Première partie

CORPUS, DÉPOUILLEMENT ET PHASES DE L'ANALYSE

Chapitre 1

Constitution et dépouillement du corpus

1. ÞAbû Íayyân at-TawÎîdî : Chronologie[61]

Homme de lettres et philosophe du IV^{ème}/X^{ème} siècle, *ÞAbû Íayyân at-TawÎdî*, *ÝAlî b.* M*u*Îa*m*m*ad b.* a*l*-*ÝAbbâ*s, dont la vie est mal connue, est probablement appelé *al-TawÎîdî* en raison du métier qu'exerçait son père en tant que marchand d'une sorte de dattes appelées *tawÎîd.* Sans aucune certitude, le lieu de sa naissance est donné comme *Nîšâpûr, Šîrâz, WâsiÔ ou BaÈdâd*, sa date de naissance devait se situer entre 310-20/922-32[62].

L'époque à laquelle vécut TawÎdî fut marquée sur le plan politique par des évènements graves et des troubles ayant provoqué la décadence et la désagrégation du califat de Bagdad. Ayant à leur tête des émirs turcs ou persans, divers États se sont érigés ici et là dans l'aire géographique du monde musulman. Parmi ces États, l'on trouve particulièrement la dynastie bouyide instaurée par une lignée d'émirs persans chiites qui gouvernèrent l'Orient musulman pendant un siècle à partir de l'an 322/933. Dans ce contexte d'agitation, pour garantir la stabilité et la pérennité de leurs états autonomes, les princes bouyides ont dû s'appuyer sur des vizirs habiles, puissants et surtout fidèles à savoir *ÞAbû* l-F*aÃl b. al-ÝAm*îd (307-360/919-970) vizir de *Rukn ad-Dawla*, aÔ-Ñ*âÎib b. ÝAbbâd (326-385/937-995) vizir de MuÞayyid ad-Dawla et plus tard de son fr*ère F*aÎ*r *ad-*D*awl*a, ou *al-Muhallabî* (291-352/903-963) vizir de *MuÝizz ad-*D*awl*a. Ces vizirs qui furent eux-mêmes singulièrement cupltivés à l'image d'*Ibn al-ÝAmîd* qui fut l'un

[61] Voir entre autres références : Stern, S.M., "*Abû Íayyân al-TawÎdî, ÝAlî b. MuÎ. b. al-ÝAbbâs.*" *Encyclopédie de l'Islam*. Nouvelle édition, p. 130-131. **Et** : Keïlani Ibrahim, *Abû Íayyân at-tawÎdî. Essayiste arabe du IV^{e} s. de l'Hégire (X^{e} s.). Introduction à son oeuvre*, Beyrouth, 1950. **Et** : Bergé Marc, *Essai sur la personnalité morale et intellectuelle d'Abû Íayyân at-tawÎdî*, Lille, 1974. Anisi que : Bergé Marc, *Pour un humanisme vécu : Abû Íayyân at-tawÎdî*, 1979

[62] Pour les dates du type D1/D2, la première date (D1) fait référence à l'ère de l'Hégire, la deuxième (D2) à l'ère Chrétienne.

des plus grands écrivains de son temps et dont on disait : « *L'art de la rédaction de la chancellerie a commencé avec ʾAbd al-Ḥamîd et s'est éteint avec Ibn al-ʿAmîd* »[63], rivalisèrent pour attirer à leurs cours érudits, savants et littérateurs ; ils devinrent ainsi les mécènes et les protecteurs que tous les hommes de lettres de l'époque cherchaient à courtiser coûte que coûte. Combiné à la décadence de la vie économique et sociale, ce mécénat marquant eut pour conséquence une instrumentalisation du rôle du *ḫadîb* auprès du vizir, plus encore que de l'émir lui-même. De ce fait, la poésie et la prose « furent adaptées à un genre où dominent les thèmes de louange, de sollicitation poussée jusqu'à une mendicité servile »[64]. Tawḥîdî ne fit pas exception.

Tawḥîdî étudia à *Bagdad* la grammaire avec ʿAbû Saʿîd as-Sîrâfî (284-367/897-977) et ʿAlî b. ʿÎsâ *ar-Rummânî (296-384/908-994)*, le droit šâfiʿite avec *ʿAbû Ḥâmid al-Marwarrûḏî* (m. en 362/972) et *ʿAbû* Bakr aš-Šâšî (291-365/903-975) ; il côtoya également de grands maîtres soufis de cette époque. *Mais c'est ʿAbû* Sulaymân as-Sijis*tânî* (m. après 391/1000), élève du célèbre logicien *Mattâ b. Yûnus* et de Yaḥyâ b. *ʿAdî*, qui « a laissé la plus profonde empreinte sur l'œuvre de *Tawḥîdî* »[65].

L'on sait qu'il était à la Mecque en 353/964 (*Al-ʿImtâʿ*, II, 79) où il écrivit *Risâlat al-Ḥanîn ilâ l-ʿawÔân* [*La nostalgie de la patrie*] (Bergé, *Essai*, 131), et à Rayy en 358/971 (*Yâqût*, *Muʿjam*, II, 292).

Son second séjour à Bagdad s'étend, selon I. Keïlani, entre 358 et 364/968-874 (Keïlani, *Essayiste*, 19). Nous savons d'après *al-Muqâbasât*, p. 156, qu'en 361/971 il assista à *Bagdad* à des cours du philosophe chrétien Yaḥyâ B. *ʿAdi*yy (m. 364/974), le plus célèbre des disciples d'Al-Fârâbî.

---

[63] Citation non attribué, citée par Frédéric Lagrange dans son avant-propos de : *La Satire des deux vizirs*, présenté, traduit de l'arabe et annoté par Frédéric Lagrange, 2004, p. 11.

[64] Keïlani, *op. cit.*, p. 28.

[65] *Ibidem*, p. 23.

La période entre 350/961 et 360/970 est une période importante dans la vie littéraire de Tawḥîdî durant laquelle il composa, comme il le dit lui-même dans la préface de ce livre, son anthologie de littérature (*þadab*), en dix volumes, intitulée *BaÒâÞir al-QudamâÞ wa bašâÞir al-ÍukamâÞ* [*Vues des anciens et pensées avant-gardistes des sages*], plus connue sous le nom d'*al-BaÒâÞir wa-Æ-Åaìâþir*.

La désorganisation administrative et les troubles sociaux qui ont accompagné le début du IV^{ème}/X^{ème} siècle, ont engendré une situation de grande misère pour les hommes de lettres vivant à l'écart de la cour. Cette situation pitoyable, *Tawḥîdî* l'a bien vécue et surtout amplement décrite dans plusieurs de ces écrits. Accablé par la misère et les soucis matériels de la vie, il tenta sa chance à Rayy, d'abord auprès du vizir *PAbû l-FaÃl b. al-ÝAmîd,* sans succès, et puis auprès du vizir *aÒ-Ñâìib b. ÝAbbâd* qui l'employa comme secrétaire en 367/977, et qui finit par le destituer trois ans après, à cause de son orgueil qui le poussa à refuser de "perdre son temps" à copier l'immense recueil des épîtres de son maître. Ces mésaventures que *Tawḥîdî* a subies à Rayy, sont à la base du très célèbre pamphlet qu'il rédigea pour se venger des deux vizirs qui l'avaient humilié : il s'agit de l'épître *þAÌlâq al-wazîrayn* ou *Ma×âlib al-wazîrayn* [*La Satire des deux vizirs*].

Dans ses célèbres *Muqâbasât* [*Entretiens*], nous trouvons plusieurs indications directes se rapportant à l'âge de *Tawḥîdî*, à cette période, ou à des évènements de sa vie littéraire, qui laissent penser qu'il composa *al-Muqâbasât* pendant une longue période s'étalant de 360/970 à 391/1000.

Après les déboires qu'il eut à Rayy, *Tawḥîdî* regagna Bagdad en 370/980 où il resta jusqu'en 400/1009 et où il retrouva la misère qu'il avait fuie, l'attendre à bras ouverts. Après quelque temps de désarroi, il fut recommandé par son nouveau protecteur et ami, le célèbre mathématicien et géomètre *PAbû l-WafâP al-Muhandis al-Buzjânî* (328-376/939-986), à *Ibn al-ÝAriÃ PAbû ÝAbd Allâh al-Íusayn b. PAÎmad b. SaÝdân* (m. en 375/985)*, vizir du prince bouyide NamÒâm ad-Dawla* (353-88/964-98), avec qui

Tawḥîdî eut, pendant les réceptions du soir que le vizir organisait, des entretiens autour de sujets très variés de nature littéraire, philosophique, scientifique, etc., qui formèrent la matière de son livre *al-ÞImtâÝ wa-l-MuÞânasa* [*La délectation et l'agrément* (trad. d'I Keïlani) ou *Le plaisir offert et la sociabilité partagée* (trad. de F. Lagrange)]. C'est à la demande de son ami *PAbû l-WafâP al-Muhandis* qu'il compila le compte-rendu de ces entretiens nocturnes avec le vizir, sous le nom qu'on connaît : *al-ÞImtâÝ wa-l-MuÞânasa*.

En répondant aux sollicitations d'Ibn *Sa'dân*, *Tawḥîdî* entreprit en 371/981, la rédaction d'*AÒ-Ñadâqa wa-Ò-Ñadîq* [*L'amitié et l'ami*] qu'il n'acheva qu'une trentaine d'années plus tard, c'est-à-dire en l'an 400/1009.

En 382-3/992-3, *Tawḥîdî* composa son livre *al-MuÎâÃarât wa-l-munâÛarât* [*les conversations et les controverses*] pour le vizir, à Chiraz, de *ÑamÒâm ad-Dawla*.

Pris, vers la fin de sa vie, par une poussée de colère et de révolte dans un contexte de détresse et de dépression, *PAbû Íayyân at-Tawḥîdî* brûla ses livres « sous prétexte du peu de considération dont il avait été l'objet pendant les vingt années précédentes »[66].

De la période de quatorze ans, entre 400/1009 et 414/1023, date de sa mort, les sources ne disent rien. Les historiens perdent sa trace. Qu'a-t-il fait durant cette période ? Où a-t-il vécu, à Bagdad, à la Mecque, à Chiraz ?

S'appuyant sur certaines données et indications à partir de quelques sources, I. Keïlani conclut que *Tawḥîdî* a dû passer la majeure partie des dernières années de sa vie à Chiraz[67].

---

[66] Stern, S.M, Encyclopédie de l'Islam, p. 131

[67] Keïlani, *Essayiste*, p. 44-45

L'œuvre, déjà éditée, d'*Abû Ḥayyân at-Tawḥîdî* se compose principalement des ouvrages et des épîtres suivants (un grand nombre d'écrits qui ne nous sont pas encore parvenus ou dont les manuscrits ne sont pas complets, sont également attribués à *Tawḥîdî*):

➢ *ÞAllâq al-wazîrayn*	أخلاق الوزيريْن
➢ *Al-ÞIšârât al-Þilâhiyya*	الإشارات الإلهيّة
➢ ***Al-ÞImtâʕ wa-l-MuÞânasa***	**الإمتاع والمؤانسة**
➢ *Al-BaÒâÞir wa-Æ-Åalâþir*	البصائر والذخائر
➢ *AÒ-Ñadâqat wa-Ò-Ñadîq*	الصداقة والصديق
➢ *Al-Muqâbasât*	المقابسات
➢ *Al-Hawâmil wa-š-Šawâmil*	الهوامل والشوامل
➢ *Ar-Risâlat l-baÈdâdiyya*	الرسالة البغداديّة
➢ *Risâlat al-Þimâma*	رسالة الإمامة
➢ *Risâlat al-Îayât*	رسالة الحياة
➢ *Risâlat as-saqîfa*	رسالة السقيفة
➢ *Risâlat fî l-ʕulûm*	رسالة في العلوم
➢ *Risâlat fî ʕilm l-kitâba*	رسالة في علم الكتابة

2. Constitution du corpus

2.1. Texte-source

Avant d'être édité dans un seul livre, *Al-ÞImtâÝ wa-l-MuÞânasa* a été publié en trois volumes édités successivement, en Egypte par *Lajnat al-TaPlîf wa-t-Tarjamat wa-n-Našr* لجنة التأليف والترجمة والنشر , en 1939, 1942 et 1944 à partir de l'édition critique تحقيق de *PAÎmad PAmîn* et PAÎmad Paz-Zayn. Ils ont été réédités ensuite en 1953, toujours en Egypte. Deux maisons d'édition ont fait une reproduction photographique en un seul volume de la totalité d'*al-ÞImtâÝ wa-l-MuÞânasa* ; la première est *Al-Maktaba al-ÝAÒriyya* المكتبة العصريّة en 1957, et la deuxième est *Dâr Maktabat al-Ŝayât* دار مكتبة الحياة en 1966[68]. Les deux rééditions ont été faites à Beyrouth. L'édition que nous avons utilisée est celle de *Dâr Maktabat Al- Ŝayât*.

Le livre comporte 3 volumes à la fin de chacun d'entre eux, les éditeurs ont établi des index concernant les noms propres de personnes, les noms propres de lieux, les anthroponymes et les noms des œuvres cités par *TawÎidî*. Seulement à la fin du deuxième volume, un index des vers cités par *TawÎidî*, classés par ordre alphabétique des rimes, et un autre des hémistiches seuls, ont été ajoutés. Dans l'introduction du livre, *PAÎmad PAmîn* et PAÎmad Paz-Zayn font une brève présentation d'*Abû Ŝayyân at-TawÎidî*, de son époque ainsi que du sujet d'*al-ÞImtâÝ wa-l-MuÞânasa*. On y trouve

---

[68] *SAYYID Ayman FuPâd, Abû Ŝayyân at-TawÎidî wa muPallafâtuhu al-maŴÒûÒa wa-l-maÒbûÝa, dans :* FuÒûl, *Majallat al-naqd al-ÞadabÎ*, n° spécial (premier d'une série de trois) sur l'anniversaire millénaire de *TawÎidî*, n° 3, vol. XIV, *Al-HayPat l-miÒriyya l-Ýâmma li-l-kitâb*, Le Caire, 1995, p. 9-29, p.24.

également la présentation de leur méthode d'édition critique et des deux seuls manuscrits qui étaient à la base de cette édition : celui d'Istanbul (Musée d'art islamique de Topkapi), complet, et celui de Milan, composé seulement de quelques fragments.

Notons que la répartition éditoriale en volumes ne correspond pas exactement à celle, originelle, de *Tawḥîdî*, même si les deux répartitions comportent 3 volumes. En effet, le 1er volume est identique dans les deux subdivisions, alors que les frontières entre le 2ème et le 3ème volumes ne coïncident pas entre le découpage éditorial et celui de *Tawḥîdî*. Le 2ème volume selon l'auteur se termine à la page 165 du volume 2 des éditeurs qui, lui, se termine à la page 205.

Les 661 pages d'*al-ÞImtâʕ wa-l-Muþânasa* se répartissent donc entre les trois volumes, selon le découpage éditorial, de la manière suivante : 206 pages forment le volume 1, le volume 2 est composé de 205 pages, quant au volume 3, il renferme 230 pages.

Au niveau de son organisation logique, *al-ÞImtâʕ wa-l-Muþânasa* est disposé, à l'image des *Mille et une Nuits*, en 40 Nuits dans lesquelles *Tawḥîdî* retranscrit, à la demande explicite de son ami et protecteur *PAbû l-Wafâ al-Muhandis*, les entretiens qu'il a eus avec *Ibn Saʕdân*, vizir de la dynastie bouyide.

Le volume 1 comporte, outre le préambule[69], 15 Nuits : de la première à la 16ème Nuit, sachant que la Nuit 12 est manquante à la fois de l'édition de référence et du manuscrit et que la Nuit 10 regroupe en fait, sans frontière claire, deux Nuits, la Nuit 10 et la Nuit 11.

Le volume 2 (répartition éditoriale) se compose de 14 Nuits complètes : de la Nuit 17 à la Nuit 30 ainsi que du début de la Nuit 31 (deux pages seulement). Si l'on

[69] Pour des raisons de commodité au moment de l'analyse et de gain d'espace dans la présentation des tableaux de données, nous avons renommé, dans les deux phases du traitement et de l'analyse, le préambule en Nuit00.

considère la répartition du manuscrit, le volume 2 est composé de 12 Nuits : de la Nuit 17 à la Nuit 28.

Le volume 3 quant à lui, il se compose de 10 Nuits : de la Nuit 31 à la Nuit 40 (édition de référence) ou de 12 Nuits : de la Nuit 29 à la Nuit 40 (répartition originelle). Après la Nuit 40, *Tawḥîdî* a inséré à la fin du volume 3 deux lettres qu'il dit avoir envoyées au vizir *Ibn Sa'dân*.

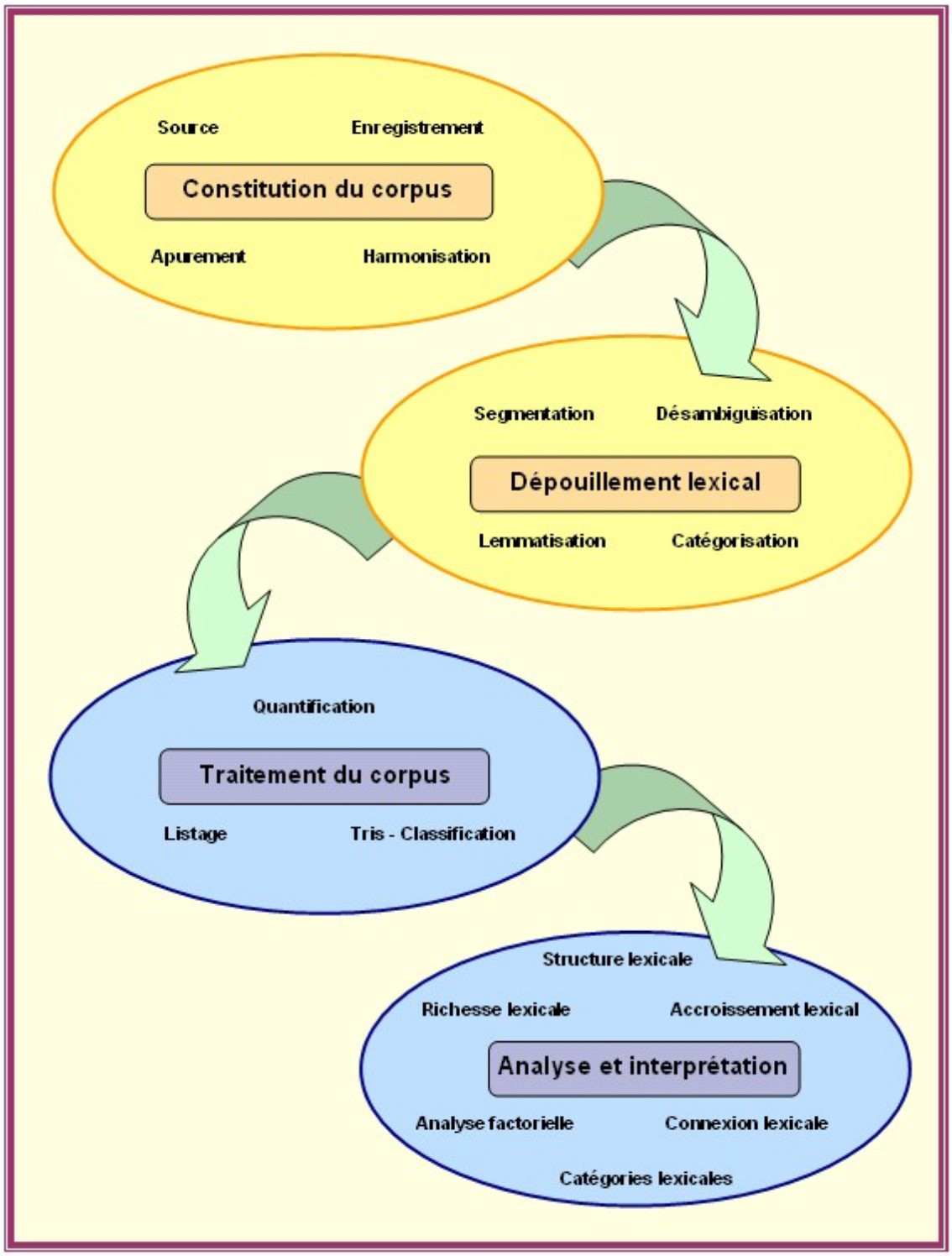


Figure 1
Les moments de l'analyse lexicométrique

2.2. Enregistrement du texte en machine

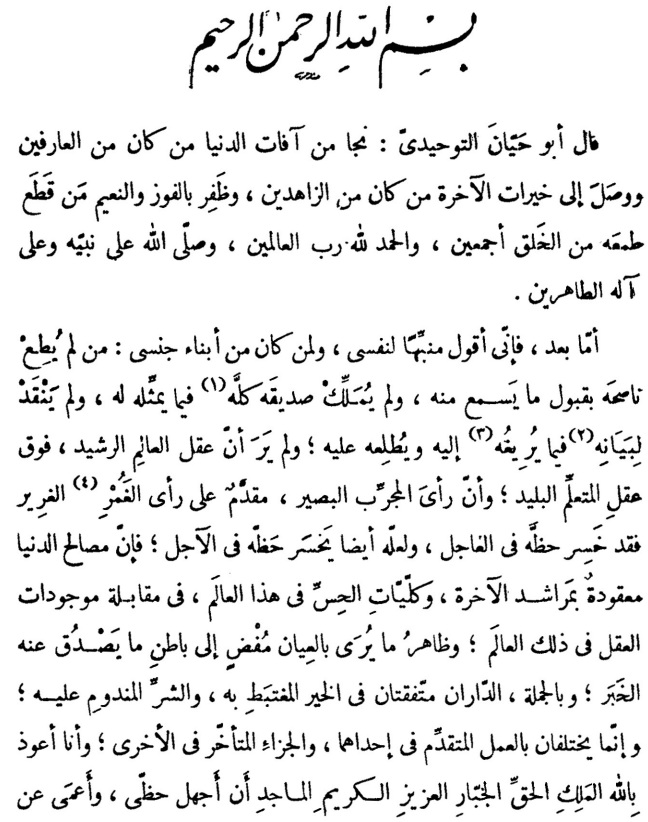
Il y a au moins deux façons d'enregistrer un texte en machine : par saisie directe sur clavier ou par numérisation, c'est-à-dire par scannage de l'édition de référence, en papier, suivi d'une opération de reconnaissance optique des caractères ROC (plus connue sous les initiales en anglais OCR, *Optical Character Recognition*).

Convaincu du fait que la saisie directe sur clavier représente un frein important à l'entrée massive de données textuelles et une cause fréquente d'erreurs, et croyant pouvoir gagner du temps dans l'enregistrement de notre corpus en machine, nous avons tenté dans un premier temps, la deuxième méthode : scannage + OCR. Pour ce faire, nous avons scanné tout *al-ÞImtâÝ wa-l-MuÞânasa* et avons constitué des fichiers d'images organisés en volumes, Nuits et pages. C'est au moment de l'opération de reconnaissance optique de caractères que nous nous sommes rendu compte de l'énorme difficulté voire même de l'impossibilité de réussir cette opération sur des textes arabes édités dans les conditions des années cinquante, à l'image de notre édition de référence.

Plusieurs causes sont à la base de cet échec :

- la qualité des OCR arabes et l'état d'avancement des recherches dans ce domaine au moment où nous avons entrepris ce travail ;
- la qualité du papier de l'édition de référence ;
- la police de caractères utilisée ;
- les caractères cassés ou tordus ;
- les ligatures ;
- les *kašîda*, etc.

Nous avons utilisé comme OCR, la version 3 du logiciel القارئ الآلي *al-Qâriþ al-þâliyy* [le lecteur automatique][70]. Il faut dire que les résultats que nous avons obtenus, comme le montre les encadrés suivants, sont extrêmement décevants même après quelques semaines d'apprentissage. L'apprentissage étant une phase importante et nécessaire dans le processus de reconnaissance optique de caractères pour améliorer les résultats d'un OCR en lui apprenant à reconnaître correctement les caractères mal reconnus ou totalement non reconnus.



بسم الله الرحمن الرحيم

قال أبو حَيّانَ التوحيدىّ : نجا من آفات الدنيا من كان من العارفين ووصَلَ إلى خيرات الآخرة من كان من الزاهدين ، وظَفِر بالفوز والنعيم مَن قَطَع طمعَه من الخَلق أجمعين ، والحمد لله رب العالمين ، وصلّى الله على نبيّه وعلى آله الطاهرين .

أمّا بعد ، فإنّى أقول منبّهًا لنفسى ، ولمن كان من أبناء جنسى : من لم يُطِعْ ناصحَه بقبول ما يَسمع منه ، ولم يُمَلِّكْ صديقَه كلَّه[1] فيما يمثّله له ، ولم يَنْقَدْ لِبَيَانِه[2] فيما يُرِيُه[3] إليه ويُطلِعه عليه ؛ ولم يَرَ أنّ عقل العالِم الرشيد ، فوق عقل المتعلّم البليد ؛ وأنّ رأىَ المجرّب البصير ، مقدّمٌ على رأى الغُمْرِ[4] الغرير فقد خَسِر حظَّه فى العاجل ، ولعلّه أيضا يَخسَر حَظّه فى الآجل ؛ فإنّ مصالح الدنيا معقودةٌ بمَراشد الآخرة ، وكلّيّاتِ الحِسِّ فى هذا العالَم ، فى مقابلة موجودات العقل فى ذلك العالَم ؛ وظاهرُ ما يُرَى بالعِيان مُفْضٍ إلى باطنٍ ما يَصْدُق عنه الخَبَر ؛ وبالجملة ، الدّاران متّفقتان فى الخير المغتبَطِ به ، والشرّ المندومِ عليه ؛ وإنّما يختلفان بالعمل المتقدّم فى إحداها ، والجزاء المتأخّر فى الأخرى ؛ وأنا أعوذ بالله المَلِكِ الحقِّ الجبّارِ العزيزِ الكريمِ الماجدِ أن أجهل حظّى ، وأعمَى عن

Figure 2 :
La première page de l'édition de référence scannée d'*al-ÞImtâÝ wa-l-MuÞânasa*.

---

[70] القارئ الآلي est un OCR de la société SAKHR : http://www.sakhr.com/Sakhr_e/Products/OCR_Off.htm?Index=2&Main=Products&Sub=OCR .

La dernière version de ce logiciel est la version 5 ; nous n'avons pas testé cette dernière version et nous ne pouvons par conséquent donner de jugement sur sa fiabilité.

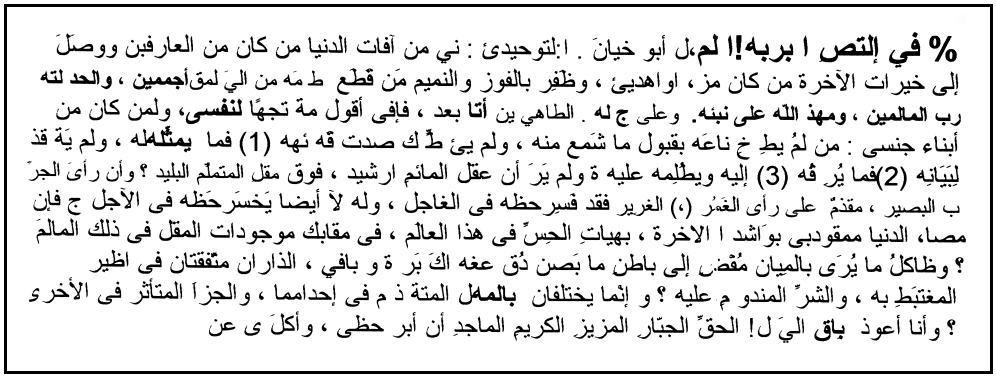


Figure 3 :
Résultat de l'OCR avant apprentissage.

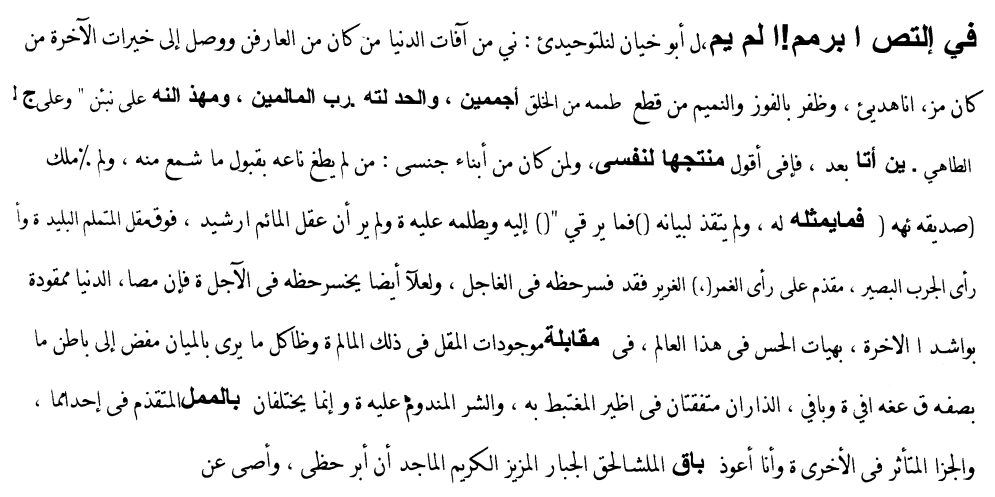


Figure 4 :
Même après apprentissage, le résultat de l'OCR est loin d'être satisfaisant.

Cette méthode ayant échoué, nous nous sommes restreint à la méthode de saisie directe sur clavier. Faute de secrétaires ou d'opérateurs de saisie en arabe, en France, capables de saisir en un temps convenable 661 pages de texte en arabe classique, nous étions obligé de faire faire cette saisie en Tunisie[71].

La saisie a été faite en un temps relativement court et était d'une qualité acceptable. Mais, comme dans tout travail de saisie, humain, des erreurs ont été

[71] Nous tenons à remercier ici M. Abdelfattah Braham de nous avoir aidé dans cette tâche en chargeant, à notre place, des secrétaires professionnelles de la saisie du corpus.

décelées dans les fichiers qui nous ont été remis sous format ".doc". Une opération de correction et d'apurement est donc nécessaire dans de pareils cas, avant de pouvoir exploiter les fichiers.

2.3. Apurement

Pour corriger les multiples erreurs de saisie contenues dans les fichiers résultant de la saisie directe sur clavier, une première correction a été faite, en Tunisie, par M. Braham qui a utilisé un correcteur orthographique s'intégrant dans Word® et élaboré par Nabil Gader, il s'agit du correcteur CorTexAr[72].

Une partie des erreurs a donc été corrigée suite à cette application du correcteur orthographique ; mais d'autres erreurs ont subsisté et ce pour deux raisons : d'abord, aucun correcteur, aussi performant soit-il, n'est fiable à 100 %. Ensuite, il arrive que le correcteur rencontre des mots correctement orthographiés mais ne correspondant en fait pas aux mots du texte saisi. De ce fait, ces mots, considérés corrects, passent entre les mailles du filet du correcteur orthographique. Le seul moyen dans ce cas, de détecter ces erreurs est un "pointage" mot par mot du texte saisi avec le texte-source.

Cette opération de pointage, longue et fastidieuse, faite, nous avons pu corriger toutes les erreurs de saisie commises. Elles sont de plusieurs types :

- **Interversion de lettres ou de signes diacritiques** : c'est le cas où l'on rencontre une permutation de deux consonnes, d'une consonne et d'une voyelle longue, de deux voyelles brèves ou d'une voyelle brève et d'un signe diacritique.

Mot erroné	Mot attendu

[72] Ce correcteur orthographique n'est pas commercialisé, il a été développé par N. Gader dans le cadre de sa thèse au sein du groupe DIINAR. Après plusieurs tests, ce correcteur orthographique a été jugé plus performant que celui de Microsoft Word 98 (pour l'arabe).

فضيلة	فضلية
عَلِمْتَ	عَمِلْتَ
كُلّ	كّلُ

- **Emplacement de signes diacritiques** : c'est un type d'erreurs où le signe diacritique est placé, non pas sur la consonne où il doit être, mais sur la précédente ou la suivante.

Mot erroné	Mot attendu
قوّية	قويّة
فظنّنا	فظننّا
عَدُّونا	عَدُوُّنا
بَوَّأَ	بَوَّأَ
كُلَّمَا	كُلَّمَا

- **Manque de signes diacritiques** : parfois on ne trouve pas un signe diacritique attendu à un emplacement bien déterminé dans un mot. Se présente le cas aussi où un signe diacritique attendu est absent et où un autre signe diacritique non attendu est ajouté ; ceci peut donner lieu à une ambiguïté non seulement lexicale mais aussi syntaxique voire même pragmatique. Ce dernier cas est illustré par le mot graphique فالحظّ segmentable en ف \ ال \ حظّ [et / la / chance], au lieu de فٱلحظ segmentable en ف \ اِلحظ [et / regarde - observe - remarque (*impératif*)].

Mot erroné	Mot attendu
لكن	لكنّ
مادة	مادّة
ماديّة	مادّيّة
كميّة	كمّيّة
فالحظّ	فٱلحظ

- **Mauvaise tradition typographique** : une des mauvaises habitudes typographiques que l'on rencontre souvent dans les livres et dans

certains journaux est l'écriture systématique d'un *ÞalifmaqÒûra*[73] ى à la place d'un *yâÞ* final ي . Une autre, en est l'écriture du *tanwîn al-fatÎa* « ً » sur le *Þalif* au lieu que ce soit sur la consonne précédant ce dernier. Ne pas écrire la hamza stable sur le *Þalif* au début du mot est aussi une mauvaise habitude typographique à corriger.

Mauvaise graphie	Graphie attendue
دوالى	دوالي
فهماً	فهمًا

[73] Même si les défenseurs de cette tradition typographique considèrent que c'est bien un *yâÞ* sans les points qui est écrit et non un *Þalif maqsûra* ; il n'en reste pas moins que c'est la même graphie et le même code machine que celui du *Þalif maqsûra*. D'autant plus que cette graphie et l'ambiguïté qu'elle engendre pose énormément de problème à l'apprentissage de l'arabe pour les débutants. Mais aussi, pour ce cas et celui du *tanwîn al-fatÎa* des conséquences désastreuses sont provoquées, en traitement automatique de l'arabe, pour le récepteur machine.

- **Ajout de lettres ou de signes diacritiques** : soit par simple mégarde en appuyant sur une touche du clavier, soit par confusion de deux formes présentant une homographie consonantique[74], il arrive que l'opérateur de saisie ajoute à un mot, une consonne, un signe diacritique ou les deux à la fois.

Mot erroné	Mot attendu
مُتبابينة	مُتباينة
أيّ	أيْ ou أي
لكنّ	لكنْ ou لكن
أنّ	أنْ ou أن

Outre ces erreurs de saisie, nous avons décelé, dans l'édition de référence, et harmonisé certaines graphies de la *hamza* liées à des habitudes typographiques (surtout égyptienne) s'écartant des règles orthographiques habituelles, à savoir le mot رؤوس [des têtes], écrit étrangement : رءوس.

Aussi, quelques erreurs liées à l'édition critique ont-elles été repérées, après l'impression du livre, par *Muḥammad* Kurd *ʿAlī* (vol. 1 et 2), par *Muṣṭafā* Jawâd (vol. 1 et 2) et par le Professeur Kraus (vol. 1, 2 et 3), et dont la liste a été ajoutée par les éditeurs, sous forme d'*errata*, à la fin du volume 1 et 2, pour le premier et à la fin du volume 3 pour les deux autres ; nous avons donc apporté ces corrections à notre corpus.

2.4. Harmonisation primaire

L'harmonisation primaire est cette opération d'harmonisation effectuée avant la segmentation et consistant à homogénéiser toutes les variantes graphiques d'un même

[74] Voir les types d'homographie au chapitre « *Norme de dépouillement* ».

mot sous une seule forme graphique d'harmonisation choisie parmi celles-ci. Mais l'harmonisation primaire n'est pas liée aux seules variantes graphiques ; un autre phénomène nécessite également le recours à cette entreprise. Il s'agit de certaines pratiques scripturales ou tapuscrites étroitement liées au système d'écriture de l'arabe caractérisé principalement par la disjonction entre l'ensemble des consonnes et celui des voyelles. En effet, dans une écriture partiellement vocalisée et en absence de toute norme de vocalisation, les voyelles placées arbitrairement sur (sous) une ou plusieurs consonnes d'un mot donnent naissance, pour le récepteur machine, à des formes graphiques, considérées comme différentes, du même mot.

Pour un mot composé, par exemple, de deux consonnes, selon que l'on place seulement la voyelle de la 1ère consonne, seulement celle de la 2ème consonne, aucune des deux ou les deux voyelles à la fois, l'on obtient 4 formes graphiques différentes pour le récepteur machine[75] (exemple : مِن – منْ – من – مِنْ) ; ce nombre est composé de la somme de la combinaison (au sens probabiliste du terme) de deux consonnes ayant une seule voyelle ($C_2^1 = \frac{2!}{1!(2-1)!} = \frac{2}{1 \times 1} = \frac{2}{1} = 2$), de celle de deux consonnes n'ayant aucune voyelle ($C_2^0 = 1$) et de celle de deux consonnes ayant deux voyelles ($C_2^2 = 1$). Pour un mot de 3 consonnes, on obtient 8 formes graphiques (exemple :بعد – بَعْدَ – بَعد – بعْد – بعدَ – بَعْد – بَعدَ – بعْدَ) ; ce nombre est composé de la somme de ($C_3^0 = 1$), de ($C_3^1 = 3$), de ($C_3^2 = 3$) et de ($C_3^3 = 1$). En général, pour un mot composé de n consonnes, les formes graphiques seront au nombre de :

$$\sum_{p=0}^{n} C_n^p = \sum_{p=0}^{n} \frac{n!}{p!(n-p)!}$$

[75] Il ne faut pas oublier à cet effet, que les voyelles, même si leur disposition dépend de celle des consonnes, ont des codes machines différents de ceux des consonnes.

En appliquant la formule du triangle arithmétique de Pascal ($C_n^p = C_{n-1}^p + C_{n-1}^{p-1}$) et sachant que $C_n^0 = 1$ et $C_n^n = 1$, l'on obtient :

$$\sum_{p=0}^{n} C_n^p = 2 \times \sum_{p=0}^{n-1} C_{n-1}^p$$

C'est ce qui explique que le nombre des formes graphiques différentes (pour le récepteur machine) est multiplié par 2, à mesure que le nombre de consonnes composant le mot augmente d'une unité (4 formes pour un mot de 2 consonnes, 8 pour un mot de 3 consonnes, 16 pour 4, 32 pour 5, etc.).

À cette étape de la constitution du corpus, notre tâche est donc d'harmoniser toutes les formes graphiques d'un même mot sous une seule forme. Nous exposons plus loin dans le chapitre consacré à la *norme de saisie et d'harmonisation*[76], toutes les graphies d'harmonisation que nous avons retenues pour les mots-outils les plus rencontrées dans notre corpus. En ce qui concerne les mots lexicaux, ce que nous exposons dans ledit chapitre, ce sont des règles d'harmonisation selon la catégorie lexicale et selon certains schèmes de telle ou telle catégorie.

2.5. Repérage

Dans cette étape, il est question de repérer puis de marquer, par des symboles, des balises ou tout autre moyen, certains éléments à la périphérie du lexique, comme les noms propres, ou faisant partie de celui-ci mais nécessitant un traitement particulier, comme les unités polylexicales ou les phrasèmes.

En ce qui concerne les unités polylexicales, qu'elles soient des noms propres composés ou des noms communs composés, une opération de ligature est nécessaire

[76] Voir p. 144-195

pour ne faire des éléments entrant en leur composition qu'une seule et unique forme. Cela est très important dans le but d'empêcher un segmenteur automatique ou semi-automatique non doté d'un dictionnaire des unités polylexicales (c'est le cas de notre segmenteur), de segmenter ces unités composées en unités simples disjointes.

Nous avons choisi le signe "+" comme symbole de ligature des éléments composant les unités polylexicales.

2.5.1. Repérage et ligature des noms propres

Au fur et à mesure de la lecture du corpus, nous avons repéré les noms propres (de personne, de lieu, d'œuvre, qualificatifs de Dieu, etc.) et les avons marqué par l'insertion juste avant et après eux d'un symbole choisi à cet effet ; il s'agit du symbole "$". Exemple : $التوحيديّ$,$الله$, etc.

Pour les noms propres composés, en plus du symbole qui les encadre, nous avons ligaturé les éléments entrant en leur composition, par le signe "+" comme précisé plus haut. Exemple : $أبو+حيّانَ+التوحيديّ$, $أبو+عبد+الله+العارض+الحسين+بن+أحمد+بن+سعدان$.

Nous aurions pu utiliser des balises, comme celles utilisées dans la phase de balisage expliquées plus loin, si nous avions choisi de garder les noms propres au sein du corpus et de les traiter comme tous les autres éléments composant le lexique. Mais notre choix a été autre : nous avons préféré suivre ce qui est devenu, à juste titre, une tradition dans les études lexicométrique : traiter les noms propres à part et leur conserver un statut à la périphérie du lexique.

2.5.2. Repérage et ligature des unités polylexicales

En ce qui concerne les unités polylexicales : noms communs composés, locutions grammaticales, etc., nous avons utilisé le même symbole "+" pour ligaturer les

éléments les composant sans avoir à les marquer par quelqu'autre symbole que ce soit, et ce parce qu'elles font partie intégrante du lexique et doivent rester à leurs emplacements respectifs dans le corpus.

Pour étudier, par exemple, les phrasèmes, on pourrait choisir d'insérer des balises, déterminées à cet effet, avant et après chacune de ces constructions pour pouvoir en étudier par la suite, la répartition, la structure, etc. ou en faire la concordance pour une étude sémantique en contexte.

3. Dépouillement lexical

Après la phase initiale de constitution du corpus (enregistrement, apurement, harmonisation, etc.) vient le moment critique du dépouillement lexical qui débouchera sur la transformation du texte de départ en une (des) liste(s) de formes, de lemmes et de catégories lexicales. Ce sont ces listes-là qui constitueront la base des traitements statistiques. Le dépouillement lexical est, en effet, un moment crucial en ce sens où les opérations, délicates, dont il se compose ne doivent pas être pratiquées arbitrairement ni sporadiquement mais nécessitent, bien au contraire, l'application constante et régulière, à tous les mots graphiques du corpus, d'un certain nombre de règles lexicologiques établies à cet effet et qui définissent ce qui est appelé en lexicométrie, la *norme de dépouillement*. Cette même *norme de dépouillement* plus la *norme de saisie et d'harmonisation*, qui intervient, elle, dans la phase initiale, constituent la *norme lexicologique* que nous présenterons en détail dans la deuxième partie de ce travail[77].

Le dépouillement lexical constitue donc le deuxième moment important du prétraitement des données textuelles avant de soumettre celles-ci au traitement et à l'analyse statistiques. Cette phase de dépouillement lexical renferme plusieurs étapes dont les plus importantes sont la segmentation, la désambiguïsation, la lemmatisation et la catégorisation.

La segmentation ou itémisation ouvre la voie aux autres étapes principales puisqu'elle se charge de découper le texte en unités minimales d'un certain type : en lexicométrie, ce sont en général les mots-formes. La lemmatisation consiste à identifier pour chaque mot-forme son adresse lexicale, que l'on appelle le *lemme*. La catégorisation, quant à elle, consiste à choisir parmi les différentes catégories lexicales possibles du lemme, la bonne catégorie à assigner à ce dernier en fonction du contexte

[77] Voir la *norme lexicologique* proposée, chapitres 4 et 5

immédiat ou médiat du mot-forme ; la tâche de l'étape de catégorisation se voit nettement facilitée si l'on procède d'abord à une opération de désambiguïsation.

Ce moment important de l'analyse lexicale est donc un processus qui, analysant le texte, va définir les frontières des unités de décompte (segmentation), assigner le lemme (lemmatisation) et la catégorie lexicale (catégorisation) à chaque unité, après avoir levé, en fonction du voisinage du mot-forme, d'éventuelles ambiguïtés (désambiguïsation). Avant de soumettre au logiciel de traitement statistique les différentes listes obtenues à la fin du dépouillement, une opération de balisage (ou encodage) est nécessaire. Ces étapes du dépouillement lexical, nous les présentons dans le schéma suivant :

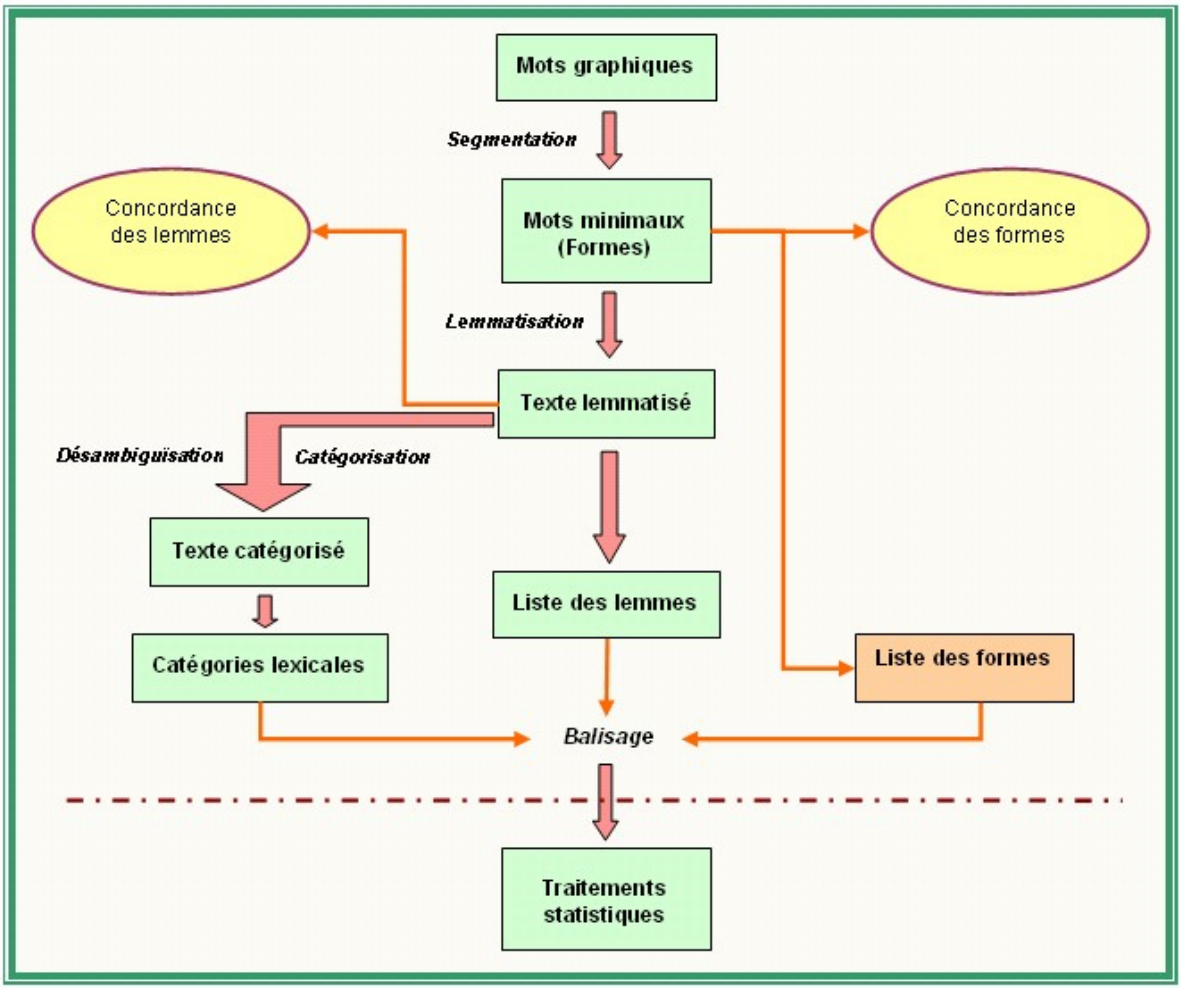


Figure 5
Les étapes du dépouillement lexical

Après la segmentation et avant de lemmatiser le texte, le fichier du texte segmenté ainsi obtenu, peut d'ores et déjà être exploité tel quel pour obtenir par exemple, des concordances de formes ou soumettre la liste des formes au traitement statistique en utilisant *Lexico3*. Une fois la segmentation faite, à la place ou en parallèle de l'exploitation du fichier des formes, c'est le fichier du texte lemmatisé qui sera exploité. Des concordances de lemmes seront alors permises et la liste des lemmes sera soumise à *Lexico3* pour le traitement et l'analyse statistiques. Après la catégorisation, le fichier final ainsi obtenu ne comporte plus que les codes des catégories lexicales ; pour

une étude statistique des catégories lexicales, c'est donc la liste de ces codes qui sera soumise à *Lexico3*.

3.1. Segmentation

Par segmentation, nous entendons la segmentation lexicale[78] ou itémisation (*tokenization* ou *word segmentation*) qui consiste à segmenter un texte en mots-formes ou items lexicaux (*tokens* en anglais).

La notion du mot graphique et de son analyse en arabe que nous présenterons plus loin, est féconde en traitement automatique de la langue arabe depuis les premiers travaux de David Cohen jusqu'aux travaux autour de la base de connaissances DIINAR, en passant par les travaux de l'équipe SAMIA. Elle est particulièrement productive en lexicométrie arabe. Selon que l'on s'attache à analyser le mot graphique, ou mot maximal, en termes de proclitiques, préfixes, base, suffixes et enclitiques, ou seulement en termes de mot minimal et de clitiques, l'on procède à une analyse morphologique pouvant aller jusqu'à la disjonction de la racine et du schème entrant en composition de la base, dans le premier cas, ou simplement à une analyse lexicale mettant en exergue les unités lexicales c'est-à-dire mots lexicaux et mots-outils, dans le second. En lexicométrie, c'est bien entendu la deuxième analyse qui est pressentie. De ce fait, segmenter un texte arabe, revient donc à analyser ses mots maximaux en mots minimaux et clitiques ; il est cependant évident que les mots graphiques qui sont dépourvus de clitiques sont des mots minimaux et leurs frontières demeurent donc inchangées.

La segmentation d'un texte est donc cette opération qui consiste à repérer les unités lexicales en délimitant bien leurs frontières. C'est ce à quoi nous nous sommes livré dans cette étape de notre travail après avoir défini, dans la *norme de segmentation*,

[78] Pour les types de segmentation, voir la norme de dépouillement, chapitre 5

les frontières entre les mots en fixant précisément les règles qui déterminent les unités segmentables et celles qui ne le sont pas. Les choix que nous avons faits et les décisions que nous avons arrêtées pour décider de ce qui est segmentable et de ce qui ne l'est pas, nous les présentons *infra*, dans la *norme de segmentation*[79]. Cette opération de segmentation de notre corpus n'a pas été faite manuellement, nous l'avons automatisée ; c'est une segmentation semi-automatique pour laquelle nous avons utilisé un segmenteur que nous avons élaboré en collaboration avec R. Zaafrani[80] et que nous présentons ci-après.

En TAL, la segmentation d'un texte consiste à structurer celui-ci en passant d'un ensemble continu de caractères à une suite discrète de segments. Mais pour que cette structuration automatique ou semi-automatique soit réalisable, il faut que l'outil assigné à cette tâche, le segmenteur, intègre une grammaire de segments (ou grammaire segmentale) regroupant des règles de combinaisons de caractères pour former des unités lexicales, des segments. De plus, des connaissances linguistiques sont nécessaires pour pouvoir segmenter du texte correctement. Une solution serait d'utiliser un dictionnaire, qui recense à la fois les exceptions et les différentes listes fermées de mots et expressions figées facilitant non seulement la reconnaissance des unités et la définition de leurs frontières mais aussi la désambiguïsation. Une autre solution serait d'utiliser des calculs statistiques qui permettraient, en étudiant les occurrences des mots dans un corpus, de mesurer la probabilité qu'une séquence d'unités lexicales soit correcte. Dans le cas où plusieurs séquences sont possibles, c'est la plus probable qui sera choisie. Pour le moment, notre segmenteur n'utilise pas cette méthode probabiliste.

---

[79] Voir la norme de segmentation, chapitre 5

[80] Voir (Zaafrani 2002)

3.1.1. Le segmenteur semi-automatique

En plus de la grammaire segmentale dont nous l'avons doté, notre segmenteur est un segmenteur à dictionnaire en ce sens où, dans la mémoire du logiciel est stocké un dictionnaire (il s'agit d'un lexique généré à partir de la base de connaissance DIINAR) et la segmentation est faite par consultation de ce lexique généré. Ce lexique généré est sous forme d'une base de données Access® qui contient, entre autres tables et relations, la table des verbes, la table des noms, la table des déverbaux, la table des mots-outils, la table des proclitique et la table des enclitiques.

Il est vrai que l'analyse de chaque séquence de caractères est basée sur le modèle d'analyse du mot graphique en arabe dans lequel un mot graphique est considéré comme une suite de constituants immédiats, l'analyseur (morphologique) ayant pour tâche d'identifier les constituants du mot en le décomposant en proclitiques, préfixes, base, suffixes et enclitiques et d'associer à chaque constituant sa ou ses catégories grammaticales ainsi que les traits qui lui sont associés ; mais ce qui nous intéresse dans notre tâche de segmentation, ce n'est pas ce niveau approfondi d'analyse morphologique : le rôle principal de ce segmenteur est l'itémisation, la segmentation lexicale. C'est pour cette raison que nous n'utilisons qu'une version allégée de cette analyse du mot graphique en restant à la surface du niveau lexical en ce sens qu'il s'agit uniquement de savoir si le mot ou ses constituants appartiennent ou non à la langue pour pouvoir proposer le cas échéant une (des) segmentation(s) envisageable(s). Cet allègement a pour conséquence de réduire considérablement la taille du lexique généré et de rendre le segmenteur plus rapide et plus performant.

Aussi, devons-nous préciser que l'allègement n'est pas seulement en profondeur ; il est également horizontal parce que nous avons décidé de procéder à ce que l'on appelle l'*analyse unifiée*. En effet, s'agissant, rappelons-le encore une fois, d'une segmentation lexicale, autrement dit d'un découpage en surface du niveau lexical, la distinction à faire entre les différentes acceptions sémantiques, référentielles, etc.

d'une unité lexicale n'est pas significative à ce stade-là. Il n'est pas capital de multiplier, par exemple, par huit les possibilités de segmenter un mot graphique commençant par le mot-outil وَ *wa* [et] pour la simple raison qu'il a en arabe huit وَ différents (coordonnant, particule de serment, particule d'accompagnement, *wâw rubba*, particule du chiffre 8, etc.) : une analyse unifié serait ici très appréciée puisque dans les huit cas possibles nous n'obtenons qu'une seule manière de découper ce mot graphique indépendamment des différentes nuances. Un autre exemple : considérons le mot graphique représentant la phrase suivante أَرِنَاهُمَا *Þarinâhumâ* [montre-les-nous]. Cette phrase est composée du verbe *montrer* conjugué à la deuxième personne du singulier à l'impératif أَرِ *Þari* [*montre*]et de deux compléments d'objet هُمَا *humâ* [les] et نَا *nâ* [nous]. Le complément d'objet هُمَا peut correspondre au pronom personnel complément de troisième personne du duel masculin ou à celui de troisième personne du duel féminin. Et le complément d'objet نَا peut correspondre au pronom personnel complément de première personne du pluriel masculin, de première personne du pluriel féminin, de première personne du duel masculin ou à celui de première personne du duel féminin. Si nous prenons en considération les variations en genre et en nombre, la combinaison de ces deux compléments nous donne huit analyses possibles que nous résumons ainsi :

Montre-les {*duel*, *masculin*}-nous {*duel*, *masculin*}.	أَرِنَا{*duel*, *masculin*}هُمَا{*duel*, *masculin*}
Montre-les {*duel*, *masculin*}-nous {*duel*, *féminin*}.	أَرِنَا{*duel*, *masculin*}هُمَا{*duel*, *féminin*}
Montre-les {*duel*, *masculin*}-nous {*pluriel*, *masculin*}.	أَرِنَا{*pluriel*, *masculin*}هُمَا{*duel*, *masculin*}
Montre-les {*duel*, *masculin*}-nous {*pluriel*, *féminin*}.	أَرِنَا{*pluriel*, *féminin*}هُمَا{*duel*, *masculin*}
Montre-les {*duel*, *féminin*}-nous {*duel*, *masculin*}.	أَرِنَا{*duel*, *masculin*}هُمَا{*duel*, *féminin*}
Montre-les {*duel*, *féminin*}-nous {*duel*, *féminin*}.	أَرِنَا{*duel*, *féminin*}هُمَا{*duel*, *féminin*}
Montre-les {*duel*, *féminin*}-nous {*pluriel*, *masculin*}.	أَرِنَا{*pluriel*, *masculin*}هُمَا{*duel*, *féminin*}

Montre-les {*duel*, *féminin*}-nous {*pluriel*, *féminin*}. {*duel*, *féminin*}هُمَا{*pluriel*, *féminin*}أَرِنَا

Dans ces huit analyses possibles, une seule et unique segmentation (découpage) résultante est opérée, quel est donc l'intérêt de faire figurer huit segmentations identiques si ce n'est le fait d'alourdir le segmenteur automatique ou semi-automatique sans raison valable pour ce genre d'entreprise ! C'est dans ces cas de figures que nous avons opté pour l'analyse unifiée : l'analyseur ne présentera alors qu'une seule façon de segmenter de tels mots graphiques. La première opération par laquelle commence le segmenteur est bien entendu le repérage des mots graphiques. Les frontières de mots graphiques ne posent pratiquement pas de problèmes dans le système d'écriture de l'arabe. Les mots graphiques sont séparés par des espaces ou par des signes de ponctuation. Une spécificité est, tout de même, à noter concernant l'espace qui fonctionne en général comme séparateur de mots sauf dans les unités polylexicales où il est considéré comme composant de mots complexes. Le repérage et le traitement des unités polylexicales posent un vrai problème aux différents segmenteurs automatiques ou semi-automatiques. Certains sont équipés de dictionnaires de mots complexes qui leur permettent de détecter ces unités dans le corpus à segmenter et de les traiter par la suite en tant qu'unités à part entière. Notre segmenteur n'étant pas doté d'inventaires des unités polylexicales, pour que le texte ne soit pas émietté à tort, nous avons procédé à un repérage manuel des unités polylexicales et avons ligaturé les éléments entrant en leur composition en utilisant le signe « + » comme nous l'avons déjà signalé au début de ce chapitre. Une règle a été alors ajoutée au programme du segmenteur pour qu'il reconnaisse comme une seule unité les suites de caractères alphabétiques séparées par le signe « + ».

Une fois le texte à segmenter chargé dans la machine et les frontières des mots graphiques repérées, le système commence à examiner, un à un, tous les mots graphiques du texte. Il consulte d'abord les tables des mots-outils, des proclitiques et des enclitiques aux entrées desquelles il compare les éventuels segments du mot à analyser, repérés suite à un découpage primaire. Le système vérifie ensuite les

compatibilités entre tous ces éléments et il propose en fin de compte la ou les segmentations possibles après avoir vérifié et appliqué les règles de la grammaire segmentale. Les segmentations sont ainsi opérées l'une après l'autre, en flot linéaire, c'est-à-dire que dans chaque suite de caractères du texte, la fin du mot segmenté sera le début du mot suivant.

D'une façon plus précise, l'analyse d'un mot graphique sur le plan lexical, c'est-à-dire la vérification de son appartenance (lui ou ses constituants) à la langue, passe par les trois phases suivantes :

Découpage primaire :

Cette opération fait appel à la table des proclitiques et à celle des enclitiques permettant de reconnaître et d'isoler les constituants immédiats du mot à vérifier. L'algorithme de découpage utilise des automates de reconnaissance construits à partir de ces tables et permettant de reconnaître un élément même s'il est totalement ou partiellement non voyellé. Le découpage fournie comme résultat trois listes, la première contient tous les proclitiques reconnus à partir du début du mot, la deuxième tous les mots minimaux reconnus à partir de la fin des proclitiques, et la troisième tous les enclitiques reconnus à partir de la fin du mot.

Vérification des compatibilités :

Considérée comme une vérification de surface, cette étape consiste d'abord à croiser deux à deux les listes obtenues à l'étape précédente pour n'en retenir que les couples compatibles, puis, par croisement, ne retenir que les combinaisons compatibles (triplets). Ceci est fait à l'aide d'une matrice de compatibilité préalablement calculée. Cette matrice est, en fait, intégrée à une matrice de compatibilité plus large utilisée dans l'analyse morphologique : la matrice de compatibilité des prébases et des postbases. Les prébases (respectivement

postbases) sont des combinaisons de proclitiques et de préfixes (respectivement suffixes et enclitiques) obtenues en respectant des règles de compatibilité et de cooccurrence issues de la grammaire des formants du mot[81].

Interrogation du lexique généré :

Étant donné que chaque mot minimal, chaque proclitique et chaque enclitique correspond à une unité du lexique, toutes les combinaisons issues de l'étape précédente sont comparées, dans cette étape, aux éléments du lexique généré et ce jusqu'à ce qu'une combinaison soit validée : une segmentation est alors proposée en sortie. Si le mot graphique présente plusieurs solutions de segmentation, toutes les possibilités sont retenues et proposée en sortie.

Figure 6
Capture d'écran : écran d'accueil du segmenteur montrant le chemin du fichier à segmenter, le bouton du lancement de la segmentation et le champ affichant le n° du mot graphique eu cours de segmentation

Comme nous venons de l'expliquer, dans le cas où un mot graphique présente plus d'une segmentation, nous avons choisi d'afficher à l'écran toutes les possibilités de découpage et c'est à l'utilisateur de choisir la bonne segmentation comme le montre la figure suivante. Dans la partie supérieure de cet écran, le logiciel présente le mot graphique sujet à une segmentation multiple (à droite), le nombre de segmentations

[81] Voir J. Dichy, *L'écriture dans la représentation de la langue : La lettre et le mot en arabe*, Lyon, 1990

proposées et le numéro de la segmentation encadré par deux flèches, droite et gauche, pour permettre de naviguer entre les différentes analyses possibles (au milieu) et enfin, le bouton qui permet d'enregistrer la segmentation choisie (à gauche). Au milieu de l'écran, est présentée la phrase contenant le mot graphique en question ou, quant celle-ci est trop longue (et c'est souvent le cas en arabe classique), le contexte le plus large possible qui permettrait à l'utilisateur de choisir en toute connaissance de cause, la bonne segmentation. Dans la partie inférieure de l'écran, le logiciel affiche la segmentation (proclitiques + mot minimal + enclitiques) correspondant au numéro choisi par l'utilisateur, parmi les analyses proposées pour le mot graphique en cours de segmentation.

Figure 7
Capture d'écran : choix de segmentation à faire parmi deux propositions présentées par le segmenteur

En plus de la méthode par règles et celle par dictionnaire sur lesquelles est basée la segmentation, nous avons doté notre segmenteur d'une méthode par apprentissage. Il ne s'agit pas ici d'apprentissage à partir de grands corpus d'entraînement préalablement segmentés comme le font certains segmenteurs pour d'autres langues, ce qui serait un

bon procédé, mais il s'agit plutôt d'un module, à la façon d'un OCR, qui reprend les mots graphiques qui n'ont pu être segmentés par le système puis l'utilisateur sélectionne, à partir d'un menu déroulant, un par un ces mots non segmentés et insère manuellement la segmentation voulue en saisissant directement sur clavier les proclitiques, le mot minimal et les enclitiques dans les champs correspondants comme le montre la figure suivante. Une fois la segmentation validée, le segmenteur enregistre cette analyse pour l'appliquer à l'avenir, à chaque fois qu'il aura à segmenter le même mot graphique.

Figure 8
Apprentissage du segmenteur : segmentation manuelle des mots graphiques non analysés par le segmenteur et que ce dernier devra enregistrer

Nous avons décidé de séparer, lors de la segmentation, par un caractère spécial différent de l'espace, non seulement les mots graphiques mais aussi les mots minimaux et les clitiques résultant du découpage des mots maximaux. C'est pourquoi le segmenteur a été programmé pour insérer une barre oblique avant et après chacune des unités lexicales obtenues à la sortie de la segmentation : un exemple de fichier de sortie est présenté ci-après, dans la figure 10.

En outre, étant donné que nous avons fait le choix de ne pas considérer la ponctuation pour notre corpus[82], nous avons programmé le segmenteur pour qu'il supprime, à chaque fois qu'il en rencontre un, les signes de ponctuation. Cette décision, bien entendu ponctuelle, ne concerne que ce corpus et peut donc être annulée à tout moment en supprimant, du code source, la ligne de commande correspondante. On pourrait même envisager de laisser l'utilisateur choisir de considérer ou non les signes de ponctuation et ce en lui permettant le paramétrage dans les options de segmentation.

Le texte soumis à la segmentation doit être sous format texte seul « .txt » ; le fichier de sortie rendu par le segmenteur et contenant le texte segmenté est aussi sous format texte mais avec l'extension « .seg », une extension que nous avons créée à cet effet.

Pour des raisons de commodité, nous avons choisi, pour segmenter le corpus, de le diviser en de petits fichiers correspondant chacun à une page de l'édition de référence.

Figure 9
Fichier à segmenter avant de le soumettre au segmenteur

[82] Voir les explication de ce choix dans la *norme de saisie et d'harmonisation*, chapitre 4

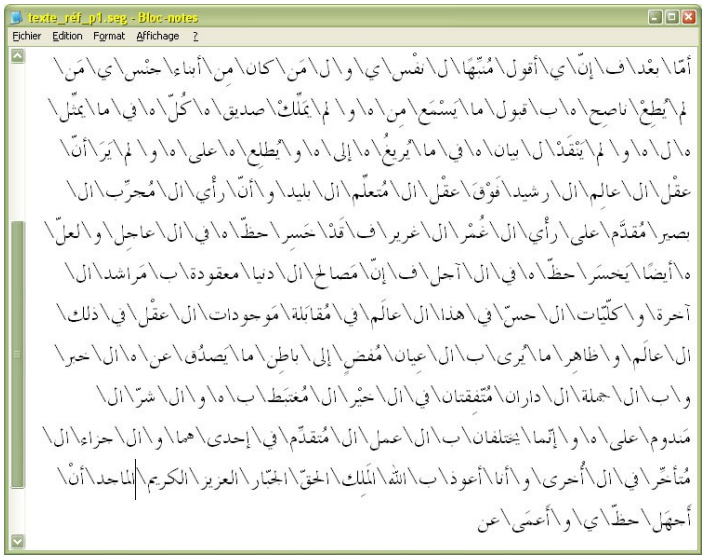


Figure 10
Le fichier de sortie après segmentation

En ce qui concerne les performances de notre segmenteur, elles sont visiblement correctes. Mais avant d'exposer les résultants obtenus en termes de pourcentage de réussite, il faut d'abord rappeler qu'une partie des mots graphiques, ceux qui présentent des analyses multiples, sont segmentés directement par l'utilisateur qui choisit, pour chaque mot graphique, la bonne segmentation parmi les différentes possibilités proposées par le segmenteur. Ce qui veut dire que pour ces mots graphiques, la segmentation validée est considérée comme totalement exacte parce que choisie en toute connaissance de cause par l'opérateur-linguiste (ou supposé comme tel). En revanche, les mots graphiques que le segmenteur a analysés et segmentés automatiquement sans demander l'avis de l'opérateur humain, peuvent comporter de toute évidence des erreurs qui peuvent être dues à plusieurs facteurs et que nous exposerons ci-après. Les mots graphiques non analysés du tout, parce que considérés comme mots minimaux ou non reconnus par le segmenteur, peuvent également renfermer des erreurs ; ils sont en effet, réinjectés tels quels à leurs places respectives dans le texte de sortie.

Sur les 37 457 mots graphiques[83] que comporte notre corpus, 34 027 ont été analysés par le segmenteur soit 90,84 % du corpus, et 3 430 mots graphiques n'ont pas pu être analysés soit 9,16 % de l'ensemble du corpus.

Pour juger de la fiabilité du segmenteur, nous avons donc calculé le taux d'erreur pour chacun des petits fichiers (correspondant aux pages) aussi bien pour les mots graphiques analysés que pour ceux que le segmenteur n'a pas pu segmenter. Le taux d'erreur général a par conséquent été calculé ; les résultats obtenus sont les suivants :

Pour les 90,84 % de mots graphiques qui sont analysées, le taux d'erreur varie entre 2,61 % et 16,86 % avec un taux moyen de 7,06 %. Rapporté à l'ensemble du corpus, ce taux d'erreur est donc de 6,41 %.

Parmi les 9,16 % de mots graphiques qui n'ont pas été analysés par le segmenteur, le taux de ceux qui auraient dû être segmentés est de 69 % (nous les considérons comme des erreurs). Rapporté à l'ensemble du corpus, ce taux d'erreur chute à 4,87 %.

Le taux d'erreur général est donc de 11,28 %. Autrement dit, l'efficacité du segmenteur est de l'ordre de 88,72 % (mots inconnus compris). Elle est tout de même de l'ordre de 93,59 % si nous ne considérons que les mots reconnus. Ce sont des performances largement correctes.

3.2. Correction des erreurs de la segmentation

Le taux d'erreur du segmenteur, nous venons de le voir, est de 11,28 %. Ces 11,28 % de notre corpus représentent quand même quelques 4 225 mots graphiques

[83] Voir le chapitre suivant *Dépouillement segmenté et dépouillement en mots graphiques : Données statistiques*.

erronément segmentés dont il faut corriger la segmentation, ou non segmentés du tout et qu'il faut segmenter manuellement. C'est ce à quoi nous nous attachons dans cette étape du dépouillement lexical du corpus.

Les erreurs rencontrées sont de plusieurs types et dues à plusieurs facteurs que nous résumons *grosso modo* dans le tableau de la page suivante :

La grande majorité des erreurs provient des ambiguïtés lexicales qu'elles soient virtuelles ou effectives[84]. Une infime partie des ces erreurs est due à un manque remarqué dans la base de donnée des mots-outils de DIINAR. Il est à noté à ce propos que les catégories lexicales que nous avons retenues pour notre travail et que nous expliquons dans la *norme de dépouillement*, apportent des modifications à la catégorisation ayant servi à la constitution de la base de donnée des mots-outils de DIINAR. De ce fait ce type d'erreurs dans la segmentation pourra disparaître dès que la BD DIINAR sera modifiée et le lexique du segmenteur régénéré. Un certain nombre d'erreurs, inattendues cette fois-ci, sont certainement dues à quelques anomalies au niveau des règles de la grammaire segmentale qu'il va falloir compléter. La particule de serment ت *ta*, par exemple, a cette particularité de n'être utilisée que devant le mot الله *allâh* [Dieu]. Dans certains mots graphiques comportant des noms commençant par la Lattre ت , le segmenteur a confondu cette première lettre faisant partie intégrante du nom avec la particule de serment en la considérant comme proclitique et a segmenté par là-même le mot graphique d'une façon incorrecte comme dans l'exemple suivant : و\ت\وصيل\ي alors qu'on s'attendait plutôt à و\توصيل\ي .

[84] Voir les types d'ambiguïté dans la norme de dépouillement, chapitre 5

Remarque : *Dans la deuxième colonne, « erreur » ne veut pas toujours dire que la segmentation proposée n'est pas correcte ; parfois elle l'est. Cela veut tout simplement dire que la bonne segmentation (que nécessite le contexte) n'est pas proposée.*

Mot graphique	Erreur de segmentation	Segmentation attendue	Remarque
إحداهما	إحدا\هما	إحدى\هما	Ces erreurs nécessitent le recours à une harmonisation régulatrice ou, mieux encore, l'ajout de quelques règles à la grammaire segmentale pour garantir la régulation de ces transformations graphiques post-segmentation
وكفاه	و\كفا\ه	و\كفى\ه	
عنّي	عنّ\ي	عن\ي	
أنّني	أنّ\ني	أنّ\ي	
أذقتني	أذقت\ني	أذقت\ي	
مُقلتي	مُقلت\ي	مُقلة\ي	
لدينا	لدي\نا	لدى\نا	
وأبوّته	و\أ\بوّة\ه	و\أبوّة\ه	
لاقيتَهم	لا\قيتَ\هم	لاقيتَ\هم	
وتوصيلي	و\ت\وصيل\ي	و\توصيل\ي	Ces erreurs inexplicables morpho-syntaxiquement sont sûrement dues à quelques anomalies au niveau de la grammaire segmentale à compléter
رقبتي	رقبتي\ي	رقبة\ي	
جاءتهم	جاءة\هم	جاءت\هم	
وكما	و\كَ\ما	و\كما	Ce type d'erreurs est dû à un manque dans la base de données des mots-outils
ولقدْ	و\ل\قدْ	و\لقدْ	
فلِمَ	فلِمَ	ف\لِمَ	
أباه	أبا\ه	أبى\ه	**On ne peut vraiment pas parler d'erreurs dans ces cas de figures puisque les anlyses faites par le segmenteur sont correctes, mais ne correspondent pas aux bonnes segmentations qui cadrent avec le contexte.** **Il s'agit bien ici de cas d'ambiguïtés effectives, pour les 6 premiers, et virtuelles, pour les 9 derniers.** (Pour le 6ème exemple de ce groupe, l'on a une ambiguïté effective quand le nom est au
أحَدُّهم	أ\حَدُّ\هم	أحَدُّ\هم	
معكَ	معكَ	مع\كَ	
جباها	جبا\ها	جبى\ها	
وأَرِنا	و\أَرِنا	و\أَرِ\نا	
وكماله	و\ك\مال\ه	و\كمال\ه	
وصولي	و\صول\ي	وصول\ي	
وأزمنته	و\أ\زمنة\ه	و\أزمنة\ه	
واستعارتها	و\استعارت\ها	و\استعارة\ها	
ولغته	و\لغت\ه	و\لغة\ه	
وصفناها	و\صفنا\ها	وصفنا\ها	

أخذته	أخذة\ه	أخذت\ه	génitif كمالِ, virtuelle dans les autres cas)
تبعته	تبعة\ه	تبعت\ه	
وأجنحتها	و\أ\جنحة\ها	و\أجنحة\ها	
دلّتنا	دلّة\نا	دلّت\نا	

Enfin, un dernier type d'erreurs est dû à des transformations graphiques anormales que subissent certains mots graphiques suite à l'opération de segmentation. En attendant l'ajout d'une panoplie de règles complétant la grammaire segmentale dont notre segmenteur est doté, ces erreurs nécessitent le recours à une harmonisation régulatrice rendant leurs graphies d'origine à ces mots.

3.3. Harmonisation régulatrice

Contrairement à l'harmonisation primaire qui intervient en amont, c'est-à-dire avant la segmentation, l'harmonisation régulatrice, elle, est effectuée en aval, une fois que la segmentation est opérée. Le recours à ce type d'harmonisation est nécessaire pour rendre leurs graphies initiales aux mots ayant subi, lors de la segmentation, des transformations graphiques dues à des considérations scripturales (écriture cursive : ة ت), orthographiques (أ ؤ), morphologiques (ي ِي), syntaxiques (ون و) ou phonétiques (عِنْدِ عِنْدَ). Nous exposons plus loin dans le chapitre consacré à la *norme de saisie et d'harmonisation*[85] tous les cas d'harmonisation régulatrice que nous avons pu répertorier, avec un tableau récapitulatif.

À ce stade de la constitution du corpus, l'harmonisation régulatrice est inévitable non seulement pour homogénéiser les graphies en réparant d'éventuels « dégâts » graphiques provoqués par la segmentation, surtout quand le segmenteur automatique n'est pas doté d'une grammaire de régulation, mais aussi pour contribuer à la

[85] Voir la *norme de saisie et d'harmonisation*, p. 144-195

désambiguïsation par le biais de certains filtres réducteurs d'ambiguïtés étroitement liés à la segmentation.

3.4. Désambiguïsation

En appliquant, avant la segmentation, les filtres réducteurs d'ambiguïtés correspondant à la voyellation et à l'harmonisation primaire, nous avons pu réduire un certain nombre d'ambiguïtés virtuelles. La segmentation et la l'harmonisation régulatrice étant faites, nous pouvons poursuivre cette tâche et aller au-devant des ambiguïtés effectives pour les réduire sinon les lever totalement pour faciliter le travail à l'étape suivante qui est la lemmatisation.

Il est vrai que la segmentation contribue, nous l'avons vu, à lever certaines ambiguïtés, surtout les ambiguïtés agglutinantes ; mais d'un autre côté, elle peut aussi engendrer d'autres ambiguïtés, segmentales cette fois-ci. Ce qui donne à cette entreprise, ici, toute sa place et sa nécessité. La lemmatisation est une opération importante et délicate, et une désambiguïsation qui la précède, ne serait-ce que manuelle, ne peut apporter à l'opérateur humain ou, surtout, au récepteur machine que du confort et de la clarté.

3.5. Lemmatisation

Les mots graphiques segmentés, les ambiguïtés virtuelles et une grande partie des ambiguïtés effectives levées, arrive maintenant une étape cruciale et délicate du dépouillement lexical : la lemmatisation. Elle peut être définie comme étant l'opération qui consiste à rassembler, sous une même forme canonique appelée *lemme*, toutes les formes fléchies d'un texte ne différant que par des modalités ou flexions grammaticales (conjugaison, déclinaison, genre, nombre, etc.).

Du côté de la langue, donc de son lexique, les lemmes servent d'adresses lexicales aux formes qui, elles, ne sont que les actualisations des premiers dans le vocabulaire du discours. Ainsi, ramener les formes fléchies à leurs adresses lexicales correspondantes et « définir les *lemmes* de manière pertinente, contraint le "lemmatiseur" décrivant le vocabulaire d'un auteur et (re)construisant progressivement le lexique d'une langue, à porter un regard sur cette langue et à l'analyser »[86].

Si la flexion est le fait de passer d'une forme générique, d'un lemme, à sa forme fléchie (par exemple, de l'adjectif جَمِيل *jamîl* [beau] à celui جَمِيلاَت *jamîlât* [belles] ou du verbe كَتَبَ *kataba* [écrire] à sa forme conjuguée يَكْتُبُونَ *yaktubûna* [Ils écrivent]) ; à l'inverse, la lemmatisation est le fait de passer d'une forme fléchie au lemme correspondant. Cette opération qui paraît simple de par sa définition, est en réalité épineuse quant à son application parce qu'elle nécessite des prises de décisions lexicologiques importantes.

Nous présenterons les différents choix que nous avons faits pour définir les lemmes ainsi que les règles et les critères de lemmatisation que nous avons établis, au chapitre consacré à la *norme de dépouillement*[87]. Mais nous pouvons d'ores et déjà, présenter sommairement ci-après les lemmes les plus notables que nous avons retenus :

- Pour les **verbes** : le verbe conjugué à la 3ème personne du singulier masculin de l'accompli actif, exception faite d'une dizaine de verbes qui ne peuvent être qu'à la voix passive comme par exemple, يُحْتَضَرُ *yuÎtaÃaru* [agoniser].

- Pour les **noms** : le singulier, exception faite des pluriels qui n'ont pas de singulier, comme أَبَابِيل *Þabâbîl* [volées/troupes d'oiseaux], ou de ceux dont le singulier n'a pas la même racine comme نِسَاء *nisâÞ* [femmes] qui est le

---

[86] B. Coulie (dir. de), *Projet de Recherche en Lexicologie Grecque*, Université catholique de Louvain, Institut orientaliste, sur le site : http://www.fltr.ucl.ac.be/FLTR/GLOR/lexico/PAGE.HTM , lien: lemmatisation, dernière consultation le 27-05-2007.

[87] chapitre 5, p. 197-302

pluriel de اِمْرَأَة *imraÞa*. Il est à noter que même le pluriel de pluriel est ramené au singulier ; sous le lemme par exemple, بَيْـت *bayt* [maison] seront rassemblés le pluriel بُيُوت *buyût* et le pluriel de pluriel بُيُوتَات *buyûtât*.

- Pour les **adjectifs** : le singulier masculin. Il faudra se garder d'un petit nombre d'adjectifs masculins se terminant par un تـاء مربوطـة *tâÞ marbûÔa* [t fermé final] (qui est le plus souvent la marque du féminin) comme فَرُوقَـة *farûqa* [peureux], et de certains intensifs, beaucoup plus nombreux, qui sont construits sue le schème فَعَّالَة *faÝÝâla* comme par exemple, رَحَّالَة *raÎÎâla* [très grand voyageur], ou عَلاَّمَـة *Ýallâma* [très grand savant] qui sont, en fait, des adjectifs masculins.

- Pour les **singulatifs** : le lemme sera le collectif اسـم الجنس الجمعي *ism al-jins al-jamÝiyy* à partir duquel a été formé le singulatif اسـم الوحـدة *ism al-waÎda* , comme par exemple, pour سَمَكَة *samaka* [un poisson] le lemme est سَمَك *samak* [(du) poisson]. Le lemme du pluriel (du singulatif) سَمَـكَات *samakât* [des poissons] est également le collectif.

- Pour les **Mots-outils** : généralement le lemme est la forme elle-même sauf pour quelques mots-outils à savoir :

  - Les **cardinaux** : le nombre cardinal qui correspond à un nombré masculin, au cas sujet et isolé (c'est-à-dire non annexé, exemple أَلْفَانِ et non أَلْفَا).

  - Les **ordinaux** : le nombre ordinal masculin (étant donné qu'ils se comportent comme des adjectifs). Exception faite des dizaines, des centaines et des milliers où c'est un cardinal qui est utilisé pour exprimer un adjectif ordinal, exemple - الليلة الأربعون–المسألة المئة

    الرجل الألف.

  - Les **verbes figés** : comme ils sont figés soit à l'accompli, soit à l'impératif, soit à l'inaccompli (un seul verbe), le lemme sera le verbe lui-même pour les premiers, le verbe à l'impératif pour les deuxièmes et, enfin, la même forme pour le dernier.

- Les **annectifs*** كِلاَ et كِلْتَا : ces deux annectifs ont la particularité d'être toujours au cas sujet quand ils sont premier terme d'une annexion, mais deviennent déclinables quand des pronoms personnels leur sont suffixés. Auquel cas, c'est toujours la forme au cas sujet qui est le lemme (كِلاَ et كِلْتَا et non كِلَيْ et كِلْتَيْ)

- Pour les **cinq noms** : la forme tronquée (أَب et non أَبَا , أَبُو ou أَبِي), au cas sujet (ذُو et non ذَا ou ذِي)

- Pour les **noms propres** : au cas sujet (أبو الوفاء et non أبا الوفاء ou ,أبي الوفاء) (... اِمْرُؤُ القَيْس.) et non اِمْرِئَ القَيْس)

- Pour certains **duelatifs***□ : le lemme est la même forme qui est au duel et ne doit surtout pas être ramenée au singulier. Le duelatif par exemple أَبَوَانِ *ƀabawâni* [les (deux) parents], littéralement en arabe « *les deux pères* » n'est pas l'addition d'un père et d'un autre père, mais bien d'un père et d'une mère.

Il faut préciser que l'opération de lemmatisation, nous l'avons faite totalement à la main. Nous avons pourtant établi, en collaboration avec R. Zaafrani, un module de lemmatisation basé sur le segmenteur semi-automatique, seulement les tests n'ont pas été concluants ; le temps que nous avons mis à utiliser le lemmatiseur était beaucoup plus important que celui d'une lemmatisation manuelle. En effet, non seulement le nombre des erreurs était important, mais surtout l'explosion combinatoire provoquée par l'analyse des unités lexicales est tellement considérable que le recours à une lemmatisation manuelle était inévitable. Les lemmes proposés par le segmenteur, même concernant les unités lexicales les plus simples pour l'opérateur humain, sont à chaque

□ et ** Nous présenterons dans la *Norme lexicologique* (dans la partie *Catégorisation*, p. 260-302), ce que nous appelons ***Annectifs*** et ***Duelatifs***.

□

□

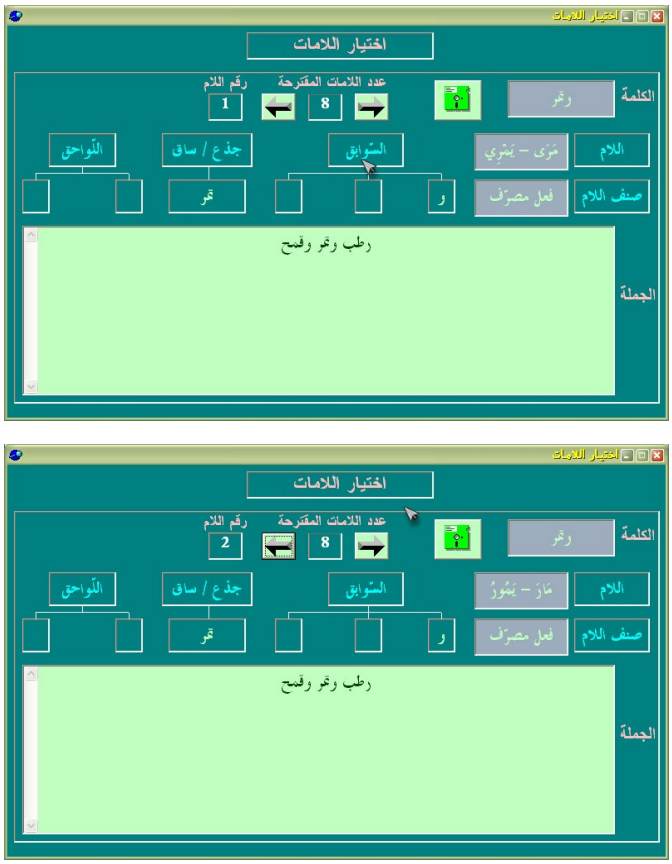
fois, en moyenne, multipliés par dix. Nous en donnons quelques exemples dans le tableau suivant :

Mot	Lemmes		Catégories
أثر	1-	ثَرَا – يَثْرُو	Verbe
	2-	وَثَّرَ – يَثِرُ	Verbe
	3-	ثَرَى – يَثْرِي	Verbe
	4-	ثَرِيَ – يَثْرَى	Verbe
	5-	ثَارَ – يَثُورُ	Verbe
	6-	أَثْرَى –يُثْرِي	Verbe
	7-	أَثَارَ – يُثِيرُ	Verbe
	8-	أثَر	**Nom**
	9-	أثِر	**Déverbal**
	10-	أَثَرَ – يَأْثُرُ	Verbe
	11-	أَثَرَ – يَأْثِرُ	Verbe
	12-	أَثِرَ – يَأْثَرُ	Verbe
ألف	1-	لَفَا – يَلْفُو	Verbe
	2-	وَلَفَ – يَلِفُ	Verbe
	3-	لاَفَ – يَلُوفُ	Verbe
	4-	لاَفَ – يَلِيفُ	Verbe
	5-	أَلْفَى – يُلْفِي	Verbe
	6-	ألف	**Nom**
	7-	أَلَفَ – يَأْلِفُ	Verbe
	8-	أَلِفَ – يَأْلَفُ	Verbe
أحد	1-	حَدَا – يَحْدُو	Verbe
	2-	وَحَدَ – يَحِدُ	Verbe
	3-	حَدَى – يَحْدِي	Verbe
	4-	حَادَ – يَحُودُ	Verbe
	5-	حَادَ – يَحِيدُ	Verbe
	6-	أَحَادَ – يُحِيدُ	Verbe
	7-	أحد	**Nom**

تبع	**1-**	بَعَا – يَبْعُو	Verbe
	2-	بَعُوَ – يَبْعُو	Verbe
	3-	بَاعَ – يَبُوعُ	Verbe
	4-	بَاعَ – يَبِيعُ	Verbe
	5-	أَبْعَى – يُبْعِي	Verbe
	6-	أَبَاعَ – يُبِيعُ	Verbe
	7-	تَبَع	**Nom - Déverbal**
	8-	تِبْع	**Déverbal**
	9-	تَبِعَ – يَتْبَعُ	Verbe
تمر	**1-**	مَرَى – يَمْرِي	Verbe
	2-	مَارَ – يَمُورُ	Verbe
	3-	مَارَ – يَمِيرُ	Verbe
	4-	أَمْرَى – يُمْرِي	Verbe
	5-	أَمَارَ – يُمِيرُ	Verbe
	6-	تمر	**Nom**
	7-	تَمَرَ – يَتْمُرُ	Verbe
	8-	تَمِرَ – يَتْمَرُ	Verbe
لبن	**1-**	لاَبَ – يَلُوبُ	Verbe
	2-	وَلَبَ – يَلِبُ	Verbe
	3-	لَبَن	**Nom**
	4-	لَبِن	**Déverbal**
	5-	لَبَنَ – يَلْبُنُ	Verbe
	6-	لَبِنَ – يَلْبَنُ	Verbe

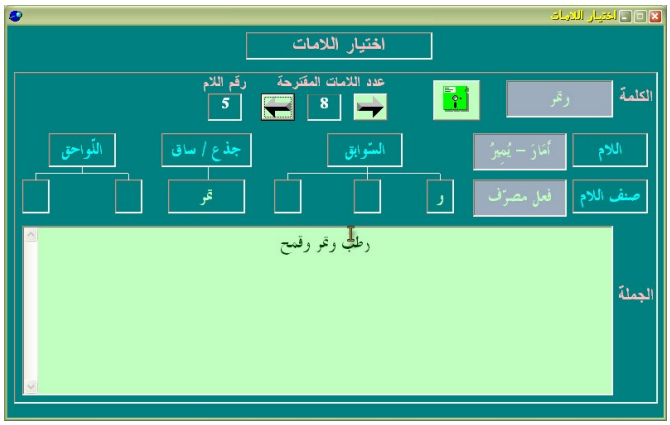
Comme pour la segmentation, à chaque fois que l'analyse repère plus d'une solution, le lemmatiseur propose tous les lemmes trouvés et c'est à l'utilisateur de choisir le lemme adéquat et correspondant au contexte. Étant donné le nombre important des lemmes proposés, l'opération de lemmatisation semi-automatique devient très pesante. Il faut dire que même pour les unités lexicales où un opérateur humain proposerait très intuitivement un lemme extrêmement évident comme, par exemple le

pronom هـم *hum* [eux] ou le démonstratif هذا *hâÆâ* [ceci/celui-ci], le lemmatiseur (de par la complétude et la robustesse du système qui, paradoxalement, se traduit ici par une lourdeur au niveau de la manipulation) propose trois lemmes pour le premier (وَهَـمَ/يَهِـمُ *wahama/yahimu* [se faire des illusions], هَامَ/يَهِيـمُ *hâma/yahîmu* [errer comme un fou] et هـم *hum* [eux]) et deux pour le second (هَـذَا/يَهْـذُو *haÆâ/yahÆû* [délirer] et هـذا *hâÆâ* [ceci/celui-ci]). La lourdeur de l'opération vient du fait qu'on est obligé, quasiment pour chaque unité lexicale, de naviguer entre les nombreux écrans proposant les divers lemmes pour choisir le lemme recherché. Comme le montre, par exemple, la figure 11 suivante, l'on est obligé d'aller jusqu'à l'écran n° 6 (sur 8) pour valider le lemme تمر *tamr* [datte] de la forme تمر .




اختيار اللامات
الكلمة
وتر
عدد اللامات المقترحة
8
رقم اللام
3
اللام
مَارَ – يَمِيرُ
صنف اللام
فعل مصرّف
السّوابق
و
جذع / ساق
وتر
اللّواحق
الجملة
رطب وتر وقمح


اختيار اللامات
الكلمة
وتر
عدد اللامات المقترحة
8
رقم اللام
4
اللام
أَمْرَى – يُمْرِي
صنف اللام
فعل مصرّف
السّوابق
و
جذع / ساق
وتر
اللّواحق
الجملة
رطب وتر وقمح


اختيار اللامات
الكلمة
وتر
عدد اللامات المقترحة
8
رقم اللام
5
اللام
أَمَارَ – يُمِيرُ
صنف اللام
فعل مصرّف
السّوابق
و
جذع / ساق
وتر
اللّواحق
الجملة
رطب وتر وقمح

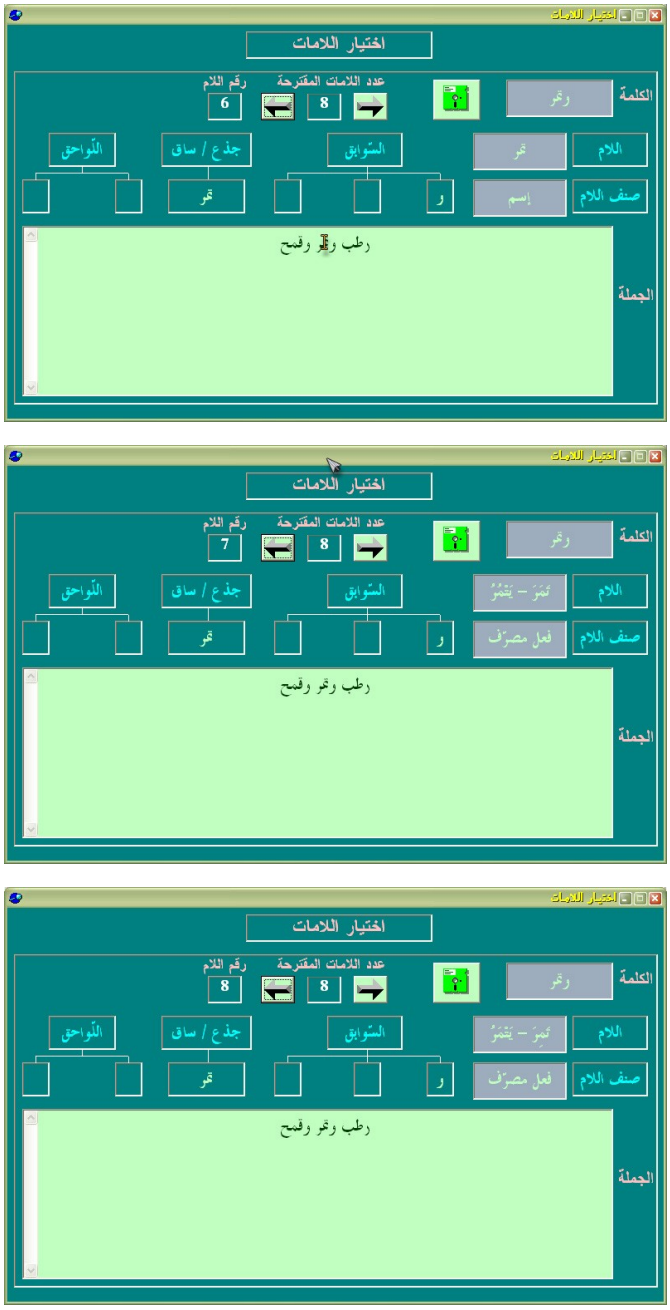

اختبار اللامات
الكلمة
رقم اللام
6
عدد اللامات المقترحة
8
8
رقر
اللام
قمر
صنف اللام
إسم
السّوابق
ر
جذع / ساق
قمر
اللّواحق
الجملة
رطب رقر وقمح
اختبار اللامات
الكلمة
رقم اللام
7
عدد اللامات المقترحة
8
8
رقر
اللام
ثمِرَ – يَثمَرُ
صنف اللام
فعل مصرّف
السّوابق
ر
جذع / ساق
قمر
اللّواحق
الجملة
رطب رقر وقمح
اختبار اللامات
الكلمة
رقم اللام
8
عدد اللامات المقترحة
8
8
رقر
اللام
ثمِرَ – يَثمَرُ
صنف اللام
فعل مصرّف
السّوابق
ر
جذع / ساق
قمر
اللّواحق
الجملة
رطب رقر وقمح

Figure 11

3.6. Catégorisation

Pouvant être considérée soit comme une opération indépendante de la lemmatisation au sens strict et succédant à celle-ci, soit comme le deuxième volet de la lemmatisation au sens large après un premier volet d'identification, la catégorisation est cette opération qui consiste à assigner à chaque lemme sa catégorie lexicale correspondante. Cette opération est d'autant plus aisée que les catégories lexicales sont clairement définies et nettement délimitées. Cependant, certains cas d'ambiguïtés mono- et polycatégorielles peuvent venir entacher le bon déroulement de cette entreprise. C'est pourquoi, la mise en place des filtres réducteurs d'ambiguïtés avant la segmentation ainsi que le recours aux procédés de désambiguïsation[88] avant la lemmatisation peuvent être d'un grand secours à l'opérateur humain ou au programme informatique, préparant le terrain à une catégorisation réussie. Réciproquement, le fait d'affecter sa catégorie lexicale à chaque lemme, contribue d'une manière efficace à la levée d'ambiguïtés effective. Mais une bonne catégorisation commence toujours par la définition des catégories lexicales. Quelles sont donc les catégories lexicales de l'arabe que nous avons retenues, quel est le code fixé pour chacune d'entre elles et comment avons-nous procédé pour l'insertion, à côté de chaque lemme, du code de la catégorie lexicale correspondante ?

Cependant, le choix des critères de classification est certes important, mais ce qui l'est encore davantage dans cette entreprise de catégorisation, c'est la stabilité d'application de ces mêmes critères de classement.

---

[88] Voir la désambiguïsation au chapitre 5 « *Norme de dépouillement* ».

3.6.1. Les catégories lexicales retenues

La présentation détaillée des catégories lexicales que nous avons retenues ainsi que leur code sera faite *infra* dans le chapitre « *Norme de dépouillement* »[89]. Par souci de simplification, nous ne présentons ici, et d'une manière très sommaire, que les catégories de base (verbes, noms primitifs, noms dérivés, adjectifs, noms composés et mots-outils sans oublier les noms propres) et les catégories qui sont associées à chacune d'entre elles (verbes simples, noms primitifs simples, diminutif, participe actif, etc. pour les mots lexicaux, et prépositions, démonstratifs, verbes figés, etc. pour les mots-outils). D'une manière générale, les classes lexicales sont hiérarchisées en quatre niveaux allant de la catégorie de base à la sous sous-catégorie, passant par la catégorie et la sous-catégorie ; c'est ce que nous schématisons dans le tableau suivant :

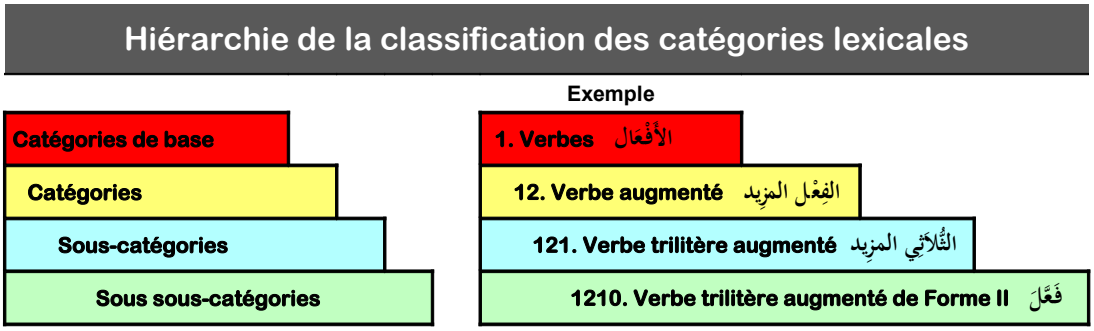


Figure 12

Sept catégories de base ont été retenues ; cinq catégories de base pour les mots lexicaux et qui représentent des listes ouvertes (Verbes, Noms Primitifs, Noms dérivés, Adjectifs et Noms Composés) et une catégorie de base pour les mots-outils qui est une liste fermée. Comme dans toutes les langues, les Noms Propres constituent en fait une classe à la périphérie du lexique ; pour la nécessité des études lexicométriques, nous les avons retenus et nous en avons fait une catégorie de base. 29 catégories se répartissent entre les 7 catégories de base : 2 catégories pour les Verbes, 6 pour chacune des catégories de base des Noms Primitifs, des Noms dérivés et des Adjectifs, 4 pour les

[89] *Idem*.

Mots-Outils et 5 pour les Noms Propres. Les sous-catégories quant elles, sont au nombre de 44 dont la part du lion (20 sous-catégories) revient à la catégorie des Particules qui appartient à la catégorie de base des Mots-Outils. Nous avons retenu enfin, 16 sous sous-catégories : 10 dans la sous-catégorie des Verbes Trilitères Augmentés, 3 dans celle des Verbes Quadrilitères Augmentés et 3 dans celle des Noms-Outils.

Voici maintenant comment se présentent, dans l'ordre et avec le code associé, les catégories de base et les catégories dont elles font partie :

- Verbes الأفعال Code : 1
    - Verbe simple الفعل المجرّد Code : 11
    - Verbe augmenté الفعل المزيد Code : 12

- Noms primitifs الأسماء الأصليّة Code : 2
    - Nom primitif simple الاسم الأصليّ المجرّد Code : 21
    - Nom d'unité اسم الوَحدة Code : 22
    - *MaÒdar* primitif المصدر الأصلي Code : 23
    - Nom d'une fois اسم المرّة Code : 24
    - Nom de manière اسم الهيئة Code : 25
    - Nom augmenté الاسم المزيد Code : 26

- Noms dérivés الأسماء المشتقّة Code : 3
    - Noms de temps et de lieu اسما الزمان والمكان Code : 31
    - Nom d'instrument اسم الآلة Code : 32
    - Diminutif اسم التصغير Code : 33
    - *MaÒdar mîmî* المصدر الميمي Code : 34
    - *MaÒdar ÒinâŶî* المصدر الصناعي Code : 35
    - *MaÒdar* dérivé المصدر المشتقّ Code : 36

- Adjectifs الصفات Code : 4

- Participe actif اسم الفاعل Code : 41
- Participe passif اسم المفعول Code : 42
- Intensif اسم المبالغة Code : 43
- Elatif اسم التفضيل Code : 44
- *Ñifa mušabbaha* الصفة المشبّهة Code : 45
- Adjectif de relation الاسم المنسوب Code : 46

- Mots-Outils الأدوات Code : 5
  - Particules الحروف Code : 51
  - Noms-Outils الأسماء الأدوات Code : 52
  - Verbes fonctionnalisés الأفعال الجامدة Code : 54
  - Mots-Outils composés الأدوات المركّبة Code : 55

- Noms propres أسماء الأعلام Code : 6
  - Noms Propres de personnes أسماء الأشخاص Code : 61
  - Noms Propres de lieux أسماء الأماكن Code : 62
  - Noms de tribus, groupes et nations أسماء القبائل والأمم والفرق Code : 63
  - Théonymes الأسماء الدينيّة Code : 64
  - Noms des œuvres أسماء الكتب Code : 65

- Noms composés الأسماء المركّبة Code : 7

3.6.2. Utilisation des macros de Word® pour insérer les codes

La macro-commande, communément appelée macro, est un ensemble d'actions que l'on utilise pour automatiser des tâches dans des applications Microsoft. Les macros

sont enregistrées dans le langage de programmation Visual Basic pour Applications (VBA)[90].

Pour insérer d'une manière semi-automatique le code correspondant à chaque catégorie lexicale pour chacun des lemmes de notre corpus, nous avons donc utilisé, dans l'environnement de l'éditeur de texte Word de Microsoft, deux types de macro-commandes :

- Une première, simple et prédéfinie, qui est l'insertion automatique et que nous avons paramétrée pour avoir, côté intitulé de l'insertion, un libellé en toutes lettres clair et significatif et, côté code à insérer, le code de la sous-catégorie concernée précédé du symbole « # » et qui seront tous deux collés au lemme en question. Les codes à insérer sont organisés en sous sous-menus (sous-catégories), sous-menus (catégories) et menus (catégories de base) ; ces derniers sont intégrés à une nouvelle barre de menus (barre des catégories lexicales) que nous avons créée et affichée sous les barres d'outils de Word® (Voir les captures d'écran ci-dessous). Pour insérer un code, il suffit donc, après avoir placé le curseur juste à la fin du lemme à coder, de dérouler le menu correspondant à la catégorie de base voulue (exemple : Verbes), le sous-menu de la catégorie souhaitée (exemple : Verbe Dérivé), le sous sous-menu concerné (exemple : Verbe Trilitère Dérivé) et cliquer enfin sur la sous-catégorie de niveau 2 correspondant au verbe dont on veut insérer le code (exemple : VI- – تَفَاعَلَ يَتَفَاعَلُ). Le verbe en question se voit alors adjoindre son code et le symbole dièse (exemple : تَضَاعَفَ#1214).

- La deuxième macro est une macro plus complexe que nous avons programmée nous-même en VBA et qui a la fonction, d'abord de chercher dans tout le corpus toutes les occurrences du lemme dont nous venons d'insérer le code avec la première

[90] VBA est une version macrolangage de Microsoft Visual Basic qui sert à programmer des applications Windows. VBA est fourni avec plusieurs applications Microsoft, notamment la suite bureautique Office.

macro, et d'insérer ensuite ce même code à toutes les occurrences trouvées dans le texte. Nous balayons ainsi au fur et à mesure tout le corpus à la recherche des lemmes non encore codés[91], c'est-à-dire dont la première occurrence n'a pas encore été rencontrée, pour procéder consécutivement à l'activation de la première puis de la deuxième macro.

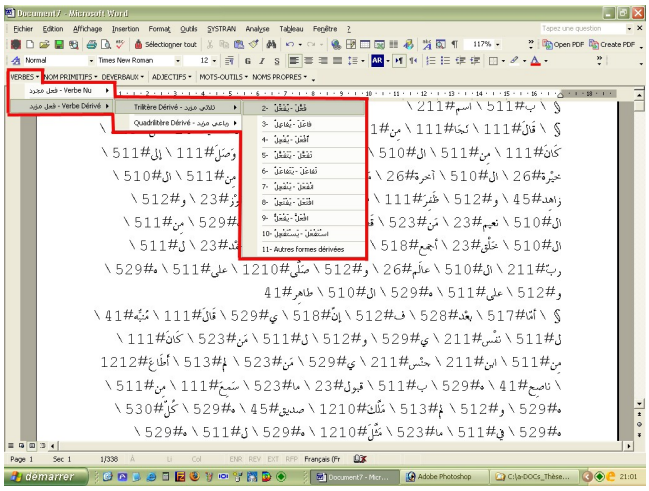


Figure 13
Capture d'écran : insertion du code des verbes

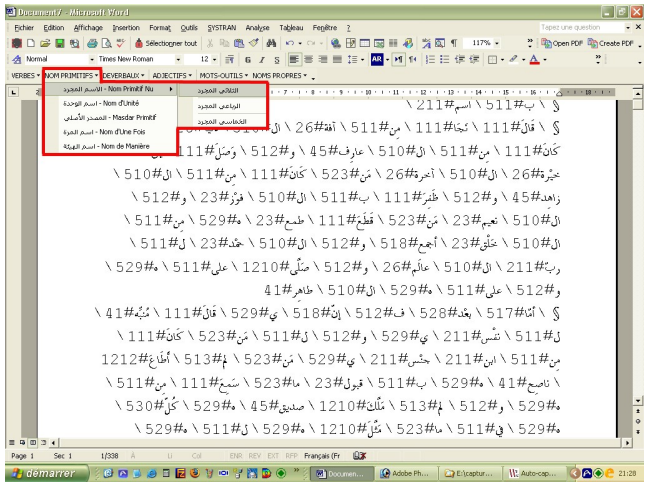


Figure 14
Capture d'écran : insertion du code des noms primitifs

[91] En plus du balayage manuel (visuel) pour rechercher les lemmes non encore codés, nous avons aussi utilisé les fonctions régulières (ou rationnelles) comprises dans la commandes « Rechercher » de Microsoft Word.

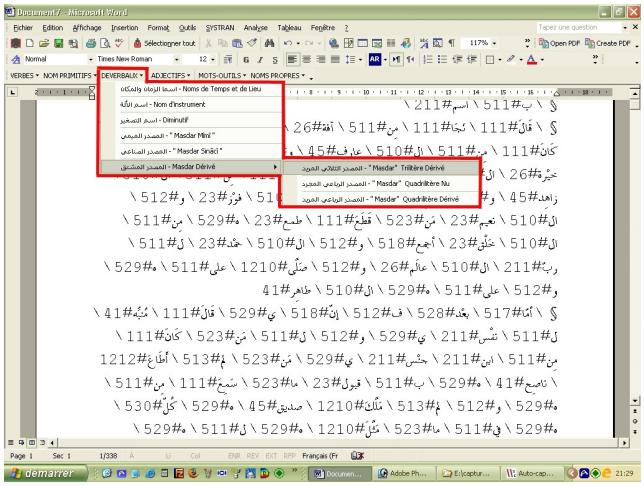


Figure 15
Capture d’écran : insertion du code des noms dérivés

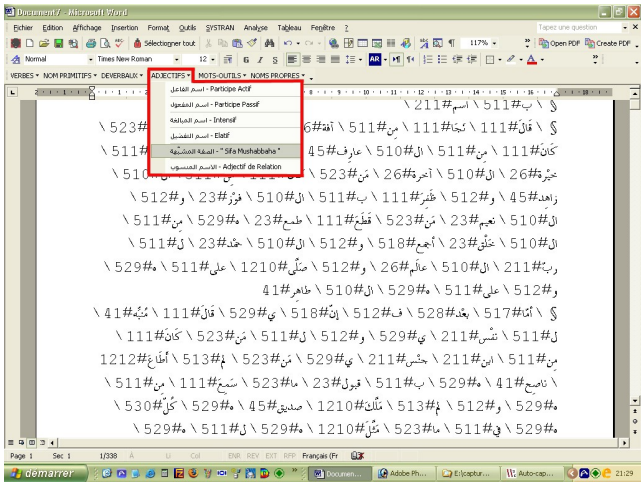


Figure 16
Capture d’écran : insertion du code des adjectifs

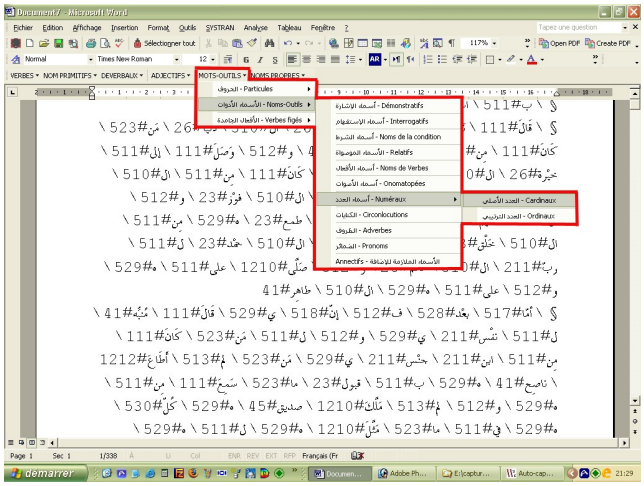


Figure 17
Capture d'écran : insertion du code des mots-outils

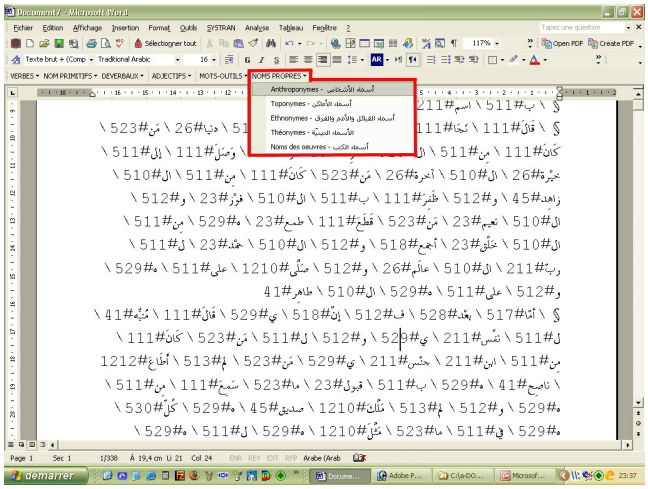


Figure 18
Capture d'écran : insertion du code des noms propres

Comme il est clair de ces captures d'écran, nous avons tenu à opérer une catégorisation directement en contexte. Cela a pour intérêt de minimiser les risques d'erreurs et de résoudre facilement les rares cas d'ambiguïté qui ont pu passer entre les mailles des filtres réducteurs d'ambiguïtés et les différentes étapes de désambiguïsation.

Ce que nous obtenons à la fin de cette catégorisation est un fichier contenant à la fois les lemmes et les catégories lexicales correspondantes. Ce type de fichier est bien

utile pour différentes études contextuelles et pour d'éventuelles corrections et retours en arrière. Mais ce dont il est question, pour la suite des étapes de traitement statistique, c'est de ramener le texte à une séquence de codes numériques en vue d'une étude quantitative des catégories lexicales. C'est pour cette raison qu'à la fin de cette opération de catégorisation, les codes vont remplacer les lemmes eux-mêmes et le fichier obtenu ne comportera que les codes numériques des catégories lexicales.

Ces codes numériques vont nous permettre de déterminer les effectifs et les fréquences des catégories lexicales aussi bien au niveau du corpus qu'au niveau de ses parties. Des comparaisons peuvent donc être faites entre ces parties sur la base des écarts relevés entre les effectifs réellement observés et les effectifs théoriques calculés par des moyens statistiques. Ces codes peuvent également nous permettre d'étudier les collocations lexicales dans un corpus par la méthode par exemple, des segments répétés[92] appliquée aux catégories lexicales et non aux formes ou aux lemmes eux-mêmes. Dans ce cas, l'analyse statistique n'est pas faite sur des codes pris séparément mais sur des séquences de n codes numériques.

[92] Voir à cet effet les travaux d'André Salem, à savoir son excellet livre *Pratique des segments répétés. Essai de statistique textuelle*, 1987.

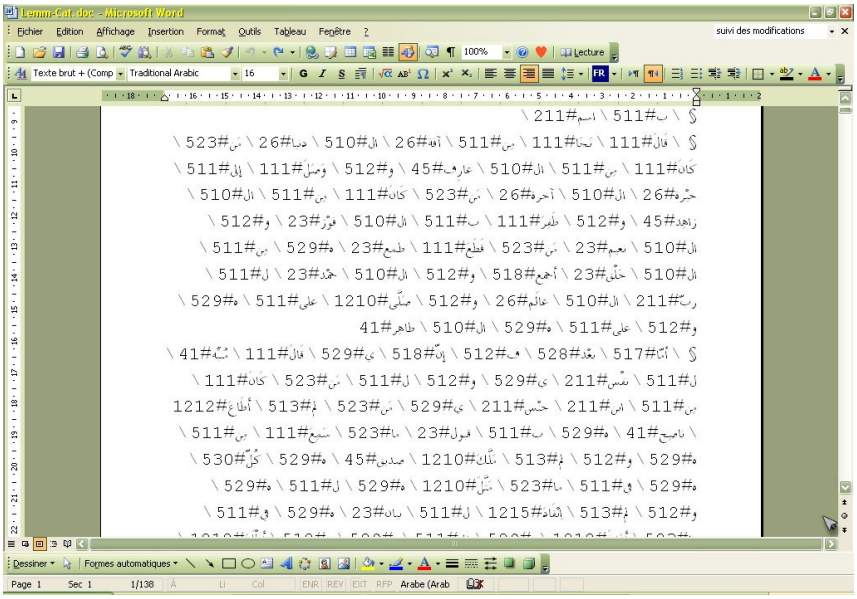


Figure 19
Capture d'écran : à la fin de l'opération de catégorisation chaque lemme est suivi du code de la catégorie lexicale correspondante

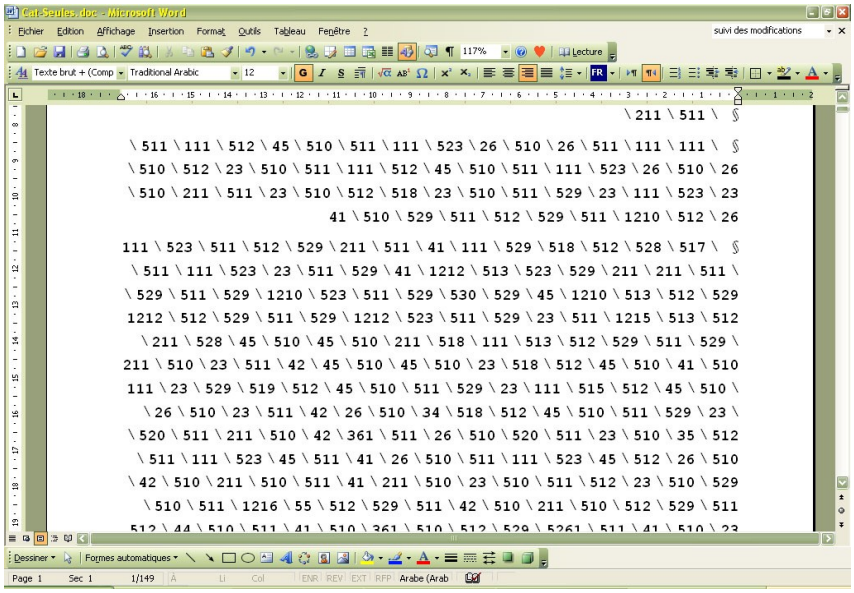


Figure 20
Capture d'écran : fichier final de la catégorisation et où les codes ont remplacé les lemmes

3.7. Balisage

Le balisage est la dernière étape du dépouillement ou prétraitement avant la phase du traitement lui-même, et avant donc de soumettre le texte au logiciel de traitement lexicométrique ***Lexico3***. Il est à noter que le balisage pourrait se faire avant même la segmentation[93] ; nous avons en effet tenté cette entreprise, mais avons découvert que les balises avaient cruellement perturbé, pour ne pas dire paralysé, l'opération de segmentation semi-automatique. Notre segmenteur ne contenant pas de règles lui permettant de détecter et de traiter ou d'ignorer toute une panoplie de balises ne faisant pas partie du texte-source à segmenter.

Lorsque l'on décide d'enregistrer en machine un texte à des fins lexicométriques, un certain nombre de problèmes de nature différente se présentent. Ils sont d'ordre orthographique, typographique, organisationnel, séquentiel, etc. Il est primordial que ces problèmes soient réglés avant d'entreprendre l'accumulation de longues séquences textuelles qui serviront de base aux comptages statistiques. Ces problèmes et difficultés naissent en fait du souci, d'un côté de fournir au chercheur en lexicométrie un texte informatisé identique ou presque à celui de l'édition de référence, et de l'autre lui donner, par le biais de son outil informatique, l'ordinateur, le pouvoir de repérer le texte-source à tous les pas du texte informatisé et d'effectuer sur celui-ci toutes les opérations que nécessite l'étude lexicométrique. Ces opération sont, entre autres, la délimitation et la localisation des unités de décompte et de leurs voisinages, l'examen des chaînes où elles sont insérées, la possibilité de séparer les phrases où elles occurrent, la capacité de calculer des fréquences, des sous-fréquences, des spécificités, etc.

---

[93] Un balisage du texte sans segmentation peut constituer un bon choix si l'on veut faire une étude lexicométrique des mots graphiques ou comparative entre texte segmenté et texte non ségmenté ; mais à partir du moment où l'on veut segmenter le texte, il est impératif soit de supprimer les balises soit de procéder à la segmentation à partir d'une autre version non balisée du texte. C'est ce dernier choix que nous avons fait.

Cette double obsession du respect formel et référentiel du texte-source et du pouvoir d'agir sur le texte informatisé est satisfaite par un dispositif de balises précises qui accompagnent les mots du texte dans leur séquence.

Sans aller jusqu'à prendre en considération un certain nombre d'aspects non-textuels de mise en page tels que : dénivellements textuels, nombre de blancs, lignes sautées, indentation de paragraphes, etc., ce dont rendent compte les balises proposées dans la norme Saint-Cloud[94], ce sont plutôt les formes typographiques majeures.

Outre les modalités typographiques que peut contenir un texte et dont le système de balisage doit rendre compte, son organisation logique (en parties, chapitre, actes, scènes, etc.), sa forme éditoriale (subdivision en volumes, pages, paragraphes, etc.) doivent également être prises en considération et respectées dans l'édition informatique du texte à traiter.

Non moins importantes et très utiles pour une éventuelle étude énonciative, analyse diachronique, thématique ou générique, les variables de situation sont, elles aussi, des informations très précieuses que le balisage doit respecter.

Les balises peuvent aussi comporter des informations multiples concernant l'archivage, l'identification du texte au sein de la base de données textuelle ainsi que les commentaires relatifs aux différentes balises, qu'elles soient prédéfinies ou supplémentaires que tout utilisateur peut ajouter.

Les balises prédéfinies, ou clés selon la terminologie de la norme Saint-Cloud, sont au nombre de cinq ; on trouve parmi elles : la balise F (comme Fichier) qui attribue un numéro d'identification à l'enregistrement réalisé, la balise C (comme Commentaire) qui signale des renseignements hors-texte, la balise L (comme Lettre) qui fournie une information relative à la typographie de l'édition-source, la balise E (comme Edition)

---

[94] Voir : P. Lafon, J. Lefevre, A. Salem, M. Tournier, *Le Machinal. Principes d'enregistrement informatique des textes*, Paris, 1985. La présentation des balises (clés selon la terminologie du *Machinal*) faite plus loin est largement basée sur ce même *Machinal*.

qui indique une information relative au découpage du texte-source et enfin, la balise S (comme Situation) qui introduit, elle, une information concernant les conditions de production du discours.

Syntaxe d'une balise[95]

Pour qu'elle soit reconnue et acceptée par ***Lexico3*** et donc opérationnelle, une balise (clé) doit se composer de cinq éléments juxtaposées (aucun espace n'est autorisé entre eux) dans l'ordre :

- un chevron ouvrant :" < " ;
- un **type** : constitué d'une espèce symbolisée par l'un de 5 majuscules, F, C, L, E ou S, suivie, dans le cas de E ou S seulement, d'un suffixe composé soit de deux minuscules soit de deux chiffres.
- le signe " = " : qui sépare le type du contenu ;
- un **contenu** : qui est un message, un numéro d'identification, un code ou toute autre valeur de variable ;
- un chevron fermant : " > "

Les balises se présentent donc sous la forme suivante : **<type=contenu>**

Exemple : <Auteur=Jaw Jidi> , <Volume=01> , <Nuit=07> ,

Il y a donc en tout cinq types de balises (clés) prédéfinies. Trois ne nécessitant pas de suffixe et deux exigent l'utilisation de suffixes.

- **Les balises sans suffixe**

  ↳ **La balise « F » :** *N° du fichier :* **<F=XXX>**

Cette balise n'entre pas effectivement dans la réalisation de l'édition électronique du corpus. Elle sert principalement à l'archivage et à l'identification du fichier au sein de la base de données textuelle. L'identification se fait par le biais d'un

[95] Nous nous sommes basé dans la présentation des balises qui va suivre, sur *Le machinal*, op. cit.

code attribué par le responsable de l'opération et représentant le contenu de cette balise. Cette balise doit être placée en tête de l'édition électronique.

- **La balise « C » :** *Commentaires :* **<C=XXX>**

Semblable à un aide mémoire, cette balise est utilisée pour présenter, résumer et transmettre à ceux qui élaborent ou utilisent l'édition électronique des commentaires concernant la signification relative aux suffixes et aux contenus des différentes balises, les décisions qui ont été prises au moment de l'enregistrement ou des messages à introduire pour fournir toutes sortes d'indications et informations utiles aux usagers de l'édition. De ce fait, cette balise peut être insérée à n'importe quel endroit du texte puisqu'elle ne marque pas un segment particulier de celui-ci.

Il est à noter que ***Lexico3*** (ainsi que les versions précédentes) « enjambe » les commentaires introduits par cette balise. Celle-ci n'a donc aucun effet sur le traitement informatique qui suivra ; mais il est quand même préférable d'ôter ces commentaires de la copie du fichier qui va être soumis à *Lexico3*.

- **La balise « L » :** *modalités typographiques :* **<L=XX>**

Le rôle de cette balise est le marquage, dans l'édition électronique, des modalités et des changements typographiques de l'édition de référence. L'intérêt de cette balise réside principalement dans une éventuelle réédition du texte original.

- **Les balises à suffixe**

- **La balise « E » :** *Découpage du texte-source :* **<Epg=XX>, <Elg=XX>, …**

Un texte est généralement subdivisé en volumes, pages, paragraphes, lignes, etc. (si l'on considère sa forme éditoriale) ou en parties, chapitres, actes, scènes, etc. (considération faite de son organisation logique ou « naturelle »). La balise « **E** » est précisément faite pour rendre compte de ce découpage du texte en fragments.

A chaque fragment on fait correspondre un suffixe. Deux suffixes préétablis sont « **pg** » et « **lg** » correspondant respectivement à la numérotation des pages et celle des lignes. On peut évidemment ajouter autant de suffixes que l'on souhaite en les choisissant dans la suite 01, …, 99 et en se gardant surtout d'omettre d'en noter la signification dans un commentaire initial.

Quant au contenu de la balise « **E** », il est généralement constitué par un numéro choisi dans une suite : numéro de page, de ligne, de chapitre, de volume, …

Cas des paragraphes

Il est certes important de reproduire, dans l'édition électronique, le découpage du texte-source en paragraphes ; mais numéroter séquentiellement ceux-ci ne présente pas un très grand intérêt quant à la recherche d'occurrences dans le texte. En effet, essayer, par exemple, de « situer une occurrence dans le 158^{éme} paragraphe d'un texte n'est pas très commode pour le retrouver ! »[96].

Pour marquer les paragraphes on ajoute devant chacun de ceux-ci un caractère réservé à cet effet qui est le signe « **§** ».

Il est important de préciser ici que ce nous entendons par *paragraphe* ce n'est pas seulement un bloc de texte délimité par deux retours chariot mais aussi tout bloc de texte compris entre le début d'une page et un retour chariot ou entre un retour chariot et la fin d'une page.

Les vers de poésie ne sont pas considérés comme des paragraphes à part entière, mais comme faisant partie du paragraphe les précédant, sauf quand ils commencent une page.

Cas des lignes

---

[96] *Le Machinal...,* op. cit. p.22

On peut, dans un souci d'alléger l'édition électronique, ne pas introduire de balises pour marquer les lignes. On les remplacera dans ce cas par une barre oblique « / » à la fin de chaque ligne. *Lexico3* se charge alors d'incrémenter le compteur de ligne chaque fois qu'il rencontre une barre oblique et de le réinitialiser à 1 à chaque page c'est-à-dire chaque fois qu'il rencontre une nouvelle balise **<Epg>**.

- **La balise « S » :** *Variables de situation :* **<S01=XX> , <Sda=XX>, …**

Cette balise dépend en fait du contexte situationnel plus que du texte lui-même. Elle sert, en effet, à introduire, dans l'édition électronique, « des informations externes, des connaissances détenues par le chercheur, qui caractérisent tout ou partie du texte. Chaque type de clé définit une variable de situation. L'ensemble des variables constitue un système, qui permet de diviser le corpus en fragments homogènes ou d'extraire du corpus des sous-ensembles textuels correspondant à des valeurs déterminées de certaines variables »[97].

Là aussi, on peut introduire autant de balise de ce type que l'on souhaite sans aucune limitation quant à la nature des variables que peut représenter la balise « **S** ». Ce sont « les objectifs d'analyses et les traitements prévus sur le texte qui définissent le système des variables externes retenues pour figurer dans l'enregistrement »[98].

Aussi, doit-on faire correspondre à chaque variable un suffixe particulier. Cependant, deux suffixes alphabétiques préétablis correspondant à deux variables retenues sont obligatoires. Les deux variables retenues sont :

- *L'attribution du texte :* jugée par rapport à son émetteur. Cette variable est représentée par le suffixe « **at** » et prend les contenus suivants :
  - **<Sat=0> :** monologue de base, texte de l'auteur.
  - **<Sat=1> :** autocitation ou dialogue, texte de l'auteur.

[97] *Idem,* p.24

[98] *Ibid* p. 24

- **<Sat=2> :** texte étranger à l'auteur, cité par lui au style direct.
- **<Sat=3, 4, …, n> :** autres valeurs, laissées au libres choix du chercheur[99].

➢ *La date d'écriture du texte :* Cette variable est représentée par les suffixes et les contenus suivants :

- **<Sda=1950> :** suffixe « **da** » (date-année) ayant comme contenu l'année d'écriture réelle ou conjecturée du texte ou, à défaut, l'année de première publication.
- **<Sdm=01, …, 12> :** suffixe « **dm** » (date-mois) ayant comme contenu le mois d'écriture ou de publication du texte. Il est à noter que ce suffixe et le suivant servent surtout pour les périodiques, facilitant ainsi une analyse diachronique fine de certains corpus à savoir des articles de presse ou autres.
- **<Sdj=01, …, 31> :** suffixe « **dj** » (date-jour) ayant comme contenu le jour d'écriture ou de publication du texte.

En plus de ces deux variables (l'attribution et la date) la balise « **S** » permet « d'introduire des indications thématiques, des genres (poésie, prose, etc.), ou encore de séparer différents niveaux du texte (titre, sous-titre, etc.). Dans tous les cas, le suffixe de la clé sera choisi dans la série 01, 02, …, 99 (en évitant bien entendu les combinaisons déjà utilisées) ; son contenu sera un code alphanumérique. »[100]

Nos choix de balises, leurs types, leurs contenus ainsi que leur signification sont présentés dans l'encadré suivant représentant l'en-tête extrait du fichier de la version informatique de notre corpus.

---

[99] Il ne faut surtout pas omettre de noter la signification de ces valeurs dans un commentaire initial.

[100] *Idem* p. 25

Dans le deuxième encadré, nous présentons quelques extraits du corpus balisé selon les règles annoncées dans l'en-tête appartenant au même fichier.

L'opération de balisage est faite dans le traitement de texte Word® qui permet de parcourir le texte par page, par section, par tabulation etc., de rechercher-remplacer, d'utiliser les caractères génériques, les expressions régulières, etc. ; ce qui facilite énormément la tâche d'insertion des balises aux endroits voulus du corpus. Après enregistrement du corpus en ".doc", une copie sous format texte seul ".txt" est créée et c'est cette version qui est soumise à *Lexico3*.

Nous ne pouvons terminer cette section consacrée au balisage sans signaler un fait important ; outre le fait que *Lexico3* utilise la norme Saint-Cloud présentée plus haut, notre choix peut également s'expliqué par la compatibilité de cette norme avec la *TEI* (*Text Encoding Initiative*) et plus largement avec XML. Notre corpus étant balisé selon la norme Saint-Cloud, son "XMLisation" ou son transfert vers la norme *TEI*, sont désormais, non seulement possibles, mais aisés. Chose qui garantit sa portabilité et son interopérabilité.

<F=0101>
<C= La clé « **F** » contient le numéro d'ordre affecté au fichier. Les deux premiers chiffres correspondent au numéro de l'auteur (01 pour *Tawḥîdî*) et les deux derniers au texte (01 pour *Al-ÞImtâÝ wa-l-MuÞânasa*)**>**
<C= L'édition de référence à partir de laquelle cet enregistrement a été réalisé est celle du *Kitâb al-ÞImtâÝ wa-l-MuÞânasa* de *PAbû Ḥayyân at-Tawḥîdî*, édition critique de *PAḥmad PAmîn* et *PAḥmad az-Zayn*, 3 volumes, *Dâr Maktabat al-Ḥayât*, Beyrouth, Liban, 1966**>**
<C= Outre les types de clé « **F** » « **C** », « **Epg** » et « **Sat** (contenu = **0**, **1** ou **2**) » utilisés avec leur contenu ordinaire selon la norme Saint-Cloud, figurent dans ce corpus : la clé « **volume** » ayant comme contenu le numéro du volume allant, pour ce corpus, de **1** à **3** ; la clé « **nuit** » ayant comme contenu le numéro de la nuit allant de **00** pour le prologue puis **01** pour la première nuit jusqu'à **40** pour la dernière nuit (Il est à noter que la nuit 10 correspond en fait aux nuits 10 et 11 et que la nuit 12 n'existe pas) ; la clé « **Sat** » qui, en plus des contenus prédéfinis, marque, d'une part, les citations religieuses avec les contenus suivants **Cor** pour les versets coraniques et **Had** pour les Hadiths, d'autre part, les interlocuteurs dans un dialogue rapporté par *Tawḥîdî* au style direct avec un contenu qui sera précisé au début de chaque nuit comme, par exemple, dans la huitième nuit, le contenu sera : **Mat** pour *Mattâ b Yûnus* et **Sir** pour *PAbû* Sa*ʿîd a*S-Sīrâfî ; la clé « **S01** » qui contient les citations poétiques ayant comme contenu **pro** pour la prose (valeur négative) ou **poé** pour poésie (valeur positive) ; la clé « **S02** » qui marque la présence de phraséologie et dont le contenu est **1** pour marquer le début de la phraséologie, et **0** la fin de celle-ci **;** la clé « **S03** » pour marquer les différents titres et sous-titres, elle a comme contenu : **1** pour les titres de premier niveau, **2** pour les titres de niveau deux, **3** pour les titres de niveau trois et **4** pour le corps de texte**>**

Encadré 1
En-tête de notre fichier résumant la signification des balises utilisées dans la version informatique du corpus

<S03=1>

ال إمتاع و ال مؤانسة

<volume=1>

<S03=2>

ال جزء ال أَوَّلٌ

<Nuit=00>

<Epg=001>

<S03=4>

<S01=pro><Sat=0>

§ ب اسم $الله$ $الرحمان$ $الرحيم$

§ قَالَ $أبو+حيّانَ+التوحيديّ$ نَجَا مِن آفة ال دنيا مَن كَانَ مِن ال عارِف و وَصَلَ إلى خيْرة ال آخرة مَن كَانَ مِن ال زاهِد و ظَفِرَ ب ال فوْز و ال نعيم مَن قَطَعَ طمع هـ مِن ال خَلْق أجمع و ال حمْد ل $الله$ ربّ ال عالَم و صَلَّى $الله$ على $نبيّ+$ هـ و على $آل+$ هـ ال طاهِر

.....

<Epg=002>

§ رشْد ي و أَلْقَى ب يد ي إلى ال تَهْلُكة و تَجَانَفَ إلى ما سَاءَ ي أَوَّلٌ و لا سَرَّ ي آخِر هذا و أنا في ذيْل ال كهولة و بادِئ ال

.....

§ على هذا ال حدّ كَانَ مَقطع كلام كَ في مَوخِدة كَ و إلى هَهُنا بَلَغَ فيْض عتْب كَ و لائمة كَ و في دون ذلك تنبيه ل ال نائم و إيقاظ ل ال ساهٍ و تقويم ل مَن قَبِلَ ال تقويم و قَدْ قَالَ ال أَوَّلٌ

<S01=poé>ألا إنّما كَفَى ال فتى عِنْدَ زيْغ هـ

مِن ال أَوَد ال بادٍ ثِقاف ال مُقوّم**<S01=pro>**

§ ف قَالَ ل كَ أنا سامِع مُطيع و خادِم شَكُور لا اِشْتَرَى سُخْط كَ ب كُلّ

<Epg=008>

.....

Encadré 2
Extraits du corpus, segmenté, lemmatisé et balisé, montrant une partie des balises utilisées

Chapitre 2

Dépouillement segmenté et dépouillement en mots graphiques : Données statistiques

Dans le premier chapitre, nous avons présenté un abrégé chronologique de la vie d'*Abû Ḥayyân at-Tawḥîdî*, le corpus *al-ÞImtâʕ wa-l-Muþânasa* et sa constitution, et le dépouillement lexical et ses étapes qui nous ont permis, entre autres, de segmenter les mots graphiques du corpus pour arriver aux unités lexicales que nous avons définies comme nos unités de décompte. Après ce premier chapitre et avant d'aborder, dans le troisième chapitre, les étapes du traitement et de l'analyse statistiques, nous avons voulu présenter, dans ce présent chapitre, des calculs statistiques que nous avons faits sur les unités résultant de la phase de restructuration du corpus qui est la segmentation. L'intérêt de ces calculs est, d'un côté de révéler l'impact de la segmentation aussi bien sur l'étendue du corpus que sur son vocabulaire, et de l'autre côté de présenter la distribution des unités lexicales résultantes eu égard aux différents types de mots graphiques. La productivité segmentale des mots graphiques est alors, de ce point de vue, un bon facteur de cette restructuration du corpus dont il est nécessaire de voir les caractéristiques.

Une remarque doit cependant être faite : quand on parle de segmentation au sens de délimitation des unités de décompte, il ne s'agit pas seulement de décompositions de mots graphiques en unités de niveau inférieur même si cela représente la quasi totalité des cas, mais il s'agit également de regroupements de certains mots graphiques en unités polylexicales : mots composés, locutions, etc.

La première constatation manifeste concernant le résultat de la segmentation est que l'étendue générale du corpus, c'est-à-dire le nombre total de ces occurrences, est passée de 37 457 à 61 177 : elle a donc presque doublé ; elle a exactement été multipliée par 1,7. Se basant sur ces résultats, et sur d'autres tests que nous avons opérés sur de petits textes pour confirmer ou infirmer cette constatation, nous pouvons d'ores et déjà affirmer que le volume d'un texte arabe (son étendue) se voit pratiquement multiplié par près de 170% après l'opération de segmentation.

Texte de référence (sans Noms Propres)		**Texte segmenté** (sans Noms Propres)	
Nombre des **mots graphiques** différents:	**15 214**	Nombre des **items** :	**10 174**
Nombre des occurrences :	**37 457**	Nombre des occurrences :	**61 177**

Inversement à cet accroissement considérable de 70 % des occurrences, le nombre des mots graphiques différents (items graphiques) a, quant à lui, subi une diminution notable de 33 % passant de 15 215 à 10 174 vocables. Cette diminution s'explique par le fait que la segmentation de mots graphiques différents engendre une redistribution des unités avec l'apparition d'unités qui ne sont pas forcément différentes. Ce phénomène est fortement observable quand on a des unités de forte fréquence, à savoir les enclitiques ou les proclitiques très fréquents en arabe, entrant en agglutination avec d'autres unités composant ainsi des mots graphiques à chaque fois différents. Pour illustrer ce phénomène, voici un exemple contenu dans ce tableau :

Au départ : **9** mots graphiques différents	Segmentation	**A l'arrivée :** **4** items différents
كلامه	كلام\ه	**كلام** **ه** **ل** **و**
وكلام	و\كلام	
لكلام	ل\كلام	
لكلامه	ل\كلام\ه	
وكلامه	و\كلام\ه	
ولكلام	و\ل\كلام	
ولكلامه	و\ل\كلام\ه	
له	ل\ه	
وله	و\ل\ه	

Quant à la portée de la segmentation concernant la distribution des éléments entrant en agglutination pour former les mots graphiques, les différents types de ces mots graphiques et leur productivité segmentale, nous résumons quelques données statistiques dans le tableau qui suit :

Nombre d'éléments par mot graphique	Formes	% des formes	Occurrences	% des occurrences	Total des éléments	% des éléments
1 élément	5 511	36,22%	17 524	46,78%	17 524	28,56%
2 éléments	6 916	45,46%	16 041	42,83%	32 082	52,29%
3 éléments	2 723	17,90%	3 825	10,21%	11 475	18,70%
4 éléments	64	0,42%	67	0,18%	268	0,44%
Totaux	**15 214**	**100,00%**	**37 457**	**100,00%**	**61 349**	**100,00%**

La différence entre le total des éléments composant les mots graphiques (61 349) et l'étendue du corpus, c'est-à-dire le nombre d'occurrences des items, (61 177) présentée dans le chapitre 9, *Les caractéristiques lexicométriques*, s'explique par le fait qu'une partie des unités obtenues après segmentation sont des noms propres qui ont d'ailleurs été écartés des comptages (plus exactement, ils ont été comptés à part) mais leurs places en tant qu'éléments de mots graphiques ont bien été comptabilisées. Par exemple, le mot graphique bipartite والله *wallâhi* [par Dieu - et Dieu] a été segmenté en deux parties و \ الله *wa* / *allâh*, et comme le nom propre الله a été supprimé donc non comptabilisé, sa place en tant qu'élément a bien été (et doit être) gardée (و \ NP) et de ce fait, comptabilisée. Ceci est justifié par le fait que ce qui importe ici, c'est de savoir en combien d'éléments est décomposable un mot graphique multipartite.

Sur les 37 457 occurrences de mots graphiques que comporte le corpus, 17 524 occurrences soit 46,78 % sont irréductibles c'est-à-dire des mots graphiques correspondant à des mots minimaux donc composés d'une seule et unique unité lexicale qui n'a par conséquent pas besoin d'être segmentée.

Toujours sur les 37 457 occurrences de mots graphiques, 19 933 occurrences soit 53,22 % sont multipartites c'est-à-dire des mots graphiques segmentables en deux, trois ou quatre parties[101].

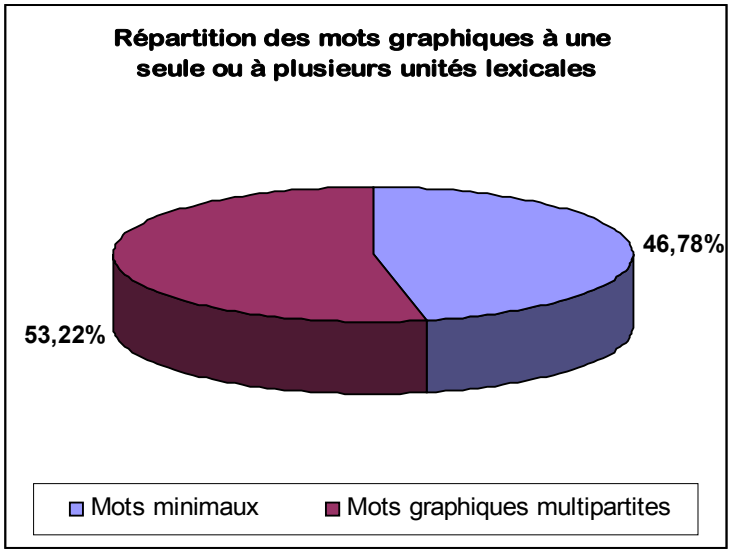


Figure 21

42,83 % de l'ensemble des occurrences des mots graphiques sont bipartites, 10,21 % sont tripartites, alors que seulement 0,18 % de l'ensemble des mots graphiques, en occurrences, sont des mots graphiques quadripartites.

[101] Bien que notre corpus ne comporte pas de mots graphiques pentapartites, ce type de mots graphiques est bien attesté dans la langue arabe. Pour illustrer ce cas de figure, nous pouvons présenter l'exemple suivant : أفبقلمك « *'afabiqalamika* » (est-ce alors avec ton crayon ?) qui est segmentable en cinq parties أ\ف\ب\قلم\ك « *'a/fa/bi/qalami/ka* ».

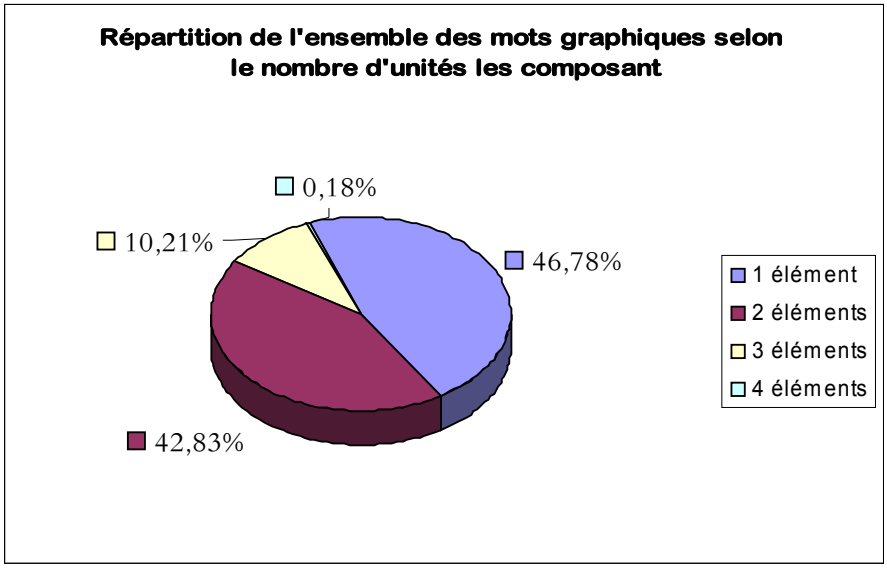


Figure 22

Au sein des mots graphiques multipartites, les mots bipartites s'octroient la part du lion avec plus de quatre cinquièmes des occurrences soit 80,47 %. La deuxième place est occupée par les mots tripartites avec moins du cinquième des occurrences soit 19,19 %. Viennent enfin, en dernière place, les mots quadripartites avec un peu plus de trois millièmes des occurrences soit 0,34 %.

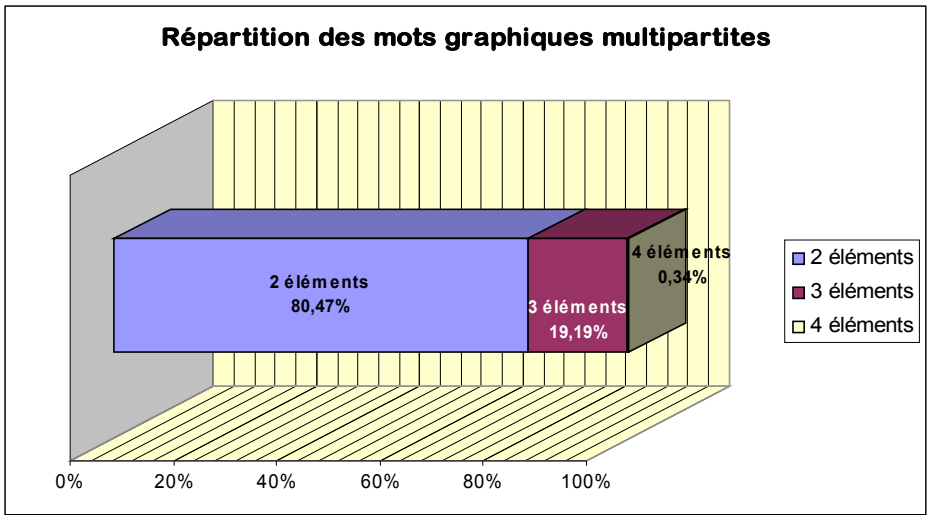


Figure 23

En ce qui concerne la provenance des unités lexicales, plus de la moitié des unités lexicales résultant de l'opération de segmentation soit 52,29% proviennent des mots graphiques bipartites, 28,56 % sont des mots minimaux, 18,70 % résultent des mots graphiques tripartites et seulement 0,44 % naissent des mots graphiques quadripartites. La figure 4 suivante montre bien la provenance des unités lexicales selon chaque type de mots graphiques.

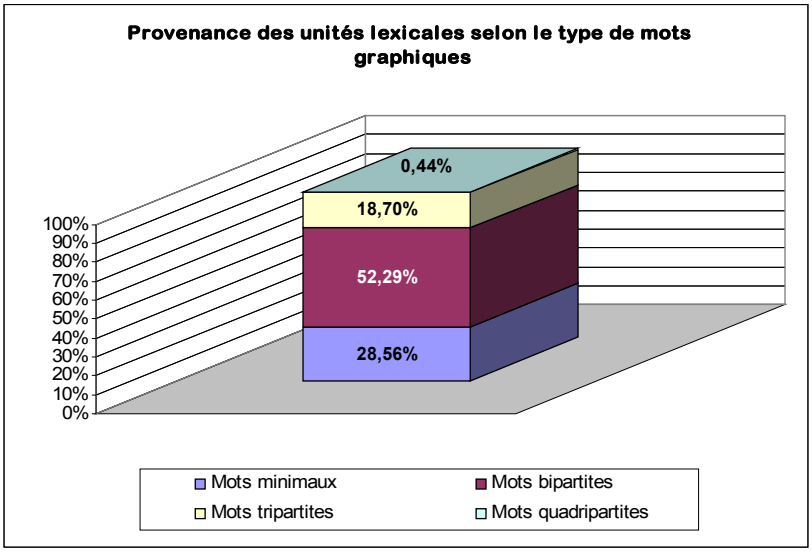


Figure 24

Parce que, de loin, les plus fréquents de tous les mots multipartites, les mots graphiques bipartites sont à l'origine de la majorité des unités lexicales. La seconde place est donc occupée par les unités lexicales provenant des mots graphiques "simples" ou "unitaires", les mots minimaux. Aussi surprenant que cela puisse paraître, ce phénomène s'explique tout simplement par la fréquence assez importante qu'ont ces mots irréductibles. Cette fréquence dépasse à peine celle des mots bipartites (46,78 % contre 42,83 %). La troisième place concernant la provenance des unités lexicales revient aux mots tripartites qui, malgré leur assez faible fréquence, sont à l'origine de près de 20% des unités lexicales. En fin, à cause de leur fréquence extrêmement basse, les mots graphiques quadripartites ne fournissent que très peu d'unités lexicales.

Mots graphiques	Coefficient de productivité
Mots minimaux	3,18
Mots bipartites	4,64
Mots tripartites	4,21
Mots quadripartites	4,19
Coeff. de productivité générale	**4,03**

Tableau 1
Coefficient de productivité de chaque type de mots graphiques et le coefficient de productivité générale

Au niveau de la productivité, en unités lexicales, des mots graphiques, les 15 214 mots graphiques ont produit 61 349 unités lexicales, ce qui représente un coefficient de productivité générale de 4,03. Si nous regardons la productivité de chaque type de mots graphiques, les mots bipartites sont les plus productifs avec un coefficient de productivité de 4,64 (soit 0,61 au-dessus de la moyenne). Le deuxième rang est occupé par les mots tripartites avec un coefficient de 4,21 (soit 0,18 au-dessus de la moyenne). Les mots quadripartites arrivent en troisième position de la productivité avec un coefficient de 4,19 (soit 0,16 au-dessus de la moyenne). Alors que les coefficients de productivité des mots graphiques multipartites sont au-delà du coefficient de

productivité générale, celui des mots minimaux accuse un écart négatif de (- 0,85) par rapport à cette moyenne ; il a une valeur de 3,18. Cette variation au-delà et en-deçà du coefficient de productivité générale est représentée dans la figure 5 suivante :

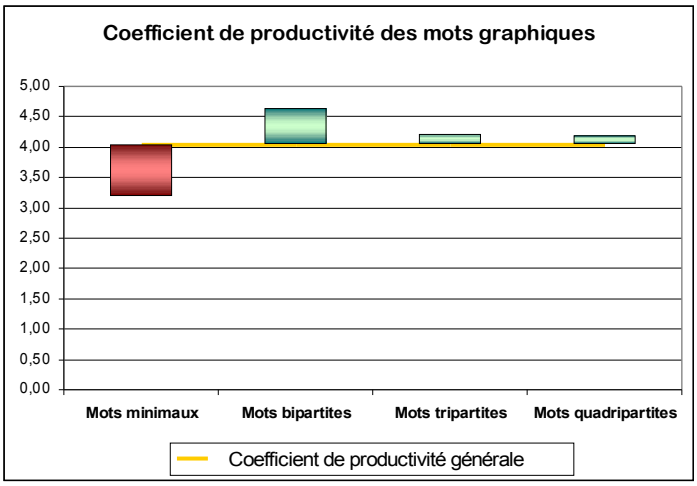


Figure 25

Nous présentons ci-après quelques exemples des mots graphiques les plus fréquents dans les différents types, bipartites, tripartites et quadripartites :

Les 3 **mots graphiques quadripartites** ***les plus fréquents***

Mot graphique quadripartite	fréquence
ف\ب\ال\عرَض	2
و\ل\ال\حقّ	2
و\ل\بعْض\ها	2

Les 20 **mots graphiques tripartites** ***les plus fréquents***

Mot graphique tripartite	fréquence
ل\أنّ\ه	47
و\ك\ذلك	34
ف\إنّ\ه	29
و\ل\هذا	29
و\ل\ذلك	20
ب\ال\عقْل	17
ل\أنّ\ها	16
و\ل\ه	16
و\مِن\ها	14
ف\إنّ\ها	13

Mot graphique tripartite	fréquence
و\ال\آخَر	12
و\لكنّ\ه	12
ف\إنّ\ي	11
و\لNP\	10
ب\ال\حقّ	9
و\ال\عقْل	9
و\في\ه	8
ل\أنّ\هم	7
ب\ال\طبْع	7
ب\ال\قوّة	7

Les 50 mots graphiques bipartites *les plus fréquents*

Mot graphique bipartite	fréquence
و\لا	331
ل\ه	200
ب\ه	156
على\ه	135
و\NP	130
في\ه	112
مِن\ه	93
و\ما	90
و\أمّا	89
و\إنْ	89
و\هو	89
ال\نفْس	86
ل\أنّ	85
ف\قَالَ	85
و\هذا	83
إلى\ه	79
ف\إذا	71
ال\إنسان	68
أنّ\ه	66
و\قد	65
عن\ه	63
و\قَالَ	63
ب\ها	61
ل\ها	60
في\ها	60

Mot graphique bipartite	fréquence
مِن\ها	60
و\لم	57
ل\NP	54
ال\عقْل	54
و\إذا	53
و\على	52
ف\هو	51
و\كان	51
ال\ناس	50
ل\كَ	49
و\مَن	49
ل\هم	46
ب\ما	46
كُلّ\ه	46
و\في	46
ل\ي	43
ف\إنْ	43
ف\إنّ	43
ف\قد	42
و\لكنْ	41
نفْس\ه	39
ال\لفْظ	39
ب\NP	39
ال\حيوان	38
ال\كلام	38

Chapitre 3

Traitement et analyse du corpus

1. Quantification du corpus

Quantifier un corpus, c'est en extraire les occurrences des unités lexicales définies comme unités de décompte en dressant la liste structurée et exhaustive de ces unités, associées à leurs fréquences, pour pouvoir les analyser par la suite et leur opérer les calculs statistiques voulus. Il s'agit donc du passage d'un texte à des listes de mots-formes, lemmes, catégories lexicales, segments répétés, contextes (concordances), etc. À l'issue de cette opération, la linéarité du texte cède sa place à la verticalité des listes et des tableaux. Cette quantification constitue la première étape nécessaire à cette phase du traitement et d'analyse du corpus en lexicométrie. C'est une opération qui trouve son intérêt dans l'approche qui caractérise la démarche lexicométrique, l'approche quantitative qui « permet seule d'accéder à la description de phénomènes textuels qui présentent un grand intérêt une fois mis en évidence et dont il aurait été difficile de cerner les contours *a priori* »[102].

Tous les traitements statistiques qui viendront par la suite pour étudier quantitativement les faits langagiers contenus dans le corpus, déceler des corrélations entre phénomènes ou mettre en évidence d'éventuelles variations/régularités des unités étudiées, reposent sur les comptages réalisés à cette étape. Il est à noter que les comptages effectués sont étroitement liés à la *norme de dépouillement* établie et aux étapes de son application que nous avons exposées dans le premier chapitre.

[102] Salem A., et al., *Les linguistiques de corpus*, 1997, p. 184

2. Un outil de traitement lexicométrique : *Lexico3*[*]

Les logiciels de traitement et d'analyse statistiques sont souvent comparables à des boites noires hermétiques : on entre les données à traiter, on sélectionne la méthode à utiliser et on obtient les résultats sans aucun regard sur les calculs intermédiaires. Ils ne permettent à l'analyste, ni de suivre les calculs effectués dans les différentes étapes de traitement, ni d'exploiter les fichiers intermédiaires ; c'est également le cas des logiciels de lexicométrie hormis *Lexico3* qui est l'un des rares logiciels qui offre la possibilité de voir et d'exploiter certains calculs et fichiers intermédiaires. Il livre un certain nombre de fichiers que l'on peut utiliser en dehors du logiciel et indépendamment de lui et de ses résultats et graphiques finaux. *Lexico3* laisse en effet à l'utilisateur le soin de dépasser les résultats fournis en approfondissant "à la main" certains niveaux d'analyse. Ceci est fortement important car ce sont les traitements de données qui guident l'interprétation à la fin du processus. Pour choisir un logiciel, il est quand même indispensable de se demander, entre autres, si celui-ci offre ou non la possibilité de contrôler et de modifier le traitement à réserver aux données que l'on se donne à analyser. Et la singularité de *Lexico3* est justement qu'elle « permet à l'utilisateur de garder la maîtrise sur l'ensemble des processus lexicométriques depuis la segmentation initiale jusqu'à l'édition des résultats finaux »[103].

L'adoption de *Lexico3* est aussi dictée par un choix pratique : le support de l'arabe. En effet, le logiciel est ouvert sur le système d'exploitation, autrement dit, quelle que soit la langue supportée par le système d'exploitation et quel que soit le

---

[*] *Lexico3* est un logiciel développé par André Salem, Serge Fleury, Cédric Lamalle et William Martinez. Le site officiel de *Lexico3* est : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/lexico3.htm

[103] C. Lamalle, W. Martinez, S. Fleury et A. Salem, ***Outils de statistique textuelle.*** B. Fracchiolla, A. Kuncova et A. Maisondieu, ***Manuel d'utilisation*****,** SYLED - CLA2T**,** Université Paris 3, Version 3.41, 2003 : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuel3.htm

codage utilisé, *Lexico3* est capable de traiter et surtout d'afficher correctement le texte et les résultats de l'analyse dans la langue et le codage de départ. Une seule petite lacune est cependant détectée au niveau de l'affichage du module de concordancier. Ce module n'affiche pas correctement les caractères arabes de la concordance du mot-pôle choisi. Mais cette petite imperfection n'est pas insurmontable comme nous le montrons plus loin.

En revanche, notons que *Lexico3* est assez limité au niveau de la représentation graphique ; c'est pour cette raison que nous avons voulu produire nous-même les différents graphiques en utilisant d'autres outils à savoir *Excel* de Microsoft et un outil complémentaire à celui-ci, *XLSTAT* comme nous l'expliquons *infra*.

Commençons tout d'abord par présenter brièvement les principales fonctionnalités de *Lexico3*, celles du moins qui sont en relation directe avec notre travail et que nous avons utilisées pour la récupération et l'exploitation des données quantitatives sur lesquelles nous avons construit nos traitements et analyses.

L'ensemble des fonctionnalités (modules) de *Lexico3* sont : Segmentation, Concordance, Segments Répétés, Groupe de Formes (TYPES généralisés), Ventilation, Partition, PCLC, AFC, Spécificités, Carte des Sections.

Pour créer une nouvelle base de données numérisée, *Lexico3* fait appel au premier module de son programme, le module SEGMENTATION[104] dont la tâche est de créer une base de données numérisée à partir du fichier texte que l'utilisateur lui a soumis. Le contenu de cette base est un dictionnaire des formes rencontrées dans le texte avec un numéro d'ordre pour chacune d'entre elles, une version codée du texte et le nombre des occurrences de chaque forme graphique du corpus.

---

[104] Étant donné que *Lexico3* opère au niveau des formes graphiques, comprises entre les délimiteurs, le terme *segmentation* ne doit pas être compris ici dans le sens de la *segmentation lexicale* telle que nous l'avons définie dans le premier chapitre et dans la norme de segmentation, mais dans le sens de l'opération qui consiste à délimiter des unités minimales dans un texte. Il s'agit tout simplement de quantification, c'est-à-dire du passage du texte aux listes. c'est un module qui effectue des décomptes sur les unités lexicales.

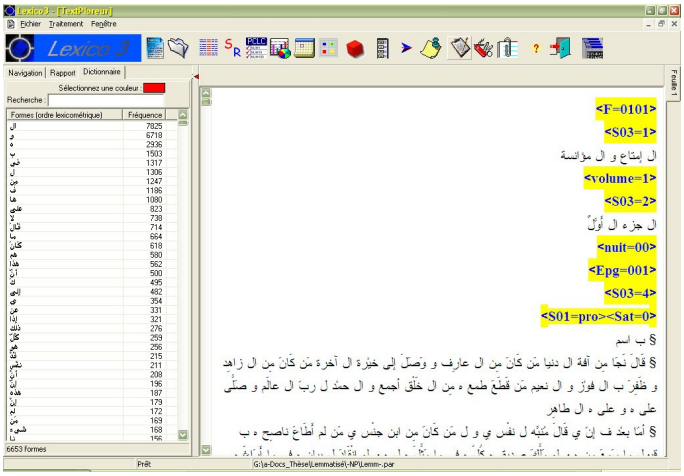


Figure 26
Ecran principal de Lexico3 après la création d'une nouvelle base

À l'issue de la création de la base, *Lexico3* crée quatre fichiers dans le même dossier que le texte-source. Trois de ces fichiers ont le même nom que la texte-source mais ayant les extensions suivantes "*.par*", "*.dic*" et "*.num*", et le quatrième nommé "*atrace.txt*".

Le fichier de sortie ayant l'extension "*.par*" renferme les principaux décomptes ; on y trouve le nombre des occurrences, celui des formes, la fréquence maximale, le nombre des hapax, le nombre des types des clés et celui des contenus des clés. Y figure aussi le rappel des caractères délimiteurs choisis lors de la segmentation.

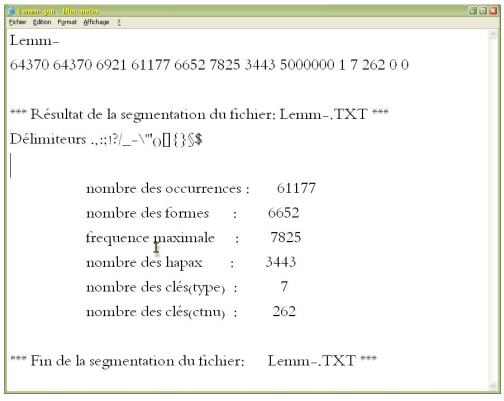


Figure **27**
Capture d'écran du fichier de sortie « *Lemm-.par* »

Le fichier de sortie "*atrace.txt*" contient un rapport sur la mémoire allouée, les paramètres pris en compte, les fichiers lus et écrits, etc. Et en cas d'erreurs dans l'encodage, par exemple, la base ne peut être créée et c'est ce fichier qui fournit les indications permettant de corriger les erreurs.

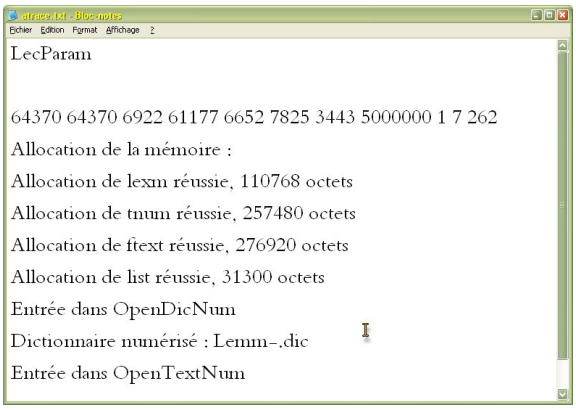
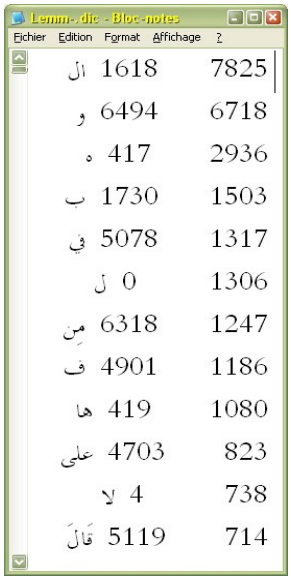
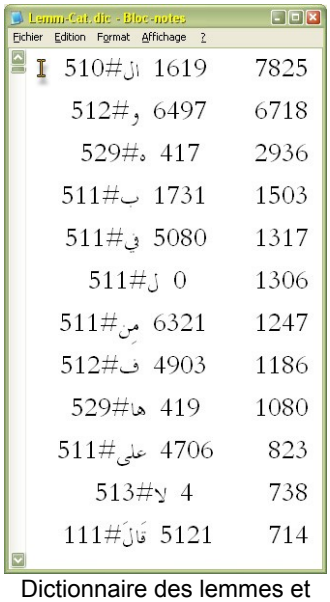


Figure 28
Capture d'écran du fichier de sortie « atrace.txt »

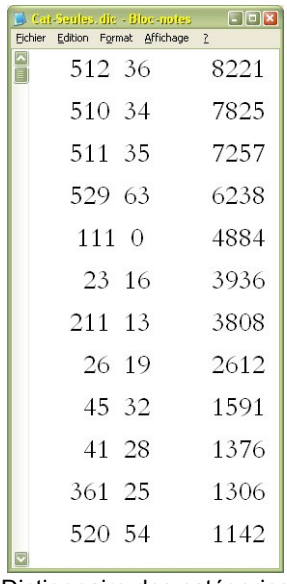
Le fichier de sortie ayant l'extension "*.dic*" contient le dictionnaire des formes classées par fréquence décroissante ainsi que le rang lexicométrique de chaque forme dans le corpus.



Dictionnaire des lemmes



Dictionnaire des lemmes et catégories lexicales



Dictionnaire des catégories lexicales seules

Figure 29
Extraits de 3 fichiers de sortie ".dic", "lemme-.dic", "lemme-Cat.dic","Cat-Seules.dic".
Dans chacun des 3 fichiers, la colonne de droite contient la fréquence, celle de milieu le rang lexicométrique et celle de gauche la forme[105] (lemme pour le premier fichier,
lemme#catégorie_lexicale pour le deuxième et catégorie lexicale seule pour le troisième)

Le fichier de sortie ayant l'extension "*.num*" contient quant à lui le texte sous une forme codée à usage interne au logiciel et ne pouvant être lu par un éditeur de texte.

La nouvelle base créée, c'est-à-dire la segmentation (au sens de *Lexico3*) faite, *Lexico3* offre à l'utilisateur trois groupes d'outils correspondant à trois tâches différentes mais pas nécessairement exclusives : explorer le texte, l'analyser statistiquement et exploiter les développements lexicométriques en se déplaçant parmi les résultats produits et entre eux et le texte de départ. Ces trois groupes d'outils sont les outils d'exploration textuelle, les outils d'analyse statistique et les outils de navigation lexicométrique.

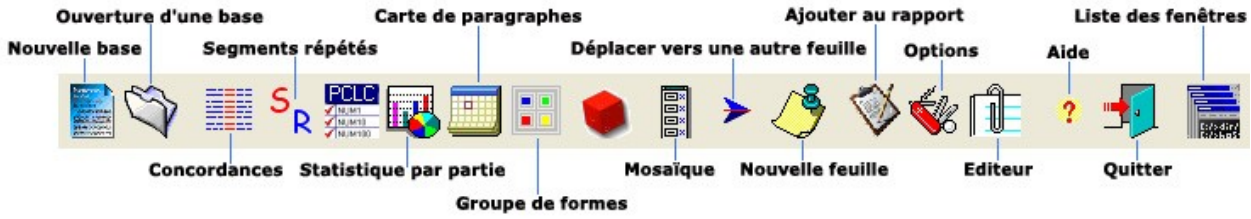


Figure 30
La barre d'outils de *Lexico3* dans laquelle figure la majorité des modules

[105] Dans le jargon de *Lexico3*, *forme* veut dire unité de décompte (mot graphique, mot-forme, lemme, lemme+catégorie lexicale, catégorie lexicale seule, etc., selon le contenu du fichier soumis à l'analyse). Il ne faut donc pas comprendre ici *forme* dans le sens de *forme fléchie* par opposition à *lemme*.

2.1. Les outils d'exploration textuelle

Outre le dénombrement des formes graphiques, *Lexico3* permet de faire l'inventaire de différents types d'unités textuelles telles que les *segments répétés* (suites de formes graphiques identiques attestées plusieurs fois dans le texte), les *cooccurrences* (couples de formes présentes dans les mêmes contextes) ou encore les *types généralisés* (groupes de formes ayant en commun une séquence de caractères).

- C'est le module CONCORDANCE qui se charge d'afficher toutes les occurrences d'une forme ou d'un type généralisé (*Tgen*) en contexte. Par la visualisation directe et contextualisée de la forme-pôle et ses cooccurrences, la concordance permet un retour systématique à l'environnement immédiat de la forme dans le corpus.

Nous mentionnions plus haut une petite lacune au niveau de l'affichage, pour les textes arabes, des concordances et qui peut facilement être contournée de la manière suivante. Étant donné que le rapport dans lequel on sauvegarde les différentes opérations d'exploration du texte dans *Lexico3*, est enregistré sous format ".html", il suffit de changer l'encodage de la page HTML ainsi générée, pour obtenir la concordance en caractères arabes comme le montre la capture d'écran suivante (figure 5) représentant la page HTML de la concordance du lemme مَسـألة *masÞala* [question, affaire, sujet] récupérée à partir du rapport de *Lexico3* et que le navigateur affiche correctement. Pour l'affichage des caractères arabes dans un navigateur il y a deux possibilités : soit le navigateur est bien configuré pour la détection automatique des caractères et dans ce cas, c'est lui qui se charge d'afficher directement la page en prenant en compte son encodage, soit il faut préciser au navigateur (l'obliger) à appliquer l'encodage de l'arabe par le biais de la commande : Affichage > Encodage des caractères (pour *Firefox* ou Codage pour *I.E.*) > Arabe (Windows-1256) ou Arabe (ISO-8859-6) ou encore Unicode (UTF-8). Dans la page présentée dans la capture d'écran suivante, nous avons également agi sur le code source de la page pour que le

sens du texte (il s'agit bien de sens d'écriture et non d'alignement de paragraphe) de la concordance soit de droite à gauche et ce en ajoutant à la balise paragraphe l'attribut « dir="rtl" » (*right to left*, en anglais). Nous avons également changé, dans le code source, la couleur de la forme-pôle (les formes-pôles sont coloriées et centrées dans le concordancier à l'intérieur de *Lexico3*, mais dès qu'on ajoute la concordance au rapport, on perd cette mise en forme).

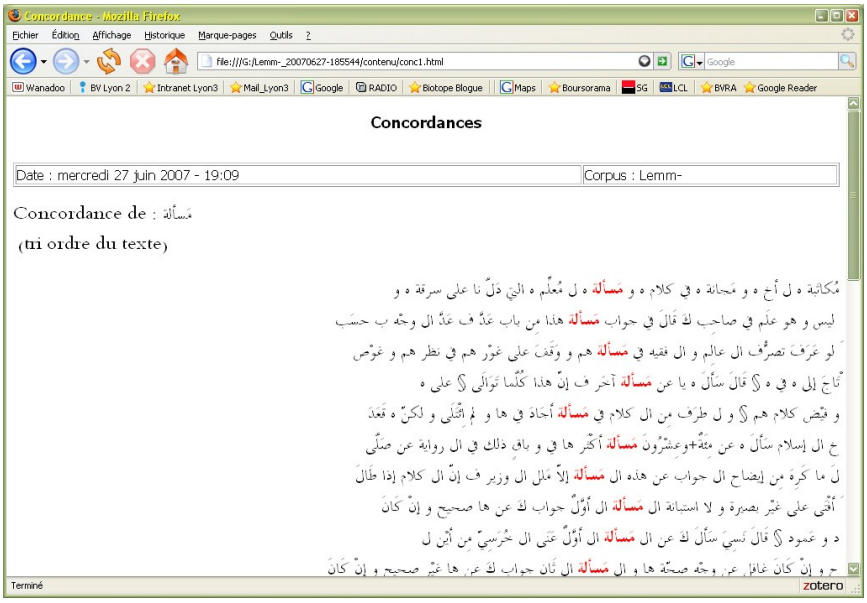
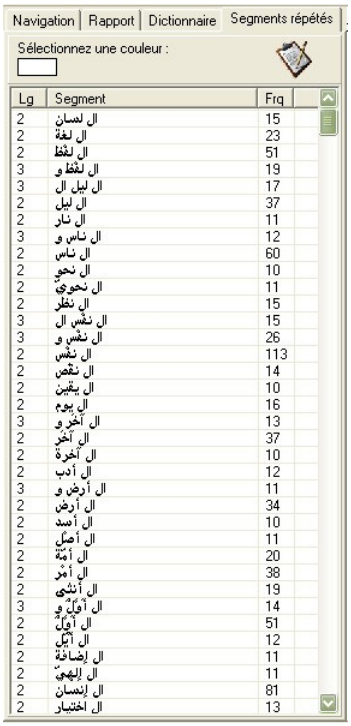


Figure 31
Concordance du lemme مَسأَلة récupérée à partir de *Lexico3*

- Le module SEGMENTS RÉPÉTÉS répertorie et calcule les segments (suites de formes) répétés du texte, dont la fréquence est supérieure à un seuil minimal fixé par l'utilisateur. Il calcule également la répartition des segments répétés dans les parties du corpus. Un dictionnaire de segments répétés est donc créé dans la fenêtre de gauche à côté du dictionnaire des formes. Nous présentons une capture d'écran du dictionnaire des segments répétés dans la figure 7 suivante (capture gauche). Dans ce dictionnaire, la première colonne (lg) indique la longueur (i.e. le nombre de formes) du segment répété qui se trouve dans la deuxième colonne (Segment), alors que la troisième colonne (Frq) contient la fréquence de chaque segment répété.



Dictionnaire des segments répétés

Types généralisés : Résultat de la recherche à partir du motif في *fi*

Figure 32

- Le module GROUPE DE FORMES, appelé aussi module TYPES GÉNÉRALISÉS, permet de chercher et de constituer des *types* rassemblant les occurrences de formes graphiques différentes mais ayant une chaîne de caractères commune. Les types ainsi regroupées peuvent constituer ensuite des entités uniques, les *Tgen*, qui peuvent être manipulées comme telles. La capture d'écran de droite dans la figure 7 ci-dessus, représente un exemple de *Tgen* ayant en commun le motif (chaîne de caractère) في *fi*. Nous remarquons dans cette capture que le motif tapé dans le champ correspondant de la fenêtre ne s'affiche pas en caractères arabes, mais la recherche dans le corpus et l'affichage du résultat se font correctement.

2.2. Les outils d'analyse statistique

- Le module PARTITION permet de créer une partition du corpus d'après les différentes balises[106] (clefs) introduites avant de soumettre le corpus à *Lexico3*. Il effectue également des calculs statistiques portant à la fois sur les formes et les segments répétés de l'ensemble du corpus. Les comptages faits dans les différentes parties du corpus permettent ensuite de construire le Tableau Lexical Entier (TLE)[107] qui servira de base aux différentes analyses statistiques et dont nous présentons un extrait dans le tableau suivant :

Lemme	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N13	N14	N15	N16	Fréq. tot.
ال	566	286	388	190	498	86	861	372	1317	777	1229	343	477	248	187	7825
و	691	256	373	218	483	107	945	319	1193	557	697	218	374	167	120	6718
ه	245	106	191	115	287	78	284	88	468	156	574	85	99	94	66	2936
ب	122	63	90	50	101	16	199	67	266	120	154	71	94	60	30	1503
في	107	50	76	32	116	20	136	51	259	92	205	49	70	35	19	1317
ل	97	64	47	41	93	23	163	38	211	96	193	61	83	61	35	1306
مِن	102	48	53	30	84	16	144	42	239	87	222	51	67	28	34	1247
ف	50	56	50	46	75	15	115	42	212	107	245	50	67	22	34	1186
ها	45	23	29	21	37	7	110	52	150	48	398	38	88	30	4	1080
على	74	25	42	27	84	23	98	31	165	57	100	18	38	23	18	823
لا	466	333	429	734	444	315	887	124	124	736	95	48	47	47	9	738

[106] Voir le premier chapitre *Constitution et dépouillement du corpus*, § 3.7.Balisage, p. 96-105

[107] Pour le TLE, voir (Salem, 1993 : 38-39) et (Lebart et Salem, 1994 : 26-27)

قَالَ	25	58	31	33	70	21	89	27	172	33	31	34	35	35	20	714

Tableau 2
Extrait du Tableau Lexical Entier (TLE) récupéré à partir de *Lexico3*

- Le module PCLC (Principales Caractéristiques Lexicométriques du Corpus) calcule et affiche les principales caractéristiques par partie selon la partition choisie, à savoir le nombre des occurrences, le nombre des formes, celui des *hapax*, la fréquence maximale et la forme la plus fréquente de chaque partie.

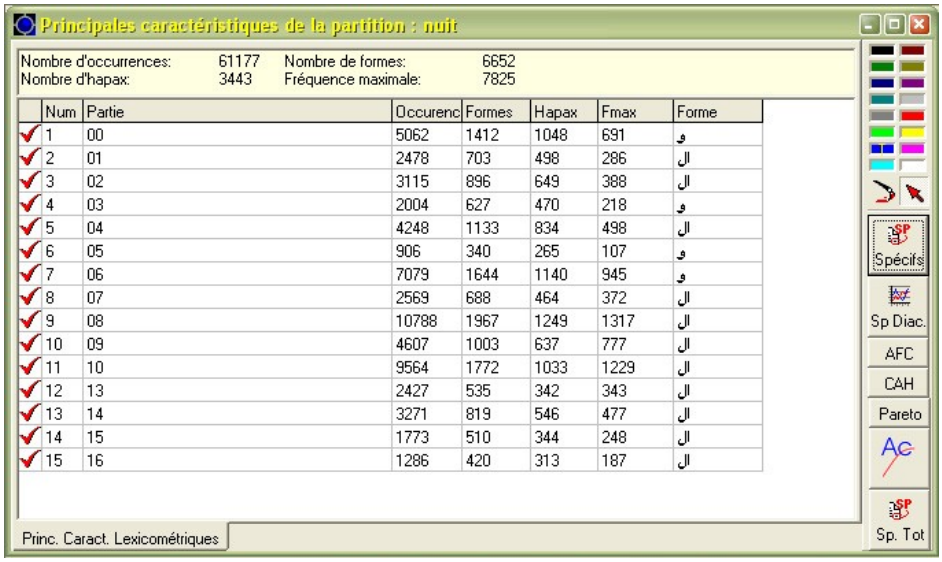
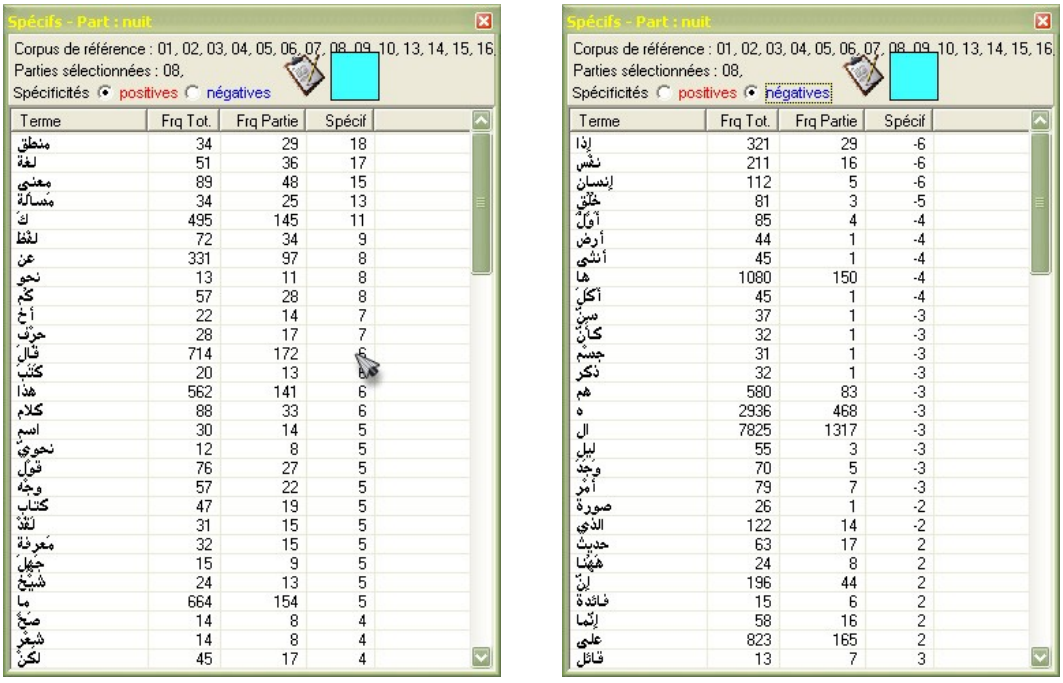


Figure 33
Capture d'écran représentant les PCLC de notre corpus selon la partition Nuit

- Le module SPÉCIFICITÉS effectue l'Analyse nécessaire pour le calcul des spécificités du corpus aussi bien sur les formes graphiques que sur les segments répétés (quand le calcul de ces derniers a été préalablement effectué) avec un paramétrage fixé par l'utilisateur, en choisissant une fréquence minimale et un seuil en probabilité. Le paramétrage que nous avons choisi pour le calcul des spécificités est le suivant : le seuil de probabilité = 5 % et la fréquence minimale = 10. Le but de cette analyse des spécificités est de pouvoir porter un jugement sur la fréquence de chaque unité textuelle dans chacune des parties du corpus. Ce module permet d'obtenir le tableau des

spécificités d'une partie sélectionnée ou d'un ensemble de parties par rapport au corpus en entier ou par rapport à un ensemble de parties plus large que celui sélectionné. Nous présentons dans la figure 9 ci-après, deux captures d'écran montrant le résultat de l'analyse des spécificités de la partie 08 (Nuit 08) de notre corpus (lemmatisé) pour la partition "Nuit". La capture de gauche représente les spécificités positives et celle de droite les spécificités négatives. Dans chaque capture, la colonne de gauche contient les lemmes spécifiques, la deuxième colonne la fréquence totale de chaque lemme dans le corpus, la troisième colonne la fréquence de chaque lemme dans la partie sélectionnée (ici la Nuit 08) et dans la dernière colonne on trouve le coefficient de spécificité correspondant à chaque lemme caractéristique de la (des) partie(s) sélectionnée(s).



Spécificités positives Spécificités négatives

Figure 34
Captures d'écran représentant les spécificités de la partie 08 pour la partition Nuit

Si le corpus est composé d'une série textuelle chronologique divisée en périodes, tel un journal, *Lexico3* est doté d'un module spécial très utile pour ce genre de corpus,

SPÉCIFICITÉS CHRONOLOGIQUES, qui calcule les spécificités chronologiques du corpus ainsi que les accroissements spécifiques pour chacune des parties.

- Le module AFC (Analyse Factorielle des Correspondances), quant à lui, permet d'effectuer une *analyse factorielle des correspondances* sur l'ensemble des parties du corpus en partant du tableau lexical entier (TLE) constitué à partir d'une partition du texte. Le but de la méthode de l'AFC est de réaliser, à partir de tableaux de données à double entrée (tableaux de contingence), en l'occurrence ici le TLE, des graphiques permettant une meilleure analyse des données et de rassembler le maximum d'informations contenues dans ces tableaux. La fenêtre de paramétrage de l'AFC dans *Lexico3* permet d'indiquer, par exemple, le nombre des unités textuelles prises en compte dans l'analyse et le nombre des facteurs à extraire, sachant que, par défaut, *Lexico3* prend en compte les unités dont la fréquence est supérieure à 10, mais ce seuil peut être revu à la hausse ou à la baisse.

Cependant, pour appliquer l'Analyse Factorielle des Correspondances à notre corpus, nous n'avons pas jugé utile d'utiliser le module AFC de *Lexico3* parce que nous le trouvons assez limité surtout au niveau de représentation graphique. En effet, en dépit de quelques outils pour agrémenter le graphique à savoir le pinceau et la boîte de couleurs permettant d'associer une couleur à un ensemble de parties, la représentation graphique, qui est tout de même le but de toute AFC, reste non paramétrable et la visualisation (même en changeant les axes) ne facilite pas l'interprétation du graphique. Nous avons préféré utiliser, comme nous l'expliquons plus bas, un outil externe, *XLSTAT* exécutant remarquablement ce type d'analyse statistique et s'intégrant d'une façon ergonomique à *MS Excel*.

2.3. Les outils de navigation lexicométrique

L'outil principal ici est le module CARTE DES SECTIONS qui permet une visualisation du corpus découpé en sections d'après les balises de section ou un caractère particulier attribué au sections, définis lors de l'étape de balisage avant de soumettre le corpus à *Lexico3*. L'une des fonctionnalités de cet outil est de permettre également de se déplacer parmi les résultats produits par les différents modules et le texte initial.

Nous présentons dans la figure 10 suivante, une capture d'écran représentant la carte des sections pour le lemme قَالَ *qâla* [Dire]. Cette carte de sections reproduit la ventilation de toutes les occurrences de ce lemme dans tous les paragraphes du corpus

Les carrés (représentant les sections) coloriés en rouge, sont les paragraphes qui contiennent au moins une occurrence de la forme choisie. Le contenu du paragraphe sélectionné (carré sélectionné) est affiché en bas de la fenêtre, dans la zone de visualisation.

Les boutons, en forme de mains, situés en bas à gauche de la fenêtre, permettent de passer à la section suivante ou précédente, pour le premier, et à l'occurrence suivante ou précédente de la forme graphique (ici le lemme) sélectionnée, pour le deuxième.

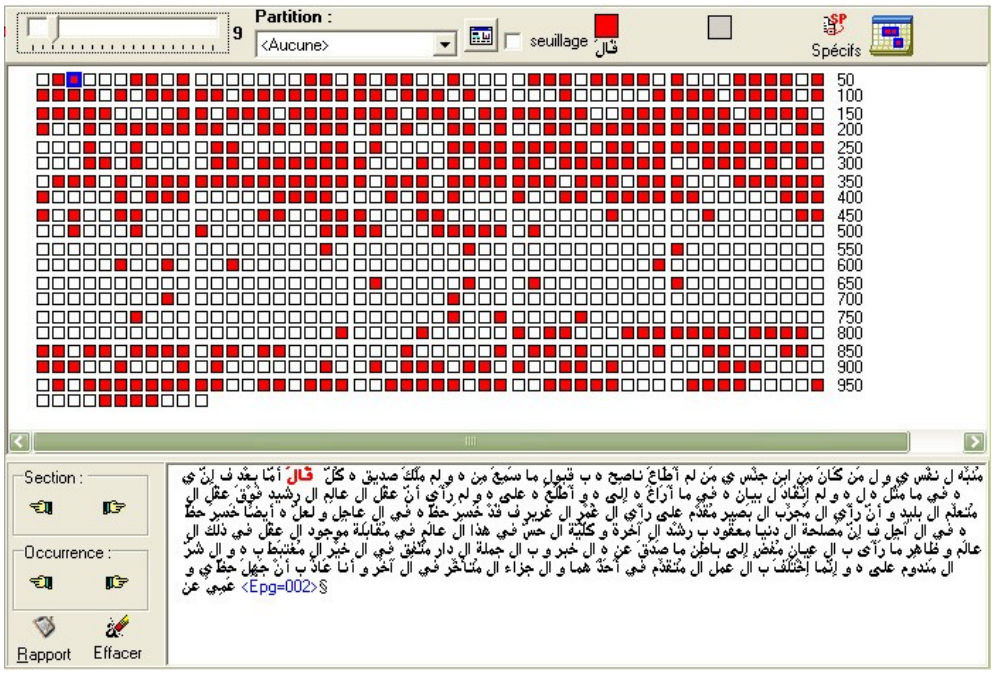


Figure 35
Capture d'écran représentant la carte des sections pour le lemme قَالَ

3. Récupération des données quantitatives

C'est donc à partir de *Lexico3* que nous avons récupéré les données quantitatives concernant les fréquences des formes, des lemmes et des catégories lexicales, globales et par partie, les coefficients de spécificité, etc. Certaines données sont récupérées directement sous format texte, d'autres sous format HTML puis convertis par nous sous format texte. Avant de les soumettre à MS *Excel*, certains fichiers ont dû être nettoyés de quelques caractères délimiteurs tels que point, astérisque, espace, etc.

Une fois récupérés dans *Excel*, ces fichiers sont soumis à quelques réorganisations comme par exemple, l'ajout de quelques en-têtes pour libeller les

colonnes, l'ordre de celles-ci, trier si nécessaire les données, etc., puis ces fichiers sont importés dans MS Access. Une base de données a donc été créée à cet effet, dont les tables ne sont autres que les feuilles constituant les classeurs *Excel* importés. Dans cette base de données, nous avons pu établir les relations nécessaires entre les enregistrements des différentes tables : nous avons ainsi pu obtenir que les formes (mots-formes) soient liées aux lemmes correspondants, les lemmes aux catégories lexicales associées, les formes aux différentes parties du corpus (Nuits, pages, ...), les lemmes aux coefficients de spécificité correspondants dans chaque partie, etc. Notons ici les données récupérées proviennent du traitement, par *Lexico3*, de plusieurs fichiers du même corpus (fichier des formes, fichier des lemmes, fichier des lemmes+catégorie_lexicales, fichier des catégories lexicales seules) qui ont été soumis séparément à *Lexico3*.

Outre les comptages, calculs, tris et autres requêtes de différentes natures qu'il nous a été aisé de manipuler à partir de cette base de données, le rôle principal de celle-ci est de nous permettre de confectionner le dictionnaire de fréquences avec deux versions différentes : un dictionnaire de fréquences alphabétique et un dictionnaire de fréquences hiérarchique.

La base de données Access ainsi créée est composée des tables suivantes :

- Table "Index_Pages" : elle contient les formes et leurs fréquences par page. À la fin de cette table ont été ajoutées deux colonnes (champs), une colonne "Lemme" qui va accueillir les lemmes correspondant aux formes et une colonne "cat" qui va accueillir leurs catégories lexicales.

- Table "Formes" : contient la liste des différents couples (forme, lemme). Cette table permet d'ajouter le lemme associé à chaque forme dans la table "Index_Pages" à l'aide d'un petit programme que nous avons nommé "Fusion". Cette table est générée par une requête de création de table nommée "generation de la table formes".

- Table "frequenceLemme" : cette table contient les lemmes et leurs fréquences totales. Elle est remplie par l'importation des données à partir de la feuille "Dic-Lemmes" du classeur "Données_Lemmes.xls".

- Table "specificites" : contient la liste des lemmes et leurs coefficients de spécificité le cas échéant.

- Table "Index_Pages_Lem" : cette table renferme les lemmes et leurs fréquences par page. Elle sert à insérer, dans le dictionnaire de fréquences, les fréquences des lemmes par page. Elle est remplie par l'importation des données à partir de la feuille "Lemmes_Pages" du classeur "Données_Lemmes.xls".

- Table "lemmesFormes" : cette table est capitale, c'est à partir d'elle que va être généré le dictionnaire de fréquences. Elle contient l'intégralité de la table "Index_Pages" plus l'intégralité de la table "frequencelemme". Cette table est générée par l'une des deux requêtes suivantes, selon le choix de l'ordre de tri à donner au dictionnaire de fréquences :

    - la requête de création de table "generation lemmesformes par ordre alphabetique"
    - la requête de création de table "generation lemmesformes par ordre de frequence".

4. Confection du dictionnaire de fréquences

Comme nous l'avons signalé plus haut, la confection du dictionnaire de fréquences a donc été possible grâce à la base de données que nous avons construite à partir des données récupérées des fichiers (intermédiaires et finaux) de *Lexico3*.

Le programme[108] qui a servi à la confection de ce dictionnaire de fréquences, est écrit en Visual Basic (VB). Le format de sortie est le format ".xls" sous *Excel* (voir la capture d'écran dans la figure 11). Le programme va chercher les données à partir de la table qui capitalise toutes les informations puis crée un classeur *Excel* dans lequel il va écrire une première ligne comme suit : fusionner les deux première cellules et y écrire le lemme ayant la plus forte fréquence, dans le cas d'un dictionnaire hiérarchique, ou le premier lemme dans l'ordre alphabétique, dans l'autre cas ; écrire le code de la catégorie lexicale correspondante au lemme dans la cellule suivante, suivie de la fréquence dans la cellule d'après ; puis dans les quinze cellules suivantes, écrire la fréquence du lemme dans chacune des quinze parties du corpus. Le programme passe ensuite à la deuxième ligne dans laquelle il inscrit à partir de la cinquième cellule (c'est-à-dire dans la colonne correspondant à la première partie du corpus) le coefficient de spécificité du lemme dans chacune des quinze parties. Passant à la troisième ligne, le programme écrit, dans la deuxième cellule, la première forme fléchie du lemme ; puis à l'intersection de chaque colonne représentant une partie, il inscrit le numéro de la page dans laquelle la forme fléchie est rencontrée suivi de " : " (deux points) suivis de la fréquence de la forme dans la page (exemple, "p4:34" veut dire que la forme fléchie est rencontrée 34 fois dans la page 4). Les fréquences par page vont être inscrites les unes en dessous des autres dans la colonne d'une partie donnée, c'est-à-dire une page plus une fréquence par ligne. Une fois la dernière fréquence par page de la première forme fléchie écrite, le programme passe à la deuxième forme fléchie du même lemme, puis la troisième, ..., puis la $n^{ième}$ le cas échéant. Après la dernière ligne dans laquelle est inscrite la dernière fréquence par page de la dernière forme fléchie du même lemme, le programme passe au lemme suivant pour réitérer la même boucle. Entre deux boucles (c'est-à-dire entre deux lemmes), nous avons demandé au programme d'insérer une ligne vide.

Après la description que nous avons faite plus haut des différentes tables que comporte la base de données qui est à l'origine de la confection du dictionnaire de

[108] Nous tenons à remercier ici Ramzi Abbès qui nous a aidé dans l'écriture de ce programme.

fréquences, nous décrivons ci-après sommairement, mais dans l'ordre exact, les différentes étapes de cette opération à l'issue de laquelle le dictionnaire de fréquences est constitué dans sa forme finale.

- Génération de la table "Index_Pages" à partir de la feuille "Formes_Pages" du Classeur "Données_Formes.xls".
- Génération de la table "frequenceLemme" à partir de la feuille "Dic-Lemmes" du Classeur "Données_Lemmes.xls".
- Intégration des catégories : l'objectif est ici de regrouper lemmes, fréquences et catégories dans la même table. Les données sont importées à partir de la feuille « Lemmes-Cat » du classeur *Excel* « Données_Lemmes.xls ».
- Génération de la table "Index_Pages_Lem" à partir de la feuille "Lemmes_Pages" du Classeur "Données_Lemmes.xls".
- Importation des données concernant les coefficients de spécificité à partir des feuilles du classeur « Spécificités.xls ».
- Lancement de la procédure "Fusion".
- Génération de la table "lemmesFormes" de l'une des deux façons décrite plus haut.
- Lancement enfin du programme de génération du dictionnaire de fréquences. Le résultat est un fichier *Excel* dont la capture d'écran est présentée dans la figure suivante :

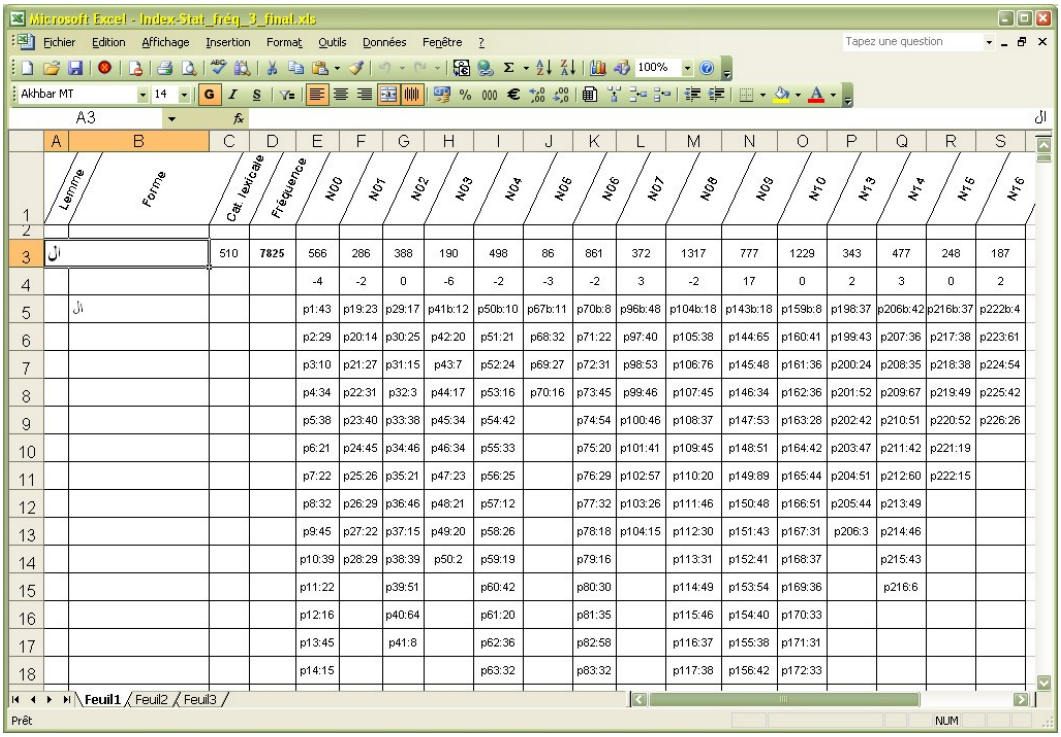


Figure 36
Capture d'écran du dictionnaire de fréquences (ici hiérarchique)

5. Utilisation d'autres outils

Dans la perspective d'une étude lexicométrique comme la nôtre, touchant à plusieurs aspects de la structure lexicale et à la trame radicale d'un corpus arabe, le recours à un certain nombre d'outils divers et complémentaires s'avère inévitable. *Lexico3* est un outil lexicométrique incontournable, utilisable pour l'arabe, il est performant, transparent et permet d'exploiter les fichiers intermédiaires et finaux. C'est pour toutes ces raisons que nous avons adopté ce logiciel. Mais un seul outil n'est cependant pas suffisant pour pouvoir mener à bien nos tâches de traitement, d'analyse et d'interprétation des données textuelles, objet de notre travail. Nous nous sommes donc trouvé dans l'obligation de compléter cet énorme parcours fait en compagnie de *Lexico3* par d'autres outils ayant la fonction de tableur, de calculateur statistique, de système de gestion de bases de données ou parfois même de "simple" traitement de texte.

5.1. *Excel* de Microsoft

Excel (ou tout autre tableur équivalent comme *Calc* de OpenOffice.org) permet d'utiliser à bon escient toutes les formules choisies, de faire des rectifications et d'effectuer facilement des changements et retours en arrière qui s'imposent quelques fois. Aussi, *Excel* est-il doté d'une panoplie de types et sous-types de graphiques bien étoffés et surtout personnalisables et paramétrables jusqu'au petit détail. L'absence, à l'heure actuelle, dans *Lexico3* de certaines fonctionnalités liées à d'autres aspects de la structure lexicale d'un corpus comme par exemple, la mesure de richesse lexicale ou de connexion lexicale, rend indispensable le recours à un outil tel qu'*Excel* dans lequel, en plus de ses fonctions intégrées, on peut poser soi-même les formules de calcul

nécessaires à cet effet et générer des représentations graphiques adéquates aux aspects et mises en forme recherchés.

5.2. XLSTAT d'Addinsoft

XLSTAT est un outil complémentaire (Add-in) à *Excel* extrêmement utile. Parce qu'intégré d'une façon ergonomique à *Excel*, il facilite énormément la manipulation et le traitement des données puisque son fonctionnement s'appuie sur Microsoft *Excel* aussi bien pour la récupération des données que pour la publication des résultats, qui peuvent au demeurant être repris et manipulés. C'est un outil d'analyse de données et de statistiques capable d'effectuer mieux qu'*Excel* certaines analyses complexes comme l'AFC (Analyse Factorielle des Correspondances), l'ACP (Analyse en Composantes Principales), l'AFD (Analyse Factorielle Discriminante), l'ACM (Analyse des Correspondances Multiples), la CAH (Classification Ascendante Hiérarchique), les "Nuées dynamiques" (k-means). Nous en avons utilisé principalement quatre fonctionnalités d'analyse de données et une fonctionnalité de visualisation :

- Pour l'analyse de données, il s'agit de

  - l'analyse factorielle des variables latentes avec rotation varimax pour comparer le classement des parties de notre corpus sur la base de la richesse lexicale selon quatre méthodes, Muller, Brunet, Guiraud et Yule-Herdan,

  - le test de corrélation des rangs de Spearman, d'un côté pour évaluer le classement des parties du corpus selon chacune des méthodes de calcul de richesse lexicale, et de l'autre côté dans l'accroissement du vocabulaire pour juger de la corrélation des rangs entre les classements des parties du corpus selon l'ordre chronologique, l'étendue relative du vocabulaire et l'accroissement lexical,

  - la matrice de corrélation de Pearson entre les Nuits sur la base des écarts réduits entre effectifs réels et effectifs théoriques des catégories lexicales pour chaque Nuit. Le but était d'évaluer le degré de corrélation ou d'opposition entre les Nuits prises deux à deux,

- l'AFC dans le cadre de l'étude des catégories lexicales. Le but principal de l'analyse factorielle des correspondances était de représenter dans le même espace les lignes et les colonnes du tableau à synthétiser - les Nuits et les catégories lexicales - de telle façon que chaque Nuit (ou chaque catégorie lexicale) ait une seule localisation résumant au mieux la distance variable entre elle et chacune des catégories lexicales (ou chacune des Nuits).

- Pour ce qui concerne la fonction de visualisation, nous avons utilisé le module complémentaire *XLSTAT-3Dplot* dans l'étude de la mesure de richesse lexicale afin de prendre en compte plus d'information concernant la dispersion des points-Nuits et visualiser, en trois dimensions, la représentation des Nuits en fonction des méthodes.

En revanche, la lisibilité de la représentation graphique de l'analyse factorielle des correspondances générée par le module AFC d'*XLSTAT*, laisse beaucoup à désirer du fait de l'encombrement du graphique créé par la juxtaposition des points-lignes ou des points-colonnes et de leurs libellés. Mais comme *XLSTAT* est intégré à *Excel*, il est bien aisé d'enrichir et de paramétrer les graphiques générés par *XLSTAT* en utilisant l'outil graphique d'*Excel*. Cette perte de lisibilité au niveau de la représentation graphique résultante de l'AFC dans *XLSTAT* pourrait aussi être comblée par le recours au module complémentaire *XLSTAT-3Dplot*.

Tous les outils de *XLSTAT* sont accessibles à partir, soit d'un menu déroulant (et des sous-menus) qui est ajouté à la barre de menus d'*Excel*, soit à partir d'une barre d'outils (et des barres d'outils dérivées) qui peut être flottante ou intégrée à la barre d'outils d'*Excel*, comme le montre la figure 12 suivante :

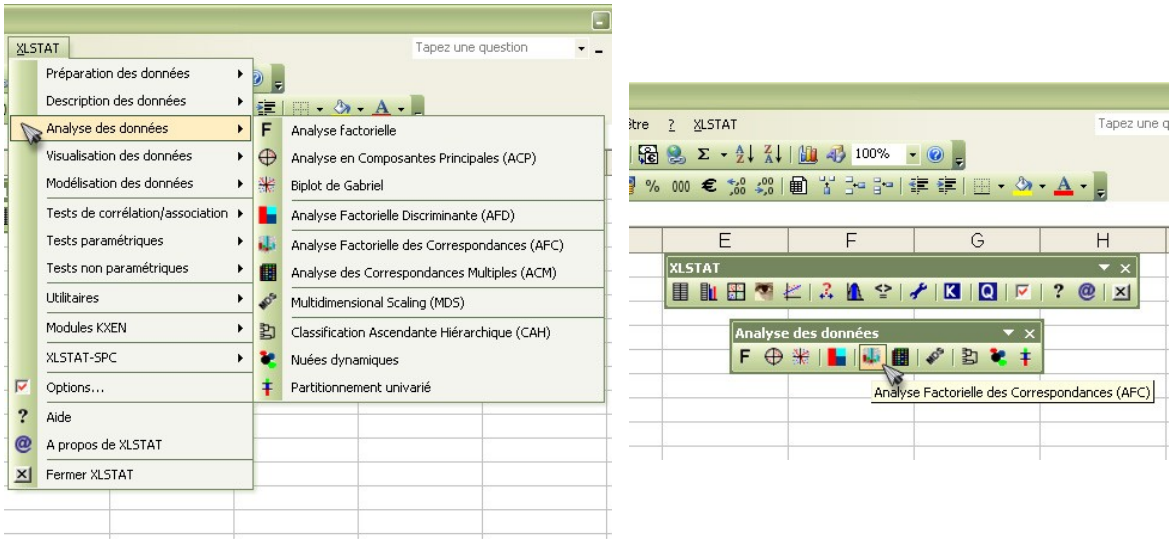


Figure 37
Captures d'écran représentant les menus déroulants de *XLSTAT* et ses barres d'outils

5.2.1. *XLSTAT-3Dplot*

XLSTAT-3Dplot[109] n'est à vrai dire pas un outil indépendant à part entière, il est un de sept modules optionnels répondant à des besoins plus spécifiques (séries chronologiques, analyse de survie, contrôle statistique des procédés, analyse des effets de doses en chimie et pharmacologie...) qui ajoutent d'importantes fonctionnalités à *XLSTAT*. Ce module permet, en effet de créer des visualisations spectaculaires en trois dimensions. Les données y sont visibles sous plusieurs angles. Nous l'avons utilisé pour visualiser la représentation des Nuits en fonction des méthodes de richesse lexicale comme le montre la figure suivante et que nous expliquerons dans le chapitre consacré à la richesse lexicale :

[109] http://www.xlstat.com/fr/products/xlstat-3dplot/

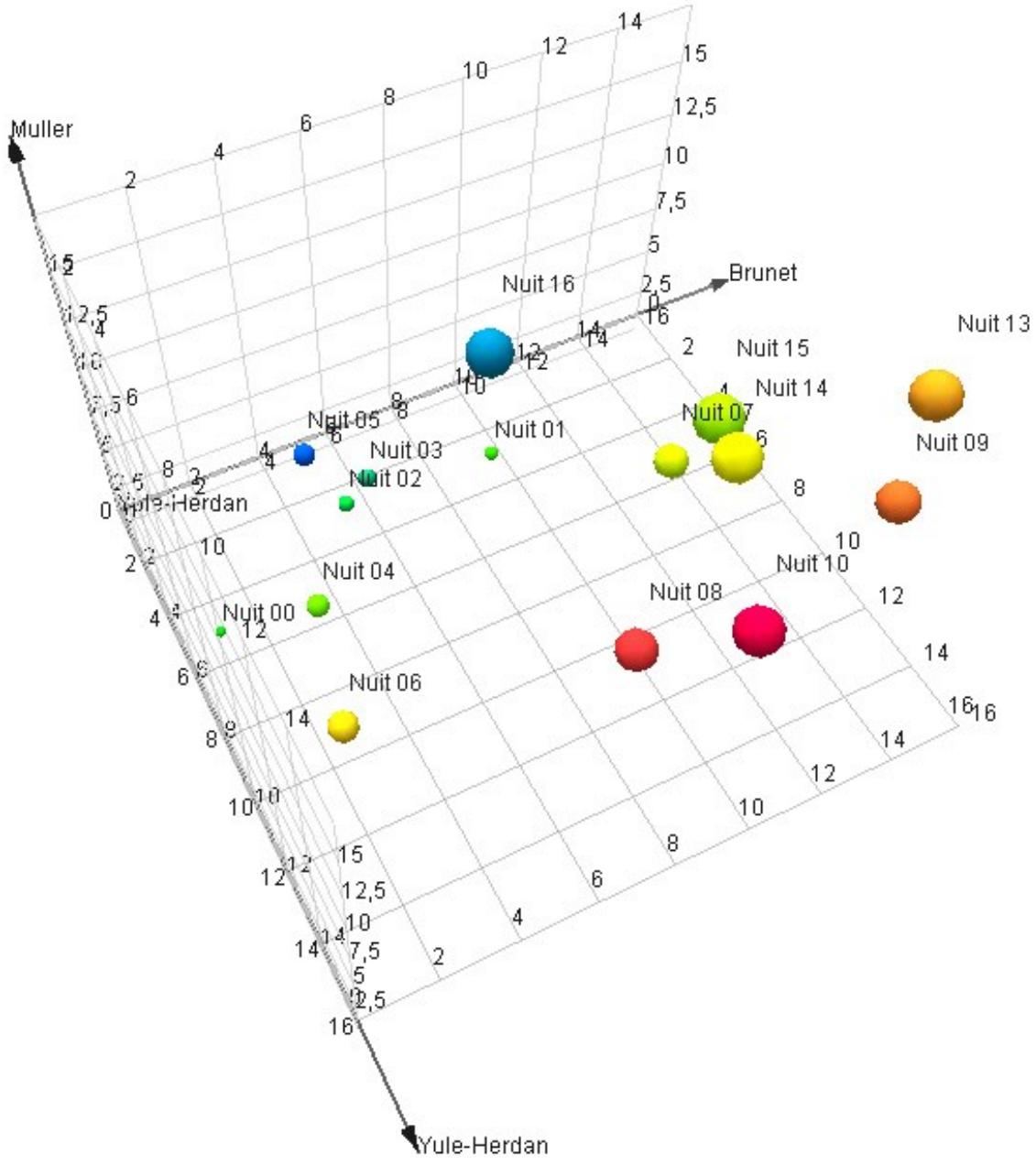


Figure 38
Représentation graphique en 3D faite avec *XLSTAT-3Dplot*