LE CORPUS CONÇU COMME UNE BOULE

Étienne BRUNET
BCL (UMR 6039), Université de Nice

SOMMAIRE

La notion de corpus semble avoir un contour précis et fixe par quoi on l'oppose à d'autres notions plus incertaines, comme le genre, le discours, la langue. Il n'est pas toujours possible de décider qu'un texte appartient à un genre ou qu'un mot appartient à la langue. Mais quand un corpus est constitué, on sait sans contestation si un mot s'y trouve ou non et si un texte y est ou non incorporé. Pourtant, à la réflexion, quand on fixe son attention assez longtemps sur cet objet dur qu'est le corpus, la vue se brouille, l'objet fond et se ramollit comme les montres de Dali.
Tout d'abord un corpus est toujours artificiel. La nature n'en produit pas spontanément. C'est une création nécessairement subjective. Pire encore, la création est orientée, conditionnée par une hypothèse, par un objectif de recherche. Quelques précautions qu'on prenne pour affiner les critères de sélection, pour les justifier et pour les appliquer, il y a toujours des choix à décider, des doutes à faire taire, des contraintes à respecter, des compromis à négocier, un ordre à établir, un terminus *a quo*, un autre *ad quem* à délimiter. Et comme l'opération de traitement est plus rapide que celle de sélection, la tentation est grande de modifier la composition du corpus, au vu des résultats d'un premier traitement, en mêlant ainsi inextricablement et illégitimement les procédures subjectives de la sélection et les procédures objectives du traitement. En tailladant à propos dans le corpus comme un sculpteur dans la pierre, on finit par lui donner la forme souhaitée, compatible avec l'hypothèse initiale et dotée de la fausse garantie d'un traitement impersonnel.
Supposons cependant qu'un corpus ait réuni tous les suffrages, qu'une commission *ad hoc* ait établi les critères, qu'une autre commission indépendante ait procédé à leur application, et que tous les recours ou appels aient été épuisés. Acceptons l'idée d'un corpus à la pureté eucharistique. Reste à l'introduire dans la salle blanche du traitement. Mais ici de nouveau le pas est suspendu devant le seuil à franchir. Il y a plusieurs salles d'opération, plusieurs technologies disponibles, plusieurs logiciels à utiliser et un choix préliminaire à faire : faut-il soumettre le corpus à un traitement purement documentaire, ou à un traitement linguistique ou à un traitement statistique ? Ces distinctions n'ont pas de barrière fixe : beaucoup de logiciels proposent à la fois des fonctions documentaires et statistiques et certaines fonctions à objectif linguistique, comme la lemmatisation, peuvent emprunter la voie statistique.
Supposons le pas franchi et le logiciel idoine. On entre alors dans un labyrinthe. Perplexité devant le trousseau de clés qui s'affiche à l'entrée. On a souvent l'impression de pénétrer dans un jeu de rôles. Heureux l'expert qui sait utiliser la panoplie des outils, et résoudre le rébus des résultats. Passe encore pour les fonctions documentaires qu'un néophyte peut maîtriser sans grand effort. Mais dès qu'intervient la statistique, le non-initié reste perplexe devant les options à choisir, les traitements à opérer, les tableaux à constituer, les graphes à commenter. C'est pire encore lorsque le doute l'épargne et qu'il se promène sans vertige sur le parapet de l'interprétation. L'évidence visuelle dont se prévaut un simple histogramme peut cacher des pièges et des incertitudes : s'agit-il de fréquences absolues, ou relatives ou réduites ? Quelle méthode a-t-on utilisée pour le calcul de l'écart, la loi normale, la loi hypergéométrique ou quelque autre ? Et sur quelle référence, interne ou externe, s'appuie-t-on ? Il arrive parfois que le seul expert apte à débrouiller les fils soit l'auteur du logiciel utilisé. Et comme j'ai cet avantage pour le logiciel Hyperbase, on me permettra

d'en profiter. Qu'on se rassure : je n'abuserai pas de la situation pour décrire en détail le fonctionnement du logiciel. Je m'en tiendrai à une seule fonction : celle qui mesure la distance intertextuelle.

Dans une première approche Hyperbase suit la méthode Jaccard qui ne se préoccupe pas de fréquence et pour un mot donné ne considère que sa présence – ou son absence – dans le texte considéré. Ou plus exactement, pour deux textes dont on cherche à apprécier la connexion, un mot contribue à rapprocher ces deux textes s'il est commun aux deux et à augmenter la distance s'il est privatif et ne se rencontre que dans un seul. La collection des données est assez lourde parce qu'il faut considérer tous les mots sans exception et que pour chacun on doit prendre en compte tous les appariements de textes deux à deux (le nombre des confrontations pour n textes étant égal à n * (n-1) / 2). Pour chaque paire considérée, la distance obtenue tient compte de l'étendue de l'un et l'autre vocabulaires, selon la formule :

d = ((a-ab)/a) + ((b-ab)/b),

où ab désigne la partie commune aux vocabulaires a et b (a-ab et b-ab recouvrant les parties privatives). Chacun des deux quotients (dont la somme constitue la mesure de la distance) est le rapport, pour un texte donné, du vocabulaire exclusif au vocabulaire total. Il évolue nécessairement entre 0 et 1. La somme a donc pour limites 0 et 2. En réalité la somme se situe autour de 1 et reste insensible aux différences d'étendue des deux textes mis en parallèle. Le calcul est en réalité un peu plus complexe dans la dernière version d'Hyperbase. Il intègre non seulement les mots communs aux textes A et B et les mots privatifs qui se trouvent dans A sans être dans B et réciproquement, mais aussi les mots du corpus qui ne se trouvent ni dans A, ni dans B. Ces mots pareillement rejetés par les deux textes contribuent dans une certaine mesure à rapprocher, même négativement, les deux textes, puisqu'ils partagent les mêmes répulsions ou les mêmes désintérêts. Ces deux variantes du calcul n'épuisent pas les possibilités de la méthode Jaccard. On a compté jusqu'à vingt autres indices, tous fondés sur les mêmes ingrédients[1].

1. Convergence de deux méthodes d'analyse : factorielle et arborée

Une fois obtenu le tableau des distances, son exploitation peut emprunter deux voies différentes mais convergentes : l'analyse de correspondance (figure 1) et l'analyse arborée (figure 2) :

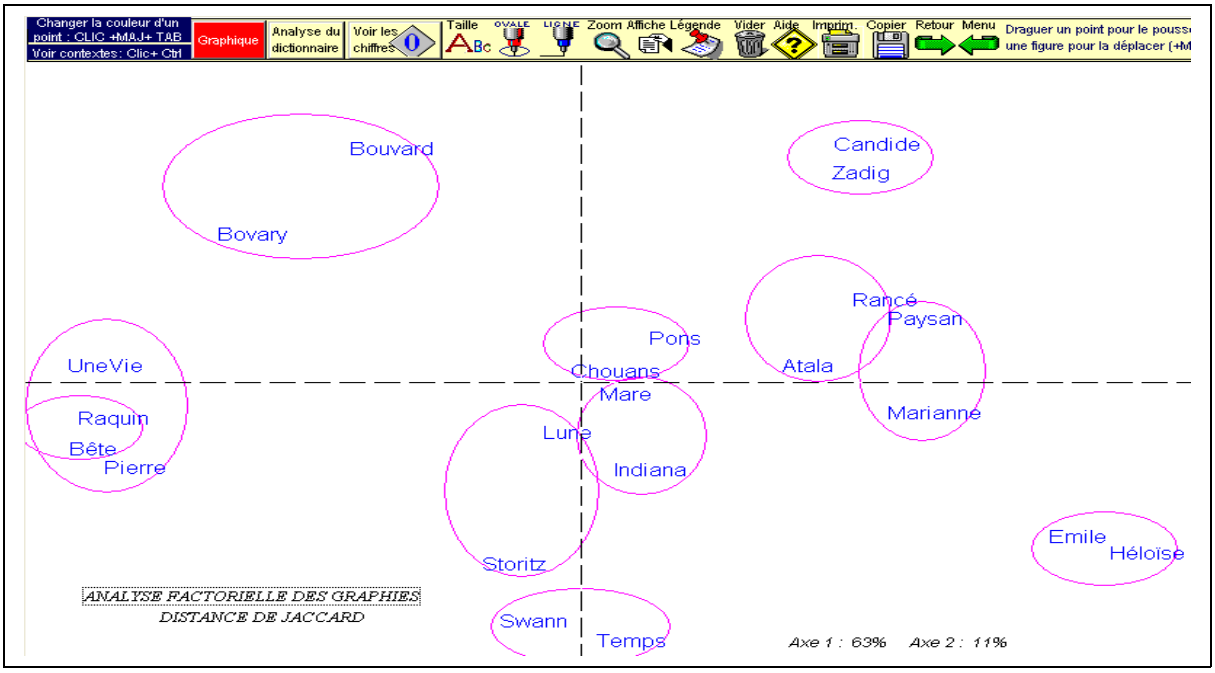
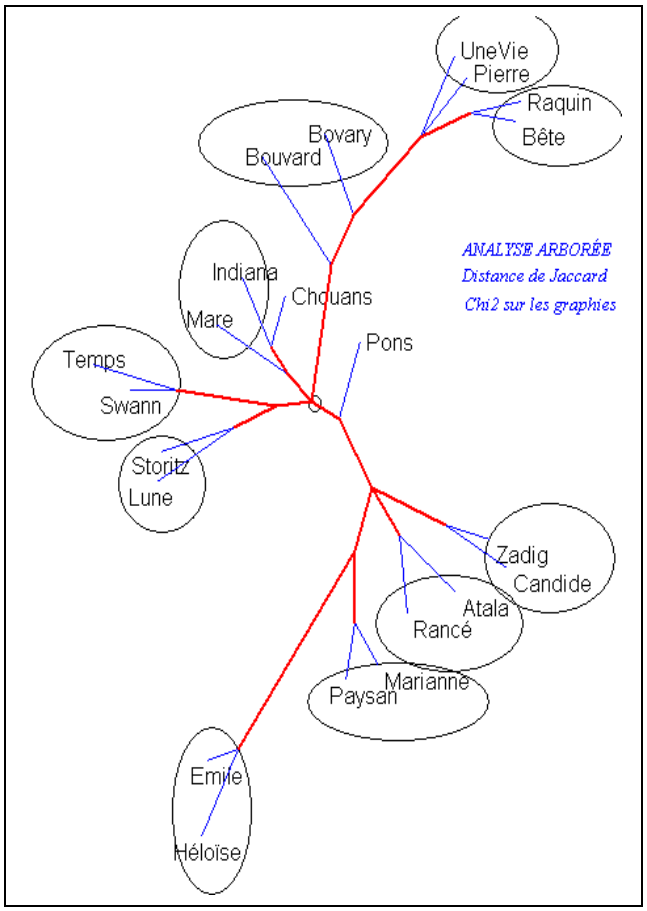


Figure 1. Analyse factorielle de la distance intertextuelle (analyse des graphies, méthode Jaccard)

[1] J.B. Baulieu, 1989, « A classification of Présence/Absence Based Dissimilarity Coefficients », *Journal of Classification*, 6, 233-246.

Qu'on ne cherche pas l'influence du genre qui est ici neutralisée : les 22 textes réunis dans ce corpus relèvent tous du genre narratif[2]. Restent deux variables qu'on a voulu croiser : la chronologie et l'écriture propre à chaque écrivain. L'analyse factorielle représentée dans la figure 1 paraît suivre la chronologie, puisque tous les textes du XVIIIe siècle et de la première moitié du XIXe se portent à droite, quand les textes les plus récents campent dans la partie gauche. Il y a pourtant l'exception remarquable de Proust et de Verne qui fuient la compagnie de Zola et des romanciers réalistes et se tiennent à cheval sur l'axe central. On voit ainsi que le tempérament d'un écrivain peut résister à la pression du temps. Pour mettre en évidence la force du lien qui lie chaque texte à son auteur, le corpus incorpore deux textes de chaque écrivain, situés au début et à la fin de sa carrière. Or l'analyse factorielle reconnaît aisément ce lien et place toujours à proximité les textes qui appartiennent à la même plume, ce qu'on peut vérifier dans la figure 1, où les binômes sont clairement identifiables : à droite Marivaux, Rousseau, Voltaire et Chateaubriand, à gauche Flaubert, Maupassant et Zola et au centre Balzac et Sand, et plus bas Verne et Proust.
Or l'analyse arborée, représentée dans la figure 2, donne, plus clairement encore, les mêmes enseignements. Aux deux bouts de la chaîne on rencontre les mêmes configurations que précédemment, avec une zone de transition indécise où flottent les textes de Balzac et de Sand.



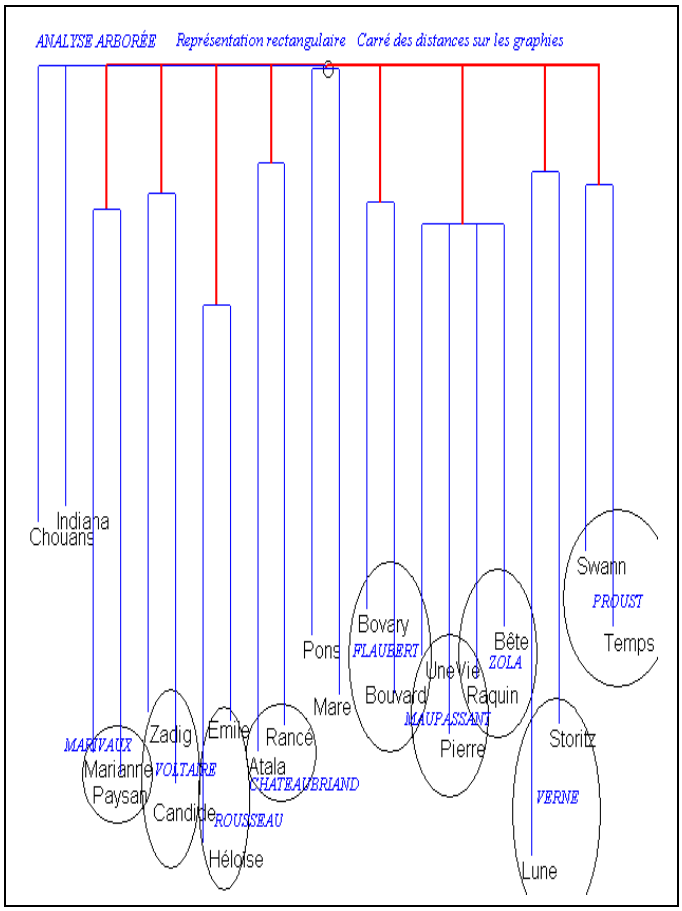


Figure 2. Analyse arborée (graphies, méthode Jaccard) Représentation radiale sur Chi2 des distances

Figure 3. Même analyse (graphies, Jaccard) Représentation rectangulaire sur le carré des distances

Et de la même façon Proust et J. Verne se tiennent à l'écart sur une branche latérale. Ici toute l'information du tableau des distances se trouve résumée au mieux, alors que dans l'analyse factorielle les deux premiers facteurs n'épuisaient pas l'inertie.

[2] Composition du corpus : Marivaux : *La Vie de Marianne* et *Le Paysan parvenu*, Rousseau : *La Nouvelle Héloïse* et *Émile*, Voltaire : *Zadig* et *Candide*, Chateaubriand : *Atala* et *La vie de Rancé*, Balzac : *Les Chouans* et *Le cousin Pons*, Sand : *Indiana* et *La mare au Diable*, Flaubert : *Madame Bovary* et *Bouvard et Pécuchet*, Maupassant : *Une Vie* et *Pierre et Jean*, Zola : *Thérèse Raquin* et *La Bête humaine*, Verne : *De la terre à la lune* et *Les secrets de Wilhelm Storitz,* Proust : *Du côté de chez Swann* et *Le Temps retrouvé*.

2. Convergence données brutes / données pondérées

La figure 3 offre une variante de l'analyse arborée. Il n'y aurait pas lieu de s'étonner qu'il y ait recoupement des deux représentations, radiale et rectangulaire, si les données étaient exactement les mêmes. Le tableau des distances absolues est bien commun aux deux analyses mais dans un cas (figure 2) il a été pondéré et transformé en profil, grâce au calcul du Chi2, dans l'autre (figure 3) il a été accentué, chaque distance étant portée au carré. Or, pondérées, amplifiées, ou brutes, les données s'organisent de la même façon. L'avantage de la transformation est toutefois de rendre plus claire la décantation et de mieux marquer oppositions et rapprochements.

Mais la méthode Jaccard fait peut-être la part belle aux raretés du vocabulaire et particulièrement aux hapax, au détriment des fréquences plus courantes. Les classes de fréquence élevée perdent ainsi tout poids dans le calcul, puisqu'elles se trouvent nécessairement dans la partie commune et inévitable du vocabulaire (*ab*). Ce calcul peut être jugé trop sensible aux artefacts que peuvent produire l'inconstance de l'orthographe, les fautes de frappe, l'abondance des noms propres, bref tous les phénomènes, parfois mineurs et négligeables, qui engendrent la multiplication des formes. Certains considèrent que c'est donner trop d'importance à l'excentricité et qu'une véritable appréciation de la distance entre deux textes doit considérer, pour un même mot, le dosage des fréquences dans les deux textes comparés.

3. Convergence de deux calculs de distance : Jaccard/Labbé

Or Dominique Labbé, a proposé un algorithme efficace qui pour chaque mot apprécie la distribution réelle des fréquences dans les deux textes A et B en comparant les fréquences observées non plus à la répartition théorique mais à l'écart maximal possible dans cette distribution : $D_{(A,B)} = \sum d_i \ / \sum dmax_i$ pour i variant du premier au dernier mot du vocabulaire.

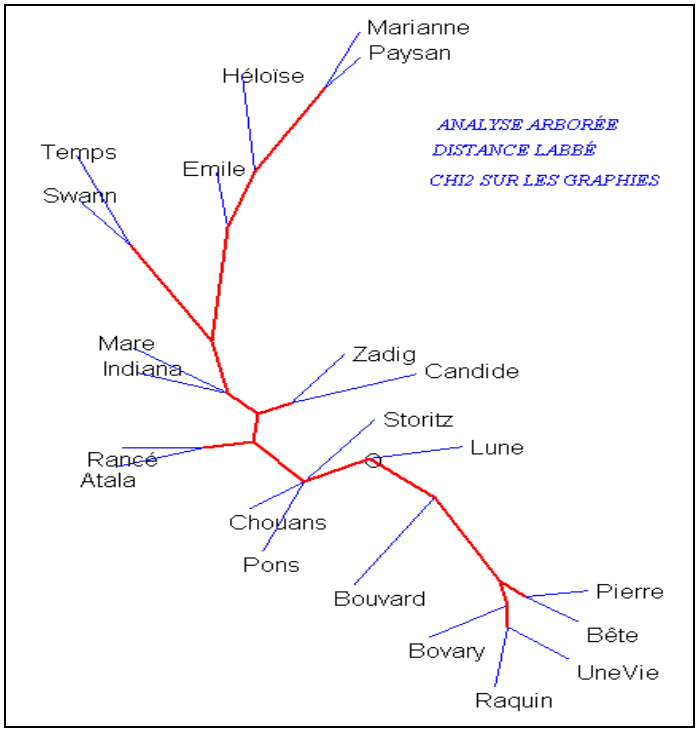


Figure 4. Analyse arborée des graphies. Méthode Labbé

On prendra la mesure de l'écart en rapprochant la figure 4 de la figure 2. Il faut reconnaître qu'en l'occurrence l'écart est faible et que les mêmes lignes de force s'y dessinent, si l'on néglige une inversion verticale qui ne tire pas à conséquence. Même s'il tient compte des moyennes et hautes fréquences, le coefficient de Labbé est surtout sensible, comme celui de Jaccard, aux mots de basse fréquence, qui sont les plus nombreux. Et l'expérience montre que, quel que soit le corpus étudié, les deux coefficients mettent en relief les mêmes influences s'exerçant dans le même sens et avec la même intensité.

Jusqu'ici la convergence des résultats n'est pas en soi un progrès car c'est toujours le même objet qu'on a soumis aux expériences méthodologiques. On a supposé acquis le tableau des distances.

On a seulement varié les éclairages, les prises de vue et le traitement de l'image en laboratoire. Mais le tableau des données peut et doit être remis en question. On n'a considéré que les graphies pour apprécier la distance intertextuelle. Mais les graphies sont un matériau dégradé et désintégré qui ne donnent qu'une image plate et déconstruite du texte. Ne pourrait-on pas affiner le produit en séparant ce qui doit l'être, les homographes, et en rassemblant ce qui doit l'être, les formes qui se rattachent à la même entrée du dictionnaire ?

4. Convergence des graphies et des lemmes

On a donc lemmatisé les deux millions de mots du corpus, en utilisant *Cordial*. Le résultat, inscrit dans la figure 5, reflète en miroir l'image de la figure 4. Le recouvrement des deux graphiques est presque parfait. Non seulement les textes d'un même auteur sont placés pareillement à proximité l'un de l'autre, mais les rares décrochages observés dans le graphique des graphies se retrouvent dans celui des lemmes : les deux textes de Flaubert, de Verne et de Rousseau restent proches mais n'ont pas un lien direct. Et dans les deux analyses les textes de Zola et de Maupassant se mêlent les uns aux autres.

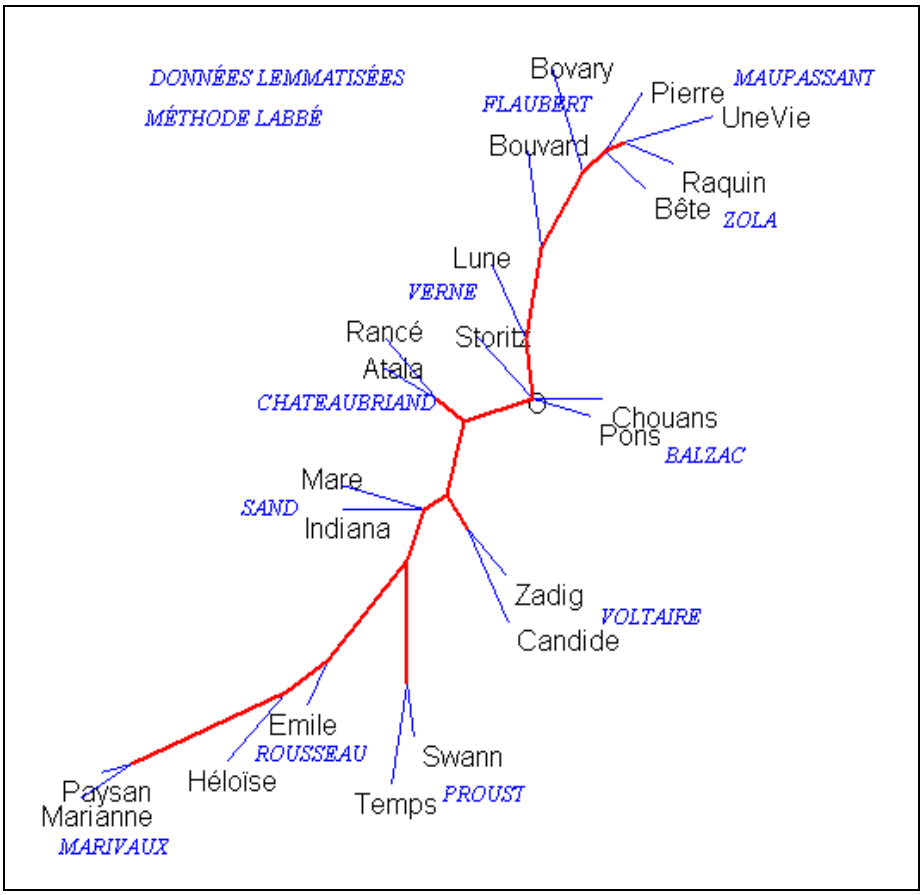


Figure 5. La distance intertextuelle fondée sur les lemmes (méthode Labbé)

5. Convergence des mots et des codes grammaticaux

Ainsi les 22 textes de notre corpus se répartissent de la même façon lorsqu'est mesurée la distance entre leurs vocabulaires, lemmatisés ou non. La distance intertextuelle peut aussi être appréciée en dehors de toute influence thématique, en observant uniquement la distribution des codes grammaticaux ou des structures syntaxiques, indépendamment des mots auxquels ces codes ou structures sont attachés. Comme chacun des quatre niveaux d'observation peut donner lieu à un calcul fondé sur la fréquence (méthode Labbé) ou sur la présence/absence (méthode Jaccard), on dispose en fin de compte de huit points de vue qui heureusement convergent. On préfèrera toutefois la méthode Labbé s'il s'agit des codes car la variété y est limitée et les effectifs importants (figure 6), et la méthode Jaccard pour la raison inverse s'il s'agit de structures (figure 7).

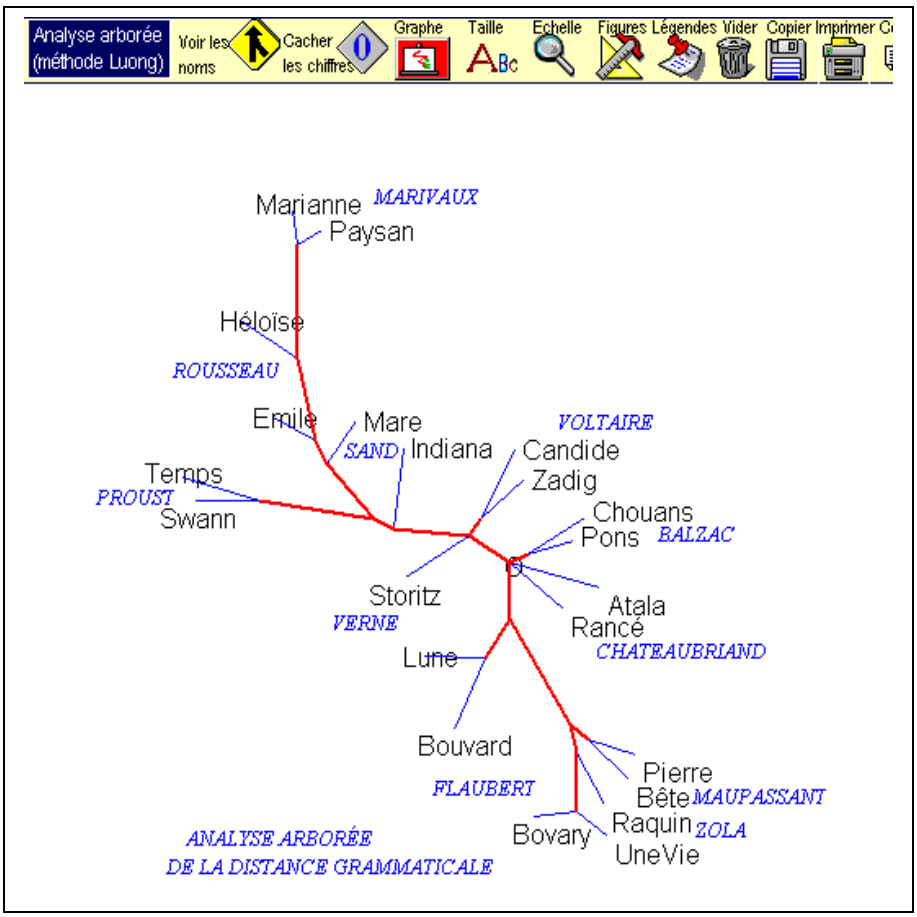


Figure 6. Analyse factorielle des codes grammaticaux (méthode Labbé)

Cette fois la convergence ne laisse pas d'étonner. Si l'on peut comprendre le parallélisme des mots-graphies et des mots-lemmes, car les premiers sont inclus dans les seconds, on ne voit pas *a priori* quel lien nécessaire pourrait être établi entre les lemmes et les codes grammaticaux. Dans l'effectif du lemme *aimer*, il y a certes l'imparfait *aimait*, mais aussi toutes les autres formes du verbe. Et dans l'effectif du code *imparfait 3e personne du singulier* il y a certes la forme *aimait*, mais aussi des centaines d'autres, comme *avait*, *était*, parfaitement étrangères au verbe *aimer*. Lemmes et codes grammaticaux se présentent en principe comme des variables indépendantes, les premiers plutôt thématiques et les seconds plutôt stylistiques. Les uns et les autres sont pourtant traversés par des courants semblables, où le tempérament des écrivains se manifeste, tantôt cédant, tantôt résistant à la dérive du temps. L'acte d'écrire ne s'exerce pas en deux temps, le choix du style succédant à celui du thème, comme on fait pour l'achat d'une voiture, la sélection de la couleur venant après celle du modèle. L'écriture implique un choix simultané, cohérent quoique souvent inconscient, des variables thématiques et stylistiques.

6. Les structures syntaxiques

La désincarnation est poussée plus loin encore dans la figure 7 qui est relative aux structures syntaxiques et où la charge sémantique des mots et des textes est complètement évacuée. Tous les schémas syntaxiques rencontrés entre deux ponctuations sont relevés dans le corpus, catalogués et cumulés. Ce qu'on prend en compte n'est plus le dosage des parties du discours, mais leur assemblage et le rythme de la segmentation. Quelques particularités apparaissent comme le déplacement de Proust et Voltaire. La longue phrase du premier le rapproche de l'époque classique, alors que le second échappe à son temps et annonce la modernité. Mais ces retouches de détail ne perturbent guère le mouvement d'ensemble qui reproduit les figures précédentes.

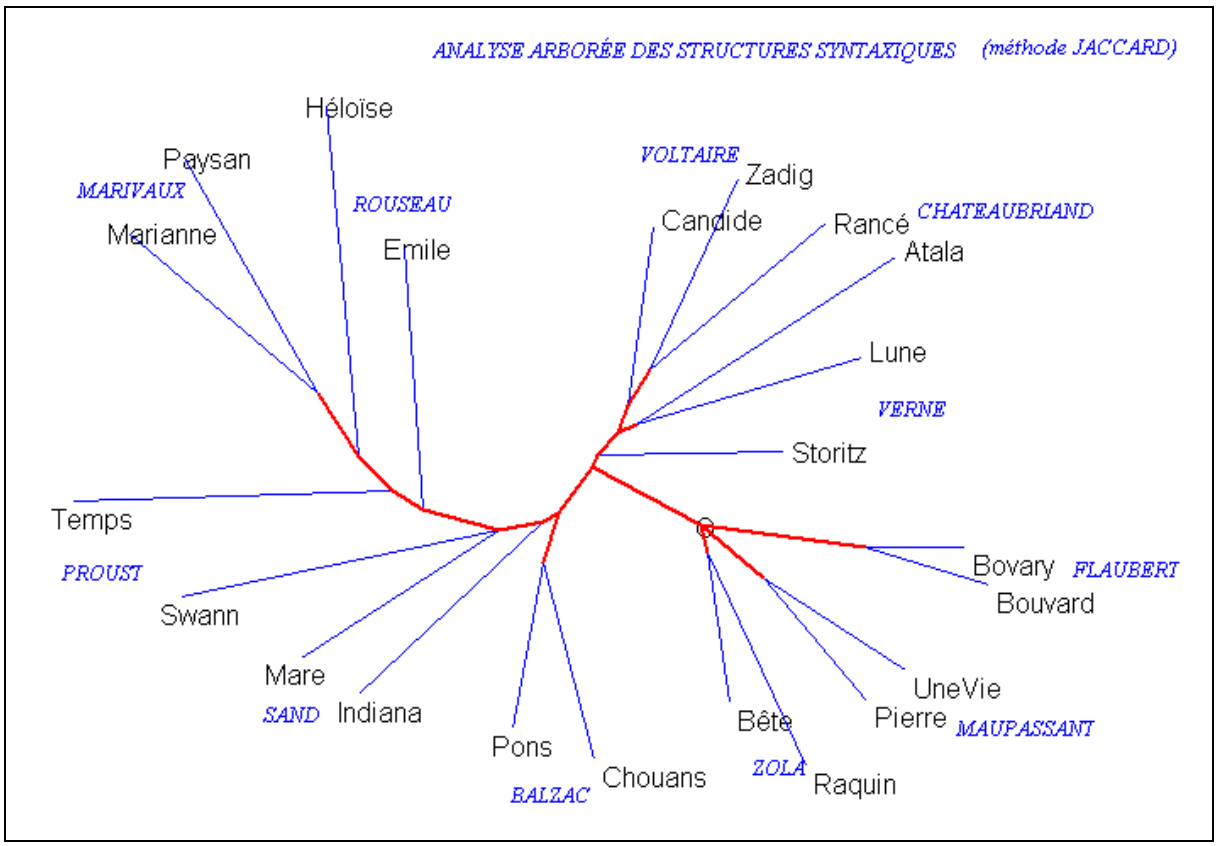


Figure 7. Analyse factorielle des structures syntaxiques (méthode Labbé)

7. Les codes sémantiques

Selon qu'on utilise les filtres appropriés, à l'échelle macroscopique ou élémentaire, un corpus peut être envisagé comme un ensemble de textes, ou de phrases, ou de graphies, ou de lemmes, ou de codes grammaticaux, ou de structures syntaxiques, ou de balises sémantiques. Ce dernier niveau est le plus difficile à atteindre, car le sens des mots échappe en grande partie à la machine. Le lemmatiseur *Cordial* propose pourtant une ontologie extérieure qui distribue des étiquettes sémantiques aux mots-pleins du corpus. Cet étiquetage est discutable, certaines balises sont curieusement nommées et leur attribution est approximative. Pourtant l'analyse factorielle appliquée à de telles données reprend pour l'essentiel la typologie observée dans le corpus. Comme précédemment, la figure 8 oppose aux autres les représentants du roman réaliste et naturaliste. Elle ajoute toutefois une information précieuse : la carte des thèmes est superposée à celle des textes et l'interprétation s'en trouve facilitée. Car la proximité d'un texte et d'un thème acquiert une signification. Ainsi les œuvres de Flaubert, Maupassant et Zola sont tournées vers la description du milieu et des réalités concrètes, matérielles, corporelles, qu'on devine derrière les thèmes qui les entourent : *concret, quotidien, production, corps, sens, santé, vivant, espace, cinétique* (= mouvement). À l'opposé la littérature classique (c'est là aussi que Proust prend place) se préoccupe davantage des réalités morales, psychologiques, religieuses, sociales ou politiques (*éthique, spiritualité, droit, volonté, esprit, homme, économie, pouvoir, société*).

Le corpus conçu comme boule

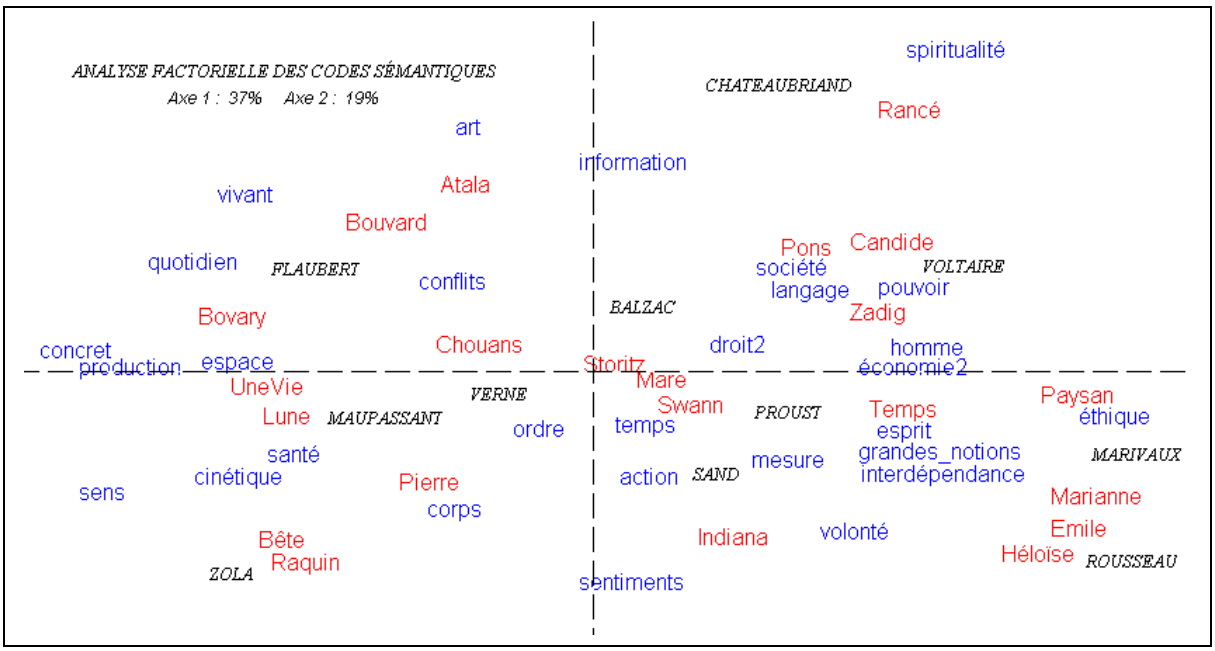


Figure 8. La carte thématique du corpus

8. L'expérience des n-grammes

Chacun des jalons sur lesquels s'appuie la segmentation peut être neutralisé ou déplacé, qu'il s'agisse de ceux qui séparent les textes, ou les phrases, et même des séparateurs qui isolent les mots. Imaginons un Champollion devant une immense pierre de Rosette où les mots n'auraient pas de frontières distinctes et qui contiendrait les dix millions de caractères de notre corpus. Et supposons que devant ce texte inconnu on ait promené une loupe de proche en proche en isolant quatre lettres à la fois, et en déplaçant la fenêtre d'une seule case à chaque pas. Ainsi le mot *fenêtre* génèrera quatre n-grammes successifs : *fenê*, *enêt*, *nêtr* et *être*, qui tiendront lieu de « mots ». Cette fois, au lieu d'enrichir le texte en le dotant de codes grammaticaux ou sémantiques, on l'appauvrit jusqu'à le rendre illisible. Le blanc ayant disparu, les mots ne sont plus reconnaissables, n'ayant ni queue ni tête. Et pourtant ce rébus opaque ne pose aucun problème au programme de reconnaissance, qui retrouve les textes issus de la même plume et dresse une carte d'attribution aussi claire que celle des lemmes. Le graphique obtenu sur ces données perverties est superposable en tous points à ceux que le matériau linguistique épuré avait produits (notamment les figures 2, 4, 5 et 6). Avant de nous interroger sur cette stabilité étonnante des traitements statistiques, poursuivons jusqu'à l'extrême notre jeu de déconstruction. Après avoir cassé les mots, cassons l'alphabet. Négligeons les accents et oublions les lettres. Procédons comme une sténo indolente ou inculte qui n'aurait à sa disposition que deux symboles pour noter ce qu'elle entend : le signe V pour une voyelle et le signe C pour une consonne.

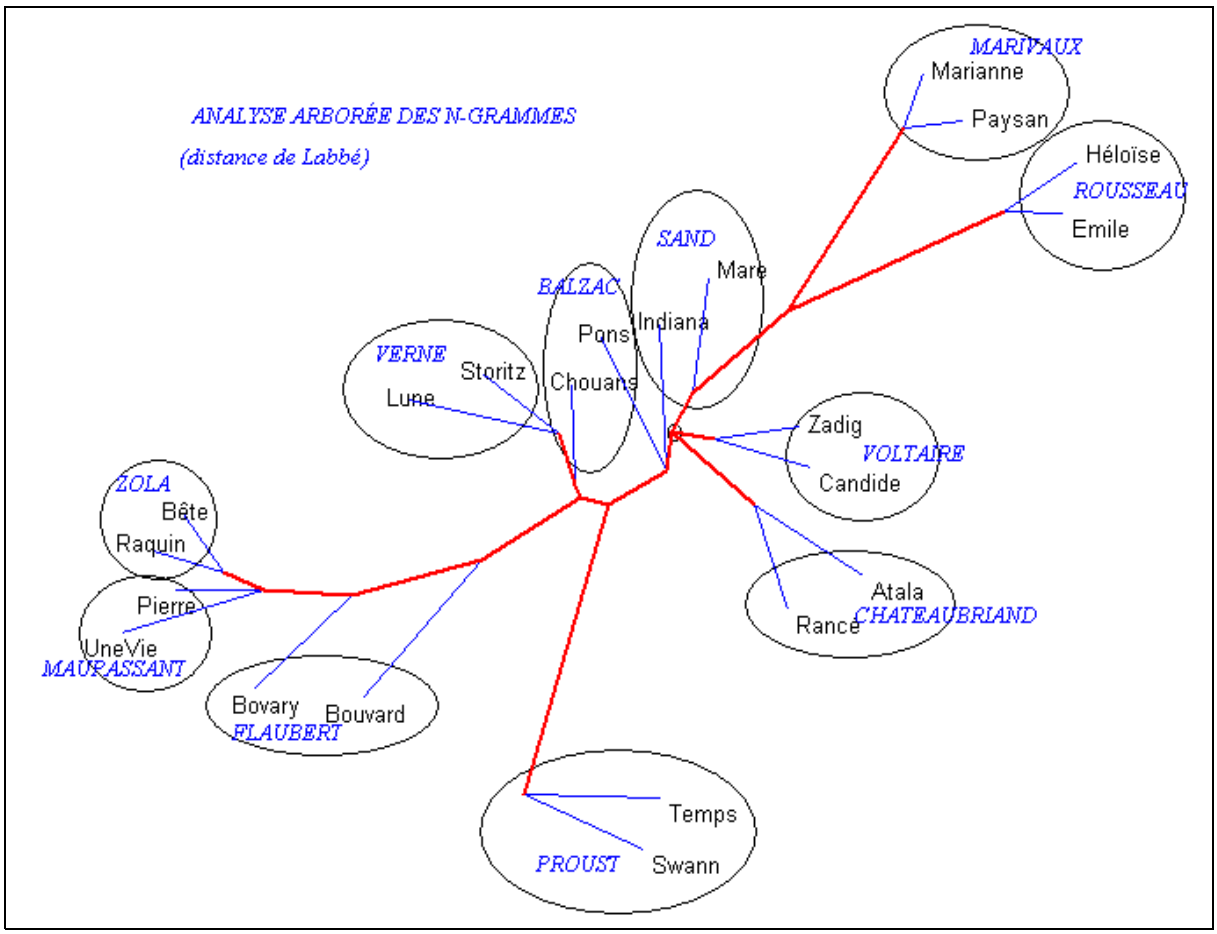


Figure 9. Analyse arborée des n-grammes

9. L'expérience ultime Consonne/Voyelle

Le résultat de cette réduction drastique apparaît ci-dessous : difficile de reconnaître la dernière ligne de la *Recherche du temps perdu* dans cette suite de CV.

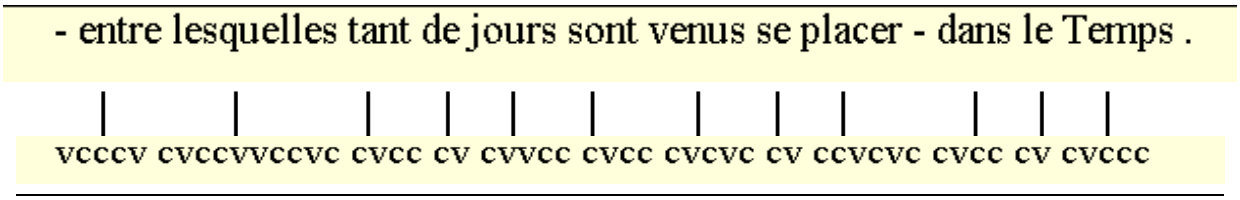


Figure 10. La dernière ligne du Temps retrouvé *réduite à une succession de voyelles et de consonnes*

La perte d'information semble irrémédiable : tous les mots de deux lettres n'ont le choix qu'entre trois combinaisons, CV, VV, et VC. Ceux de trois lettres ont un choix à peine plus ouvert. C'est dire que tous les mots-outils sont quasiment confondus. À elles seules les trois premières combinaisons, à savoir CV, CVC et VC, représentent le tiers de la surface imprimée. Et inversement, avec un alphabet aussi pauvre, les combinaisons rares se raréfient encore. Il n'y a plus que 607 hapax, contre 19156 dans le texte original. Et pourtant le miracle se produit : imperturbable, la machine arrive à démêler le nœud gordien et à proposer une typologie des textes qui s'écarte à peine de celles qu'on a obtenues avec un matériau cent fois plus riche et plus précis. La plupart des binômes ont un lien direct, ce qu'on observe pour Marivaux, Voltaire, Chateaubriand, Balzac, Maupassant et Proust. Ailleurs la liaison est courte même si elle n'est pas immédiate. Le mouvement d'ensemble est grossièrement respecté. Les textes du XVIIIe forment un bloc, ceux du roman réaliste un autre, et l'irrédentisme de Proust, à l'écart sur une branche latérale, est bien visible.

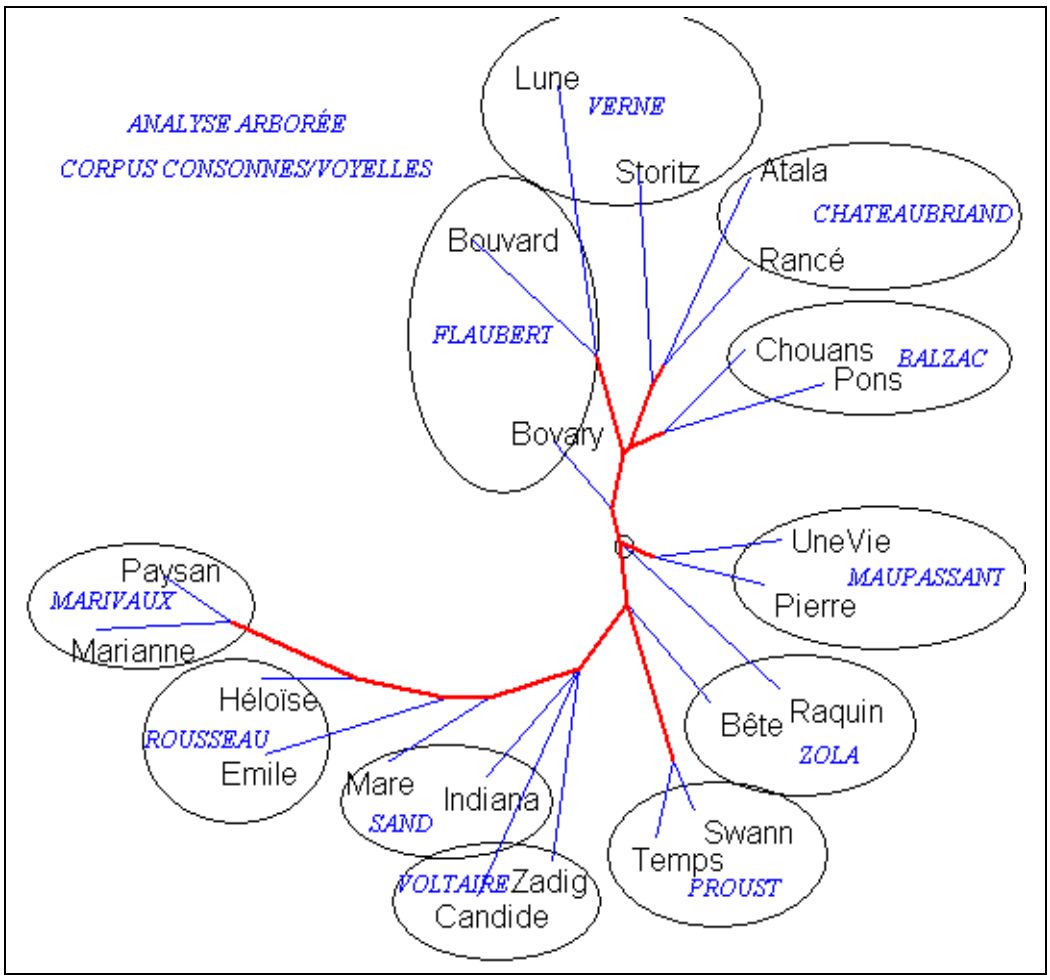


Figure 11. Analyse arborée du corpus consonnes/voyelles

Réduit à une combinaison de trois éléments – consonne, voyelle et blanc – le corpus prend l'allure du génome et les mêmes méthodes de décryptage pourraient s'y appliquer. Jean-Pierre Anfosso[1] dans sa thèse a montré qu'on pouvait y parvenir, quand l'emploi d'un alphabet restreint autorise le traitement des chaînes de Markov.

L'appareillage statistique, appliqué au langage, apparaît ainsi d'une remarquable stabilité, jusqu'à provoquer le soupçon que c'est toujours la même chose qu'on mesure. Il produit des résultats convergents à des niveaux très éloignés et très variés, depuis les regroupements ontologiques les plus larges – et les plus flous – jusqu'aux analyses les plus microscopiques des molécules et des atomes du langage. Les mêmes lignes de force s'y reconnaissent, quelle que soit la focale utilisée ou l'éclairage ou l'angle de la prise de vue. Le corpus est comme une boule : qu'on le considère d'en haut ou d'en bas, de la droite ou de la gauche, de l'avant ou de l'arrière, l'image est la même.

[1] J.P. Anfosso, *Contribution à une modélisation statistique du langage et à sa mise en œuvre informatique*, Nice, nov. 2002.