Proximités segmentales

André Salem

EA2290 SYLED – CLA2T, Université de la Sorbonne nouvelle - Paris 3

Ilpga, 19 rue des Bernardins, 75005 Paris

Résumé

La littérature textométrique propose différentes manières de calculer des proximités relatives entre textes, à partir de la comparaison *globale* de stocks lexicaux. On propose ici de considérer également des proximités fondées sur l'emploi simultané par deux textes de séquences textuelles identiques. On analyse, du point de vue de ce type de répétition, les rapports entretenus par le texte d'un congrès syndical avec un classique du mouvement ouvrier qui date du 19e siècle (§2). On compare ensuite les résultats obtenus à l'aide des mêmes méthodes à partir d'autres textes syndicaux de l'époque récente (§3), puis avec des textes d'une même organisation produits sur un laps de temps plus long (§4).

Abstract

Various ways of computing a distance between texts, mainly based on comparison between global amount of vocabulary among them, can be found in textometric studies. In this paper, we consider measures of similarity, based on the computation of the frequencies of identical sequences of words among the texts to be compared. The method is firstly used to compare a text of a French worker trade-union with a classical text of the 19th century (§2). Then we compare various texts of different unions published in the 70ies (§3). Finally a comparison of texts produced by one single union during a larger period is provided (§4).

1. Introduction

Comment mesurer la distance entre deux textes ? Comment mesure-t-on la distance entre deux textes ? Au-delà des procédures de comparaison habituelles entre vocabulaires constitués de formes ou de lexèmes, la première de ces questions constitue une invitation à imaginer de nouvelles mesures de proximité entre textes. La seconde nous invite à une enquête parmi les productions qui se donnent pour objet la comparaison quantitative des textes.

2. Proximités textuelles

Lorsqu'on réunit en corpus un ensemble de textes, afin de les comparer, il est pratique de se donner une distance entre chaque paire de textes[1]. Le calcul d'une distance permet de rapprocher certains des textes rassemblés, d'en opposer d'autres, de dresser une typologie des éléments textuels rassemblés, d'en faire apparaître les dimensions de variation les plus importantes.

[1] Une distance d sur un ensemble I est une fonction à valeurs non négatives qui associe à chaque couple d'éléments i et i' un nombre d(i,i'). Elle satisfait en outre aux trois axiomes suivants :

- d(i,i) = 0, et pour i' différent de i on a d(i,i') > 0 (identité et positivité) ;
- d(i,i') = d(i',i) (symétrie) ;
- d(i,i') est inférieur ou égal à d(i,j) + d(j,i') (inégalité triangulaire).

Les opérations qui permettent de résumer ainsi, à l'aide d'un simple calcul numérique, les différences de tous ordres qui existent nécessairement entre deux textes sur les plans : lexical, syntaxique, sémantique, pragmatique, etc., constituent une simplification extrêmement grossière et forcément partiale de la diversité inhérente à toute collection de textes. Cependant, la pratique de l'analyse textométrique montre que ce type de calcul permet souvent d'organiser les parties du corpus en sous-ensembles qui révèlent a posteriori une certaine cohérence, de rapprocher des textes qui relèvent d'un même genre, des textes dus au même auteur, des textes écrits à des périodes proches dans le temps, pour peu que le corpus de départ permette de mettre en évidence de telles oppositions.

2.1. Les distances lexicales

L'examen des différents indices proposés dans la littérature textométrique pour mesurer la distance lexicale[2], ou la proximité, entre textes fait apparaître que ces indices obéissent à deux constantes :

- ils s'appuient tous sur des décomptes de formes isolées de leur contexte immédiat ;
- ils opèrent des comparaisons entre l'ensemble des stocks d'unités lexicales attestées dans chacun des textes.

L'analyse des procédures employées pour construire les différentes distances entre éléments d'un ensemble de textes permet de mettre en évidence les principes, souvent implicites, qui ont présidé au choix de telle ou telle formule de calcul. Ainsi, par exemple, la distance du chi-deux[3], sur laquelle repose, entre autres méthodes, l'analyse factorielle des correspondances, tend à rapprocher les textes dont les *stocks distributionnels*[4] sont proportionnels. L'indice de Jaccard[5] rapproche pour sa part les textes dont les vocabulaires montrent une forte intersection, ce qui constitue une toute autre manière de construire le concept de proximité. Notons que la pratique de l'analyse textométrique montre que les typologies appuyées sur différentes méthodes ne se révèlent pas toujours aussi opposées que les différentes approches pouvaient le faire penser au départ.

Ces caractéristiques communes aux indices de distance utilisés dans la recherche textométrique, souvent nées dans le cadre des études stylistiques sur le vocabulaire des *grands auteurs*, portent la marque d'une conception déjà ancienne de la genèse du discours ; celle de la sélection du vocabulaire par un scripteur, puisant à l'occasion du processus d'écriture des formes lexicales dans un lexique qu'il partage avec ses contemporains.

2.2. Intertextualité et coïncidence segmentales

Dans la période récente, l'étude des grands corpus de textes politiques rassemblés autour de périodes temporelles relativement étendues attire au contraire l'attention des chercheurs sur la circulation, à l'intérieur des communautés produisant des textes dans le domaine sociopolitique, d'unités textuelles plus étendues que la forme ou le lexème. Dans ce cadre,

[2] Il est à noter que certains travaux tentent d'évaluer la distance intertextuelle à partir de la distribution des catégories grammaticales, en prenant en compte à la fois leur fréquence et leur séquentialité ; voir, par exemple Longrée & Mellet (2004).

[3] Dans les applications aux dépouillements textuels, la distance du chi-deux rapproche les textes dont les profils d'utilisation du vocabulaire sont proportionnels. Cf., par exemple, Lebart & Salem (1994).

[4] Le *stock distributionnel* d'un de texte est constitué par la liste des formes de son vocabulaire munie de la fréquence de chacune des formes.

[5] Pour deux textes, dont on notera les vocabulaires par A et B, l'indice de Jaccard rapporte l'intersection des vocabulaires à leur union. J(A, B) = AB/(A+B).

l'attestation simultanée dans deux textes de séquences identiques dont la taille dépasse la forme induit le sentiment que les textes se réfèrent à des concepts communs, qu'ils ont peut-être été produits dans des conditions proches, sinon par des *formations discursives* proches, même si leurs stocks lexicaux, pris dans leur totalité, ne présentent pas de similitudes particulières[6].

Dans ce qui suit, nous illustrerons l'approche proposée à travers l'analyse des rapports de répétition segmentale entretenus par le texte d'une résolution syndicale avec un des textes qui ont posé des concepts fondamentaux en matière de lutte politique il y a près de deux siècles. Nous comparerons ensuite les résultats obtenus avec ceux que l'on obtient par les mêmes méthodes à partir d'autres textes syndicaux de l'époque récente (§3), puis avec des textes de la même organisation produits à des époques différentes (§4).

3. Le congrès CFDT (1973) et le Manifeste

Les textes des résolutions adoptées au cours du congrès de 1973 de la CFTD, désormais ***DT73***, emploient de manière remarquable des séquences textuelles comme : *lutte de classe, classe dominante, moyens de production et d'échange, rapports de production, etc.*, couramment utilisées dans les textes contemporains des partis politiques qui se réclament d'une analyse marxiste de la société, mais dont l'utilisation est nettement moins courante dans l'expression syndicale de l'époque concernée[7].

> § pour la cfdt la conquête du pouvoir politique et économique est une condition nécessaire, mais elle est aussi insuffisante. La construction du socialisme démocratique et autogestionnaire exigera des transformations radicales visant à la propriété sociale des principaux moyens de production et d'échange ; la transformation des rapports de production ; le développement économique fondé sur la satisfaction des besoins ; /…/
>
> *Extrait de la résolution générale du congrès CFDT de 1973*

Plusieurs des concepts convoqués dans cet extrait ont une histoire qui prend sa source dans celle des affrontements politiques en France et en Europe au cours des deux derniers siècles. Parmi les textes qui ont été produits au cours de cette histoire, l'un des plus importants est sans doute le *Manifeste du Parti communiste*, désormais ***Manifeste***[8]**,** publié en 1848 par K. Marx et F. Engels qui fera, par la suite, l'objet de très nombreuses éditions. Les concepts mis à l'œuvre dans le *Manifeste* n'ont pas tous été élaborés à l'occasion de cette publication. Cependant, leur organisation d'ensemble, le retentissement et la diffusion de l'ouvrage, les

[6] Les problèmes de l'intertextualité ont été largement étudiés par des chercheurs appartenant à différentes communautés scientifiques. Parmi les textes fondateurs, on consultera par exemple Foucault (1971), Bakhtine (1984). On trouvera quelques réflexions relatives à l'approche automatisée des corpus textuels dans Pêcheux (1969).

[7] Dans le contexte sociopolitique de la France des années 70, une division des rôles s'est établie entre les organisations politiques et les syndicats. Les partis politiques critiquent l'ordre économique existant, organisent la contestation politique du pouvoir en place et prônent l'instauration de modes de fonctionnement sociaux en rupture avec les structures existantes. Les organisations syndicales affichent, de leur côté, la volonté de rassembler l'ensemble des (*prolétaires/ouvriers/ travailleurs/ salariés*) pour l'amélioration de leurs *conditions de vie, de travail* etc., n'intervenant que de manière indirecte dans les luttes politiques. Cette prise de position de la CFDT (Confédération Française Démocratique du Travail) sur des questions dépassant le cadre traditionnel de la lutte syndicale sera remise en cause par la centrale syndicale lors de ses congrès ultérieurs, aboutissant à la formulation de la nécessité d'un *recentrage* de la centrale sur les questions plus proprement syndicales.

[8] Rédigé par Karl Marx et Frédéric Engels, le *Manifeste du parti communiste* a d'abord été publié en allemand en 1847, une première édition en français parut à Paris dès le début de l'année 1848. Nous avons utilisé pour cette expérience le texte établi par l'Association des Bibliophiles Universels (ABU), sur le site : http://abu.cnam.fr/.

ont constitués en unités saillantes dans la mémoire collective de nombreuses générations de militants et, par conséquence, comme des objets discursifs importants pour les chercheurs confrontés à l'étude des textes sociopolitiques.
Peut-on objectiver le sentiment que, malgré les années qui les séparent, malgré le genre textuel différent, malgré le fait qu'ils ont été écrits au départ dans des langues différentes, ces textes présentent une forme de *parenté* repérable par l'emploi récurent de séquences textuelles identiques ? Est-il possible de mesurer l'importance de ce phénomène de *résonance* entre les deux textes ? Pour tenter d'obtenir quelques éléments de réponse à ces questions, nous allons réunir en un même corpus expérimental ***Manif&DT73*** les textes que nous souhaitons comparer**.** Le dépouillement quantitatif de ce corpus montre que les textes mis en présence présentent des caractéristiques textométriques comparables, notamment en ce qui concerne leur taille.

Partie	*occurrences*	*formes*	*hapax*	*Fréq. Max*	
DT73	12636	2344	1300	775	de
Manifeste	11992	2444	1530	688	de

Tableau 1 : *Principales caractéristiques lexicométriques du corpus* ***Manif&DT73***

L'analyse des spécificités[9] appliquée à chacune des deux parties met en évidence des sous-ensembles de vocabulaire pratiquement exclusifs pour chacun des deux textes, ce qui ne constitue pas une surprise s'agissant de textes produits à des périodes aussi éloignées dans le temps. Parmi les formes fréquentes, le texte du ***Manifeste*** est le seul à employer les formes : *bourgeoisie, bourgeois, bourgeoise, prolétariat, communistes, ouvriers.* De manière symétrique, la résolution ***DT73*** emploie de manière presque exclusive les formes : *travailleurs, cfdt, syndicale, emploi,* etc. Cette répartition contrastée concerne également un grand nombre de formes de faible et de moyenne fréquence qui trouvent toutes leurs occurrences dans l'un ou l'autre des deux textes.

Manifeste				***DT73***			
forme	**Total**	**partie**	**spécificité**	**forme**	**total**	**partie**	**spécificité**
bourgeoisie	93	93	30	*travailleurs*	126	124	34
la bourgeoisie	82	82	27	*la cfdt*	74	74	22
bourgeois	62	62	20	*les travailleurs*	52	52	16
prolétariat	59	59	19	*action*	94	84	16
bourgeoise	43	43	14	*des travailleurs*	53	51	13
propriété	62	57	13	*l information*	35	35	11
n	85	70	11	*doit*	48	43	9
ils	77	65	11	*des*	561	354	9
plus	137	101	10	*l action*	44	40	9
communistes	29	29	10	*d action*	30	29	8
classes	49	44	10	*de la cfdt*	25	25	8
production	66	55	9	*organisation*	50	43	8
ouvriers	27	27	9	*l organisation*	30	28	7

Tableau 2 :

Spécificités positives majeures pour chacun des deux volets du corpus ***Manif&DT73***

[9] Pour chaque partie d'un corpus, l'analyse des spécificités permet de dégager : a) des ensembles de formes et de segments répétés dont la fréquence est relativement élevée dans la partie considérée, par rapport aux autres parties du corpus ; b) des formes et segments dont la fréquence est au contraire particulièrement faible. Cf., par exemple, Lafon (1981) et Lebart & Salem (1994).

Ces circonstances suffisent à laisser prévoir que les indices classiques de connexion lexicale mentionnés plus haut diagnostiqueront une faible proximité entre les deux textes, comme nous pourrons le vérifier au paragraphe 3.

Pour tenter de repérer les séquences communes aux deux textes, nous allons recenser tous les segments composés de 4 formes au moins et répétés dans le corpus ***Manif&DT73.*** Nous éliminerons ensuite les segments qui trouvent toutes leurs occurrences dans un seul des volets du corpus (par ex : *de la société féodale, de la vieille société, la communauté des femmes* pour le ***Manifeste*** et *tous les aspects de, le socialisme démocratique et autogestionnaire* pour la CFDT73)[10] .

segments répétés	*DT73*	*Manifeste*
de la classe ouvrière	9	9
moyens de production et d'échange	5	6
de plus en plus	9	2
de la classe dominante	6	2
de tous les pays	5	2
la lutte des classes	4	2
le développement de l'	3	2

Tableau 3 :
Les segments répétés de longueur et de fréquence supérieures à 4 attestés dans chacun des deux volets du corpus ***Manif&DT73***[11]

Les séquences présentées au tableau 3 ont été retenues sur le double critère de leur longueur et de leur présence dans chacun des volets du corpus. Certaines d'entre elles correspondent à des constructions syntaxiques, ou des fragments de construction syntaxiques, courants dans les textes rédigés en français (*de plus en plus, etc.)*, d'autres séquences (*moyens de production et d'échange, la lutte des classes, etc.)* sont constituées de groupes nominaux qui concernent plus directement le domaine de l'analyse des rapports politiques et économiques et le rôle des acteurs sociaux dans ces rapports.

Notons que dans la liste des séquences de longueur plus faible qui présentent les mêmes caractéristiques de répartition, on relève également, au milieu des séquences les plus courantes du français (*de la, et des, et le,* etc..), des associations qui semblent relever de ce second registre (*action politique, conditions sociales, organisation sociale, pouvoir politique, propriété privée, mouvement ouvrier, rapports sociaux, travail accumulé, entre les peuples, sociale et politique, rapports de production*). Le problème est que la différence de nature entre ces deux catégories de segments est extrêmement difficile à formaliser de manière simple et satisfaisante. En nous bornant au recensement des segments répétés composés d'au moins 4 formes, nous sélectionnons, par des moyens entièrement automatisés et dont l'énoncé est particulièrement simple, des séquences susceptibles de correspondre à des constructions syntaxiques un peu complexes éliminant du même coup des segments très fréquents et beaucoup moins intéressants dans la mesure où ils sont présents dans la plupart des textes rédigés en français.

[10] Les segments répétés du corpus calculés par le programme Lexico3 ont réordonnés et sélectionnés après importation dans un tableur afin de produire les sélections présentées ici.
[11] On a retenu pour cette sélection les segments les plus longs en éliminant à la fois les plus redondants (ex : *de la lutte des classes*) et ceux qui renvoient à des expressions de la langue courante (ex : *au fur et à mesure).*

4. Analyse synchronique : textes syndicaux des années 70

Au paragraphe précédent, nous avons mis en évidence l'existence de séquences textuelles dont la taille dépasse la forme isolée, attestées à la fois dans ***Manifeste*** et dans le congrès ***CFDT73.*** Pour apprécier ces résultats, il nous faut les replonger dans un ensemble de comparaisons avec d'autres textes produits dans le même type de circonstance et à la même époque.

Trois autres textes nous serviront pour cette comparaison. Ils correspondent aux congrès de centrales syndicales concurrentes survenus à peu près à la même époque :

- CGT72 : congrès de la Confédération Générale du Travail en 1972
- CFTC73 : congrès de la Confédération Française des Travailleurs Chrétiens en 1973
- FO74 : congrès de la Confédération Générale du Travail – Force Ouvrière en 1974

Dans la mesure où ces textes sont de longueur très différentes, nous les avons ramenés, pour l'expérience qui suit, à la taille des deux premiers, soit environ douze mille occurrences chacun[12]. Nous appellerons le corpus ainsi constitué ***Manif&Synd70***.

Partie	*occurrences*	*formes*	*Hapax*	*fréq. max*	
Manifeste	11 992	2 444	1 530	688	de
DT73	12 636	2 344	1 300	775	de
TC73F	10 609	2 250	1 226	709	de
GT72F	11 717	2 304	1 254	671	de
FO74F	12 221	2 733	1 604	744	de

Tableau 4 : *Principales caractéristiques lexicométriques du corpus* ***Manif&Synd70***

4.1. Distances lexicales

À partir de la segmentation en formes graphiques des occurrences du corpus, nous commencerons par calculer les indices de distance lexicale les plus classiques :

- l'indice de Jaccard, qui est un indice de proximité ;
- la distance du chi-deux entre les stocks distributionnels de ces textes.

Comme on le voit sur ces tableaux, quel que soit le type de mesure *lexicale* utilisée à partir des recensements du stock distributionnel des formes graphiques isolées de leur contexte, la partie ***Manifeste*** ne présente d'affinité particulière avec aucune des autres parties du corpus.

Guide de lecture pour le tableau 5

- la colonne de gauche du tableau 5 donne les valeurs de l'indice de Jaccard calculé pour chaque couple de parties *(i, j)* du corpus ;
- la colonne de droite donne la distance du chi2 entre chaque couple de textes ;
- dans le bas du tableau, on trouve le premier plan d'une analyse factorielle des correspondances réalisée à partir du tableau (formes x parties), c'est une projection sur ce plan de points dont les distances mutuelles figurent dans la colonne de droite.

[12] Les fragments: TC73F, GT72F et FO74F dont le nom est affecté de la lettre finale *F* (comme fragment) ont été réduits en prélevant les premières occurrences du texte.

Indice de Jaccard *(x 10**2)*					***Distance du chi2*** *(x 10**6)*				
1 M:					***1 M***: :				
2 DT73:	18				***2*** DT73:	37			
3 :TC73F	16	29			***3*** TC73F :	44	29		
4 :GT72F	18	29	27		***4*** GT72F:	38	25	32	
5 :FO74F	17	28	31	27	***5*** FO74F :	41	27	27	29
	1	***2***	***3***	***4***		***1***	***2***	***3***	***4***

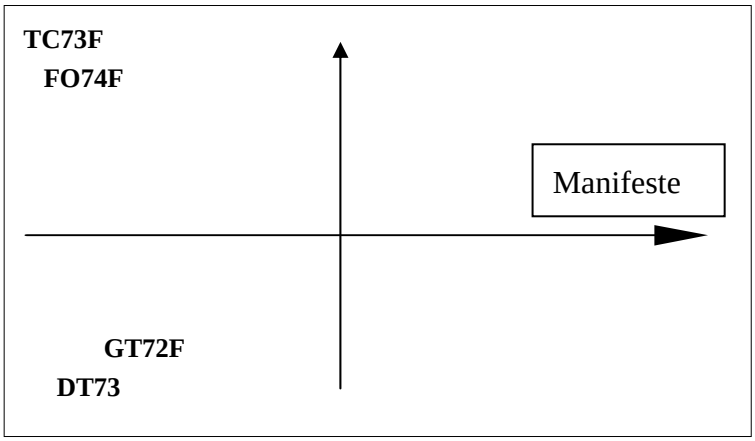


Tableau 5 : *Indice de Jaccard et distance du chi2 pour les parties du corpus* ***Manif&Synd70***

4.2. Un point de vue segmental

Une analyse de ces mêmes données appuyée cette fois sur le repérage des séquences de formes répétées dans le corpus, ou *segments répétés* du corpus, nous conduit, comme nous allons le voir, à des conclusions très différentes.

segment	***Manifeste***	***DT73***	***TC73***	***GT72***	***FO74***
de plus en plus	9	2	2	6	5
de la classe ouvrière	9	9	0	1	0
moyens de production et	5	6	0	0	0
le développement de l	3	2	0	4	2
de tous les pays	5	2	0	1	1
de la classe dominante	6	2	0	0	0
la lutte des classes	4	2	0	1	0
dans le cadre de	1	2	0	0	4
des moyens de production	1	4	1	0	0
le développement de la	2	2	2	0	0
de la grande industrie	4	0	0	1	0

Tableau 6 : *Quelques segments répétés du corpus* ***Manif&Synd70***
attestés dans la partie ***Manifeste*** *et dans un des textes syndicaux au moins*

Pour l'expérience dont nous rendons compte ci-dessous, nous avons repéré les segments composés de 4 formes au moins, attestés au moins une fois dans le texte du ***Manifeste*** et trouvant également une ou plusieurs occurrences dans les textes syndicaux qui constituent le reste du corpus. On trouvera au tableau 6 quelques exemples de segments de ce type choisis parmi les plus fréquents du corpus.

La figure 1 montre la répartition de la surface couverte par les occurrences qui composent les segments de ce type dans chacun des textes syndicaux du corpus. Pour mesurer cette surface, et pour tenir compte de la différence de taille qui subsiste entre les parties, on a utilisé une représentation en spécificités[13].

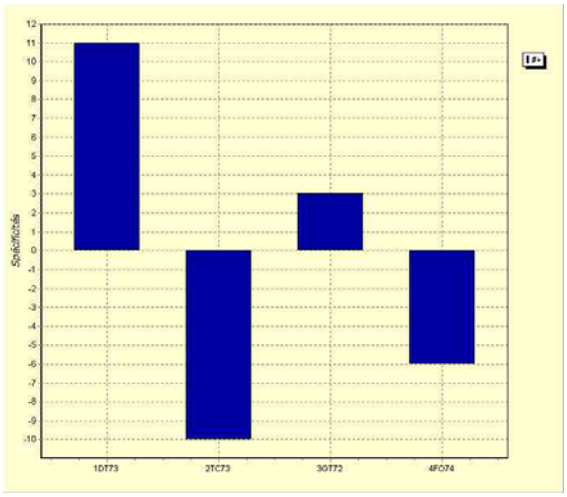


Figure 1 : *Surface couverte par les segments de longueur 4 présents dans le* ***Manifeste*** *pour 4 textes syndicaux (exprimée en spécificité)*

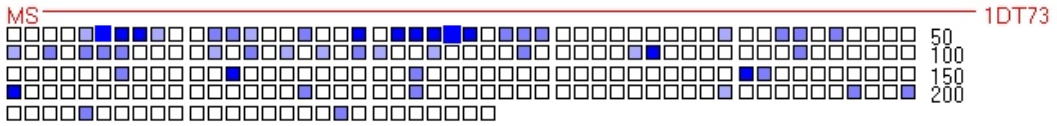
Comme on le voit sur cette figure, le calcul signale une abondance relative, dans la partie ***DT73***, de la surface couverte par des séquences de formes attestées dans le ***Manifeste***. La partie ***GT72*** montre également un suremploi de ces mêmes séquences, bien que d'ampleur nettement moindre. À l'inverse, les deux parties ***TC73*** et ***FO74*** ne contiennent ces mêmes séquences que dans une faible proportion.

Pour préciser ce diagnostic, on se reportera à la figure 2 qui décline en paragraphes le constat fait à propos de l'ensemble des textes de congrès ***DT73***. Le même calcul statistique permet de signaler les paragraphes dans lesquels les séquences présentes dans le *Manifeste* sont particulièrement représentées[14]. Les carrés de couleur sombre correspondent à des paragraphes particulièrement remarquables de ce point de vue. On trouve sur la figure 2 quelques exemples de paragraphes qui se rapportent au type mentionné plus haut pour lesquels la surface également attestée dans le texte du ***Manifeste*** a été détachée en caractères gras.

[13] Notons que, dans ce cas particulier, le comptage de la surface couverte par ces segments en « pourcentage de la surface totale » conduit sensiblement aux mêmes résultats.

[14] Par souci d'homogénéité, nous avons utilisé, ici encore, un calcul hypergéométrique qui apprécie la surface couverte par de telles séquences à la longueur du paragraphe.

L'examen de ces textes montre qu'ils recèlent en outre d'autres unités de longueur inférieure (*organisation sociale*, *appropriation*, etc.), écartées de nos résultats par le mode de sélection que nous avons adopté pour cette expérience, mais qui peuvent être aisément localisées dans le texte du ***Manifeste***.



***DT73* - section 6**

§ la logique du profit et de la puissance a progressivement gagné tous les aspects de la vie quotidienne. **ce n'est pas** seulement la situation de travail et la vie économique, mais **l'ensemble des rapports sociaux**, individuels et collectifs, qui subissent aujourdhui les effets et la marque du capitalisme.

***DT73* - section 25**

§ elle exige non seulement l'appropriation sociale **des moyens de production et d'échange** mais également une transformation de l'organisation sociale et du mode de vie. elles permettront ainsi à chacun de déterminer librement et d'acquérir les capacités créatrices pour l'**organisation de la société**.

Figure 2 : *Repérage des paragraphes spécifiques*

5. Analyse diachronique : la série CFDT (1973-2002)

Quelle est la portée du phénomène constaté à propos des textes du congrès de 1973 de la CFDT ? S'agit-il d'un moment particulier de la vie de cette organisation ou doit-on conclure que cette forme d'expression constitue son expression régulière ?

Pour répondre à cette dernière question, nous allons appliquer la procédure décrite plus haut à la *série textuelle chronologique*[15] constituée par l'ensemble des congrès de la centrale CFDT tenus entre 1973 et 2002.

On voit sur la figure 3 des comptages réalisés, toujours à partir des segments répétés composés de 4 formes attestés à la fois dans le ***Manifeste*** et dans au moins une résolution d'un congrès CFDT appartenant à cette période.

Le phénomène constaté à propos du congrès de 1973, reçoit un éclairage plus complet dans cette perspective diachronique. L'importance relative de la surface couverte par les segments de longueur 4, également attestés dans le *Manifeste*, est encore plus forte dans le texte du congrès de 1976 que dans celui de 1973. Cependant, dès le congrès de 1979, la plupart de ces séquences cesseront d'être utilisées pour disparaître pratiquement dans les derniers congrès.

[15] On appelle *séries textuelles chronologiques* des corpus de textes produits par une même source textuelle au cours du temps et répondant à des caractéristiques lexicométriques comparables cf. par exemple Salem (1993).

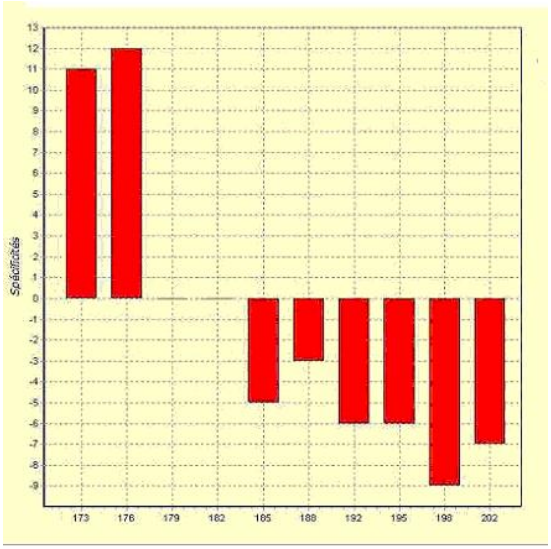


Figure 3 : *Surface couverte par les segments de longueur 4 présents dans le* ***Manifeste*** *pour la série chronologique* ***CFDT*** *(1973-2002)*

6. Conclusion

À partir de l'impression, qui aurait pu se révéler subjective, que certaines séquences textuelles employées dans une résolution syndicale (CFDT 1973) ressemblaient à des séquences également présentes dans le ***Manifeste***, nous avons dressé l'inventaire systématique de toutes les séquences de longueur 4 communes aux deux textes. Le fait de ne considérer que des séquences relativement longues nous a permis d'écarter dans un premier temps de très nombreuses séquences, plus courtes et plus fréquentes, inévitablement présentes dans tous les textes rédigés en français (*de la*, *et des*, etc..).

Par comparaison avec d'autres textes syndicaux, les distances et indices de proximité lexicaux calculés sur les formes isolées de leur contexte n'ont pas montré d'affinités particulières entre les deux textes sur le plan du vocabulaire. Par contre, les calculs effectués à partir du repérage des segments répétés montrent que ces textes partagent un nombre relativement élevé de séquences longues, par comparaison avec un ensemble de textes syndicaux comparables.

Une analyse diachronique sur les congrès de la seule centrale CFDT montre que cette reprise régulière de séquences attestées dans le ***Manifeste*** ne concerne que les périodes 1973 et 1976 du corpus et que les congrès qui se dérouleront à partir de 1979 éviteront de faire appel aux séquences de ce type.

La méthode proposée devrait permettre d'étudier avec profit toute une série de phénomènes liés à la circulation d'unités textuelles plus longues que la forme de vocabulaire isolée de son contexte immédiat à l'intérieur de grands corpus de textes produits par différentes formations discursives.

Références

Bakhtine M. (1984). *Esthétique de la création verbale.* Paris, Gallimard.

Luong X. (dir.) (2003). La distance intertextuelle. *Corpus*, n°2, http://revel.unice.fr/corpus/sommaire.html?id=52.

Hetzel A.-M., Lefèvre J., Mouriaux R., Tournier M. (1998). *Le syndicalisme à mots découverts. Dictionnaire des fréquences (1971-1990).* Paris, Syllepse.

Foucault M. (1971). *L'ordre du discours.* Paris, Gallimard.

Lamalle C., Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. in *Actes des 6èmes Journées d'analyse statistique des données textuelles,* Université de Rennes.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Paris, Slatkine-Champion.

Lebart L., Salem A. (1994). *Statistique textuelle,* Paris, Dunod.

Longrée D., Mellet S. (2004). Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatize. In G. Purnelle, C. Fairon & A. Dister (éds), *Le poids des mots, actes des 7èmes Journées d'analyse statistique des données textuelles*. Louvain-la-Neuve, UCL, Presses universitaires de Louvain.

Rajman M., Lebart L. (1998). Similarités pour données textuelles. In *Actes des 4es Journées internationales d'Analyse statistique des Données Textuelles*, Université de Nice.

Pêcheux M. (1969). *Analyse automatique du discours*. Paris, Dunod.

Salem A. (1987). *Pratique des segments répétés.* Publications de l'INaLF, Paris, Klincksieck.