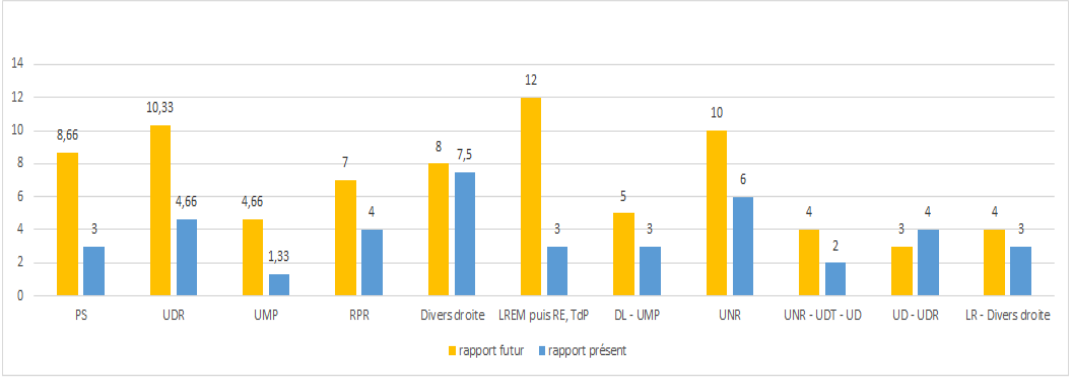
[74] Nous avons compté comme aynt la valeur temporelle du futur les verbes conjugués au futur et les semi-auxiliaires *aller*, *entendre*, conjugués au présent.



Cependant l'analyse serait faussée si nous ne prenions pas en considération le fait que le Parti socialiste compte à lui seul neuf (9) premiers ministres, autant que l'UDR, l'UMP et le RPR réunis. Le rapport du nombres des occurrences du futur et du présent change si les locuteurs sont comparés entre eux, ce qui donne les statistiques suivantes :

Parti Politique	Nmbr de locuteurs	Occurrences futur	Occurrences présent	rapport futur	rapport présent
PS	9	78	27	8,66	3
UDR	3	31	14	10,33	4,66
UMP	3	14	4	4,66	1,33
RPR	3	21	12	7	4
Divers droite	2	16	15	8	7,5
LREM puis RE, TdP	1	12	3	12	3
DL - UMP	1	5	3	5	3
UNR	1	10	6	10	6
UNR - UDT - UD	1	4	2	4	2
UD - UDR	1	3	4	3	4
LR - Divers droite	1	4	3	4	3

Le graphique suivant fait apparaître que réellement c'est Elisabeth Borne de *La République En Marche* qui utilise le plus le *futur* avec le nom *gouvernement* (12 occurrences)



Bibliographie

Bednarek, M. et Carr, G. (2021). « Analyse de texte numérique assistée par ordinateur pour la recherche en journalisme et en communication : introduction de techniques linguistiques de corpus qui ne nécessitent pas de programmation ». Médias internationaux Australie, Incorporating Culture & Policy, Volume 181 (1), Publications SAGE, p. 131-151.

Bilger, M., Debaisieux J.-M., Deulofeu J. et Sabio F. (2013). « Analyses linguistiques sur corpus : le cadre descriptif ». Analyses linguistiques sur corpus. Subordination et insubordination en français. D. Jeanne-Marie, Lavoisier, Hermes Sciences : p. 61-98.

Bourion, E. (2001). *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat, Sciences du langage, Université de Nancy II.

Brunet, É. (2003). « Lexicométrie et étude du vocabulaire ». *A la recherche des Illusions perdues*, edité par P. Hubert de Nizet.

Équipe DELIC (2004). « Présentation du Corpus de référence du français parlé », in P. Cappeau (éd.), Autour du corpus de référence du français parlé, Recherches sur le français parlé 18, Université de Provence. p. 11-42.

François, J. et Ghérissi, Y. (2012). *Pour une linguistique orientée outils. La polysémie du verbe compter et les genres textuels*. Cahiers du CRISCO 34, 10 mai 2012.

Gasiglia, N. (2004). « Faire coopérer deux concordanciers-analyseurs pour optimiser les extractions en corpus », in B. Habert (dir.), Linguistique et informatique : nouveaux défis, Revue Française de Linguistique Appliquée, volume IX – 1, p. 45-62.

Gasiglia, N. (2005). « Stratégie de constitution de corpus oraux transcrits : arguments pour un corpus plurithématique à haut rendement », in G. Williams (éd.), *La linguistique de corpus en France ou en français*. Presses Universitaires de Rennes.

Gendner, V. Adda-Decker, M. (2002). « Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques ». XXIVèmes Journées d'Etude sur la Parole, Nancy, 24-27 juin 2002.

Habert, B., Fabre, C. & Issac, F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. Paris : InterEdition.

Habert, B., Nazarenko, A. et Salem, A. (1997). *Les linguistiques de corpus*, Paris, Armand Colin.

Henkel, D. (2016). *Initiation à la linguistique de corpus : Créer un corpus étiqueté en français, anglais* (*ou d'autres langues*). Séminaire de recherche CeLiSo (EA7332) du 25/03/2016. Université Paris IV-Sorbonne.

Legallois, D. -2015). *Didacticiel Antconc. Présentation du logiciel AntConc*. Cours complet : https://uoh.fr/front/noticefr/?uuid=6e4ddf71-f934-4c37-8ac3-16d1e01f717f

Prunet, A. (2018). *Les littéracies en Français sur Objectifs Universitaires. Etude d'un corpus contrastif de productions écrites argumentées et perspectives didactiques*, thèse de doctorat, Universite Sorbonne Nouvelle Paris 3.

Prunet, A. (2019). « L'acculturation aux littéracies universitaires : exemple de l'apprentissage et enseignement de l'emploi de "on" à l'aide du logiciel AntConc », Alsic [En ligne], Vol. 22, n° 1 | 2019, mis en ligne le 27 décembre 2019, consulté le 16 avril 2023. URL : http://journals.openedition.org/alsic/4118

Roiné, Ph., Blasco, M. et Auriac-Slusarczyk, E. (2021). « Rôles et valeurs des emplois en « c'est » dans le corpus Philosophèmes In : Des corpus numériques à l'analyse linguistique en langues de spécialité ». [en ligne]. Grenoble : UGA Éditions, 2021 (généré le 16 avril 2023). Disponible sur Internet : http://books.openedition.org/ugaeditions/24285

Silberztein, M., Poibeau, Th. & Balvet, A. (2001). « Intex et ses applications informatiques ». Tutoriel, Actes de la huitième conférence TALN, 2-5 juillet 2001, volume II. p. 145-174.

Tigziri, N. (2016). « Analyse textuelle à l'aide du concordancier ANTCONC d'une oeuvre de Belaïd Ait-Ali ». Iles d Imesli, 8(1), pp. 163-172. Disponible en ligne : https://www.asjp.cerist.dz/en/article/34838

Tintin, A. (2022). *S'appuyer sur les logiciels d'analyse textuelle AntConc et TreeTagger afin de jauger la qualité d'un écrit réflexif - une étude de faisabilité*. (Unpublished master's thesis). Université de Liège, Liège, Belgique. Disponible sur Internet : https://matheo.uliege.be/handle/2268.2/15938

Vuillemin, A. (1990). *Informatique et littérature* (*1950-1990*), Genève-Paris, Slatkine-Champion.

Williams, G. (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. Thèse de linguistique soutenue à Nantes.

Table des matières