Formalisation et quantification des textes. Le domaine français

Étienne Brunet

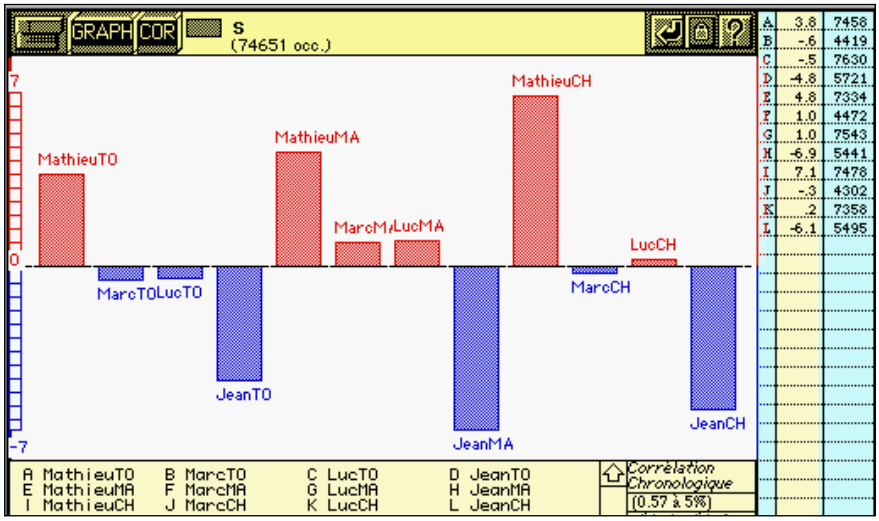
Formalisation et quantification des textes. Le domaine français.

Etienne Brunet

La statistique, c'est le cousin Pons des mathématiques : un parent pauvre qui ne fait guère honneur à la famille. La noblesse de la profession est dans la théorie, dans la chevauchée logique et déductive qui conquiert le terrain et impose sa loi à la réalité, au lieu que la démarche rampante et inductive des mathématiques appliquées et particulièrement de la statistique est le lot des esprits faibles qui ne s'aventurent jamais au-dessus du sol et restent esclaves des observations. Une loi physique est une prédiction, presque une prescription, à quoi le réel doit se conformer. Une loi statistique est au mieux une prévision qui n'offre pas de garantie certaine, et qui n'a guère d'emploi que dans les sciences molles, là où le jeu de la liberté humaine et l'enchevêtrement inextricable des paramètres politiques, économiques, sociaux et culturels empêchent d'isoler clairement les causes et les effets. Même dans ces disciplines moins favorisées, la statistique n'a pas toujours grand crédit. Elle est certes un passage obligé pour la recherche médicale, la maîtrise de l'économie ou l'intelligence des rapports sociaux, mais son rôle y apparaît comme auxiliaire ou ancillaire : cantonnée dans la cuisine pour éplucher les légumes et les données, elle est rarement invitée au salon et au débat. À plus forte raison s'il s'agit d'un débat linguistique ou littéraire. Dans le domaine des lettres et des arts, on suspecte a priori l'intervention des chiffres et le chercheur imprudent qui se livre à ce jeu peut s'attendre aux quolibets que Zola réserve à l'un de ses héros : « Vous n'imagineriez jamais à quoi Mouret passe son temps dans la pièce où il s'enferme ? (…) Eh bien ! Il compte les *s* qui se trouvent dans la Bible. Il a craint de s'être trompé, et il a recommencé trois fois son calcul (…) Ma foi ! Vous aviez raison, il est fêlé du haut en bas, ce farceur-là! »[1]

[1] *La Conquête de Plassans*, édition de la Pléiade, tome 1, p. 1127.

La fêlure s'étant propagée au siècle suivant, je suis en mesure de vous livrer le résultat du calcul besogneux de Mouret. À vrai dire, on s'est contenté de l'Évangile, dans trois traductions françaises récentes. Zola n'avait probablement pas imaginé que la répartition des S puisse varier significativement selon qu'on a affaire à Saint Matthieu ou à Saint Jean. Comme nous retrouverons l'Évangile plus loin dans cet exposé, nous n'irons pas plus avant dans le commentaire de cette découverte dérisoire. Nous observerons seulement, à partir de cet exemple, que la statistique a des vertus heuristiques insoupçonnées. Si son usage ordinaire est de confirmer ou d'infirmer des hypothèses, il arrive qu'elle mette son doigt inquisiteur sur des faits que le soupçon n'avait pas effleurés. L'anomalie constatée n'est souvent que la trace lointaine, indirecte et négligeable d'un phénomène sous-jacent qui reste à élucider. Mais on est lancé sur une piste, et d'indice en indice, on peut poursuivre la chasse.



Graphique 1. La chasse aux S dans trois traductions des quatre évangiles

1. Les Textes

1- Mais avant d'aborder le domaine quantitatif, il convient d'explorer les ressources textuelles qui s'offrent à la lecture et à la mesure. Internet les dispense à foison et il est devenu aisé, dans la limite du copyright, d'y chercher et d'y trouver des textes de toutes sortes, dans l'un des formats courants (DOC, TXT, RTF, PDF, EPS, SGML, XML, HTML, etc), sans parler du mode « image » qui ne permet aucune

manipulation du texte et que nous n'envisagerons pas. Le format le plus simple – mais aussi le plus pauvre – correspond à l'option « texte seulement » que les traitements de texte proposent pour l'enregistrement d'un document. C'est aussi le format le plus ouvert qui s'offre innocemment à la lecture et au téléchargement, sans cryptage, sans clé, sans balises ni table de description. Il reste toutefois à l'utilisateur la charge de préparer à sa convenance le texte ainsi transmis et de le soumettre au logiciel de son choix. Beaucoup de sites cependant ne se contentent pas de proposer des textes à la lecture et à la copie. Les meilleurs constituent des bases textuelles qu'on peut interroger à distance sans rapatrier le corpus, des formulaires permettant la sélection. Le plus généralement il s'agit de bases CGI qui communiquent avec le client en utilisant pour le dialogue le protocole lié au langage HTML. Mais il peut s'agir aussi de bases purement HTML, uniquement constituées de pages et de liens. Nous avons eu l'occasion, ici même, il y a quelques années, de montrer ces deux types de bases, que nous avons réalisées, l'une autour de Rabelais, l'autre de Balzac. Depuis lors des progrès ont été enregistrés et le meilleur exemple qui puisse illustrer la puissance des techniques hypertextuelles peut être trouvé en Italie sur le site *Intratext* (www.intratext.com).

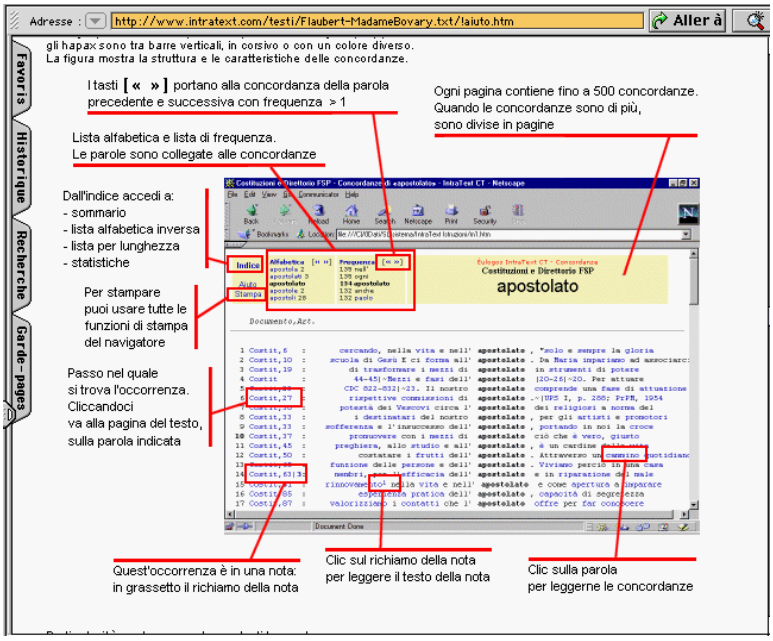


Figure 2. Un exemple d'hypertexte intégral : le site Intratext

Toutes les pages proposées à la lecture, et qui plus est, tous les mots de la page, sont sensibles au clic de la souris, chaque mot étant pourvu d'un lien qui renvoie à l'endroit ainsi désigné de la concordance générale, dont chaque élément dispose à son tour d'adresses multiples pour retourner au texte ou à quelque autre endroit de la concordance. Il est difficile d'imaginer une circulation plus rapide et plus souple, puisqu'à chaque pas les deux directions, verticale et horizontale, sont disponibles à la fois.

Les vertus documentaires que le site *Intratext* a empruntées au logiciel EULOGOS ont certes un prix : le texte transmis sur Internet a un poids quintuplé, que la concordance décuple encore. Mais cet encombrement reste sous-jacent et, l'écran n'en laissant aucune trace, seul le temps de transmission en porte témoignage. En revanche les informations statistiques sont d'une grande sobriété et leur intérêt reste limité, comme on peut le voir dans l'exemple ci-dessous qui rend compte de *Madame Bovary*[2].

- **Occurrences**: 120334
- **Words**: 13202
- **Occ./Words**: 9,11
- **Empty occurrences**: 26097
- **Empty words**: 47
- **Average occurrence length**: 4,49
- **Average word length**: 7,78

Figure 3. Les informations quantitatives dans *Intratext* à propos de *Madame Bovary*

2- Les renseignements d'ordre quantitatif sont radicalement absents de la base *Gallica* que la *Bibliothèque Nationale de France* a constituée en puisant dans les fonds de l'INaLF et de certaines maisons d'édition (Bibliopolis, Acamédia, Champion). Comme on peut l'espérer d'un organisme officiel, l'offre est abondante : plus de 2 000 textes en mode texte et beaucoup d'autres en mode image. Si ces derniers restent inaccessibles au moteur de recherche[3], mis à part leur table des matières et la légende de certains documents iconographiques, la sélection s'exerce pleinement, non seulement sur les critères bibliographiques mais même dans le corps du texte intégral. Une fois rempli le formulaire (étape

2 Il est vrai que quelques histogrammes rudimentaires sont aussi fournis qui rendent compte des classes de fréquences et de la longueur des mots.

3 La BNF utilise le moteur *Trevi* de la société Bibliopolis.

1 de la figure 4), les références obtenues apparaissent sur l'écran (étape 2), où l'utilisateur opère son choix (étape 3).

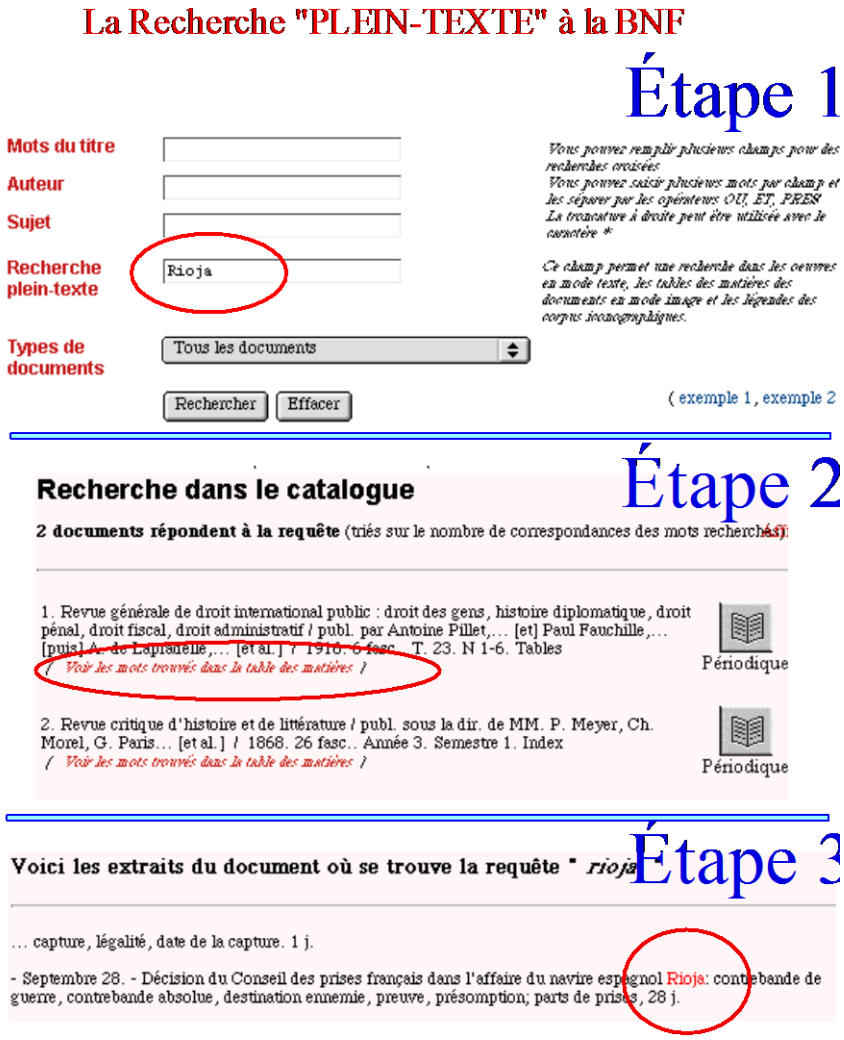


Figure 4

3- Il y a bien d'autres sources pour alimenter le réseau et distribuer des textes français. Citons parmi beaucoup d'autres les sites ABU, ATHENA, CLINET, la Bibliothèque de Lisieux. Aucun cependant ne peut rivaliser avec la BNF pour la capacité du réservoir et la puissance du débit. La plupart se contentent de proposer leur catalogue sans valeur ajoutée. N'ayant pas de moteur de recherche, ils ne jouent qu'un rôle

d'entrepôt intermédiaire. Il en va tout autrement de *Frantext* qui a hérité des données du *Trésor de la Langue Française* et les a fait fructifier au centuple. *Frantext* a toujours répugné à assumer la simple fonction de distributeur automatique de textes, d'autant que son stock est constitué de nombreux titres trop récents pour échapper aux droits d'auteur. *Frantext* n'est pas un entrepôt, mais une banque, et la base textuelle qui est mise en œuvre a longtemps passé pour un modèle, non seulement pour l'étendue, mais aussi pour la cohérence et l'homogénéité du corpus, et pour la puissance et la rapidité du logiciel d'interrogation et d'exploitation. Comme le créateur de cette base, Jacques Dendien, va prendre la parole dans un instant, il est plus qualifié que moi pour en expliquer les fonctions. Mais je ne puis taire l'admiration que la communauté scientifique voue à son produit et qui tient à son extraordinaire puissance. Depuis dix ans, chacun a pu faire l'expérience du logiciel et reconnaître sa robustesse, sa souplesse, sa vitesse, l'étendue de son champ d'action et la richesse des résultats obtenus. Il suffit d'une seconde pour interroger un ensemble de 180 millions de mots ou n'importe quel sous-ensemble de ce corpus, la question pouvant porter sur une forme, une expression, une cooccurrence, un lemme, une liste, une alternative, ou n'importe quelle combinaison de ces objets régie par les opérateurs booléens. La figure 5 qui détaille les entrées possibles ajoute une nouveauté : les catégories grammaticales. Une grande partie du corpus a en effet bénéficié d'un étiquetage, ce qui multiplie les critères de sélection et permet de cibler l'objectif en alliant variables et constantes, portée et précision.

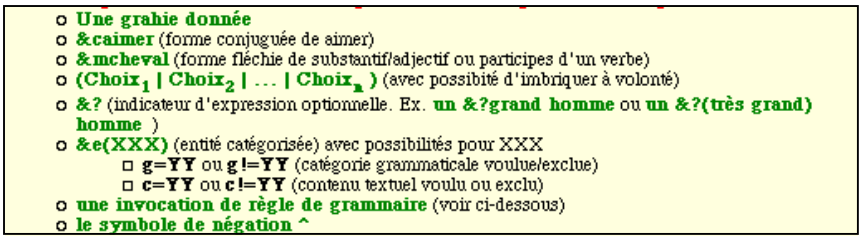


Figure 5. Objets traités par *Frantext* (toute combinaison de ces objets étant valide)

Pour pallier la complexité de l'interrogation et prévenir l'effroi que trop de parenthèses peuvent provoquer chez les littéraires, sinon les linguistes, une grammaire est disponible dont on fera usage lorsque les tâches sont répétitives. Le dialogue s'établit alors sur des règles convenues dont le contenu est à la charge de l'utilisateur. L'exemple ci-dessous (tableau 6) illustre l'ordre générique qu'on peut donner lorsqu'on

souhaite extraire les constructions où intervient un verbe pronominal, qu'il s'agisse d'un énoncé affirmatif ou négatif. Il serait facile d'ajouter quelques règles supplémentaires pour traiter les énoncés interrogatifs ou pour faire une place aux formes LA, LE, LES, EN ou Y qui peuvent s'intercaler dans la chaîne.

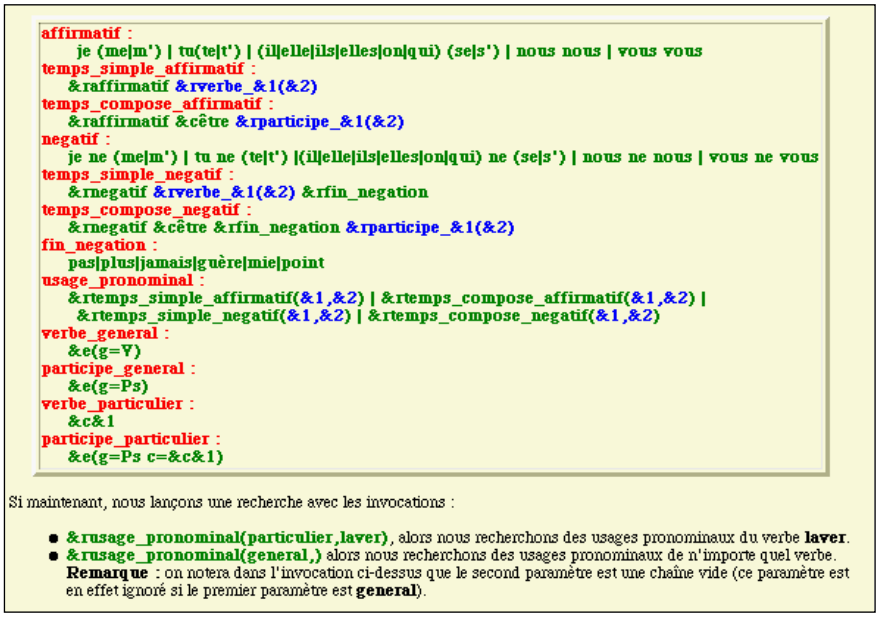


Figure 6. La grammaire de *Frantext*

2. Corpus annotés, étiquetés, lemmatisés, arborés

Ainsi se trouve comblé le retard de la France en matière de lemmatisation. Les premiers apôtres dans ce domaine avaient pourtant prêché en faveur des corpus désambiguïsés et pourvus de codes grammaticaux. Muller n'a cessé de recommander de soumettre les données à cette contrainte préalable. Et certains chercheurs scrupuleux comme Dominique Labbé, dans ses travaux sur de Gaulle et Mitterrand, ont maintenu, sabre au clair, cette exigence. Mais le gros de la troupe n'a pas suivi. Les corpus devenus très gros auraient imposé un effort jugé surhumain. Et surtout manquaient pour le français les logiciels d'étiquetage qui auraient diminué la part manuelle des corrections. Si les données de *Frantext* ont reçu comme on vient de le voir un traitement

approprié, le logiciel utilisé[4] exige un environnement Unix et n'est pas encore disponible à l'extérieur du laboratoire.

1- Il en existe certes un autre qui est adapté à l'environnement Windows et que l'on obtient facilement en s'adressant à l'INaLF. Dérivé des travaux de Brill au MIT, il a été adapté au français par deux chercheurs de Nancy, J. Lecomte et G. Souvay. Et nous l'avons intégré dans une version spéciale du logiciel HYPERBASE. On trouvera ci-dessous le détail du jeu d'étiquettes, dont certaines pourront créer la surprise. Aussi avons-nous opéré après coup des regroupements de façon à obtenir une liste de codes simplifiés et plus traditionnels qui facilitent l'interrogation et la quantification. Comme leur nombre est réduit à une douzaine, la combinatoire est moins dispersée et les effectifs s'accroissent d'autant.

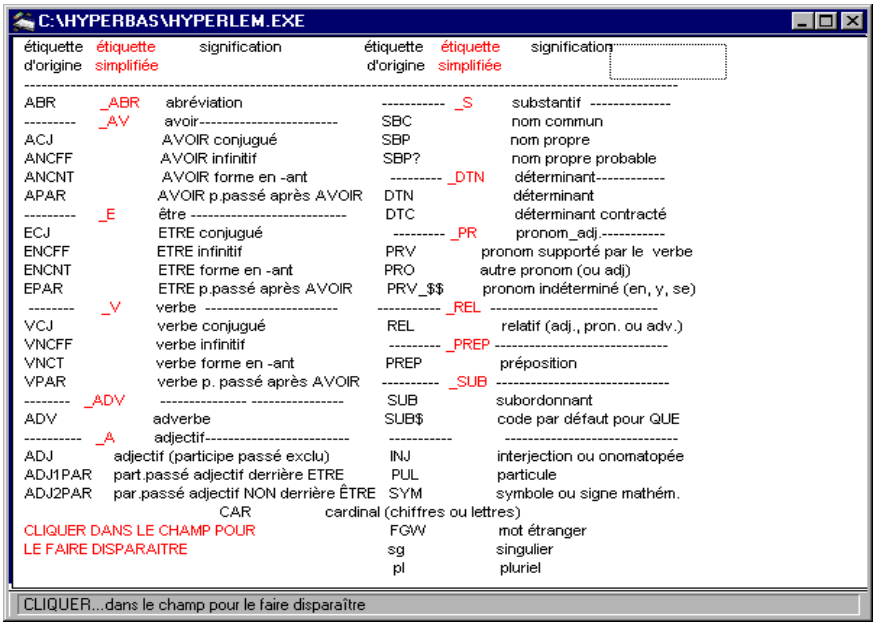


Figure 7. Les codes grammaticaux de *Winbrill*

Le relevé des structures lexicales s'opère avec la souris en choisissant une séquence qui peut mêler les codes et les mots, selon une procédure analogue à celle qu'on vient d'observer dans *Frantext* (avec

[4] On l'a parfois désigné sous l'appellation de « Maupin », du nom de ses deux créateurs : Maucourt et Papin.

une puissance moindre car le choix de la séquence est limité à cinq éléments maximum). Tout élément peut être non plus une classe grammaticale, mais aussi une forme particulière, une liste ou un lemme. Les structures lexicales peuvent être comptabilisées et donner lieu à des courbes ou analyses factorielles, ce qui ouvre un nouveau champ à la statistique linguistique, trop souvent cantonnée dans l'analyse des formes individuelles[5]. Voir figure 8.

Figure 8. Le relevé des structures lexicales

Cette version étiquetée d'HYPERBASE a été appliquée à divers corpus littéraires, généralement des monographies d'écrivain[6]. On découvre que le dosage des parties du discours n'est pas constant suivant les époques, les auteurs et les genres et une voie se dégage qui donne enfin à la quantification un accès à cette composante essentielle de

[5] Le traitement des segments répétés proposé par A. Salem permet aussi d'échapper à cette limitation.

[6] Rabelais, Montaigne, Racine, Corneille, Molière, Rousseau, Sand, Maupassant, Baudelaire, Rimbaud, Verne, Proust, Saint-John Perse, Mammeri, Chraïbi, Gracq, Le Clézio.

l'écriture qu'est la syntaxe, jusqu'ici ignorée de la lexicométrie. On observe généralement (c'est le cas dans la figure 10 ci-dessous) un champ clos où deux camps s'affrontent : le clan du verbe et de ses associés (subordonnants, relatifs, pronoms et adverbes) s'oppose à la classe nominale qui réunit autour du substantif les adjectifs, les déterminants, les prépositions et souvent les coordinations.

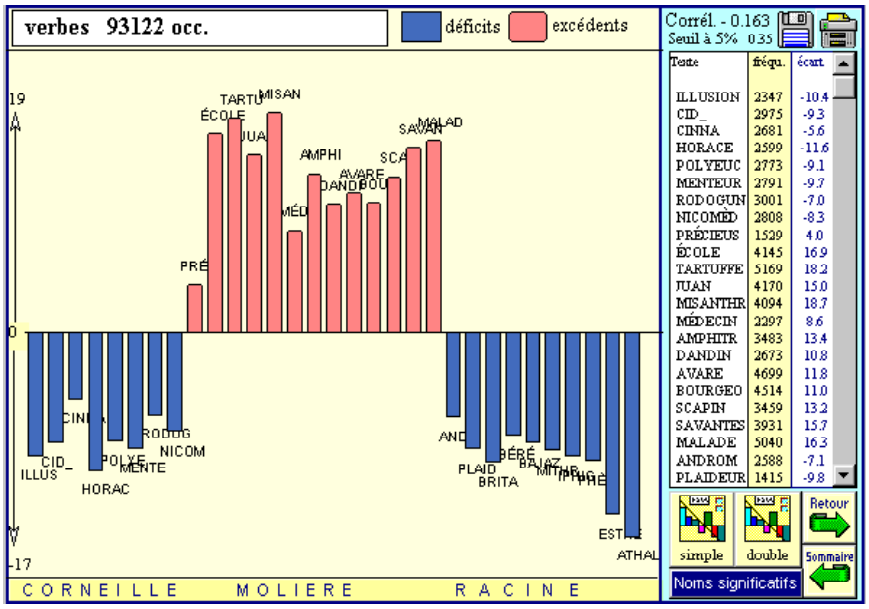


Figure 9. Les verbes dans le théâtre classique

L'exemple de la figure 9 ci-dessus est révélateur. On y voit Molière s'opposer à Racine et à Corneille relativement à l'emploi des substantifs. On observera que la différence des genres ne suffit pas à expliquer cet écart, puisque le *Menteur* et les *Plaideurs* ne rejoignent pas les comédies dans la zone des excédents. De même, en reprenant le corpus qui nous a servi d'exemple dans la figure 1 et en soumettant à l'analyse factorielle la répartition des parties du discours qu'on y observe, on voit que si le premier facteur rend compte des différences qui opposent l'évangéliste Jean aux trois autres, c'est du côté des traducteurs qu'il faut chercher l'explication du second facteur, la version de Chouraqui s'opposant à celles de la TOB et de Maredsous.

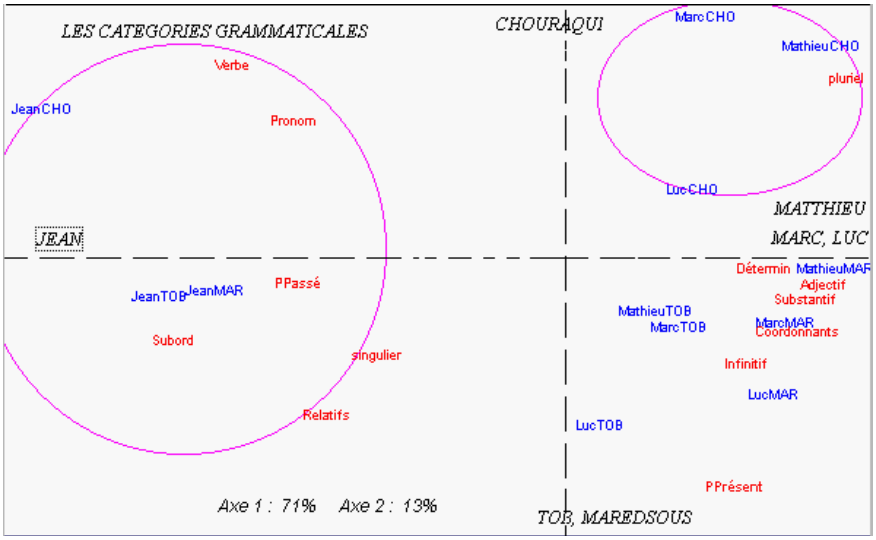


Figure 10. Les catégories grammaticales dans trois traductions évangéliques

2- Il ne faut pas cacher pourtant les faiblesses de *Winbrill* qui sont aussi paradoxalement sa force. Comme l'étiquetage est entièrement automatique, les ponts peuvent être jetés d'un corpus à l'autre, donnant quelque valeur aux comparaisons. La réduction du traitement aux formes graphiques avait au moins cet avantage d'offrir un plancher où se retrouvaient des corpus comparables, quoique issus de sources diverses. Ce que Benoît Habert appelle « l'inévitable éparpillement des étiquetages »[7] représente un danger grave de babélisation. L'entreprise d'unification menée à bien par la TEI pour le codage des textes devrait être poursuivie pour l'étiquetage. Certes des résistances et des inerties ont empêché que cette norme soit établie partout, et l'émergence de standards nouveaux complique sans cesse le problème de la standardisation[8]. Mais en matière de codage grammatical on se heurte à la spécificité des langues, chacune ayant sa grammaire et ses codes. Reste, au moins pour le français, à fédérer les efforts qui ont été entrepris ici ou là, dans des laboratoires publics ou privés, et auxquels sont associés, à des titres divers, les noms de Charles Muller, Maurice Gross, Pierre Lafont,

[7] B. Habert, A. Nazarenko, A. Salem, *Les linguistiques de corpus*, p.23. Sur ces questions de formalisation, nous suivons cet ouvrage, en regrettant que B. Habert n'ait pu se libérer pour les traiter ici même.

[8] On peut cependant parier pour le codage XML, compromis acceptable entre la nudité de format TXT, la pauvreté du HTML et la complexité du SGML.

Laurent Romary, Max Reinert, Jean Véronis, Dominique Labbé. Ce dernier en particulier a créé un lemmatiseur dont les résultats sont remarquables, surtout lorsque les doutes de la machine ont été levés par l'homme (cela représente 2 à 3% du total). Je lui aurais volontiers cédé la parole sur ce sujet. Mais il a préféré s'abstenir pour se consacrer à la traduction, dans un langage mieux adapté aux machines modernes, des 50 000 lignes de son programme[9]. D'autres lemmatiseurs sont proposés sur le marché, dont certains sont intégrés à des logiciels d'analyse textuelle et n'ont parfois qu'un but pratique de simplification : en utilisant la troncature, certains obtiennent à moindres frais un regroupement des mots par famille et, par là, des effectifs plus élevés et plus exploitables. Nous avons pu ainsi tester les ambitions et les résultats de quelques produits connus : *Alceste*, *Sphinx*, *Tropes* et *Cordial*. La palme revient sans contexte à *Cordial* dont la compétence linguistique a été reconnue par de nombreux prix (et de nombreux clients, dont *Microsoft*) et dont rend compte ici même François Rastier, à l'issue d'une fructueuse collaboration.

3- L'idéal serait d'obtenir à la sortie du lemmatiseur un fichier dont chaque ligne rendrait compte d'un mot du texte à traiter en précisant à tout le moins la forme, le lemme et le code grammatical (sous forme numérique, hexadécimale ou symbolique). À ces informations *Cordial* 6 (dans sa version *Universités*) ajoute de nombreux renseignements relatifs au traitement des expressions, à la fonction dans la phrase, à la place hiérarchique du mot dans l'arbre syntaxique, et même à la classe sémantique à laquelle le mot se rattache. Le choix des paramètres, détaillé dans la figure 12, donne en sortie le fichier suivant :

[9] Il est envisagé d'intégrer ce lemmatiseur à notre logiciel HYPERBASE.

```
N°  §  Phrase  Forme  lemme  amb.  Typegra  CodeHexa  Codegram  Syntagme  Fonction  Num  Sens
      ==== DEBUT DE PHRASE ====
1   1   1      Je     je      36    0xE480    Pp1.sn    1    S    1
2   1   1      crois  croire  A3    101       -         Vmip1s    2    V    1
3   1   1      que    que     A3    21        0x0000    Cs        -    -    2
4   1   1      la     le      A3    15        0x6000    Da-fs-d   5|5  T    2
5   1   1      langue langue  26    0x6080    Ncfs      5|5       T    2    forme
6   1   1      est    être    A3    103       -         Vmip3s    6    V    2
7   1   1      l'     le      A4    14        0xE000    Da-ms-d   8|8  B    2
8   1   1      outil  outil   24    0xA010    Ncms      8|8       B    2
9   1   1      de     de      23    0x0000    Sp        10|8      B    2
10  1          communication  communication  26  0x6010  Ncfs  10|8  B  2
11  1   1      par    par     A2    23        0x0000    Sp        12|12  H  2
12  1   1        excellence  excellence  26  0x6020    Ncfs      12|12  H  2  perfection
13  1   1      .      .       209   -         Yps       -    -    0
===== FIN DE PHRASE =====
```

Figure 11 - Un fichier lemmatisé par *Cordial* 6

Un tel fichier pourrait servir d'entrée aux logiciels destinés à l'exploitation documentaire ou statistique des corpus textuels. On peut imaginer que le traitement ultérieur mettrait en parallèle les formes, les lemmes et les codes, comme dans l'exemple de la figure 13 où la littérature latine dépouillée au LASLA de Liège et exploitée par Sylvie Mellet apparaît en champs juxtaposés, les lemmes à gauche et les formes à droite.

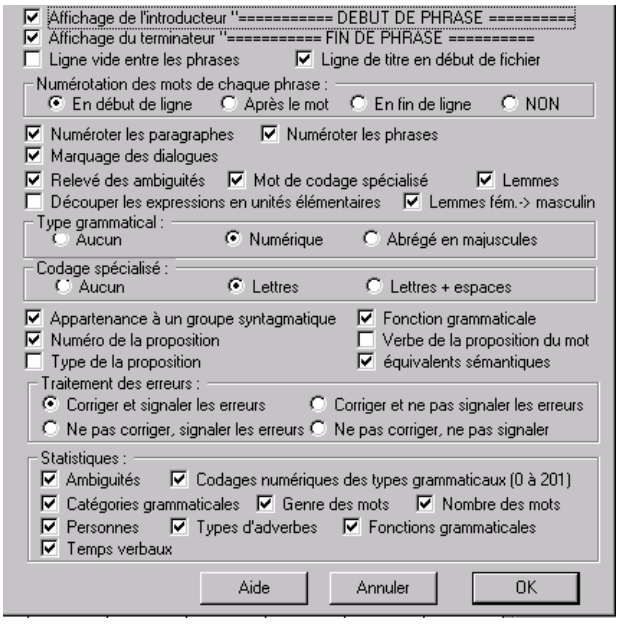


Figure 12. Les options de lemmatisation proposées par *Cordial* 6

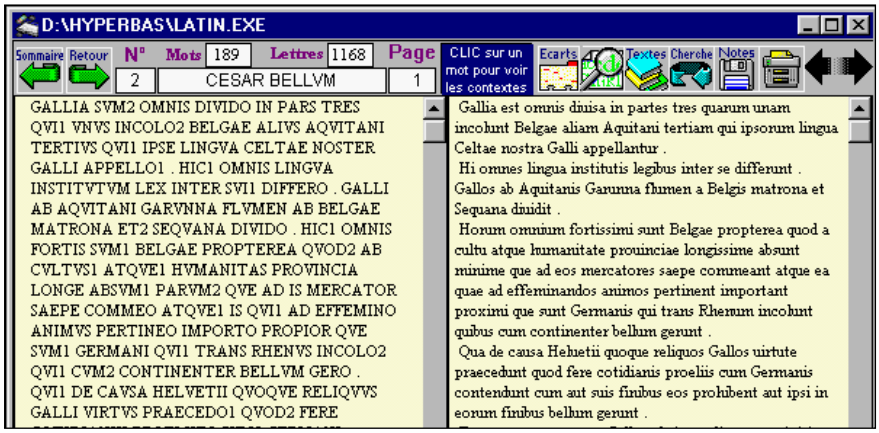


Figure 12. La littérature latine sur CD-Rom

3. Les traitements statistiques

1- *Cordial 6* qu'on vient de proposer pour modèle de formalisation pourrait aussi revendiquer la première place pour la quantification. On y trouve en effet le relevé de quelque 200 codes grammaticaux différents et de quelques classes sémantiques, ce qui donne lieu à une multitude d'indices et de pourcentages, dont on essaie de tirer des conclusions d'ordre stylistique ou thématique en s'appuyant sur une typologie des genres et des domaines constituée à partir d'un large corpus de référence (de 2 000 ouvrages). C'est là prendre quelque risque, en l'absence d'une véritable théorie des genres, en l'absence aussi d'une classification universelle des représentations. Là encore, nous proposerons l'Évangile (traduction Maredsous) à l'analyse de *Cordial*, en restituant l'un des quatre tableaux de résultats qu'il fournit, celui qui rend compte des types grammaticaux. On notera sur la marge gauche la légende qui précise l'échelle des observations sur la carte des genres.

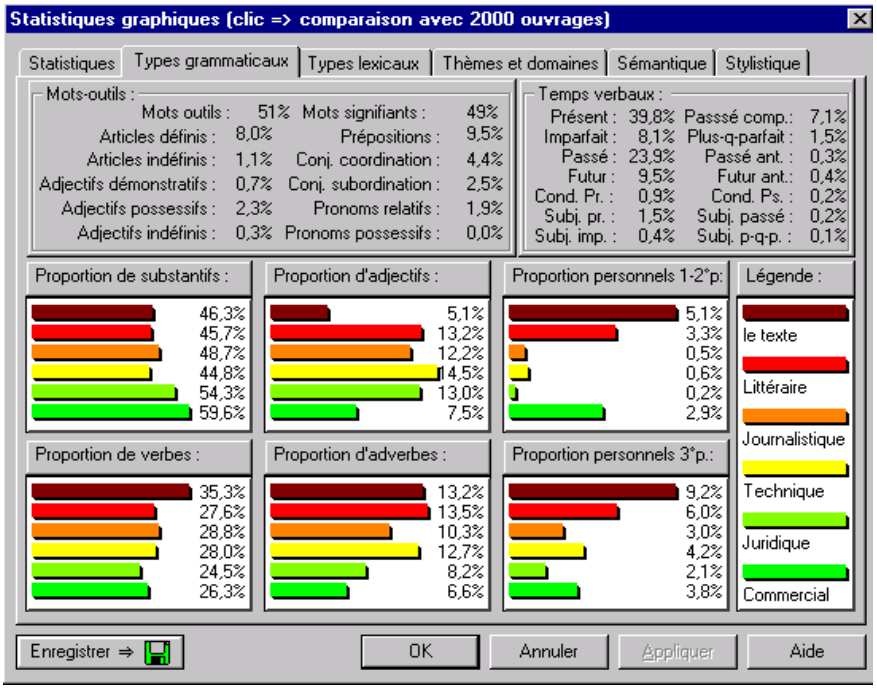


Figure 13. Les types grammaticaux analysés par *Cordial* dans l'Évangile

Parmi les genres distingués, l'Évangile semble se rapprocher du genre littéraire avec une propension marquée pour le verbe et ses acolytes, adverbes et pronoms personnels, au détriment du substantif et de l'adjectif. Ce constat peut paraître satisfaisant mais l'appréciation synthétique qui résume la leçon des comptages l'est nettement moins :

> *Ce texte est accessible. Le vocabulaire est courant mais comporte quelques raretés. La complexité sémantique est plutôt élevée. Les expressions figées sont peu nombreuses. On relève une proportion de noms propres importante. Les phrases ont une longueur habituelle. Le nombre de phrases par paragraphe est très réduit. Si ce texte ne comporte pas de listes ou de titres et de sous-titres, vous devriez grouper certains paragraphes. Les phrases ont une structuration grammaticale simplifiée. Ce texte offre un niveau d'abstraction très élevé. Le langage utilisé est efficace mais peu descriptif. Votre texte comporte quelques mots grossiers ou injurieux. Si ce n'est pas voulu, mieux vaudrait les supprimer! Ce texte comporte quelques mots argotiques ou populaires. Sont-ils utiles ici ? Ce texte comporte quelques mots ou tournures familières. Avec une proportion réduite de*

mots très usuels, ce texte est plutôt difficile; le nombre élevé de mots rares enlève de la lisibilité, sauf s'il s'adresse à un public spécialisé. La très faible proportion de noms communs rares améliore encore la lisibilité. La proportion très réduite d'adjectifs de ce texte indique une volonté d'objectivité et de non description. La proportion de verbes, nettement supérieure à la moyenne, dénote un style efficace et dynamique.

Figure 14. L'appréciation stylistique de *Cordial*

On souhaite que le Saint Esprit – qui passe pour avoir inspiré le texte sacré – ne prenne pas connaissance des appréciations que la machine a portées sur sa copie.

L'appréciation relative au contenu n'est pas plus heureuse, même si les thèmes qui parcourent l'Évangile sont reconnus grossièrement. Mais les termes pour les caractériser sont mal choisis : le mot « agriculture » convient mal pour désigner les coutumes pastorales de l'époque, la « chirurgie » intervient peu dans les guérisons miraculeuses et la « cinétique » est une notion étrange pour caractériser le mouvement du récit. Il ne suffit pas non plus qu'un texte appartienne à l'Écriture sainte, pour qu'on parle à son propos de « grammaire » et de « littérature ». Sans doute l'impropriété des termes tient-elle à l'anachronisme et au décalage entre un univers antique et une terminologie moderne. Mais des bévues semblables sont à redouter pour bien des textes, même contemporains.

*Le domaine « religion » est un second domaine prédominant. Le domaine « agriculture » est le troisième domaine saillant. Le domaine « chirurgie » est un quatrième domaine remarquable. Dans la thématique de ce texte, l'individu, par opposition à l'univers et à la collectivité, occupe une place particulière. D'une façon plus précise, l'analyse des thèmes généraux de ce texte indique une prédominance des thèmes suivants : « La spiritualité », « Le langage », « La société », « Le pouvoir » et « La cinétique ». Une analyse plus fine encore de la thématique de ce texte fait apparaître comme thèmes centraux : * judaïsme (catégorie : Les religions) * mot (catégorie : La grammaire) * famille (catégorie : La famille) * parole (catégorie : Le discours) * littérature (catégorie : La littérature) * arrivée (catégorie : Le mouvement) * divination (catégorie : La sacralité) * croyance (catégorie : Les croyances). Parmi les noms propres, on relève une nette prédominance des mots « Jésus », « Jean » et « Galilée ».*

Figure 15. L'appréciation thématique de *Cordial*

Cet échec relatif de la quantification dans *Cordial* est en réalité imputable à la communauté scientifique qui n'a pas su créer une théorie stylistique appuyée sur des dénombrements[10], non plus qu'un modèle sémantique réellement exploitable. L'imprudence des uns est lié à la timidité des autres. On note pourtant des avancées dans certains domaines de la discipline. Par exemple dans la mesure du temps. Sur ce point je laisse le champ libre à A. Salem qui a consacré sa thèse au temps lexical. En mesurant texte après texte le renouvellement du vocabulaire, les ajouts, les abandons, les retours, on peut saisir le mouvement d'une pensée, l'évolution d'une œuvre ou d'une société. Avec les précautions requises, on peut aborder ainsi les questions de datation ou d'attribution. Aucun indice quantitatif n'est en soi une preuve, mais une présomption, au moins quand les indices convergent.

2- Il en est un toutefois qui acquiert une force particulière en cela qu'il envisage tous les mots d'un texte pour aboutir à une mesure unique et globale. En réalité, cette mesure met en rapport deux textes dont on souhaite apprécier la distance. Un raisonnement simple, proposé il y a bien longtemps par Ch. Muller, présume que deux textes sont d'autant plus proches qu'ils ont plus de mots en commun. Le rapport entre la part privative ou exclusive de chacun et la zone partagée du vocabulaire donne la mesure de cette distance. Certes cette mesure pourrait être faussée quand les deux textes sont de longueur inégale, le quotient pour une même paire se rapprochant de 0 quand il s'agit du plus petit et de 1 pour le plus étendu. Mais il suffit de faire la somme ou la moyenne des deux quotients pour corriger automatiquement la distorsion. On obtient ainsi un tableau des distances des textes deux à deux. Un tel tableau se prête à une analyse globale qui projette sur une carte l'ensemble des points comme ferait une carte géographique à partir d'un relevé des distances de ville à ville. Le résultat de cette analyse factorielle est illustré dans la figure 16.

[10] Saluons pourtant les travaux de D. Biber qui contribuent grandement à éclairer les registres et les genres (*Variations across speech and writing*, 1988; *Dimensions of register variation : a cross-linguistic comparison*, 1995, *Corpus linguistics*, 1998).

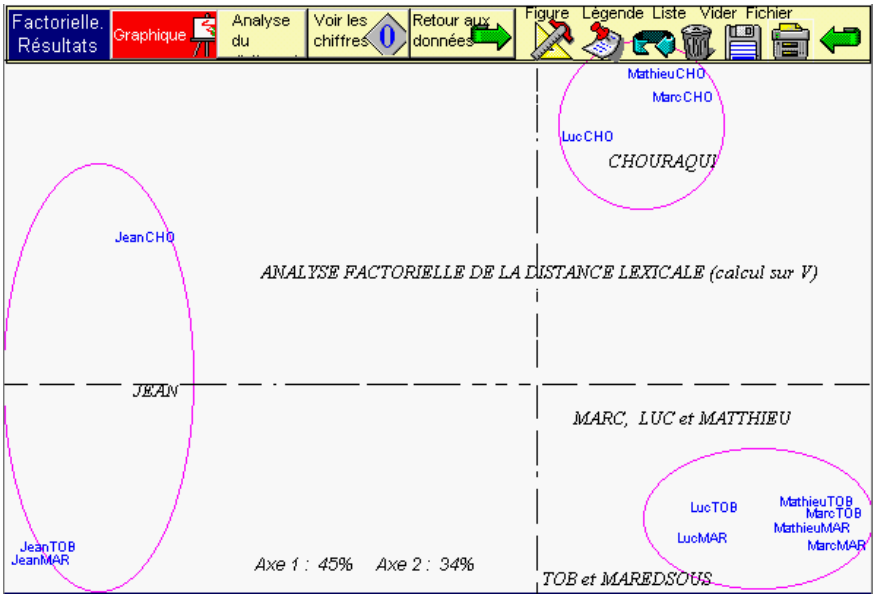


Figure 16. Analyse factorielle de la distance lexicale

La hiérarchie des variables mises en cause est ici clairement établie : l'auteur compte plus que le traducteur. Le premier facteur oppose en effet l'évangile de Jean, qui réunit sur la gauche les trois traductions de ce texte, aux autres évangiles, tous situés sur la droite. Le second facteur pourrait isoler un autre évangile, si l'influence des sources était souveraine. Or tel n'est pas le cas : ce qui distingue le haut et le bas du graphique fait référence à la traduction : celle de Chouraqui impose sa marque puissante aux textes auxquels elle s'applique (*MathieuCH*, *MarcCH*, *LucCH* et même *JeanCH*), tandis que les autres traductions sont reléguées au bas du graphique.

L'analyse arborée (méthode Luong) rend encore plus nette la représentation des forces d'attraction qui s'exercent dans le champ lexical. Ici les distances s'apprécient directement en parcourant le chemin qui mène d'un point à un autre. L'analyse (figure 17) souligne l'irrédentisme de Jean qui s'écarte violemment des autres textes, concentrés sur la branche opposée. Cependant à mi-chemin une déviation conduit, du côté de Chouraqui, tous les textes qu'il traduit et qui portent sa marque.

Le calcul des distances s'appuie ici sur la présence ou l'absence des mots, sans considérer leur fréquence. Or Dominique Labbé a proposé

récemment un nouvel algorithme qui tient compte, pour chaque mot, de la fréquence réelle et théorique dans chacun des deux textes considérés[11]. Les enseignements sont à peu près les mêmes : la branche la plus excentrique du graphique 18 isole les trois traductions de Jean, dont le message apparaît irréductible aux autres évangiles. Les trois autres évangiles apparaissent peu différenciés et laissent le champ libre à l'influence du traducteur, les groupements proposés s'ordonnant autour de Chouraqui, Maredsous et Tob respectivement. On doit observer toutefois que même en considérant tous les mots sans exception ni filtrage, les deux mesures ne leur donnent pas le même poids. La première donne l'avantage aux mots de faible fréquence, la seconde aux mots courants. La première est plus sensible aux variations thématiques, la seconde aux particularités stylistiques.

Figure 17
Calcul sur présence-absence

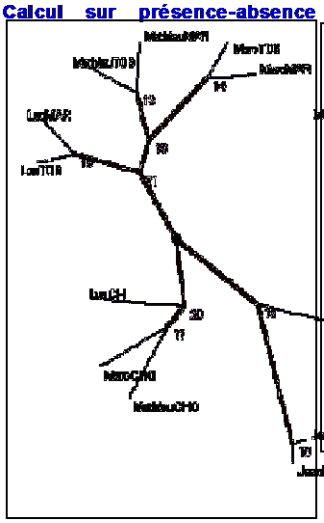
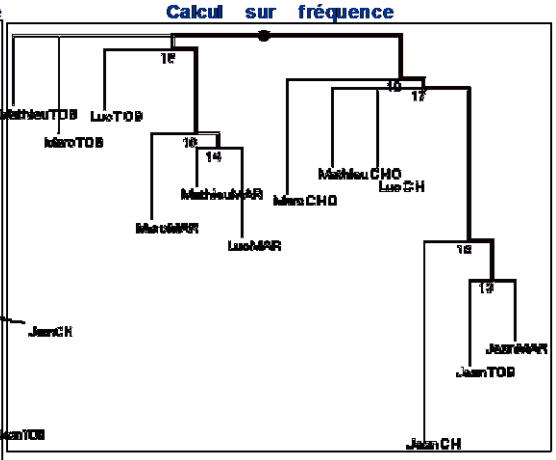


Figure 18
Calcul sur fréquence



D'autres mesures encore peuvent rendre compte de la distance intertextuelle. Bénédicte Pincemin a consacré une bonne partie de sa thèse à cette question et fera le point ici même. Comme elle connaît les dessous d'*Internet*, les moteurs et méta-moteurs mis en place sur les serveurs du *Web*, elle peut mettre en lumière leurs défauts et proposer une approche nouvelle qui ne reposerait plus seulement sur les mots-clés proposés par l'utilisateur mais sur le profil qui caractérise ce dernier à

[11] D. Labbé, D. Monière, *La connexion intertextuelle,* in *JADT 2000,* École Polytechnique de Lausanne, p.85-94.

travers les textes définis par lui comme représentant ses préoccupations. Il s'agit ici d'une vraie révolution, qui établit un calcul de distance sémantique entre deux textes, et qui n'est pas un rêve, puisque l'application en a été faite, en *Intranet*, à *Électricité de France*, au bénéfice de quelques centaines de chercheurs.

3- On peut regretter la multiplicité des formats et des codages et craindre que la statistique, s'appliquant aux uns puis aux autres, aboutisse à des résultats incohérents. En réalité, la statistique est bonne fille et s'accommode de ce qu'on lui donne. Quand on considère un corpus d'une certaine étendue pour un examen global, il importe assez peu que le texte soit lemmatisé ou non, que l'objet d'étude porte sur les formes (V) ou les occurrences (N), qu'on utilise telle ou telle méthode d'analyse multidimensionnelle, ou qu'on fasse appel à un logiciel plutôt qu'à un autre. On vient de voir la convergence des graphiques 16 et 17 qui partagent le même objet (la distance intertextuelle calculée sur V), mais non la même méthode (analyse de correspondance vs analyse arborée). Même accord des graphiques 17 et 18 qui partagent la même méthode (analyse arborée), appliquée à des objets différents (V vs N).

Plus surprenante est l'indifférence des résultats aux variations du codage. Le même texte évangélique traité successivement avec et sans lemmatisation donne la même image des distances intertextuelles. Qu'on traite 9 622 formes graphiques ou 5 014 vocables, les résultats restent stables, et les deux représentations de la figure 19 sont superposables.

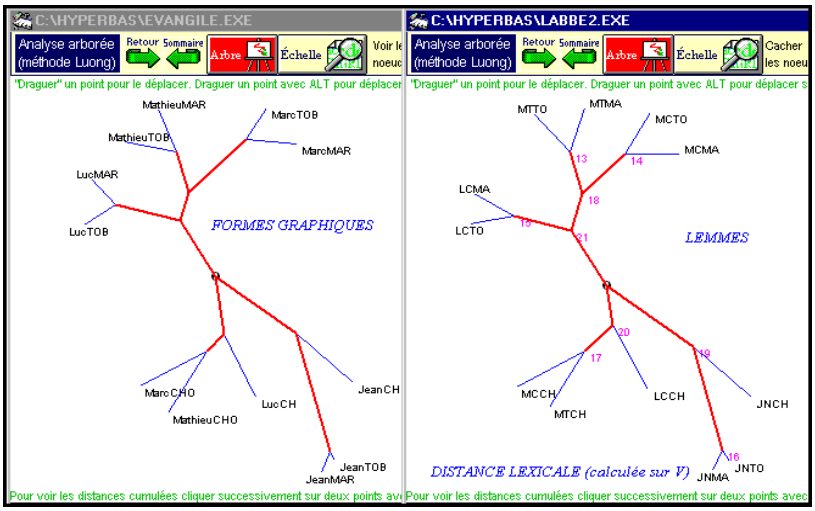


Figure 19. La convergence des codages (avec et sans lemmatisation)

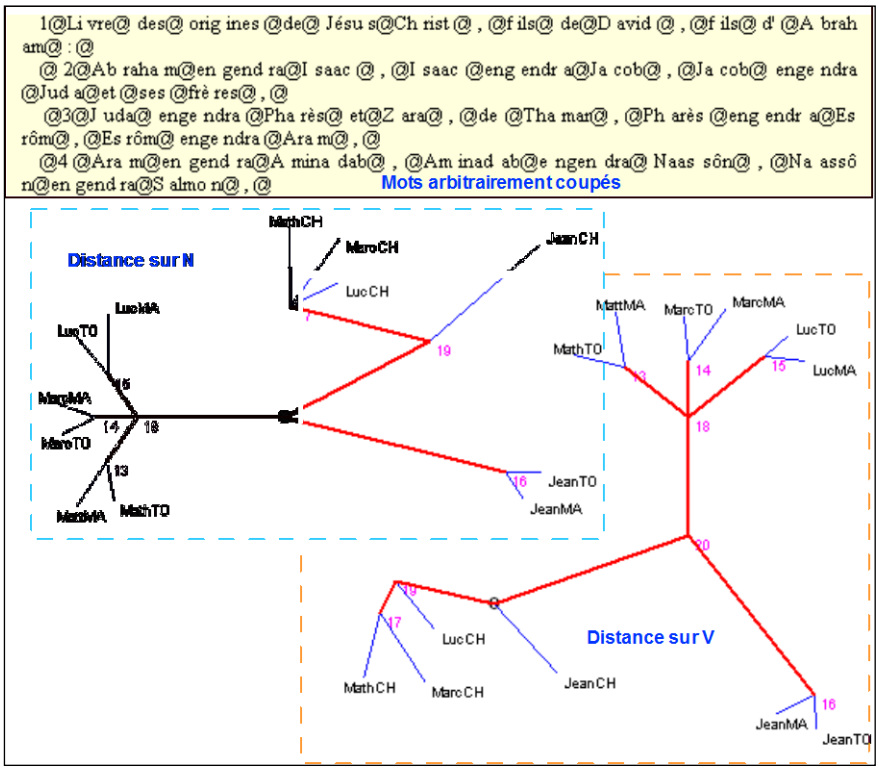


Figure 20. Découpage du texte en chaînes arbitraires

Cette stabilité ne laisse pas d'étonner lorsqu'on dénature le texte en faisant éclater la structure du mot. Remplaçons tous les blancs par un caractère arbitraire, par exemple le signe @, et découpons la chaîne en tronçons de quatre lettres, qu'on va considérer comme des « mots », même s'ils n'ont ni queue ni tête, ni forme ni sens. En cela, nous reprenons la démarche que A. Lelu applique aux N-grammes. Au lieu d'ajouter une information (ce que fait l'étiquetage), nous retranchons un élément essentiel : la segmentation en unités lexicales. Pis encore, le continuum graphique[12] est rompu et perverti par de fausses coupures, comme si on voulait crypter le texte. On peut en juger à partir des premières lignes de l'Évangile qu'on a ainsi transcrites au haut de la figure 20. Et pourtant dans cette eau boueuse où aucun mot n'est

[12] Les conditions se rapprochent de celles de l'oral et du continuum sonore. L'invention du blanc à l'écrit est d'ailleurs relativement récente et beaucoup d'inscriptions anciennes se présentent sans espace entre les mots.

reconnaissable, la décantation des évangélistes et des traducteurs se fait limpide, qu'on envisage les 400 000 occurrences (partie gauche du graphique 20) ou les 16 000 « mots » découpés (partie droite). La conclusion de cette expérience est encourageante : à l'heure d'*Internet*, où circulent tant de textes de qualité médiocre, les méthodes multidimensionnelles sont assez puissantes et robustes pour souffrir sans dommage les impuretés, et les erreurs[13].

Dernière expérience, positive aussi : la stabilité des résultats ne dépend guère des logiciels utilisés. Le même texte évangélique a été proposé à plusieurs logiciels que leurs auteurs ont bien voulu me confier. On montrera ci-dessous (figure 21) l'analyse que fournit *Sphinx* à partir des 50 mots les plus fréquents. La disposition des textes y reproduit celle du graphique 16 qui avait été obtenu avec HYPERBASE, en attribuant à Jean le premier axe, et à la traduction de Chouraqui le second. On a essayé successivement *Lexico* et *Alceste* et constaté la même convergence. Il est vrai que tous ces logiciels (on aurait pu ajouter SPADT de L. Lebart) partagent un module commun d'analyse de correspondance dont l'origine remonte à J.P. Benzécri.

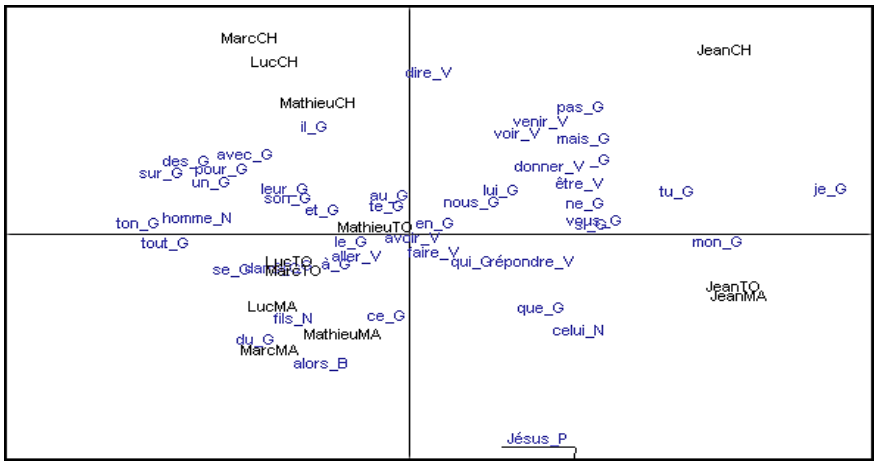


Figure 21. Analyse factorielle des 50 mots les plus fréquents (logiciel *Sphinx*)

L'algorithme utilisé par *Alceste* est cependant original, en cela qu'il ne repose pas sur une segmentation préétablie. L'analyse constitue

[13] Bien entendu ce n'est pas une invite à la paresse : on tirera toujours plus d'un texte propre et enrichi. Là où manquent les codes, l'information qu'ils véhiculent manquera toujours.

d'abord des classes, indépendamment des grandes divisions du corpus. Celles-ci ne prennent place que lorsque le cadre a été établi. On voit dans le graphique 22 que ces classes recouvrent ce que recouvraient les facteurs dans l'analyse précédente : la première est dévolue à l'évangéliste Jean (triangle à droite), la seconde au traducteur Chouraqui (triangle à gauche), tandis que Marc s'impose dans la classe 3 (triangle inférieur).

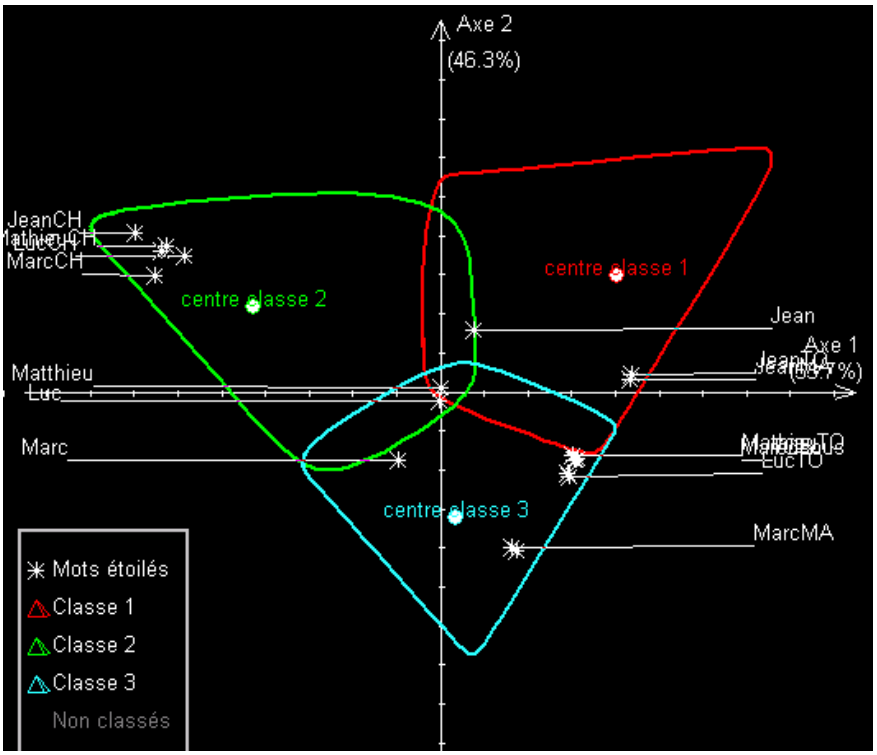


Figure 22. Les classes distinguées par Alceste

On n'insistera pas sur le fait que la liste des spécificités[14] est semblable dans tous ces logiciels, qu'il s'agisse de comparaison interne ou externe, même lorsque le corpus de référence est différent. JESUS étant omniprésent dans le texte sacré, c'est lui qui arrive en tête de liste, le corpus externe de *Cordial* confirmant celui de *Frantext*.

[14] Quand le corpus a une taille suffisante, la loi normale et la loi hypergéométrique se rejoignent.

En conclusion, il reste à expliquer pourquoi le texte évangélique nous a servi de prétexte, au risque de passer pour sacrilège. Ce n'était pas seulement pour répondre à la plaisanterie peu respectueuse de Zola[15]. Nous ne cherchions pas non plus à apporter quelque lumière nouvelle sur un texte dont chaque verset a été abondamment commenté, et particulièrement dans cette abbaye qui nous reçoit[16]. Au reste dès 1979 une analyse factorielle fondée sur le texte grec avait paru dans les *Cahiers d'analyse* qui concluait pareillement à l'irrédentisme de Jean, seul face aux trois synoptiques[17] :

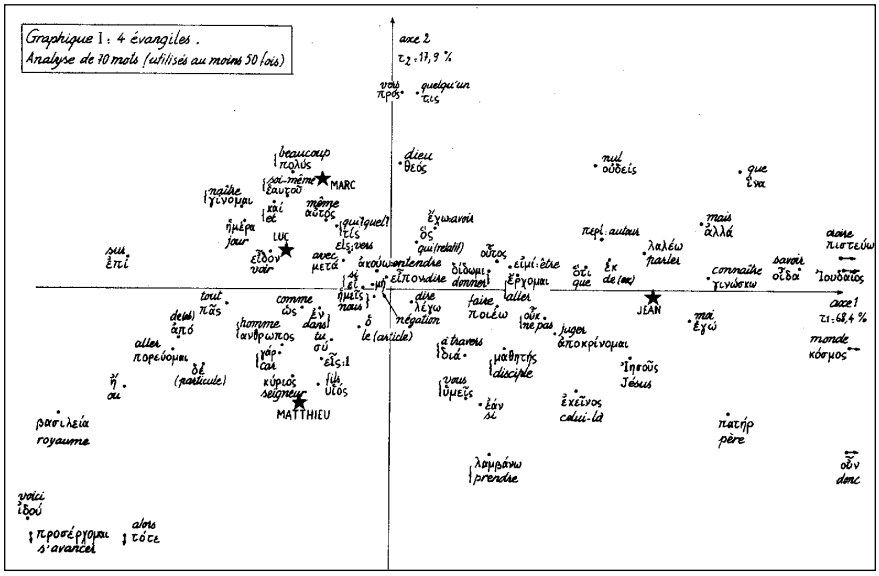


Figure 23. Analyse des mots employés au moins 50 fois dans le texte grec de l'Évangile

Notre intention était autre. Partant du fait que les textes sont difficilement comparables quand trop de variables les séparent (l'auteur,

---

[15] Il est temps d'expliquer la signification du graphique 1, laissée en suspens. L'excédent systématique des S observé dans Mathieu et le déficit symétrique constaté chez Jean sont à mettre en relation avec la catégorie du nombre. Là où Mathieu considère les groupes et les foules assemblés autour du Christ, Jean (« le disciple que Jésus aimait ») maintient une intimité personnelle et mystique avec Jésus. Le graphique 10 confirme cette distribution orientée du singulier et du pluriel.

[16] C'est une autre abbaye, celle de Maredsous, en Belgique, qui m'a communiqué le texte.

[17] B. de Solages et J.M. Vacherot, *Le vocabulaire des Évangiles*, analyse des similitudes entre chapitres de Jean.

le sujet, le genre, la taille, la date, le public, la langue), nous avons estimé que l'Évangile pouvait offrir l'occasion d'une expérience de laboratoire, où toutes ces variables seraient neutralisées. Comme la culture est ce qui reste, dit-on, quand on a tout oublié, la variation que nous voulions mesurer est ce qui reste quand on a tout enlevé, ou presque. Dans le cas présent les sources sont semblables, comme aussi la matière racontée, et le public visé. Peu de différences quant à l'étendue, l'état de langue ou – ce qui est plus dangereux – le genre littéraire. Car de toutes les forces qui s'exercent sur un texte, le genre semble la plus pesante et la plus pressante. Ne restait qu'une variable à mesurer : la double signature de l'auteur et du traducteur. Et les conditions exceptionnelles offertes par un texte familier, contrôlable et calibré, pouvaient servir à étalonner les méthodes et les instruments.