Surface et Profondeur : des arbres pour éclairer la forêt

Juillard Michel

Pour citer cet article
Juillard Michel, « Surface et Profondeur : des arbres pour éclairer la forêt », *Cycnos*, vol. 8. (Apparences textuelles et réalité linguistique), 2008, mis en ligne en juillet 2008.
http://epi-revel.univ-cotedazur.fr/publication/item/393

Lien vers la notice http://epi-revel.univ-cotedazur.fr/publication/item/393
Lien du document http://epi-revel.univ-cotedazur.fr/cycnos/393.pdf

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

AVERTISSEMENT

Les publications déposées sur la plate-forme épi-revel sont protégées par les dispositions générales du Code de la propriété intellectuelle. Conditions d'utilisation : respect du droit d'auteur et de la propriété intellectuelle.

L'accès aux références bibliographiques, au texte intégral, aux outils de recherche, au feuilletage de l'ensemble des revues est libre, cependant article, recension et autre contribution sont couvertes par le droit d'auteur et sont la propriété de leurs auteurs. Les utilisateurs doivent toujours associer à toute unité documentaire les éléments bibliographiques permettant de l'identifier correctement, notamment toujours faire mention du nom de l'auteur, du titre de l'article, de la revue et du site épi-revel. Ces mentions apparaissent sur la page de garde des documents sauvegardés ou imprimés par les utilisateurs. L'université Côte d'Azur est l'éditeur du portail épi-revel et à ce titre détient la propriété intellectuelle et les droits d'exploitation du site. L'exploitation du site à des fins commerciales ou publicitaires est interdite ainsi que toute diffusion massive du contenu ou modification des données sans l'accord des auteurs et de l'équipe d'épi-revel.

EPI-REVEL
Revues électroniques de l'Université Côte d'Azur

Surface et Profondeur : des arbres pour éclairer la forêt

Michel Juillard

CNRS (URL 9) et Université de Nice-Sophia Antipolis

L'une des tâches essentielles de la linguistique actuelle est d'éclairer, par delà le foisonnement des énoncés réalisés, répétés et différents, multiples et disparates en apparence, les modes de fonctionnement, rarement explicites, de l'activité langagière. Quel que soit le nom donné aux concepts sous-tendant la démarche, il convient toujours de passer du plan de l'observable, du donné objectif, à une construction théorique abstraite, voire peut-être subjective, par définition. Il importe de mettre en rapport tour à tour langue et parole pour Saussure, compétence et performance dans la grammaire générative transformationnelle de Chomsky, système et procès selon la glossématique de Hjelmslev. Cela constitue même, si l'on en croit un linguiste contemporain, G. Garnier, l'un des acquis sûrs de la linguistique, tant il est vrai, en dépit des nombreuses interrogations et des conflits théoriques animant les diverses écoles, que l'énoncé terminal, le dictum, la surface, ne saurait être le lieu où se manifestent les secrets du fonctionnement du langage[1]. C'est aussi ce que rappelait déjà Gustave Guillaume en opposant langue et discours :

> "La différence d'ordre est essentiellement que le discours a pour matière -pour matière d'expression- le pensé, et la langue pour matière -pour matière de représentation- le pensable. Le discours est une forme prise, pour expression, par le pensé ; la langue une forme prise, pour représentation, par le pensable."[2]

La linguistique quantitative, parce qu'elle prend en compte à travers ses dénombrements exhaustifs l'intégralité des énoncés, réclame, plus que toute autre saisie de la réalité langagière immédiate, des procédures heuristiques pour dépasser le stade de la simple collecte des faits de surface, quel qu'en soit par ailleurs l'intérêt intrinsèque, et accéder au système qui les sous-tend. Ce système, selon Saussure, ne recèle que des différences alors que la parole, pourrait-on dire, est surtout faite de répétitions. L'analyse arborée, avec la représentation à laquelle elle aboutit, constitue une méthode à la fois puissante et subtile pour ordonner la forêt des phénomènes de surface et les structurer de façon lumineuse. On se propose de décrire brièvement les bases théoriques de ce nouveau type de représentation et d'exploration des faits linguistiques puis d'illustrer son application à des données réelles.

L'utilisation de structures arborées pour représenter l'organisation d'un domaine conceptuel, d'un champ de connaissances ne date pas d'aujourd'hui, ni même d'hier. Le philosophe néoplatonicien Porphyre (233-301) représentait déjà sous cette forme (fig. 1) la hiérarchie des concepts allant de la substance à l'homme en passant par une série de choix binaires à condition d'emprunter toujours la branche de gauche[3].

Figure 1

[1] Garnier, G., Linguistique et traduction, Paradigme, Caen, 1985, p. 2.

[2] Guillaume, G., Principes de Linguistique théorique, Québec, Presses Universitaires Laval & Paris, Klincksieck, 1973, p. 160.

[3] L'illustration est empruntée à J.-P. Barthélemy, "Les arbres, de la matière pour bricoler...", CUMFID, CNRS URL 9, Nice, 1989, 16, p. 5.

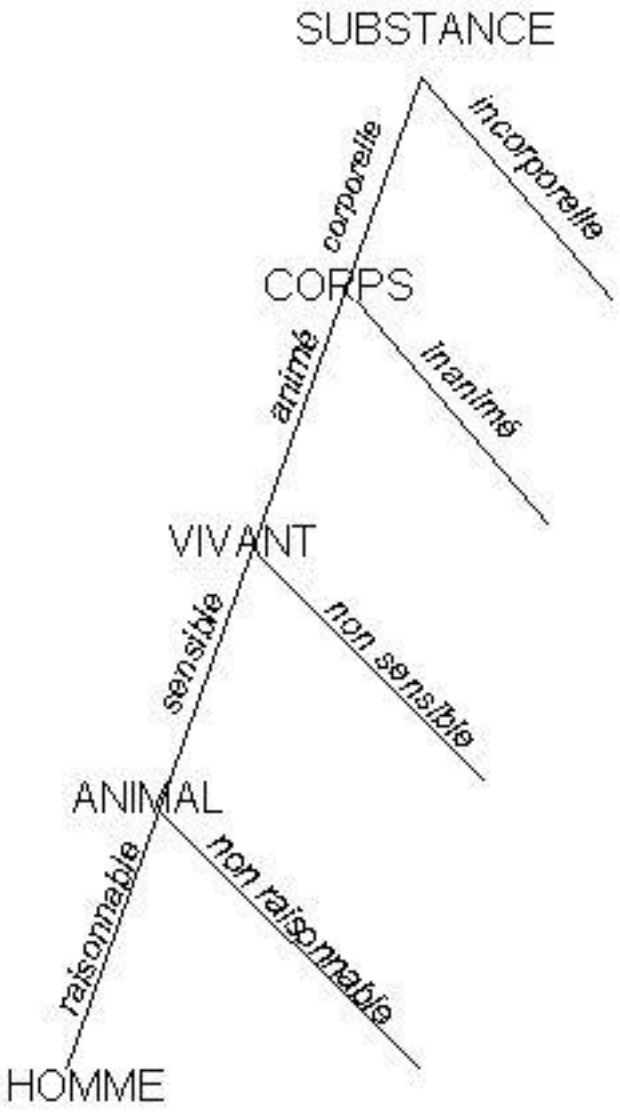

Figure 1

Les grands domaines de la taxinomie font depuis longtemps appel à des représentations arborées : les arbres généalogiques et les arbres décrivant la classification des espèces animales ou végétales. Si les arbres de Chomsky ont fait le tour du monde, on connaît peut-être moins les schémas arborés ou stemmas au moyen desquels le linguiste français Louis Tesnière eut l'idée de rendre compte de la hiérarchie des éléments d'une phrase (Fig. 2).

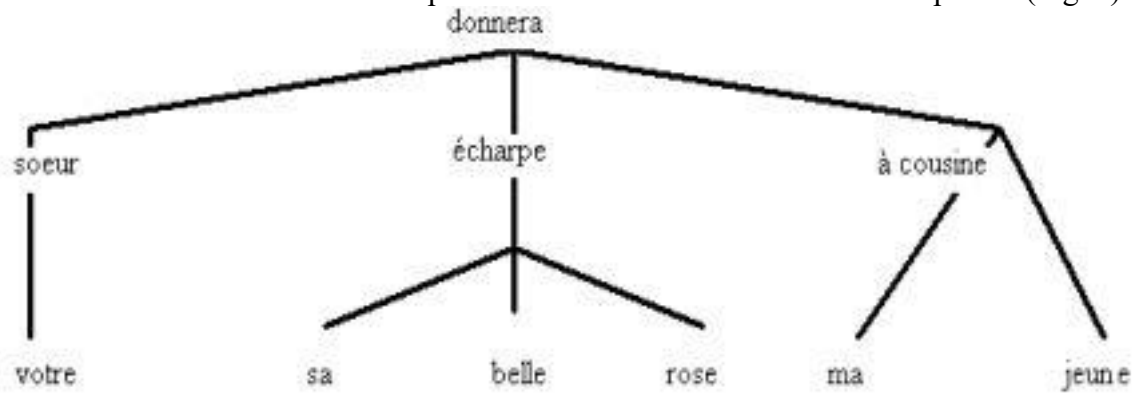

Figure 2

Les scientifiques contemporains quant à eux ont recours à des schémas arborescents pour rendre explicites la structure et l'organisation des données dans les réseaux de communication.

Lorsqu'il s'agit essentiellement de constituer des classes à partir des objets étudiés, le modèle le plus souvent utilisé a recours aux arbres dits plantés puisque caractérisés par une racine à partir de laquelle on atteint, en suivant les branches ou arêtes, les sommets terminaux ou feuilles en passant par des sommets intermédiaires (figure 3). Ces dendrogrammes ou arbres ultramétriques ont les propriétés suivantes :

Toutes les feuilles sont équidistantes de la racine.

Tous les sommets d'une classe donnée, c'est-à-dire les sommets dominés par un même noeud, sont équidistants entre eux.

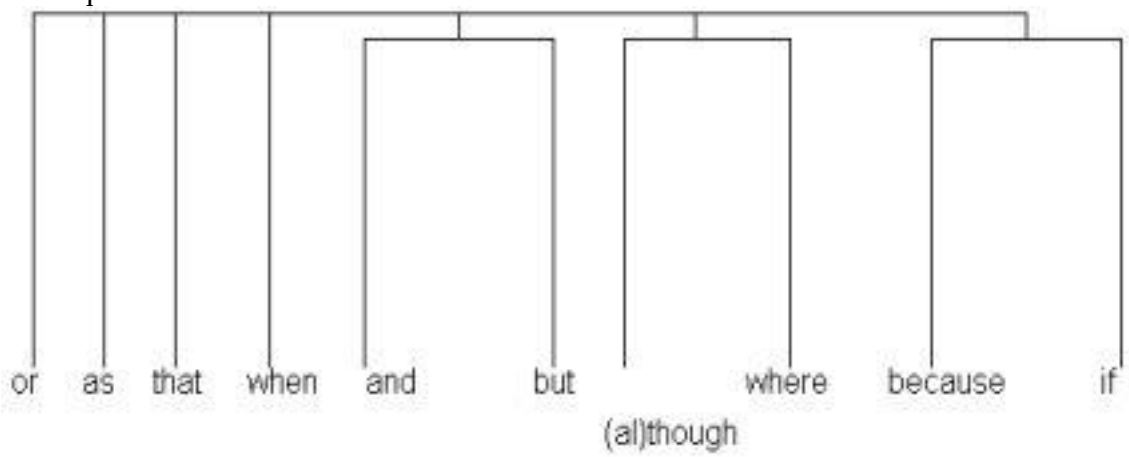

Figure 3

Figure 3

La distance d entre les feuilles vérifie alors l'inégalité ultramétrique suivante: pour toutes les feuilles x, y, z de l'ensemble X on a :

$$d(x,y) \leq \text{Max } \{d(x,y), d(y,z)\}$$

Le modèle ultramétrique présente au moins deux avantages :

- a) la notion de classe y est définie sans ambiguïté :

on passe de classe en classe, du plus grossier groupement aux plus fines associations en parcourant l'arbre de nœud en nœud, de la racine aux feuilles.

- b) la distance d'une feuille à un nœud constitue un indice du niveau de formation d'une classe.

C'est précisément cette notion rigide de classe qui, pour certaines applications, fait préférer au modèle ultramétrique le modèle général de l'analyse arborée, plus propre à rendre compte de la ressemblance, de l'air de famille (Wittgenstein, 1953), qu'à enfermer les objets dans des compartiments étanches[4].

L'arbre produit en utilisant cette méthode offre alors non plus une structure hiérarchique (c'est-à-dire plantée), mais se présente sous la forme d'un arbre quelconque, non orienté, un arbre dit non-planté semblable aux arbres phylogénétiques des naturalistes. Deux sommets pris au hasard sont reliés par un chemin et un seul. On distinguera les sommets terminaux ou "feuilles" de l'arbre des sommets intérieurs ou "noeuds". Les feuilles de l'arbre (et non les sommets intérieurs) correspondent aux objets étudiés. Le rapport entre proximité (critère de distance) et classification (critère de structure) institué par ce modèle de représentation fournit davantage d'information, même si cette dernière se paye parfois d'une certaine ambiguïté.

[4] Wittgenstein, L., Philosophische Untersuchungen, oxford, Blackwell, 1953.

Si l'on considère les éléments (ou feuilles) de l'ensemble X (matrice de dissimilarités obtenue à partir du tableau chiffré des données), pour toutes les feuilles x, y, z, t de X on aura :

$$d(x,y)+d(z,t) \leq \text{Max } \{ d(x,t)+d(y,z), d(x,z)+d(y,t) \}$$

C'est une condition nécessaire et suffisante pour qu'un ensemble de dissimilarités soit représentable par un arbre. On la connaît sous le nom de condition de Buneman et de Dobson, deux mathématiciens contemporains responsables d'avancées décisives dans le domaine de la topologie[5].

La démarche qui sous-tend l'algorithme fait appel aux trois notions simples fondamentales d'*opposition*, de *voisinage* et de *groupement.*

- Première notion : opposition :

sur un arbre donné (fig. 4), la paire (x,t) est opposée à une paire (u,v) si (x,t) se trouve sur une branche ne contenant pas (u,v).

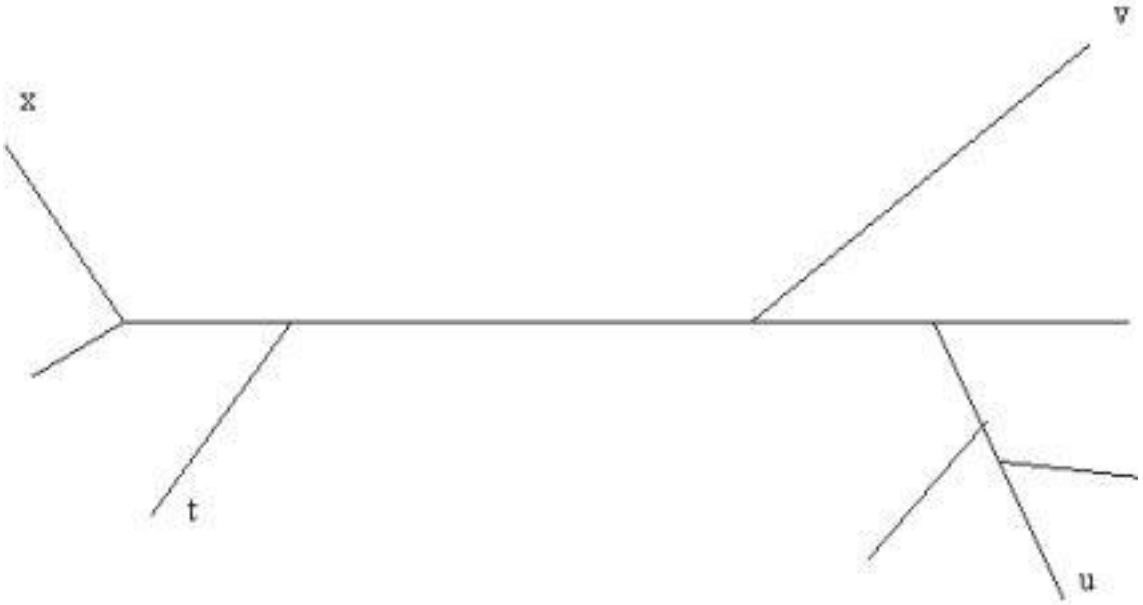

Figure 4

- Deuxième notion : voisinage :

x et y sont voisins (fig. 5) s'ils appartiennent à l'ensemble A tel que toute paire opposée à (x,y) se trouve dans le complémentaire de A, c'est-à-dire dans X-A.[6]

[5] Topologie : domaine de la mathématique consacré à l'étude des propriétés qualitatives et des positions relatives des êtres géométriques, abstraction faite de leur forme et de leur grandeur.

[6] Cette partie de notre étude est le fruit d'une collaboration avec un mathématicien chercheur de l'URL 9, spécialiste de topologie, X N. Luong, que nous remercions ici de son apport. Les résultats de nos travaux initiaux ont été présentés pour la première fois à l'Université de Liège (Congrès International "Le nombre et le texte", mai 1987) et à l'Université de Göteborg à l'occasion du XIVème Congrès International de l'ALLC (Association for Literary and Linguistic Computing) en juin1987.

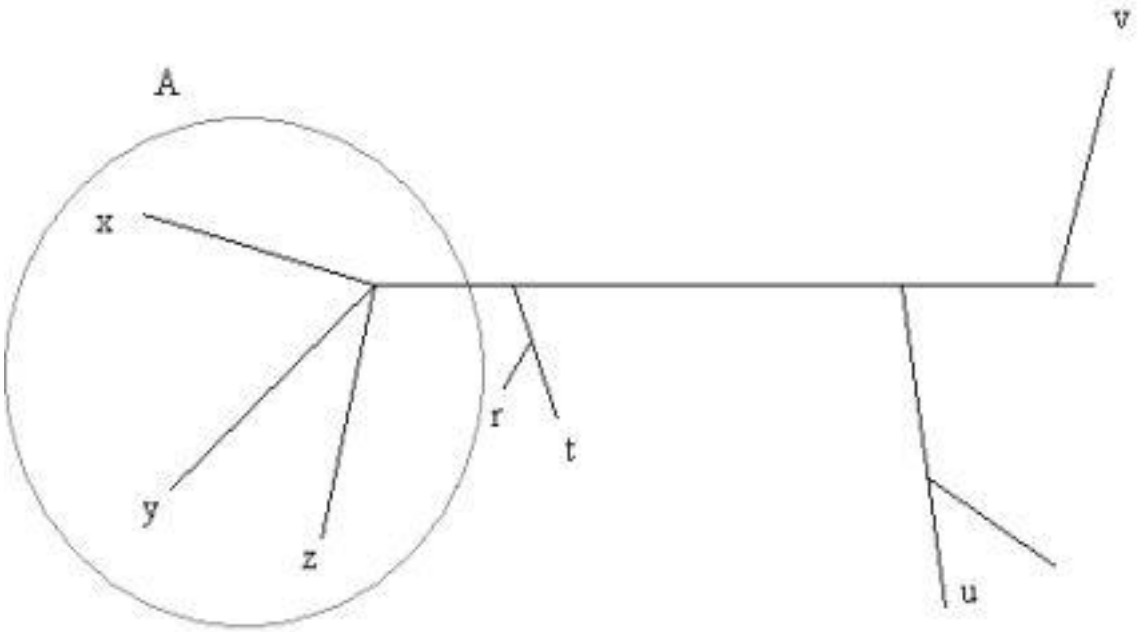

Figure 5

- Troisième notion : groupement :

On appelle groupement un ensemble d'objets considérés comme voisins selon les deux notions précédentes.
Ainsi, dans la figure 6, les élements a, x et y forment un groupement. Il en va de même pour e, b et pour c, f.

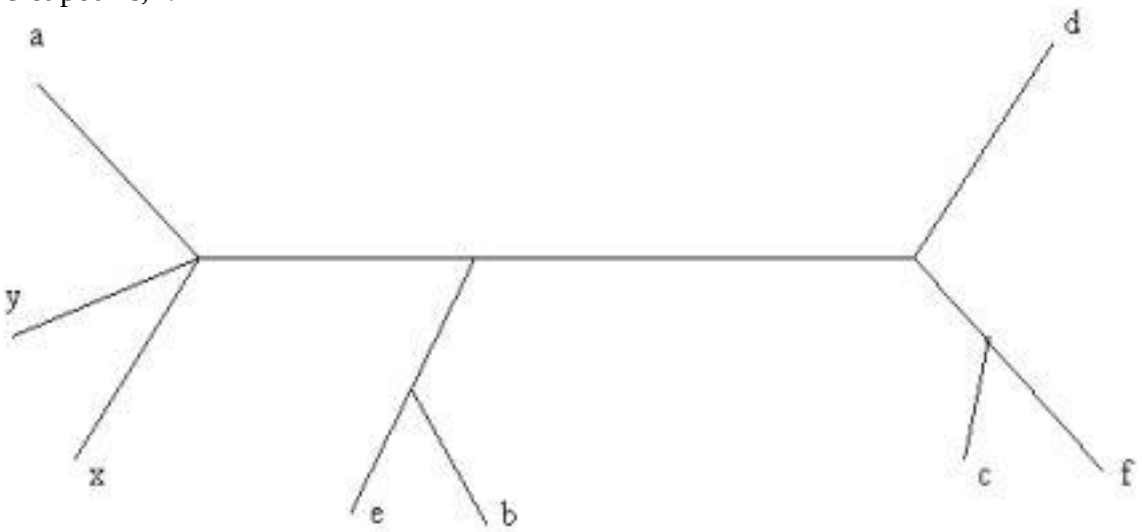

Figure 6

Construire un arbre revient donc, au terme d'une procédure d'approximation prenant comme point de départ le tableau de dissimilarités (données chiffrées), à construire un tableau de distances. A chaque pas d'itération, on réunit les éléments qui sont proches d'un point de vue topologique, c'est-à-dire appartenant à un groupement.
L'algorithme se déroule en 3 étapes :

- 1) mise en évidence des groupements et calcul de la longueur de leurs arcs.
- 2) identification de chaque groupement par un "objet moyen".

- 3) répétition du processus appliqué aux objets moyens et aux objets restants jusqu'à obtention de l'arbre complet[7].

Quelques exemples d'applications à des données linguistiques :

Pour mettre à l'épreuve la méthode et illustrer son fonctionnement, nous l'avons appliquée à un corpus de treize textes de poésie anglaise contemporaine, soigneusement dépouillés et codés syntaxiquement, totalisant 100 000 mots (Day Lewis 1929-1970). Ce corpus a déjà été partiellement analysé au moyen d'autres procédures statistiques plus traditionnelles (Juillard 1983, 1985, 1989).

On a constitué à partir des données du corpus des matrices d'occurrences, c'est-à-dire des tableaux rectangulaires croisant colonnes (propriétés ou objets) et lignes (objets ou propriétés). Dans ce premier exemple (fig. 7), les colonnes du tableau représentent la distribution des 100 000 occurrences des diverses catégories grammaticales (noms, verbes, adjectifs, adverbes descriptifs, auxiliaires et modaux, déterminants, pronoms, adverbes autres que descriptifs, prépositions et particules, conjonctions de subordination, conjonctions de coordination, interjections). Les lignes correspondent aux 13 textes du corpus ordonnés chronologiquement.

TxtCat	N	V	ADJ	ADD	AUXM	DET	PRON	AND	PREPA	SUB	COOR	INT
TP	1651	1051	458	35	198	1025	315	228	842	299	375	2
FFI	1222	902	385	23	135	716	155	172	643	173	212	4
TMM	2123	1525	586	33	312	1196	488	321	1038	287	397	25
ATTD	1646	1269	500	42	145	1115	356	210	920	273	340	5
NATW	350	233	135	12	19	248	55	53	215	67	75	0
OTD	1879	1358	678	35	175	1355	441	269	1038	305	389	15
WOA	1530	1051	473	51	144	1025	333	200	799	295	307	12
43-47	2301	1806	825	83	262	1716	672	396	1325	456	526	44
AIV	3075	1980	1034	163	284	2013	729	462	1721	520	608	18
PEG	2173	1689	812	108	226	1608	641	404	1177	433	467	20
GATE	2484	1812	867	101	336	1651	813	369	1398	461	463	39
ROOM	1687	1240	634	56	157	1145	470	247	939	275	361	9
TWR	1697	1058	645	57	145	1055	344	240	850	207	265	7

Figure 7

Tel quel, ce tableau peut donner lieu à divers tests statistiques (écarts réduits, Khi-deux) portant sur chacun de ses éléments envisagé isolément, sur une ligne ou une colonne, sur la comparaison entre deux colonnes ou deux lignes (corrélation), mais jamais sur la totalité.

L'intérêt du logiciel présenté ici est de rendre compte globalement du comportement dans les textes de tous les objets du tableau au moyen d'un arbre. Cet arbre représente les affinités de ces objets, ici des catégories grammaticales, en fonction de leurs propriétés, c'est-à-dire de leur appartenance à tel ou tel texte du corpus.

Concrètement, le premier stade du travail consiste à saisir au clavier le tableau de données pour le transformer en un tableau de dissimilarités[8]. L'algorithme fait appel à la distance du Khi-deux, méthode de calcul introduite par J.P.Benzécri[9], pour produire un tableau de distances auquel on applique l'analyse arborée ; c'est-à-dire que l'on détermine la structure de la représentation en définissant les sommets et les longueurs des arêtes qui donneront sa physionomie, son port caractéristique à l'arbre achevé, à l'issue de la dernière étape de l'algorithme de représentation.

---

[7] Les mathématiciens trouveront un exposé plus détaillé de la théorie de la représentation arborée dans : Juillard, M. & Luong, X. , "Unrooted Trees Revisited : Topology and Poetic Data", Computers and the Humanities , 23, 1989, pp. 215-225.

[8] Le logiciel existe en deux versions : Apple Macintosh ou IBM PC et compatibles.

[9] Benzécri, J. P. et al., L'analyse des données, 2 volumes, Paris, Dunod, 1973.

Figure 8

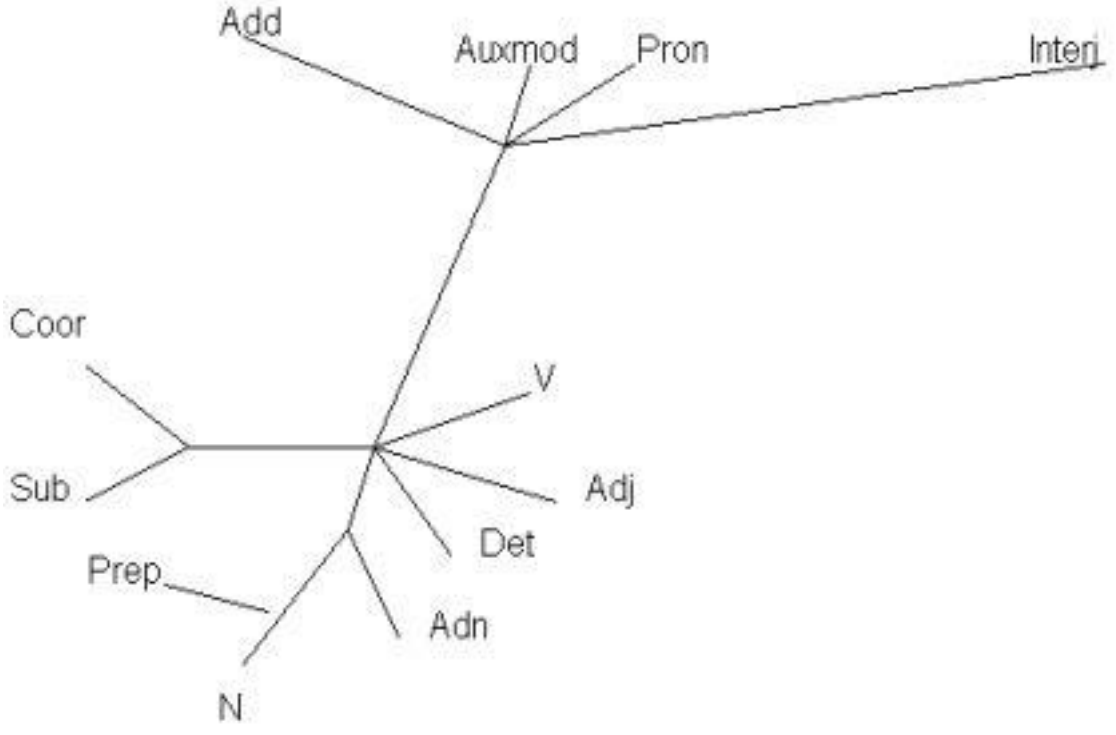

Figure 8

La figure 8 présente la version définitive de l'arbre obtenue en associant le logiciel à une imprimante à laser[10]. Il est aussi possible de tracer l'arbre au moyen d'une imprimante plus sommaire et même de reprendre le dessin à l'aide d'un logiciel de graphisme. La tâche est aisée et le résultat fidèle aux calculs puisque la méthode fournit au chercheur la liste complète des arêtes (c'est-à-dire des branches ou rameaux feuillus de l'arbre) avec leur sommet et leur longueur.

Les catégories grammaticales devenues les feuilles de la partie supérieure de l'arbre sont assez distantes les unes des autres. L'écart le plus grand distingue les adverbes descriptifs et les interjections (longueur des arêtes respectives : 46,29 et 24,96). Dans ce groupe, il faut remarquer l'affinité plus appuyée entre les modaux et les auxiliaires d'une part et les pronoms personnels d'autre part (longueur des arêtes respectives : 12,79 et 7,14).

Le linguiste rapprochera avec intérêt cet aspect de l'arbre et les variations de sens affectant les modaux selon la personne à laquelle ils sont conjugués à la surface du discours. Le chercheur curieux se reportera aux textes et constatera en outre que les modaux, associés à tel ou tel pronom personnel font souvent partie du vocabulaire caractéristique (Juillard 1983, 1986).

Le bas du graphique présente des couples très unis : la subordination et la coordination, le nom et la préposition. Les déterminants, bien qu'associés directement aux adverbes non descriptifs, ne sont pas très loin du nom; les adjectifs non plus, même si le verbe est leur partenaire favori.

La méthode de l'analyse arborée appliquée à des données linguistiques offre la possibilité précieuse de soumettre à l'algorithme de représentation un sous-ensemble de la population totale. L'arbre de la figure 9 illustre ainsi le comportement au sein du corpus considéré des seules conjonctions de coordination et de subordination. On distingue aisément trois grands groupes très nettement individualisés.

---

[10] Cet arbre ainsi que tous ceux de cet article ont été produits en utilisant la nouvelle version, compilée, du logiciel d'analyse des données.

Figure 9

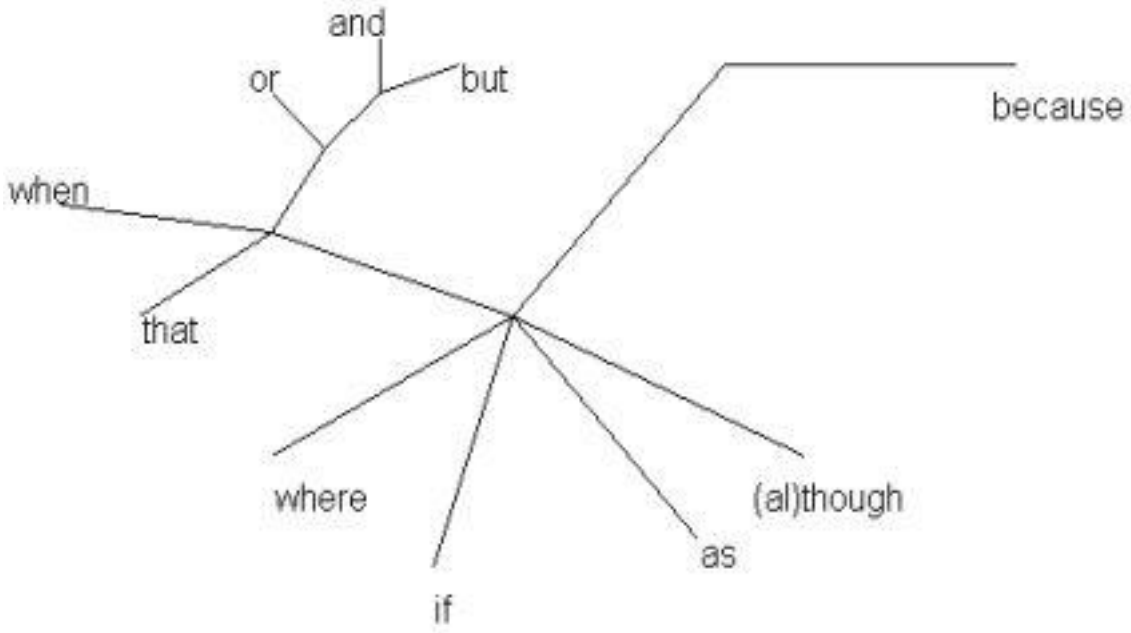

Figure 9

Les rameaux les plus serrés regroupent les principales conjonctions de coordination (*and*, *but* et *or*) au subordonnant à tout faire *that*, lui-même rapproché de *when*. Le deuxième groupe réunit, mais à distance respectable, *where*, *if*, *as* et *(al)though*. Le dernier ensemble ne comporte qu'un élément, s'opposant comme un long gourmand de pure logique discursive aux autres entités grammaticales, plus essentielles à la poésie et au parler ordinaire, poussées à partir de la même racine.

Pour illustrer la souplesse de cette technique nouvelle de représentation des données, nous l'avons appliquée non plus, comme précédemment, à l'ensemble d'une catégorie grammaticale mais à la seule population des pronoms personnels (figure 10).

Figure 10

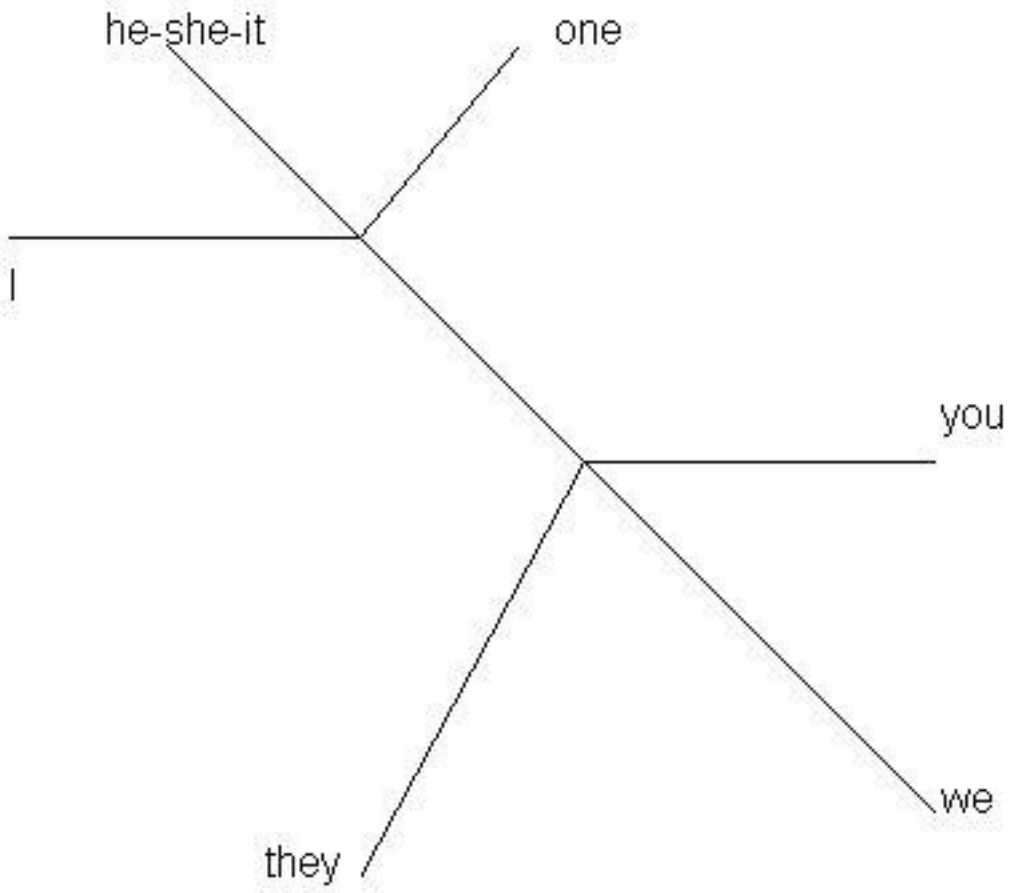

Figure 10

On saisit au premier coup d'oeil l'opposition péremptoire entre les deux personnes fondamentales du discours, en particulier poétique, *I* et *you*. L'une est portée par le rameau horizontal de gauche, l'autre par le rameau homologue de droite, à la manière des feuilles que le botaniste décrit comme alternées. L'autre grande dichotomie qui structure cet arbre très limpide oppose le haut et le bas de la figure, c'est-à-dire le singulier et le pluriel, même si pluriel grammatical ne signifie pas toujours pluralité arithmétique.

Si l'on dissocie les pronoms de troisième personne, on obtient un nouvel arbre et des informations complémentaires (figure 11) : la spécificité de *I* éclate en pleine lumière : seuls s'approchent de cet élément unique des entités neutres ou impersonnelles (*it* et *one*). comme dans la figure précédente, *I* et *you* appartiennent à des ensembles opposés.

Figure 11

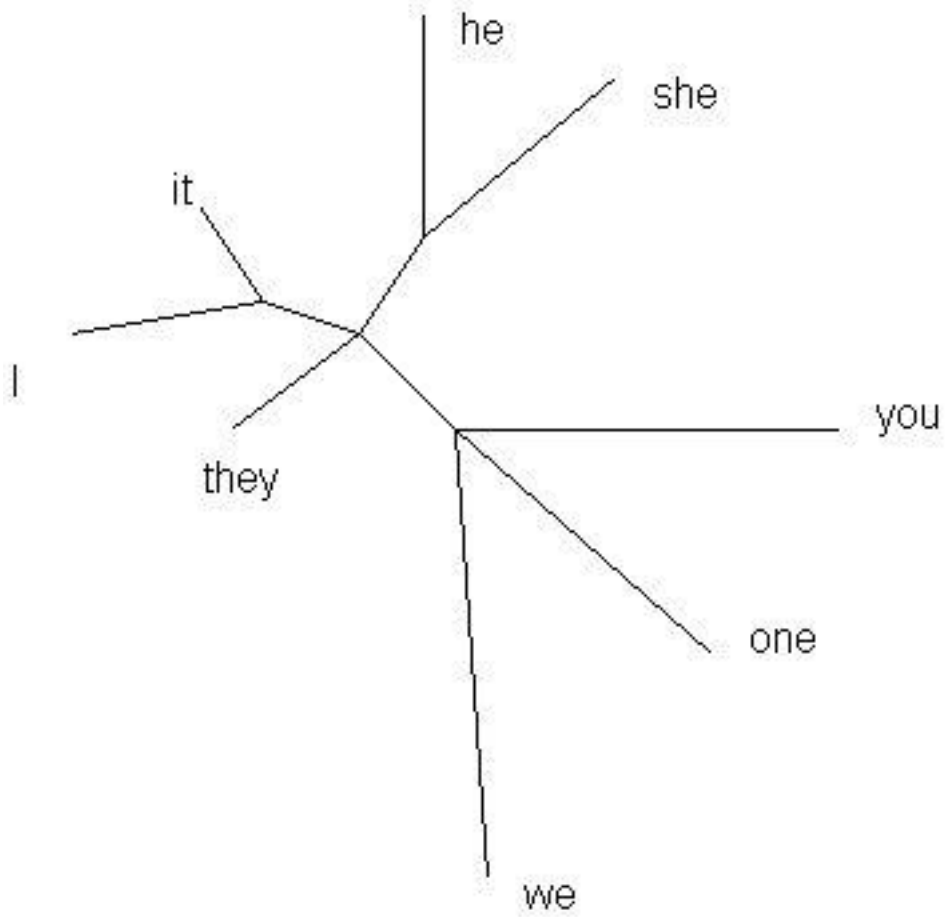

Figure 11

On peut voir dans cette représentation comme l'équivalent et la confirmation objective de l'analyse élaborée par Emile Benveniste de leur unicité et de leur complémentarité :

> Quand je sors de "moi" pour établir une relation vivante avec un être, je rencontre ou je pose nécessairement un "tu", qui est, hors de moi, la seule "personne" imaginable. Ces qualités d'intériorité et de transcendance appartiennent en propre au "je" et s'inversent en "tu". On pourra donc définir le "tu" comme la *personne non-subjective* , en face de la *personne subjective* que "je" représente ; et ces deux "personnes" s'opposeront ensemble à la forme de la non-personne (= "il").[11]

Avec le pronom, le verbe est la seule variété de mots qui se trouve soumise à la catégorie de la personne.[12] Aussi s'imposait l'étude des entités les plus fréquentes par lesquelles l'homme s'inscrit dans la langue en fonction de la modalité et de la polarité, du positif au négatif et du possible au nécessaire.[13] Pour un modal donné par exemple, on sait que le contenu sémantique variera en fonction des rapports entre sujet et prédicat, locuteur et énoncé, sujet grammatical et locuteur. En dépit de tous ces paramètres, l'arbre obtenu à partir d'un substantiel tableau d'occurrences comportant 13 lignes et 13 colonnes (fig. 12) est d'une lumineuse lisibilité. Comme toujours, nous nous contenterons d'esquisser les grandes lignes de sa lecture et son interprétation.

Figure 12

[11] Benveniste, E., Problèmes de linguistique générale, Paris, Gallimard, 1966, p. 232.
[12] Benveniste, E., ibid., p. 224.
[13] Halliday, M.A.K., An Introduction to Functional Grammar, London, E. Arnold, 1985, pp. 85-89.

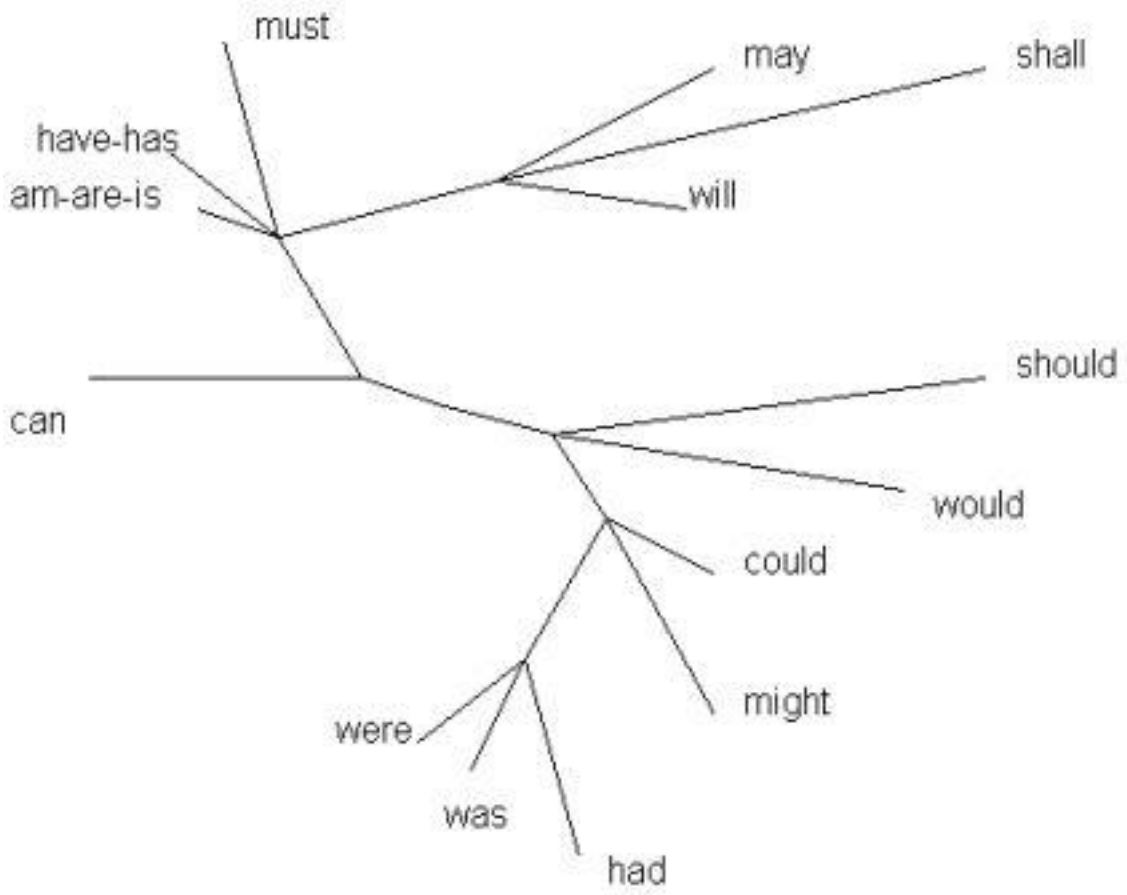

Figure 12

Parcouru dans le sens vertical, de haut en bas ou de bas en haut, l'arbre oppose nettement les deux temps qui structurent le système verbal de l'anglais, le présent et le passé. La lecture dans l'autre dimension, de gauche à droite ou de droite à gauche, fait ressortir la solidarité de *be* et *have* par rapport aux modaux de stricte obédience, quel que soit le temps auxquels ils sont employés. Parmi les modaux, il faut souligner la distance importante qui sépare *must* et *can* et reflète la forte spécificité de *can* dans l'expression de la modalité dynamique et de *must* dans ses emplois déontiques. La spécialisation de *may* dans des rôles épistémiques le range aux côtés de *shall* et *will*, les principaux marqueurs du futur. Les emplois spécifiques de *should* et *would*, leur rôle dans la surmodalité, leur association fréquente au *present perfect* expliquent qu'ils se trouvent sur des rameaux rapprochés, à mi-chemin entre le présent et le prétérit.

Il était tentant d'ajouter aux données précédentes sur les auxiliaires et les modaux, les cinq colonnes des effectifs des pronoms personnels (*I, you, he-she-it, we, they*), afin d'aller plus loin dans la saisie objective de l'activité langagière à partir de ses manifestations à la surface du discours. Il est en effet primordial de chercher à mettre en évidence scientifiquement les forces qui orientent et structurent la parole, la performance, et reflètent des tropismes inhérents au système. C'est donc un nouveau tableau de treize lignes et dix-huit colonnes qui a servi à produire l'arbre de la figure 13.

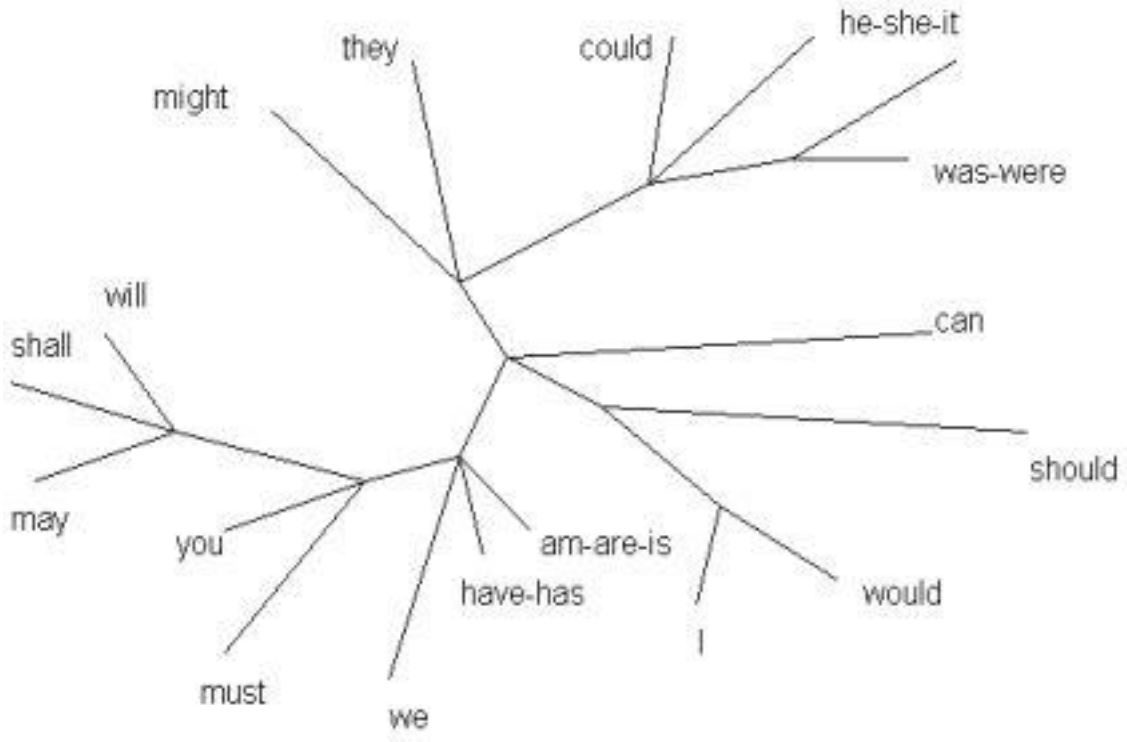

Figure 13

Là encore, le tracé du schéma, le port de l'arbre corroborent et prolongent les acquis fondamentaux de la linguistique du discours et éclairent singulièrement les données brutes initiales. La figure reproduit l'information déjà portée par l'arbre précédent (oppositions présent-passé, polarité-modalité) et met en pleine lumière les oppositions fondamentales entre d'une part le locuteur et l'interlocuteur et l'intérieur et l'extérieur de la sphère allocutive d'autre part. En outre, la première personne a pour partenaires favoris les formes désactualisantes *would* et *should* contrebalancées par le présent d'un modal dont l'importance n'a fait que croître en anglais moderne et contemporain (*can*).[14] Le pronom de deuxième personne, à la fois singulier et pluriel, objet et sujet, est attiré par les mêmes modaux au présent de l'indicatif, *may* venant cette fois faire pendant à *can*. Entre ces deux éléments fondamentaux du système des pronoms, *we* occupe un place intermédiaire qui traduit parfaitement son ambiguïté foncière : *we* n'est pas le véritable pluriel de *I*; c'est *I* associé à l'une, l'autre ou plusieurs des autres personnes.

Pour montrer la fécondité de cette procédure mathématique d'analyse des données, nous allons quitter le domaine privilégié de la syntaxe pour lui soumettre des unités lexicales choisies selon des critères multiples : les adjectifs les plus fréquents (plus de 30 répétitions dans le corpus). Il s'agit bien entendu d'adjectifs communs présents dans toutes les variétés d'anglais, des monosyllabes sans éclat, peu poétiques a priori. La poésie anglaise du vingtième siècle préfère à l'adjectif de nature des éléments rares ou des composés originaux. L'arbre proposé par le logiciel (figure 14) présente une structure rayonnée, arbre en boule ou buisson, qui semble la contrepartie graphique de leur indépendance et de leur inertie sémantique.

Figure 14

[14] On peut légitimement se demander dans quelle mesure ce phénomène ne ne s'expliquerait pas, au moins partiellement, par la nature du corpus . C'est l'une des multiples voies que nous nous proposons d'explorer à l'occasion de travaux ultérieurs sur des corpus variés en cours d'élaboration.

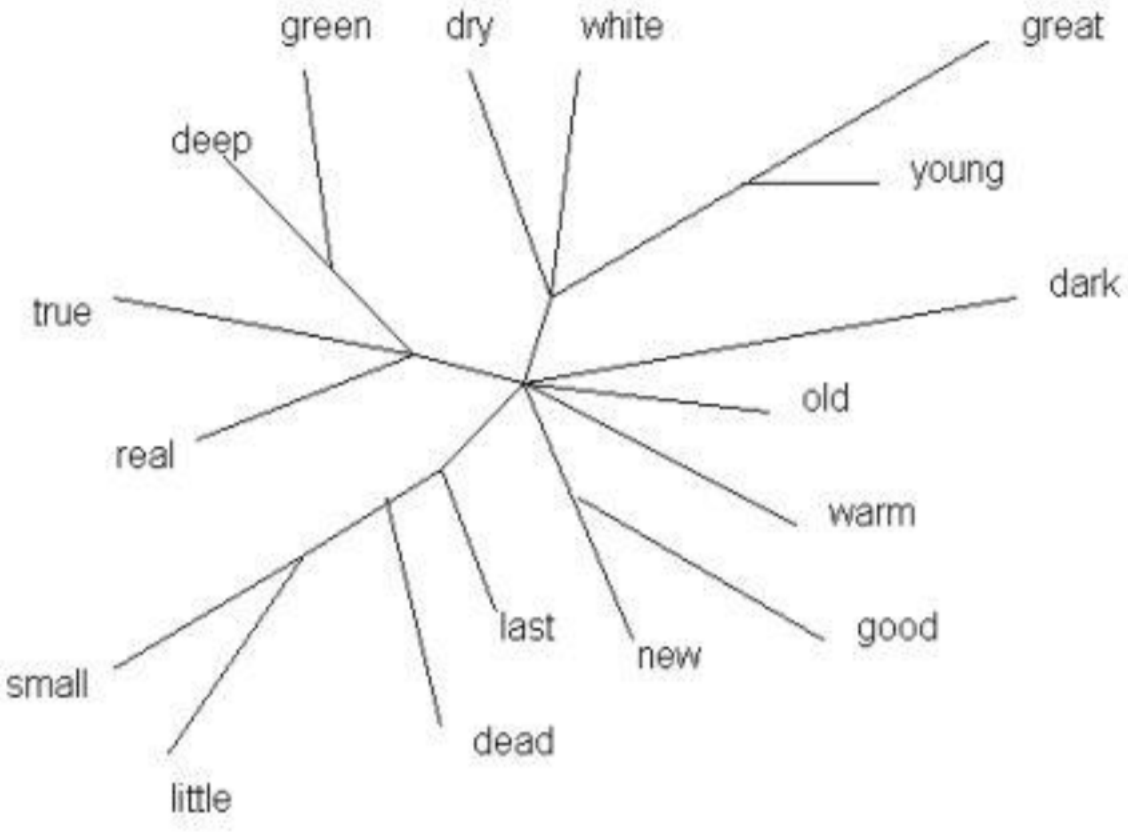

Figure 14

Cependant, à y regarder de plus près, l'arbre est loin de manquer de structure. Le bouquet des adjectifs communs s'y organise en quatre groupes bien individualisés :

1	2	3	4
new	young	green	small
good	great	deep	little
warm	white	true	dead
old	dry	real	last
dark			

Dans tous les cas, un couple nettement particularisé s'oppose aux autres éléments du sous-ensemble. Ainsi, *new* et *good* se séparent de *old*, *warm* et *dark*, tandis que s'accouplent pareillement *young* et *great*, *green* et *deep*, *small* et *little*. Aucun des groupements de l'arbre n'est aléatoire. Ils soulignent toujours une parenté qui peut être sémantique (*small-little*, *real-true*) ou phonique (*green-deep*, *dry-white*). Le rapprochement entre *new* et *good* a de quoi intriguer. La proximité sur l'arbre de ces deux feuilles suggère des distributions semblables dans le corpus. Or, des travaux antérieurs[15] nous ont montré qu'un texte, *The Magnetic Mountain*, faisait un emploi anormalement élevé de ces adjectifs (écarts réduits de 5,31 et 4,89 respectivement). Dans ce poème qui orchestre des thèmes publics et des motifs privés sous le signe de Marx et de Lawrence, les deux adjectifs répétés sont porteurs des mêmes connotations messianiques. Tous deux sont suremployés et font donc partie du vocabulaire caractéristique du texte, mais c'est *new* qui est l'élément marqué du couple, qui reprend en les amplifiant, toutes les valeurs sémantiques de *good*. Ce dernier se manifeste en effet dans des

[15] Juillard, M., L'expression poétique chez Cecil Day Lewis, vocabulaire, syntaxe, métaphore. Etude stylostatistique, Genève, Slatkine, 1983, p. 181.

collocations à très forte probabilité, des clichés par lesquels le poète rapproche son écriture des rythmes du parler ordinaire :

"Line was a ***good*** line, ballasted on grit,
Surveyors weren't fools, platelayers didn't quit"
(*Collected Poems*, p. 84)
"Go mad in ***good*** company, find a ***good*** country,
Make a clean sweep or make a clean end."
(*Ibid*., p. 85)
"Simple that world, of two dimensions,
Of stone mansions and ***good*** examples ;
(*Ibid*., p. 88)
"Don't tease the keepers, eat up your kippers,
And you'll have a treat one day if you're ***good*** boys."
(*Ibid*., p. 100)

"Those whom winter has wasted, not worsted,
Good at their jobs for a break-down gang"
(*Ibid*., p. 112)
"Broad let our valleys embrace the morning
And satisfied see a ***good*** day dying"
(*Ibid*., p. 112)
"When the land is ours, these springs shall irrigate
Good growing soil until it teems"
(*Ibid*., p. 114)
"You who like peace, ***good*** sorts, happy in a small way
Watching birds or playing cricket with schoolboys"
(*Ibid*., p. 115)
"All you fine ladies, once you were flowers
England was proud of, rich blooms, ***good*** growers"
(*Ibid*., p. 116)

New au contraire, figure dans des collocations hardies, au sein de puissantes métaphores, comme en témoignent ces autres extraits de notre concordance :

"Warm in my walled garden the flower grew first,
Transplanted it ran wild on the estate.
Why should it ever need a ***new*** sun ?"
(*Collected Poems*, p. 87)
"Will you tighten the belt and shrug the shoulder
Or plough up the playing-fields, sow ***new*** soil,
Build a reservoir and bore for oil ?"
(*Ibid*., p. 90)
".....their dynamos chant
Canticles of a ***new*** power : my holy land is blasted,"
(*Ibid*., p. 92)
"This one shall hear, though from afar,
The clear first call of ***new*** life,.....
(*Ibid*., p. 104)
"Collect your forces for a counter-attack,
New life is on the way, the relief train."
(*Ibid*., p. 110)
"Out of that dark a ***new*** world flowers."
(*Ibid*., p. 111)
"Make us the wind from a ***new*** world ! "
(*Ibid*., p. 115)
"You shall be leaders when zero hour is signalled,
Wielders of power and welders of a ***new*** world."
(*Ibid*., p. 116)

Nous avons là de la poésie lyrique par excellence, car, si l'on en croit P. Servien , que cite G. Deleuze, il existerait deux langages : le langage des sciences, dominé par le symbole

d'égalité, et où chaque terme peut être remplacé par d'autres; le langage lyrique, dont chaque terme, irremplaçable, ne peut être que répété.[16]
Tous les arbres présentés et analysés jusqu'ici étudiaient des objets (les mots des colonnes du tableau 7 par exemple) par rapport à des propriétés, l'appartenance à tel ou tel texte du corpus (les lignes du tableau 7) qui manifeste la différence entre les éléments répétés. Le logiciel utilisé offre la possibilité de traiter comme objet ce qui était propriété et vice-versa, en inversant lignes et colonnes par le biais d'une transformation générale du tableau de départ. Il devient ainsi possible d'apprécier les affinités et les antagonismes entre textes par rapport à tel ou tel groupe de mots. La figure 15 provient de la transformation du tableau concernant les adjectifs de forte fréquence.

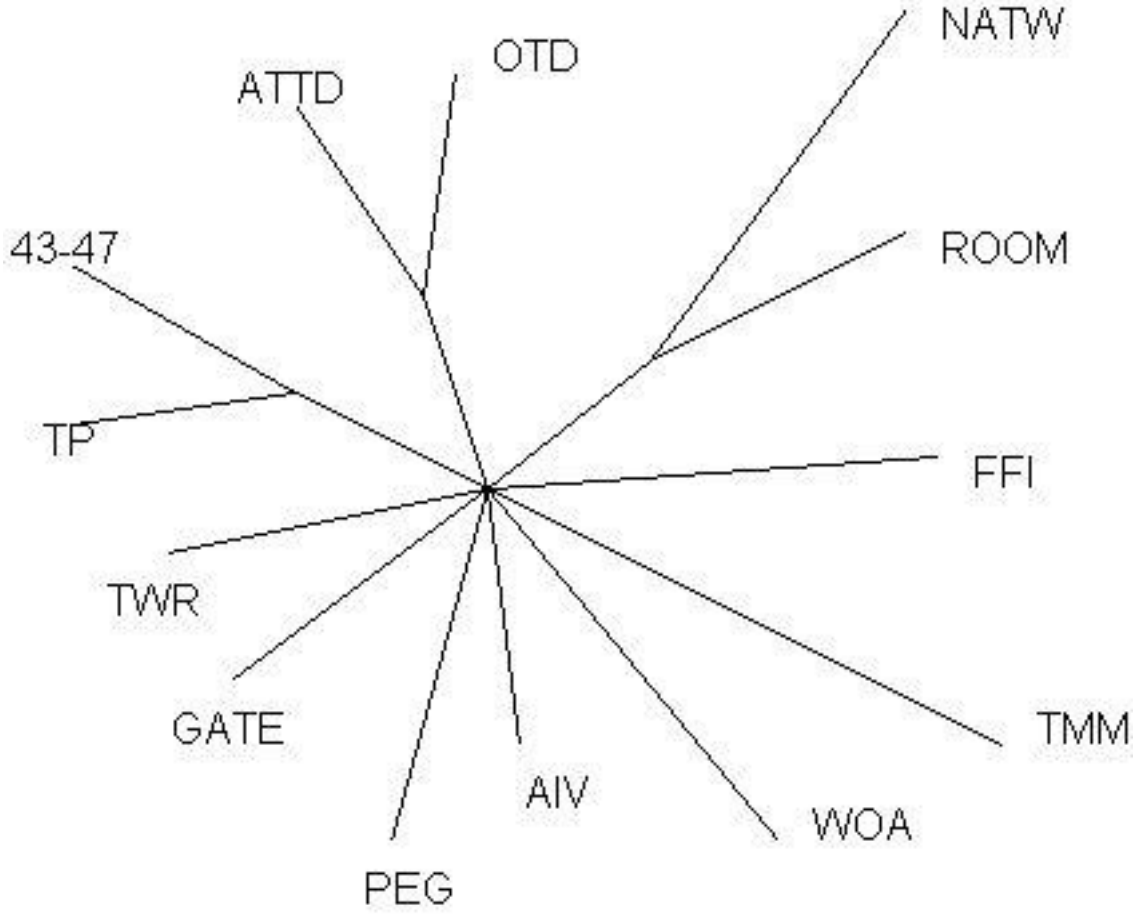

Figure 15

L'allure d'ensemble de la figure, la convergence en un même point de la plupart des branches traduit le peu d'attirance des textes pour ces éléments au faible potentiel poétique. Seuls se regroupent deux à deux quelques oeuvres où le poète revient à des thèmes déjà explorés ou reprend des formes poétiques et des schémas rythmiques temporairement délaissés.
Les rapprochements sont plus éclatants lorsque l'on fait porter la même transformation sur les pronoms personnels (figure 16).

[16] Servien, P., Principes d'esthétique, Paris, Boivin, 1935, pp. 3-5 ; Science et poésie, Paris, Flammarion, 1947, pp. 44-47, cités par Deleuze, G., Différence et répétition, Paris, P.U.F., 1968, p. 8.

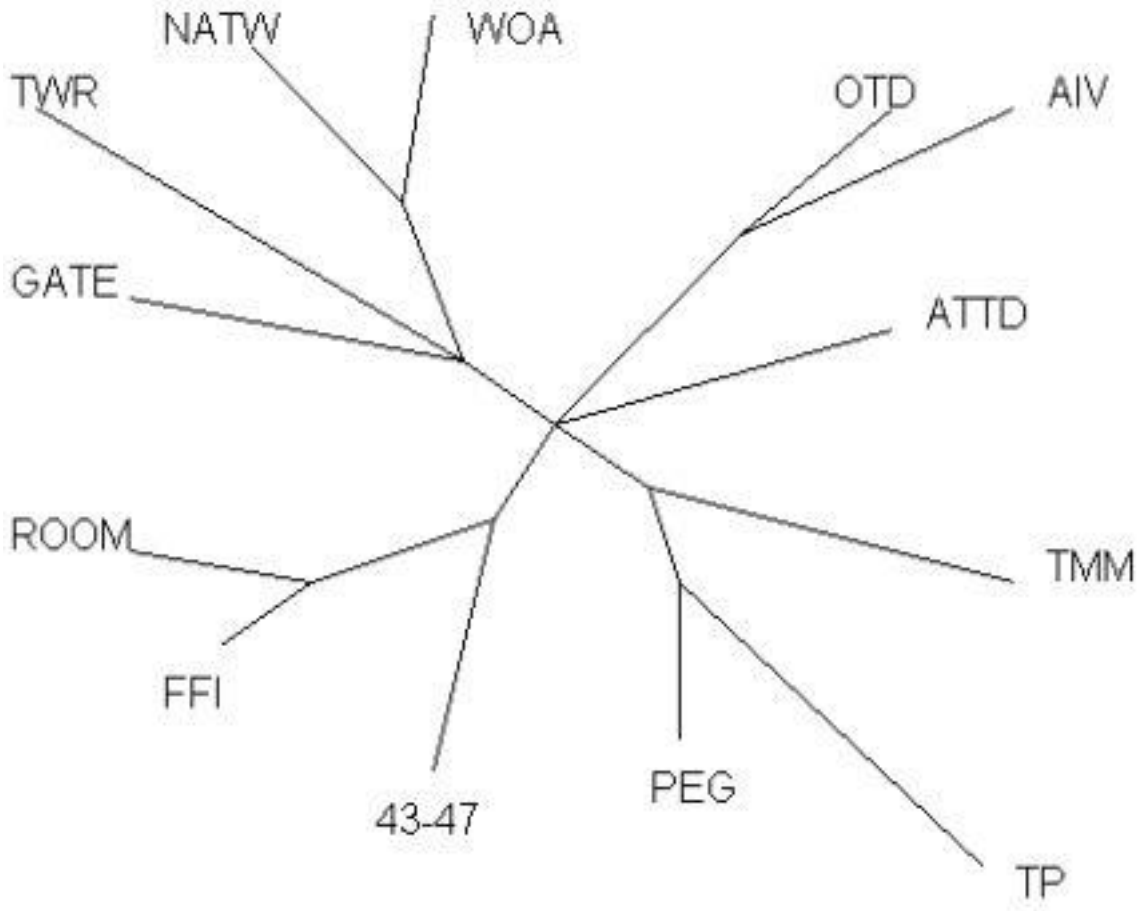

Figure 16

La structure d'ensemble est beaucoup plus finement ramifiée et les textes se regroupent ou se repoussent selon qu'y domine tour à tour le dialogue, le soliloque ou la description au gré des variations dans la manière, lyrique, dramatique ou descriptive, épique ou intimiste, d'aborder des thèmes différents ou répétés. C'est ainsi que le premier texte et le dernier chronologiquement (*TP, Transitional Poem* et *TWR, The Whispering Roots*) occupent des positions diamétralement opposées. Il n'est donc pas possible de parcourir l'arbre de feuille en feuille en reproduisant la succession des oeuvres dans le temps. Ceci n'a rien d'étonnant, mais au contraire reflète et corrobore l'opinion de l'auteur qui décrivait son parcours poétique comme une série de perpétuels recommencements.

Si, comme le pensent de nombreux linguistes, le but primordial de la linguistique est de démontrer que la langue est un système, de dépasser le chaos des phénomènes de surface pour mettre au jour l'ordre secret qui les régit, alors l'application à des données chiffrées exhaustives de l'algorithme de représentation arborée présenté ici est un outil précieux pour parvenir à cette fin. Alors que les méthodes plus classiques de la linguistique quantitative, en dépit de leur grande valeur mainte fois prouvée, risquent parfois d'aboutir à un éparpillement des données, donc des résultats, et imposent au chercheur un effort de synthèse à la faveur duquel peut se réintroduire le subjectif et l'arbitraire, l'analyse arborée se recommande par son exhaustivité et la clarté des représentations qu'elle propose pour organiser lumineusement des amas touffus de données. Elle est aussi supérieure aux méthodes de l'analyse hiérarchique ou multidimensionnelle, parce qu'elle n'impose pas à la réalité de cloisonnement en catégories étanches mais considère essentiellement les parentés, l'air de famille cher à Wittgenstein, en s'appuyant uniquement sur les notions complémentaires de structure et de

distance.[17] Elle est la mise en oeuvre de la subtile dialectique entre la différence et la répétition, non pas celle que G. Deleuze appelle la répétition du Même, mais celle, positive et dynamique, qui éclaire et fait avancer :

> "La première répétition est répétition du Même, qui s'explique par l'identité du concept ou de la représentation; la seconde est celle qui comprend la différence, et se comprend elle-même dans l'altérité de l'Idée, dans l'hétérogénéité d'une "apprésentation". L'une est négative, par défaut du concept, l'autre, affirmative, par l'excès de l'Idée. L'une est hypothétique, l'autre catégorique. L'une est statique, l'autre dynamique. L'une est répétition dans l'effet, l'autre dans la cause. L'une, en extension, l'autre intensive. L'une ordinaire, l'autre, remarquable et singulière. L'une est horizontale, l'autre verticale. L'une est développée, expliquée; l'autre est enveloppée, et doit être interprétée. L'une est révolutive, l'autre d'évolution. L'une est d'égalité, de commensurabilité, de symétrie; l'autre fondée sur l'inégal, l'incommensurable ou le dissymétrique....."[18]

BENVENISTE, E. *Problèmes de Linguistique Générale*, I et II Paris, Gallimard, 1966 et 1974.

BENZECRI, J. P., L'Analyse des Données, 2 vols, Paris, Dunod, 1973.

BARTHELEMY, J.P., LUONG, N.X., "Représentation arborée des mesures de dissimilarités", *Statistique et Analyse des Données*, Paris, 1986.

BUNEMAN, P., "The recovery of trees from measures of dissimilarity", in HODSON & al., *Mathematics in the Archaelogical and Historical Sciences*, Edinburgh, Edinburgh University Press, 1971.

---------- "A note on the metric properties of trees", *Journal of Combinatorial Theory*, 17(b), 1974, pp. 48-50.

DAY LEWIS, *Collected Poems*, London, Jonathan Cape, 1954.

DOBSON, J., "Unrooted tree for numerical taxonomy", *Journal of Applied Probability*, 11, 1974, pp. 32-42.

GUILLAUME, G., *Langage et Science du Langage*, Québec, Presses de l'Université Laval ; Paris, Nizet, 1964.

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

JUILLARD, M., *L'expression Poétique chez Cecil Day Lewis, Vocabulaire, Syntaxe, Métaphore. Etude Stylostatistique*, Genève, Slatkine, 1983.

---------- "A quantitative approach to semantic and morphosemantic fields in a literary work", *ALLC Journal*, vol. 6, 1 & 2, 1985, pp. 14-23.

---------- "Linguistique et linguistique quantitative", *Procceedings of the ALLC XIIIth Conference*, Norwich, 1986, Slatkine, Genève, 1989.

JUILLARD, M. & LUONG, X.N., "Unrooted trees revisited : Topology and poetic data", *Computers and the Humanities*, 23, 1989, pp. 215-225.

LUONG, X.N., "Using a Tree-model in Textual Analysis", *ICCH*, University of South Carolina, April 1987.

---------- *Méthode d'analyse arborée. Algorithmes, applications*. Thèse de Doctorat d'Etat, Université Paris V, 1988.

WITTGENSTEIN, L., *Philosophische Untersuchungen*, Oxford, Blackwell, 1953.

Abréviations

TP : *Transitional Poem*
FFI : *From Feathers to Iron*

---

[17] Cette nouvelle méthode de représentation arborée offre en outre la possibilité d'une lecture dynamique de la genèse de l'arbre. Pour sa mise en oeuvre sur des données se reporter à Juillard, M., "La Dynamique de l'arbre", CUMFID, CNRS, Nice, 16, juin 1989, pp. 67-86.

[18] Deleuze, G., Différence et répétition, Paris, P.U.F., 1968, p. 36.

TMM : *The Magnetic Mountain*
ATTD : *A Time to Dance*
NATW : *Noah and the Waters*
OTD : *Overtures to Death*
WOA : *Word over All*
43-47 : *Poems 43-47*
AIV : *An Italian Visit*
PEG : *Pegasus*
GATE : *The Gate*
ROOM : *The Room*
TWR : *The Whispering Roots*.