Guide d'utilisation du logiciel Lexico3 appliqué à la langue arabe

Catherine Pinon

GUIDE D'UTILISATION DE LEXICO3

Application à la langue arabe

Catherine PINON

Ce guide s'adresse aux utilisateurs débutants qui n'ont jamais travaillé sur un logiciel de textométrie. Il s'inspire fortement du manuel d'utilisation de Lexico3.

Le logiciel Lexico3 ainsi que la documentation associée (corpus d'essai, tutoriels, manuels) sont disponibles à cette adresse : http://www.lexi-co.com/

On conseille la lecture d'ouvrages portant sur la statistique textuelle, notamment celui de L. Lebart et A. Salem [En ligne] http://ses.telecom-paristech.fr/lebart/ST.html

Pour ce qui est de la constitution des corpus et des considérations techniques associées, nous renvoyons à B. Habert, C. Fabre et F. Issac : *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, InterEditions, Masson, Paris, 1998, 320 p.

Il existe beaucoup d'autres références et ressources en ligne pour s'initier à la linguistique de corpus.

Les exemples de ce manuel sont principalement inspirés de recherches en syntaxe, mais tous les chercheurs qui travaillent sur des sources arabes peuvent être intéressés, quelle que soit leur discipline. Le but de ce manuel est de donner les clés pour poursuivre seul la réflexion en fonction de l'objectif de ses recherches.

Lorsqu'on recourt à la transcription, on recourt au système *arabica*.

Table des matières

Introduction : qu'est-ce que Lexico3 ?

Lexico3 est un logiciel qui a plusieurs fonctionnalités. Il permet :

- d'obtenir la **liste des formes** présentes dans un document (nombre de formes, *hapax* , fréquence des mots, *etc*.),
- d'établir le **dictionnaire** de ces formes (classement lexicographique, i.e. par ordre alphabétique, ou lexicométrique, i.e. par fréquence décroissante),
- d'afficher la **concordance** d'un mot précis (en offrant la possibilité de trier le contexte à gauche ou à droite par ordre alphabétique, ce qui permet de trouver des expressions ou autres collocations),
- de rechercher des **segments répétés** dans le texte (suite de deux mots ou plus qui se répète au moins deux fois),
- d'établir des **partitions** du texte (en fonction des auteurs, de la chronologie, de la source, du genre, de la pagination originale, *etc*., ce qui permet d'effectuer par la suite des comparaisons entre parties).
- de déterminer des **sections** à travers lesquelles il est possible de naviguer
- d'effectuer différents **calculs statistiques**.

Toutes les manipulations sont consignées dans un fichier généré automatiquement et tous les résultats peuvent être consignés dans un rapport disponible au format htm.

Les personnes qui découvrent ce type de logiciel peuvent être momentanément désorientées, c'est pourquoi nous leur conseillons de suivre pas à pas ce guide. Ce dernier a été élaboré pour permettre à l'utilisateur de *comprendre* le fonctionnement du logiciel et les différentes manipulations qu'il permet de réaliser. On se placera du point de vue pratique en fournissant des exemples généraux puis appliqués à la langue arabe. Il apparaît indispensable d'avoir un objectif clairement défini (une hypothèse de recherche à vérifier, un calcul à effectuer, *etc*.) avant de se lancer dans l'utilisation du logiciel. Ce guide vise aussi à montrer les possibilités offertes par le logiciel pour aller plus loin dans la recherche lexicométrique.

I. Préparer la ressource textuelle à la segmentation

Un corpus est constitué de plusieurs éléments, dont les deux principaux qu'il convient de différencier sont :

1. La matière textuelle du corpus, que l'on nomme ressource textuelle , sur laquelle porte la recherche.

2. Les informations recueillies sur cette matière, ou méta-information (date de production, genre, auteur, pagination, sujet, *etc*.), qui vont permettre d'établir des champs de comparaison (chronologiques, génériques, stylistiques, *etc*.).

Plusieurs opérations sont nécessaires avant de commencer à travailler statistiquement sur le texte. Elles consistent à définir les normes permettant au texte d'être segmenté en formes (unités textuelles) qui deviennent des objets manipulables par la statistique. Ainsi, selon les normes de segmentation programmées, on peut travailler sur le "mot" (graphique ou lexical), le "morphème" (affixes par exemples), le radical ou la racine, *etc*.

Cette étape doit être faite avec une grande attention, car la précision des informations codées permet d'améliorer les résultats du logiciel. Certaines règles visant à déjouer les problèmes dus au codage doivent alors être définies. En fonction de la nature du corpus, mais aussi des objectifs de la recherche, elles seront redéfinies et réajustées.

I. 1. Le statut des signes

L'opération de **segmentation** est donc l'opération fondamentale : au cours de la segmentation, le texte est transformé en unités manipulables, sur lesquelles tout calcul statistique peut être effectué. En quelques sortes, la machine "traduit" le texte dans un code qu'elle comprend et qu'elle peut ensuite manipuler à sa guise pour effectuer toutes les sortes de calcul imaginables.

Un bref rappel sur la technique permettant au texte d'être transformé en donnée statistiquement mesurable aidera à pénétrer la logique élémentaire de la lexicométrie . Cette logique est très simple : pour permettre tout calcul sur les mots, il faut marquer chaque mot d'une étiquette permettant à la fois de l'identifier par rapport aux autres (ce mot est-il unique dans le texte ? Sinon, combien de fois apparaît-il ?) et de le localiser dans le texte (où le mot apparaît-il dans le texte ? Ceci permet d'établir une topographie textuelle).

Il faut d'abord que la machine puisse séparer les mots les uns des autres : pour cela, il suffit simplement de ranger les caractères dans deux catégories différentes. Soit un caractère est délimiteur (il sépare les mots les uns des autres), c'est le cas de l'espace et plus généralement des signes de ponctuation, soit il ne l'est pas. En conséquence, une suite de caractères non délimiteurs contenue entre deux caractères délimiteurs est une forme textuelle analysable. Nous reparlerons de ces caractères délimiteurs/non délimiteurs plus loin, mais bien évidemment, il suffit simplement de fournir au logiciel la liste des caractères délimiteurs, beaucoup moins nombreux que les caractères non délimiteurs.

Pour prendre un exemple concret, si vous rentrez dans la machine la phrase (en ayant au préalable déterminé l'espace, rendu ici par le tiret bas pour plus de clarté, et le point comme des caractères délimiteurs) :

Je _ ne _ sais _ pas _ pourquoi _ il _ ne _ mange _ pas.

Vous voyez que les caractères délimiteurs (ici en vert) permettent de découper le texte en "mots" (formes graphiques plus justement).

Ensuite, la machine va assigner à chaque forme originale un numéro, dans l'ordre d'apparition :

Je	_	*ne*	_	*sais*	_	*pas*	_	*pourquoi*	_	*il*	_	*ne*	_	*mange*	_	*pas*	.
1		2		3		4		5		6		2		7		4	

Les mots "ne" et "pas" figurent deux fois dans le texte, c'est pourquoi le numéro qui leur est assigné est répété. C'est ce passage de la forme graphique à la forme numérique qui va permettre d'effectuer des calculs de fréquence et de repérer la répartition des formes à l'intérieur du texte. Dans cet exemple, le mot n°1 = "je" n'apparaît qu'une fois dans le texte (dans ce cas, on parle d'*hapax*), le mot *n*°2 = "ne" apparaît deux fois dans le texte, *etc*.

Ainsi, l'opération de segmentation se voit répondre à un principe extrêmement simple : chaque signe typographique reçoit un statut (délimiteur ou non délimiteur) qui lui est assigné au tout début du processus. On fournit à la machine la liste des délimiteurs uniquement. Tous les autres caractères sont considérés comme non-délimiteurs.

Concrètement, voilà donc ce qu'il faut faire :

▶ Il faut **affecter à chaque signe du texte un statut en établissant la liste des délimiteurs**

Attention, cependant :

💣 Le choix des caractères délimiteurs doit être mûri, car ces derniers n'ont pas toujours qu'une seule valeur. Par exemple, le trait d'union peut marquer un mot composé (*après-midi*) ou une liaison (*dit-il*). Dans le premier cas, il faudrait que le caractère soit non délimiteur, pour que la suite *après-midi* soit considérée comme une occurrence ; dans le second cas il faudrait qu'il soit délimiteur pour obtenir d'une part le verbe *dit* et d'autre part le pronom *il*. On peut rencontrer ce même problème avec l'apostrophe.

💣 Il faut **supprimer les majuscules**, sinon un même mot présent dans le texte une fois avec une minuscule et une fois avec une majuscule serait traité comme deux formes différentes par le logiciel. Mais cette manipulation peut poser un problème pour les noms qui sont à la fois

propres et communs. Il faut alors désambigüiser les formes par un signe qui doit figurer dans la liste des délimiteurs (un tildé, un astérisque, *etc*.) collé au mot sans espace (avant ou éventuellement après pour plus de lisibilité). L'arabe étant dépourvu de majuscules, le problème ne va pas se poser de la même manière. Si nous voulons marquer les noms propres d'un texte arabe, il suffit de leur assigner un caractère non délimiteur. Ceci permettra par exemple d'isoler les adjectifs ou les verbes des prénoms : *farīd* (adj.) / Farīd (nom propre) ; *yazīd* (verbe) / Yazīd (nom propre).

💣 Avec les textes arabes, les caractères délimiteurs sont plus nombreux, car on trouve souvent dans les textes des signes de ponctuation empruntés à différentes langues et dans deux sens de lecture (c'est le cas des points d'interrogation par exemple : ? ou ؟).

Bien évidemment, le choix des caractères délimiteurs et, plus généralement, la préparation du texte à l'opération de segmentation présuppose d'avoir une idée des cas ambigus qui vont se présenter pour les traiter préalablement. Il est judicieux d'établir et de tenir une liste des cas rencontrés et des traitements proposés. Après une première segmentation, il faut observer la liste des formes fournies par le dictionnaire pour détecter les aberrations et les éventuels caractères délimiteurs oubliés. Il est souvent nécessaire de procéder à plusieurs segmentations, jusqu'à ce que le rendu soit acceptable.

Pour l'arabe, voici quelques précautions à prendre :

☞ La nature concaténative de la graphie arabe va poser problème. Comment doit-on traiter les mots graphiques qui contiennent plusieurs unités ? Doit-on systématiquement ajouter un délimiteur entre chaque unité (par exemple entre les particules monolitères, l'article ou les pronoms suffixes et le nom auquel ils sont graphiquement collés) ? Tout dépendra bien entendu de l'objectif visé par la recherche, mais nous pouvons déjà dire que l'outil "groupe de formes" va permettre de résoudre certains problèmes. On en fera une présentation plus précise par la suite, mais pour résumer, il s'agit d'un outil recherchant dans le texte un segment déterminé (une suite de caractères, ou une suite comprenant certains caractères) et permettant d'associer comme une seule forme différentes formes.

☞ La vocalisation sera aussi un facteur d'ambigüité, car il n'est pas envisageable de vocaliser entièrement un texte : non seulement ce serait trop long, mais en plus faudrait-il le faire très scrupuleusement[1]. En aucun cas le traitement de la vocalisation peut être systématique : il est nécessaire d'envisager les différents problèmes qui vont se poser en fonction des objectifs de la recherche, pour les devancer (par exemple, comment différencier le prénom *ʿalī* de la préposition suivie d'un suffixe *ʿalayya*, si ce n'est en

[1] D. Kouloughli avait recourt au stratagème suivant pour contourner ce problème : ayant constaté que les textes en arabe, lorsqu'ils sont vocalisés, génèrent de nombreux problèmes parce qu'ils sont mal vocalisés ou que la machine inverse l'ordre entre la voyelle et son support, il utilisait un système de transcription en caractère latin, qu'il avait mis au point lui-même, et vocalisait manuellement le texte. Il pouvait ensuite le repasser en caractères arabes. Il existe plusieurs systèmes sembables, comme celui de Buckwalter (https://en.wikipedia.org/wiki/Buckwalter_transliteration).

marquant bien la *šadda* dans le second cas ?). Il existe en arabe de nombreux mots ou groupes de mots qui ont la même graphie sans avoir le même sens ni recevoir la même analyse. Il convient d'en dresser la liste à mesure que l'on progresse.

🏳 Pour une recherche portant sur l'ordre des mots dans la phrase, on peut imaginer enlever systématiquement tout signe de lecture. Mais pour toute étude touchant au lexique, supprimer les signes de lectures génèrera trop de problèmes : quel statut donner à la *šadda* ? Doit-on l'ajouter systématiquement ? La présence de cette marque de gémination désambigüiserait de nombreux cas, mais elle est rarement notée et la rajouter prendrait aussi beaucoup de temps. On peut aussi imaginer un *motus vivendi* concernant l'utilisation des voyelles, de manière à affiner la distinction de formes graphiques semblables ayant des analyses différentes (par exemple, mentionner tous les passifs par la voyelle *u*, *etc*.).

🏳 Il ne faut pas oublier non plus les problèmes dus à une graphie "fautive" : les *yāʾ* finaux auxquels il manque souvent les points, ce qui pour un mot comme *fī* donnerait deux formes (*fī* et *fā*) et donc fausserait les calculs statistiques ; pour certains mots cela génère même une confusion (par exemple, entre *ʿalā* la préposition et *ʿalī* le prénom).

➲ ***Chaque chercheur, en fonction de ses objectifs, doit élaborer ses propres règles. L'important étant de les respecter scrupuleusement.***

I. 2. Le balisage du corpus

L'intérêt de Lexico3 est de pouvoir faire des comparaisons entre différentes partitions d'un même corpus. On peut donc imaginer comparer des discours appartenant à des époques, des genres ou des auteurs différents. Il faut donc intégrer dans le corpus ce que nous avons appelé plus haut "méta-information" : ainsi, des renseignements pourront être pris en compte par la machine sans être traités comme du texte à part entière. Pour cela, il faut les intégrer sous forme de balises , ou encore de clés .

Il faut introduire à ce stade (avant de lancer la segmentation du texte par le logiciel) les différents délimiteurs du corpus. Le choix de ces délimiteurs permet de rassembler en une seule base textuelle exploitable différents sous-corpus, ou tout simplement de mentionner différentes informations telles que la date, l'auteur, le chapitre, la pagination d'origine, le genre du texte, *etc.* Ces méta-informations sont repérables par le logiciel et compréhensibles grâce aux clés qui fonctionnent comme des balises informatives codées. Elles doivent figurer dans le fichier sans pour autant influer sur les comptages statistiques, i.e. sans être prises pour de la donnée textuelle par le logiciel. C'est pourquoi on les introduit sous forme de clés isolées du corpus par les chevrons.

Les clés, qui peuvent être alphanumériques, ont la forme suivante : ***<type=contenu>.*** Chaque ***type*** particulier de balise (partie située avant le signe « = ») permet de définir une partition du corpus. Pour un type fixé, si on ignore tous les autres types, les différents *contenus* (partie située après le signe « = ») correspondent à autant de parties différentes dans le corpus. On peut donc imaginer délimiter ainsi les différents articles d'un journal (par des clés du type : <article=1>, <article=56>, *etc*.) ou associer des textes de différents auteurs (<auteur=x>, <auteur=y>, *etc*.).

I. 3. La délimitation des zones textuelles

Pour pouvoir s'appuyer sur une division du texte en phrases, en paragraphes, *etc*., i.e. en parties "naturelles" du texte, il est possible d'assigner à un caractère délimiteur le rôle de ***délimiteur de section***. Le fait de coder ainsi le découpage initial du texte permet ensuite d'étudier la répartition des formes à l'intérieur des sections ainsi constituées (par exemple, on peut aisément repérer que tel connecteur s'utilise systématiquement en début de phrase).

Pour prendre l'exemple d'une délimitation de zones textuelles en paragraphe, il suffit de coder chaque paragraphe en ajoutant en début de paragraphe un caractère réservé à cet effet et déclaré dans la liste des délimiteurs (par exemple, §). Ce remplacement peut être effectué de manière automatique grâce à un logiciel de traitement de texte : on remplace le caractère "retour-chariot" par la séquence "retour-chariot" suivi de §. Avec le logiciel Word, par exemple, on cherche ^p et on le remplace par ^p §.

En vue d'établir des cartes des sections, il convient d'insérer des caractères délimiteurs de section dans le fichier texte d'origine.

I. 4. Enregistrement du texte source

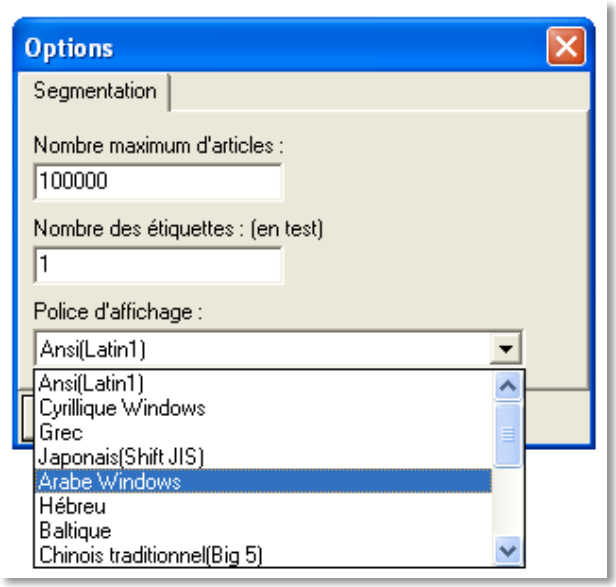
▶ Il faut enregistrer la matière textuelle balisée en texte seul (extension .txt).

▶ L'idéal est de placer ce fichier dans un dossier particulier, car le logiciel génère des fichiers qu'il enregistre directement dans le même dossier de provenance que celui du texte segmenté (chargé au départ pour l'analyse).

II. Du fichier au logiciel : la segmentation du texte

Il est indispensable de sélectionner l'arabe comme police d'affichage **avant** de charger un texte en caractères arabes.

Options > Police d'affichage > Arabe Windows



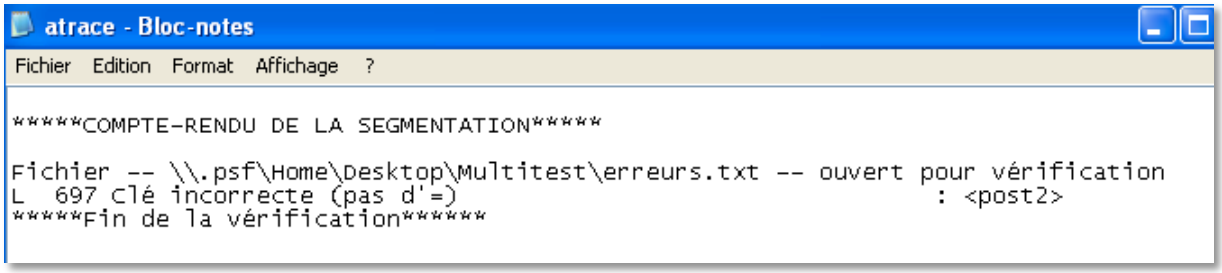
Attention : dans la version actuelle du logiciel, il n'est pas possible de changer la langue par défaut. Il faut donc prendre l'habitude de sélectionner l'arabe dès l'ouverture du logiciel.

II. 1. La segmentation

Cliquer sur l'icône *nouvelle base (segmentation)*, la première icône à gauche : 

Sélectionner un fichier **.txt**

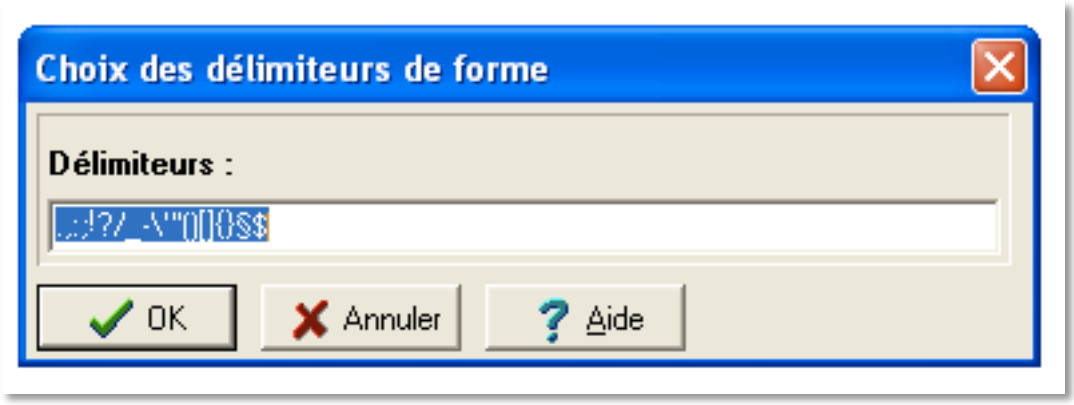
Avant de segmenter, la machine va vérifier les clés. En cas d'erreur dans la saisie des clés, un message apparaîtra :



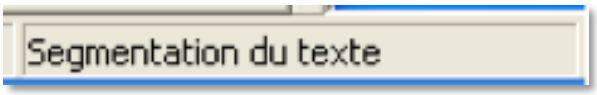
À l'affichage de ce message, cliquez sur *ok*, puis ouvrez le fichier **atrace.txt**, généré automatiquement par le logiciel (il se trouve dans le même dossier que celui d'où provient le fichier .txt que vous avez donné à segmenter). Le fichier atrace.txt est le fichier où sont mentionnées toutes les actions lancées par le logiciel pour une source donnée.

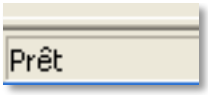
Si des erreurs figurent, il faut les corriger puis relancer la segmentation. 

Si toutes les erreurs ont été corrigées, une fenêtre de dialogue apparaîtra pour vous permettre de choisir les caractères délimiteurs :



Par défaut, une liste figure déjà. Choisir les caractères délimiteurs en modifiant cette liste si besoin est. Il n'est pas encore possible de modifier la liste des délimiteurs par défauts. Il peut être judicieux d'en établir la liste pour l'arabe et de la conserver pour simplement la copier-coller dans la fenêtre des délimiteurs.

Pendant la segmentation, vous pourrez lire *segmentation du texte* dans la barre inférieure du logiciel. 

Quand la segmentation est achevée, vous lirez au même endroit : 

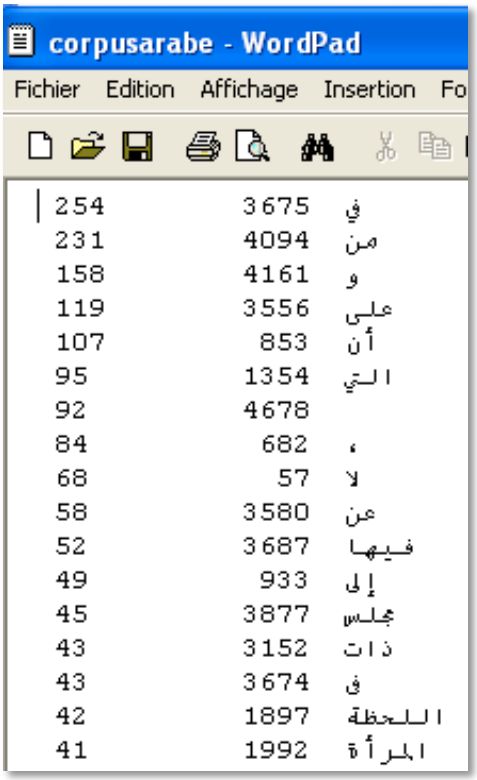
⌛ Attention, en fonction de la taille du fichier, la segmentation peut prendre du temps.

III. Génération de fichiers : paramètres, trace de l'utilisation et dictionnaires

À l'issue de la segmentation, le logiciel va créer différents fichiers qui seront automatiquement enregistrés sur le même site que le fichier .txt d'origine.

1. un dictionnaire (nomducorpus.dic) : il s'agit de toutes les formes contenues dans le document, classées par ordre de fréquence décroissante, et pour une même fréquence, par ordre alphabétique. Le dictionnaire classe les formes graphiques, la ponctuation, les types de clés, et les contenus de clés. Pour consulter ce fichier, un simple éditeur de texte suffit.

🔔 si votre PC est entièrement arabisé, les fréquences et rangs seront donnés en chiffres indiens.

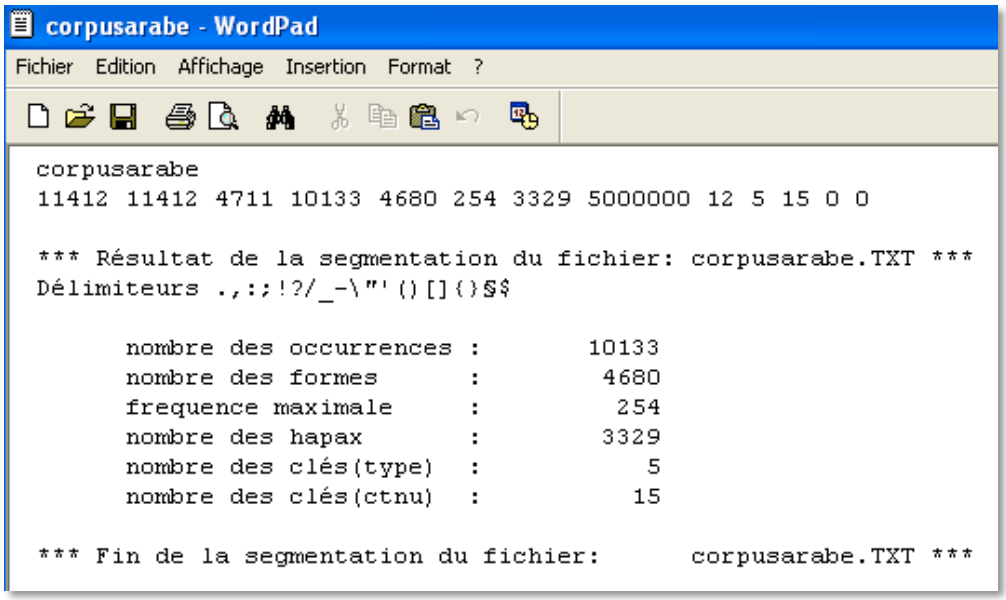


La colonne de gauche indique la fréquence : dans ce corpus, le mot *fī* est celui qui apparaît le plus avec 254 apparitions, suivi de *min* (231 fois) et de *wa* (158 fois). La colonne du milieu indique le rang lexicographique du mot en question, c'est-à-dire le numéro qui lui a été assigné dans la liste des formes classées par ordre alphabétique.

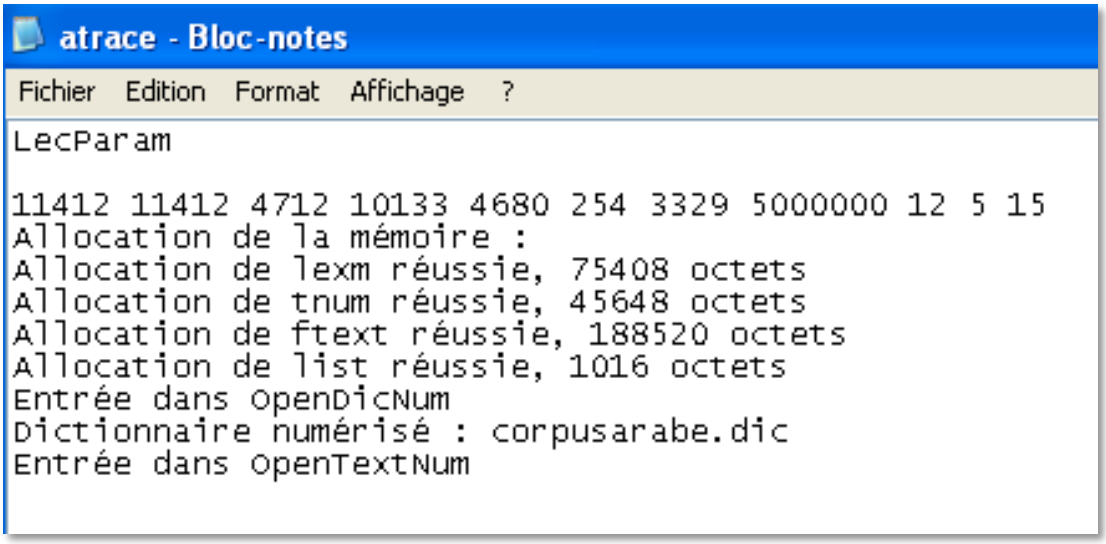
Attention : l'ordre alphabétique arabe produit par Lexico3 ne correspond pas à l'ordre alphabétique réel. Cette lacune sera comblée dans une prochaine version.

2. un fichier de paramètres (nomducorpus.par) : y sont consignés les paramètres utilisés pour ce fichier et les décomptes élémentaires des formes. On y trouve donc le rappel des caractères délimiteurs choisis lors de la segmentation, le nombre des occurrences, le nombre des formes, la fréquence maximale, le nombre d'*hapax* (une seule occurrence d'une forme donnée), le nombre de types de clés et le nombre de contenus de clés.

Le fichier paramètre s'ouvre aussi sur un quelconque éditeur de texte :



3. un fichier *atrace.txt* : il contient un rapport détaillé des opérations effectuées par le programme (mémoire allouée, paramètres pris en compte, fichiers lus et écrits, *etc*.). Il témoigne des opérations effectuées par l'utilisateur sur le fichier d'origine. Ainsi, en cas de difficulté, on peut le consulter pour trouver la cause du problème.



4. un fichier .num (nomducorpus.num) : c'est un fichier interne au logiciel, auquel on ne peut accéder avec un simple éditeur de texte. Il contient le texte numérisé, c'est-à-dire que tous les éléments du corpus que l'on retrouve dans le dictionnaire ou dans le fichier paramètres (occurrences, formes, ponctuation, clés, *etc*.) sous une forme codée de façon compacte.

Ouvrir une base déjà segmentée

Une fois constituée, une base peut être réutilisée pour différentes recherches. Il est inutile de recommencer tout le travail de segmentation à chaque fois (sauf bien évidemment si l'on veut modifier certains critères) : il suffit d'ouvrir une base déjà créée grâce à l'icône *Ouverture d'une base*. 

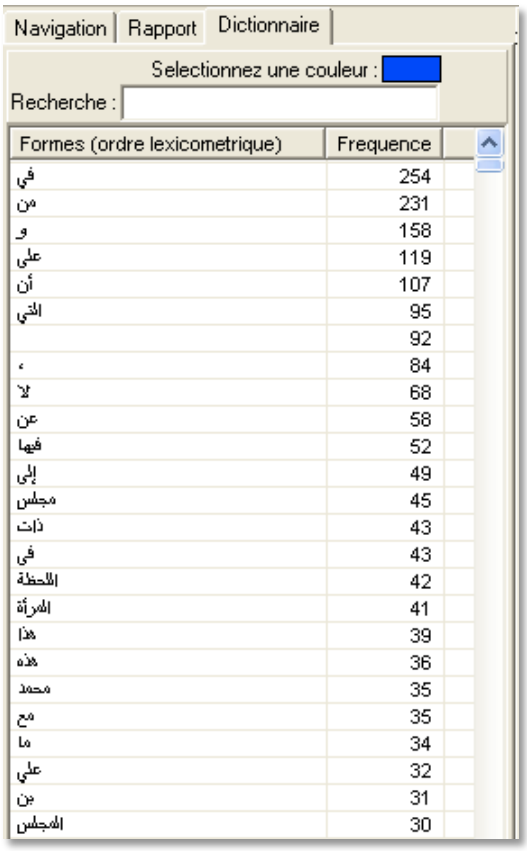
IV. Les outils de Lexico3

Lexico3 met de nombreux outils à dispositions de l'utilisateur. Certains sont assez basiques et d'autres beaucoup plus complexes. Nous allons progresser pas à pas dans les possibilités offertes par le logiciel.

IV. 1. Outils d'exploration textuelle

IV. 1. 1. Les dictionnaires

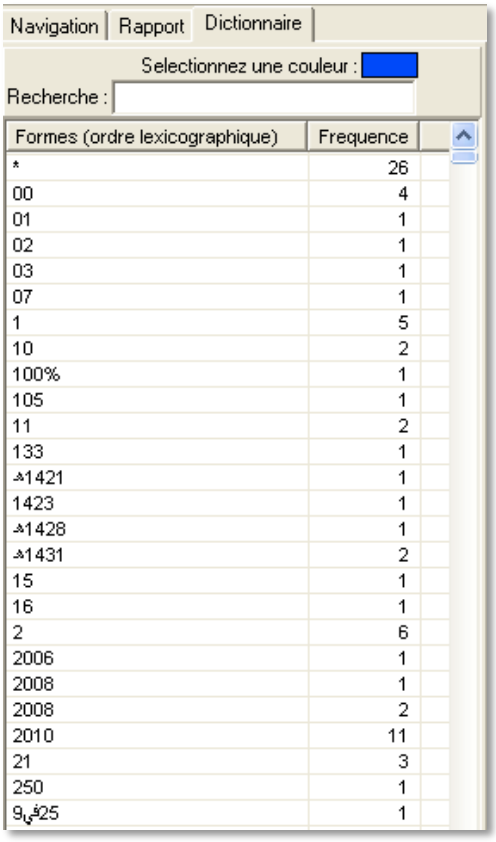
Nous avons vu que le logiciel génère automatiquement un fichier.dic que l'on peut ouvrir sur un éditeur de texte. La liste des formes présentes dans le texte est toujours consultable dans la colonne de gauche du logiciel, sous l'onglet *dictionnaire*, à côté des onglets *navigation* et *rapport*.


Navigation | Rapport | Dictionnaire

Selectionnez une couleur :

Recherche :

Formes (ordre lexicometrique)	Frequence
في	254
من	231
و	158
على	119
أن	107
التي	95
	92
،	84
لا	68
عن	58
فيها	52
إلى	49
مجلس	45
ذات	43
فى	43
اللحظة	42
المرأة	41
هذا	39
هذه	36
محمد	35
مع	35
ما	34
علي	32
بن	31
المجلس	30


Navigation | Rapport | Dictionnaire

Selectionnez une couleur :

Recherche :

Formes (ordre lexicographique)	Frequence
*	26
00	4
01	1
02	1
03	1
07	1
1	5
10	2
100%	1
105	1
11	2
133	1
هـ1421	1
1423	1
هـ1428	1
هـ1431	2
15	1
16	1
2	6
2006	1
2008	1
2008	2
2010	11
21	3
250	1
9في25	1

Dans le dictionnaire de gauche ci-dessus, les formes sont présentées dans l'*ordre lexicométrique* (fréquence décroissante). Si on clique sur l'onglet *Formes (ordre lexicographique*), on obtient le classement des formes du texte par ordre lexicographique (ordre alphabétique) : dictionnaire de droite ci-dessus.

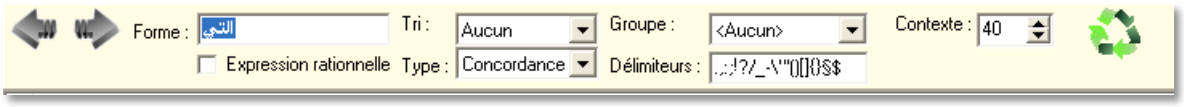
Les chiffres apparaissent en premier, puis les mots, mais attention au fait que l'ordre alphabétique n'est pas celui de l'arabe.

IV. 1. 2. Établir une concordance

Une concordance est une liste alignée (sous forme de trois colonnes distinctes) de toutes les occurrences en contexte d'une même forme (ou d'un même *type généralisé* dont on parlera un peu plus bas). L'occurrence en question figure au centre, en couleur (une couleur différente pour chaque mot dont on souhaite établir une concordance). De part et d'autre, on accède à son contexte à gauche et à droite, en noir. La concordance est un outil essentiel qui permet non seulement d'accéder au contexte, mais aussi de classer les occurrences en fonction de ce contexte. En effet, on peut classer le contexte avant ou après le mot choisit par ordre alphabétique, ce qui permet de repérer des segments répétés.

Pour établir une concordance, cliquez sur l'icône  . La fenêtre de travail s'ouvre et la

barre suivante apparaît :



Entrez la forme que vous souhaitez rechercher (ici *allatī*), puis cliquez sur entrer. Vous obtenez la concordance suivante :

Le mot recherché apparaît en rouge, au milieu. À chaque nouvelle concordance, la couleur du mot changera.

Différents menus déroulants figurent sur la barre :

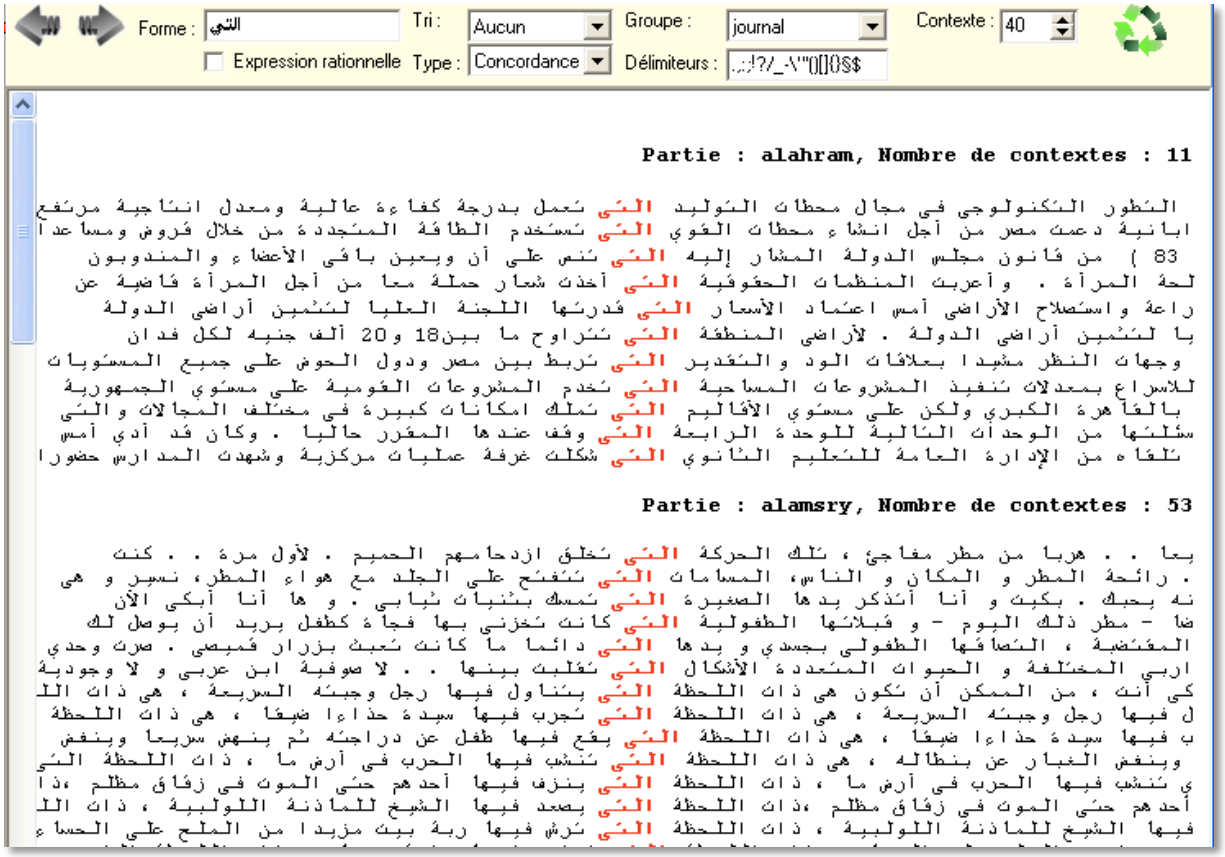
Tri : permet de choisir l'ordre de tri des contextes (aucun, avant, après) :
- par défaut, si aucun ordre n'est mentionné, la concordance affiche les occurrences dans leur ordre d'apparition dans le texte.
- si l'on sélection le **tri avant**, les occurrences seront classées par ordre alphabétique du mot précédant la forme choisie
- si l'on sélectionne le **tri après**, les occurrences seront classées par ordre alphabétique du mot qui suit la forme choisie.

Ces deux modes de tri permettent de repérer des expressions ou des structures syntagmatiques récurrentes.

L'exemple de tri du contexte avant met en évidence l'expression récurrente *ḏāt al-laḥẓa llatī* qu'on ne soupçonnerait peut-être pas si redondante :

Le tri du contexte suivant le mot fait apparaître qu'il s'agit souvent d'un verbe à la troisième personne suivit d'une préposition et d'un pronom de rappel.

Groupe : permet de regrouper les contextes en fonction d'une partition (par exemple, par locuteur, mois ou année). Une fois une partition créée (voir plus bas), on peut aussi établir une concordance par parties. Pour cela, sélectionner la partie voulue dans le menu défilant *groupe*. Dans l'exemple ci-dessous, on a choisi la partie "journal" : les occurrences sont donc triées par journaux. On a toujours la possibilité d'effectuer un tri avant ou après, à l'intérieur des partitions :



Contexte : permet de choisir le nombre de caractères (espaces inclus) qui doivent apparaître avant et après chaque pôle (au minimum, 10 caractères, de 5 en 5). On peut le modifier après une première recherche, en cliquant sur rafraîchir  pour obtenir la nouvelle concordance.

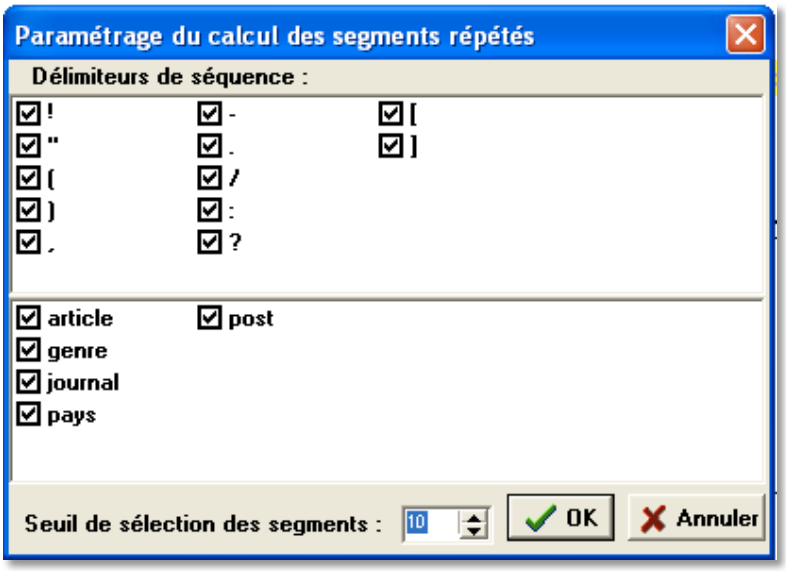
Pour revenir à une concordance antérieure ou postérieure, utiliser les flèches :  . Chaque nouvelle concordance utilise une nouvelle couleur.

Astuce : pour établir une concordance, il est aussi possible de glisser/déposer un mot depuis le dictionnaire ou à partir du garde mot (*cf.* IV. 1. 5). Il est possible de modifier la liste des caractères délimiteurs à tout moment lorsqu'on utilise l'outil concordance.

IV. 1. 3. Recherche de segments répétés

Il s'agit de trouver des groupes de formes qui se suivent toujours dans le même ordre et qui apparaissent au moins deux fois. Cliquez sur la touche *segments répétés* dans la barre des outils 

. Une fenêtre de dialogue intitulée "paramétrage du calcul des segments répétés" apparaît :



Dans la partie supérieure (*délimiteurs de séquence*), il est possible de désélectionner des signes préalablement enregistrés comme étant des délimiteurs de séquence. Cette fenêtre permet donc de fixer le statut des caractères délimiteurs du texte. Le statut par défaut est celui de délimiteur de séquence. Pour le modifier, il faut retirer la coche en regard du caractère correspondant. Les segments répertoriés ne chevaucheront pas ce type de délimiteur.

La partie inférieure permet de décider du statut des clés rencontrées dans le corpus : si la clé est cochée, le segment ne peut pas contenir des termes se trouvant de part et d'autre de cette clé. Au contraire, si la clé est décochée, c'est comme si elle n'existait pas (un segment pourra contenir des termes situés de part et d'autre de la clé). C'est intéressant car, dans le cas où une clé marque la pagination originale du document, celle-ci peut ne pas être prise en compte pour accéder aux segments qui se trouveraient à cheval entre deux pages.

En bas, il faut indiquer le *seuil de sélection des segments* qui, par défaut, s'élève à 10 formes.

On lance la recherche en cliquant sur *OK*.

Attention : les résultats ne s'ouvrent pas dans une fenêtre, mais figurent dans la colonne à gauche. Il faut cliquer sur l'onglet *segments répétés* qui s'est ouvert automatiquement au lancement de la recherche, à côté des onglets *navigation*, *rapport* et *dictionnaire* toujours

présents :

Lg	Segment	Frq
2	في هذا	11
2	و لا	11
4	ذات اللحظة التي ،	30
2	مجلس الدولة	12
2	مجلس الوزراء	14
3	ذات اللحظة التي	40
2	اللحظة التي	41
2	المجلس الخاص	12
2	المحكمة الدستورية	10

La colonne de gauche *Lg* indique la longueur du segment (en nombre de formes), la colonne du milieu donne le segment répété en question, et la colonne de droite *Frq* indique sa fréquence, i.e. le nombre de fois que le segment répété apparaît dans le corpus.

À chaque nouvelle recherche de segments répétés, un nouvel onglet s'ouvre. Pour les supprimer, il suffit de glisser/déposer l'onglet à l'extérieur de la colonne de gauche, puis de cliquer sur la croix en haut à droite de la fenêtre pour la fermer.

Attention : pour ajouter les résultats au rapport, il faut cliquer sur l'icône qui se situe non pas sur la barre des outils, mais dans l'onglet-même « *segments répétés* ».

Astuce : on peut aussi établir une concordance des segments répétés en glissant déposant le segment dans l'outil concordance ou en passant par le garde-mot (*cf.* IV. 1. 5).

IV. 1. 4. Constitution d'un groupe de formes

Si vous voulez rechercher en même temps un mot au singulier et au pluriel, un verbe dans toutes ses flexions, plusieurs mots ayant un sens proche, *etc*., alors il faut constituer ce que l'on appelle un groupe de formes ou encore un type (un type rassemble les occurrences de formes graphiques différentes liées par une propriété commune que détermine l'objectif de la recherche : l'utilisation d'un mot (au singulier et au pluriel), celle d'un verbe (quelle que soit sa conjugaison), l'emploi de synonymes (lien sémantique), *etc*.).

Les types généralisés, abrégés en Tgen, sont manipulés comme des entités uniques. Ils permettent de lancer simultanément une recherche sur plusieurs formes regroupées au préalable par le chercheur.

Cliquez sur l'icône *groupe de formes* :  La fenêtre suivante apparaît :

Elle vous permet de constituer et d'enregistrer votre groupe de forme.

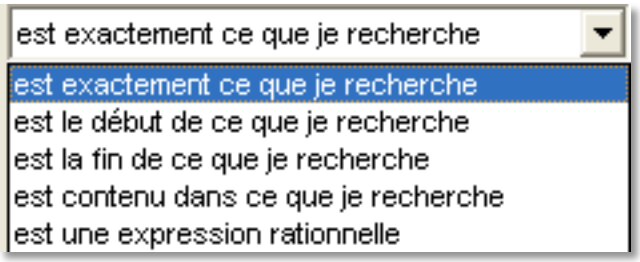
1. Tapez le nom que vous voulez donner à votre groupe.

2. Entrez le motif de la recherche. Cette étape nécessite de la réflexion pour que la recherche soit pertinente et efficace (qu'elle produise un résultat maximal à partir d'une manipulation minimale). L'outil groupe de formes ne sera pas manipulable avec la même facilité en français et en arabe, puisque le système morphologique des deux langues diffère notablement. Il faut comprendre les possibilités offertes par le logiciel pour arriver à créer un groupe de forme en faisant le moins de manipulations possible, mais sans omettre une occurrence potentielle.

Nous avons plusieurs commandes à notre disposition :

- le menu déroulant qui propose de considérer la forme rentrée comme étant

1. exactement ce qu'on recherche,
2. le début de ce qu'on recherche,
3. la fin de ce qu'on recherche,
4. contenu dans ce qu'on recherche,
5. une expression rationnelle :



La graphie concaténatoire de l'arabe, qui offre la possibilité de réunir en une seule forme graphique prépositions monolitères antéposées au mot et suffixes postposés au mot, s'avère

ici être un handicap. Si on veut chercher un nom commun, les trois premières commandes ne sont pas valables : 1. est exactement exclurait toutes les formes graphiques où ce nom est agrégé à une préposition antéposée ou supporte un suffixe postposé, 2. et 3. excluant respectivement l'une et l'autre des formes graphiques potentielles énoncées en 1.

Il apparaît donc qu'il faille le plus souvent utiliser la 4ème commande « est contenu dans ce que je recherche ». La commande choisie doit dans tous les cas faire l'objet d'une réflexion préalable.

On dispose aussi, pour pouvoir englober un maximum de formes en une seule commande, d'expressions régulières (aussi appelées expressions rationnelles). *Cf.* le manuel d'utilisation Lexico3 pour une liste.

Le sens de la commande pour l'arabe est le même que le sens d'écriture (cf. exemples ci-dessous).

🔔 **Attention**, pour lancer une nouvelle recherche il faut relancer l'outil groupe de mots, sinon les résultats s'ajoutent les uns aux autres.

Prenons différents exemples :

Il n'y a évidemment aucun intérêt à chercher un mot unique qui ne peut apparaître que sous une seule forme graphique.

Remarque : notre texte n'est pas vocalisé, donc nous chercherons des mots non vocalisés, mais pour la transcription, on donne les mots cherchés, pas leur translittération caractère pour caractère (*kbīr* pour *kabīr* par exemple, par souci de clarté à la lecture).

A. Recherche d'un adjectif : nous travaillons sur un adjectif et voulons toutes ses occurrences au singulier, duel et pluriel, au masculin et féminin. Prenons l'adjectif *kabīr* : les différentes formes possibles seront : *kabīr, kabīra, kabīrāni, kabīratāni, kibār, kabīrat*. Ces mots sont réductibles à deux formes différentes : *kabīr* (que comprennent les formes du singulier et du duel masculin et féminin ainsi que le féminin pluriel) et *kibār*.

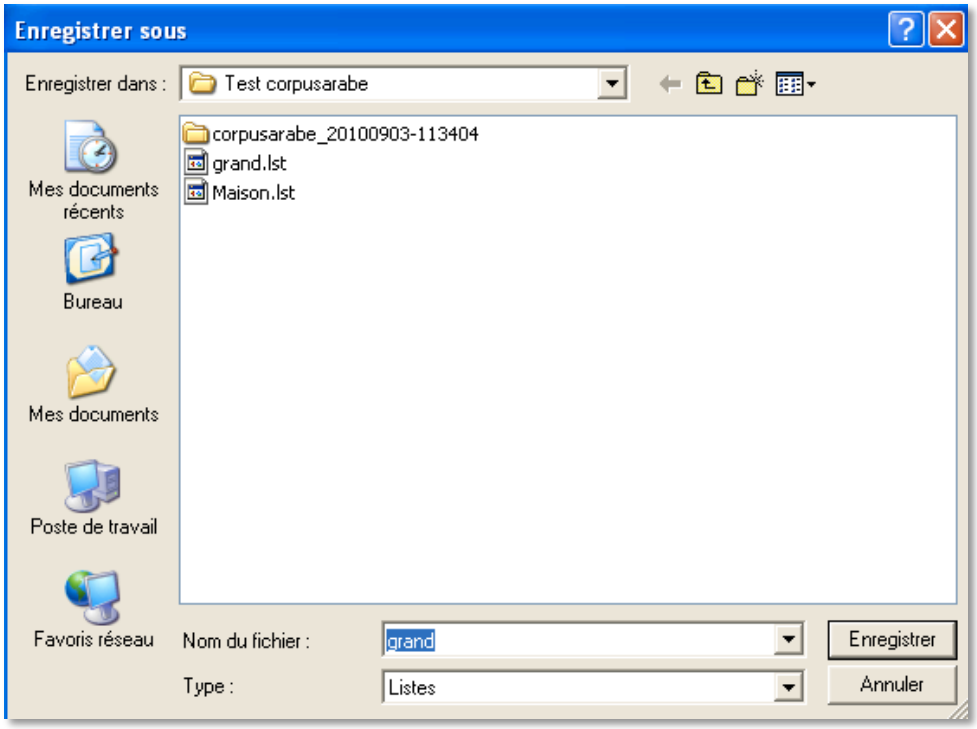
1. On lance la recherche de *kabīr* et avec la commande "est contenu dans ce que je recherche" du menu déroulant, et on obtient quatre formes :

Si aucun nom du groupe n'a été saisi, il en apparaît un automatiquement (comme ci-dessus), sous la forme +motif. Il est possible de le modifier.

2. On lance, dans la même fenêtre, la recherche de *kibār* avec la commande "est contenu dans ce que je recherche". Les résultats sont ajoutés aux précédents, pour permettre justement de créer le groupe de formes :

3. Il est possible de choisir une couleur à appliquer au Tgen en cliquant sur le carré blanc qui se situe à droite de la commande enregistrer.

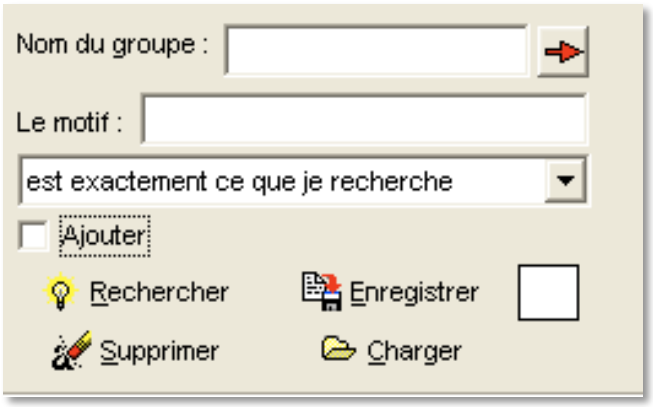
4. Enregistrer la liste ainsi constituée en cliquant sur enregistrer. Par défaut, le fichier sera enregistré dans le dossier d'où provient le texte source, en extension.lst (liste) :



Ici par exemple, nous nommons notre fichier « grand », qui apparaît dans le dossier sous le nom de grand.lst.

En cas d'erreur, pour relancer une nouvelle recherche qui ne s'ajoute pas à la précédente sans forcément relancer l'outil groupe de forme, il suffit de décocher « ajouter » au-dessous du

menu déroulant. Le nouveau résultat effacera la précédente recherche :



Avant de prendre d'autres exemples, il est intéressant de voir ce que la recherche aurait donné en modifiant le choix du menu déroulant. Nous avions demandé *kabīr* « est contenu dans ce que je recherche » et avions obtenu quatre formes.

Voyons ce que cela donne avec les autres commandes :

- *kabīr* est exactement ce que je recherche donne 1 seule forme :

- *kabīr* est le début de ce que je recherche donne 2 formes :

- *kabīr* est la fin de ce que je recherche donne 3 formes :

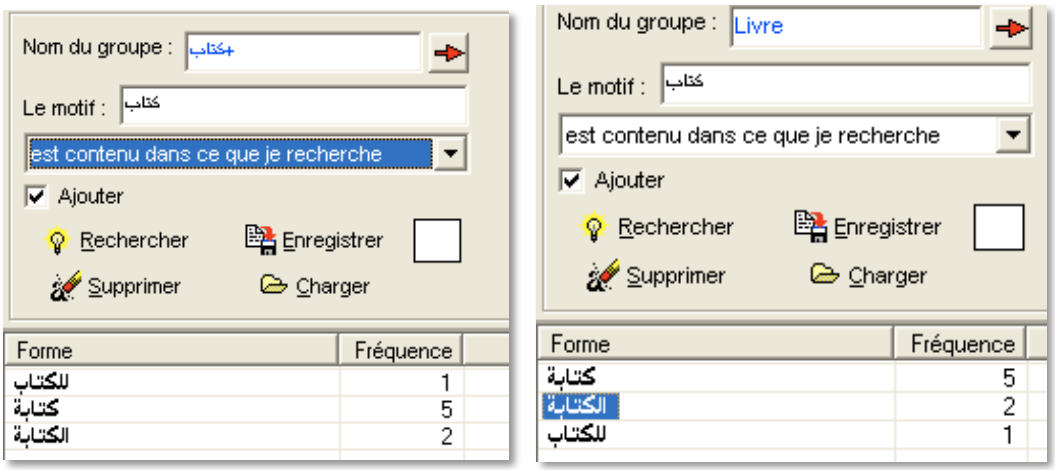
Pour cet exemple, c'est donc bien la commande « est contenu dans ce que je recherche » qui nous fournit le meilleur résultat.

B. Recherche d'un nom : nous voulons trouver toutes les occurrences du mot « livre » dans notre corpus arabe. Il nous faut donc créer un groupe de formes avec le singulier *kitāb* et le pluriel *kutub*. Notre texte n'étant pas vocalisé, on peut imaginer trouver un peu de bruit (*kuttāb*, s'il figure dans le corpus, sera analysé comme *kitāb*, sauf mention de la *šadda*, comme discuté tout au début de ce guide, mais l'analyse du contexte, donc le retour au texte sera inévitable si la *šadda* n'est pas notée ; *katab* a le même *ductus* que *kutub*). Cet exemple est équivalent au premier, mais nous allons cette fois-ci tester les expressions régulières et voir comment effacer le bruit (les mots trouvés qui ne nous intéressent pas).

On a plusieurs techniques pour constituer notre groupe de formes : soit faire deux recherches consécutives (*kitāb* puis *kutub*, en ajoutant l'une à l'autre et en supprimant le bruit), soit trouver l'expression rationnelle qui va nous permettre de trouver toutes les formes potentielles en une seule commande.

Première technique :

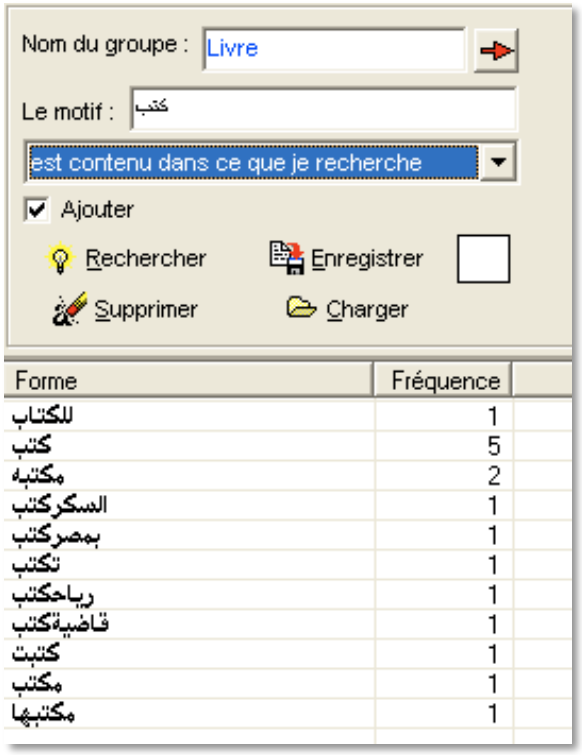
1. On lance une recherche simple, le motif *kitāb est contenu dans ce que je recherche*.



On obtient du bruit, puisqu'on trouve deux formes du mot *kitāba* qui ne nous intéresse pas ici. Pour les supprimer de la liste, c'est très simple, il suffit de cliquer sur la forme (elle est surlignée en bleu pour signaler qu'elle est bien sélectionnée) puis de cliquer sur *supprimer*.

Il est possible de sélectionner plusieurs formes en même temps, en maintenant la touche majuscule gauche du clavier enfoncée.

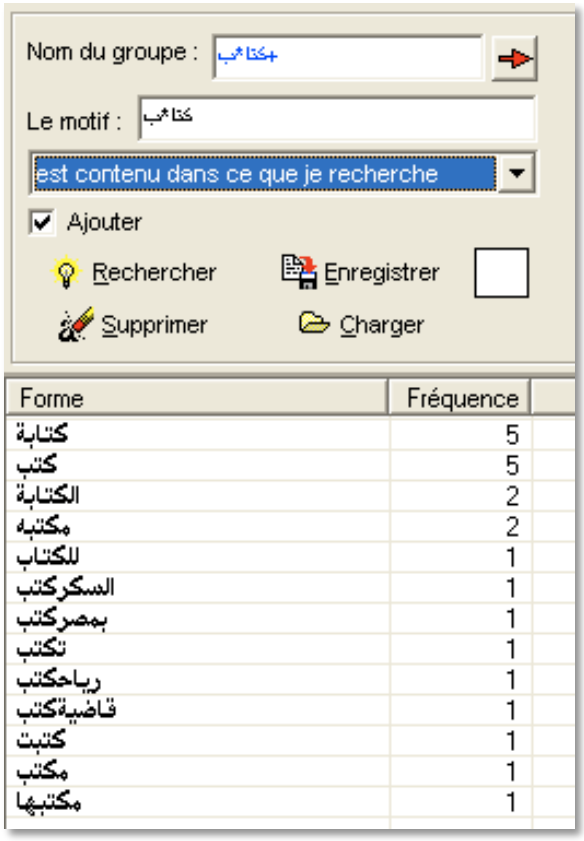
2. On lance une seconde recherche, motif *kutub est contenu dans ce que je recherche*. On obtient alors 10 nouvelles formes ajoutées à *kitāb* conservé de l'étape précédente. Sur ces 10 formes, une seule nous intéresse. La recherche a généré beaucoup de bruit, qu'il faut supprimer.



Deuxième technique :

On passe par les expressions régulières pour obtenir le même résultat en une seule manipulation.

Parmi les expressions régulières mentionnées ci-dessus dans un tableau, une seule peut satisfaire notre demande : *kitāb*, l'astérisque signifiant 0, 1 ou plusieurs fois le caractère précédant, ici le *alif*. Le logiciel cherchera 1. *ktb* (0 fois *ā*), le *ductus* de *kutub*, 2. *kitāb* (1 fois *ā*), puis en théorie 3. *kitāāb*, écriture impossible en arabe. Il s'agit donc de la commande idéale :



Forme	Fréquence
كتابة	5
كتب	5
الكتابة	2
مكتبه	2
للكتاب	1
السكركتب	1
بمصركتب	1
تكتب	1
رياحكتب	1
قاضيةكتب	1
كتبت	1
مكتب	1
مكتبها	1

Sur les 13 résultats, seuls deux nous intéressent (en partant du principe que le ductus *ktb* transcrit le nom *kutub* et pas le verbe *kataba*, ce qui n'est pas vérifiable autrement qu'en retournant au texte en l'absence de vocalisation).

Forme	Fréquence
كان	23
يكون	8
كانت	8
تكون	6
لتكون	4
لإمكانية	2
ستكون	2
مكان	2
وتتكون	2
وكان	2
وكانت	2
لإمكانية	1
لكنني	1
لمكان	1
ليكون	1
نكون	1
يتحركون	1
يتمسكون	1

Les expressions régulières permettent donc de créer des groupes de formes en faisant un minimum de recherches, mais elles semblent plus pertinentes au système morphologique français qu'arabe.

C. Recherche d'un verbe : nous voulons étudier l'emploi du verbe *kāna* : il nous faut regrouper toutes les formes possibles de sa conjugaison. En fait, elles peuvent être réduites à trois suites graphiques : *kun*, *kūn* et *kān*. Rechercher ces trois formes simultanément générera beaucoup de bruit, car on les trouvera dans de nombreux mots autre que le verbe *kāna*, mais cette recherche peut être faite assez rapidement. Deux étapes sont tout de même nécessaires :

1. On cherche *k.n* (expression régulière

permettant de trouver en une seule commande toutes les occurrences de *kān* et de *kūn*). On obtient alors 44 formes. Puis on cherche *kn* (après avoir vérifié que la commande *ajouter* est bien cochée). On obtient un total de 77 formes (respectivement 44 et 33 formes pour la nouvelle recherche *kn*).

Les résultats de la deuxième commande apparaissent après ceux de la première, ce qui est visible dans la fenêtre suivante :

En faisant défiler les résultats, on repère facilement le passage de la première à la seconde commande en regardant la fréquence des formes. Il faut ensuite nettoyer le bruit, seules 20 formes sont pertinentes.

D. Recherche de synonymes : nous voulons étudier la répartition de mots sémantiquement liés, par exemple : *bayt*, *dār* et *manzil*. On reprend les étapes précédentes :

1. Cliquez sur l'outil *groupe de formes*.

2. Nommez le groupe *maison*.

3. Cherchez d'abord *bayt est contenu dans ce que je recherche*, puis **en vérifiant que l'option *ajouter* est bien cochée** (sous le menu déroulant), on lance une nouvelle recherche en tapant *dār* dans le motif (toujours *est contenu dans ce que je recherche).* Cette nouvelle recherche va générer du bruit. On lance la troisième recherche, *manzil*, qui ne donne aucun résultat. On obtient le groupe suivant :

Astuce : il n'est pas nécessaire, lorsque l'on ajoute plusieurs recherches, de nettoyer le bruit à chaque fois : autant le faire une bonne fois pour toute à la fin.

Associer des groupes de formes :

Si l'on veut associer deux groupes de formes déjà enregistrés pour en créer un nouveau, il faut cliquer sur *charger*. Ici, par exemple, nous étions sur le groupe *grand*, et nous avons chargé le groupe *maison*, ce qui nous donne un nouveau groupe avec toutes les formes de l'adjectif *grand* et les trois synonymes de *maison*. L'exemple en question n'est pas des plus pertinents, mais il permet de comprendre le fonctionnement de cette possibilité.

Réutiliser les groupes de formes :  On peut utiliser la flèche rouge à droite du nom du groupe de formes pour le glisser/déposer :

IV. 1. 5. Le garde mots

Le garde-mots se trouve au milieu de la barre des outils : 
Il permet de mémoriser formes, segments ou TGen pour une utilisation ultérieure. Pour stocker un TGen dans le garde-mots il suffit de le faire glisser sur l'icône du cube rouge (*cf.* glisser/déposer supra). Pour utiliser un TGen stocké dans le garde-mots on le glisse à partir du cube rouge jusqu'à la fenêtre de travail (concordance, ventilation des fréquences, carte des sections, *etc.*) dans laquelle il doit être visualisé.

IV. 2. Outils d'analyse statistique

Le logiciel Lexico3 permet d'accéder à des fonctionnalités basiques (description statistique élémentaire : comptages, histogrammes, *etc.)* ainsi qu'à des analyses multidimentionnelles des données textuelles beaucoup plus poussées (analyses factorielles des correspondances, classification automatique, analyse des séries textuelles chronologiques).

On présentera ces fonctionnalités une à une, en expliquant à quoi elles correspondent au niveau de la statistique.

IV. 2. 1. Le découpage en parties

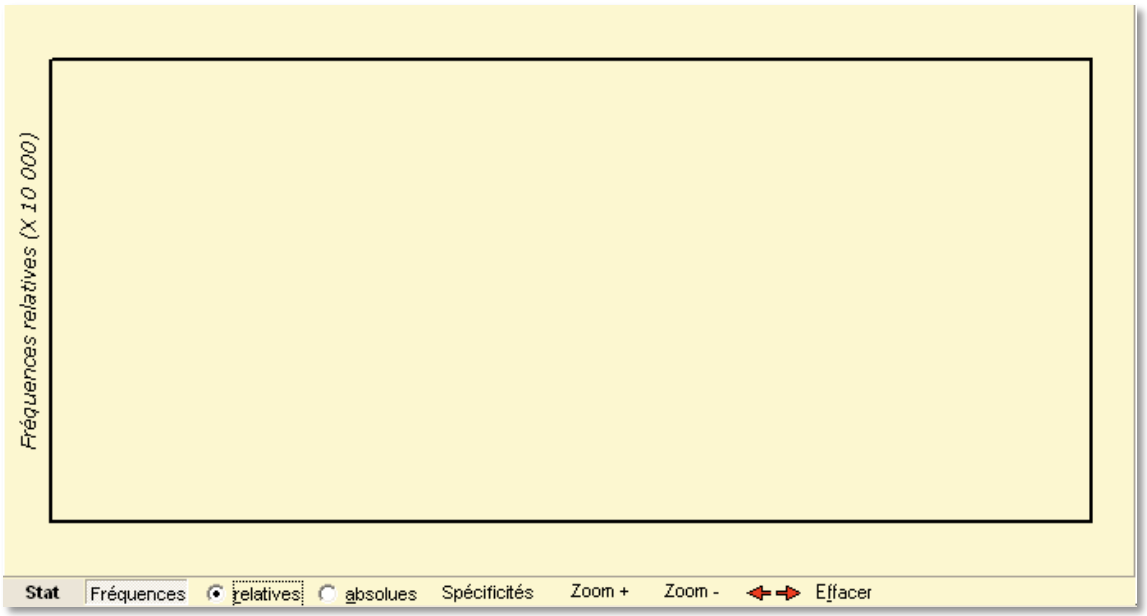
Avant de segmenter le corpus, nous y avions introduit des balises, indiquant par exemple le genre, l'auteur, la date, *etc*. du texte ainsi balisé. Ceci va nous permettre de constituer des sous-corpus en quelque sorte, i.e. de regrouper les différentes parties du corpus en fonction d'une caractéristique donnée. Ainsi, on pourra étudier un phénomène selon une partition préétablie, par exemple : la fréquence d'utilisation d'un mot en fonction de la date ou de l'auteur du texte. De ce fait, il est important de réfléchir, dès la constitution du corpus (lors de l'étape de son balisage), aux différentes clés qui doivent être introduites.

Pour réaliser un découpage en parties (une partition du corpus), il faut sélectionner une clé : tous les textes balisés par cette clé seront regroupés en un ensemble. Ainsi, vu les clés introduites dans notre *corpusarabe*, nous pouvons opérer des partitions par genre (presse ou blog), pays (Arabie Saoudite, Yémen, Syrie, Égypte), par journal, article ou post. Au niveau technique, les différents contenus affectés au type de clé sélectionné découpent le corpus en autant de parties différentes.

Cliquez sur l'outil *Partitions* :  . Une boîte de dialogue apparaît, où figurent toutes les clés contenues dans le corpus. Pour créer une partition, il suffit de cliquer sur la clé

en question, puis sur *créer* :

Un graphique des fréquences relatives s'ouvre alors automatiquement, pour permettre de comparer les fréquences des unités textuelles dans l'ensemble de la partie constituée (du sous-corpus *article*, regroupant tous les articles du corpus, si nous venons de créer cette partition) :



IV. 2. 1. a. Ventilation d'une forme

La **ventilation** des occurrences d'une forme dans les parties du corpus, c'est la suite constituée par la succession des sous-fréquences prises dans l'ordre des parties.

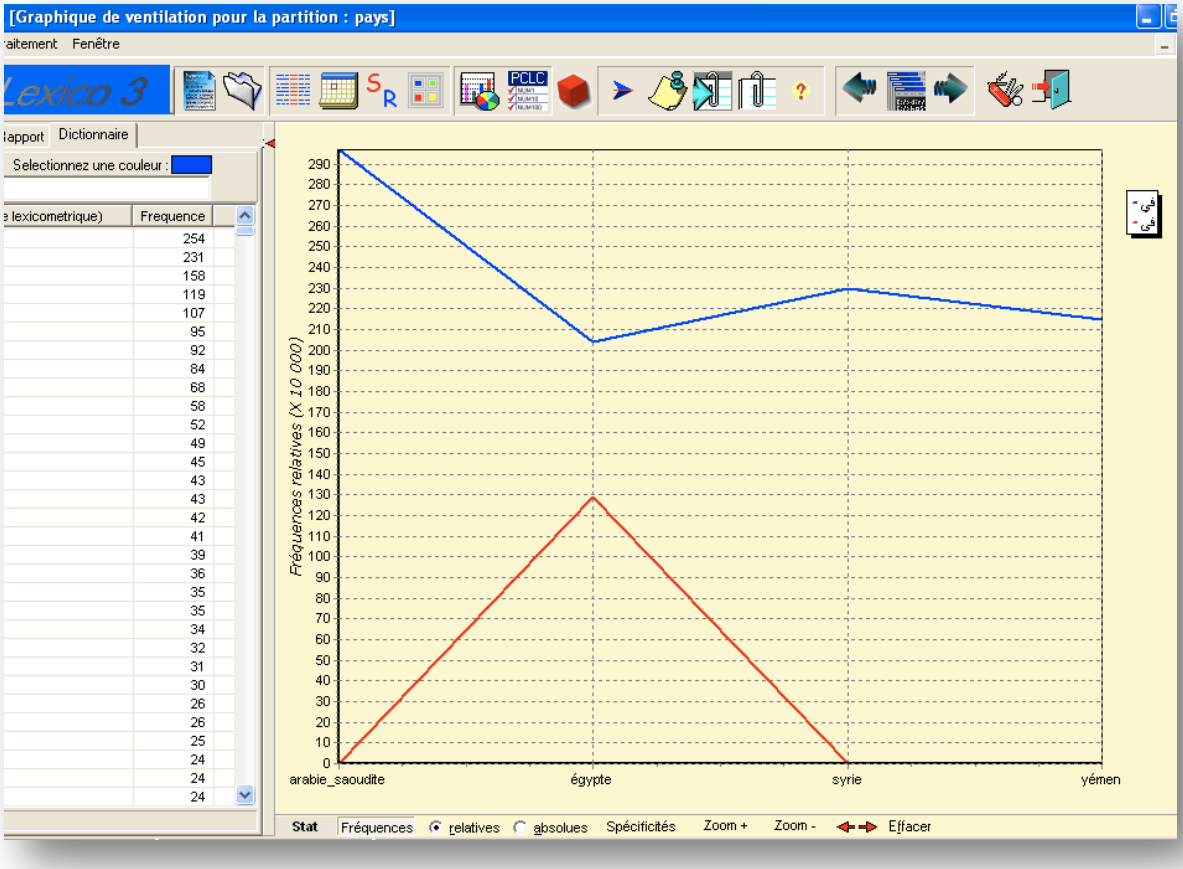
Il faut donc commencer par créer une partition. Une fois la partition créée, c'est la fenêtre de ventilation qui s'ouvre automatiquement.

Ensuite, on glisse le mot (ou un T-gen du garde-mot) dans la fenêtre des résultats. On peut glisser plusieurs unités textuelles (pour comparaison), chacune aura une couleur et la légende se fera automatiquement dans la marge.

On peut visualiser la ventilation dans les parties du corpus :

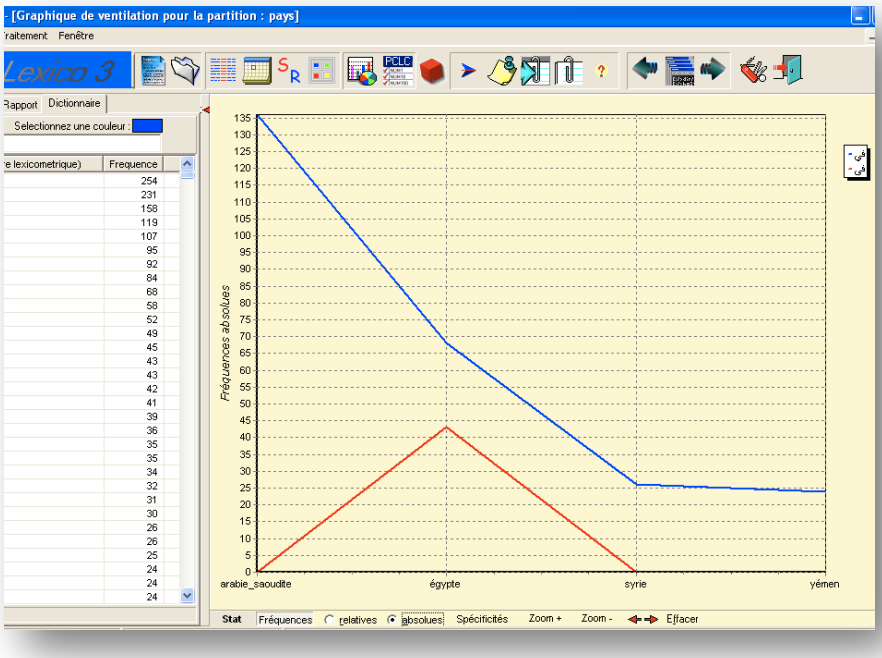
- en fréquences absolues (nombre d'occurrences dans la partie)
- en fréquences relatives (nombre d'occurrences rapporté à la longueur de la partie)
- en termes de spécificités (résultat d'un calcul statistique, *cf. infra*).

Prenons l'exemple suivant : on veut connaître la fréquence d'utilisation du mot *fī* écrit avec ou sans les points sur le *yā'*, pour savoir si cet usage est propre à l'Égypte. On crée une partition par pays, puis on glisse depuis le dictionnaire les mots *fī* avec points et sans points. On obtient le graphique suivant :



Le titre apparaît dans la barre tout en haut : [Graphique de ventilation pour la partition : pays]. Par défaut, le graphique mentionne les fréquences relatives (choix coché dans la barre tout en bas). Le mot *fī* avec points est en bleu, le mot *fā* sans points en rouge. En ordonnée, on a la fréquence relative, en abscisse les différents pays selon la partition effectuée. On constate que l'emploi du mot *fā* sans les points ne figure qu'en Égypte.

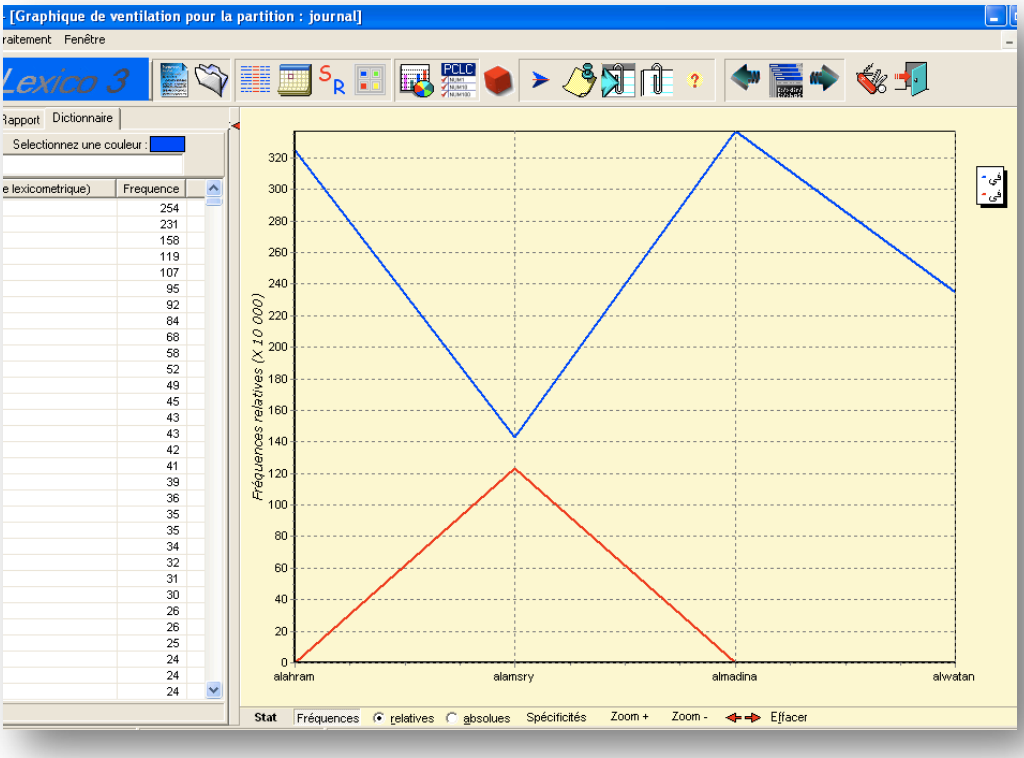
Il est possible d'obtenir le même graphique selon les fréquences absolues. Il suffit de cocher *absolues* dans la barre tout en bas. On obtient alors le graphique suivant :



Le zoom (toujours dans la barre en bas) permet d'agrandir le graphique pour lire le nombre de fréquences avec plus de précision.

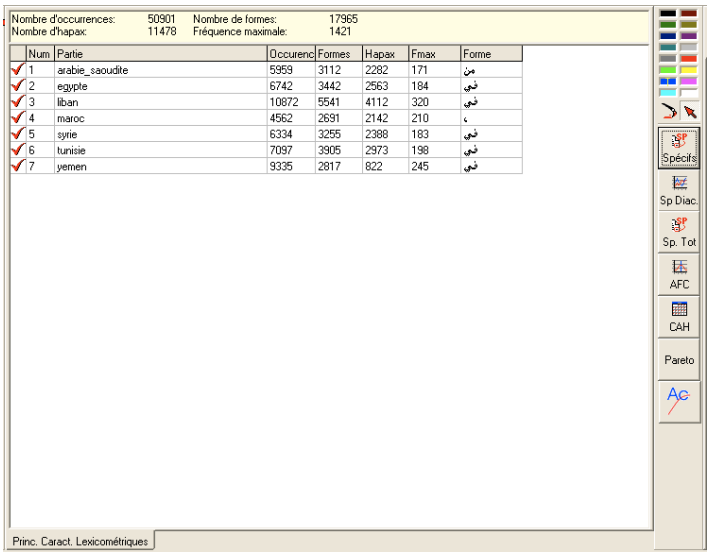
Lorsque l'on clique sur effacer, le graphique de ventilation est de nouveau vierge : on peut obtenir une nouvelle ventilation.

Si, maintenant, on crée une partition par journal et qu'on veut obtenir la ventilation de ces deux mêmes formes, on peut visualiser le résultat suivant où l'on constate donc que dans Al-Ahram, tous les *yāʾ* ont des points :



IV. 2. 1. b. Principales caractéristiques lexicométriques

Lorsque l'on réduit la fenêtre (en haut à droite : dans le gris et non pas dans le bleu, qui figure la fenêtre du logiciel Lexico 3), on voit apparaître une autre fenêtre, celle des principales caractéristiques de la partition créée.



On peut aussi obtenir cette fenêtre en cliquant sur l'icône PCLC 

, signifiant

Principales Caractéristiques Lexicométriques du Corpus et de la partition, et en sélectionnant la partie créée qui nous intéresse.

Le tableau fait apparaître les principales caractéristiques par partie suivant la partition choisie.

- une coche rouge dans la colonne la plus à gauche indique que la partie est sélectionnée pour le décompte des fréquences globales dans le corpus.
- la deuxième colonne numérote les parties.
- la troisième colonne donne les noms des différentes parties (ici il s'agit de la partition par pays, on a donc le nom des différents pays).
- la colonne ***occurrences*** indique le nombre des occurrences des formes répertoriées.
- la colonne ***formes*** indique le nombre des formes graphiques présentes dans chaque partie.
- la colonne ***hapax*** indique, pour chaque partie, le nombre des formes qui n'apparaissent qu'une fois dans la partie.
- la colonne ***fréquence maximale*** indique le nombre des occurrences de la forme la plus fréquente, que l'on retrouve dans la dernière colonne.

Ce tableau permet une comparaison visuelle rapide des parties en fonction de leurs caractéristiques lexicométriques les plus importantes.

Voici la signification des icônes de droite :

: Calcul des spécificités pour la partition et le corpus sélectionnés.

: Calcul des accroissements spécifiques et des spécificités chronologiques.

: Calcul des spécificités par partie, chronologiques et évolutives pour toutes les parties et sur la totalité du corpus.

: Analyse factorielle des correspondances.

:

: Diagramme de Pareto

: Accroissement du vocabulaire

IV. 2. 1. c. Spécificités (présentation générale)

Un peu de théorie *(Lebart et Salem, chapitre 6, p. 170 et suivantes)*.

« Les représentations spatiales fournies par l'analyse des correspondances gagnent à être complétées par quelques paramètres d'inspiration plus probabiliste : les *spécificités* ou *formes caractéristiques*. Il existe en effet des outils statistiques qui permettent de décrire chacune des classes d'une partition en exhibant tout à la fois les unités qu'elle contient en grand nombre par rapport aux autres classes, et au contraire, les unités qu'elle contient en très petit nombre. Dans le cas des unités de base, qu'il s'agisse de formes, de segments ou de lemmes, on parlera d'éléments caractéristiques ou de spécificités.

L'analyse des correspondances crée une typologie portant à la fois sur l'ensemble des parties du corpus et sur l'ensemble des formes attestées dans celui-ci. Cependant, il peut être utile de compléter ces analyses globales par des calculs probabilistes effectués séparément à partir de chacune des sous-fréquences du tableau lexical.

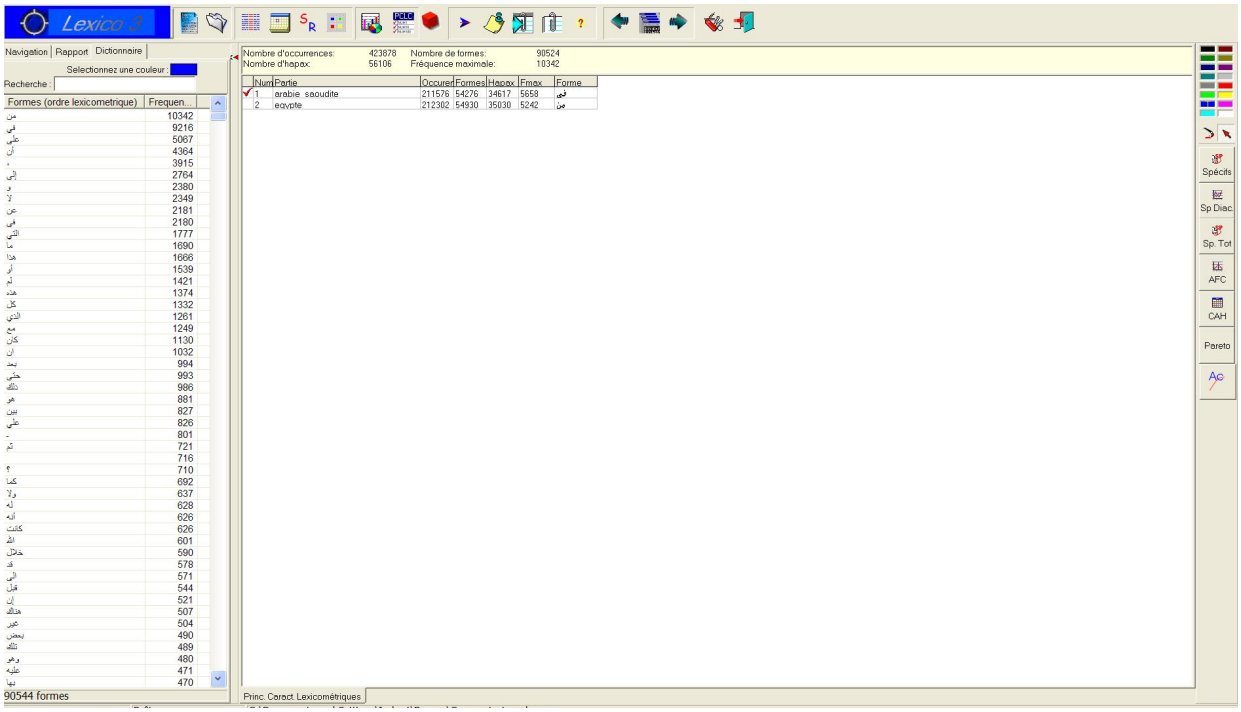
accroissement spécifique - (sp) spécificité calculée pour une partie d'un corpus par rapport à une partie antérieure
analyse factorielle (stat) - famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des « facteurs » résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.
analyse des correspondances (stat) - méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou c2). »

IV. 2. 1. d. Calcul des spécificités

Attention : il faut toujours faire les spécificités totales avant le reste, sinon rien n'apparaît.

On commence par créer une partition : la fenêtre de ventilation s'ouvre, qu'on peut réduire ; en-dessous figure la fenêtre des principales caractéristiques lexicométriques. On l'agrandit.

On coche la partition que l'on souhaite observer : sur l'exemple suivant, on voit que j'ai coché la partie 1 arabie_saoudite. Il suffit de cliquer à gauche du numéro de la partition pour sélectionner ou désélectionner une partie.

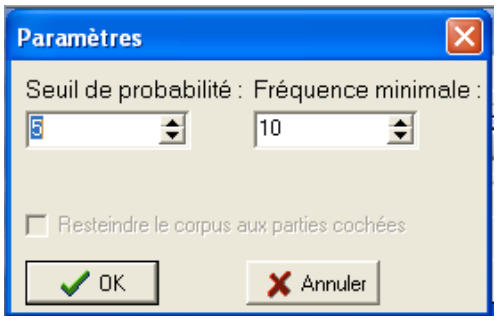


Calcul des spécificités totales

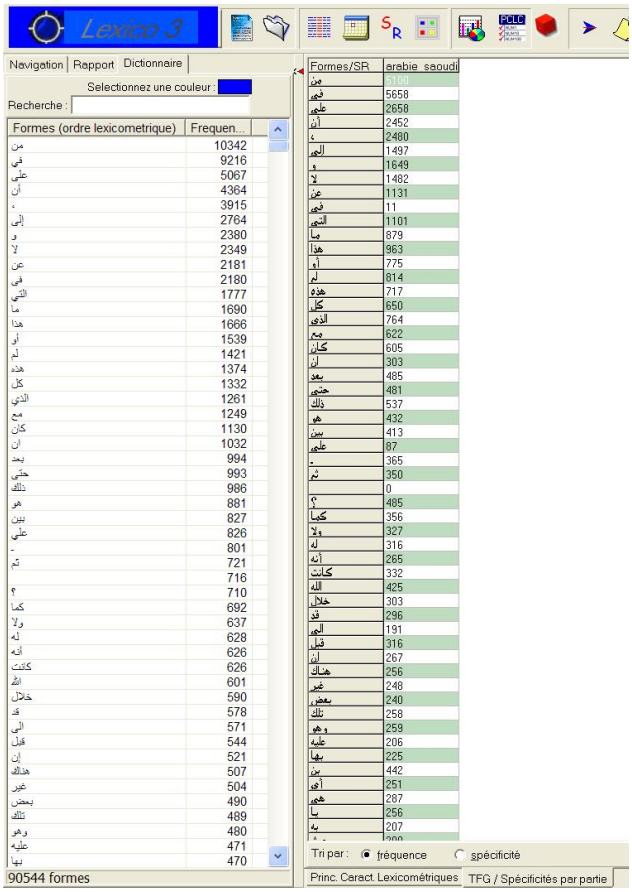
Ensuite, je clique sur les **spécificités totales** : 

. Une fenêtre de paramétrage s'ouvre :

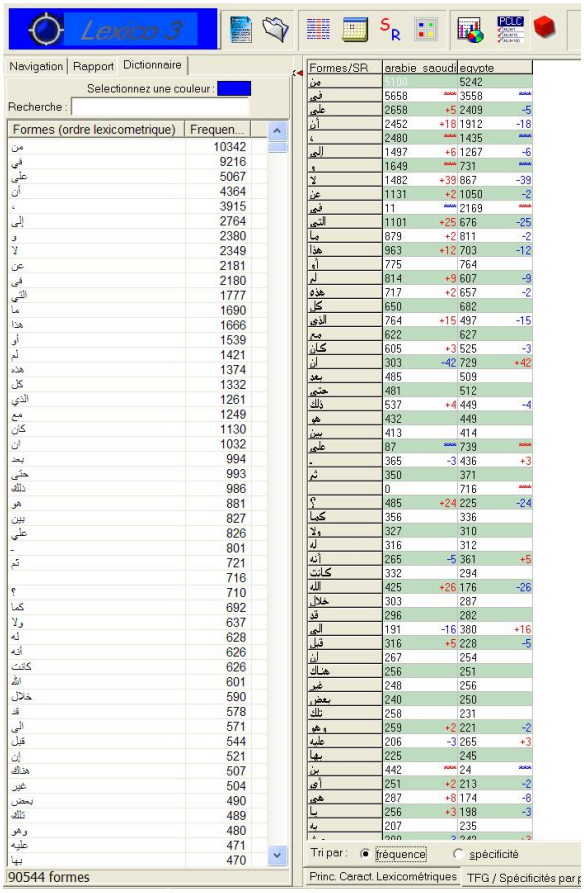
Ces réglages par défaut signifient que l'indice de spécificité est calculé pour toutes les unités dont la fréquence est supérieure à 10, avec un *seuil de probabilité* fixé à 5 %. On peut modifier ces paramètres et décocher la mention « restreindre le corpus aux parties cochées ».



Si l'on garde la coche « restreindre le corpus aux parties cochées » et qu'on clique sur ok, on va obtenir le tableau suivant : colonne de gauche, les formes et segments répétés classés par ordre de fréquence, colonne de droite le nombre d'occurrences de cette forme dans la partie sélectionnée.



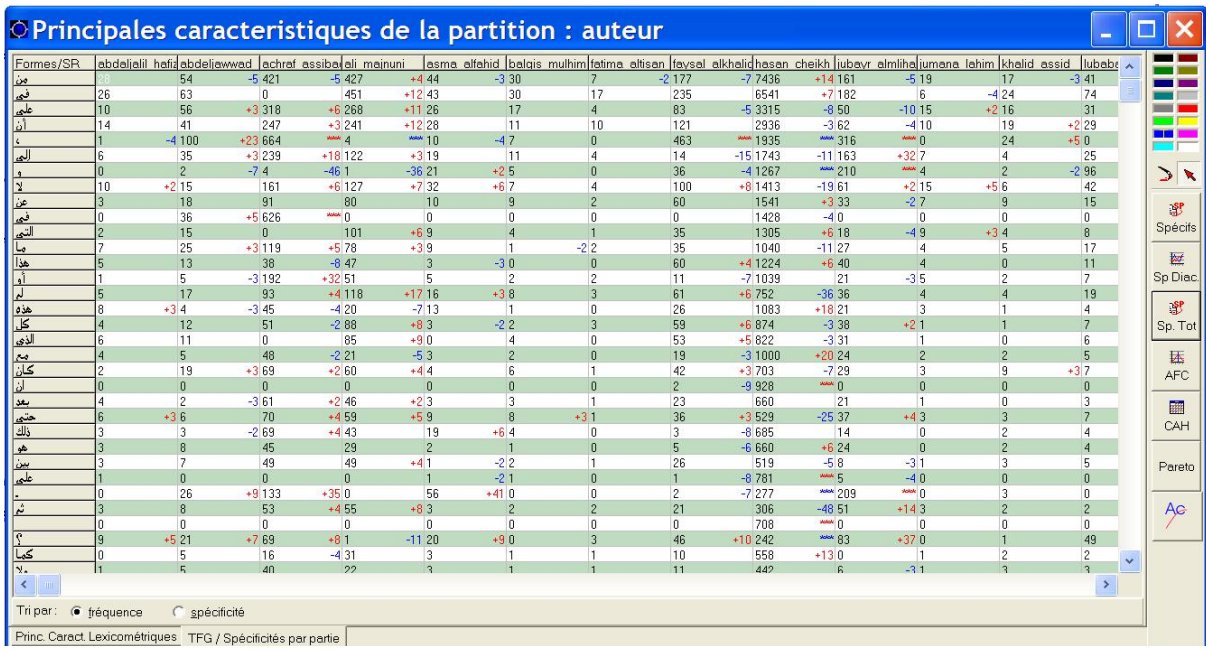
Si au contraire, on a décoché « restreindre le corpus aux parties cochées » dans les paramétrages, on obtient un tableau contenant l'ensemble des parties :



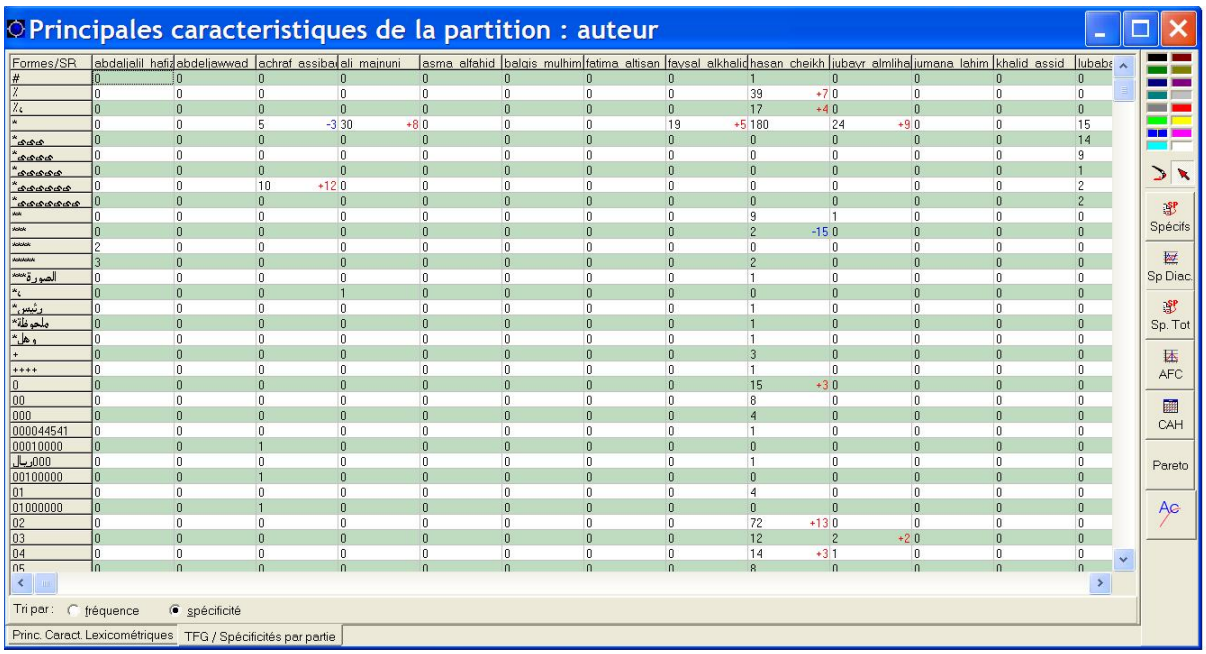
Si l'on clique sur le Formes/SR sur le haut du tableau, les formes vont être présentées par ordre de fréquence croissant. De même, si l'on clique sur le nom des parties dans le haut du tableau, les formes vont apparaître par fréquence décroissante dans cette partie en particulier, puis par fréquence croissante toujours dans cette partie si l'on reclique.

On observe aussi en bas du tableau une ligne *Tri par* : o *fréquence* o *spécificités*.
Par défaut, c'est « fréquence » qui est coché.

Dans l'exemple précédent, il n'y avait que deux parties. Je prends un autre exemple avec une partition plus fournie. Les spécificités totales classées par fréquence donnent le tableau suivant :



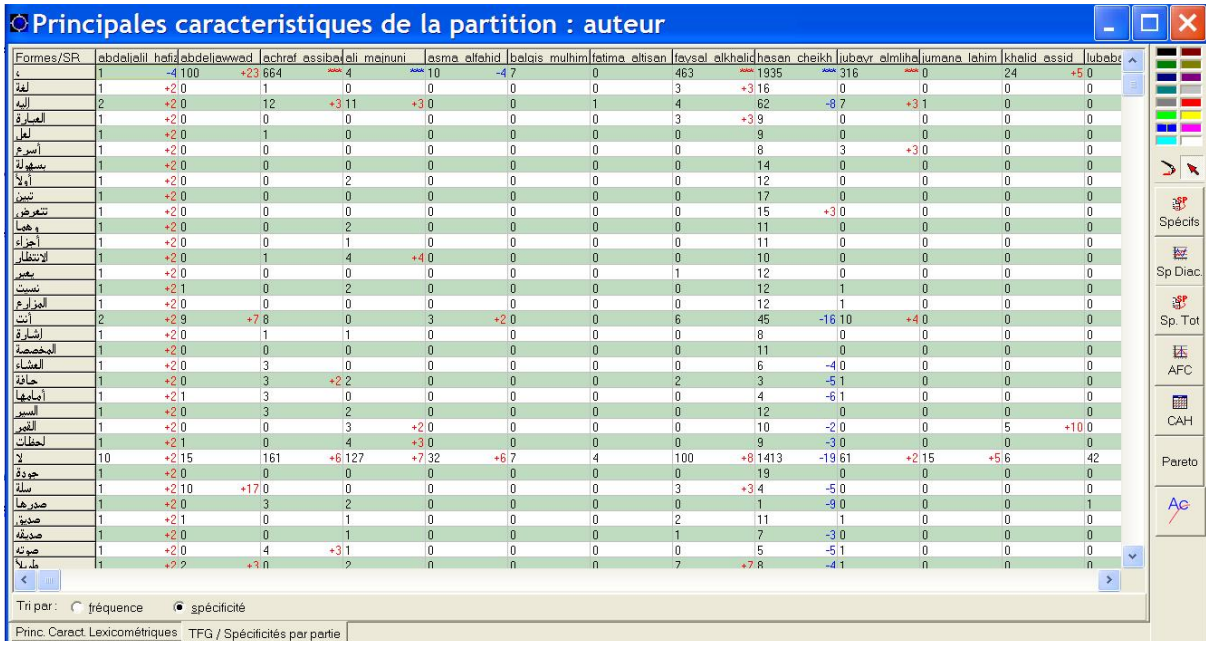
Voilà ce qui se passe si l'on coche « spécificités » en bas (quand on coche « spécificités », il faut patienter un peu) : par défaut, ce qui s'affiche est un tableau des spécificités classées par ordre lexicographique des formes (colonne de gauche).



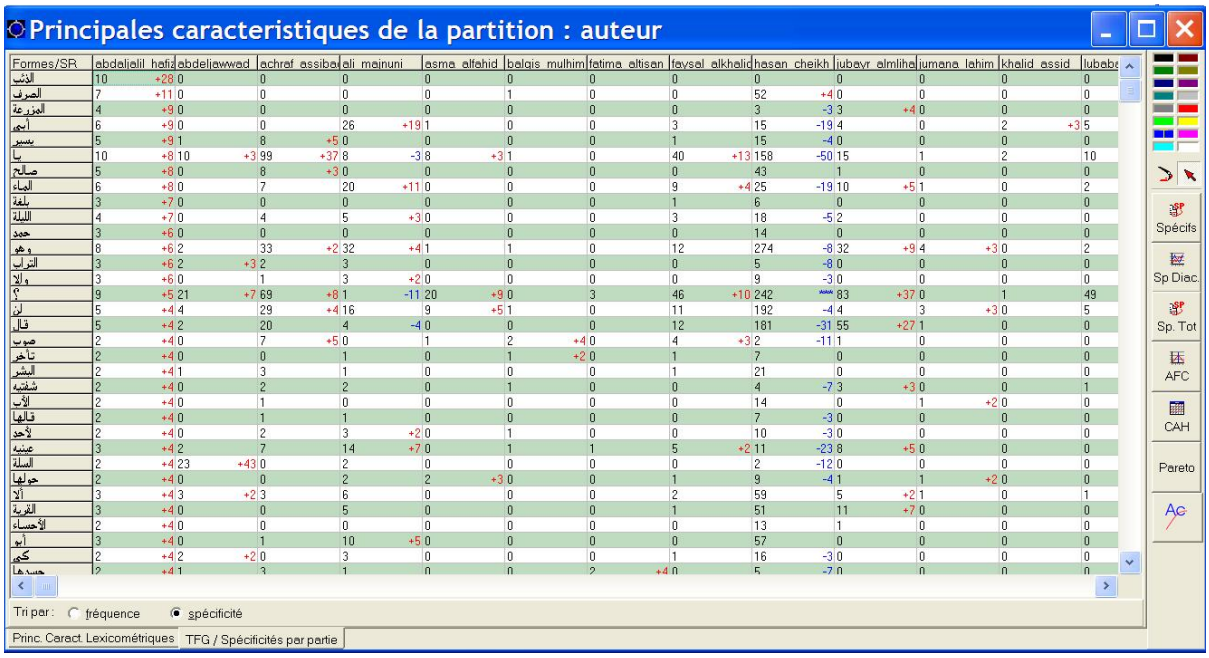
De la même manière, si on reclique sur « Formes / SR », on aura la liste des formes spécifiques par ordre croissant, si l'on clique sur une partition, on aura la liste des formes spécifiques de cette partition par ordre décroissant puis croissant si l'on reclique.

Exemple de spécificités totales par spécificité de la première partie (on a cliqué sur

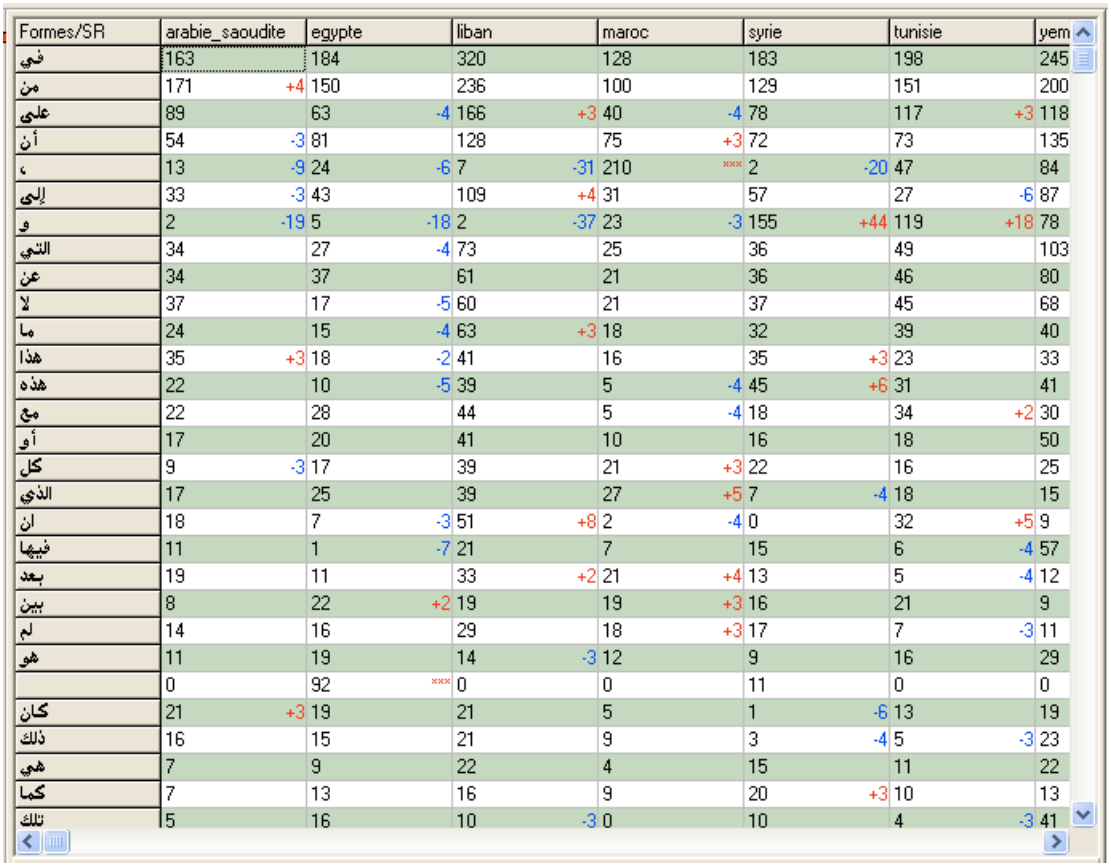
« abdekjalil_hafiz », deuxième colonne, qui est notre première partie) : le premier chiffre est négatif (-4 en bleu), suivi des spécificités positives en rouge. Le classement est donc par spécificités négatives à positives.



Si l'on reclique sur le nom de cette partie, les spécificités totales seront classées des spécificités positives (c'est-à-dire les suremplois d'une forme) aux spécificités négatives (les sous-emplois de cette forme) de cette première partie :



Regardons de plus près un tableau des spécificités totales :



Formes/SR	arabie_saoudite	egypte	liban	maroc	syrie	tunisie	yem
في	163	184	320	128	183	198	245
من	171 +4	150	236	100	129	151	200
على	89	63 -4	166 +3	40 -4	78	117 +3	118
أن	54 -3	81	128	75 +3	72	73	135
،	13 -9	24 -6	7 -31	210 ***	2 -20	47	84
إلى	33 -3	43	109 +4	31	57	27 -6	87
و	2 -19	5 -18	2 -37	23 -3	155 +44	119 +18	78
التي	34	27 -4	73	25	36	49	103
عن	34	37	61	21	36	46	80
لا	37	17 -5	60	21	37	45	68
ما	24	15 -4	63 +3	18	32	39	40
هذا	35 +3	18 -2	41	16	35 +3	23	33
هذه	22	10 -5	39	5 -4	45 +6	31	41
مع	22	28	44	5 -4	18	34 +2	30
أو	17	20	41	10	16	18	50
كل	9 -3	17	39	21 +3	22	16	25
الذي	17	25	39	27 +5	7 -4	18	15
ان	18	7 -3	51 +8	2 -4	0	32 +5	9
فيها	11	1 -7	21	7	15	6 -4	57
بعد	19	11	33 +2	21 +4	13	5 -4	12
بين	8	22 +2	19	19 +3	16	21	9
لم	14	16	29	18 +3	17	7 -3	11
هو	11	19	14 -3	12	9	16	29
	0	92 ***	0	0	11	0	0
كان	21 +3	19	21	5	1 -6	13	19
ذلك	16	15	21	9	3 -4	5 -3	23
هي	7	9	22	4	15	11	22
كما	7	13	16	9	20 +3	10	13
تلك	5	16	10 -3	0	10	4 -3	41

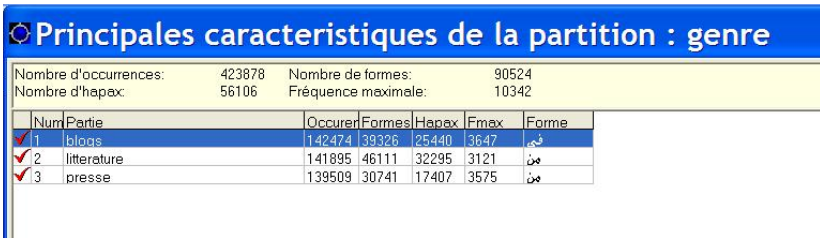
Ce tableau contient deux indications :
1) le signe + (suivi d'un nombre en rouge) ou le signe – (suivi d'un nombre en bleu) qui indiquent respectivement le sur-emploi ou le sous-emploi de la forme dans la ou les partie(s) sélectionnée(s) par rapport à l'ensemble du corpus.
2) les nombres en rouge ou en bleu sont des exposants qui rendent compte du degré de significativité de l'écart constaté.
Si l'on regarde la deuxième ligne de notre tableau (forme *min*), dans la partie arabie_saoudite (première colonne), on lit 171 +4. Cela signifie que la forme *min* apparaît 171 fois dans la partie arabie_saoudite, et qu'elle est plus fréquente que ce que laissait prévoir une répartition « au hasard ».

Concrètement, un exposant égal à *x*, indique que la probabilité d'un écart de répartition supérieur ou égal à celui que l'on a constaté était, au départ de l'ordre de 10^{-x}.

Remarque : Si le calcul des segments répétés a été préalablement effectué, les segments spécifiques apparaissent également dans la liste des unités spécifiques.

Calcul des spécificités

L'analyse des spécificités permet de porter un jugement sur la fréquence de chacune des unités textuelles dans chacune des parties du corpus.

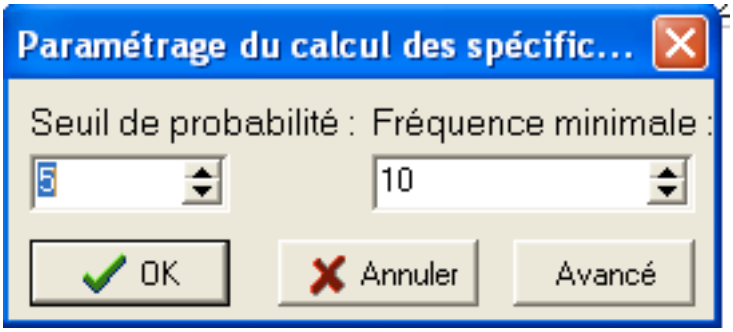


Allez dans les principales caractéristiques de la partition créée, puis sélectionnez en cliquant sur le nom de la partie celle pour laquelle vous voulez observer les spécificités. La ligne est surlignée en bleu. Pour sélectionner un ensemble de parties, il suffit de cliquer sur le nom d'une partie puis d'ajouter d'autres parties à celle déjà sélectionnée en appuyant simultanément sur la touche *Ctrl*.

Ensuite, cochez en cliquant à gauche du numéro des parties celles auxquelles vous voulez la comparer. Pour que cela fonctionne, il faut sélectionner au moins une partie (en cliquant sur le nom) et cocher au moins une autre. Vous décochez celles que vous ne voulez pas comparer.

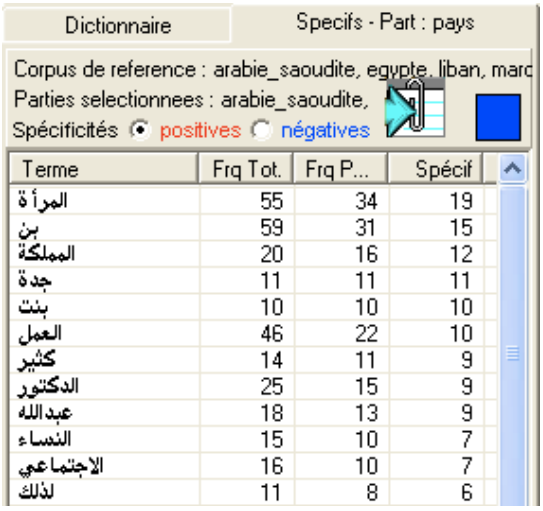
Une fois votre sélection effectuée, vous cliquez sur le bouton 

. Une fenêtre de paramétrage du calcul des spécificités s'ouvre :



Elle signifie que par défaut, l'indice de spécificité est calculé pour toutes les unités dont la fréquence est supérieure à 10, avec un *seuil de probabilité* fixé à 5 %.
Cliquez sur ok.

Il apparaît alors à gauche un tableau des spécificités en fonction de la partition choisie. Par exemple, si l'on opère une partition par pays, on obtient le tableau suivant à gauche (sous les onglets navigation, rapport, dictionnaires) :

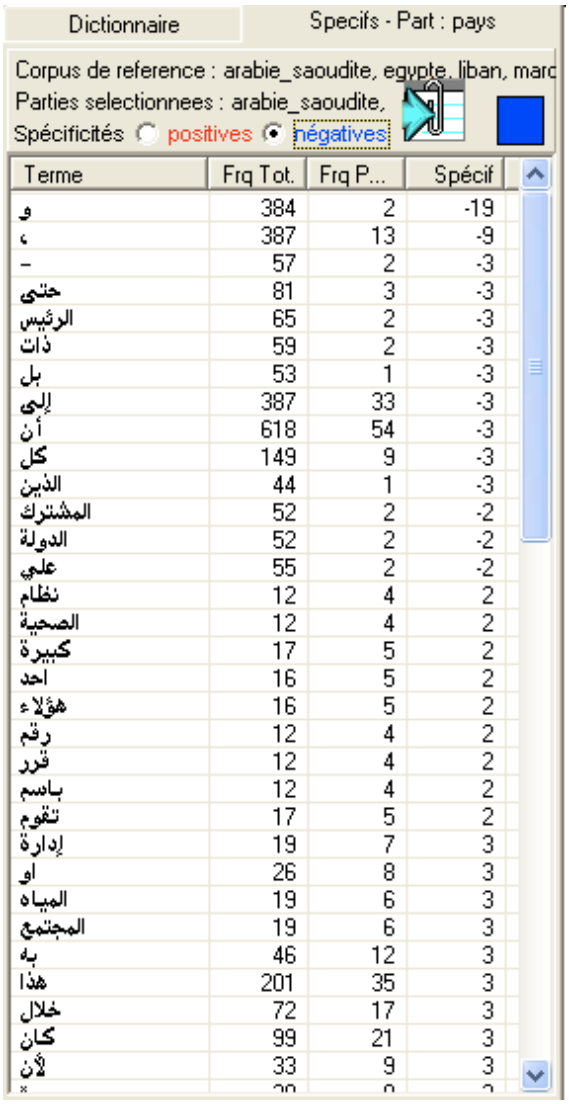


Terme	Frq Tot.	Frq P...	Spécif
المرأة	55	34	19
بن	59	31	15
المملكة	20	16	12
جدة	11	11	11
بنت	10	10	10
العمل	46	22	10
كثير	14	11	9
الدكتور	25	15	9
عبدالله	18	13	9
النساء	15	10	7
الاجتماعي	16	10	7
لذلك	11	8	6

Le tableau qui apparaît se nomme ***diagnostic de spécificité***. Dans la première colonne, on trouve les unités spécifiques classées par ordre décroissant de spécificité. Les colonnes suivantes varient selon la commande : soit elles indiquent la fréquence totale de la forme dans l'ensemble du corpus et la fréquence de la forme dans la partie sélectionnée, soit la fréquence de la forme dans les différentes parties sélectionnées. Les boutons ***positives*** et ***négatives*** de l'onglet des spécificités permettent d'inverser l'ordre de présentation de la liste qui s'ouvre par défaut sur les spécificités positives.

Pour chaque forme, on donne donc la fréquence totale dans le corpus, la fréquence dans la partie et les spécificités positives classées par ordre décroissant (c'est-à-dire la sur-apparition de cette forme dans cette partie par rapport à son apparition dans l'ensemble du corpus).

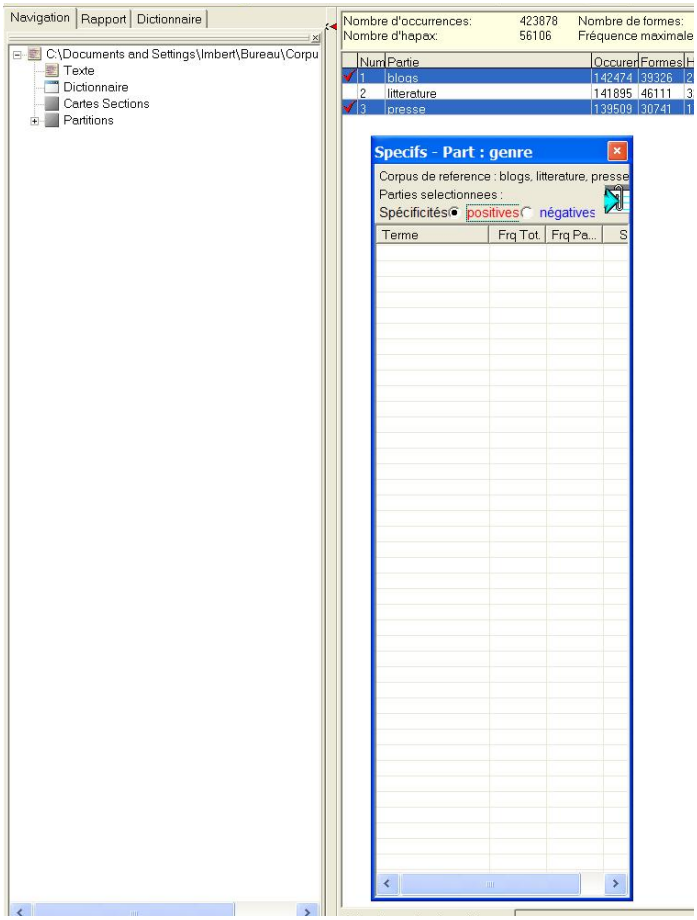
On peut aussi cocher « négatives ». On obtient alors le même tableau présentant les formes par spécificités négatives (toujours par ordre décroissant) :



Terme	Frq Tot.	Frq P...	Spécif
و	384	2	-19
،	387	13	-9
–	57	2	-3
حتى	81	3	-3
الرئيس	65	2	-3
ذات	59	2	-3
بل	53	1	-3
إلى	387	33	-3
أن	618	54	-3
كل	149	9	-3
الذين	44	1	-3
المشترك	52	2	-2
الدولة	52	2	-2
علي	55	2	-2
نظام	12	4	2
الصحية	12	4	2
كبيرة	17	5	2
احد	16	5	2
هؤلاء	16	5	2
رقم	12	4	2
قرر	12	4	2
باسم	12	4	2
تقوم	17	5	2
إدارة	19	7	3
او	26	8	3
المياه	19	6	3
المجتمع	19	6	3
به	46	12	3
هذا	201	35	3
خلال	72	17	3
كان	99	21	3
لأن	33	9	3

Supprimer des spécificités

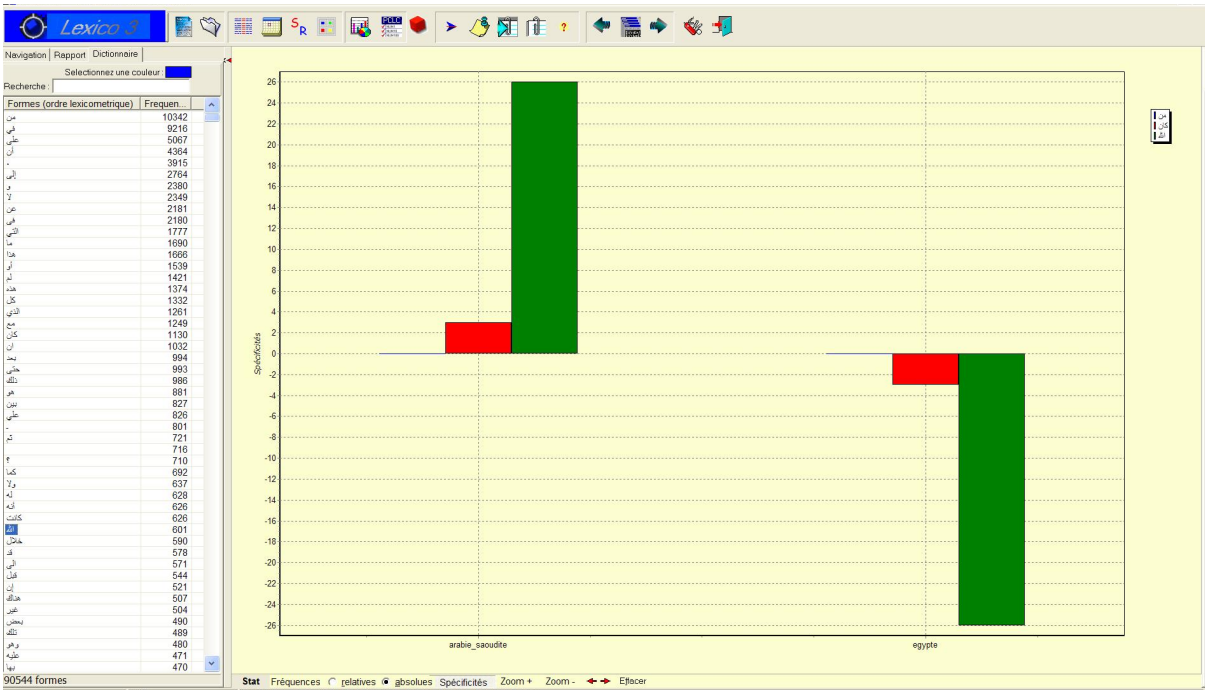
On a la possibilité de supprimer ces tableaux tout simplement. On fait glisser l'onglet dans la fenêtre des résultats. Il va s'afficher indépendamment. On peut ensuite le fermer en cliquant sur la croix rouge en haut à droite.



Diagrammes en barres des spécificités

On peut aussi accéder aux spécificités quand on crée une partition et que le graphique de ventilation s'ouvre. On glisse les formes depuis le dictionnaire, puis on peut cliquer dans la barre du bas sur « spécificités ». On a alors un diagramme en barres. Si on glisse une seconde forme, elle prend une autre couleur et la légende se fait automatiquement en haut à droite du diagramme.

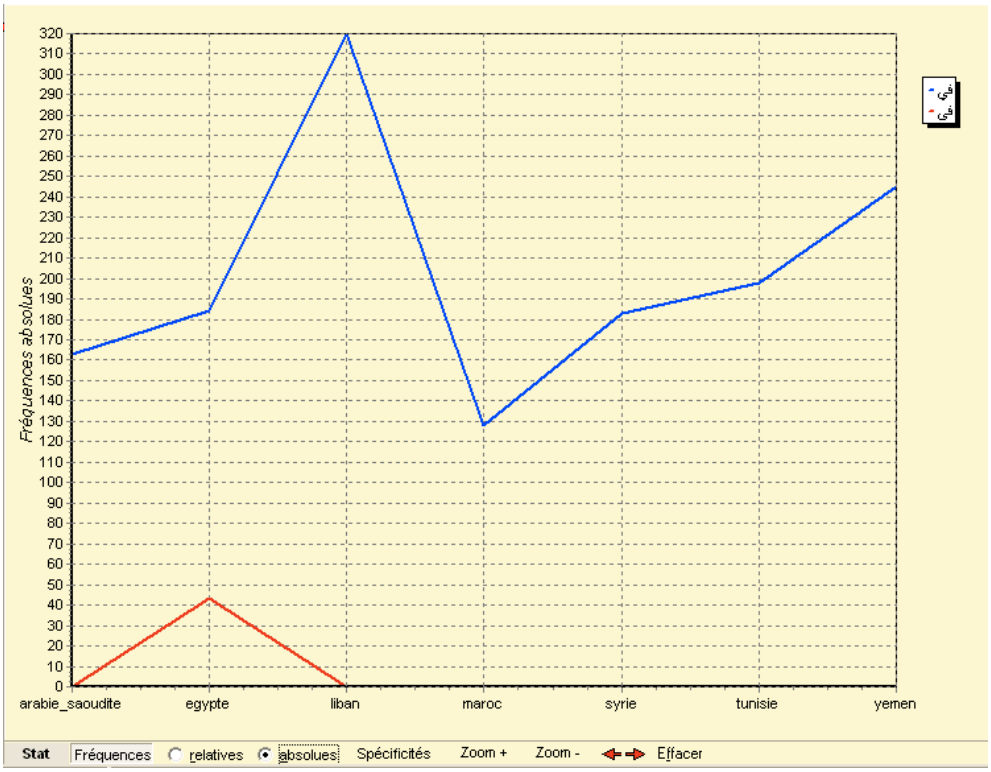
Ici, exemple des mots *min* (en bleu), *kāna* (en rouge) et *Allāh* (en vert).



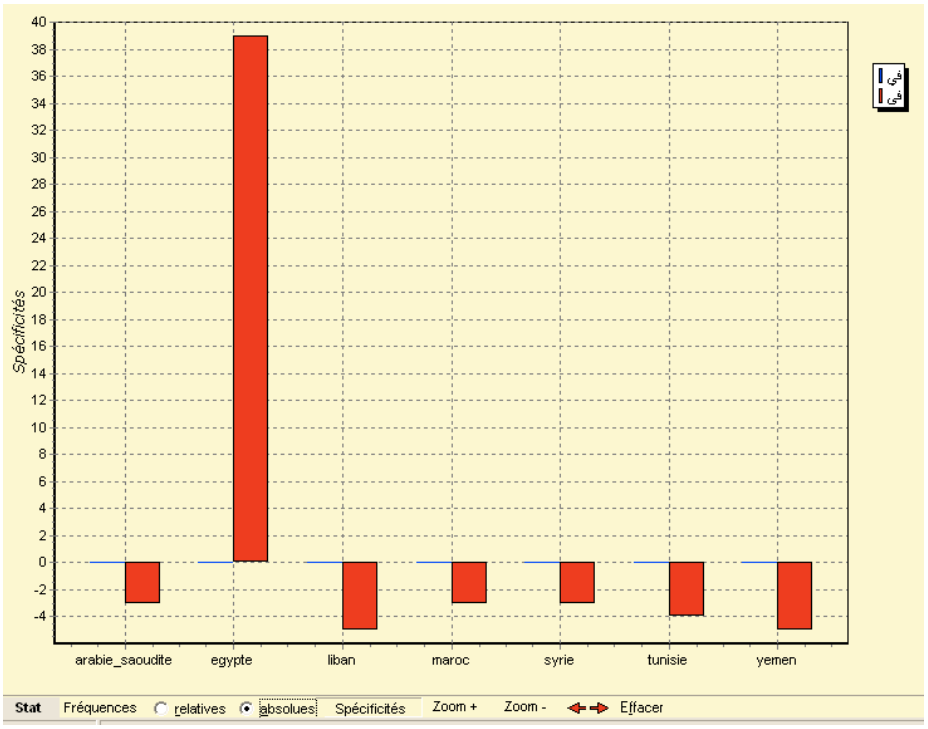
On comprend que la préposition *min* n'a pas d'emploi statistiquement original dans les deux parties. Par contre, le verbe *kāna* apparaît plus en Arabie Saoudite qu'en Égypte (d'où sur-emploi dans l'un et sous-emploi dans l'autre). Plus remarquable encore, la répartition du mot *Allāh* en Arabie Saoudite par rapport à l'Égypte.

Si l'on clique sur « effacer » en bas, on retrouve notre graphique vide prêt à recevoir de nouvelles formes.

Autre exemple à partir de la ventilation des formes *fī* et *fā* dans notre corpus :



Là, il s'agit des fréquences. Si l'on clique en bas sur « spécificités », on obtient le diagramme en barres suivant :



Comme *fā* n'apparaît qu'en Égypte, par comparaison cette forme est sous-représentée dans les autres pays. En revanche, *fī* est distribué de manière statistiquement égale dans l'ensemble

des parties du corpus.

Si l'on se penche sur un tableau des spécificités totales (on a vu tout à l'heure comment l'obtenir), *fī* ne présente aucune originalité mais *fā* une spécificité positive en Égypte.

Formes/SR	arabie_saoudite	egypte	liban	maroc	syrie	tunisie	yemen
فيا	0	1	0	0	0	0	0
في25	1	0	0	0	0	0	0
في24	0	1	0	0	0	0	0
في22	0	2	0	0	0	0	0
في2014	0	1	0	0	0	0	0
في15	0	1	0	0	0	0	0
في	163	184	320	128	183	198	245
فى	0	43 +39	0	0	0	0	0
فولت	0	0	0	0	0	1	0

IV. 2. 1. e. Calculs pour les séries textuelles chronologiques

Si vous utilisez une série de textes produits par la même source sur un certain laps de temps, vous pouvez utiliser deux autres outils : les spécificités chronologiques et les accroissements spécifiques.

Spécificités chronologiques

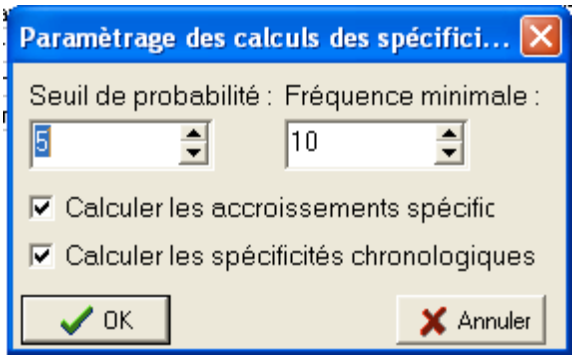
Elles concernent les *séries textuelles chronologiques*, à savoir une série de textes produits par une même source textuelle et régulièrement espacés dans le temps. À côté de l'analyse des spécificités de chacune des parties du corpus, l'analyse des spécificités chronologiques met en évidence le vocabulaire particulier de périodes plus larges formées de parties consécutives.

Accroissements spécifiques

Pour une partie sélectionnée, le bouton ***SpEvol***, permet de calculer les spécificités (ou ***accroissements spécifiques***) de cette partie par rapport à l'ensemble des périodes précédentes (en excluant momentanément du corpus les périodes postérieures). Le résultat de ces calculs est fourni sous la forme d'un tableau de spécificités identique à celui présenté plus haut.

Remarque : La partie négative des accroissements spécifiques met en évidence des unités textuelles qui ont tendance à être sous-utilisées dans la période considérée par rapport aux périodes qui précèdent.

Paramétrage du calcul des spécificités diachroniques :



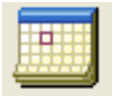
Analyse factorielle des correspondances, histogrammes des valeurs propres, diagramme de Pareto, courbes d'accroissement du vocabulaire : Si vous en arrivez là, vous n'avez certainement plus besoin de moi. Reportez-vous au manuel en ligne.

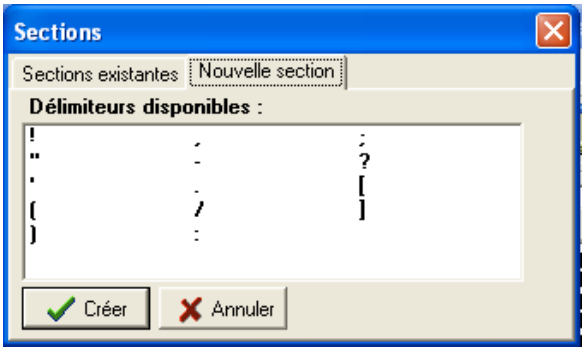
IV. 3. Outils de navigation lexicométrique

Les outils de navigation permettent de se déplacer parmi les résultats produits par les différentes méthodes lexicométriques et dans le corpus initial.

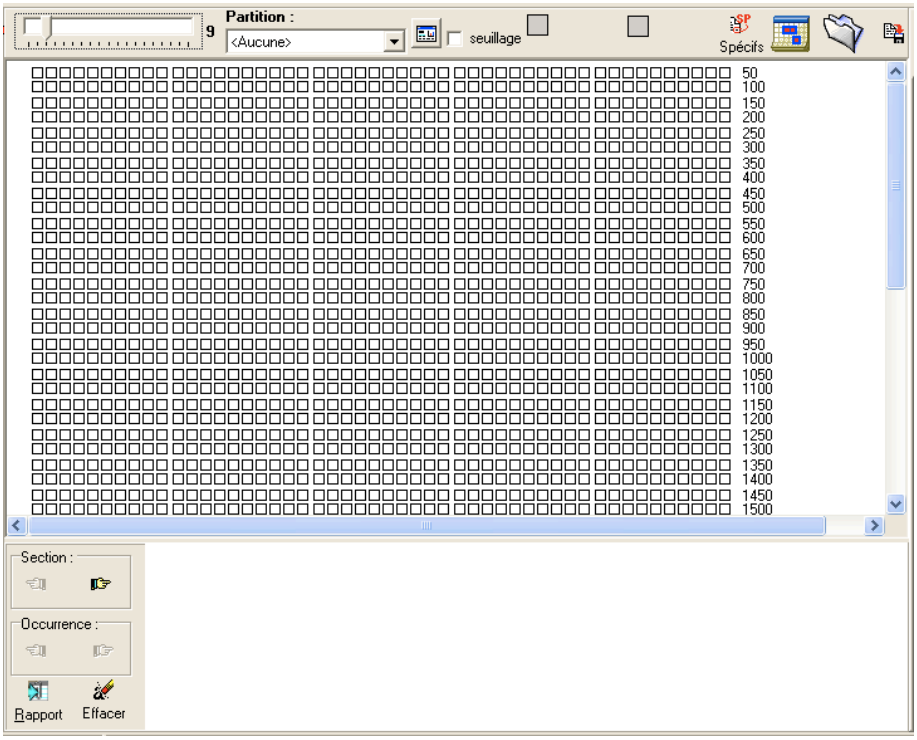
Carte des sections

La carte des sections présente le corpus découpé en sections. Il faut auparavant choisir un caractère particulier comme ayant le statut de délimiteur de section (on peut par exemple insérer automatiquement le caractère § dans le corpus avant la segmentation).

On clique sur l'icône « carte des sections » :  . Une fenêtre s'ouvre où il s'agit de choisir le délimiteur de section. Vous choisissez votre délimiteur et vous cliquez sur « créer ».



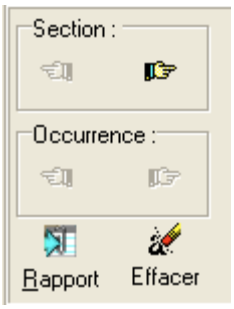
Par exemple, je choisis le point et j'obtiens la carte des sections suivante (où un carré représente une phrase puisque mon délimiteur de section est le point) :



Les nombres en colonne à droite des sections indiquent le nombre de sections. Il est possible de zoomer (curseur en haut à gauche).

Naviguer dans la carte des sections

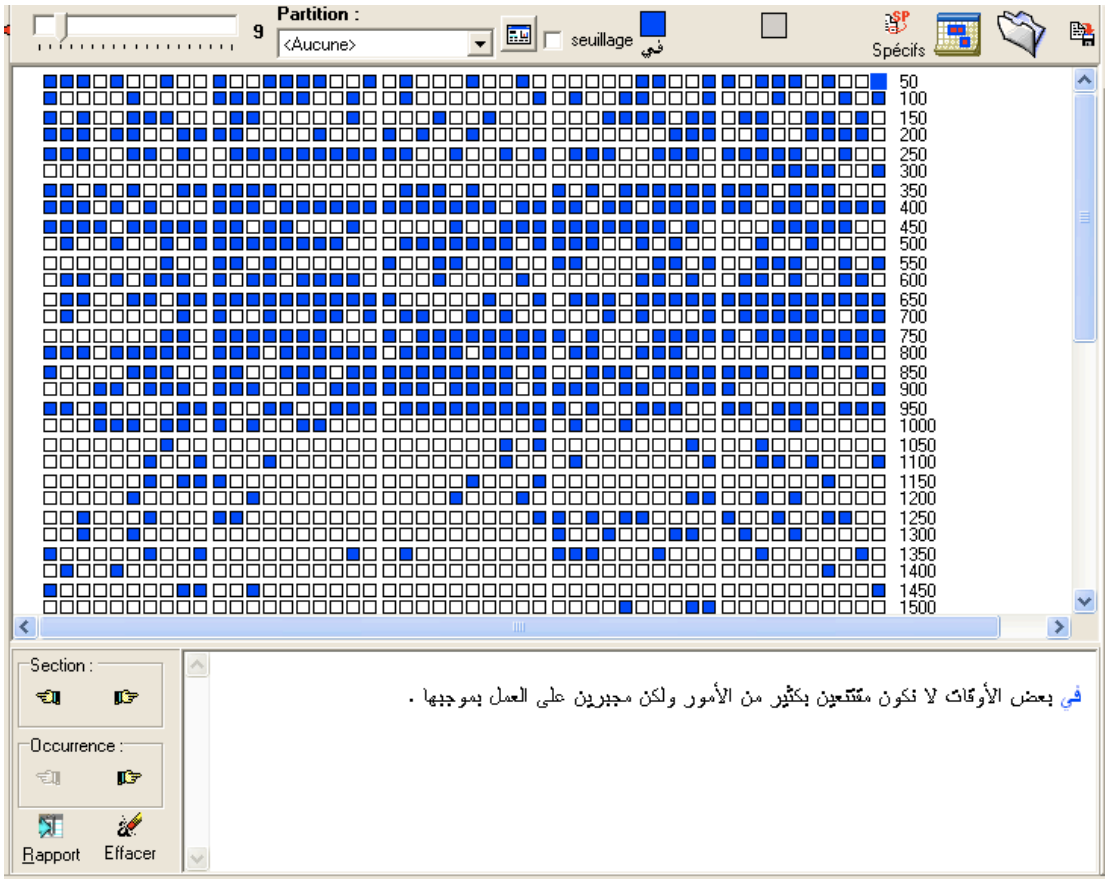
En bas à droite apparaît le bloc suivant :



Il permet de naviguer dans la carte des sections, soit d'une section à une autre, soit d'une occurrence à une autre (en cliquant sur la main). Il permet aussi d'effacer la carte obtenue ou d'enregistrer les résultats dans le rapport (on ajoute au rapport la section visualisée dans la fenêtre du bas).

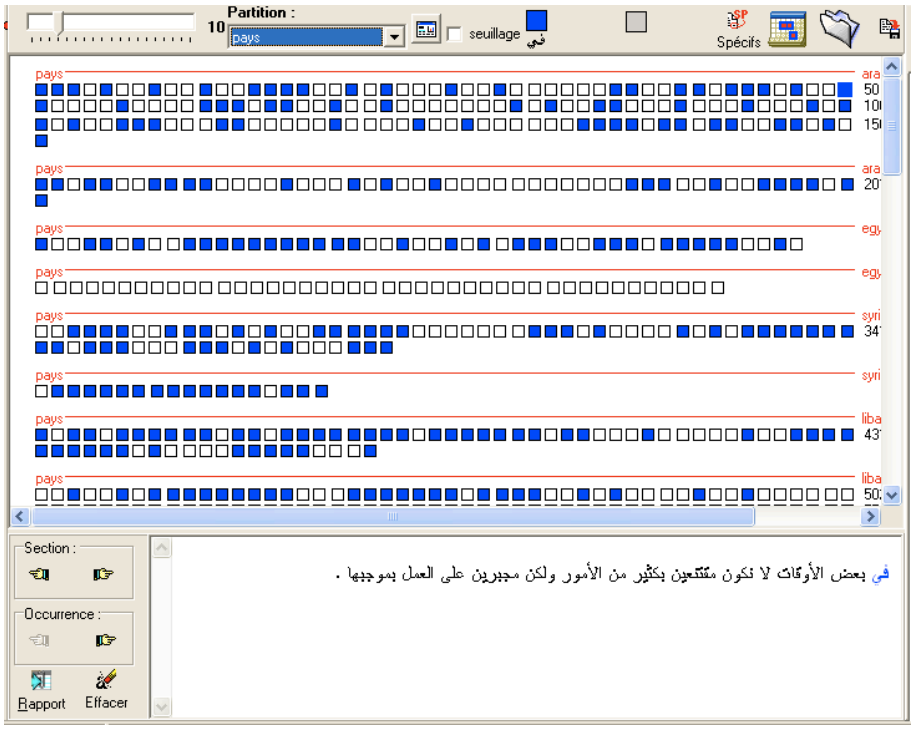
Pour faire une carte des sections d'une occurrence, d'un segment répété ou d'un groupe de formes (type généralisé), il suffit de sélectionner l'objet en question (à partir du dictionnaire, du garde-mots, de la liste des segments répétés, *etc*.) et de le faire glisser sur la carte (on maintient le clic gauche de la souris enfoncé).

Par exemple, si je glisse la forme *fɪ̄*, j'obtiens la carte des sections suivante :



Quand on veut visualiser dans la fenêtre du bas à droite la section, il suffit de cliquer sur le carré qui la représente. Dans cet exemple, on peut observer la cinquantième section. Dans la carte des sections, on voit que le tout premier carré en haut à droite est sélectionné (un gros carré bleu ciel, à ne pas confondre avec le seuillage qu'on détaille plus bas).

On peut activer une partition déjà créée en la sélectionnant dans la liste située en haut, à droite du curseur. Par exemple, si je sélectionne la partition « pays », j'obtiens la carte des sections suivantes :



La même carte des sections pour la forme *fā* montre clairement que cette forme n'est présente que dans une partie du corpus, en Égypte :

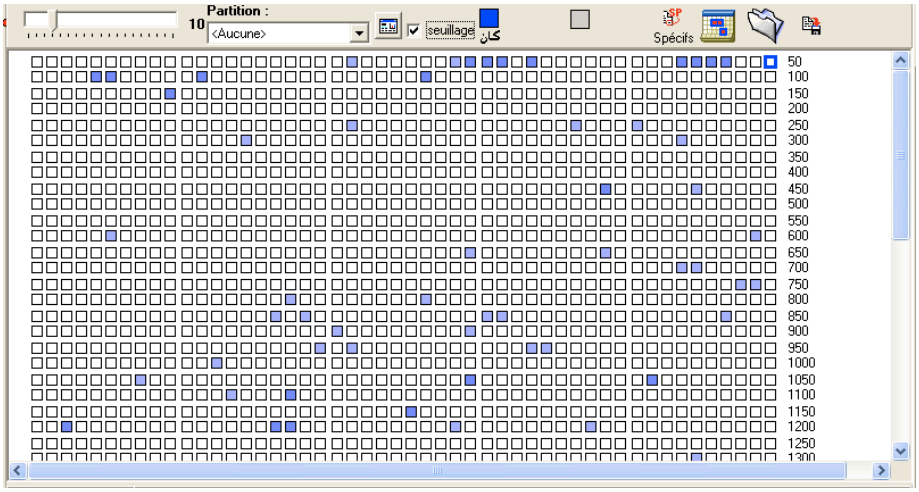
Seuillage

Le seuillage indique, par un jeu de couleur, les spécificités d'une forme donnée au sein de chaque section. La couleur devient plus dense en fonction de l'accroissement de la spécificité

de la forme dans la section.

Pour ce faire, on coche la case « seuillage ». L'icône qui précède immédiatement à gauche le coche permet de régler deux seuils en probabilités qui entraîneront un coloriage (plus ou moins sombre) des sections. Paramétrage du seuillage :

Voici le seuillage de la forme *kāna* dans mon corpus :



Plus la section est foncée, plus les spécificités sont importantes pour cette forme dans la section en question.

Pour une représentation simultanée de deux *Tgens*, ce processus peut être réitéré. Il ne faut pas oublier de changer la couleur dans la boîte correspondante (dans la fenêtre de paramétrage du seuillage). Il faut maintenir, dans ce cas, la touche *Ctrl* appuyée lors du second glisser/déposer.

Outils statistiques de la carte des sections

En haut à droite de la carte des sections figurent deux icônes situées qui permettent de repérer les types caractéristiques d'un ensemble de sections (les spécificités des sections sélectionnées, *cf.* le manuel d'utilisation).

- l'icône *Cooccurences* constitue automatiquement une sélection des sections dans lesquelles le groupe de formes (ou *Tgen*) étudié est présent. C'est cet ensemble de sections que l'on compare à l'ensemble du corpus.
- l'icône *Spécificités* permet à l'utilisateur de constituer une sélection arbitraire de sections dont on étudiera ensuite le vocabulaire spécifique (on sélectionne les sections une à une en maintenant le bouton *Ctrl* enfoncé ; la touche majuscule permet de sélectionner un groupe de sections consécutives).

Comme toujours, les listes de spécificités sont affichées dans la fenêtre de gauche. Le nombre des sections concernées par la sélection apparaît en haut de la fenêtre ; un bouton *ajouter au rapport Section* placé en bas de la fenêtre permet de sauvegarder les résultats.

V. Le rapport

Tout au long des recherches effectuées à partir du logiciel, il est possible de conserver les résultats dans un rapport sauvegardé au format htm.

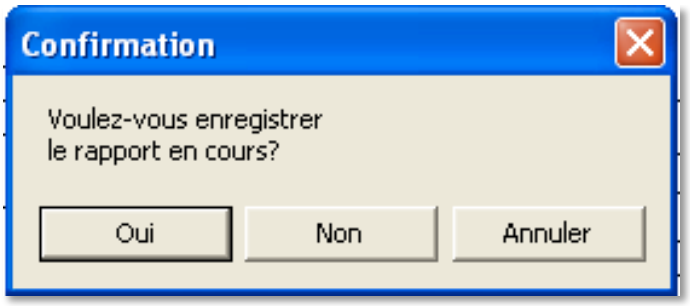
V. 1. Présentation et consultation du rapport

Le dossier ***rapport*** figure toujours dans la colonne de gauche du logiciel, entre *fichier* et *dictionnaire* :

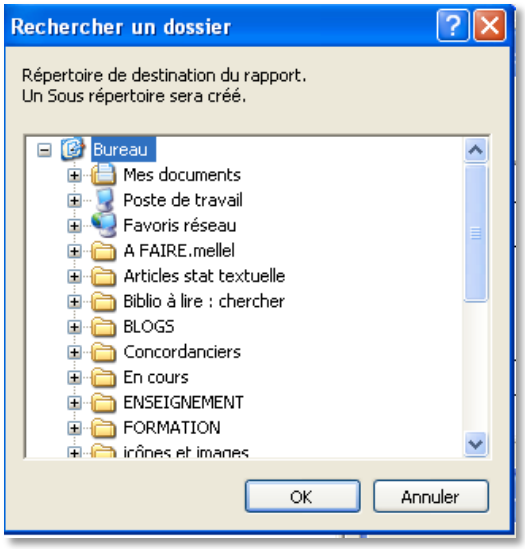
Lorsqu'il est vide, le rapport n'indique que le titre *Éléments du rapport*, comme ci-dessus.

Quand on consigne une recherche dans le rapport, elle apparaît sous *Éléments du rapport*. Il est possible de modifier le titre et d'ajouter des mémos (en cliquant sur l'icône précédent le titre). Ces mémos pourront être ouverts lorsque l'on consultera le rapport ultérieurement.

Deux commandes sont disponibles à partir de cet onglet : *enregistrer*, pour que les informations consignées dans le rapport soient sauvegardées, et *nouveau*, pour créer un nouveau rapport. Si vous cliquez sur *nouveau*, une fenêtre de confirmation apparaîtra :

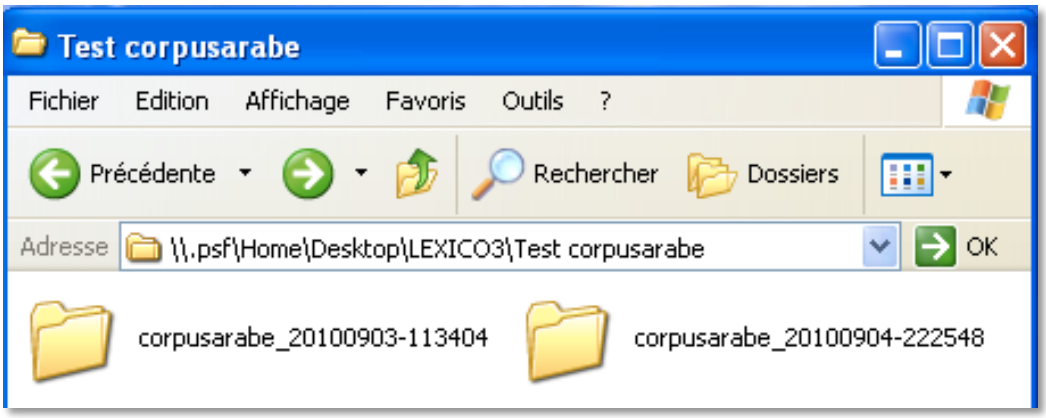


Si vous voulez sauvegarder le rapport en cours, cliquez sur *oui* : la même fenêtre que lorsque vous cliquez directement sur *enregistrer* s'affiche :



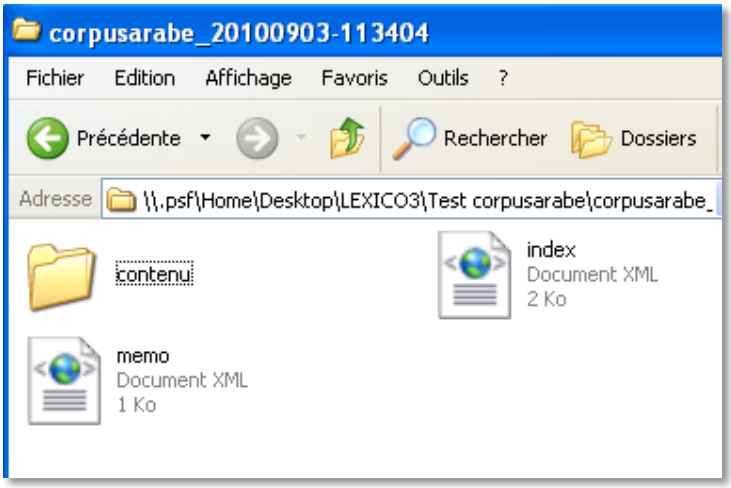
Vous pouvez spécifier l'endroit où le rapport sera enregistré, mais par défaut, il figurera dans le dossier *Rapport* à l'intérieur du dossier *Lexico3* créé par l'installation du logiciel.

Comme précisé dans le message, un sous-répertoire sera créé, avec pour titre le nom du texte source et une série de chiffres :



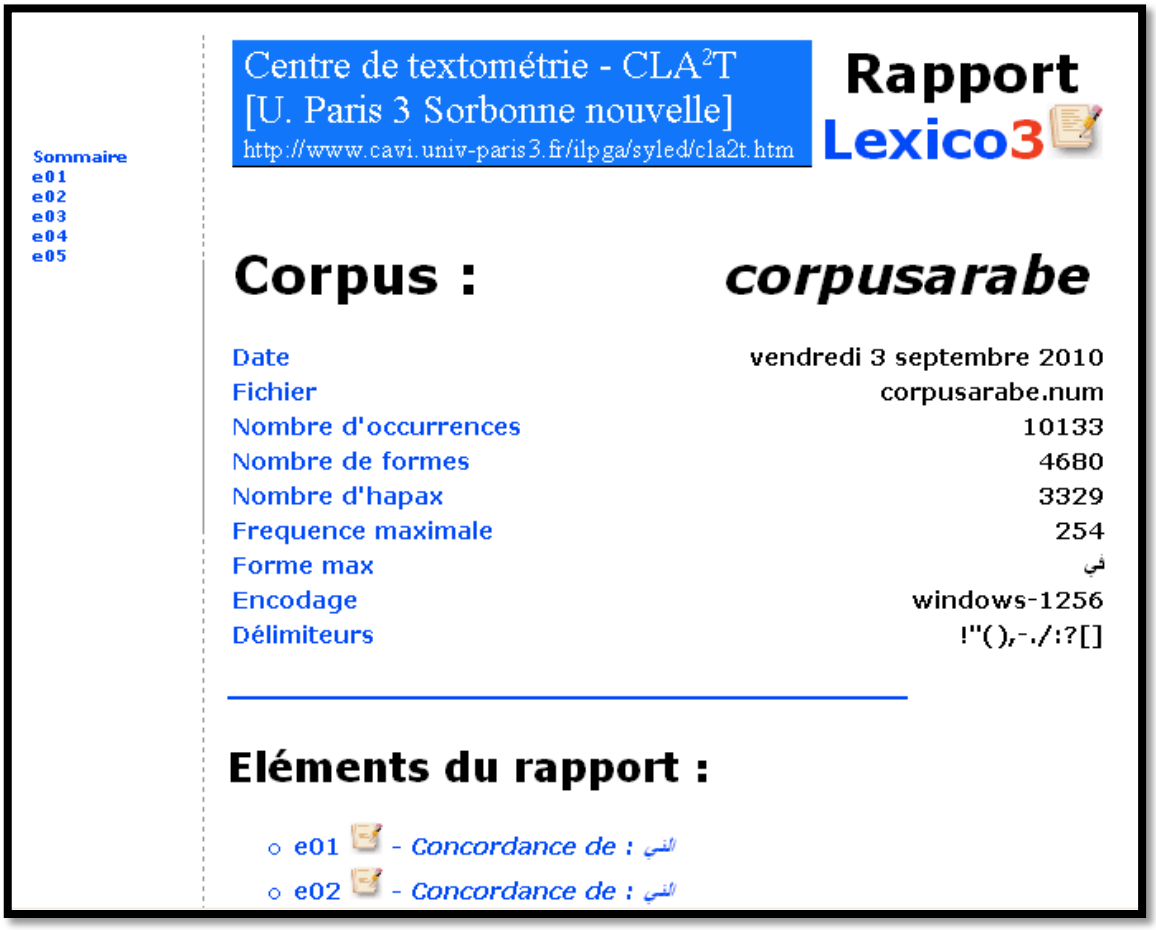
Ici, dans le dossier intitulé *Test corpusarabe* où je conserve tout ce qui se rapporte au petit corpus d'essai élaboré pour servir d'exemple à ce guide, deux rapports ont été conservés.

Lorsqu'on ouvre un rapport, on observe trois éléments :



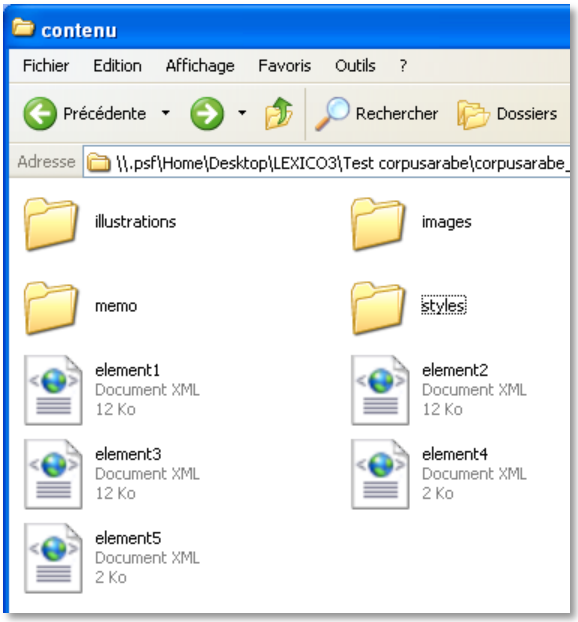
- un dossier *contenu*, qui contient comme son nom l'indique tout ce qui est consigné dans le rapport
- une page *memo*, sorte de résumé des éléments contenus dans le rapport
- une page *htm* intitulée *index*, qui fournit l'index du rapport. S'ouvrant à l'aide d'un navigateur web (Internet Explorer, Netscape, *etc.*), il permet la navigation parmi les résultats.

Lorsqu'on ouvre l'index, on arrive sur une page comme la suivante :



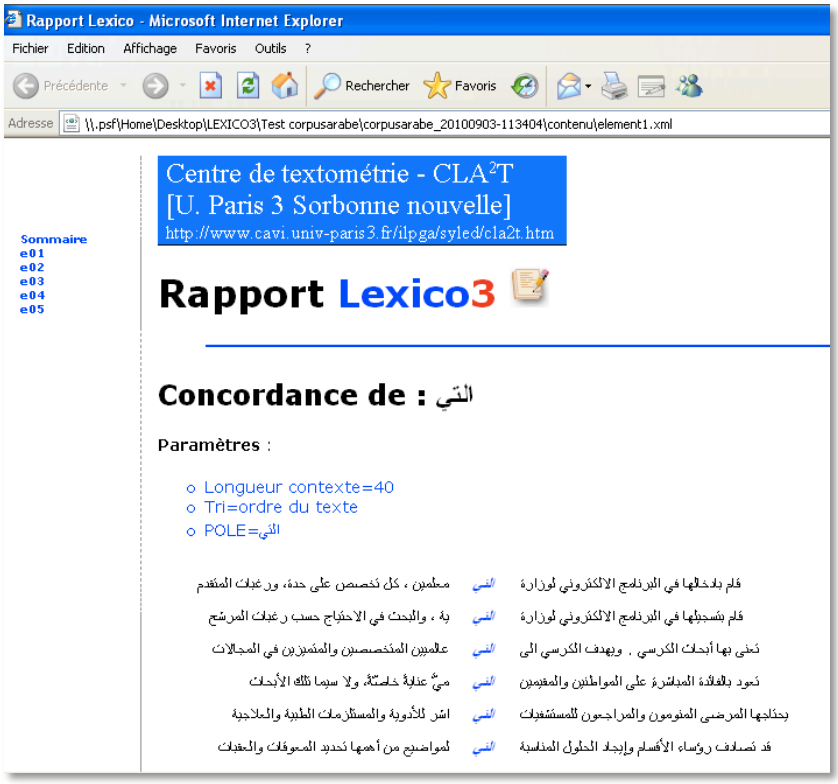
L'index mentionne le laboratoire où a été développé Lexico3, à côté du titre **Rapport**. Au dessous, **corpus** indique le texte source utilisé (ici, notre fichier test *corpusarabe*). Des indications générales figurent (date du rapport, fichier source, statistique générale, encodage et délimiteurs). En dessous, les **Éléments du rapport** listent les différents ajouts qui ont été faits au rapport. À gauche, le sommaire permet d'accéder directement à ces différents éléments. Il suffit de cliquer sur un élément pour y accéder.

Dans le dossier *contenu*, on va trouver différents dossiers générés automatiquement ainsi que chaque élément ajouté au rapport enregistré en fichier indépendant (*element1, element2, etc.*) :



Il est possible de renommer les éléments de façon plus détaillée.

Quand on ouvre un élément, par exemple ici *element1*, une concordance, on retrouve la même présentation que dans le fichier *index* : à gauche, le sommaire permet d'accéder aux autres éléments (même s'ils n'ont pas été ouverts au préalable), le titre en gras rappelle la recherche effectuée, les paramètres en sont donnés et les résultats suivent.



Il est tout à fait possible de copier les résultats et de les coller dans un éditeur de texte (dans l'exemple de la concordance, on aura un tableau en trois colonnes).

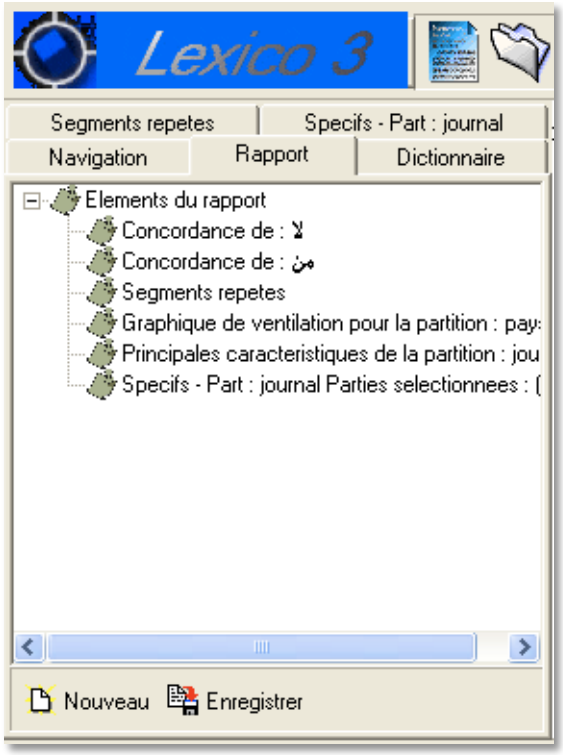
Attention : si les concordances s'affichent dans le sens arabe dans le logiciel (contexte avant l'occurrence dans la colonne de droite et contexte après l'occurrence dans la colonne de gauche), lors du passage au rapport une inversion s'opère : le contexte avant se retrouve dans la colonne de gauche et le contexte après dans la colonne de droite. Il faut rétablir manuellement.

Le rapport peut être consulté à tout moment, à condition que l'utilisateur l'ait préalablement enregistré (bouton Enregistrer au bas de l'onglet Rapport, comme présenté ci-dessus).

V. 2. Ajouter des éléments au rapport : sauvegarder les résultats de ses recherches

Pour ajouter un résultat au rapport, il suffit de cliquer sur l'icône *Ajouter au rapport* qui figure dans la barre des outils :

Il est possible de suivre les ajouts en cliquant sur l'onglet *rapport* dans la colonne de gauche :



Ceci permet d'accéder rapidement à des résultats préalablement enregistrés et permet de se souvenir des recherches effectuées.

Dans certains cas, il faut utiliser la même icône qui figure dans la colonne de gauche, sous l'onglet de la recherche en cours. C'est le cas par exemple des segments répétés :

Tout comme on consulte les résultats de la recherche des segments répétés en cliquant sur

l'onglet *segments répétés* situé dans la colonne gauche, on clique sur l'icône *ajouter au rapport* qui se trouve dans la fenêtre des résultats plutôt que sur celui se trouvant dans la barre des outils.

V. 3. Éditer les résultats

Il est possible de visualiser directement les résultats enregistrés au cours d'une recherche. Tout rapport enregistré est consultable par l'éditeur. Il suffit pour cela de cliquer sur l'icône *éditeur* : 

Si vous n'avez rien sauvegardé, un message d'erreur vous en informera :

Nous **conseillons de renommer systématiquement** les rapports ainsi que les éléments de façon détaillée. Ainsi, il sera beaucoup plus aisé de consulter les résultats pour un travail ultérieur.

V. 4. La feuille

La feuille sur laquelle on travaille est appelée *nouvelle feuille*, les autres sont énumérées. Il suffit de cliquer sur la feuille désirée pour y basculer.

Cette étape peut être faite directement en cliquant sur l'onglet correspondant à la feuille que l'on souhaite ouvrir, à droite toute de la fenêtre.

Feuille 1
Feuille 2
Feuille 3

Glossaire

accroissement spécifique : spécificité calculée pour une partie d'un corpus par rapport à une partie antérieure.

analyse factorielle : famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des « facteurs » résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

analyse des correspondances : méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou c2).

balise : clé qui permet d'entrer des méta-données dans le corpus. Elle a la forme suivante : <type=contenu> et permet d'opérer des partitions du corpus.

caractère : signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

caractères délimiteurs / **non-délimiteurs** : distinction opérée sur l'ensemble des caractères qui entrent dans la composition du texte, permettant aux procédures informatisées de segmenter le texte en occurrences (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs :

- les caractères **délimiteurs d'occurrence**s (encore appelés **délimiteurs de forme**) qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.

- les caractères **délimiteurs de séquences** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.

- les caractères **séparateurs de phrase** : sous-ensemble des délimiteurs de séquence qui correspondent, en général, aux seules ponctuations fortes.

concordance : l'ensemble de lignes de contexte se rapportant à une même forme-pôle. La concordance est une liste alignée (sous forme de trois colonnes distinctes) de toutes les occurrences d'une même forme.

cooccurrence : présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, *etc*.) des occurrences de deux formes données.

corpus (ling) : ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique. (lexicométrie) : ensemble de textes réunis à des fins de comparaison servant de base à une étude quantitative.

clé : v. **balise.**

facteur : variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.

forme ou **forme graphique** : archétype correspondant aux occurrences identiques dans un

corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

forme banale : pour une partie du corpus donnée, forme ne présentant aucune spécificité (ni positive ni négative) dans cette partie.

forme caractéristique (d'une partie) : synonyme de spécificité positive.

forme commune : forme attestée dans chacune des parties du corpus.

forme originale (pour une partie du corpus) : forme trouvant toutes ses occurrences dans cette seule partie.

fréquence (d'une unité textuelle) : le nombre de ses occurrences dans le corpus.

fréquence d'un segment (ou d'une polyforme) : le nombre des occurrences de ce segment, dans l'ensemble du corpus.

fréquence maximale : fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition « de »).

fréquence relative : la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

hapax : signifie « chose dite une seule fois ». C'est une forme dont la fréquence est égale à 1 dans le corpus (*hapax* du corpus) ou dans une de ses parties (*hapax* de la partie).

lexicométrie : ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes.

longueur (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, *etc*.) : le nombre des occurrences contenues dans ce corpus (respectivement : partie, fragment, *etc*.). Synonyme : *taille*. On note : T la longueur du corpus, t j celle de la partie (ou tranche) numéro j du corpus.

occurrence : suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs de forme.

ordre lexicographique

1. pour les formes graphiques : l'ordre selon lequel les formes sont classées dans un dictionnaire (ordre alphabétique)

NB : Les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple rangées dans cet ordre les formes : *mais, maïs, maison, maître*.

2. pour les polyformes : ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante. Les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, *etc*.

ordre lexicométrique

1. pour les formes graphiques : ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes ; les formes de même fréquence sont classées par ordre lexicographique.

2. pour les polyformes : ordre résultant d'un tri par ordre de longueur décroissante des

segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.

partie (d'un corpus de textes) : fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition

1. d'un corpus de textes : division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

2. d'un ensemble, d'un échantillon : division d'un ensemble d'individus ou d'observations en *classes* disjointes dont la réunion est égale à l'ensemble tout entier.

périodisation : regroupement des parties naturelles du corpus respectant l'ordre chronologique d'écriture, d'édition ou de parution des textes réunis dans le corpus.

polyforme : archétype des occurrences d'un segment ; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.

répartition (des occurrences d'une forme dans les parties du corpus) : nombre des parties du corpus dans lesquelles cette forme est attestée.

section : portion de texte comprise entre deux délimiteurs de section.

segment : toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur de séquence est un segment du texte.

segment répété (ou polyforme répétée) : suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentation : opération qui consiste à délimiter des unités minimales dans un texte.

segmentation automatique : ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales.

séquence : suite d'occurrences du texte non séparées par un délimiteur de séquence.

seuil : quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, *etc*.).

sous-fréquence (d'une unité textuelle dans une partie, tranche, *etc*.) : nombre des occurrences de cette unité dans la seule partie du corpus.

spécificité chronologique : spécificité portant sur un groupe connexe de parties d'un corpus muni d'une partition longitudinale.

spécificité positive : pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

spécificité négative : pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée

est inférieure au seuil fixé au départ.

tableau des segments répétés (TSR) : tableau à double entrée dont les lignes sont constituées par les ventilations des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique des segments (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).

taille d'un corpus : sa longueur mesurée en occurrences (de formes simples).

type : un type rassemble les occurrences de formes graphiques différentes liées par une propriété commune que détermine l'objectif de la recherche : l'utilisation d'un mot (au singulier et au pluriel), celle d'un verbe (quelle que soit sa conjugaison), l'emploi de synonymes (lien sémantique), *etc*.).

types généralisés (Tgens) : unités de dépouillement définies par l'utilisateur à l'aide d'outils permettant d'effectuer automatiquement des regroupements d'occurrences du texte (ex : les occurrences des formes qui commencent par la séquence de caractère *patr : patrie, patriotes, patriotisme, etc.*).

ventilation des occurrences d'une unité dans les parties du corpus : La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire : ensemble des formes attestées dans un corpus de textes.

Quelques outils de textométrie en ligne

Concordanciers

AConCorde : http://www.andy-roberts.net/coding/aconcorde
Qamus (Tim Buckwalter) : http://www.qamus.org/

Logiciels, analyseurs morpho-syntaxiques, extracteurs de racines

Lexico3 : http://www.tal.univ-paris3.fr/lexico/
mkAlign (aligneur de corpus parallèles) : http://www.tal.univ-paris3.fr/mkAlign/
Qamus (différents outils) : http://www.qamus.org/
Arabic Treebank (génération d'arbres) : http://www.ircs.upenn.edu/arabic/
Kawâkib (différents outils) : http://www.ifao.egnet.net/kawakib/
Extracteur de racines : http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py

Corpus

Arabicorpus (Dilworth Parkinson) : http://arabicorpus.byu.edu/index.php
Qamus (Tim Buckwalter) : http://www.qamus.org/
An-Nahar news paper : http://catalog.elra.info/product_info.php?products_id=767
Sources diverses : https://www.ldc.upenn.edu/
Coran : http://corpus.quran.com/contact.jsp
Al-maktaba al-Šāmila : http://shamela.ws/

Éditeurs de gros corpus

BabelPad : http://www.babelstone.co.uk/Software/BabelPad.html
Notepad++ : http://notepad-plus-plus.org/release/5.9

Mais aussi...

TEI (*text encoding initiative*), codage des corpus : http://www.tei-c.org/index.xml Version simplifiée : http://www.tei-c.org/Vault/P4/Lite/teiu5_fr.html
Automates arabes : http://www.ifao.egnet.net/axes/ecritures-langues/tal-arabe/automatesarabes/
http://www.comp.leeds.ac.uk/eric/latifa/ArabicCorporWebConc.htm

Bibliographie indicative

BENZECRI Jean-Paul et *alii*. (1981) : *Pratique de l'analyse des données 3. Linguistique et lexicologie*. Paris, Dunod, 565 p.

GUIRAUD Pierre (1960) : *Problèmes et méthodes de la statistique linguistique*. Paris, PUF, 145 p.

HABASH Nizar (2010) : *Introduction to Arabic Natural Language Processing*. Coll. Synthesis Lectures on Human Language Technologies #10, Morgan & Claypool publishers, 167 p.

HABERT Benoît, **NAZARENKO** Adeline et **SALEM** André (1997) : *Les linguistiques de corpus*. Paris, Armand Colin, 240 p.

HABERT Benoît, **FABRE** Cécile et **ISAAC** Fabrice (1998) : *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. Paris, InterEditions, Masson, 320 p.

HAMDANI Abdelfattah, **LACHHAB** Khalid et **ERRADI** Mohammed (éds.) (2007) : *Traitement automatique de la langue arabe / Arabic Language Processing, Actes du colloque Proceedings, juin 2006*, Université Mohammed V - Souissi. Institut d'Études et de Recherches pour l'Arabisation, 335 p + 103 p.

KOULOUGHLI Djamel (2004) : « Initiation pratique à la constitution et à l'exploitation de corpus électroniques en langue arabe (1ère partie)», *in* Langues et Littératures du Monde Arabe 5, pp. 231-293. [En ligne] http://w3.ens-lsh.fr/llma/

KOULOUGHLI Djamel (2007) : « Initiation pratique à la constitution et à l'exploitation de corpus électroniques en langue arabe (2ème partie)», *in* Langues et Littératures du Monde Arabe 6, pp. 97-114. [En ligne] http://w3.ens-lsh.fr/llma/

KOULOUGHLI Djamel (2008) : « Initiation pratique à la constitution et à l'exploitation de corpus électroniques en langue arabe (3ème partie)», *in* Langues et Littératures du Monde Arabe 7, pp. 75-93. [En ligne] http://w3.ens-lsh.fr/llma/

KOULOUGHLI Djamel (2009) : « Initiation pratique à la constitution et à

l'exploitation de corpus électroniques en langue arabe (4ème partie)», *in* Langues et Littératures du Monde Arabe 8, pp. 117-133. [En ligne] http://w3.ens-lsh.fr/llma/

LEBART Ludovic et **SALEM** André (1994) : *Statistique textuelle*. Paris, Dunod, 335 p. [En ligne] http://www.tal.univ-paris3.fr/lexico/lectures.htm

MÜLLER Charles (1968) : *Initiation à la statistique linguistique*. Paris, Larousse, collection langue et langage, 247 p.

MÜLLER Charles (1977) : *Principes et méthodes de statistique lexicale*. Paris, Classiques Hachette, 205 p.

MÜLLER Charles (1993) : *Initiation aux méthodes de la statistique linguistique*. Paris, Champion, 185 p.