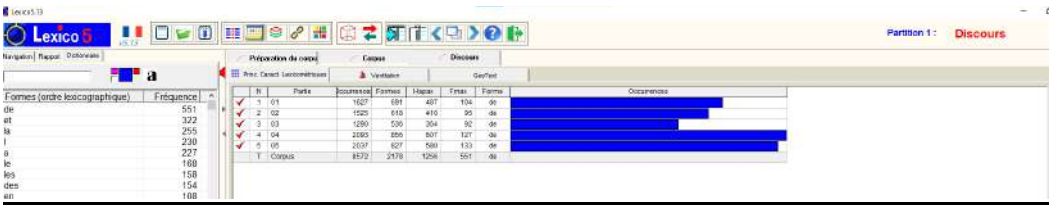
République du Sénégal
Ministère de l'Enseignement Supérieur, de la Recherche et de L'innovation
UNIVERSITÉ CHEIKH ANTA DIOP de DAKAR
Centre de Linguistique Appliquée de Dakar

FICHE TECHNIQUE

UTILISATION DE METHODES ET D'OUTILS LEXICOMETRIQUES POUR ANALYSER DES CORPUS DE DISCOURS INSTITUTIONNELS



Conçue par

Dr Augustin NDIONE
Chargé de recherche en linguistique descriptive et didactique des langues
Dr Noël Bernard BIAGUI
Chargé de recherche en linguistique descriptive et didactique des langues

MARS 2021

Résumé
Cette fiche présente la lexicométrie comme un outil pour l'analyse des discours institutionnels mais aussi l'analyse de tous types de corpus volumineux construits en respectant deux éléments essentiels, l'homogénéité afin de produire des résultats sur un phénomène et un seul, mais aussi l'hétérogénéité, pour assurer une clef de comparaison permettant d'objectiver les résultats car tenant compte d'une diversité que représente les différents sous-corpus étudiés. Il est présenté ici, une approche permettant d'allier les méthodes quantitatives aux analyses qualitatives menées notamment en analyse du discours.

Mots-clés : Linguistique de corpus, lexicométrie, Analyse de discours, statistique lexicale.

Introduction

Nous nous intéressons parmi les méthodes de statistique textuelle à un ensemble de méthodes relevant d'un champ d'analyse diversement nommé : statistique linguistique, statistique lexicale, linguistique quantitative, ou encore, terme relativement récent mais maintenant bien implanté, lexicométrie. On identifie actuellement plusieurs logiciels susceptibles de traiter un corpus selon une optique lexicométrique, entre autres Alceste commercialisé par la société Image et conçu par Reinert, Hyperbase, Spad T développé par Brunet, il est le plus ancien de ces outils et enfin, Lexico 5 précédé des versions 1 2, et 3 il est développé par Salem à l'Université de Paris 3 et est très largement utilisé par le centre de St cloud. Dans cette présentation nous nous limitons à la présentation d'illustration et d'exemples issus de l'utilisation de Lexico 3.6. Nous avons proposé de l'utiliser avec un corpus que nous avons constitué autour de 22 discours (soit un total de 37605 occurrences) du président actuel du Sénégal Macky Sall. Il ne sera pas question de proposer une analyse et une interprétation linguistique des observations, cette tâche est réservée à une exploitation ultérieure lors d'une publication à partir du corpus et en utilisant un outil de lexicométrie comme le montre cette fiche technique.

La linguistique de corpus

L'approche, dont nous parlons, relève de ce qu'on appelle la linguistique de corpus. Il s'agit là d'une discipline récente puisqu'elle date des années 80 mais elle a vraiment pris son essor avec la démocratisation des ordinateurs personnels tout en restant un domaine peu connu surtout dans le monde des lettres où elle fait peur. Toutefois, la nouvelle dynamique des humanités numériques, tend à donner une place plus importante aux travaux et recherches utilisant les outils informatiques. L'outil informatique était, jusque-là, considéré comme purement quantitatif et donc pas assez noble notamment pour des recherches de littérature et civilisation. La linguistique de corpus s'intéresse à la langue en contexte sous la forme de grands ensembles de textes, les corpus. C'est une discipline qui est très liée à l'informatique mais qui reste une discipline des sciences humaines et non de l'informatique. D'un point de vue linguistique, elle ne cherche pas nécessairement les formalismes mais plutôt à révéler les choix linguistiques opérés par les locuteurs dans des contextes réels. Elle relève de la linguistique appliquée, elle cherche à comprendre les mécanismes de communication et éventuellement à apporter des solutions à des questions pratiques. La linguistique de corpus s'est fait une place dans l'enseignement des langues, la lexicographie, la traduction et plus récemment dans la terminologie. La linguistique de corpus n'est pas seulement une méthodologie ; elle s'est révélée comme une discipline en soi avec ses propres présupposés. Deux approches fondamentales se côtoient, l'une déductive, corpus-based en anglais, qui utilise le corpus pour confirmer ou infirmer une hypothèse, et une linguistique inductive, corpus-driven, qui cherche à explorer les données a priori. Les deux sont complémentaires, les deux ont besoin de corpus électroniques.

Du point de vue du matériau, le dit corpus est à distinguer naturellement de l'archive, la linguistique de corpus porte sur des textes réels produits en situation de communication. Les productions artificielles produites par l'introspection du linguiste n'ont pas de place ici. On ne travaille que sur des unités réellement produites qui dépassent la phrase, l'exemplier du syntacticien par exemple est automatiquement exclu. Et, plus que le texte, le matériau d'analyse se constitue d'ensembles de textes choisis soigneusement et ordonnées selon des critères précis. Pour Rastier (2005), pour passer de l'archive au corpus, il faut distinguer 4 niveaux :

- L'archive contient l'ensemble des documents accessibles. Elle n'est par un corpus, parce qu'elle n'est pas constituée pour une recherche déterminée.
- Le corpus de référence est constitué par ensemble de textes sur lequel on va contraster les corpus d'étude.
- Le corpus d'étude est délimité par les besoins de l'application.

- Enfin le sous-corpus de travail en cours varie selon les phases de l'étude et peut ne contenir que des passages pertinents du texte ou des textes étudiés.

Linguistique et méthodes statistiques

L'utilisation des méthodes statistiques dans le domaine des données textuelles a été très tardive. Alors que ces méthodes trouvaient des applications dans d'autres disciplines, la linguistique ne se tournait que très peu vers ses potentialités de sorte qu'en 1950 M. Cohen (Cité par Lebart et Salem (1994, p.16) fait le constat suivant : « Il me semble pouvoir affirmer que ce serait entraver le développement de la linguistique que de continuer à tant se désintéresser des nombres quand nous parlons des phénomènes linguistiques ».
La pratique qui consiste à mesurer des données lexicales est certes ancienne mais ce dont souffrait la linguistique, c'est de l'absence d'une méthodologie propre. Or les nouveaux moyens informatiques permettent tant aux chercheurs qu'aux professionnels d'accumuler un stock d'informations qui ne peut plus rester inexploité. Et ce surtout si on considère avec de nombreux linguistes exploitant le traitement quantitatif que ces outils constituent une assistance très efficace à l'analyse des données et garantissent une systématicité d'analyse. Ils permettent d'objectiver les intuitions que l'on peut avoir vis-à-vis d'un corpus, tout en mettant en évidence certains aspects qu'une analyse manuelle ne relèverait sans doute pas. C'est une aide à la lecture et à l'analyse des textes utile pour différents types d'analyses et dans différents domaines d'analyse.
C'est donc à une période assez tardive, première moitié du XXème siècle que les travaux linguistiques exploitant les méthodes statistiques vont se développer. On compte parmi les premiers, les études menées par Brunet sur le Trésor de la langue française et les recherches menées par le laboratoire « Lexicométrie et textes politiques » de l'École normale supérieure de Fonthenay-St-Cloud.

Une perspective nouvelle pour la lexicométrie

La lexicométrie proclame, à l'encontre des méthodes quantitatives passées développées par les analyses textuelles traditionnelles ou la lexicologie, son statut scientifique. Elle rejette tout apriorisme et s'inscrit, refusant de privilégier quelque que données que ce soit dans son traitement, dans un souci d'exhaustivité des données étudiées. De façon générale, on peut dire qu'il existe dans la pratique deux méthodes de traitement quantitatif. D'une part, selon une terminologie empruntée à Salem, la lexicomancie, d'autre part la lexicométrie. Leurs propriétés sont successivement définies par Bonnafous (1991) **:**
« Une étude de vocabulaire peut relever de la lexicologie aussi bien que de la lexicométrie. La lexicologie s'appuie sur des relevés manuels de mots choisis lors de la lecture de textes souvent assez hétérogènes et s'inscrit dans une perspective de comparaison et de datation des emplois. Travail d'historien de la langue effectué sur un corpus ouvert, toujours enrichissable. La lexicométrie, en revanche, se donne comme objectif d'analyser de façon objective et systématique le vocabulaire de corpus clos constitué autour de variables et d'invariants déterminés. L'analyse est automatisée et porte sur des critères quantifiés ».

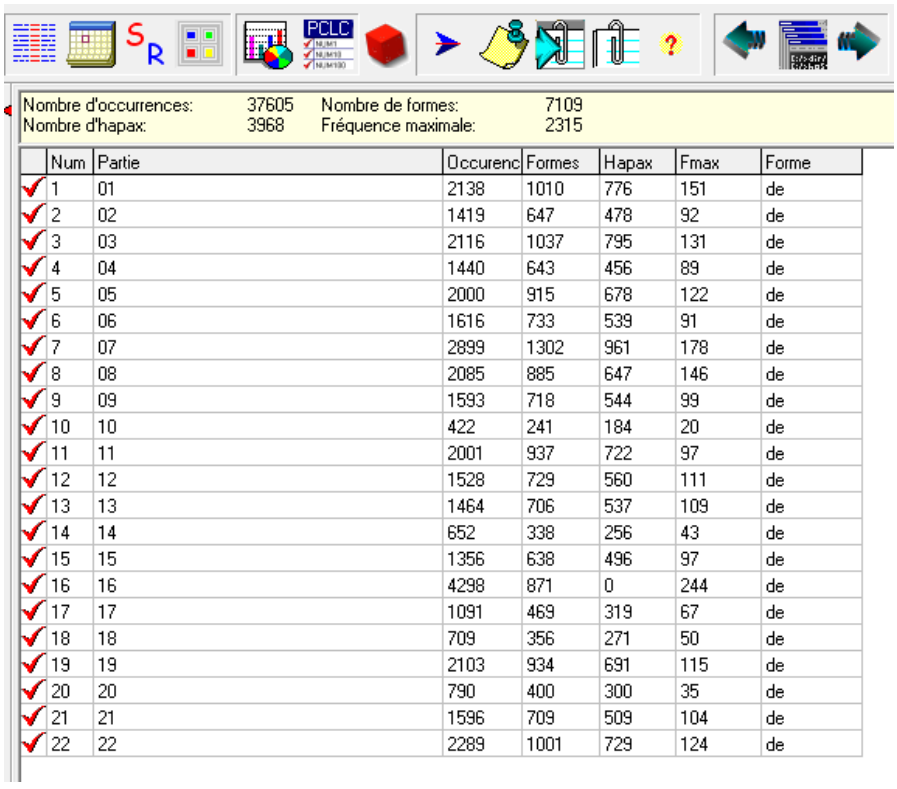
Segmentation de la matière textuelle

Quelles unités choisir ?
Par définition, tout logiciel d'analyse textuelle implique des comptages effectués sur les textes, autrement dit tout logiciel de ce type implique une segmentation de la matière textuelle. Il revient donc au concepteur de logiciel de définir cette unité textuelle et naturellement la viabilité de l'analyse repose sur l'invariabilité de cette unité. C'est-à-dire que pour être étudiable statistiquement les unités qui segmentent la matière textuelle ne doivent jamais changer au cours de la recherche. Or, un texte est composé de formes qui

peuvent s'envisager sous plusieurs angles. On peut isoler des unités lexicales, des unités sémantiques, des unités morpho-syntaxiques. La lexicométrie repose sur l'unité lexicale.

Les unités lexicales

L'unité d'analyse des méthodes lexicométriques est donc l'unité lexicale conçue comme unité minimale et indivisible. La reconnaissance de cette unité relève du travail de segmentation du texte : le texte est découpé en formes graphiques définies comme une suite de caractères, lettres, chiffres, symboles, délimités par deux blancs et la ponctuation. C'est donc l'écriture qui fournit une première définition de l'unité graphique. Les logiciels qui reposent sur une analyse en unités lexicales intègrent un ensemble de caractères appelés caractères délimiteurs opposés à caractères non-délimiteurs qui leur permettent de segmenter le texte et d'en identifier les formes et les occurrences : toute séquence définie par des caractères délimiteurs est une occurrence et deux suites de caractères identiques sont deux occurrences d'une même forme. L'ensemble des formes d'un texte définit son vocabulaire et l'ensemble de ses occurrences sa taille. A l'issue de la segmentation, un texte contient donc toujours moins de formes que d'occurrences. Les délimiteurs intégrés au logiciel lexico 3 et qui fonctionnent donc par défaut sont les suivants : .,:;!?/_-\'"()[]{}§$. Il est toujours possible pour l'utilisateur de les modifier. C'est sur la base de ces indices qu'est découpé le texte, à l'issue du traitement nous sommes informés des principales caractéristiques de ce texte



Nombre d'occurrences: 37605 Nombre de formes: 7109
Nombre d'hapax: 3968 Fréquence maximale: 2315

Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
1	01	2138	1010	776	151	de
2	02	1419	647	478	92	de
3	03	2116	1037	795	131	de
4	04	1440	643	456	89	de
5	05	2000	915	678	122	de
6	06	1616	733	539	91	de
7	07	2899	1302	961	178	de
8	08	2085	885	647	146	de
9	09	1593	718	544	99	de
10	10	422	241	184	20	de
11	11	2001	937	722	97	de
12	12	1528	729	560	111	de
13	13	1464	706	537	109	de
14	14	652	338	256	43	de
15	15	1356	638	496	97	de
16	16	4298	871	0	244	de
17	17	1091	469	319	67	de
18	18	709	356	271	50	de
19	19	2103	934	691	115	de
20	20	790	400	300	35	de
21	21	1596	709	509	104	de
22	22	2289	1001	729	124	de

Par rapport à ce travail de segmentation un certain nombre d'interrogations peuvent apparaître : un tiret peut renvoyer au signe moins, à un mot composé, à la marque d'une parenthèse ; l'apostrophe peut signifier la disparition d'un « e » muet ou d'une autre voyelle, elle peut aussi ne pas fonctionner comme délimiteur : « aujourd'hui ». Ces aspects, lexico et les autres y apportent des réponses dans le traitement notamment par l'identification des segments répétés et la construction de concordances de formes.

La lemmatisation

La définition des unités soumises à l'analyse ouvre un débat entre les logiciels exploitant des lemmatiseurs et ceux n'en exploitant pas. Ce débat oppose les partisans d'une analyse qui, refusant tout apriorisme, colle à la surface du texte et ceux qui appuient leur analyse sur une perspective plus lexicographique que lexicométrique et acceptent donc une intervention sur le matériau avant traitement. La lemmatisation consiste à ramener toutes les formes dérivationnelles et flexionnelles d'une unité à un représentant unique, à une forme canonique correspondant en général à l'entrée du dictionnaire.
Ainsi pour lemmatiser un texte on ramène en général :

- les formes verbales à l'infinitif
- les substantifs au singulier
- les adjectifs au masculin singulier
- les formes élidées à la forme sans élision
- les formes en majuscule à la forme sans

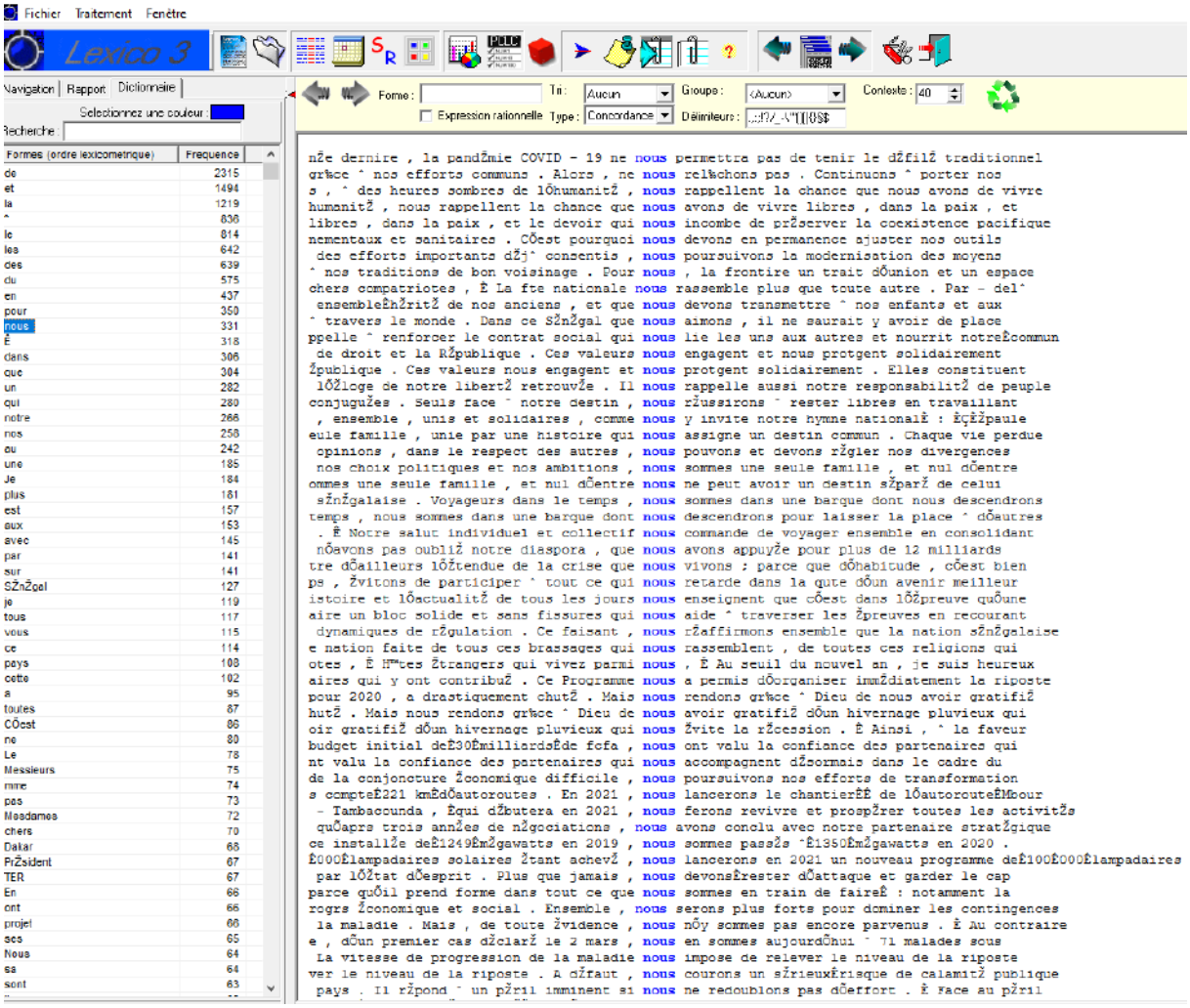
En général, ce traitement s'effectue sur la base de dictionnaires intégrés au logiciel, l'un des principaux problèmes que l'on rencontre est celui du traitement des formes ambiguës qui disparaît nécessairement dans l'entreprise de lemmatisation. Par ailleurs, dans l'étude de certains corpus, il peut être très précieux par exemple d'observer les différents temps utilisés.

Ex : un président qui semble résigné dans le traitement de certains problèmes de société et qui utilise très majoritairement le conditionnel présent : création d'un univers de rêve.

Le logiciel Alceste repose sur une lemmatisation du corpus, Lexico 3 en revanche n'y recourt pas mais il propose un outil de compensation à travers la possibilité de construire des groupes de formes. On peut par exemple sur requête de l'utilisateur regrouper les différentes formes de la personne 1 : Je, J, je, j, me, moi, mon, mes, ma, etc. Il est donc à noter que lexico 3 n'exclut pas la lemmatisation, toutefois, celle-ci repose en quelque sorte sur une lemmatisation non automatique, donc une lemmatisation manuelle.

Les segments répétés
Nous avons vu précédemment que notamment l'apostrophe et le tiret utilisés comme caractères délimiteurs mais également l'espace peuvent occasionner en particulier la segmentation des mots composés (marqués par un tiret mais également parfois sans) ainsi désolidarisés et non identifiés en tant que formes graphiques une. Afin de pallier cet inconvénient, les logiciels lexicométriques permettent l'identification, sur requête ou non de l'utilisateur selon le logiciel, de ce qu'on appelle les segments répétés. Ces unités complètent les formes graphiques, elles correspondent à des séquences de formes graphiques non séparées par un délimiteur de séquences (la ponctuation) qui apparaissent plus d'une fois dans le corpus. Un inventaire des segments répétés permet d'isoler des mots composés mais également des segments figés plus ou moins stables dont le sens n'est pas réductible à la somme du sens des éléments constitutifs. Il permet également de lever en partie les problèmes de polysémie mais à cet effet on aura également et plutôt recours aux concordances.

Les concordances
Les concordances permettent de rassembler toutes les occurrences d'une forme dite forme-pôle en l'accompagnant de son contexte de gauche et de droite. Il revient à l'utilisateur de définir la longueur de ces contextes. Pour connaître le contexte des formes on peut se reporter à leur adresse, travail particulièrement fastidieux surtout pour le non initié. Les concordances permettent de faire la même opération en quelques secondes et fournissent une vision synthétique des formes étudiées.
Par exemple : corpus du président Macky Sall, les contextes du pronom personnel « nous » désignant le locuteur (figure ci-dessous)



Une approche typologique

La phase de segmentation de la matière textuelle ne présente pas réellement d'originalité, de nombreux logiciels de traitement statistique reposent en effet sur cette opération. C'est dans les phases suivantes du traitement statistique qu'apparaît l'originalité des logiciels de lexicométrie.

Ceux-ci s'articulent à une approche typologique des corpus qui vise à établir des différences de profils entre textes soumis au traitement. La lexicométrie trouve son fondement sur des opérations de comparaison : le texte ou le corpus soumis à l'analyse est divisé en parties, ou sous-corpus, obtenues par des regroupements de textes élémentaires et ce sont les différences entre ces parties qui seront analysées. On est dans une perspective qui vise l'objectivité et l'on considère que seule la confrontation à d'autres valeurs de même type permet d'interpréter des valeurs obtenues.

Le corpus d'étude

La méthode lexicométrique ainsi définie montre qu'une étape primordiale à toute analyse est la constitution du corpus celui-ci devant se caractériser par plus d'homogénéité que d'hétérogénéité. Cette dernière est néanmoins absolument indispensable dans le cadre de méthodes lexicométriques, elle permet la comparaison. Afin de mener une analyse, on opère sur un corpus dit clôt au moins pour le moment de l'analyse puisque ne sont comparables que des unités stabilisées et partitionnées en parties qui forment la base de la comparaison. La définition de la partition relève d'une variable propre au corpus et en général sélectionnée selon les objectifs de l'analyse. Elle correspond en général à une des données de la situation d'énonciation : temps, espace, locuteur, destinataire, canal, etc. On peut ainsi dans une

perspective diachronique partitionné le corpus selon la date de production des textes (ex : textes publicitaires), ou encore en synchronie selon l'instance à l'origine de la production (exemple textes patronaux). Le corpus est divisé en N parties au regard desquelles chaque forme est étudiée pour définir sa fréquence sur l'ensemble du corpus et ses sous-fréquences sur chaque partie. L'ensemble des sous-fréquences d'une forme est appelé ventilation des occurrences de cette forme. Ainsi, la somme des sous-fréquences correspond à la fréquence totale de la forme sur l'ensemble du corpus. Une forme appartenant au vocabulaire de chaque partie étudiée est dite forme commune, en revanche si elle n'est occurrente que dans une seule des parties, elle dite originale.

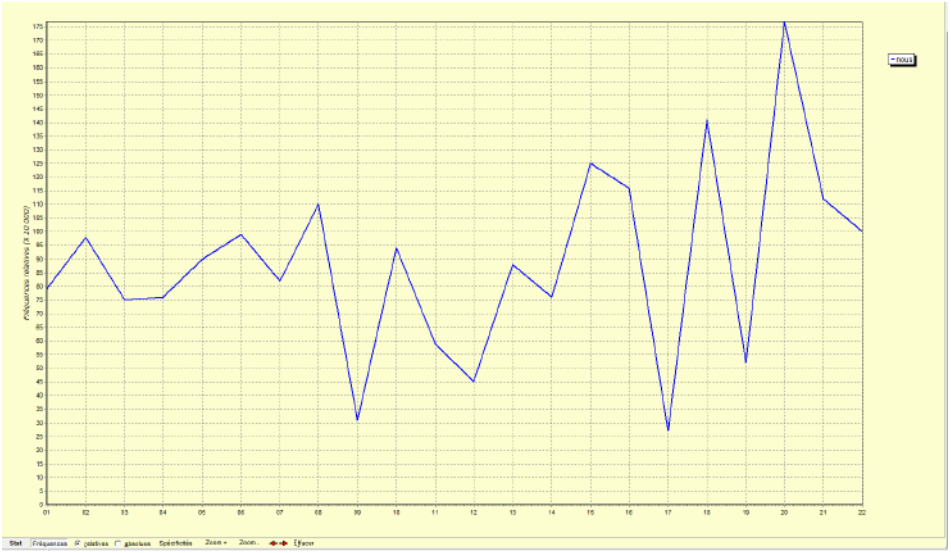
- Le vocabulaire est l'ensemble des formes attestées dans un corpus de textes
- Le vocabulaire commun est l'ensemble des formes attestées dans chacune des parties du corpus
- Le vocabulaire original est pour une partie du corpus, l'ensemble des formes originales, c'est-à-dire l'ensemble des formes qui trouvent leurs occurrences que dans cette partie du corpus.

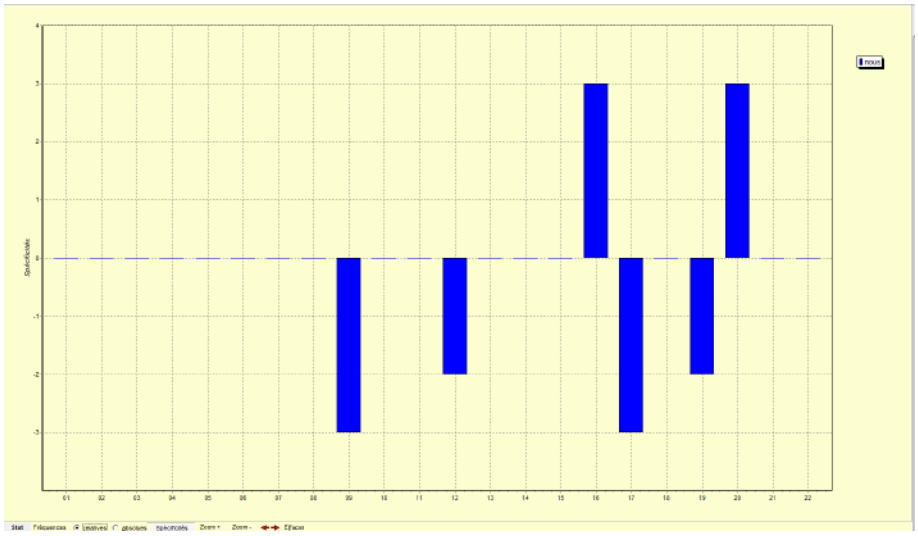
Les clés de partition

Concrètement, la partition du corpus, nécessairement enregistré sous le format txt, se réalise à l'aide de clés de partition introduites dans le matériau soumis à l'analyse. Ces clés doivent être présentées entre crochets au sein desquels on n'utilise pas d'espace et dans lesquels peuvent entrer des informations alphanumériques. L'information comprise entre ces crochets n'entre pas dans la segmentation et le décompte des unités : par exemple <Discours=01>, dans le cas de notre corpus support, la clef est <Macky=01>, <Macky=02>, <Macky=03>…<Macky=22> : chacun des 22 sous-corpus recueillis apparaît comme une base pour la comparaison et est une variable.

On peut introduire plusieurs clés dans le corpus et les clés peuvent se répéter, le traitement statistique les réunira alors.

Si l'étape de segmentation échoue, c'est bien souvent parce qu'il y a une erreur dans la formulation des clés. Le logiciel indique l'échec par un message qui invite l'utilisateur à consulter l'Atrace, document rangé dans le fichier d'origine du corpus qui indique notamment les erreurs dans la formulation des clés.

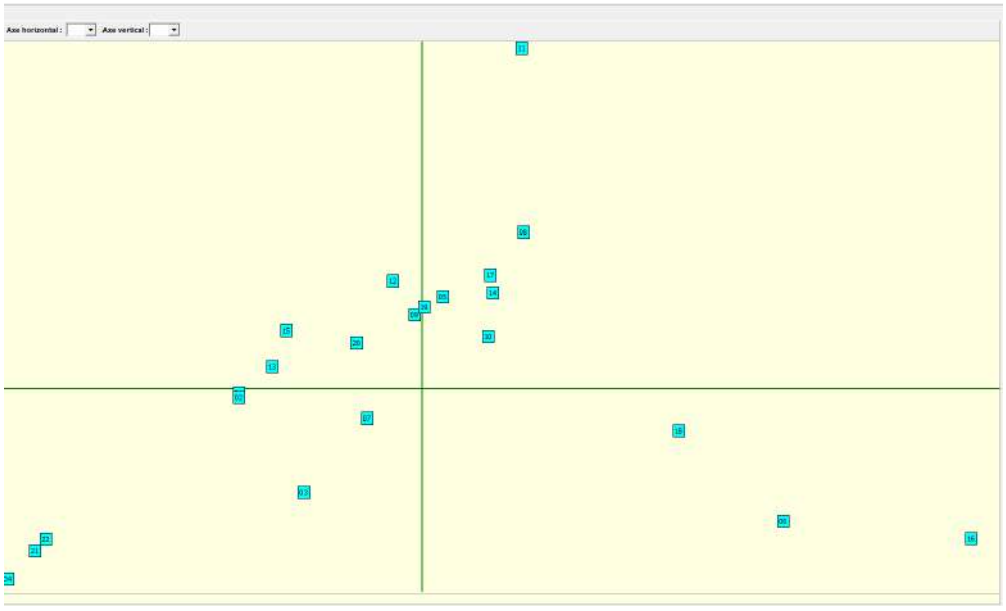




L'analyse factorielle des correspondances

A la base des traitements statistiques de données, on trouvait traditionnellement des calculs de pourcentage et des moyennes croisant plusieurs variables et construisant des tables de contingences ou des tableaux croisés. Appel au chi-2

L'analyse factorielle des correspondances repose sur ces principes statistiques de base, nous n'enterons pas dans les détails ici dans la mesure où ces principes sont automatisés dans les logiciels. L'objet de cette analyse est de comparer l'ensemble de données numériques contenues par exemple dans le tableau suivant pour en donner ce qui nous intéresse prioritairement une représentation géométrique qui permettra de décrire simultanément les similitudes de profils-lignes et de profils-colonnes :



On peut vouloir de la sorte croiser les valeurs de toutes les variables lignes et colonnes, ce qui n'est pas possible sur une telle visualisation. C'est ce que nous permet en revanche l'analyse factorielle des correspondances qui exploite également une représentation sur des axes orthogonaux auxquels néanmoins, on ne peut plus affecter une valeur particulière puisque ceci résultent du croisement de l'ensemble des valeurs, ils figurent un espace non plus à deux dimensions mais à n dimensions dans lequel ce qui est soumis à l'interprétation, c'est seulement des distances et des proximités de point.

L'AFC fournit une typologie des sous-parties du corpus, qui vise à rapprocher entre elles celles qui emploient les mêmes mots dans les mêmes proportions.

L'analyse factorielle des correspondances est une analyse globale, de l'ordre d'un panorama, qui doit nécessairement être complétée par une analyse des sous-fréquences, dite analyse des spécificités, qui repose sur des calculs probabilistes intégrés au logiciel. Pour apprécier la répartition d'une unité linguistique à l'intérieur d'un corpus, il est nécessaire d'établir des comparaisons sur la valeur de cette unité avec la valeur de l'ensemble des unités de même type contenues dans ce corpus. Mais cette comparaison ne peut être pertinente que si elle se réalise en prenant en compte la taille des textes étudiés, du corpus dans son entier et des sous-parties. Or espérer travailler sur des corpus réunissant des sous-parties de même taille est naturellement un idéal rarement atteint et on ne peut pas tout simplement couper une sous-partie pour l'atteindre. Ceci poserait la question, à laquelle il n'y a pas de réponse, de quoi couper dans les textes rassemblés pour atteindre la taille idéale.
Pour contourner cette difficulté, l'analyse des spécificités intègre des calculs probabilistes qui permettent pour chaque forme graphique de définir un seuil de probabilité attendu, un seuil type, par rapport auquel les seuils effectifs de réalisation sont jugés. Autrement dit, il s'agit de calculs de proportionnalité et de pondération qui définissent ce que serait une répartition uniforme de la forme sur l'ensemble des sous-corpus par rapport à leur taille. A cette répartition, est confrontée la fréquence réelle pour obtenir une information exploitable sur chacune des formes étudiées.
On peut ainsi obtenir 3 types d'information pour une forme :
Si la sous-fréquence d'une forme dans une sous-partie est par rapport au seuil de spécificité fixé anormalement élevée, on dit qu'elle est spécifique positive. Cette forme est en sur-représentation dans la partie, à nous d'interpréter ensuite cette spécificité positive. Elle est notée S+X, X indiquant l'indice de spécificité.
Si la sous-fréquence d'une forme dans une sous-partie est par rapport au seuil de spécificité fixé anormalement faible, on dit à l'inverse qu'elle est spécifique négative. Cette forme est alors en sous-représentation dans la partie, aussi importante pour l'analyste dans ce cas que dans le cas précédent. Elle est notée S-X, X indiquant l'indice de spécificité.
Si une forme pour une sous-partie donnée ne présente aucune spécificité, ni positive, ni négative, elle est dite banale. L'ensemble des formes banales pour chaque sous-partie du corpus, c'est-à-dire l'ensemble des formes ne présentant pour un seuil fixé aucune spécificité dans aucune des parties du corpus constitue le vocabulaire de base du corpus étudié.

Conclusion

C'est ce matériau qui constitue une clé de lecture nouvelle des textes, l'interprétation qui en découle est naturellement étroitement solidaire des hypothèses de départ et du corpus que l'on analyse. Ces spécificités permettent de construire des profils-types de chacune des sous-parties, on pourrait par exemple imaginer construire le profil du discours politique, ou encore du discours politique anti-système, ou celui panafricain etc.

Références bibliographiques

Bonnafous, Simone, 1991, *L'immigration prise aux mots,* Paris, Éditions Kimé.
Lebart, Ludovic et Salem André, 1994, *Statistique textuelle,* Paris, Dunod.
Rastier, François, 2005, « Pour une sémantique des textes théoriques », *Revue de sémantique et de pragmatique,* N°17, pp. 151 – 180.