Hyperbase et Lexico 3, outils lexicométriques pour l'historien

Emmanuel Bonin et Alain Dallo

Édition électronique

URL : <https://journals.openedition.org/histoiremesure/840>
DOI : 10.4000/histoiremesure.840
ISSN : 1957-7745

Éditeur

Éditions de l'EHESS

Édition imprimée

Date de publication : 2 décembre 2003
Pagination : 389-402
ISBN : 2-222-96740-6
ISSN : 0982-1783

Référence électronique

Emmanuel Bonin et Alain Dallo, « Hyperbase et Lexico 3, outils lexicométriques pour l'historien », *Histoire & mesure* [En ligne], XVIII - 3/4 | 2003, mis en ligne le 03 avril 2007, consulté le 21 avril 2022.
URL : <http://journals.openedition.org/histoiremesure/840> ; DOI : <https://doi.org/10.4000/histoiremesure.840>

Ce document a été généré automatiquement le 21 avril 2022.

© Éditions de l'EHESS

Hyperbase et Lexico 3, outils lexicométriques pour l'historien

Emmanuel Bonin et Alain Dallo

- 1 L'historien désireux d'utiliser la lexicométrie au sein de sa démarche historique dispose à ce jour d'outils logiciels essentiellement développés au sein de laboratoires de linguistique. Nous étudierons la palette des fonctions offertes par deux d'entre eux, Hyperbase 5.3 conçu par Étienne Brunet dans le cadre de l'UMR 6039 CNRS, Université de Nice, et Lexico 3.45 réalisé par l'équipe universitaire SYLED-CLA2T sous la responsabilité d'André Salem¹. Une approche comparative se justifie pour démontrer que si les deux logiciels diffèrent sur nombre de points importants, ils ne s'opposent pas mais se complètent, leur usage conjoint étant la meilleure solution à adopter.
- 2 Pour étudier ces deux programmes, nous nous fonderons sur l'intégration et l'étude d'un corpus de sept grands discours de Saint-Just prononcés à la Convention entre le 8 juillet 1793 et le 9 thermidor an II². Ce corpus servira simplement de support aux exemples et n'aura dans le peu de place dont nous disposons aucune vocation heuristique. Nous suivrons cependant une méthode allant d'une première approche des textes par un lecteur qui à la limite, n'aurait pas lu le corpus (lecture des textes, spécificités externes, comptage des formes, occurrences et hapax) à l'analyse de thèmes transversaux (concordances, thème, accroissement lexical, distribution) en passant par le positionnement des textes les uns par rapport aux autres (essentiellement grâce à la méthode de l'AFC et au calcul des spécificités internes).
 1. Les logiciels et l'intégration des données
- 3 Hyperbase et Lexico 3 sont développés sous le système d'exploitation commercial Windows vendu par Microsoft (Hyperbase fonctionne également sous Mac OS), mais il faut tout de suite noter qu'ils sont, comme de très nombreux programmes, assez mal adaptés aux versions réellement multi-utilisateurs de ce logiciel. Celui qui voudra les utiliser dans une architecture administrateur-utilisateur devra soit créer des failles de sécurité en ouvrant des droits à l'utilisateur sur certains répertoires, sous peine de ne pas pouvoir utiliser tout ou partie des fonctions des logiciels, soit se connecter en mode administrateur pour disposer de tous les droits. C'est bien évidemment regrettable,

mais monnaie courante pour des programmes dont les premières versions ont été développées sous des systèmes mono-utilisateurs comme Windows 95 ou les anciennes versions de Mac OS.

- 4 L'intégration des données dans la base lexicométrique nécessite leur préparation, une mise en forme propre à chaque logiciel. Sous Hyperbase, elle consiste à découper le corpus rassemblé au sein d'un fichier texte unique à l'aide de balises du type `&&text1&&`, propres à chaque niveau de découpage (texte, page, paragraphe). Un seul découpage des textes est possible au sein du corpus, limité à 76 textes ; il faut créer une nouvelle base si l'on veut l'aborder avec un découpage différent³. Lexico 3 utilise un découpage en balises du type `<mois=01>`, qui permet cette fois-ci de superposer plusieurs niveaux de découpage (par exemple par mois, semaine et jour : `<mois=01><semaine=03><jour=17>`) au sein du même corpus, sans avoir à modifier le fichier texte et ingérer à nouveau les données.
- 5 Hyperbase ne se préoccupe pas de la casse pour le découpage automatique des formes⁴, contrairement à Lexico 3 qui oblige donc à enlever les majuscules en début de phrase pour éviter de distinguer les formes « Je » et « je », par exemple. Dans les deux cas, l'intégralité du texte situé hors des balises de découpage est intégré à la base ; il est donc impossible de conserver des notes ou des commentaires au sein du fichier ingéré.
- 6 Le travail de création de la base est sensiblement plus long sous Hyperbase et monopolise l'activité sur le poste. Le logiciel effectue, dès ce moment, la quasi-intégralité des travaux et calculs (création du tableau lexical entier, spécificités, distance lexicale, etc.), à l'exception des AFC, ce qui rend le fichier exécutable assez lourd à transporter (5,8 Mo dans le cas du fichier `saintjust.exe`, contre 248 ko pour le fichier `saintjust.num` généré par Lexico3). Lexico 3 est plus rapide, mais les calculs sont renvoyés à plus tard.

2. Première perception des textes et du corpus

- 7 L'interface de lecture des textes sous Hyperbase est performante, et associée à l'index permet de retrouver très rapidement des occurrences d'une forme donnée dans le corpus. Lexico 3 invente une sorte de cartographie avec la représentation graphique colorée des occurrences d'une (ou de plusieurs) forme(s), d'un groupe de formes, de segments répétés au sein du corpus. La lecture se fait ensuite par phrases, avec une navigation possible entre les différentes occurrences de la forme ou du segment. Ce système met immédiatement en valeur la distribution des formes au sein du corpus, et invite donc dès la lecture à passer à l'interprétation⁵.
- 8 Afin de situer le corpus par rapport à d'autres écrits, Hyperbase permet seul de calculer les spécificités externes du corpus par rapport à celui de Frantext ou à un découpage chronologique au sein de celui-ci. Ainsi, les spécificités positives des discours de Saint-Just par rapport au vocabulaire des textes de la période 1780-1820 mettent en avant le vocabulaire politique de l'orateur, et plus particulièrement ce qui caractérise les luttes de la période (le problème de la corruption et de la lutte des factions notamment, avec des formes comme « factions » ou « Fabre »). Les spécificités négatives singularisent un langage qui recourt peu aux pronoms personnels en comparaison avec un corpus essentiellement littéraire (« je », « j », « me », « moi », « mon »).
- 9 Les deux logiciels rendent compte de la longueur des textes, du nombre de formes, occurrences et hapax au sein de chaque partie du corpus. Les différences de comptage parfois sensibles (ainsi Hyperbase dénombre 622 occurrences de la forme « vous », Lexico 3 seulement 555) s'expliquent par la définition des délimiteurs de forme,

différente d'un programme à l'autre. Lexico 3 permettant de déterminer finement les délimiteurs, nous obtenons des résultats quasi identiques avec les deux logiciels si nous utilisons les 17 délimiteurs d'Hyperbase dans Lexico 3 (622 occurrences dans les deux cas pour la forme « vous »). Lexico 3 précise la forme la plus fréquente dans chaque texte.

- 10 Il n'y a pas de différence dans le décompte du nombre total d'hapax pour le corpus (3 599 hapax dans les deux cas). Le décompte des hapax par texte repose, en revanche, sur des principes différents : Hyperbase donne le nombre des hapax du corpus présents dans chaque texte (ainsi T1 comporte 719 hapax sur les 3 599 hapax du corpus), alors que Lexico 3 considère chaque texte indépendamment des autres et compte les formes qui n'apparaissent qu'une fois dans ce texte (par exemple, T1 comprend 1 449 hapax sur 2 348 formes du texte T1). Hyperbase permet une représentation graphique de tous ces résultats. Ici, les textes 1 et 5 sont les plus longs et présentent le plus d'hapax.

3. Positionnement des textes les uns par rapport aux autres

- 11 La méthode de l'analyse factorielle des correspondances, qui facilite la lecture de grands tableaux de données et leur interprétation, est adaptée à l'étude lexicométrique d'un corpus de textes. Les deux logiciels offrent donc la possibilité d'effectuer des AFC.

Analyse factorielle des correspondances

- 12 Hyperbase permet de faire une AFC à partir du tableau lexical des formes les plus fréquentes dans le corpus, le seuil de fréquence variant en fonction de la taille du document, pour des raisons de mémoire. Ce seuil n'est pas indiqué de manière transparente ; il semble que, dans le cas de Saint-Just, il soit de 5. Le tableau soumis à l'analyse est un tableau d'écarts réduits calculé selon la loi hypergéométrique. Avec Lexico 3, l'utilisateur peut choisir la fréquence minimum nécessaire à l'intégration d'une forme dans le tableau analysé, et en choisissant la fréquence 1, réaliser une AFC sur l'ensemble du tableau lexical entier⁶.
- 13 Hyperbase met en opposition trois textes sur les facteurs 1 et 2, T1, T2 et T5, qui, seuls, ont une contribution supérieure à la moyenne sur les deux facteurs. Le facteur 1 a une valeur propre de 24 % et le facteur 2 de 22 %. Lexico 3, pour une AFC portant sur les vocables ayant une fréquence minimum de 5 dans le corpus, donne un graphique comparable avec une valeur propre de 26 % pour le facteur 1 et de 20 % pour le facteur 2.
- 14 Seul Hyperbase permet de récupérer le tableau lexical entier pour l'exporter au format texte et le soumettre à une AFC avec d'autres outils (des logiciels statistiques comme R ou SPAD, des macros Excel). Dans la mesure où l'on ne peut afficher l'environnement lexical dans les AFC de ces deux logiciels, c'est-à-dire placer les formes dont les contributions pour les facteurs représentés sont supérieures à la moyenne, le traitement complémentaire par des outils externes semble inévitable⁷. De plus, il n'est pas possible de fixer un seuil de contribution pour la représentation des vecteurs colonnes (les textes), ce qui peut rendre le graphique difficile à lire en cas de découpage important (plusieurs dizaines de textes). Saluons, cependant, les nettes améliorations des dernières versions des deux logiciels, qui permettent désormais par un simple clic ou passage de souris sur un point d'en connaître les contributions pour les facteurs représentés.
- 15 Hyperbase permet également d'effectuer des AFC sur les distances lexicales⁸. Mais, dans la version 5.3, l'option « Analyse sur V » donne des résultats qui semblent peu fiables.

Elle travaille sur la matrice distance V multipliée par 1 000, mais quand nous regardons les tableaux de résultats, ces derniers sont différents pour les contributions et coordonnées des lignes et des colonnes, alors que l'on est face à une matrice carrée symétrique ! La mise en valeur des textes 4 et 7 sur les facteurs 1 (83 %) et 2 (15 %) est donc à prendre avec précaution.

- 16 Les options « Analyse factorielle sur distance V » et « Analyse factorielle sur distance N » sont à préférer, qui calculent les AFC sur les distances V et N sans coefficient multiplicateur. Dans le premier cas, cinq textes ont des contributions supérieures à la moyenne, T1, T2, T3, T5 et T7, pour un facteur 1 à 77 % et un facteur 2 à 13 %. Dans le deuxième, six textes ressortent (tous sauf T4), avec un facteur 1 à 64% et un facteur 2 à 22%. Les graphiques sont proches, et facilitent l'interprétation des tableaux des distances lexicales.
- 17 Dans les deux analyses de vocabulaire (Hyperbase et Lexico 3), comme dans les deux analyses de distance lexicale (Hyperbase), les textes 1, 2 et 5 sont situés dans des quarts du graphique nettement singularisés, les différentes méthodes donnent donc des résultats similaires.
- 18 Hyperbase permet une analyse arborée radiale et rectangulaire sur les distances N et V, mais nous renvoyons à l'article publié dans ce volume pour une étude précise de ces méthodes.
- 19 Nous avons vu avec l'AFC qu'Hyperbase a comme particularité de permettre de calculer une distance lexicale entre les textes et d'en donner un tableau et des représentations graphiques. Le calcul de ces distances entre textes a évolué dans les dernières versions. Améliorant une ancienne fondée sur un calcul simple à partir du vocabulaire commun et exclusif (dont les hapax), Étienne Brunet en a intégré une nouvelle méthode tenant compte de la différence de fréquence pour une forme apparaissant dans les deux textes en se fondant sur l'algorithme de Dominique Labbé⁹. Complémentaire de l'AFC du vocabulaire, l'étude des distances lexicales est un outil précieux pour positionner les textes les uns par rapport aux autres¹⁰.

Les spécificités internes

- 20 Hyperbase permet aussi d'extraire des « phrases-clés », « séquences de paragraphes caractéristiques par quoi on tente de donner une idée du contenu d'un texte ». Pour cela, le logiciel calcule un quotient déterminant pour ne retenir que les paragraphes contenant les spécificités les plus nombreuses et significatives, complétant l'étude des spécificités lexicales par celle de spécificités phrastiques¹¹.
- 21 Les deux logiciels permettent de travailler sur l'évolution du vocabulaire : Hyperbase par un calcul de l'évolution lexicale fondé sur un coefficient de corrélation entre les valeurs de l'écart réduit et le rang de chaque élément¹² ; Lexico 3 par un calcul sur les spécificités chronologiques¹³ ou évolutives (calcul des spécificités uniquement par rapport aux parties précédentes dans le corpus). Il met ainsi en évidence des formes ou segments qui ont tendance à être sur et sous-utilisées dans la période considérée par rapport aux périodes qui précèdent.
- 22 Très complémentaire de l'AFC pour l'analyse des particularités de chaque texte, l'étude des spécificités internes au corpus est une des applications lexicométriques majeures. Si les résultats sont là encore similaires entre les deux logiciels (ainsi les onze premières spécificités positives du texte 1 sont dans les deux cas, même si l'ordre diffère un peu : « on », « détenus », « roi », « anarchie », « convention », « avait », «

sections », « ils », « valazé », « maire », « Buzot »), Lexico 3 offre plus de possibilités. Il permet ainsi de voir les spécificités négatives de chaque texte, et crée un grand tableau de toutes les spécificités de toutes les formes du dictionnaire. Et, surtout, il autorise à regrouper plusieurs textes en un ensemble pour lequel on calcule les spécificités négatives et positives. Ainsi, les textes 3, 4 et 6 pris ensemble depuis le graphique de l'AFC sur le vocabulaire ont parmi les spécificités positives les plus fortes les formes « bonheur », « peuple », « veut », « étranger », « aristocratie », « corruption », « ennemis »¹⁴. Si l'on a généré une liste des segments répétés, elle est prise en compte dans le calcul des spécificités.

4. Analyse de thèmes transversaux

- 23 Une fois le corpus pris en main et les textes singularisés, le travail consiste à analyser les thèmes qui traversent le corpus.
- 24 Pour cela, Lexico 3 comme Hyperbase autorisent un travail sur les concordances d'une forme, avec un tri chronologique ou un tri sur son contexte gauche ou droit. Hyperbase et Lexico 3 permettent de rentrer une forme, une chaîne de caractères initiale ou finale d'une forme ainsi qu'une liste de formes créée au préalable. En outre, Hyperbase autorise l'entrée d'un lemme ou d'une expression. Pour Hyperbase, l'accès au texte depuis la liste des concordances se fait par un simple clic sur l'extrait. Lexico 3 permet de choisir la longueur de l'extrait de concordance (ce qui est l'objet de l'option contexte sous Hyperbase), de regrouper les concordances par texte et de garder ce découpage lorsqu'on les trie sur le contexte gauche ou droit. Lexico 3 propose un inventaire distributionnel, c'est-à-dire un regroupement des concordances permettant de repérer des expressions répétées incluant la forme ou le segment étudié. Hyperbase permet lui l'étude de la cooccurrence possible entre deux formes distantes dans le cadre d'un contexte un peu trop large, celui de la page. Il est dommage que l'on ne puisse paramétrer finement la taille de ce contexte de cooccurrence.
- 25 L'originalité d'Hyperbase est l'option Thème, qui exécute un calcul de spécificités entre les formes de l'ensemble des contextes considérées comme un texte et l'ensemble du corpus. Cela donne ainsi un ensemble de formes que l'on peut considérer comme thématique : par exemple, les contextes de faction(s) donnent une spécificité positive pour ces deux formes, mais aussi pour les formes « indulgents », « partis », « criminelles », « criminelle », « dynastie », « Orléans », « étranger », « Necker », « aristocratie ».
- 26 Celle de Lexico 3 est de pouvoir travailler non seulement à partir d'une forme, mais aussi à partir de segments répétés, chers aux travaux d'André Salem. Ceux-ci sont extraits dans un tableau en fonction d'un seuil de fréquence minimum du segment choisi par l'utilisateur, 10 par défaut. Ainsi nous retrouvons 13 occurrences du segment « ennemis du peuple » et 11 de celui « ennemis de la révolution ».
- 27 Hyperbase permet de générer une liste en saisissant manuellement les formes, expressions ou chaînes de caractères, puis de la traiter au sein des menus concordance, contexte, thème, analyse arborée ou AFC (en ce cas, le tableau analysé est construit à partir des écarts réduits en utilisant la loi hypergéométrique comme pour le dictionnaire).
- 28 Lexico 3 rend possible la constitution de groupes de formes ou de segments pour étudier un thème ou travailler sur des graphies différentes d'un même lemme, et met à disposition un système de troncature très efficace pour les constituer. Nous pouvons

afficher les concordances du groupe, enregistrer ce groupe, établir une carte de section, mais pas réaliser d'AFC.

- 29 Le graphique des distributions (en histogrammes) d'une forme ou plusieurs formes au sein du corpus permet l'étude d'une corrélation entre ces formes. Hyperbase représente un calcul d'écarts réduits pour chaque forme, et limite la représentation à deux formes. Lexico 3 affiche la fréquence relative ou absolue en courbes, ou bien les spécificités de la forme en histogrammes. Il permet de superposer les distributions d'un nombre virtuellement illimité de formes ou de segments, ce qui semble un atout important¹⁵.
- 30 L'étude de l'accroissement du vocabulaire sous Hyperbase se fait à partir du calcul d'un écart pondéré pour chaque texte et permet de constater que le renouvellement chronologique du vocabulaire est de plus en plus ralenti, notamment grâce à une représentation graphique en histogramme. Nous pouvons mesurer l'évolution du premier discours vers le dernier, ou du dernier vers le premier, démarche purement théorique mais qui peut avoir un intérêt. Sous Lexico 3, l'accroissement du vocabulaire est mesuré sous forme graphique pour l'ensemble du corpus, pour un texte ou pour un ensemble de textes. Mais, surtout, l'établissement du diagramme de Pareto dans le seul Lexico 3 permet de comparer la répartition du vocabulaire à la loi de Zipf, selon laquelle le produit rang*fréquence est constant. Cela n'est valable que pour des corpus de très grande taille ; nous n'avons donc pu en vérifier l'intérêt dans le cas du corpus des textes de Saint-Just.

5. Sauvegarde des résultats

- 31 Sous Hyperbase, les résultats sont des documents textes non rassemblés, qu'il faut sauvegarder au fur et à mesure pour éviter qu'ils ne disparaissent. Ainsi, les AFC sont sauvegardées avec les mêmes noms de fichier, on risque donc d'écraser l'AFC précédente si l'on n'y prend garde. Sous Lexico 3, les documents choisis par l'utilisateur sont rassemblés dans un dossier de fichiers HTML, mais le système hypertexte reste à améliorer : il y a encore des erreurs d'enregistrement, ainsi nous ne pouvons conserver l'inventaire distributionnel.
- 32 Au terme de cet exposé technique comparatif des logiciels Hyperbase et Lexico 3, nous constatons donc que, malgré quelques menus défauts à corriger dans les mises à jour à venir, ces deux programmes présentent une vaste palette de fonctions permettant à l'historien de constituer et d'exploiter une base lexicométrique. Les principales fonctions (spécificités, AFC, concordance) sont prises en charge par les deux logiciels, de manière parfois sensiblement différente, notamment en termes de représentation graphique, mais généralement complémentaire. Chacun a ses particularités, ses spécialités propres (distance lexicale pour Hyperbase par exemple, étude des segments répétés pour Lexico 3), ce qui ne fait que confirmer le caractère indispensable d'une utilisation conjointe et simultanée de ces deux outils.

Annexe 3. Graphique de distribution des formes « je » et « j » (Hyperbase)

Annexe 4. Graphique des fréquences absolues des formes « je », « j », « me » et « moi » (Lexico 3)

Annexe 5. Graphique des spécificités des formes « je », « j », « me » et « moi » (Lexico 3)

NOTES

1. Site de l'UMR 6039 : <http://ancilla.unice.fr/default.html>. Site de lexico3 : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW>
2. T1 : Rapport sur les 32 membres de la Convention détenus en vertu du décret du 2 juin, présenté à la Convention nationale dans la séance du 8 juillet 1793. T2 : Rapport sur la nécessité de déclarer le gouvernement révolutionnaire jusqu'à la paix, présenté à la Convention nationale dans la séance du 19^e jour du 1^{er} mois de l'an II (10 octobre 1793). T3 : Rapport sur les personnes incarcérées, présenté à la Convention nationale dans la séance du 8 ventôse an II (26 février 1794). T4 : Rapport au nom du Comité de salut public sur les factions de l'étranger, présenté à la Convention nationale dans la séance du 23 ventôse an II (13 mars 1794). T5 : Rapport sur la conjuration ourdie depuis plusieurs années par les factions criminelles, pour absorber la révolution française dans un changement de dynastie, présenté à la Convention nationale dans la séance du 11 germinal an II (31 mars 1794). T6 : Rapport sur la police générale, sur la justice, le commerce, la législation et les crimes des factions, présenté à la Convention nationale dans la séance du 16 germinal an II (5 avril 1794). T7 : Discours du 9 thermidor an II (27 juillet 1794).
3. Le découpage des textes au sein du corpus est parfois quasi « naturel », il préexiste au travail de l'historien. C'est le cas ici des sept discours de Saint-Just, qui forment sept textes bien distincts. Mais dans le cas de l'étude d'un corpus d'éditoriaux quotidiens, par exemple, on peut vouloir découper le corpus en jours, en semaines, en mois, et considérer chacune de ces entités comme un texte dans les travaux statistiques.
4. Sauf si nous lui demandons de prendre en considération les noms propres : en ce cas, il considère les mots commençant par une majuscule comme des noms propres, à l'exception de ceux se situant en début de phrase, ce qui n'est pas sans présenter quelques difficultés.
5. Cf. Annexe 1, carte de section de la forme « je » dans le corpus.
6. Notons que le calcul de l'AFC n'est pas transparent sous Lexico 3 ; nous ne savons pas exactement à quoi ressemble le tableau initial analysé. Nous ne pouvons donc affirmer avec certitude que le logiciel réalise une AFC sur l'ensemble des formes du corpus.
7. Un seul exemple : en exportant le tableau lexical et en réalisant, grâce à une macro sous Excel, l'AFC des 200 formes les plus fréquentes, nous obtenons un graphique des deux premiers facteurs (VP = 34.4 % et 24.4 %) mettant en valeur deux nuages : un présentant une forte conjonction entre T5 (76 % sur le facteur 1) et les formes tu, te, toi, Dumouriez, Danton, Fabre, Orléans, parti ; l'autre présentant une forte conjonction entre T7 (81 % sur le facteur 2) et les formes je, j', me, m', ai, membres et comités. Ainsi, l'AFC insiste de façon spectaculaire sur le revirement du discours de Saint-Just, comminatoire dans son utilisation de la deuxième personne du singulier en germinal an II, beaucoup plus justificatif avec celle de la première personne en thermidor.
8. Voir un peu plus bas pour l'explication du calcul des distances lexicales entre textes.
9. Nous renvoyons à la page 3 de la mise à jour du guide d'utilisation d'Hyperbase pour l'explication des deux algorithmes de calcul.
10. Ainsi, la proximité de vocabulaire entre les textes 3, 4 et 6 observée sur l'AFC se retrouve très nettement dans l'étude des distances lexicales entre ces trois textes. Cf. Annexe 2, graphique de la distance lexicale entre le texte 4 et les autres textes.
11. Mise à jour du guide d'utilisation d'Hyperbase, p. 22 pour l'explication du calcul de ce quotient.

12. Mode d'emploi d'Hyperbase, p. 46 pour le calcul de la corrélation, Mise à jour du guide d'utilisation d'Hyperbase, p. 17 pour le calcul des écarts réduits à partir de la loi hypergéométrique.
13. Pour ces spécificités chronologiques, nous renvoyons à L. LEBART & A. SALEM, 1988, p. 197. On en trouvera une application dans l'article de J.-C. DEROUBAIX, 2001.
14. L'AFC permet de repérer les textes à regrouper en un seul pour en calculer les spécificités, et ici l'interface de Lexico 3 qui permet immédiatement de faire le lien entre AFC et spécificités est particulièrement performante.
15. Cf. Annexe 3 : graphique de distribution des formes « je » et « j » dans Hyperbase ; Annexe 4 : graphique des fréquences absolues et Annexe 5 : graphique des spécificités des formes « je », « j », « moi » et « me » dans Lexico 3.

RÉSUMÉS

À ce jour, aucun outil lexicométrique n'a été développé par des historiens pour des historiens. Il faut donc utiliser des logiciels mis au point au sein de laboratoires de linguistique. Cet article se propose de comparer les fonctions offertes par deux d'entre eux, Hyperbase et Lexico 3. Au terme d'une démarche d'étude d'un corpus de discours de Saint-Just, permettant d'aborder l'ensemble des étapes utiles au traitement quantitatif de données textuelles (lecture des textes, spécificités externes, comptage des formes, occurrences et hapax, analyse factorielle des correspondances, spécificités internes, concordances, thème, accroissement lexical, distribution), les auteurs montrent que les deux programmes, loin de s'opposer ou de se recopier, se complètent. Chacun propose des fonctionnalités propres (la distance lexicale sous Hyperbase, le traitement des segments répétés sous Lexico 3 par exemple). L'historien a donc intérêt à utiliser Hyperbase et Lexico 3 simultanément.

Hyperbase and Lexico 3, Lexicometrical Tools for the Historian. No lexicometrical tool has been worked out by historians for historians. For this reason, we have to use software developed by laboratories specialized in linguistical sciences. This contribution compares the fonctions of two among them, Hyperbase and Lexico 3. Proceeding through the study of a corpus of Saint-Just's speeches, allowing them to approach all of the steps useful to quantitative treatment of textual data (reading of the texts, external specificities, counting the forms, occurrences and hapax, factorial analysis, internal specificities, correspondences, theme, lexical growth, distribution), the authors show that both softwares, far from copying or opposing each other, are complementary. Each one proposes its own fonctions (lexical distance under Hyperbase, treatment of repeated segments with Lexico 3 for exemple). Historians should use both simultaneously.

INDEX

Mots-clés : histoire politique, analyse textuelle, informatique

Index chronologique : Révolution française

Index géographique : France

AUTEURS

EMMANUEL BONIN

Université de Paris 1-Panthéon Sorbonne, UFR d'histoire, 1 rue Victor Cousin,
75005 Paris.
bonin@univ-paris1.fr

ALAIN DALLO

Université de Paris 1-Panthéon Sorbonne, UFR d'histoire, 1 rue Victor Cousin,
75005 Paris.
adallo@univ-paris1.fr