Nouvelles méthodes statistiques. L'exemple de Rabelais

Etienne Brunet

Nouvelles méthodes statistiques
L'exemple de Rabelais[1]

Etienne Brunet

Nous avons eu l'occasion d'aborder l'oeuvre de Rabelais au moment du cinquième centenaire célébré en 1994. Sur l'initiative de Marie-Luce Demonet et sous son contrôle, nous avions alors assuré la réalisation technique d'un cédérom, consacré à l'auteur de *Gargantua*[2]. Les fonctions documentaires et statistiques[3] se donnent libre cours dans cet hypertexte, dont la version Internet est toujours disponible, sur le serveur ancilla.unice.fr (page d'accueil reproduite dans la figure 1).

Si nous ravivons aujourd'hui cette recherche, c'est qu'en quelques années d'étonnants progrès ont vu le jour, non seulement dans l'appareillage informatique, mais aussi dans les approches méthodologiques. Jusqu'ici les résultats dont la lexicométrie pouvait s'honorer ne concernaient guère que le vocabulaire, entendu dans son acception la plus pauvre et la plus fruste : un ensemble de graphies rencontrées dans un texte. Malgré les recommandations et l'exemple de Charles Muller, rares étaient les études qui traitaient des données lemmatisées et qui ouvraient ainsi l'accès à la syntaxe, à la sémantique et aux faits de style. De telles données existent maintenant, et FRANTEXT, par exemple, permet des interrogations qui tiennent compte des parties du discours et de la structure de la phrase, au moins lorsque les textes sont modernes. La lemmatisation est en effet plus épineuse quand les textes sont anciens et le présent colloque en porte témoignage.

---

1. Article publié dans *Ancien et moyen français sur le Web*, édité par Pierre Kunstmann, les éditions David, Ottawa, 2003, p. 33-44, disponible sous le titre « Statistique et lemmatisation. L'exemple de Rabelais » in *Astrolabe*, Ottawa, 2003, http://www.uottawa.ca/academic/arts/astrolabe/articles/art0041/Rabelais1.htm.

2. *Les Électro-chroniques de François Rabelais*, éditions *Les Temps qui courent*, Paris, 1995. Un autre cédérom sur le même auteur, pour lequel notre concours a été sollicité, est paru en 1998, aux éditions Champion, Paris.

3. L'exposé de ces méthodes figure dans notre article « Le CD-Rom Rabelais » *Travaux du Cercle linguistique de Nice*, n°16, 1994, p. 43-79.

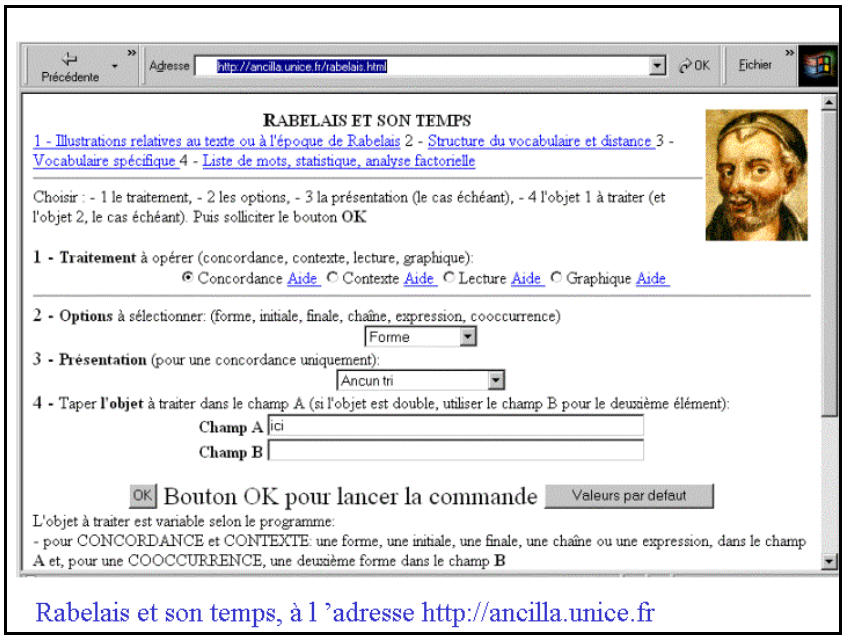


Figure 1. La base RABELAIS sur Internet

Comme je ne suis pas spécialiste de la langue du XVI^e^ siècle (et encore moins des siècles antérieurs), je ferai comme si Rabelais vivait à notre époque (il serait moins surpris que beaucoup d'autres) et je proposerai à la machine l'orthographe moderne de son texte. Quand on considère en effet l'ossature syntaxique d'une phrase, il importe assez peu que les éléments constitutifs de la structure soient ou non habillés à l'ancienne. Quand les accents sont instables, le mot *élève* peut se présenter de multiples façons selon la distribution des graves et des aigus (on a compté 9 variantes du mot chez Rousseau, parmi beaucoup d'autres combinaisons). Mais au niveau syntaxique, on n'aura le choix qu'entre le substantif et le verbe, quelle que soit la forme orthographique.

Le texte ainsi normalisé peut être soumis à un logiciel de lemmatisation prévu pour les textes modernes. Le marché des correcteurs d'orthographe a produit des outils d'analyse suffisamment élaborés pour réaliser un étiquetage à peu près correct. Sans doute un codage automatique génère automatiquement quelques erreurs. Mais la statistique est faite pour les situations de ce genre où l'entropie trouble un peu la transparence. Mieux vaut ce flou uniforme, qu'un défaut dans l'optique, qui produirait une image déformée – ce qui peut arriver quand le codage est humain et que la conscience des faits linguistiques varie

d'un individu à l'autre et parfois même d'un moment à l'autre chez le même analyste.

Après avoir essayé – sans grande satisfaction – le lemmatiseur *Winbrill*, nous avons utilisé *Cordial*[4]. *Cordial* est un correcteur d'orthographe, qui est reconnu comme le meilleur sur le marché français[5], et dont une version particulière est destinée aux professionnels des industries de la langue. Cette version – nommée *Analyseur* – décompose un texte à raison d'une ligne par mot, chaque ligne précisant la graphie du mot analysé, le lemme de rattachement, un codage grammatical aussi précis que possible, la fonction dans la phrase et même une étiquette sémantique qui classe le mot dans le catalogue des concepts. Notre logiciel HYPERBASE reprend le fichier créé par *Cordial* et distribue les données dans des champs appropriés, dévolus aux graphies, aux lemmes, aux codes et aux structures syntaxiques (figures 2 et 3).

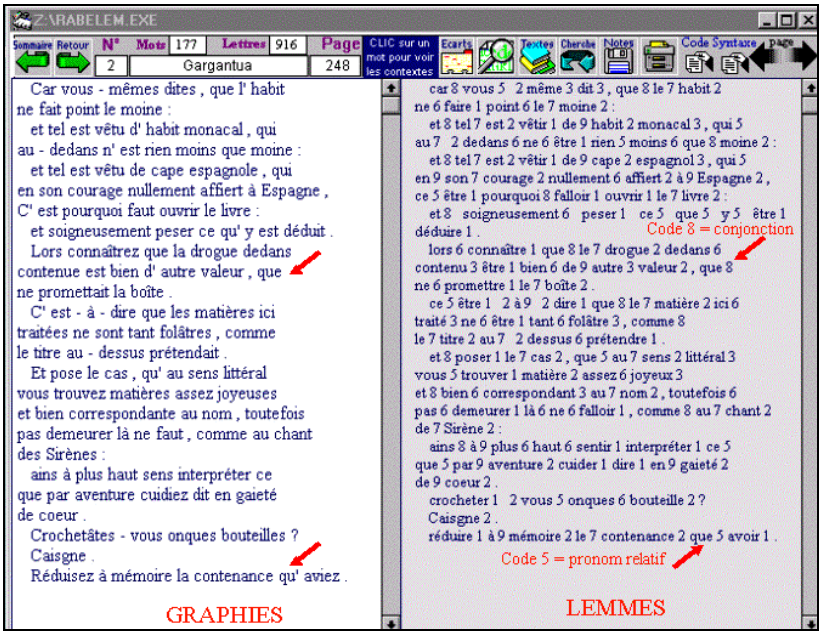


Figure 2. Graphies, Lemmes

4. Nous disposons aussi du prototype réalisé en 2002 par Dominique Labbé. L'analyse y est moins détaillée et les codes moins précis, mais comme l'étiquetage peut être semi-automatique, la correction manuelle apporte une sécurité accrue.

5. Le succès de *Cordial* est sans doute moindre au Québec, où des produits locaux comme *Correcteur 101* et *Hugo plus* lui font concurrence.

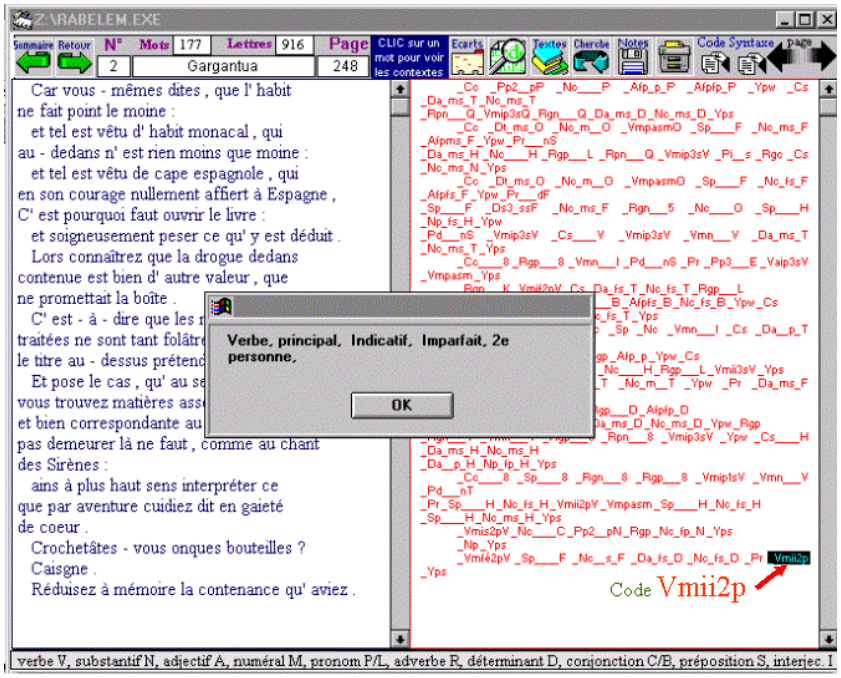


Figure 3. Codes grammaticaux

La figure 2 qui met en regard les graphies et les lemmes montre que l'analyse de *Cordial* ne recule pas devant la forme *que*, qui décourage tant de lemmatiseurs. Non seulement les formes *que* et *qu'* sont regroupées, mais aussi leurs emplois sont différenciés selon qu'il s'agit de la conjonction (code 8) ou du pronom relatif (code 5). Quant à la profondeur d'analyse elle est maximale dans le cas des verbes, où sont précisés le statut (auxiliaire ou non), le mode, le temps, la personne et le nombre. Ainsi le dernier mot de la figure 3 (*aviez*), est analysé sous la forme *Vmii2p*, c'est à dire verbe non-auxiliaire indicatif imparfait deuxième personne du pluriel.

Cordial est capable de produire l'analyse en arbre de toute phrase qu'on lui propose. Mais cette possibilité n'est pas étendue à l'ensemble du texte et le fichier d'étiquetage perd cette information, dont, au reste, la statistique tirerait peu de profit. Car les combinaisons étant si nombreuses, on trouverait peu de schémas de phrase répétés, et l'effectif pour chacun serait très faible et donc inexploitable. On a préféré réduire la profondeur d'analyse et, en prenant appui sur la ponctuation, segmenter la phrase en unités plus courtes, où chaque mot est représenté par un code

grammatical simplifié (un seul caractère symbolisant la catégorie concernée). On obtient ainsi, non pas un arbre hiérarchique, mais une séquence horizontale dans la chaîne du discours, comme dans l'exemple de la figure 4, où la structure *brvdn* représente le texte « *que ne promettait la boîte* », soit la combinaison *subordination + adverbe + verbe + déterminant + substantif.*

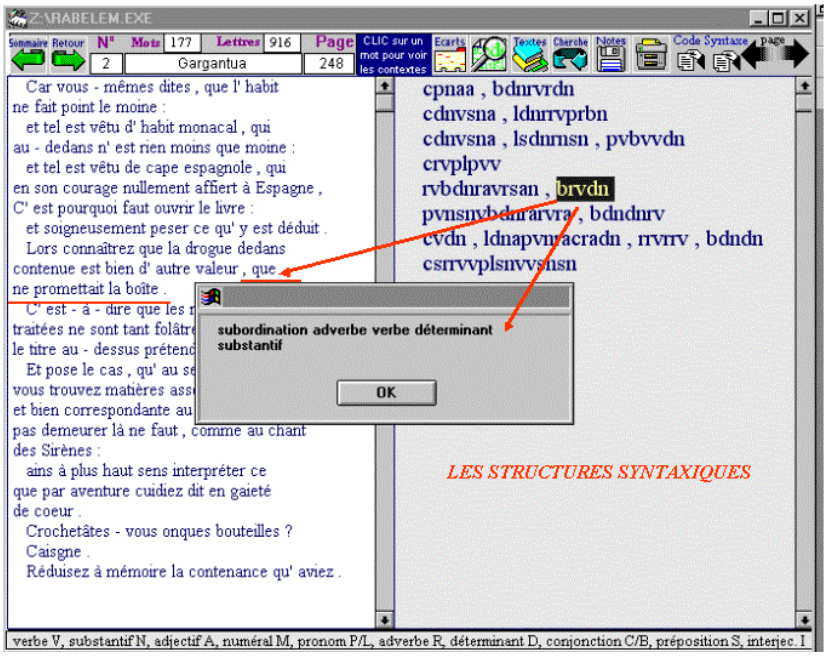


Figure 4. Les structures syntaxiques

Le logiciel HYPERBASE – dont le menu principal apparaît ci-dessous (figure 5) – met en rapport ces champs rigoureusement alignés de telle sorte qu'en cliquant sur un élément on puisse atteindre ceux qui se situent au même niveau horizontal (la graphie, le lemme, le code ou la structure alignés) ou vertical (les occurrences du même objet dans le corpus), ou même les objets d'un champ qui ont la même définition dans un autre champ. Par exemple, en reprenant l'exemple de la figure 4, un clic sur la structure *brvdn* fait défiler tous les passages du corpus qui relèvent de cette structure.

Nous nous bornerons à cette évocation des possibilités documentaires que la lemmatisation apporte à l'exploitation du corpus et qui correspondent aux fonctions disponibles sur la marge horizontale et supérieure de l'écran (*index, concordance, contexte, lecture,* etc..). Ce

qu'on voudrait montrer c'est la valeur ajoutée que la lemmatisation vaut à l'exploitation statistique (fonctions apparaissant sur la marge droite).

Figure 5. Le logiciel HYPERBASE. Menu principal

Il s'agit d'abord d'une confirmation des résultats acquis sur les simples graphies, quand on mesure la distance entre les différents textes du corpus. La distance entre les textes A et B s'obtient par le rapport entre les mots communs à A et B et ceux qu'on ne rencontre que dans l'un des deux textes. En réalité le calcul diffère, selon qu'on se contente d'une observation présence/absence (méthode Jaccard) ou qu'on tient compte de la fréquence du mot dans les deux textes (méthode Labbé). On observe la convergence des deux méthodes, mais aussi des quatre analyses qui prennent pour objet successivement les graphies, les lemmes, les structures syntaxiques et les codes grammaticaux.

Précisons qu'on a artificiellement coupé en deux chaque livre du corpus, afin de mettre à l'épreuve la sagacité de la machine. Or l'algorithme (il s'agit ici de l'analyse arborée) n'a pas été troublé par ce piège et les deux parties d'un même livre sont facilement recollées. Mieux, on voit partout les trois derniers livres se séparer des deux

premiers, soulignant ainsi le changement d'inspiration que tous les commentateurs ont observé chez Rabelais.

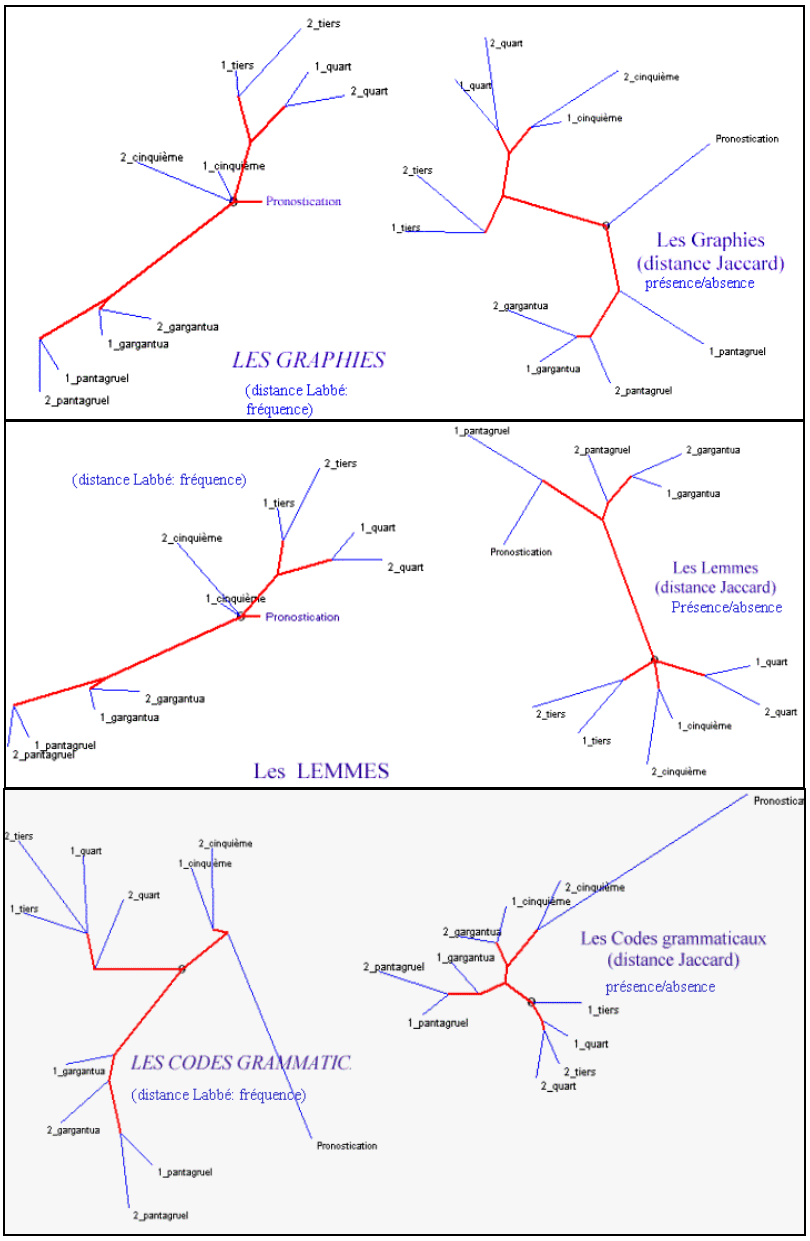


Figure 6. La distance lexicale (graphies, lemmes et codes)

Le même résultat a été obtenu avec d'autres méthodes de calcul, et notamment l'analyse factorielle. La répartition des textes sur le plan factoriel met en relief la même configuration. C'est cette méthode que nous emploierons pour examiner la carte thématique du corpus. Rappelons que les mots rencontrés dans le texte, surtout lorsqu'il s'agit de substantifs, reçoivent un code sémantique que leur attribue un thesaurus des concepts et connaissances, pris pour référence externe. Malheureusement le champ des connaissances a beaucoup évolué depuis Rabelais, et les étiquettes retenues pour désigner les disciplines ou les champs du savoir manifestent parfois un caractère anachronique quand on les applique au XVI[e] siècle (par exemple on aurait préféré *mouvement* à *cinétique*). Mais ce décalage n'empêche pas la décantation des thèmes. Ainsi voit-on le *Tiers Livre* se cantonner dans les problèmes moraux, juridiques et sentimentaux, et plus généralement les trois derniers livres (dans la partie supérieure) s'opposer à *Pantagruel* et à *Gargantua*, comme l'esprit s'oppose au corps.

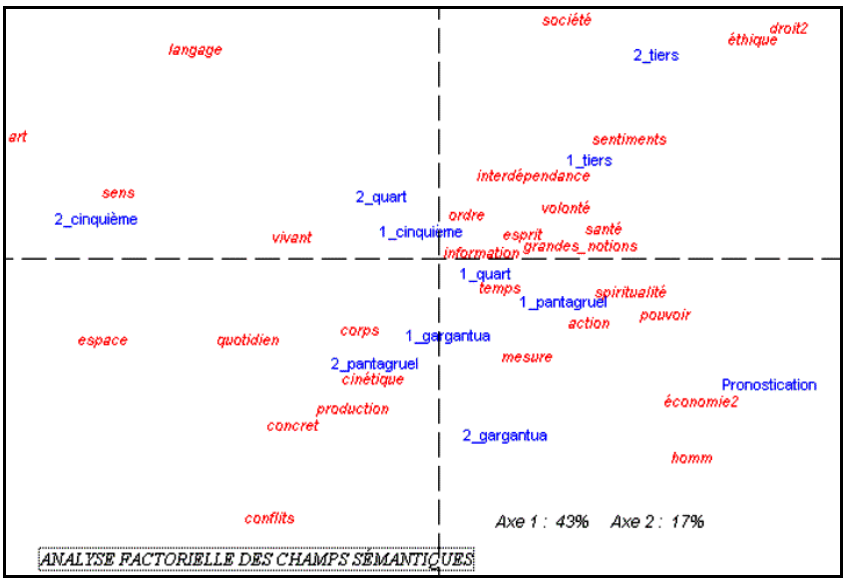


Figure 7. Analyse factorielle des champs disciplinaires

Une image semblable est obtenue dans la figure 8, quand on substitue à cette série une liste plus importante, constituée également de thèmes ou champs sémantiques. Il faut toutefois reconnaître que, pour donner le change à la censure et laisser croire que son propos est d'amuser, Rabelais maintient dans les derniers livres des épisodes

burlesques qui prolongent la veine populaire de *Gargantua*, et qu'il en résulte un peu de confusion dans les développements thématiques.

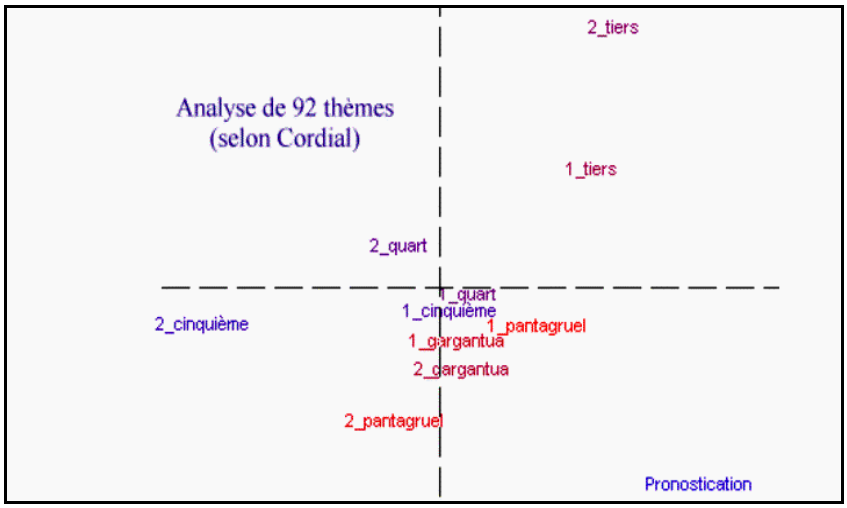


Figure 8. Autre analyse factorielle des thèmes relevés chez Rabelais

Les parties du discours offrent à l'analyse un terrain plus sûr, non seulement parce que leur repérage n'a nul besoin d'une référence externe, mais aussi parce que l'environnement syntaxique des homographes suffit à dissoudre les ambiguïtés, alors qu'il faut envisager le sens global d'un passage pour déterminer si le mot *nature* est à rattacher aux arbres et au paysage (*se promener dans la nature*), ou bien aux propriétés essentielles de tel ou tel objet (*la nature humaine*).

La statistique se repaît de tableaux, notamment ceux qui rendent compte de la totalité des données, distribuées en lignes (ici les parties du discours) et en colonnes (les 11 textes du corpus). Celui de la figure 9 est sensible aux clics de la souris quand ils s'exercent sur les marges. Ainsi en sollicitant la ligne « coordination », on obtient un histogramme qui rend compte des coordonnants dans le corpus (figure 10) et qui montre la désaffection croissante de Rabelais pour cette catégorie[6]. De façon symétrique le profil grammatical d'un texte apparaît quand on clique sur la colonne correspondante. *Pantagruel* qu'on a représenté dans la figure

6. On ignorera, dans la figure 10, la position du dernier texte, *Pronostication*, qui appartient à un autre genre et qui n'est pas à sa vraie place dans la chronologie. Le graphique 10 est en réalité double car on a projeté un élément de la série (*mais*) sur la courbe de la catégorie.

11, y est certes favorable aux coordonnants – ce que montrait déjà le graphique 10 – mais aussi aux subordonnants, aux pronoms, aux verbes et aux adverbes et se range délibérément du côté du verbe, du style oral et du ton populaire. Une galerie de portraits semblables pourrait être constituée en épuisant la liste des catégories et celle des textes.

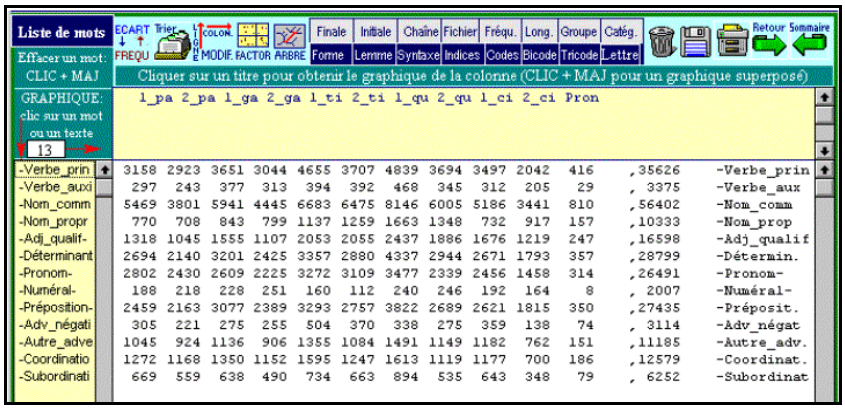


Figure 9. Relevé des parties du discours

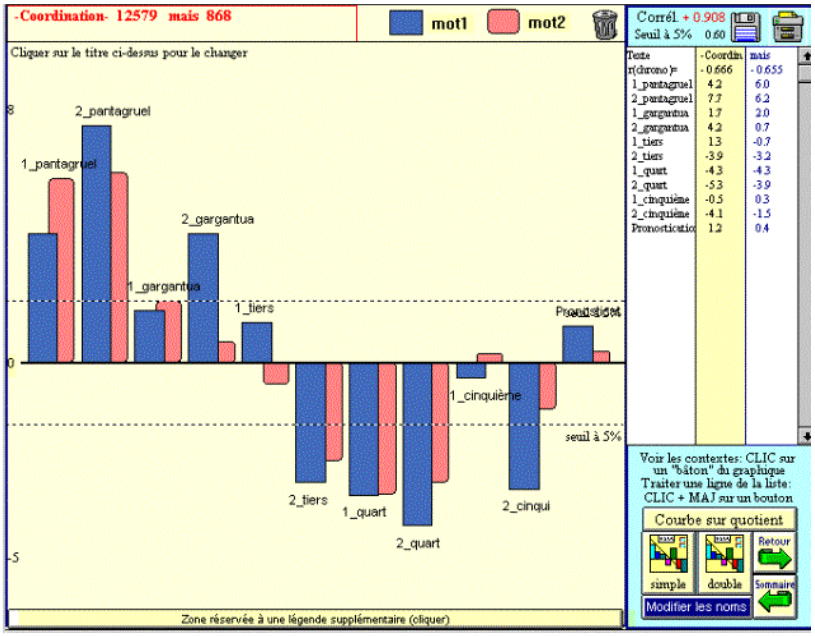


Figure 10. Les conjonctions de coordination

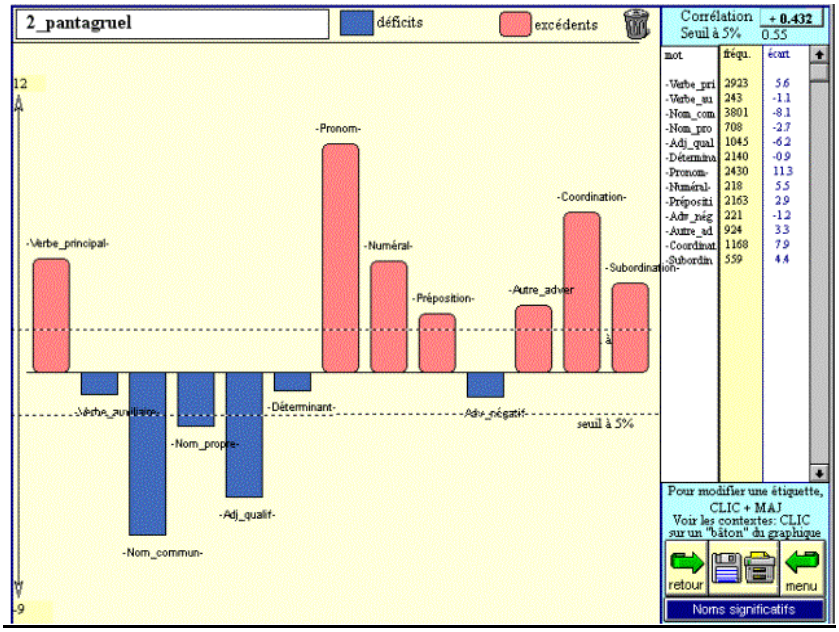


Figure 11. Répartition des parties du discours dans *Pantagruel*

Mais il y a un moyen plus rapide d'obtenir une synthèse : c'est l'analyse (on devrait dire synthèse) factorielle. Observons en effet la suite des textes dans la figure 12. Ils suivent, dans l'ordre chronologique, une ligne en forme de croissant qui part du quadrant supérieur droit, passe dans la moitié inférieure, et après avoir franchi l'axe médian vertical, remonte vers le haut. Un tel trajet est typique des données sérielles, celles qui suivent un ordre[7]. Le calcul, auquel cet ordre chronologique est caché, le découvre dans les fluctuations orientées des catégories : celles qui sont en faveur dans les premiers textes appartiennent à la classe verbale et à ses acolytes, pronoms, adverbes et conjonctions. Si ces ingrédients sont encore présents dans les nombreux dialogues du *Tiers Livre*, ils font place progressivement aux éléments de la classe nominale : substantifs et adjectifs, qui se concentrent dans la partie gauche, dans la sphère d'influence des derniers textes.

Le schéma est identique lorsqu'on envisage non plus seulement les catégories isolées, mais les combinaisons qu'elles forment dans la chaîne du discours. La figure 13 met en relief les combinaisons à trois termes, que nous appelons tricodes (par exemple *dna* représente la séquence

7. Il y a cependant une anomalie, relative à la première moitié du *Cinquième livre*. Peut-être s'explique-t-elle par l'attribution incertaine de ce dernier livre à Rabelais.

déterminant + nom + adjectif et *pvr* la séquence pronom + verbe + adverbe). Là aussi la décantation est nette entre la droite où s'installent *Pantagruel* et *Gargantua*, et la gauche où s'établissent les trois derniers livres. À droite dominent les combinaisons où entrent les verbes ou les pronoms, à gauche celles où se trouve un nom ou un adjectif.

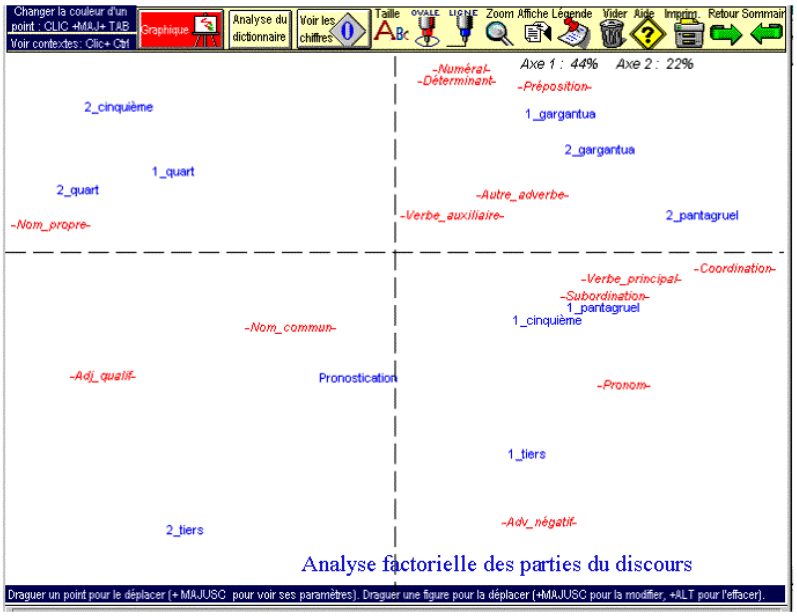


Figure 12. Analyse factorielle des parties du discours

Reste la voie royale où s'est engagée depuis l'origine la lexicométrie et qui exploite le gisement des spécificités[8]. On appelle ainsi les mots qui ont un excédent ou un déficit difficilement explicable par les lois du hasard. La figure 14 juxtapose la liste des graphies spécifiques de la première partie de *Gargantua* et celles des lemmes du même texte. Rien ne change évidemment pour les noms propres que la lemmatisation ne concerne pas. Les déplacements sont faibles également pour les noms communs, surtout lorsque l'emploi au singulier (*soif, cul, joie*) ou au pluriel (*cloches, aunes*) est dominant. Mais s'il s'agit d'un verbe et, à un moindre degré, d'un adjectif, les conclusions sont mieux assurées lorsque

8. Muller appliquait la loi normale à ce calcul – ce qui suffit généralement. Mais dans les corpus de faible dimension l'écart réduit, qui en résulte, est moins fiable que le résultat auquel conduit le calcul hypergéométrique. C'est ce dernier calcul qui a été appliqué dans le présent corpus.

les regroupements sont faits : ainsi en est-il des verbes *boire* et *torcher* dans la colonne des lemmes, à droite.

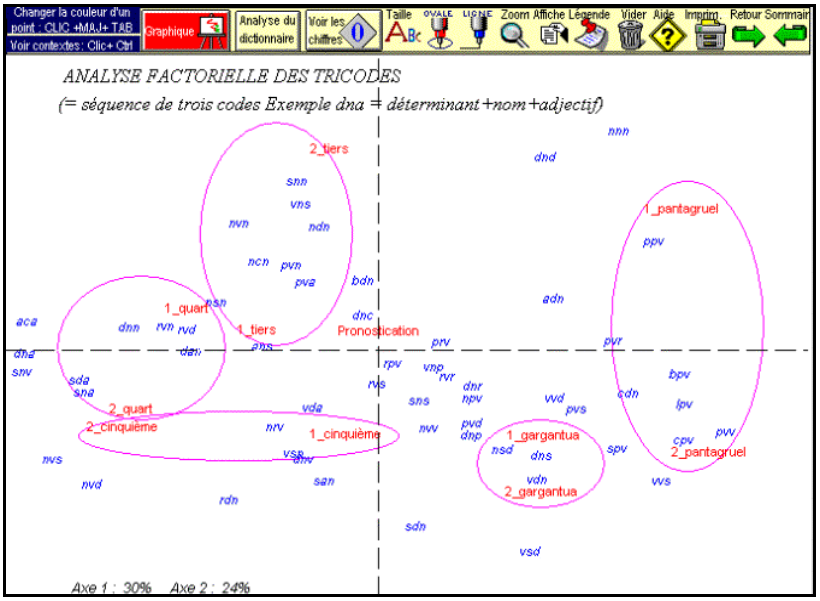


Figure 13. Analyse factorielle des structures syntaxiques à trois éléments

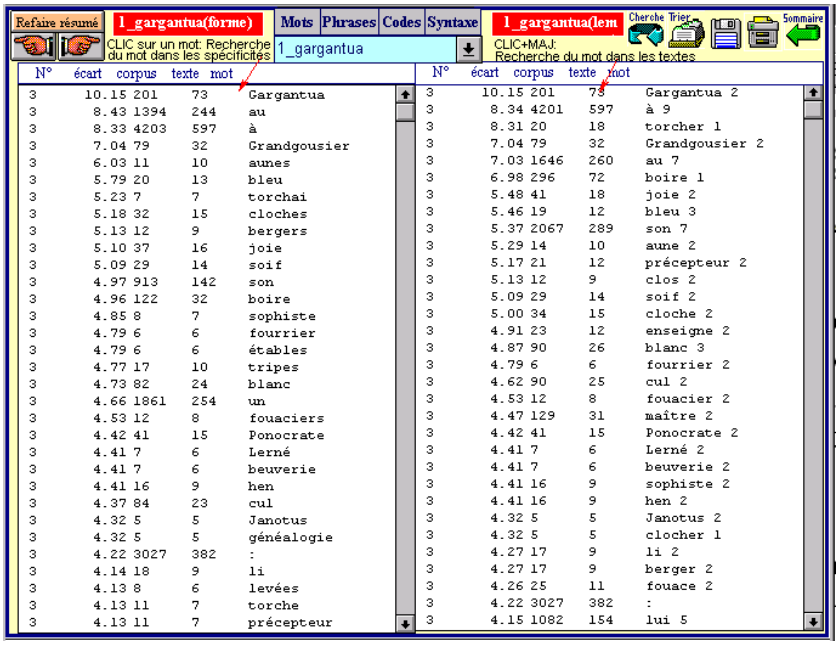


Figure 14. Les spécificités de Gargantua (graphies à gauche, lemmes à droite)

Le calcul des spécificités sert également à extraire les passages caractéristiques d'un texte du corpus. Le principe est simple : seront retenus les extraits qui contiennent le plus grand nombre de termes spécifiques. Mais diverses pondérations rendent le calcul complexe. Car l'effectif absolu ne suffit pas : encore doit-on considérer si les écarts sont faibles ou importants (la valeur de l'écart réduit entre en ligne de compte), si la phrase est courte ou longue (les phrases extrêmes sont rejetées et les autres pondérées), et s'il s'agit de noms propres ou de mots grammaticaux (catégories éliminées). On pourrait aussi diminuer l'effet des répétitions qui favorisent l'extrait classé premier dans la figure 15 : « *Je bois pour la soif advenir. Je bois éternellement. Ce m'est éternité de beuverie, et beuverie d'éternité* ». Mais ce redoublement des termes *bois* et *beuverie* est en soi une mise en relief d'une force singulière, dont le calcul a sans doute raison ici de tenir compte. Mais dans l'exemple 4, il n'est pas certain que la litanie des *Saint* ait la même valeur d'insistance. La lemmatisation n'est pas sans effet sur le classement des extraits : on observe certes que l'extrait 2 des graphies est le même que l'extrait 3 des lemmes, mais les autres diffèrent, car le calcul sur les lemmes est plus sensible aux spécificités des verbes.

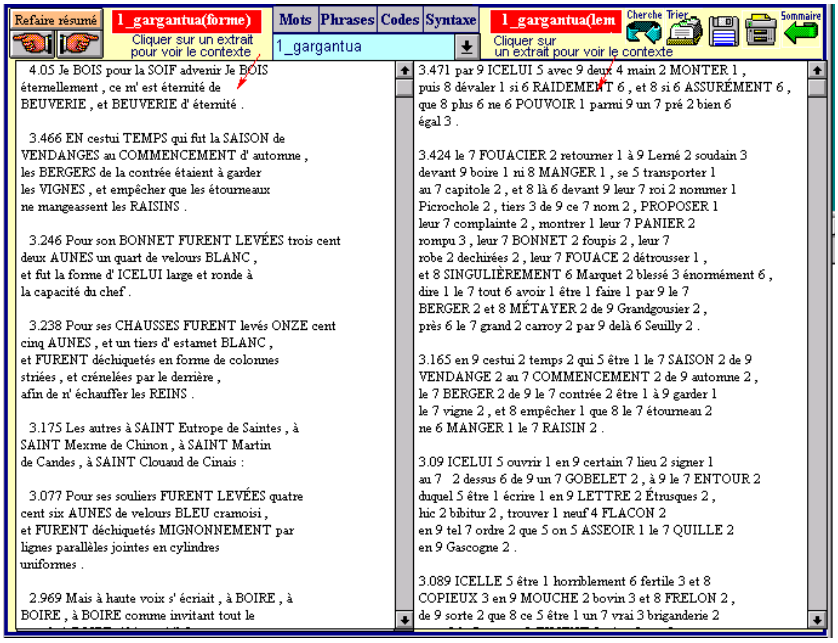
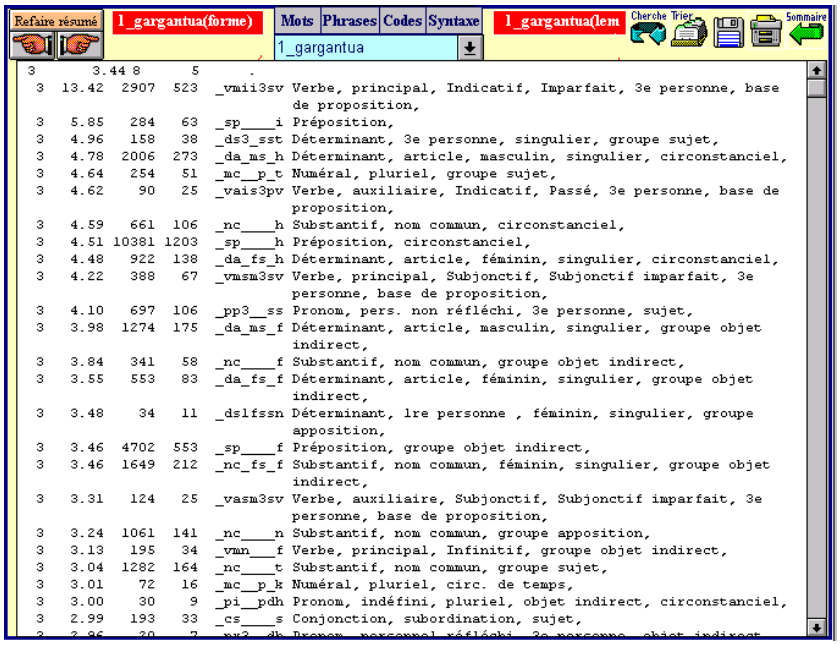


Figure 15. Les extraits spécifiques de *Gargantua* (graphies à gauche, lemmes à droite)

Le calcul des spécificités ne s'arrête pas là. Rien n'empêche de l'étendre aux codes grammaticaux et aux structures syntaxiques. Dans le même *Gargantua* le calcul s'exerçant sur les codes relève la récurrence des verbes à l'imparfait ou au passé simple à la troisième personne, ou encore celle des déterminants ou les pronoms appartenant à la troisième personne. D'autres informations moins triviales sont consignées dans la figure 16, qui aident à préciser les caractéristiques narratives du texte.



3	3.44	8	5	.	
3	13.42	2907	523	_vmii3sv	Verbe, principal, Indicatif, Imparfait, 3e personne, base de proposition,
3	5.85	284	63	_sp____i	Préposition,
3	4.96	158	38	_ds3_sst	Déterminant, 3e personne, singulier, groupe sujet,
3	4.78	2006	273	_da_ms_h	Déterminant, article, masculin, singulier, circonstanciel,
3	4.64	254	51	_mc__p_t	Numéral, pluriel, groupe sujet,
3	4.62	90	25	_vais3pv	Verbe, auxiliaire, Indicatif, Passé, 3e personne, base de proposition,
3	4.59	661	106	_nc____h	Substantif, nom commun, circonstanciel,
3	4.51	10381	1203	_sp____h	Préposition, circonstanciel,
3	4.48	922	138	_da_fs_h	Déterminant, article, féminin, singulier, circonstanciel,
3	4.22	388	67	_vmsm3sv	Verbe, principal, Subjonctif, Subjonctif imparfait, 3e personne, base de proposition,
3	4.10	697	106	_pp3__ss	Pronom, pers. non réfléchi, 3e personne, sujet,
3	3.98	1274	175	_da_ms_f	Déterminant, article, masculin, singulier, groupe objet indirect,
3	3.84	341	58	_nc____f	Substantif, nom commun, groupe objet indirect,
3	3.55	553	83	_da_fs_f	Déterminant, article, féminin, singulier, groupe objet indirect,
3	3.48	34	11	_dslfssn	Déterminant, lre personne , féminin, singulier, groupe apposition,
3	3.46	4702	553	_sp____f	Préposition, groupe objet indirect,
3	3.46	1649	212	_nc_fs_f	Substantif, nom commun, féminin, singulier, groupe objet indirect,
3	3.31	124	25	_vasm3sv	Verbe, auxiliaire, Subjonctif, Subjonctif imparfait, 3e personne, base de proposition,
3	3.24	1061	141	_nc____n	Substantif, nom commun, groupe apposition,
3	3.13	195	34	_vmn___f	Verbe, principal, Infinitif, groupe objet indirect,
3	3.04	1282	164	_nc____t	Substantif, nom commun, groupe sujet,
3	3.01	72	16	_mc__p_k	Numéral, pluriel, circ. de temps,
3	3.00	30	9	_pi__pdh	Pronom, indéfini, pluriel, objet indirect, circonstanciel,
3	2.99	193	33	_cs____s	Conjonction, subordination, sujet,
3	2.86	20	7	_pp3__dh	Pronom, personnel réfléchi, 3e personne, objet indirect

Figure 16. Les codes spécifiques de *Gargantua*

Enfin le calcul des spécificités peut être tourné vers l'extérieur : au lieu d'opposer les textes des uns aux autres à l'intérieur du corpus, il prend appui sur une référence externe, ici les textes du XVI^{e} siècle disponibles dans FRANTEXT, pour mettre en relief les mots propres à Rabelais, qu'il en ait un usage exclusif (c'est le cas de certains noms propres) ou privilégié (tableau 17). Chez aucun auteur de cette époque on ne trouve autant de *diables*, de *moines*, de *braguettes* et de *cocus*. Le tableau 17 montre aussi les silences de Rabelais, les zones du vocabulaire qu'il exploite moins que d'autres : dans la colonne de droite, réservée au vocabulaire « négatif », nombreux sont les éléments de la première personne (*je*, *j'*, *me*, *m'*, *ma*, *mes*, *mon*) qu'on trouve à foison dans des genres peu familiers à Rabelais, comme la correspondance ou la poésie

lyrique. Deux substantifs, *amour* et *coeur*, figurent dans la liste négative pour des motifs semblables. Beaucoup d'autres, de même tonalité, s'y trouvent qui échappent à notre extrait trop court.

La comparaison externe souffre toutefois de l'absence de lemmatisation dans le corpus de référence. Le dictionnaire de fréquences dont nous nous servons a en effet été constitué à partir des graphies, et même si FRANTEXT se trouve présentement étiqueté, du moins en grande partie, les outils statistiques qu'il offre aux chercheurs, contrairement aux fonctions documentaires, ignorent cette valeur ajoutée. On espère une mise à jour des dictionnaires de fréquences qui prenne en compte les lemmes et non plus seulement les graphies.

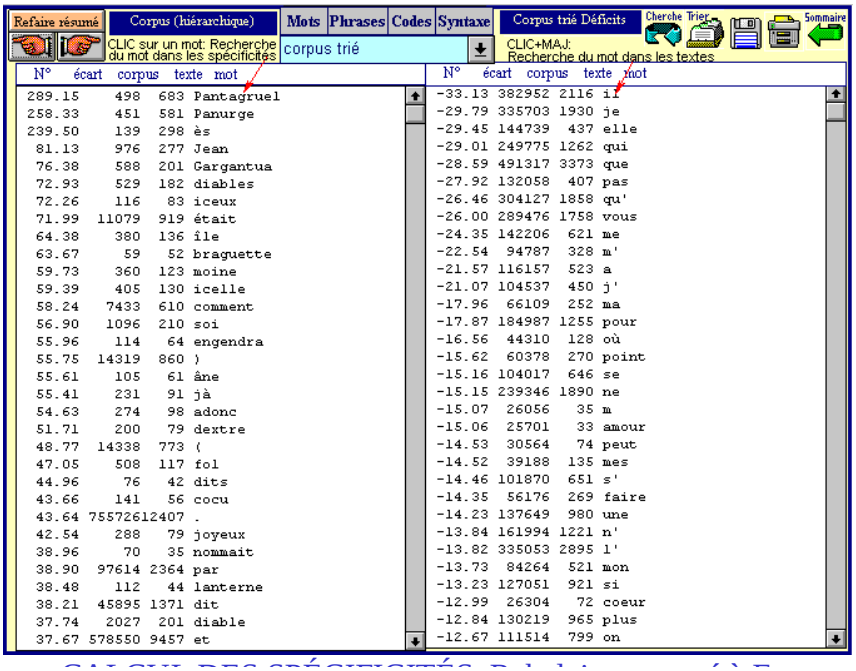


Figure 17. Le vocabulaire spécifique de Rabelais

En s'enfermant dans le corpus et en ignorant l'extérieur, il est possible d'ignorer aussi les frontières internes qui séparent les textes les uns des autres et de faire comme si le corpus était d'un seul tenant. Dans ce bloc compact et opaque, la statistique peut-elle jeter de la lumière ? La segmentation par texte étant abolie, installons une fenêtre étroite dans laquelle on fera défiler le texte en continu, en observant les constellations lexicales qu'on y rencontre. Cette démarche qui est celle du logiciel

Alceste permet de relever les associations de mots dans un environnement immédiat, qui peut être celui de la phrase ou de la page. Nous avons choisi le cadre de la page, en limitant la recherche aux adjectifs et aux substantifs dont la fréquence n'est ni trop grande, ni trop faible et qu'on appellera corrélats. L'effectif retenu comporte quelque 200 vocables, qu'on devrait retrouver en ligne et en colonne dans un tableau carré de 200 x 200 = 40000 éléments, chaque cellule du tableau indiquant le nombre de rencontres dans la même page des mots lus dans la ligne i et la colonne j. En réalité pour des raisons de dépassement de la mémoire, le tableau a été réduit, et si tous les vocables sont effectivement présents, certains, au nombre de 168, sont portés en ligne, les autres (29) en colonne. L'analyse factorielle (figure 18) rend compte d'un tel tableau, que l'on interprétera à l'aide du principe : qui s'assemble se ressemble.

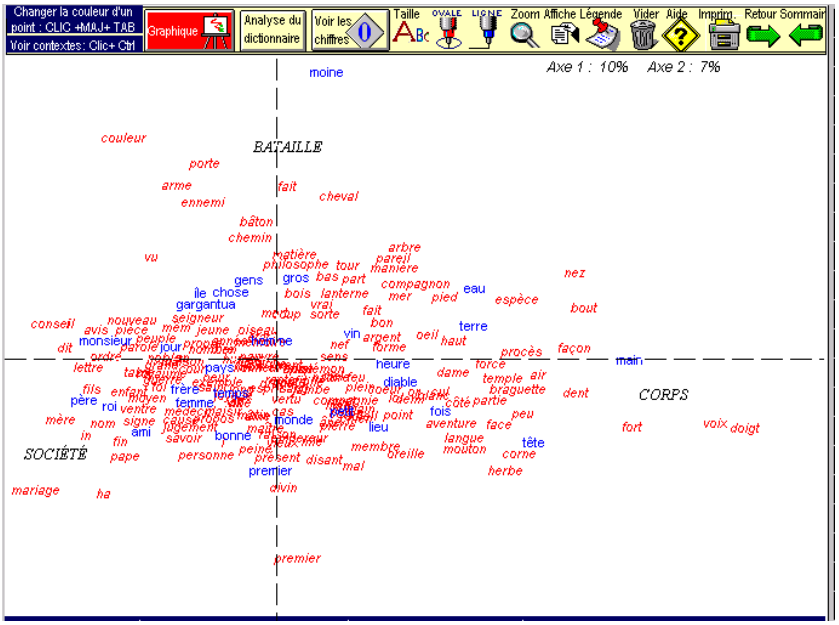


Figure 18. Analyse factorielle des corrélats

On trouve effectivement un air de famille dans la constellation située à l'est où les éléments sont unis comme les *doigts* de la *main.* Ces deux mots apparaissent sur la marge extrême et voisinent avec le *nez*, la *tête*, la *face*, les *oreilles*, la *langue*, les *cornes*, les *dents*, le *pied* et *l'oeil*. Il faudra attendre le naturalisme pour que le corps prenne autant de place dans la littérature. À l'opposé, sur la marge gauche, ce sont les acteurs sociaux qui se sont donné rendez-vous : ceux du cercle familial : *père*, *mère*,

frère, fils, enfant, femme, mariage, mais aussi ceux du pays ou de la chrétienté : *personne, médecin, maître, seigneur, roi, pape*. Enfin le quartier nord est un champ clos où l'on bataille ferme (*arme, ennemi, cheval, bâton*) et où il est facile de reconnaître Frère Jean dans la figure du *moine* qui trône au haut du graphique.

Nous arrêterons là la visite, qui sans doute apportera peu de surprises aux familiers de Rabelais. Mais le but était moins de découvrir un auteur, que d'explorer les méthodes de la lexicométrie, et particulièrement les approches nouvelles qu'autorise la lemmatisation.