Les lexèmes dans l'espace du texte : analyses élaborées et bases de voisinage

Juillard Michel

Cycnos, études anglophones
revue électronique éditée sur épi-Revel à Nice
ISSN 1765-3118 ISSN papier 0992-1893

Les lexèmes dans l'espace du texte : analyses arborées et bases de voisinage

M. Juillard*

Mon propos s'inscrit tout naturellement dans le thème officiel du congrès, l'esprit des lieux, puisque je présenterai quelques résultats de recherches en cours au laboratoire CNRS bases-corpus-langage (ex URL9) mettant en oeuvre les concepts fondamentaux de la topologie, partie de la mathématique consacrée à l'occupation de l'espace : plan, volume ou texte, par ses entités constitutives : points, figures ou lexèmes. Pierre Guiraud, disciple de Guillaume et homme libre, qui fut professeur de linguistique dans ces lieux, déclarait :

"La linguistique est la science statistique type; les statisticiens le savent bien; la plupart des linguistes l'ignorent encore" (*Problèmes et méthodes de la statistique linguistique*, 1960).

Les choses ont bien changé depuis, en tout cas à Nice où la mise en garde de Guiraud a été entendue. La statistique classique utilisait certes déjà des outils fort efficaces : écarts-réduits, Khi-deux, coefficients de corrélation, analyse des correspondances mais la panoplie était encore limitée et surtout le bon matériau était rare.

On dispose aujourd'hui de données abondantes et de qualité, lorsque ces données linguistiques ont été codées statistiquement avec finesse, rigueur et pertinence.

La puissance accrue des outils informatiques permet en outre de soumettre ces donneés textuelles à de nouveaux algorithmes fondés sur les mathématiques les plus subtiles, sans les dénaturer, mais encore en éclairant leur comportement et en respectant la qualité essentielle des énoncés : la linéarité, c'est-à-dire la dynamique du langage en situation.

Après avoir montré comment on peut offrir une image synthétique parlante des données grâce à l'analyse arborée, je consacrerai l'essentiel de cette présentation à un autre aspect de la topologie appliquée aux textes : l'utilisation de la notion de base de voisinage.

Toutes les données utilisées ici proviennent de la version codée syntaxiquement du ***LOB corpus of English***, ensemble de plus d'un million

* CNRS Nice.

de mots représentant tous les domaines d'emploi de la langue écrite. Ce corpus dont la genèse remonte aux années soixante-dix a été conçu comme l'équivalent pour l'anglais d'Europe de ce que le Corpus de Brown University est à l'anglais d'Amérique. Le projet a pris forme à l'Université de Lancaster sous la direction de Geoffrey Leech jusqu'en 1976, date à laquelle l'Université d'Oslo et le Centre de Calcul pour les Lettres et Sciences Humaines (Norwegian Computing Centre for the Humanities) de Bergen ont pris le relais sous la direction de Stig Johansson[1]. A l'image de son homologue américain, le LOB Corpus se compose de 500 échantillons d'environ 2500 occurrences représentant la plupart des domaines de la langue écrite et tous provenant d'oeuvres et de documents divers publiés au cours de l'année 1961. Voici, avec leurs appellations anglaises, comment se répartissent les quinze grandes catégories de textes et le nombre d'échantillons de chacune :

Catégorie A	*Press : reportage*	(44 échantillons)
Catégorie B	*Press : editorial*	27 échantillons)
Catégorie C	*Press : reviews*	(17 échantillons)
Catégorie D	*Religion*	(17 échantillons)
Catégorie E	*Skills, trades and hobbies*	(38 échantillons)
Catégorie F	*Popular lore*	(44 échantillons)
Catégorie G	*Belles lettres, biography, essays*	(77 échantillons)
Catégorie H	*Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)*	(30 échantillons)
Catégorie J	*Learned and scientific writing*	(80 échantillons)
Catégorie K	*General fiction*	(29 échantillons)
Catégorie L	*Mystery and detective fiction*	(6 échantillons)
Catégorie N	*Adventure and western fiction*	(29 échantillons)
Catégorie P	*Romance and love story*	(29 échantillons)
Catégorie R	*Humour*	(9 échantillons)

1 On consultera pour plus de détails Johansson (Stig) : (1980), "The LOB Corpus of British Texts : Presentation and Comments", *ALLC Journal* 1, 25-36 ainsi que l'introduction de Hofland (Knut) et Johansson (Stig) : (1982), *Word Frequencies in British and American English*, the Norwegian Computing Centre for the Humanities, Bergen.

Les méthodes d'échantillonnage, la composition, l'équilibre interne et les sources de chacune des catégories se trouvent décrits dans un manuel rédigé à l'intention des utilisateurs.[2]

La version codée est apparue au cours des années quatre-vingt. Le codage a été effectué avec une rigueur, une intelligence et une exhaustivité exemplaires par une équipe réunissant, sous la direction de Stig Johansson, Geoffrey Leech et Roger Garside une équipe de chercheurs linguistes et informaticiens, notamment Eric Atwell, Ian Marshall, Mette-Cathrine Jahr et Knut Hofland. Il suffira de dire que ce travail enrichit d'informations considérables la base textuelle de départ et lui confère une nouvelle dimension qu'on pourrait appeler celle de la profondeur syntaxique.

Les décisions préalables au codage s'appuient sur la théorie linguistique, de tradition essentiellement firthienne, la plus sûre et la plus cohérente. Le codage résultant de l'analyse fine de toutes les entités linguistiques et graphiques des textes invite à des explorations variées en ouvrant de nombreuses voies à l'imagination du chercheur.

Ainsi, aux vingt-trois étiquettes grammaticales de départ (base tags) peuvent s'adjoindre des suffixes qui rendent compte tour à tour du nombre, du genre, du cas, de la personne, du temps, du degré de l'unité considérée[3]. La seule catégorie du nom, par exemple, se subdivise en 31 sous-catégories selon les appartenances aux divers types de noms propres ou au nom commun et par le jeu combiné du genre et des flexions (*Users' Manual, op. cit*, pp. 144-146).

Ces options de codage assurent une grande souplesse en permettant des regroupements de sous-catégories et en autorisant des comparaisons avec d'autres corpus codés différemment et dont on peut en quelque sorte aisément "émuler" la norme. Les auteurs reconnaissent néanmoins l'existence d'incontournables zones à problèmes (*Users' Manual, op. cit.*, p. 27).

Même si quelques décisions peuvent inévitablement paraître arbitraires, les difficultés identiques ont reçu des solutions identiques

---

2 Johansson (Stig), Leech (Geoffrey N.), Goodluck (Helen) : (1978) *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*, Department of English, University of Oslo.

3 On trouvera une description très claire et complète des options de codage dans Johansson (Stig), Atwell (Eric), Garside (Roger), Leech (Geoffrey) : (1986) *The Tagged Lob Corpus, Users' Manual*, Norwegian Computing Centre for the Humanities, Bergen. Nous souhaitons remercier Stig Johansson de nous avoir spontanément et généreusement offert un exemplaire de ce précieux document.

assurant par là même simplicité et constance, qualités primordiales et garantes de l'efficacité de toute norme selon Charles Muller.[4]

Le plus souvent, les données chiffrées du linguiste utilisant ce genre de corpus se présentent sous la forme d'un tableau à n lignes et m colonnes.

Les lignes correspondent à des **objets**, pour nous des textes, les colonnes correspondent à des **attributs** ou **variables** de ces objets, pour nous l'appartenance à telle ou telle catégorie grammaticale ou classe syntaxique.

Le tableau 1 est un exemple classique.

Txt	DET	AuxM	Coor	SUB	Prep	Adj	N	Pron	Adv	V	Interj
T01	233	169	65	37	224	114	447	279	179	278	1
T02	205	225	66	51	225	77	420	297	122	310	1
T03	321	116	74	33	237	147	530	164	145	262	1
T04	240	152	71	27	245	125	430	304	164	270	7
T05	216	226	46	48	222	125	436	295	136	274	7
T06	237	207	107	37	203	142	377	259	212	267	7
T07	215	209	77	45	243	107	307	374	176	283	8
T08	269	190	84	58	227	101	387	278	163	230	6
T09	348	106	93	17	305	171	492	138	138	199	3
T10	261	128	53	30	251	138	448	296	142	282	10
	DET	AuxM	Coor	SUB	Prep	Adj	N	Pron	Adv	V	Interj

Tableau 1 : les classes grammaticales dans les textes

Ce tableau peut tel quel donner lieu à de nombreux tests statistiques fondés sur la comparaison de cette réalité à un modèle, par exemple celui de la loi normale.

Souvent, on recherche une vue synthétique du comportement des données au gré des textes; on a alors tout intérêt à recourir à une représentation sous forme d'arbre.

L'idée première de recourir aux schémas arborescents les plus simples est bien antérieure à Chomsky ou même Louis Tesnière puisque c'est le philosophe grec néoplatonicien Porphyre (233-301) qui inaugura l'arbre

---

4 Muller (Charles) : (1977) *Principes et Méthodes de Statistique Lexicale*, Paris Hachette, p. 28.

pédagogique et heuristique pour représenter la hiérarchie des concepts conduisant de la substance inanimée à l'homme, en opérant une suite de choix binaires.

Notre représentation arborée quant à elle s'appuie sur la théorie mathématique et fait appel à des algorithmes mis en oeuvre au moyen de programmes informatiques originaux. Cette représentation traduit la distance entre les textes, chacun d'eux se caractérisant par sa distance à chacun des autres, les textes étant envisagés comme les points d'un espace à m dimensions. L'algorithme conçu par mon collègue mathématicien et informaticien Xuan Luong fait appel à la distance du Khi-deux pour obtenir un indice de proximité.

Le modèle d'arbre produit est "un graphe connexe et sans circuit caractérisé par l'ensemble des distances entre ses éléments". Les distances sur un arbre satisfont à la propriété dite des quatre points de Buneman et Dobson (pour plus d'information, consulter les ouvrages de X. Luong, en particulier sa thèse de Doctorat d'Etat, *Méthodes d'analyse arborée*, Paris V 1988).

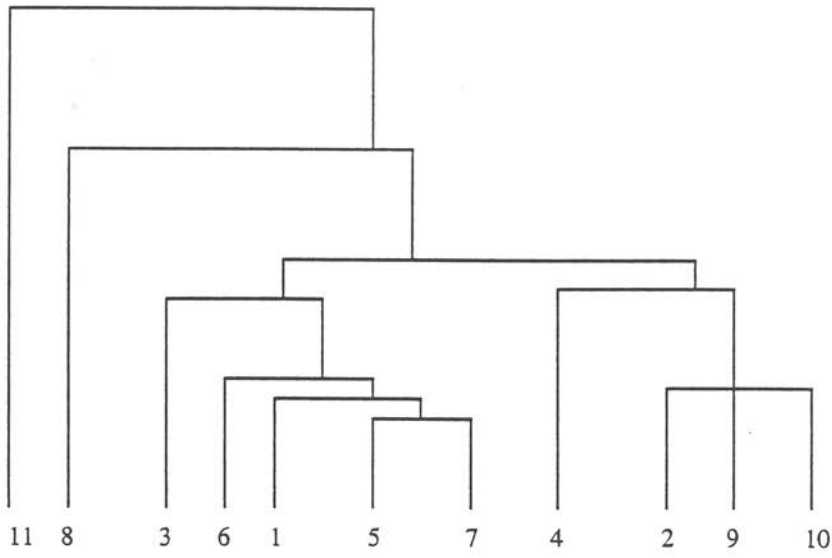

Figure 2 : arbre planté

La lecture de cet arbre en remontant des feuilles vers la racine met en évidence des classes distinctes qui s'emboîtent les unes dans les autres.

La distance d'une feuille à un noeud constitue un indice du niveau de formation d'une classe.

Tous les éléments d'une classe sont à égale distance d'un noeud : ainsi les catégories grammaticales 2, 9 et 10 (respectivement auxiliaires et modaux, verbes, adverbes) se trouvent associées entre elles avant de s'adjoindre la catégorie 4 (les subordonnants). La partie gauche de l'arbre est encore plus étroitement compartimentée.

Cet arbre est satisfaisant à l'oeil épris d'ordre parce qu'il sépare les objets par des cloisons étanches, sans pourtant nous renseigner finement sur le degré de leurs affinités, l'air de famille cher à Wittgenstein.

Tout autre est l'arbre radial, non planté, non hiérarchique, obtenu à partir des mêmes données :

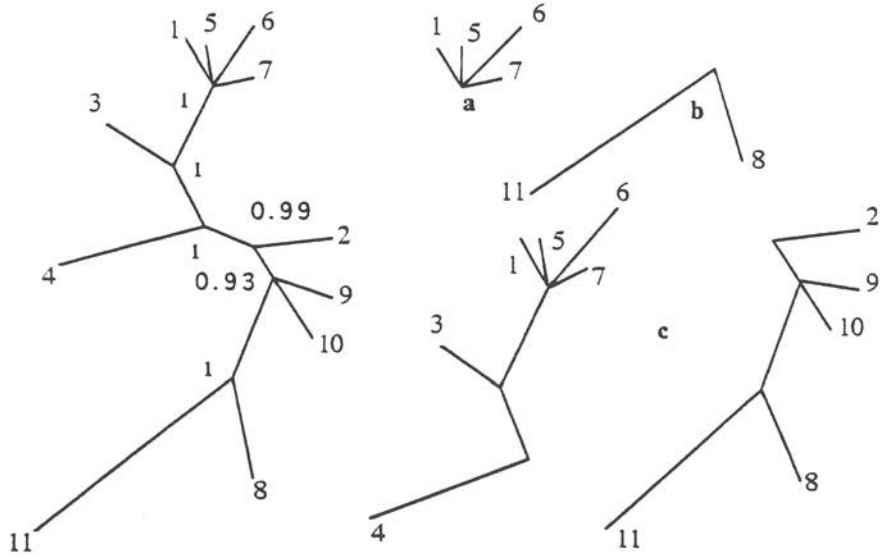

Figure 3 : l'arbre radial de L, P, H, J, K, N. Groupements **a** et **b**, opposition entre deux familles (**c**)

L'arbre est donc un arbre non planté, semblable aux arbres phylogénétiques des naturalistes. Les feuilles de l'arbre correspondent aux catégories grammaticales étudiées. La distance entre deux objets est la distance lue sur l'arbre. Les objets 1, 5 et 7 sont très proches. Dans une moindre mesure, 9 et 10 le sont aussi. On dégage ainsi un ***premier critère, de proximité***.

Un deuxième critère est d'ordre topologique (les objets se regroupent par leur proximité mais aussi par leur opposition commune à tel ou tel groupe) : 1-5-6-7 constituent un *groupe serré*, 11-8 un *groupe lâche*.

En enlevant une arête quelconque, on obtient une bipartition de l'arbre. Si l'on coupe l'arbre en son milieu, par exemple, on a une *opposition* entre deux familles : 1-5-6-7+3+4 d'une part et 2+9-10+(8-11) de l'autre.

Un *groupement* est un ensemble composé des feuilles reliées à un sommet commun. On peut le considérer comme une classe à laquelle chaque constituant ne contribue pas de manière identique.

- A chaque sommet intérieur se trouve associé un indice de formation de l'arbre, appelé *indice d'agrégation* (en gras sur la figure). Remarquons ici la qualité de cette représentation. (Si les données sont obtenues à partir des distances lues sur un arbre, ces indices sont alors tous égaux à 1).

- Soit un ensemble de n objets liés entre eux par un tableau de distances. Pour former une classe au sens de la classification hiérarchique ascendante on cherchera la distance la plus petite parmi les (n-1)(n-2)/2 distances.

Définissons maintenant la notion de *proximité arborée* :

On appelle bipartition significative d'un arbre une bipartition dont chacune des parties contient au moins deux éléments. Deux objets sont dits proches lorsque toutes les bipartitions significatives les regroupent toujours dans une même partie de l'arbre. Le lecteur peut vérifier que cette notion de proximité est très stable, car elle fait intervenir de très grandes familles d'ensembles et leur nombre est nettement plus important que celui des distances évoquées précédemment. L'algorithme utilisé ici se fonde sur cette notion de proximité arborée.

Pour parler en image (et de manière imprécise) on plonge les objets dans un espace muni d'une distance arborée pour en dégager les objets qui sont proches. On forme ainsi les premiers groupements. Ensuite, l'algorithme ajuste les données pour se conformer à la nouvelle structure. Chaque groupement représentera un nouvel objet et le processus peut se réitérer jusqu'à la formation complète de l'arbre.

(Le lecteur désirant un exposé plus rigoureux et plus exhaustif de la méthode dispose d'une étude complète des procédures d'analyse arborée dans une thèse récente).[5]

Le premier ensemble que nous avons traité en vue d'une exploitation linguistique est constitué d'une trentaine d'échantillons de deux mille cinq mots environ chacun, provenant des sections L, P et N du corpus (Lob corpus of English texts, version codée syntaxiquement), constituées respectivement de romans policiers, de romans d'amour et de romans d'aventures.

La figure produite au terme de l'analyse de ces soixante quinze mille occurrences rangées dans onze grandes catégories grammaticales (Figure 4) s'avère bien formée et présente une structure nettement arborée.

---

5 LUONG (Xuan N.) : 1988 *Méthodes d'analyse arborée. Algorithmes, applications,* doctorat d'Etat, Université Paris V René Descartes.

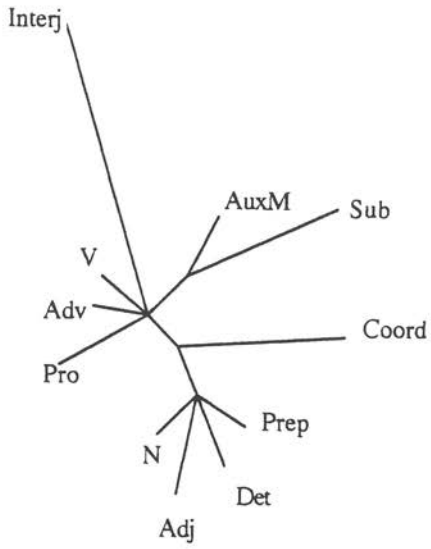

Figure 4 : les catégories grammaticales dans L, P, N.

S'il est possible de penser avec R. Quirk et de nombreux linguistes contemporains que la proposition est une unité syntaxique aux contours mieux définis que la phrase[6], l'application de la topologie arborée aux données codées grammaticalement du ***LOB corpus*** reflète clairement l'opposition fondamentale sujet-prédicat.

Un tronc trapu relie deux groupes de branches bien individualisés, l'un occupant le bas de la figure, l'autre la partie supérieure.

Le premier faisceau associe étroitement les noms, les adjectifs, les déterminants et les prépositions, qui sont les satellites naturels du groupe nominal.

S'oppose à ce groupe le noyau verbal qui réunit les verbes, les adverbes et les pronoms d'une part, les auxiliaires et modaux et les conjonctions de subordination d'autre part.

Le pronom se trouve dans le camp du verbe, manifestant ainsi ses spécificités par rapport au substantif et son rôle dans l'anaphore qui le place à distance pratiquement égale des verbes et des auxiliaires et modaux. Deux éléments ont du mal à s'agréger étroitement à l'arbre : la conjonction de coordination hésite entre le pôle verbal et le pôle nominal, ce qui reflète son ubiquité et sa portée variable dans les textes, reliant tantôt des groupes

6 Quirk (Randolph), Greenbaum (Sidney), Leech (Geoffrey), Svartvik (Jan) : 1985, *A Comprehensive Grammar of the English Language*, London, Longman, p. 47.

nominaux, tantôt des propositions; le rejeton démesuré qui surgit du coeur de l'arbre côté verbe est l'interjection; elle vient s'intégrer au groupe prédicatif mais à bonne distance, manifestant ainsi sa faible distribution, sa forte spécificité et son manque d'affinité avec toute autre partie du discours.

D'autres analyses d'autres unités (par exemple les signes de ponctuation) sur d'autres corpus (le ***TLF*** ou les textes de l'***Oxford Computer Archive of Modern English***) ont maintes fois prouvé la fécondité de la méthode.

Il est temps de parler d'un autre apport de la topologie à l'étude des énoncés réels. Le problème de la répétition des occurrences dans la linéarité du texte a donné lieu à quelques recherches pour mettre en évidence la distribution en ***rafales*** (Pierre Lafon ENS St Cloud) ou les "***effets de bloc***" (Philippe Thoiron et Dominique Sérant à Lyon). Nous avons, Luong et moi, conçu une approche différente, fondée sur la notion de ***base de voisinage***, soit pour chaque lexie considérée, un contexte de longueur variable, quelques mots, le groupe, la phrase, le paragraphe, le texte etc... Dans le premier exemple dont je vais parler, nous avons choisi une base étroite de six unités de part et d'autre du terme pivot, le mot ***there*** et ses deux valeurs, existentielle (pronom, inaccentué, faible) ou adverbiale (accentuée, forte) dans le sous-corpus de presse (88 échantilllons sur 500, 228 698 occurrences, 18 733 formes ou vocables).

Nos résultats pour ce premier essai, vont bien au-delà de l'image de cet humble donnée par les lexicographes. Certains dictionnaires (le ***C.O.D.***) ne font même pas la différence entre les deux variétés de ***there*** (existentielle ou adverbiale). seul le dictionnaire de Longman (***LDOCE***) dans sa dernière édition propose une analyse syntaxique acceptable de ce mot.

Tous les dictionnaires, sans exception, sont très loin de représenter l'usage contemporain tel qu'il est reflété dans la version codée syntaxiquement du corpus ***LOB*** (codes EX et RN respectivement pour ***there1*** et ***there2***).

Alors que les lexicographes donnent le même poids aux deux variantes de ***there***, les exemples issus de nos voisinages montrent que l'on rencontre 6 emplois existentiels pour à peine un seul emploi adverbial.

L'énoncé non-marquée procède de l'information donnée ou thème (le sujet...) à l'information nouvelle ou rhème (parfois appelé propos). L'emploi de ***there*** existentiel permet au journaliste de faire passer tout l'énoncé dans la partie rhématique.

Ainsi la phrase syntaxiquement acceptable, *a fly is in the ointment* devient : *there's a fly in the ointment.*

there is doubt - there is evidence - there is something -
there is much (talk, liberty, leeway, prophecy etc..)

la plupart de ces expressions standard récurrentes sont à la forme négative, sémantiquement ou syntaxiquement :

there is little doubt - there is no doubt - there is no evidence
there is no need - there is no reason - there is no hope -
there is no time - there is no reason - there is no question -
there is no truth (in)

Les suites les plus remarquables sont celles qui combinent *there* existentiel et un modal :

there will/would be - there can/could be

De nombreuses occurrences de ***there*** dans ces contextes sont suivies par un quantifieur subjectif ou le déterminant négatif ***no*** :

there is little doubt - there is little illusion - there is much to ... -
there is no need - there is no question - there is no evidence -

Il n'est pas rare que la structure soit renforcée par un modal :

there can be no doubt - there can be no reason
there can be no room -

Le journalisme apparaît ainsi comme l'art de brouiller les lignes entre la réalité, son expression et une opinion.

Les emplois adverbiaux de ***there*** sont beaucoup moins fréquents dans le corpus (environ un sixième de la totalité des occurrences de ***there***). Un examen de la totalité des contextes de ***there*** adverbial dans le sous-corpus de la presse nous a permis de mettre en évidence une tendance intéressante de l'usage anglais contemporain. ***There*** adverbial tend de plus en plus à occuper la place extrême dans les propositions et les phrases tandis que ***there*** existentiel apparaît nécessairement en tête tous les syntagmes. Cette différence de comportement syntaxique des deux formes ***there***, leurs particularités sémantiques et phonétiques suggère, en dépit de leur origine unique, l'évolution vers une distribution complémentaire au sens le plus strict.

Seul le recours à un ensemble de textes aussi finement codé que le LOB corpus combiné à l'analyse linguistique a pu rendre possibles ces observations.

Je souhaite maintenant (et enfin) montrer une exploitation plus systématique et plus puissante de la notion de base de voisinage réalisée en faisant varier la taille de la base et en recourant à des procédures mathématiques originales permettant de dépasser le stade de la concordance spécialisée. Une base ayant ainsi été définie, la distribution d'un lexème se caractérise par une suite d'adresses correspondant aux divers voisinages. On transforme cette suite d'adresses en un vecteur V appelé ***vecteur***

caractéristique, dont les composantes v_i représentent pour un i donné le nombre d'occurrences du lexème considéré dans une unité significative.

Exemple :

Mot y, base de voisinage : la phrase.

Adresses : 7-15-17-17-19-20-23-54 .....

vecteur caractéristique V_1 :

0000001000000102010010*(30 fois)1 .....

Le comportement d'un ensemble de lexèmes dans la linéarité des énoncés peut s'observer à travers l'étude des ***vecteurs caractéristiques*** correspondants. Ces vecteurs sont porteurs d'informations synthétiques précieuses et peuvent être exploités mathématiquement et linguistiquement de diverses manières, ce que nous illustrerons.

Il est nécessaire de présenter à l'aide de la figure 5 quelques notions avant de faire paraître les résultats de la méthode d'exploration linéaire appliquée au corpus.

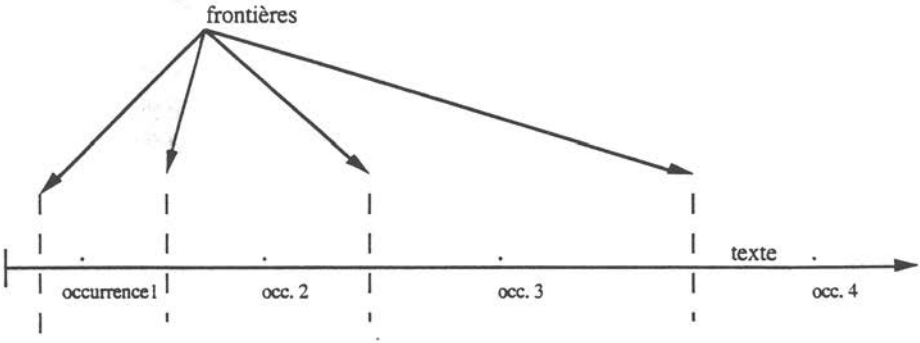

Figure 5 : sphères d'influence

Les points représentent la suite des occurrences du même lexème L1. Chaque occurrence détermine une zone ou sphère d'influence qui s'étend d'une frontière à la suivante, la première étant le début du texte, la suivante le point à mi-chemin entre la première occurrence et la suivante, et ainsi de suite.

Comme ces sphères d'influence sont nécessairement variables, nous avons pris soin de les explorer à l'intérieur d'un voisinage de longueur fixe afin de rendre tous les résultats comparables. Nous avons ainsi choisi une base de voisinage N1 de longueur 30, soit 15 mots avant et 15 mots après la lexie considérée.

Ce choix n'est pas totalement arbitraire, étant donné que dans le sous-corpus utilisé (la presse), la longueur moyenne de la phrase est d'environ 20

mots. Le programme tient évidemment compte des éventuels chevauchements et chaque élément n'est bien sûr compté qu'une fois.

Le déroulement du programme est alors assez simple : le premier pas consiste à explorer le voisinage du lexème à l'intérieur de la base N1, comme définie ci-dessus. Au pas suivant la base de voisinage est étendue de 15 mots de part et d'autre de N1 en excluant évidemment la zone précédemment explorée.

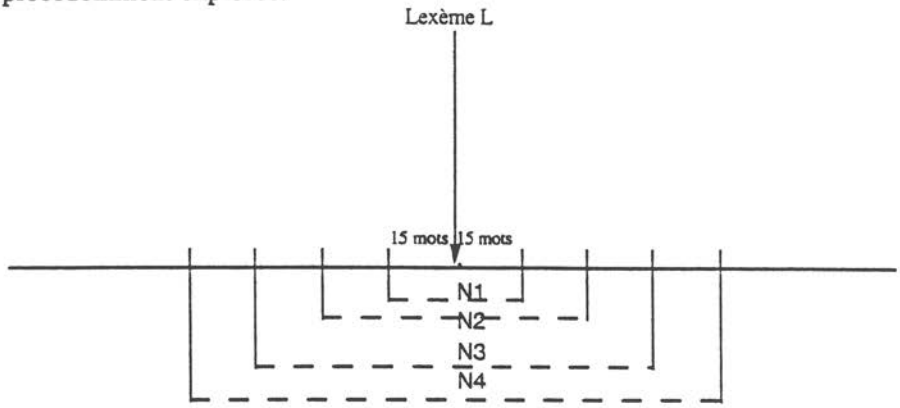

Figure 6 : bases de voisinage

On répète la procédure, si bien que N3 résulte de l'addition de deux tranches symétriques de 15 mots chacune. Nous nous sommes arrêtés à N4.

On peut aussi donner une idée du déroulement de l'algorithme qui sous-tend l'exploration en représentant les bases de voisinage au moyen de cercles concentriques (Figure 7).

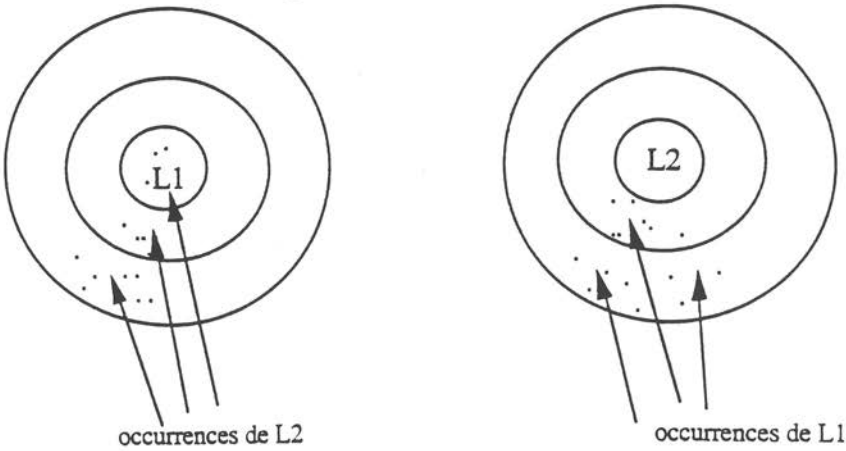

Figure : 7 voisinages successifs

Il est temps maintenant de montrer les résultats obtenus concernant les pronoms personnels et les modaux.

Voici les résultats rendant compte de la proximité des pronoms sujets de première et troisième personne ***I*** et ***he*** dans le sous-corpus des éditoriaux de presse (catégorie B du corpus, environ 25000 mots).

On s'est posé la question suivante :

Le lexème L_1 (***he***, fréquence totale dans B : 216) appartient-il au voisinage du lexème L_2 (***I***, fréquence totale dans B : 76) ?

Voici les résultats pour les bases successives N1 à N4 :

(*I*, 76)	N1	N2	N3	N4
he (216)	20	19	17	15

Tableau 8 : *He* au voisinage de *I*

Ainsi le pronom sujet de troisième personne ***he*** se trouve 20 fois dans le voisinage un(portée totale 30, c.-à-d. 15 de chaque côté) 19 fois dans le voisinage deux (portée 60 moins la zone centrale N1) etc...

Les chiffres correspondant à la question inverse L_2 appartient-il au voisinage de L_1 (he) sont les suivants :

(*He*)	N1	N2	N3	N4
I	18	21	15	10

Tableau 9 : *I* au voisinage de *he*

Bien que d'ordres de grandeur comparables, les chiffres sont différents car les voisinages de L_1 et L_2, bien qu'ayant beaucoup d'éléments en commun, n'appartiennent au même ensemble.

Il est par conséquent intéressant de combiner en les additionnant les résultats des deux explorations (L_1 au voisinage de L_2, L_2 au voisinage de L_1)

	N1	N2	N3	N4
I/he, He/I	38	40	32	25

Tableau 10 : méthode additive

Les très faibles variations entre les zones N1 à N4 sont la contrepartie chiffrée de l'indépendance mutuelle des deux éléments choisis pour ce premier exemple.

Tout autres sont les résultats obtenus (méthode additive) pour le pronom sujet de première personne et les modaux ***shall***, ***should*** et ***can***.

(*I*,76)	N1	N2	N3	N4
shall (6)	4	3	2	1
should (36)	9	0	2	1
can (21)	16	5	5	1

Tableau 11 : *Shall, should, can* au voisinage de *I*

Les résultats, on le voit, sont très différents de ceux de l'exemple précédent, non seulement en valeurs absolues, mais surtout par la concentration des cooccurrences dans la zone de voisinage étroit N1. Ceci reflète bien sûr la plus grande dépendance entre première personne et auxiliaire modal dans les textes d'éditorialistes.

De la même façon, on remarque les bonnes valeurs obtenues par ***will*** et ***would*** au voisinage des pronoms de dialogue ***I*** and ***you*** avec cependant une préférence marquée des modaux pour ***I***..

(*I*,76)	N1	N2	N3	N4
will (65)	17	6	8	10
would (56)	18	9	11	15
(***you***,15)	N1	N2	N3	N4
will (65)	2	2	5	1
would (56)	3	2	0	3

Tableau 12 : *Will* et *would* au voisinage de *I* et *you*

You se distingue en outre de ***I*** par son indifférence relative à ***can***, ***could*** et surtout ***should***.

(***you***,15)	N1	N2	N3	N4
can (21)	4	0	0	2
could (31)	4	1	1	0
should (36)	0	0	0	3

Tableau 13 : *Can, could, should* dans les voisinages de *you*

Ces résultats ne valent provisoirement bien sûr que pour le corpus étudié. Il serait nécessaire d'examiner les contextes eux-même pour isoler, par l'analyse classique, les paramètres proprement linguistiques des facteurs situationnels qui peuvent expliquer le comportement de ces éléments dans ces parages.

Le pronom de troisième personne (***he***) manifeste une présence constante et régulière dans tous les voisinages de ***I***, probablement en raison de l'emploi du style indirect avec un verbe introducteur à la première personne et aussi en raison de son inertie relative qui ne l'éloigne d'aucun autre pronom personel de ces textes.

En réalité, ***he*** se comporte à la manière d'un élément étranger au système, sans tropisme affirmé. On aimerait évoquer ici une contrepartie chiffrée à l'appellation de non-personne si souvent reprochée à Benveniste à propos du statut énonciatif particulier de la troisième personne.

Ces résultats concernant ***I*** et ***he*** sont encore plus éclatants dans la partie reportage du sous-corpus de la presse (section A, 44 textes).

(***I***, 23)	N1	N2	N3	N4
(*he*, 149)	**21**	**13**	**18**	**2**

Tableau 14 : *He* au voisinage de *I* (reportage)

Le pronom neutre sujet de troisième personne ***It*** affiche un comportement assez semblable, tandis que le pronom féminin sujet de troisième personne ***she*** a trop peu d'occurrences (7 au total dans ce sous-corpus) pour permettre des conclusions sûres.

Quant aux pronoms pluriels, ***we*** et surtout ***they***, tous deux manifestent un tropisme marqué pour ***can*** et davantage encore pour ***will*** et ***would*** dans les voisinages serrés (N1) du sous-corpus éditorial (B1).

(***They***, 70)	N1	N2	N3	N4
(*will*, 65)	**14**	**3**	**8**	**5**

(***They***,70)	N1	N2	N3	N4
(*would*, 56)	**12**	**9**	**6**	**4**

Tableau 15 : *Will* et *would* au voisinage de *they*

Les attirances entre modaux eux-même ne manquent pas d'intérêt. ***Can*** a des affinités, purement spatiales peut-être, avec ***will***, ***would*** et, à un moindre degré, avec ***should***.

Could, quant à lui, se sent attiré par ***should***, alors que ***may*** fréquente les voisinages de ***will*** et ***should***.

Cette affinité se retrouve par ailleurs dans les cooccurrences de ***will*** et ***should*** avec ***may***

(***Can*, 28)**	**N1**	**N2**	**N3**	**N4**
(*will*, 57)	**4**	**2**	**4**	**1**
(*would*, 58)	**5**	**8**	**4**	**9**
(*should*, 24)	**1**	**1**	**0**	**2**

(***Could*, 16)**	**N1**	**N2**	**N3**	**N4**
(*might*, 9)	**0**	**0**	**0**	**0**
(*should*, 24)	**4**	**3**	**0**	**0**

(***May*, 12)**	**N1**	**N2**	**N3**	**N4**
(*will*, 57)	**4**	**2**	**5**	**0**
(*should*, 58)	**0**	**2**	**0**	**0**

(***Will*, 57)**	**N1**	**N2**	**N3**	**N4**
(*should*, 24)	**0**	**4**	**6**	**0**
(*may*, 12)	**4**	**2**	**5**	**0**

Tableau 16 : Modaux : attractions mutuelles

Afin d'éprouver la méthode et le corpus, nous avons soumis à l'algorithme d'exploration dynamique deux champs lexicaux, celui de la vie politique et celui de l'enseignement.

Comme on pouvait s'y attendre, la fréquence des éléments était trop basse pour donner lieu à des résultats décisifs et complets. Les seules affinités que nous avons détectées furent entre ***country*** d'une part et ***party***, ***government***, ***election*** et ***majority*** de l'autre...

Si l'on reprend les résultats très complets obtenus pour les modaux et les pronoms personnels, il est possible d'en offrir une image globale synthétique très parlante en fonction de leurs fréquences dans les différentes bases de voisinage successives. Pour ce faire, on construit un index statistique de ces fréquences en pondérant leurs valeurs par des facteurs de 1, 1/2, 1/4, 1/8 respectivement pour les chiffres de N1, N2, N3 et N4.

Nous avons procédé à cette pondération des données afin de privilégier la faible distance (zone N1) en vue d'une représentation topologique.

On peut alors dresser un tableau de proximités, matériau de choix pour l'analyse arborée et l'analyse multidimensionnelle.

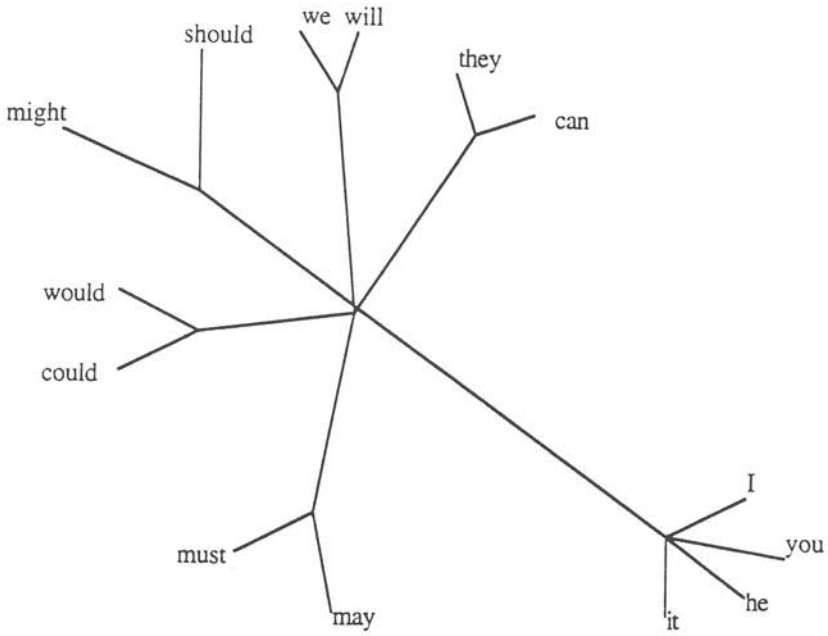

Figure 17 : Analyse arborée des modaux et des pronoms

Cette représentation arborée confirme et complète les résultats présentés analytiquement de l'exploration du texte par la méthode nouvelle des bases de voisinage.

Nous avons ici une représentation plus globale, synthétique, plus puissante et plus élégante aussi des données, où les notions de ***proximité***, d'***opposition*** et de ***groupement*** sont également significatives.

On voit de la sorte se regrouper les pronoms de dialogues (***I***, ***you***) et les pronoms singuliers de troisième personne dans la partie droite de l'arbre, ce qui signifie qu'ils partagent entre eux plus d'affinités textuelles qu'avec tout autre élément des données, que ce soit un modal ou un autre pronom personnel.

Les formes passées de modaux se regroupent en couples, ***would-could***, ***might-should***, tandis que ***must*** et ***may*** se retrouvent associés, probablement davantage en raison de leur caractère partagé de basse fréquence, que de leur valeur commune déontique ou épistémique.

On notera aussi avec intérêt les affinités des occurrences nombreuses du présent de ***will*** and ***can***, non avec d'autres modaux, mais avec leurs sujets favoris dans le corpus, les pronoms ***we*** et ***they***.

Nous avons enfin soumis ces mêmes données pondérées et le tableau de similitudes correspondant à l'***analyse multidimensionnelle***, en utilisant les contraintes du moindre carré dans un espace à deux dimensions.

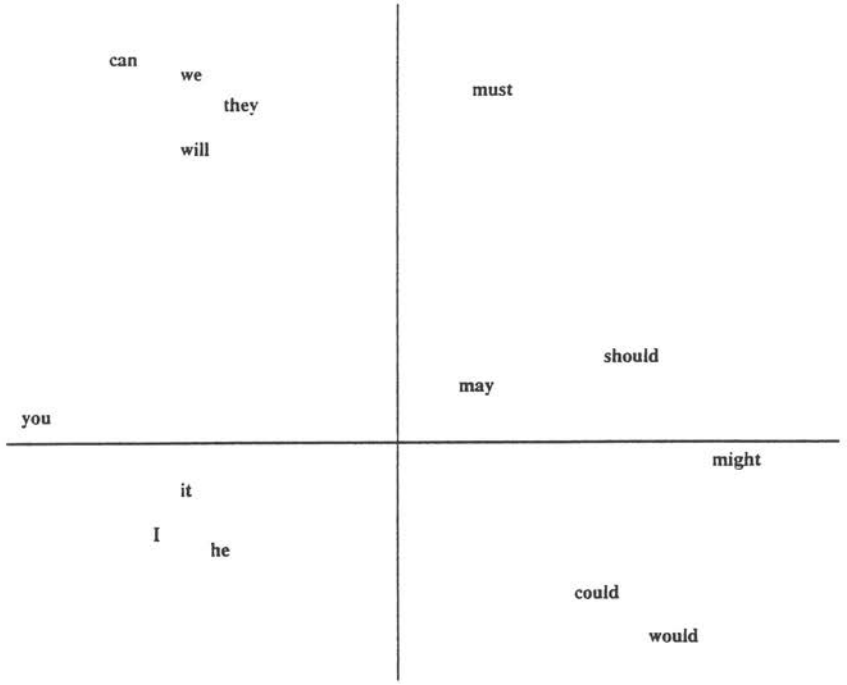

Figure 18 : Analyse multidimensionnelle des modaux et des pronoms

L'information acquise au terme de l'analyse arborée se trouve confirmée, même si elle peut-être moins facile à percevoir pour l'oeil profane.

L'axe vertical oppose d'une part l'ensemble de tous les pronoms et les modaux ***can*** et ***will*** au reste des données d'autre part.

L'axe horizontal oppose ou sépare les pronoms personnels singuliers (***I***, ***it***, ***he***) et les modaux les plus fréquents du passé (***could***, ***would***, ***might***) au reste des données. Tandis que la proximité de ***may*** à l'intersection des axes reflète son rôle marginal en anglais contemporain et donc ses faibles tropismes.

Cette nouvelle méthode d'exploration du texte nous a permis de dépasser les approches statistiques traditionnelles et de rendre compte pleinement de la nature linéaire et dynamique du texte ainsi que des solidarités entre éléments, qui fondent l'ordre des mots, cette "entité abstraite", selon Saussure, "mais qui ne doit son existence qu'aux unités concrètes qui la contiennent et qui courent sur une seule dimension". L'apport combiné de la topologie et de la linguistique appliquée aux corpus fait ainsi écho aux paroles de Saussure en nous rappelant l'importance primordiale des données et des textes, qui préexistent à toute théorie.

BIBLIOGRAPHIE

BOLINGER, Dw. (1980), *Language - The loaded weapon*, London, Longman.

DELEUZE, G. (1968), *Différence et répétition*, Paris, P.U.F..

HALLIDAY, M.A.K. (1985), *An Introduction to Functional Grammar*, London, Arnold.

HOFLAND, K., & JOHANSSON, S. (1982), *Word Frequencies in British and American English*, Norwegian Computing Centre for the Humanities, Bergen.

HUDDLESTON, R. (1984), *Introduction to the Grammar of English*, Cambridge, C.U.P.

JOHANSSON, S., ATWELL, E., GARSIDE, R., & LEECH, G.(1986), *The tagged LOB Corpus, Users' manual*, Norwegian Computing Centre for the Humanities, Bergen.

JUILLARD (1994), M., "Regard quantitatif sur la coordination", Strasbourg, *RANAM*, XXVII.

JUILLARD, M., & LUONG, N.X.(1989), 'Unrooted Trees Revisited : Topology and Poetic Data', *Computers and the Humanities*, 23, pp. 215-223.

JUILLARD, M., & LUONG, N.X.(1997), Words in the hood : a new look at the distribution of words in texts, *Literary and Linguistic Computing*, 12, 2, à paraître.

LAFON, P. (1981), "Statistiques des localisations des formes d'un texte", *Mots*, 2, pp.157-187.

LUONG, N.X.(1988) Méthodes d'analyse arborée. Algorithmes, applications, doctorat d'Etat, Université Paris V René Descartes.

QUIRK, R., et al. (1985), *A Comprehensive grammar of the English language*, London, Longman.

QUIRK, R., et al. (1985), *A Comprehensive grammar of the English language*, London, Longman.

SAUSSURE, F. De (1915), *Cours de Linguistique Générale*, Paris, Payot;

SERANT, D., THOIRON, Ph. (1988) "Topographie des formes répétées", Le nombre et le texte : Hommage à Etienne Evrard, *Revue Informatique et Statistique dans les Sciences humaines*, 24, pp. 333-343.

SERANT, D., THOIRON, Ph. (1988) "Topographie des formes répétées", Le nombre et le texte : Hommage à Etienne Evrard, *Revue Informatique et Statistique dans les Sciences humaines*, 24, pp. 333-343. BOLINGER, Dw. (1980), *Language - The loaded weapon*, London, Longman.