Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels

Cédric Lamalle, André Salem

EA 2290 SYLED, Université de la Sorbonne nouvelle – Paris 3 – 19, rue des Bernardins – 75005 Paris – France – cedric@cpac.embrapa.br, salem@msh-paris.fr

Abstract

The *type/token* relationship is considered under a generic point of view, based on a wider definition of lexicometrical units. The distribution of units within the text corpora is then considered from a "spatial" approach.

Résumé

On considère une généralisation de la relation type/occurrence en étudiant la répartition d'unités dont la définition est plus générique. La localisation à l'intérieur du corpus des unités ainsi constituées est ensuite étudiée d'un point de vue « spatial » à l'intérieur des textes.

Mots-clés : segmentation automatique, statistiques textuelles, topographie textuelle

1. Introduction

La question du choix des unités de décompte les mieux adaptées à l'étude statistique des textes a souvent fait l'objet de controverses au sein des communautés scientifiques qui utilisent les différentes approches quantitatives de corpus textuels. Dans ce débat déjà ancien, des arguments hétérogènes interviennent tour à tour, appuyés sur des considérations de type linguistique (Quelles sont les unités qui circulent dans le corpus des textes que l'on étudie ?) ou statistique (La prise en compte de tel ou tel type d'unités a -t-elle une influence quelconque sur les résultats produits par les analyses lexicométriques ?)[1].

L'intérêt de ce débat est, avant tout, de mettre en lumière l'existence de préoccupations très diverses dans le monde des analyses automatiques appliquées aux textes. On réunit des corpus de textes à des fins très différentes : comparer les moyens d'expressions employés par différents auteurs, étudier l'évolution du vocabulaire d'une source textuelle au cours du temps, accéder à des « contenus » véhiculés par les productions textuelles d'individus répondant à une enquête socio-économique.

En l'état actuel des choses, la question du choix des unités ne saurait être tranchée une fois pour toutes et pour tous les types d'études à venir. Il nous semble au contraire que cette question constitue une piste de recherche parmi les plus intéressantes dans le domaine des

[1] Pour un exposé des différents points de vue exprimés dans ce débat, on consultera par exemple (Muller 1977 ; Lafon 1981 ; Labbé 1990 ; Brunet 2000 ; Salem 1993 ; Bolasco 1992 ; Lebart & Salem 1994). Un numéro en cours d'élaboration de la revue électronique *Lexicometrica* (http://cavi.univ-paris3.fr/lexicometrica) coordonné par D. Labbé tente également de faire le point sur ce sujet.

études textuelles automatisées. Nous tenterons dans ce qui suit de définir un type d'unités génériques différent de celui couramment employé dans les études lexicométriques.

Les unités que nous proposons sont avant tout destinées à permettre au chercheur de faire des expériences sur les corpus de textes et d'avancer vers des réponses plus satisfaisantes aux questions posées plus haut. Les fonctionnalités logicielles décrites ci-dessous ont été implémentées dans la récente version d'un logiciel d'analyse lexicométrique (***Lexico3***) élaboré dans le cadre de notre équipe[2].

2. Les types généralisés

Dans ce qui suit, nous appellerons ***occurrence*** *(angl. token)* chacun des éléments découpés par un algorithme de segmentation automatique au fil d'un corpus de texte et ***type*** *(angl. type)* les divers regroupements de ces unités que l'on peut opérer sur la base de leur identité ou de leurs ressemblances[3].

On peut définir le *type généralisé* ***TGen*** comme « ensemble d'occurrences sélectionnées parmi les occurrences du texte ». Cette définition très large permet de généraliser le concept de *type* (ou de *forme*) habituellement utilisé dans le domaine lexicométrique. Certains des types ainsi définis sont susceptibles de recevoir une description *en compréhension* (c'est à dire qu'il est possible dans ces cas d'énoncer une propriété commune à toutes les occurrences du texte qui relèvent du type considéré et de définir par la même cet ensemble – par exemple : les occurrences de la forme graphique *liberté* ou les occurrences du texte dont les trois premiers caractères sont : *lib*).

Le marquage au fil du texte, éventuellement guidé par des considérations difficiles à formaliser, d'un sous ensemble d'occurrences pour lequel aucune définition en compréhension n'est possible (– exemple : le vocabulaire « politique » du texte) détermine au contraire un type qu'il sera plus aisé de décrire *en extension* (i.e. en fournissant, par exemple, une liste, peut-être provisoire, des adresses de chacune des occurrences qui constituent le type ainsi créé). Ce deuxième type d'unité lexicométrique constitue l'exemple même de l'unité que le chercheur est souvent amené a prendre en compte au cours d'une expérience, parce qu'elle correspond à un phénomène qu'il commence à percevoir, sans pour autant être capable d'en tracer les limites précises.

3. Traitement informatisé des *TGen(s)*

La définition donnée plus haut permet de considérer un ensemble assez vaste de nouvelles unités dont on peut recenser les occurrences au fil des textes. Sur des ensembles de textes étendus, la question de l'automatisation de ces comptages se pose assez rapidement. Remarquons que les comptages peuvent être aisément automatisés pour tout une série de types. Ainsi, on peut recenser systématiquement au-delà des occurrences des formes graphiques dont on a parlé plus haut :

---

[2] Cf. *Lexico3* – Outils d'analyse lexicométrique (Lamalle et al. 2001)

[3] Dans les études lexicométriques, on commence en général par segmenter le texte en *occurrences* de formes graphiques (chaînes de caractères non-délimiteurs bornées par deux caractères délimiteurs) pour regrouper ensuite les occurrences identiques sous un même *type*. Si l'on effectue un dépouillement en *lemmes* la définition, pour chaque occurrence d'un type de rattachement est une opération plus complexe, bien que largement automatisable, qui nécessite à la fois le concours d'un dictionnaire et un retour au contexte dans certains cas.

- les occurrences d'un segment répété[4] ou d'un quasi-segment[5] (suite d'occurrences de longueur donnée reproduite plusieurs fois dans un corpus de texte – exemple *: liberté d'expression*).
- la rencontre de deux formes (ou *cooccurrence*) à l'intérieur d'une phrase, d'un paragraphe, ou d'une fenêtre de x-mots graphiques[6].
- le type constitué par les occurrences d'un ensemble de formes graphiques défini en raison de la parenté lexicale de ces dernières (exemple *liberté, libertés, libre, etc.)*, sémantique, etc.[7]

Enfin, comme on vient de l'évoquer ci-dessus, on peut envisager de soumettre aux mêmes décomptes statistiques des types constitués par le marquage d'un thème ou d'un registre prédéfini par un expérimentateur annotant, de manière difficilement formalisable, certaines des occurrences au fur et à mesure d'une lecture érudite.

4. La fragmentation du texte

Les comparaisons statistiques entre textes constituent la partie centrale de la démarche lexicométrique. Dans la pratique, la plupart des corpus de texte se laissent d'autant plus aisément décomposer en « parties naturelles » qu'ils ont été fabriqués à partir de ces mêmes textes réunis à des fins de comparaison.

Dans certains domaines (études littéraires par exemple), l'œuvre analysée a été, dès sa création, divisée en chapitres ou en volumes qui donnent au texte une partition « naturelle » au corpus. Dans d'autres au contraire (analyse lexicométrique des questions ouvertes), de petits fragments de textes constitués par des réponses à une même questions sont regroupés en parties artificielles qui constituent des agrégats homogènes du point de vue des variables sociologiques qui font l'objet de l'enquête.

Au delà de ces divisions du corpus en parties, l'expérimentateur recourt souvent, dans les cas ou la division des textes réunis en corpus constitue elle-même une partie du problème posé, à des divisions plus régulières, réputées plus « neutres », en pages, en tranches de 100 mots, en paragraphes, en phrases ou encore en fenêtres de longueur fixe se limitant à quelques mots à droite et à gauche d'un pôle retenu[8]. La division en phrases, peut-être réalisée approximativement par la sélection d'un certains nombre de caractères que l'on dotera du statut de délimiteurs de phrases[9]. Cependant, le paragraphe que l'on peut souvent identifier au

[4] On décide arbitrairement dans ce cas que c'est la première des occurrences du segment qui fournit l'adresse de la séquence répétée. Sur les segments répétés on consultera (Lafon et Salem, 1983), (Salem 1984).

[5] Sur les quasi segments, voir par exemple (Bécue et Peiro, 1993).

[6] Sur la méthode des cooccurrences (Lafon 1984).

[7] Ces types peuvent être constitués assez facilement à l'aide de procédure informatisées permettant l'accès au langage des expressions régulières.

[8] Sur ces questions on se reportera, par exemple, à (Lebart et Salem, 1994) ainsi qu'à (Reinert, 1993).

[9] On sait que, dans la pratique, cette division du texte en phrases est assez difficile à réaliser automatiquement du fait de l'ambiguïté inhérente à certains signes quant à leur statut de délimiteurs (ex. le point, tantôt séparateur de phrase, marquant parfois l'abréviation [M.], délimitant les sigles [S.N.C.F], également présent dans les points de suspension [...]).

caractère *retour-chariot* dans les textes encodés sur support lisible par un ordinateur constitue une division privilégiée du texte, à la fois relativement fiable (cette division peut souvent être portée au crédit de l'auteur du texte qui a signifié par là qu'il abordait un sujet légèrement différent de celui traité dans le paragraphe précédent) et qui présente dans de nombreux cas une certaine régularité du point de vue de la taille des ensembles qu'elle découpe.

5. Spécificités, spécificités chronologiques, fragments caractéristiques

Le repérage du *vocabulaire caractéristique* ou *spécificités*[10], permet de sélectionner pour chaque partie du corpus une série de types plutôt sur-représentés dans cette partie de texte par rapport à l'ensemble du corpus. Cette méthode permet à la fois de sélectionner des unités qui illustrent tout particulièrement la spécificité de la partie textuelle considérée et d'esquisser une typologie des parties qui s'appuie simultanément sur le sur–emploi et le rejet de certaines formes.

Dans le cas d'une *série textuelle chronologique* (corpus de textes produits par une source homogène, calibrés du point de vue de leur longueur et étalés dans le temps), il est utile d'établir, à côté des jugements statistiques portant sur la ventilation de chaque unité textuelle dans chacune des périodes du corpus, des diagnostics portant sur la répartition des mêmes unités dans les groupes de périodes consécutives (*spécificités chronologiques*) et de mettre en évidence pour chaque unité de décompte, les groupes de périodes consécutives au cours desquelles la fréquence de l'unité connaît des variations importantes dans son utilisation (*accroissements spécifiques*).

Une fois repérées les principales variations fréquentielles, dans l'utilisation d'une unité statistique pour une portion de texte donnée (par exemple, l'abondance relative des occurrences d'un type donné dans une des périodes du corpus) la question se pose de préciser les limites de la portion de texte affectée par le phénomène mis en évidence, et d'exhiber de manière automatique des fragments de textes tout particulièrement significatifs par rapport au phénomène constaté[11]. Dans ce qui suit, et en nous appuyant sur un exemple concret, nous montrons qu'il est possible d'affiner encore ce genre de description par le recours systématique à une représentation topographique de la répartition des unités textuelles à l'intérieur d'un corpus de textes.

[10] Sur la méthode des spécificités on consultera (Lafon, 1981), sur les séries textuelles chronologiques et les spécificités chronologiques (Salem 1993), sur les réponses caractéristiques (Lebart, 1982) ainsi que des rappels sur l'ensemble de ces questions dans (Lebart et Salem, 1994).

[11] Dans le cas de corpus de textes constitués par les réponses à une question ouverte recueillies lors d'une enquête socio-économique, on parle de *réponses caractéristiques* d'un individu pour l'ensemble d'un groupe de sujets pour dire que sa réponse contient un grand nombre de mots fréquemment utilisés « en moyenne »par l'ensemble du groupe (Cf. Lebart et Salem, 1994).

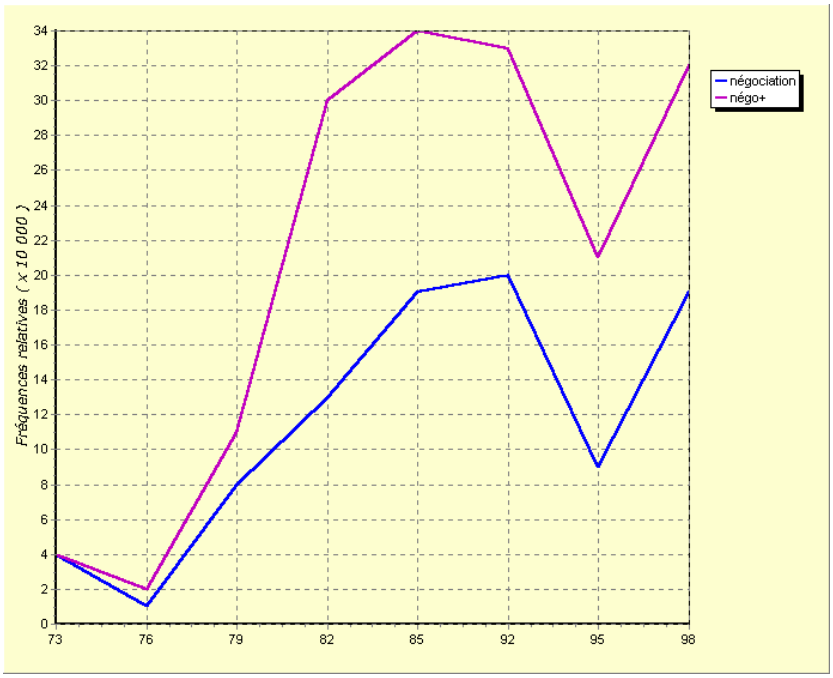

Figure 1 : La ventilation de la forme négociation *et du type* ***négo+*** *dans les 8 congr*ès

6. Application à un corpus de texte : le thème de la *négociation* dans le corpus *CFDT 1973-1998*

Pour illustrer notre propos, nous emprunterons des exemples à un corpus de textes syndicaux ***CFDT1973-1998***[12]. Ce corpus regroupe l'ensemble des résolutions votées par une même centrale syndicale - la Confédération Française Démocratique du Travail (CFDT) - au cours des huit congrès consécutifs qu'elle a tenu entre 1973 et 1998.

6.1. Le thème de la négociation

Dans ce corpus qui a suscité de nombreuses études de caractère socio-politique ainsi que des études à caractère méthodologique, le thème de la *négociation* subit de profondes variations au plan fréquentiel tout au long de la période considérée[13], ce qu'on peut vérifier en se reportant à la figure 1.

Sur ce même graphique, on a transcrit simultanément la ventilation du type ***négo+*** constitué par l'ensemble des occurrences des formes graphiques qui relèvent de la même famille morphologique :

```
négociable(1 occ.), négociant(1 occ.), négociateur(1 occ.), négociation(139
occ.), négociations(33 occ.), négocie(2 occ.), négocié(6 occ.), négociée(15
```

[12] Ce corpus a été réuni au laboratoire de l'ENS de Fontenay-Saint-Cloud dans le cadre d'une recherche portant sur l'ensemble des centrales syndicales françaises pendant la période (1971-2000). (Hetzel et al., 1998) propose une synthèse des recherches effectuées à partir de ce corpus.

[13] On trouvera dans (Salem 1993) des exemples d'études chronologiques réalisées à partir de corpus de ce type.

```
occ.), négociées(11 occ.), négocier(20 occ.), négociera(1 occ.), négociés(7
occ.)
```

Comme on le voit sur la figure 1, la ventilation du type ***négo+*** est assez semblable dans ses grandes variations à celle des occurrences de la forme graphique *négociation.* On note que la prise en compte d'un plus grand nombre d'occurrences liées au concept de négociation permet de mieux en apprécier l'évolution chronologique. L'accroissement chronologique constaté sur la ventilation de la forme graphique *négociation* dans les périodes correspondant aux congrès de 1982 et de 1985 est encore plus net si l'on prend en compte les variations de l'ensemble de la famille morphologique liée à cette forme.

L'étude des cooccurrences (recensement des unités textuelles souvent présentes dans les mêmes phrases ou dans les mêmes paragraphes que les occurrences du type considéré) apporte elle aussi des résultats plus riches que dans le cas de cooccurrences calculées à partir de la seule forme *négociation*[14].

6.2. Cartographie textuelle

Une analyse plus précise de la localisation des occurrences du type ***négo+*** dont on a noté plus haut qu'il présente des variations chronologiques importantes dans certaines périodes du corpus peut être envisagée si l'on entreprend de donner une représentation topographique du texte.

La figure 2 montre un type particulier de localisation de l'ensemble des occurrences du type ***négo+*** dans l'ensemble du corpus. Pour cette représentation on a choisi de fragmenter le texte en paragraphes (on aurait pu choisir une représentation analogue fondé sur un découpage en phrases).

Sur cette figure, chacun des 2319 paragraphes que compte le corpus ***CFDT 1973-1998*** est représenté par un carré de taille fixe (cette taille ne varie pas en fonction de la longueur du paragraphe, mesurée en nombre d'occurrences). Chaque ligne compte 50 paragraphes regroupés par paquets de 10. La division du texte en périodes (ici les différents congrès tenus par l'organisation syndicale) est matérialisée quant à elle par des lignes horizontales.

Dans cette représentation particulière en *présence-absence*, les carrés de couleur sombre signalent la présence, au sein du paragraphe concerné, d'une occurrence au moins du type cartographié (ici le type ***négo+***)[15].

Ce type de représentation permet de préciser la description esquissée à partir de la ventilation des occurrences du type dans les périodes du corpus. L'accroissement spécifique constaté pour les occurrences du type ***négo+,*** à propos de la période qui correspond au congrès de 1985, résulte en fait, pour l'essentiel, comme on peut le vérifier sur la figure 2, de la présence d'occurrences de ce type dans des séries de paragraphes consécutifs aisément localisables sur le graphique. Cela est encore plus probant pour la période qui correspond au congrès de 1998 pour lequel les occurrences du type se concentrent presque toutes sur deux zones facilement

[14] Ces résultats qui sortent du cadre de la présente étude seront développés dans un travail à venir.

[15] On peut encore préciser cette description en affectant à chacun des carrés - paragraphes de cette représentation d'une couleur plus ou moins foncée selon le degré de spécificité attaché au nombre des occurrences relevant du type étudié dans le paragraphe considéré (i.e. pratiquer un seuillage statistique pour manifester que la forme est significativement présente dans le paragraphe considéré).

repérables. Le problème de la délimitation automatique du contour de ces zones fait l'objet d'une étude à venir.

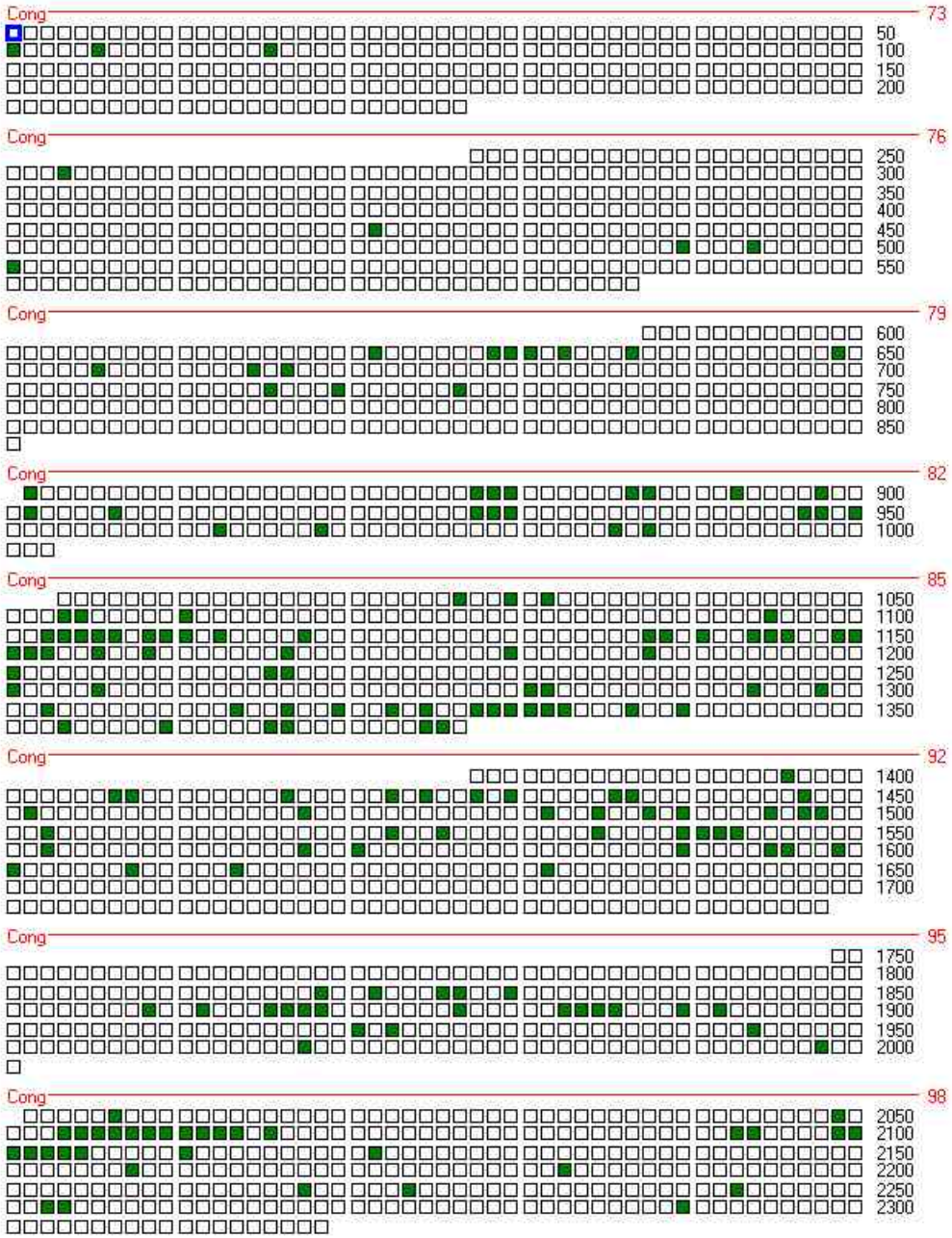

*Figure 2 : Les occurrences du type **négo+** dans les 2319 paragraphes des 8 congrès CFDT*

Dans les cas les plus favorables, ces paragraphes délimitent des zones vers lesquelles le chercheur intéressé par ce type de phénomènes aura tout intérêt à diriger son attention. Pour rendre cette dernière tâche plus aisée, il est possible de réaliser des éditions des zones textuelles ainsi mises en évidence dans lesquelles les occurrences qui relèvent du type que

l'on a choisi d'étudier seront distinguées, en en caractère gras, par exemple, comme cela a été fait sur la figure 3,

Sur cette figure, on a sélectionné quelques-uns des paragraphes qui sont à l'origine de l'abondance relative du type ***négo+*** dans la dernière période du corpus (1998). Sous réserve d'une formalisation à venir, on peut avancer que, dans cette période, plus encore que dans les périodes 1985 et 1992, les occurrences du type ***négo+*** sont regroupées dans des séquences de paragraphes consécutifs qui traduisent certainement le traitement de ce thème particulier à ces endroits précis du corpus.

Congrés CFDT de 1998 - §§ 2051-2054

§ dans les **négociation**s d'entreprise et de branche, dans les fonctions publiques et les entreprises publiques, la CFDT lie de manière dynamique et diversifiée les salaires et l'emploi.
§ le choix de l'emploi par la RTT fait de la compensation salariale un des éléments de la **négociation**, sans a priori dans un sens ou dans l'autre. dans ces **négociation**s, les équipes syndicales prennent en compte le volume d'emplois créés, l'ampleur de la RTT, la participation de l'entreprise et le niveau des salaires et de ses éléments accessoires (intéressement, participation, actionnariat...).
§ dans ces **négociation**s, la CFDT porte l'exigence du maintien du pouvoir d'achat des salaires les plus bas. elle revendique pour les temps partiels contraints la possibilité d'augmenter leur durée de travail et leur rémunération.
§ l'amélioration des plus bas salaires emprunte pour la CFDT différentes voies. le SMIC doit évoluer au-delà de l'obligation légale actuelle en intégrant la totalité de l'augmentation du pouvoir d'achat de la moyenne des salaires. les minima conventionnels doivent être décrochés du SMIC. la **négociation** des classifications doit permettre à tous les salariés de bénéficier d'un déroulement de carrière. la CFDT poursuit son action pour l'amélioration des basses rémunérations dans les fonctions publiques.

Figure 3 : Quelques paragraphes caractéristiques du congrès CFDT de 1998

7. Conclusion

Au terme de cette étude, centrée sur la question des unités de décompte lexicométrique et de la représentation de leur « organisation spatiale » à l'intérieur d'un corpus de textes, nous proposons au lecteur de nouveaux moyens d'investigation pour aborder ce type de tâche.

a) la possibilité de constituer des ensembles d'unités sur la définition desquelles le chercheur peut agir plus aisément le temps d'une expérience ;

b) la possibilité de cartographier l'extension spatiale d'une unité à travers un découpage du corpus dont le grain peut être réglé à volonté devrait permettre, à travers un retour au texte beaucoup plus précis d'affiner les constats produits dans le cadre des études lexicométriques.

Les applications ouvertes par cette approche concernent plusieurs types de recherche au premier plan desquelles se trouvent le domaine de l'étude des séries textuelles chronologiques, celui de l'étude des corpus multilingues et la réalisation de systèmes de veille technologique.

Références

Benzécri J-P. et coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*, Paris, Dunod.

Bécue M. Peiro R.. (1993). «Les quasi-segments pour une classification automatique des réponses ouvertes», in *Actes des secondes JADT*, Montpellier, ENST.

Bolasco S. (1992) «Criteri di lemmatizzazione per l'individuazione di coordinate semantiche. » in *Atti del Convegno internazionale "Ricerca qualitativa e computer nelle Scienze Sociali"*, Roma, ENEA.

Brunet E. (2000) « Qui lemmatise, dilemme attise ». in *Lexicometrica n°2,* revue électronique sur le web.

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

Labbé D., (1990) *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

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

Lafon P., Salem A. (1983) «L'Inventaire des segments répétés d'un texte*»*, in *Mots N° 6*, Paris p..161-177.

Lamalle C., Martinez. W, Fleury S., Salem A., Kuncova A., Maisondieu A. (2001*) Lexico3 - Outils de statistique textuelle* - http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/

Lebart L. (1982). « *L'analyse statistique des réponses libres dans les enquêtes socio-économiques»*.in *Consommation* N°1, Dunod. p. 39-62.

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

Muller Ch., (1977) *Principes et méthodes de statistique lexicale*, Paris, Hachette.

Reinert M.(1993). « Les "mondes lexicaux" et leur logique. » in *Langage et société,* 66 : 5-39, Paris..

Salem A. (1987) Pratique des segments répétés, Paris, Klinksieck.

Salem A. (1993) *Méthodes de la statistique textuelle*, Thèse pour le Doctorat d'Etat, Université Paris3.