De nouveaux arbres pour un nouveau corpus

Michel JUILLARD et Xuan LUONG

Abstract. This article implements the latest version of software for tree-representation developed in the Nice-Sophia Antipolis C.N.R.S. laboratory for linguistic research U.P.R. 6861. The new programmes are applied to new data from the coded version of the LOB Corpus of English texts. The results presented here, although fragmentary and provisional, corroborate current modern linguistic hypotheses, while pointing to both the versatility of the method and the richness and reliability of the corpus.

Keywords: Tree topology, tree-analysis, computational linguistics, english syntax, LOB Corpus of British English.

Mots-clés : Topologie sur un arbre, analyse arborée, linguistique quantitative, syntaxe de l'anglais, Corpus LOB d'anglais britannique.

Depuis longtemps les sciences naturelles, les sciences humaines et la linguistique se préoccupent de représenter leurs volumineuses informations sous une forme synthétique. Le but commun est souvent de faire apparaître des regroupements et des oppositions propres à éclairer le chercheur sur la configuration de ses données. Il existe divers modèles d'analyse qui produisent selon le cas des arbres hiérarchiques, des arbres plantés, des arbres phylogénétiques ou des arbres de Chomsky. Sans prétendre lui faire toucher du doigt l'arbre de la science du bien et du mal, nous convions le lecteur à une promenade sous des frondaisons nouvelles. Nous appliquerons la dernière version du logiciel d'analyse arborée, conçu à l'université de Nice-Sophia Antipolis au sein de l'U.P.R. 6861 du C.N.R.S., à des données linguistiques sur lesquelles nous n'avons pas encore travaillé : la version codée grammaticalement du LOB Corpus de textes d'anglais britannique.

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

✉ Institut National de la Langue Française; Statistique linguistique (UPR 6861, CNRS); UFR Lettres, Arts et Sciences Humaines; 98, boulevard Herriot; B.P. 209; F-06204 Nice Cedex 3 (France).
Fax : + 33 93 37 54 45

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]. À l'image de son homologue américain, le LOB Corpus se compose de 500 échantillons d'environ 2 500 occurrences représentant la plupart des domaines de la langue écrite et tous provenant d'œuvres et de documents divers publiés au cours de l'année 1961. Le tableau 1 donne la répartition, avec leurs étiquettes anglaises, des quinze grandes catégories de textes avec le nombre d'échantillons de chacune.

Tableau 1

Catégorie		Nombre d'échantillons
A	*Press: reportage*	44
B	*Press: editorial*	27
C	*Press: reviews*	17
D	*Religion*	17
E	*Skills, trades and hobbies*	38
F	*Popular lore*	44
G	*Belles lettres, biography, essays*	77
H	*Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)*	30
J	*Learned and scientific writing*	80
K	*General fiction*	29
L	*Mystery and detective fiction*	24
M	*Science fiction*	6
N	*Adventure and western fiction*	29
P	*Romance and love story*	29
R	*Humour*	9

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 dans le cours des années quatre-vingt. Le codage a été effectué avec une rigueur, une intelligence et une exhaustivité exemplaires par une équipe de chercheurs réunissant, sous la direction de Stig Johansson, Geoffrey Leech et Roger Garside des linguistes et des informaticiens, notamment Eric Atwell, Ian Marshall, Mette-Cathrine

[1] On consultera pour plus de détails Johansson : 1980, 25–36 ainsi que l'introduction de Hofland et Johansson : 1982.

[2] Johansson, Leech, Goodluck : 1978.

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

C'est sur ce nouveau corpus considérablement enrichi par son codage très fin que porte notre travail. L'approche qui s'impose est d'observer la distribution des diverses catégories grammaticales — le codage en distingue près d'une centaine — à travers les textes. Nous les avons regroupées en onze grandes classes : déterminants, auxiliaires et modaux, conjonctions de coordination, conjonctions de subordination, prépositions, adjectifs, noms, pronoms, adverbes, verbes, interjections. Nous étudions d'abord la distribution des principales catégories dans un premier sous-ensemble de textes. On trouvera en annexe les tableaux des occurrences de ces 11 catégories grammaticales dans des textes d'environ 2 500 mots chacun, provenant des sections L, P, H, J, N, K du corpus soumises à cette première expérience. Chaque tableau se présente ici sous forme d'un comptage. On connaît ainsi, pour chaque « objet » i, le nombre d'occurrences p_{ij} de la « caractéristique » j. Chaque tableau peut donner lieu à divers tests statistiques (écarts réduits, χ^2, corrélations...) portant sur tel ou

---

[3] On trouvera une description très claire et complète des options de codage dans Johansson, Atwell, Garside, Leech : 1986. Nous souhaitons remercier Stig Johansson de nous avoir spontanément et généreusement offert un exemplaire de ce précieux document.

[4] Muller : 1977, p. 28.

tel de ses éléments. Afin d'étudier globalement le comportement des objets (ou de leurs caractères), notre choix se portera dans un tout premier temps sur des méthodes d'analyse multidimensionnelle. Il est aisé de transformer les tableaux d'occurrences de départ en tableaux de fréquences. Chaque objet se trouve alors résumé par un profil décrivant la distribution de la probabilité des caractères correspondant à cet objet. La distance du χ^2 est l'indice de proximité reconnu comme le mieux adapté à cette situation. Avec les textes de l'annexe (sous-ensembles L, P, H, J, N et K), nous utilisons en première approche la classification hiérarchique. La figure 1 est l'arbre de la classification (en appliquant un algorithme ascendant, par la méthode du lien moyen).

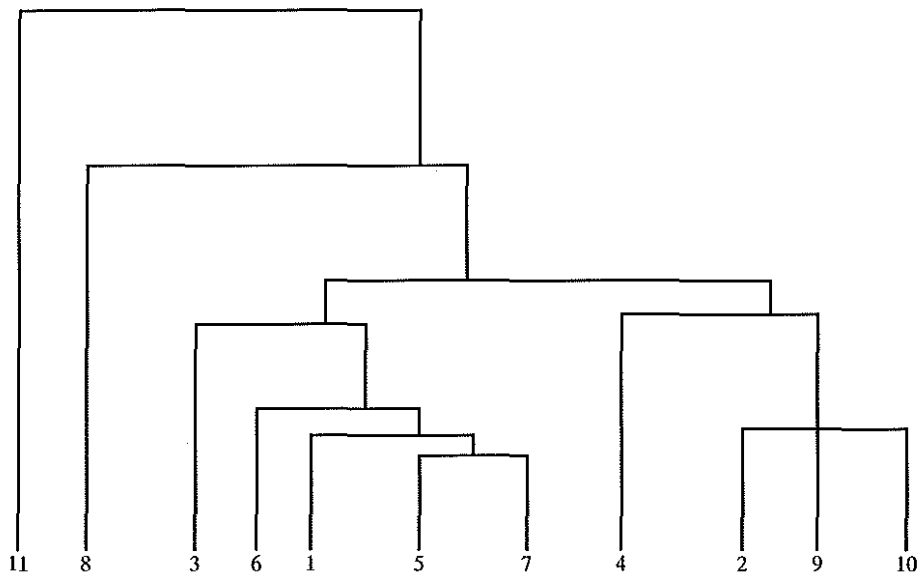

Fig. 1.– Arbre planté

La lecture de l'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 nœud fournit un indice du niveau de formation d'une classe. Les éléments d'une classe sont à égale distance d'un nœud; ainsi les catégories grammaticales 2, 9 et 10 se trouvent associées entre elles avant de s'adjoindre la catégorie 4. La partie gauche de l'arbre est encore plus étroitement compartimentée. Cet arbre, satisfaisant à l'œil épris d'ordre sépare les objets par des cloisons étanches et ne nous renseigne pas sur leurs affinités, leur air de famille plus ou moins prononcé. Il faut se tourner vers d'autres figurations si l'on

souhaite mettre en évidence des ensembles où des nuances peuvent déterminer la représentativité de chacun des éléments, nuances qui constituent pour les psychologues un « gradient de représentativité »[5]. La nature de nos indices de proximité nous a conduits à utiliser aussi des techniques de « *Multidimensional Scaling* ». La figure 2 offre une représentation dans un plan des mêmes données.

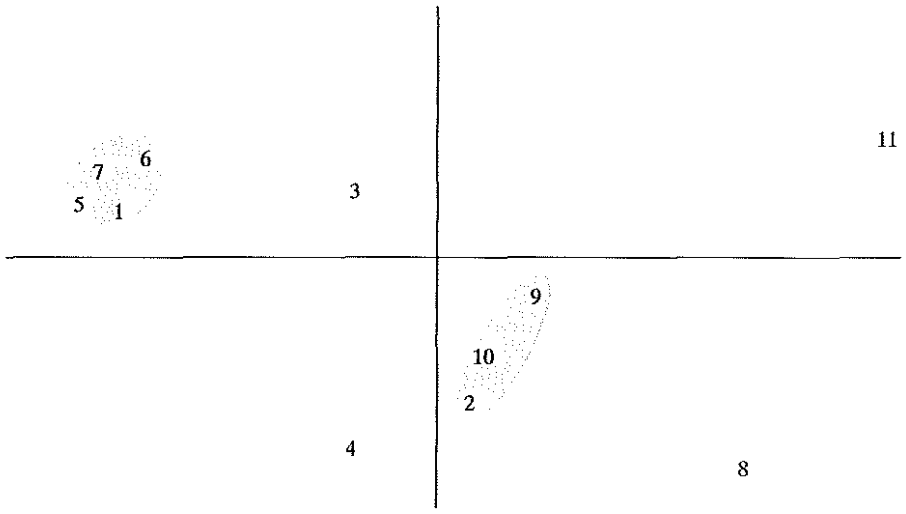

Fig. 2.– Analyse multidimensionnelle (*Multidimensional Scaling*)

L'axe vertical oppose deux groupes, l'un étroitement constitué (5, 1, 7, 6), l'autre à la structure plus lâche (2, 10, 9). Quatre catégories sur onze restent à l'écart (4, 3, 8, 11) et l'axe horizontal est encore plus difficile à interpréter. La carte obtenue ci-dessus positionne les objets en fonction de la distance euclidienne et le « *Multidimensional Scaling* » favorise une interprétation des données en termes de dimension. Une troisième approche pour analyser ces données consiste à les représenter sur un arbre. On obtient l'arbre de la figure 3. Nous l'utilisons d'abord pour illustrer quelques considérations générales de la méthode avant de passer à son exploitation proprement linguistique. Le lecteur de *Revue, Informatique et Statistique dans les Sciences humaines* a déjà eu un aperçu de la démarche dans un précédent article[6]. Nous évoquerons ici quelques idées forces.

[5] Rosh : 1983.

[6] Juillard et Luong : 1988, pp. 222–240.

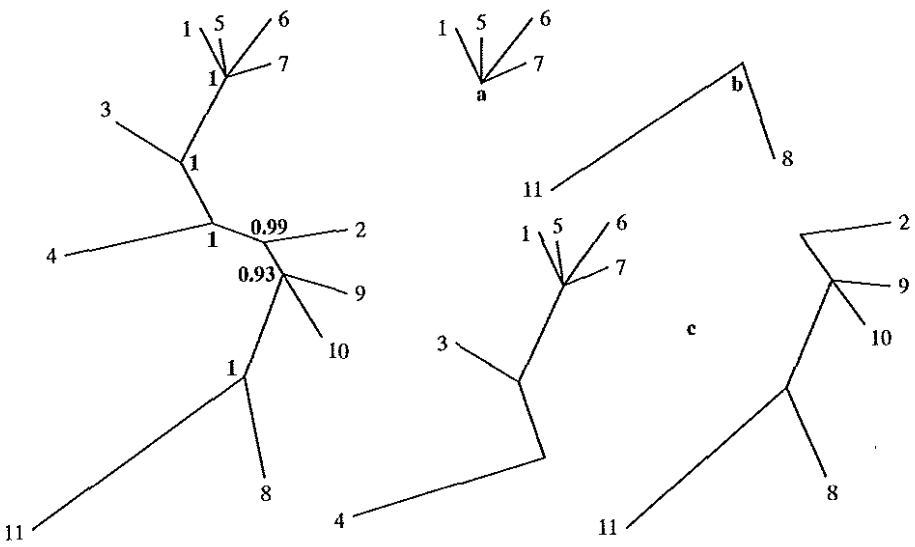

Fig. 3.– L'arbre radial de L, P, H, J, K, N. Groupements a et b, opposition entre deux familles (c).

L'arbre est un arbre non planté, semblable aux arbres phylogénétiques des naturalistes. Les feuilles de l'arbre correspondent aux objets étudiés. 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 : 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.

- À chaque sommet intérieur est 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 bipartions 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 est basé 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[7].

Le premier ensemble que nous avons traité en vue d'une exploitation linguistique était constitué d'une trentaine d'échantillons de 2 500 mots environ chacun, provenant des sections L, P et N du corpus, constituées respectivement de romans policiers, de romans d'amour et de romans d'aventures. La figure produite au terme de l'analyse de ces 75 000 occurrences rangées dans onze grandes catégories grammaticales s'avère bien formée et présente une structure nettement arborée.

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[8], 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

[7] Luong : 1988.

[8] Quirk, Greenbaum, Leech, Svartvik : 1985, p. 47.

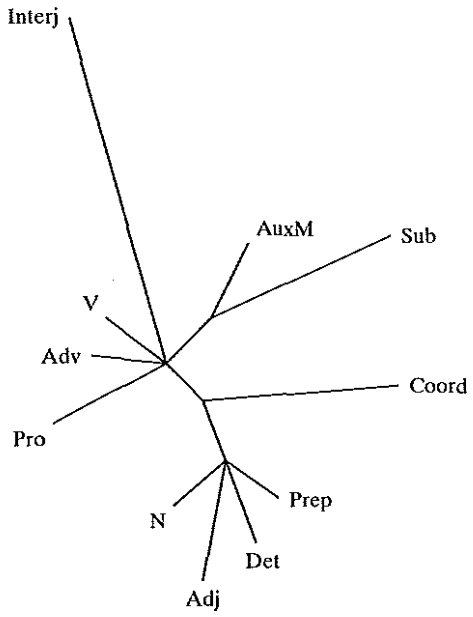

Fig. 4.– Les catégories grammaticales dans L, P, N

portée variable dans les textes, reliant tantôt des groupes nominaux, tantôt des propositions; le rejeton démesuré qui surgit du cœur 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.

Il nous a paru intéressant de prendre le risque de mettre en cause la représentation des données linguistiques décrites ci-dessus et donc nos conclusions provisoires, en accroissant le volume et la variété des textes traités. Aux trois groupes de textes de L, P et N nous avons adjoint trois passages de taille égale (dix échantillons d'environ 2 500 mots) tirés des catégories textuelles H (textes divers, documents officiels et rapports), J (écrits savants et scientifiques) et K (fiction générale) pour les soumettre à nouveau au logiciel de représentation arborée (figure 5).

C'est au fruit que l'on connaît l'arbre. Le nouvel arbre oppose toujours deux ensembles à noyaux verbal et nominal. Leur position dans l'espace, rappelons-le, n'est pas en soi déterminante; seules sont signifiantes les distances entre éléments terminaux ou feuilles et les groupements et oppositions. La stabilité de cette image d'ensemble n'en rend que plus intéressantes les variations par rapport au précédent schéma. Les distances entre les composantes du groupe

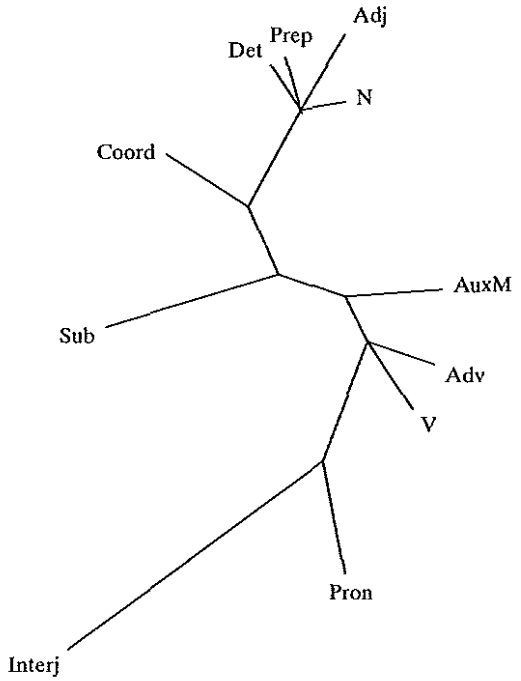

Fig. 5.– Les catégories grammaticales dans L, P, H, J, K, N

nominal se sont amenuisées, signe de leur plus grande cohérence au gré des textes soumis à l'analyse. Les éléments du groupe verbal ou du groupe prédicatif voient au contraire leurs affinités s'estomper et leur autonomie s'accentuer : le verbe et l'adverbe restent proches et solidaires, tandis que le pronom s'éloigne pour contracter une union lâche, insolite, par défaut, avec l'interjection. Les phénomènes les plus intéressants concernent la relative indépendance des auxiliaires et modaux et surtout des conjonctions de subordination. Il est difficile de donner une explication univoque, définitive de ce phénomène d'ensemble touchant le groupe prédicatif. Sa survenue est intéressante en soi car elle prouve à la fois la sensibilité de la méthode d'analyse et l'homogénéité de chaque catégorie textuelle. On pourrait suggérer que les textes ajoutés ici à L, P, N sont principalement des documents à valeur référentielle (catégories textuelles H et J) où domine la troisième personne et où s'effacent les acteurs de la communication intersubjective, marginalisant de la sorte la catégorie du pronom. Une confirmation empirique et péremptoire de cette hypothèse s'offre d'emblée si l'on compare les tableaux d'effectifs des catégories grammaticales ayant servi de point de départ à cette analyse arborée. Pour des tranches de 2 500 mots

environ, l'effectif des pronoms varie dans L et P entre 138 et 355, dans K et N entre 323 et 492 alors que pour H et J les bornes correspondantes ne sont que de 17 et 114. L'effectif de la tranche de H et J la mieux pourvue en pronoms personnels est inférieur à la tranche la plus pauvre de L, P, K ou N (cf. Annexe). Ces effectifs absolus sont sans appel et découragent tout recours à un test statistique. L'analyse arborée prouve ici sa grande valeur heuristique. Une autre façon de mettre en évidence le rôle joué par les textes de H et J dans la nouvelle physionomie de l'arbre consiste à étudier la représentation arborée duale des textes à travers la distribution des unités de la syntaxe (figure 6). La forte spécificité des textes de H et J, liée en partie au mode d'énonciation qu'ils mettent en œuvre, éclate sous forme de chiffres dès le cycle d'itérations préalable à la construction de l'arbre : toutes les premières itérations (un bon tiers) proposent comme feuilles de l'arbre en devenir, avec d'excellents indices d'agrégation, des textes de ces deux sous-ensembles ajoutés[9]. Ces tropismes se trouvent confirmés de façon lumineuse dans l'arbre final dont toute la partie gauche ploie sous les seuls rameaux portant les feuilles de H et de J.

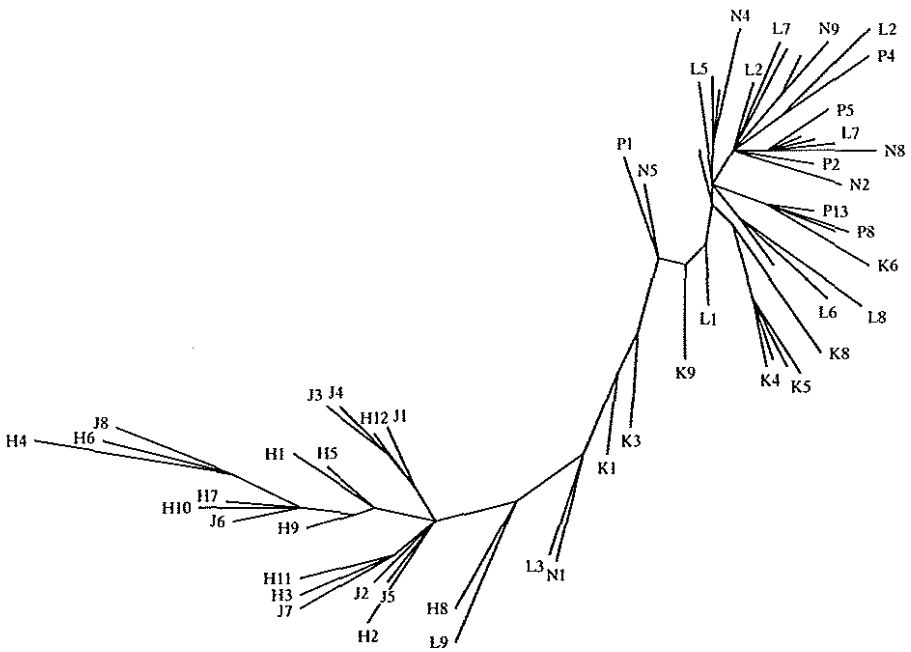

Fig. 6.– LPHJNK

[9] Pour une lecture cinétique des arbres, se reporter à Juillard : 1986.

Soucieux d'éprouver le logiciel tout autant que la qualité de ces données nouvelles si soigneusement codées grammaticalement, nous avons procédé à une réduction du corpus précédent en revenant à notre ensemble initial de trente échantillons composé de dix sous-ensembles de chacune des tranches L, P et N, tout en concentrant notre attention sur cinq grandes catégories grammaticales seulement : *Be*, *Do*, *Have*, les modaux et enfin les pronoms. La figure 7 est fort instructive.

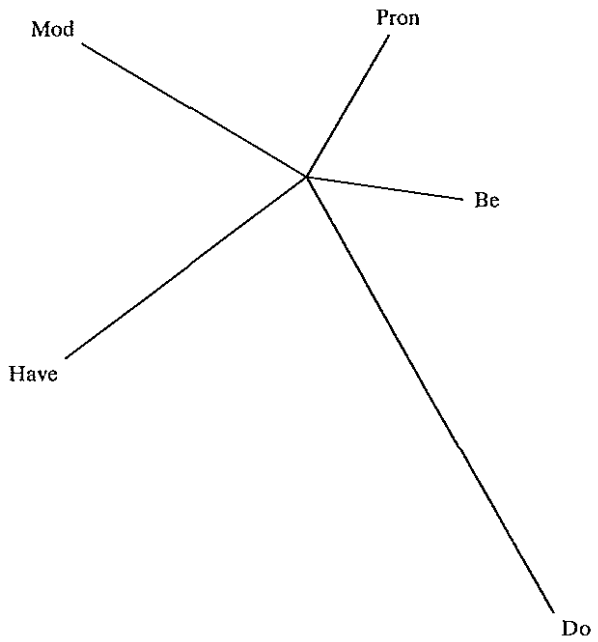

Fig. 7.– L'arbre de LPN5 (*be*, *do*, *have*, mod, pron)

La structure étoilée de l'arbre reflète à la fois l'omniprésence de ces éléments clés de la syntaxe dans des textes de fiction (romans policiers, romans d'amour, romans d'aventures) où l'on rencontre tour à tour descriptions et dialogues, interventions de l'auteur omniscient et représentations d'échanges en discours direct, qui ne sauraient se passer ni de l'une ni de l'autre de ces catégories grammaticales. Les distances prennent toute leur signification et apportent des informations que la structure de l'arbre montre de façon moins éclatante. Les diverses formes de *Be* et les pronoms constituent un groupe tandis que *Do*, opérateur fondamental de la prédication en anglais, fait preuve de la plus grande spécificité en refusant toute association. L'indépendance des différents

éléments du groupe auxiliaire de la phrase anglaise apparaît nettement sur le schéma et traduit la complémentarité de leurs rôles syntaxiques. Il est possible enfin de voir dans la distance séparant *Do* et les modaux la traduction imagée de la concurrence que les théories linguistiques de l'énonciation instaurent entre les modaux de stricte obédience et *Do* considéré comme un opérateur modal à part entière[10]. L'opposition de *Be* et *Have* traduirait alors leurs valeurs aspectuelles particulières au sein du groupe auxiliaire. D'autres explorations seront nécessaires pour tester plus exhaustivement ces hypothèses surgies de la représentation arborée des données. On peut, en attendant, se demander si les réactions des textes à ces unités fondamentales, tropismes positifs ou négatifs, sont visibles lorsque l'on transpose le tableau des effectifs catégoriels afin de tracer l'arbre regroupant les textes selon leurs hypothétiques affinités avec tel ou tel de ces éléments ou dans leur gestion de l'ensemble de ces unités syntaxiques. Il est parfois intéressant, en choisissant un sommet particulier, de planter un arbre résultant de l'analyse arborée. La figure 8, l'arbre planté des textes LPN, a une certaine ressemblance avec un dendrogramme. Nous avons privilégié le centre des regroupements (LUONG, *op. cit.*) comme racine. Rappelons que du point de vue algorithmique, les éléments proches, les structures stables, sont détectés en premier et ce centre est le point où l'analyse se termine.

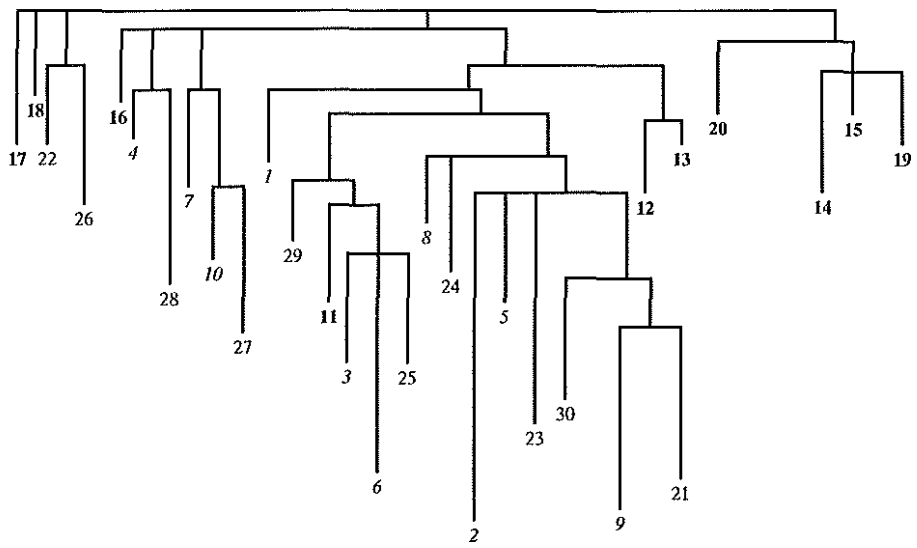

Fig. 8.– Arbre planté des textes LPN

[10] Voir par exemple Bouscaren et Chuquet : 1987, p. 66 et *passim*.

Les textes figurent en bout de branche, ceux de L (romans policiers) numérotés de 1 à 10, en italiques, ceux de P (roman d'amour) de 11 à 20, en gras sur la figure, et ceux de N (romans d'aventures) de 21 à 30. Les textes de L se regroupent vers le cœur de l'arbre, mais à bonne distance les uns des autres, tandis que ceux de P sont à la fois opposés à tous les autres et portés dans leur majorité par des rameaux voisins et très courts. Les textes de N manifestent des spécificités moins marquées qui leur permettent d'occuper des branches dans toute la couronne de l'arbre. Cette répartition évoque la forte personnalité stylistique de L et surtout de P, ainsi que les contraintes moins fortes que le genre N ferait peser sur ses auteurs. Il s'agit encore d'une hypothèse révélée par l'analyse arborée qu'il conviendra de mettre à l'épreuve sur d'autres textes avant de la présenter comme définitivement vérifiée.

Dans un article récent, Jean-Pierre Barthélemy évoque Claude Lévi-Strauss opposant, dans *La Pensée Sauvage*, le bricoleur à l'homme de l'art[11]. Le bricoleur, nous disent-ils, dispose d'une gamme d'outils et de matériaux limitée, souvent sans rapport avec la tâche qu'il s'est assignée et sa règle revient toujours à «s'arranger avec les moyens du bord», alors que le technicien utilise des outils et des matières premières adaptées d'emblée à son projet. À l'occasion de ce premier travail sur un corpus nouveau, nous avons tenté de combiner les qualités scientifiques du technicien et l'empirisme bien compris du bricoleur, en appliquant à des données linguistiques une méthode de représentation conciliant les exigences du regroupement et de l'opposition. Nous ne pensons pas nous être livrés à de plus coupables détournements d'arbres que ceux de nos devanciers, proche et lointain, Porphyre et Buneman. En combinant nos aptitudes, sinon nos talents, de mathématicien et de linguiste, nous avons cherché à montrer à la fois la pertinence de la méthode et la richesse d'un corpus moderne, codé avec finesse, cohérence et exhaustivité, au terme d'une analyse syntaxique rigoureuse. Il nous restera à continuer de mettre à l'épreuve l'un de l'autre l'outil et le matériau, à poursuivre la réalisation d'un vaste chantier, dont nous nous sommes contentés ici de jeter les bases, en nous efforçant de ne déroger ni à la rigueur du technicien ni au génie sporadique du bricoleur.

[11] Barthelemy : 1989.

Bibliographie

BARTHELEMY (Jean-Pierre) : 1989, « Les arbres, de la matière pour bricoler... », in LUONG (Xuan), éd. *Analyse Arborée des Données Textuelles* (Nice : CUMFID, 16, C.N.R.S.).

BARTHELEMY (Jean-Pierre) & GUENOCHE (Alain) : 1991, *Trees and Representation of Proximity Data* (New York : Wiley and Son).

BENVENISTE (Émile) : 1974, *Problèmes de Linguistique Générale* (Paris, Gallimard, vol. 2).

BENZÉCRI (Jean-Pierre) : 1974, *L'Analyse des Données*, 2 volumes (Paris, Dunod).

BOUSCAREN (Janine) et CHUQUET (Jean) : 1987, *Grammaire et Textes anglais, guide pour l'analyse linguistique* (Paris : Ophrys).

BUNEMAN (P.) : 1971, « The recovery of trees from measures of dissimilarity », in HODSON (J.) *et al.*, eds *Mathematics in the Archaelogical and Historical Sciences* (Edinburgh : E.U.P.)

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

HOFLAND (Knut), JOHANSSON (Stig) : 1982, *Word Frequencies in British and American English* (Bergen : Norwegian Computing Centre for the Humanities).

HUDDLESTON (Rodney) : 1984, *Introduction to the Grammar of English* (Cambridge : Cambridge University Press).

JOHANSSON (Stig) : 1980, « The LOB Corpus of British Texts : Presentation and Comments », *ALLC Journal* 1, pp. 25–36.

JOHANSSON (Stig), ATWELL (Eric), GARSIDE (Roger), LEECH (Geoffrey) : 1986, *The Tagged Lob Corpus, Users' Manual* (Bergen : Norwegian Computing Centre for the Humanities).

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* (Oslo : Department of English, University of Oslo).

JUILLARD (Michel) : 1986, « La dynamique de l'arbre : linguistique et représentation arborée », in LUONG (Xuan), ed. 1986, *Analyse arborée des données textuelles* (Nice : CUMFID, 16, C.N.R.S.).

JUILLARD (Michel) et LUONG (Xuan) : 1988, « Des feuilles aux racines : du discours à la langue », *Le nombre et le texte*, Hommage à Étienne Évrard, *RISSH* (Liège : C.I.P.L.-L.A.S.L.A.), pp. 222–240.

JUILLARD (Michel) & LUONG (Xuan N.) : 1989, « Unrooted Trees Revisited: Topology and Poetic Data », *Computers and the Humanities*, 23, pp. 215–223.

KUCERA (Henry) & FRANCIS (William Nelson) : 1967, *Computational Analysis of Present-Day American English* (Providence R.I., Brown University Press).

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

LUONG (Xuan N.) : 1989, éd. « L'analyse arborée des données textuelles » (Nice : CUMFID, 16, CNRS, URL 9).

MULLER (Charles) : 1973 [1re éd. 1967 (Paris : Larousse)], *Initiation aux Méthodes de la Statistique Linguistique* (Paris : Hachette).

MULLER (Charles) : 1977, *Principes et Méthodes de Statistique Lexicale* (Paris, Hachette).

QUIRK (Randolph), GREENBAUM (Sidney), LEECH (Geoffrey), SVARTVIK (Jan) : 1985, *A Comprehensive Grammar of the English Language* (London : Longman).

ROSH (Edward) : 1983, « Prototype classification and logical classification: the two systems », in SCHOLNIK (E.K.), ed., *New trends in conceptual representation* (Hillsdale : Erlbaum).

STRANG (Barbara M. H.) : 1962, 1969, *Modern English Structure* (London : Arnold).

WITTGENSTEIN (Ludwig) : 1953, *Philosophische Untersuchungen* (Oxford : Blackwell).

Annexe

	DET	AuxM	Coor	SUB	Prep	Adj	N	Pron	Adv	V	Interj
L01	233	169	65	37	224	114	447	279	179	278	1
L02	205	225	66	51	225	77	420	297	122	310	1
L03	321	116	74	33	237	147	530	164	145	262	1
L04	240	152	71	27	245	125	430	304	164	270	7
L05	216	226	46	48	222	125	436	295	136	274	7
L06	237	207	107	37	203	142	377	259	212	267	7
L07	215	209	77	45	243	107	307	374	176	283	8
L08	269	190	84	58	227	101	387	278	163	230	6
L09	348	106	93	17	305	171	492	138	138	199	3
L10	261	128	53	30	251	138	448	296	142	282	10
P01	231	165	86	22	235	150	471	233	149	254	7
P02	194	174	86	62	213	105	380	293	185	312	8
P03	225	202	65	29	222	97	387	324	175	281	15
P04	210	228	90	51	199	76	360	382	141	293	2
P05	171	197	87	60	209	115	345	355	176	313	12
P06	207	200	57	39	207	96	365	351	166	319	7
P07	232	201	41	40	245	91	372	308	159	304	5
P08	198	168	60	31	214	108	422	296	193	295	13
P09	203	196	72	41	225	86	362	315	201	298	18
P10	204	194	70	37	231	122	338	353	181	280	11
H01	334	135	101	55	333	152	582	50	94	175	0
H02	334	158	76	42	342	172	522	90	60	194	0
H03	323	163	88	42	319	233	483	72	144	169	0
H04	426	65	110	28	417	193	646	17	62	115	0
H05	354	165	75	39	337	152	550	52	111	171	0
H06	362	102	64	15	395	218	607	19	80	168	0
H07	317	106	90	28	363	175	617	34	101	163	0
H08	295	196	58	50	352	135	490	114	109	202	0
H09	361	127	63	27	340	164	571	92	83	178	0
H10	327	116	98	23	339	169	630	77	96	123	0
H11	358	156	72	57	276	207	510	65	101	193	0
H12	282	147	74	59	336	151	560	71	85	233	0
J01	306	173	63	42	298	164	563	57	98	215	0
J02	308	150	60	41	341	165	519	89	134	185	0
J03	326	164	60	79	317	162	575	68	79	180	0
J04	328	149	45	52	351	151	583	84	87	161	0
J05	329	156	083	069	303	148	512	086	113	193	0
J06	325	110	075	043	351	220	571	071	094	158	0
J07	384	176	052	041	324	175	490	048	158	180	0
J08	458	111	091	039	387	118	561	033	076	174	0

	DET	AuxM	Coor	SUB	Prep	Adj	N	Pron	Adv	V	Interj
N01	298	176	50	48	288	116	441	194	142	244	1
N02	191	166	61	59	272	111	377	344	159	291	4
N03	274	176	59	50	250	123	455	197	187	220	11
N04	267	209	62	41	217	122	466	269	146	248	5
N05	237	211	57	43	189	116	458	252	172	278	9
N06	241	149	76	25	219	82	412	317	215	317	13
N07	232	161	40	35	205	140	448	282	155	313	15
N08	268	122	73	34	253	101	374	303	194	312	3
N09	296	159	54	33	220	112	439	247	182	278	4
N10	237	215	63	55	223	107	427	258	133	300	8
K01	284	121	76	39	305	170	492	190	145	180	4
K02	198	157	85	53	243	118	323	319	220	270	7
K03	246	142	72	31	241	142	432	284	127	305	9
K04	263	146	61	30	280	126	457	261	98	277	6
K05	206	181	64	35	230	142	484	256	157	245	3
K06	221	217	54	62	200	112	305	340	168	307	4
K07	197	196	96	39	226	120	339	347	198	303	10
K08	230	178	90	38	209	94	333	384	172	271	13
K09	245	201	40	26	221	130	378	339	153	280	3
K10	229	194	91	34	237	115	378	296	167	288	6