Analyse textométrique du Cycle du Monde réel

Véronique Magri-Mourgues

Véronique Magri-Mourgues
Université Nice Sophia-Antipolis
UMR 7320 Bases, Corpus, Langage

Analyse textométrique du Cycle du Monde réel

Le corpus

▪ *Les Cloches de Bâle* [1934] (Paris: Gallimard, Coll. Folio 1972)

▪ *Les beaux Quartiers* [1936] (Paris: Gallimard, Coll. Folio 1972)

▪ *Les Voyageurs de l'Impériale* [1942] (Paris : Gallimard, Coll. Folio, 1972) [réimpression de la version de 1965]

▪ *Aurélien* [1944] Paris: Gallimard, Coll. Folio, 1996)

▪ *Les Communistes* [1949] Version originale (février 1939-juin 1940), Stock, 1998. Cette version ne correspond pas à la version réécrite en vue de l'édition des *Œuvres romanesques croisées*. Il s'agit là de la première édition plus authentique par opposition à la version « autorisée », si on suit l'argumentation de l'éditeur scientifique, B. Leuilliot.

Pour les besoins de l'analyse, l'ouvrage a été séparé en deux parties parce qu'il est plus long que les autres. *Communistes 1* : deux premières parties (février-septembre 1939 et septembre-novembre 1939). *Communistes 2* : trois dernières parties (novembre 1939-mars 1940, mars-mai 1940, mai-juin 1940)

L'objectif de ce travail est de cerner les divergences d'écriture éventuelles entre les œuvres réunies sous le titre général de *Cycle du Monde réel*. Par delà les points de rencontre qui justifient le rapprochement sous un même titre – ce « commun dénominateur » en lequel réside « la décision réaliste, la conscience du réel », autrement dit la conception du roman comme « machine inventée par l'homme pour l'appréhension du réel dans sa complexité » (*C'est là que tout a commencé*, Paris : Gallimard, 1972, p. 10-12) - peut-on discerner des sous-systèmes ? Des composantes qui distinguent sur le plan morphosyntaxique et lexical les œuvres peuvent-elles être décelées ?

Méthode et outil : Le repérage de traits distinctifs éventuels passe par l'utilisation de l'outil de recherche hypertextuelle, Hyperbase, qui permet une approche textométrique des œuvres. Les œuvres ont été au préalable scannées puis traitées avec ce logiciel qui intègre désormais un étiqueteur Cordial[1], qui procède à la lemmatisation du corpus textuel, autrement dit à l'étiquetage morpho-syntaxique des unités graphiques d'un texte et à leur regroupement sous un lemme ou entrée lexicale de rattachement. Les flexions verbales de même que les variations en genre

[1] Cordial a été mis au point par la Société Synapse Développement de Toulouse : Voir le site : http://www.synapse-fr.com

et en nombre sont neutralisées pour réduire les unités graphiques à leur forme canonique : l'infinitif pour les formes verbales, le singulier pour les substantifs, le masculin singulier pour l'adjectif. Le logiciel atteint une fine granularité ; il fournit en somme pour chaque forme la graphie, le lemme de rattachement, le codage grammatical autrement dit sa catégorie grammaticale, sa fonction dans la phrase, une information d'ordre sémantique qui classe la forme dans un champ lexical. Hyperbase redistribue ces données dans les champs appropriés et procède à leur dénombrement. Le texte se trouve ainsi redistribué en autant de paradigmes qui privilégient tantôt le versant lexical ou thématique du texte – si on travaille sur les lemmes par exemple – tantôt le versant grammatical ou morpho-syntaxique – si ce sont les codes grammaticaux qui sont dénombrés et comparés.

La distance lexicale fondée sur les lemmes

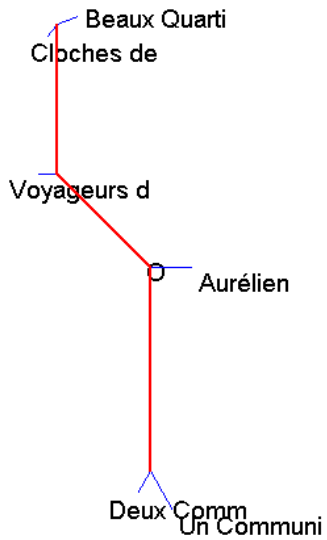

Figure 1 – la distance lexicale (les lemmes)

On a procédé d'abord à des calculs de distance intertextuelle[2]. Celle-ci permet de confronter les textes entre eux et d'obtenir une première image de leurs proximités.

Cela revient à comparer la distribution de tous les mots ; un mot commun à deux textes contribue à rapprocher les deux textes considérés ; il ne peut que les éloigner l'un de l'autre s'il n'est présent que dans l'un des deux ; plus l'intersection des textes est grande, autrement dit plus le nombre de mots en commun est important, plus la connexion des textes est forte.

[2] Voir *Corpus*, n°2, *La Distance intertextuelle*.

Les calculs ont porté sur la fréquence des unités graphiques (*méthode Labbé*[3]), afin de mettre en valeur les éventuelles particularités syntaxiques. La représentation graphique proposée est celle de l'analyse arborée développée par X. Luong. La disposition dans l'espace du graphe, l'orientation ou les directions des branches ne sont pas interprétables pour l'évaluation de la distance intertextuelle ; seule compte la distance physique entre les points du graphique, en suivant les lignes, ici les titres des œuvres, ainsi que les ramifications qui rapprochent les œuvres ou, au contraire, les éloignent.

L'étiqueteur Cordial implémenté dans Hyperbase permet des comparaisons sur diverses variables, graphies, lemmes, codes grammaticaux et structures syntaxiques. On peut voir ici le graphique qui prend en compte les lemmes, mettant ainsi en valeur les proximités lexicales entre les œuvres, la cohérence thématique : le schéma esquisse un rapprochement attendu entre les deux parties des *Communistes* qui s'éloignent cependant du groupe formé par les deux premiers écrits du corpus, *Les Cloches de Bâle* et *Les beaux Quartiers*, tandis que Les *Voyageurs* s'écarte de ces deux ensembles et qu'*Aurélien* garde une position rattachée au nœud de l'arbre, par conséquent peu interprétable. Ces rapprochements sont assez simples à expliquer par le chronotope des récits : l'action des deux premiers textes prend place dans le même contexte de la première guerre mondiale – se concluant en novembre 1912 pour *Les Cloches de Bâle* et en juillet 1913 pour *Les beaux Quartiers*.

La seconde figure qui met en évidence les proximités des configurations morpho-syntaxiques, analysant les codes grammaticaux établit d'autres rapprochements. Si les *Communistes* conservent leur proximité, les *Cloches de Bâle* s'isole en s'éloignant du premier roman tandis que des affinités s'esquissent entre ce dernier et *Les Voyageurs*, affichant un emploi similaire des codes grammaticaux et ouvrant sans doute davantage à une manière d'écrire comparable.

[3] Voir C. Labbé et D. Labbé (2003).

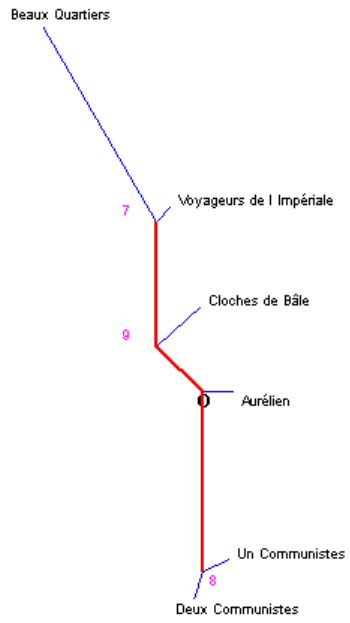

Figure 2 – la distance lexicale (les codes)

Les parties du discours

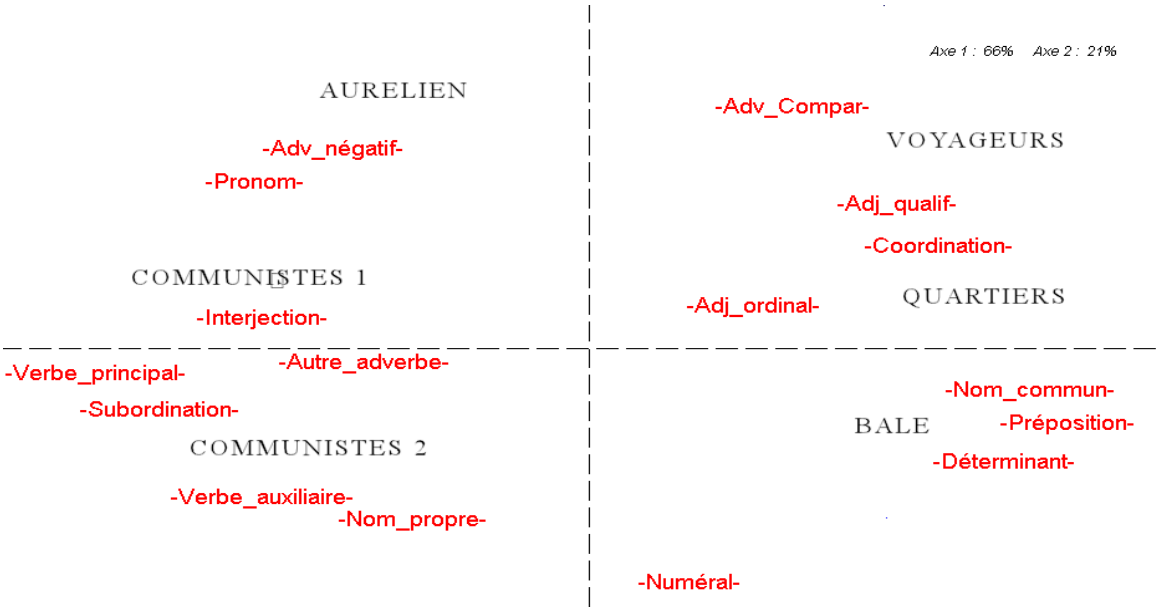

Figure 3 – les parties de discours

L'observation du vocabulaire et des parties du discours présentées sous la forme d'une analyse factorielle fournit une approche plus précise de leur distribution dans les œuvres (l'analyse factorielle permet la présentation simultanée des titres des œuvres et des catégories

grammaticales). On peut ainsi observer quelles sont les catégories grammaticales qui sont éventuellement privilégiées par chacune d'entre elles. Le codage utilisé est celui simplifié de Cordial qui propose dix-huit catégories grammaticales incluant les signes de ponctuation (ces derniers ont été cependant exclus pour cette présentation).

Les œuvres se répartissent dans les quadrants gauche et droit, L*es Communistes* et *Aurélien* d'une part et trois autres œuvres, *Bâle*, les *Quartiers*, les *Voyageurs*, d'autre part. Parallèlement, les parties de discours se distribuent selon leur appartenance à la classe du verbe et à celle du nom, de part et d'autre de l'axe vertical. La seule exception est celle du nom propre qui paraît se rapprocher de la seconde partie des *Communistes*. La classe du verbe (en emploi auxiliaire ou principal) attire à elle deux classes qui lui sont attachées, l'adverbe (adverbe négatif ou autre) et le pronom, toutes catégories confondues. L'adverbe comparatif cependant rejoint le quadrant droit et la classe nominale. A la classe du verbe se joignent les interjections et la subordination, révélant une complexité plus grande des structures phrastiques.

La proximité de tous ces éléments avec les textes sur le graphique peut autoriser une interprétation quant aux affinités entre les œuvres et les parties de discours. Les interjections par exemple qui se rapprochent de la première partie des *Communistes* peuvent être interprétées comme un indice de la transcription du discours oral. De même, le rapprochement entre les deux parties de ce roman et la classe du verbe signale sans doute une narration plus dynamique au sens où elle privilégie moins le niveau statique de la langue, engagé par des séquences descriptives.

Les deux histogrammes qui proposent la répartition de la classe du verbe et parallèlement de la classe nominale (nom et adjectif) selon les œuvres confirment ce que laisse pressentir l'analyse factorielle - L'échelle des ordonnées correspond aux écarts réduits affectés à la catégorie grammaticale sollicitée. Dans la zone positive, les écarts révèlent les titres pour lesquels la classe grammaticale est en excédent ; dans la zone négative, ceux qui en présentent, au contraire, un déficit - Le privilège accordé par *Les Communistes* à la classe du verbe tandis que les trois premiers romans affichent un penchant pour celle du nom.

La solidarité attendue entre nom et adjectif est respectée avec des nuances intéressantes : par exemple la place prépondérante accordée à l'adjectif sur le nom pour *Les Voyageurs*, qui peut être en relation avec un souci de précision accru, tandis que la deuxième partie des

Communistes voit l'écart négatif se creuser pour cette classe grammaticale. La raréfaction de l'adjectif peut-elle étayer cette tendance à accentuer ce qui est, au contraire, relié à l'action ?

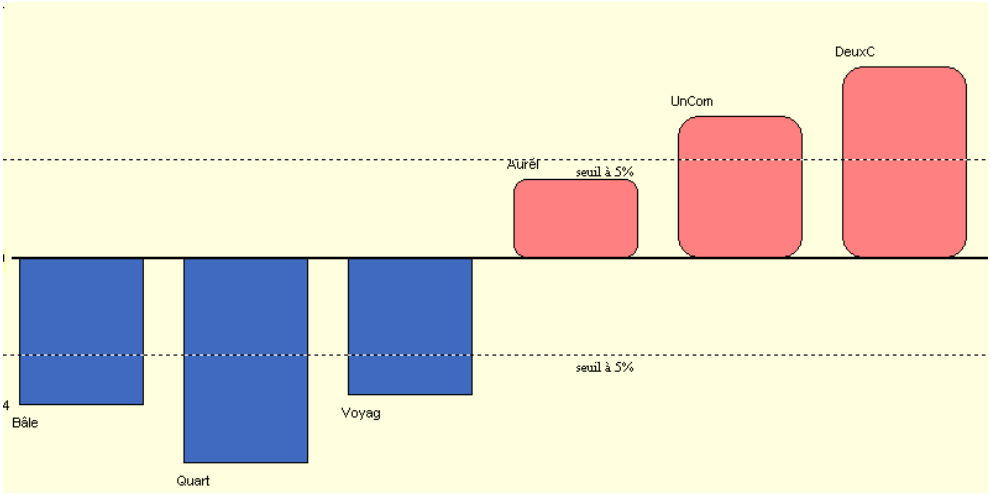

Figure 4 – la classe du verbe

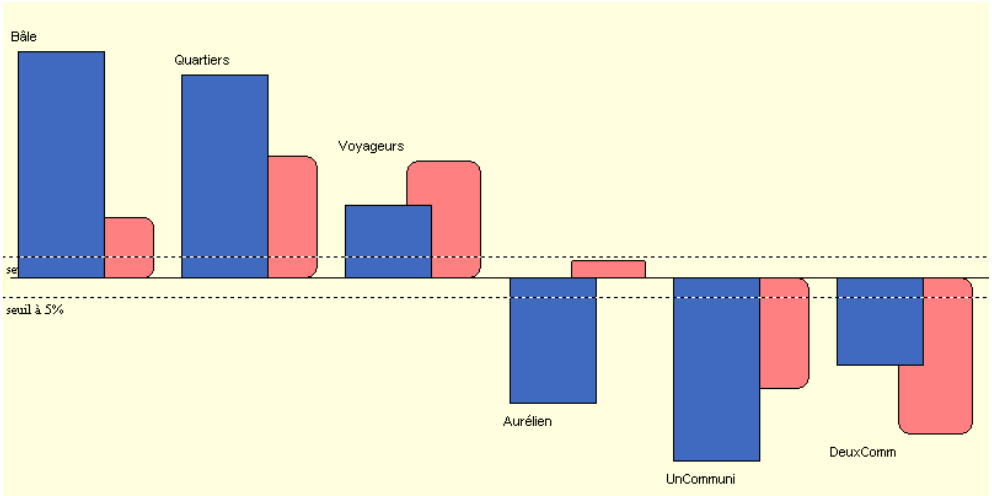

Figure 5 – la classe du nom et de l'adjectif

Le lemmatiseur permet aussi de comparer de manière globale la répartition selon le genre masculin et féminin. Toutes les catégories susceptibles de porter une telle variation sont regroupées et étudiées sous cet angle. L'histogramme qui présente le quotient entre le code masculin et le code féminin permet d'opposer assez clairement *Les Communistes* – qui présente un excédent du code masculin - aux premières œuvres du Cycle, notamment *Les Cloches de Bâle*. Cependant, ce partage qui prend en compte les déterminants comme les pronoms personnels de troisième personne croise deux distinctions, une distinction sexuée entre personnages masculins et féminins et une distinction grammaticale. L'interprétation

peut dès lors être double : une prédominance des personnages masculins pour *Les Communistes* confirmée par le privilège du pronom « ils » aussi, et peut-être une tendance à la concrétisation, si on se souvient que la plupart des noms abstraits sont du genre féminin en français. Cette observation irait ainsi à l'encontre des propos tenus par Aragon lui-même qui précisait que le titre devait s'entendre au féminin : *Les Femmes communistes*, en écho à la volonté de chanter la « femme des temps modernes » affirmée à la fin des *Cloches de Bâle*[4].

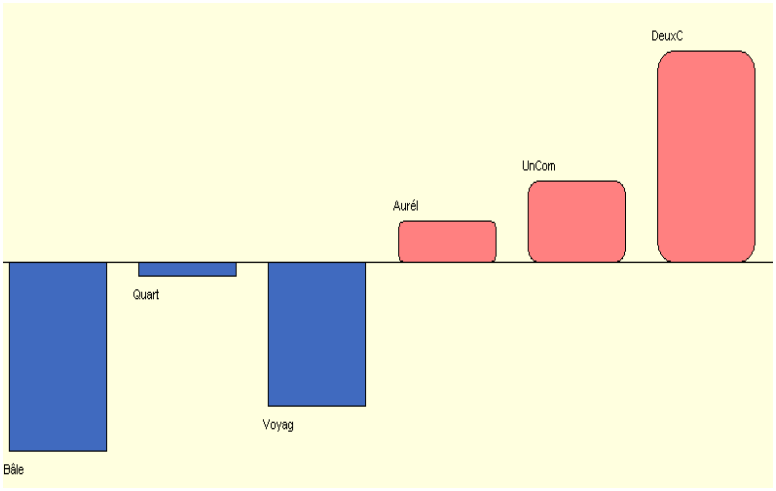

Figure 6- Quotient code masculin / code féminin

Modes, temps, personnes

[4] Voir l'introduction de B. Leuilliot, *Les Communistes*, Paris, Stock, 1998.

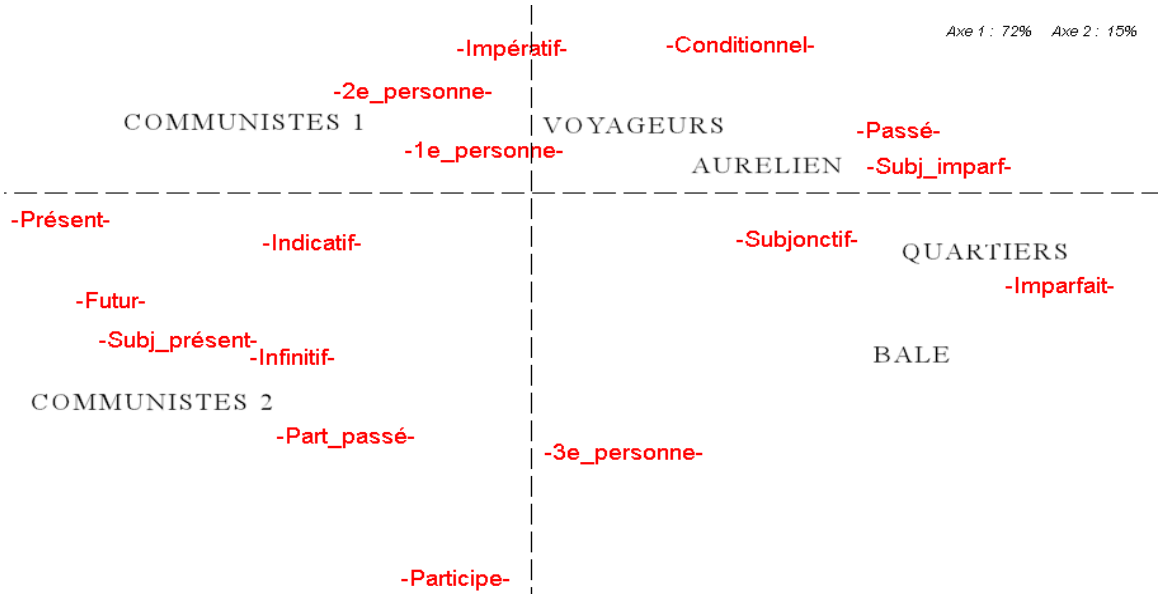

Figure 7 – Modes, temps, personnes

En centrant les observations sur la classe du verbe, on peut espérer une explication plus fine des affinités intertextuelles constatées sur le plan grammatical. La forme verbale conjuguée véhicule des informations sur le mode, le temps ou la personne, facilement identifiables par le repérage automatique (seul l'aspect lui échappe).

Là encore, c'est la présentation par une analyse factorielle qui a été choisie : l'opposition est nette de part et d'autre de l'axe vertical entre *Les Communistes* et les autres œuvres ; le couple déterminant pour *Les Communistes* paraît être l'association entre la deuxième personne du singulier et le présent de l'indicatif qui se rallie aux formes composées qu'on peut voir représentées au travers du participe passé et au futur simple : un système centré autour du présent d'énonciation se dessine comme caractéristique de ce roman, ce qui peut s'expliquer par le plus grand nombre de scènes dialoguées, tandis que les autres récits affichent un système narratif attendu pour des récits rétrospectifs.

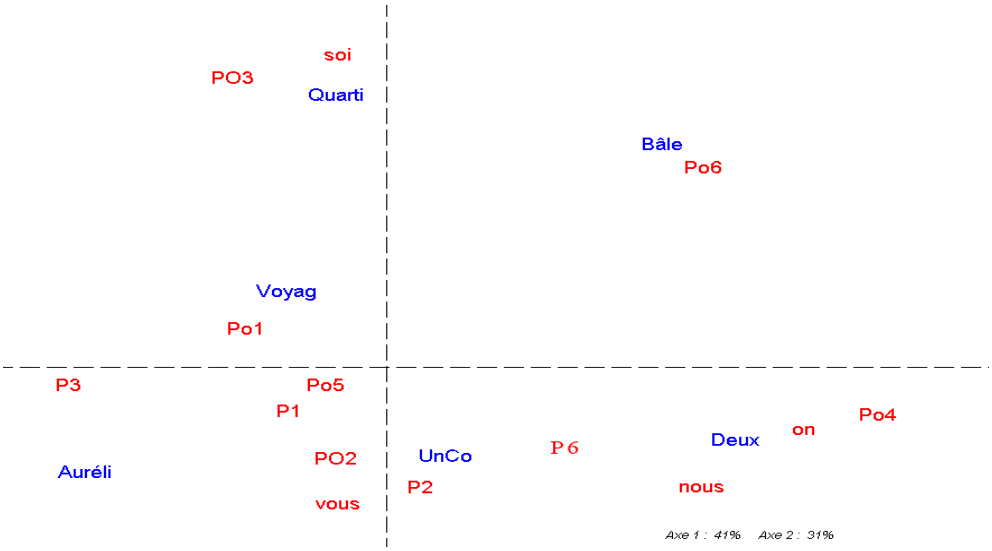

Figure 8 – les personnes (pronoms personnels, pronoms et déterminants possessifs)

L'analyse factorielle plus ciblée autour du système des personnes confirme cette interprétation ; Les marques des personnes ont été regroupées en pronoms personnels d'une part (P1, P2, P3, P4, P5, P6) et pronoms et déterminants possessifs d'autre part (Po1, Po2, Po3, Po4, Po5, Po6). Autour des deux segments des *Communistes* gravitent les marques de troisième personne du pluriel (regroupant les pronoms « ils, elles, eux ») mais aussi toutes les formes de première personne du pluriel et les pronoms personnels de deuxième rang, interprétables comme marqueurs de l'interaction verbale. Le « nous » signale une prise de parole au nom d'une collectivité relayée par le pronom « on » qui peut en être l'équivalent familier, comme dans cet exemple : « Nous, on a un papier de la préfecture, on peut te faire passer » (*Les Communistes*).

L'analyse arborée fondée sur la distribution de ces trois paramètres que sont le mode, le temps et la personne affiche des regroupements concordants : la distance intertextuelle la plus grande se laisse lire entre les deux premiers romans et *Les Communistes*.

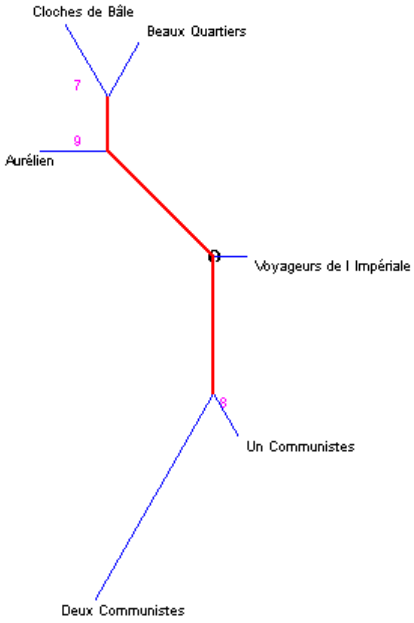

Figure 9 – la connexion des œuvres selon la distribution des modes, temps, personnes

La richesse lexicale

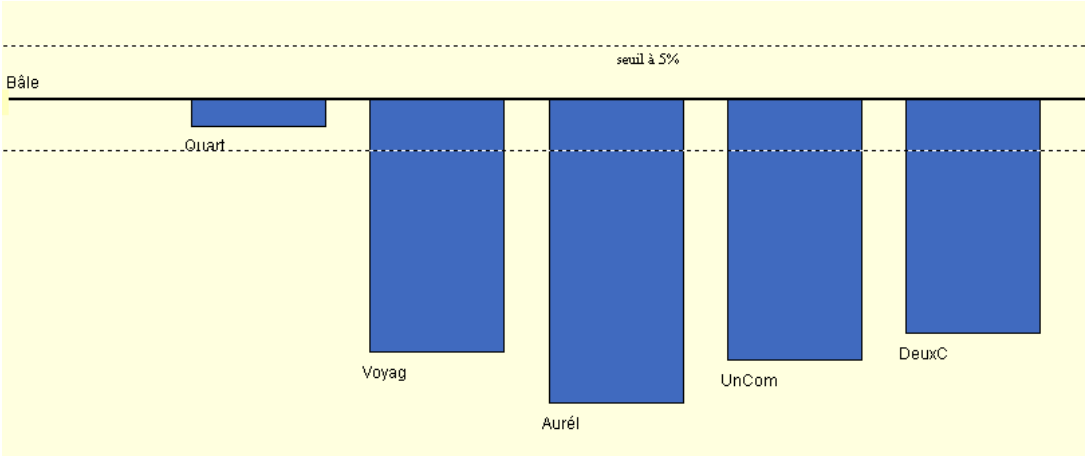

Figure 10 – la richesse lexicale

Le corpus peut enfin être observé indépendamment du contenu du discours. L'appréciation de la richesse lexicale s'appuie sur le décompte des classes de fréquence. Cette version d'Hyperbase permet d'apprécier la richesse lexicale d'un corpus soit sur les formes, soit sur les lemmes, ce qu permet de différencier la variété lexicale proprement dite et la variété grammaticale ; la prise en compte de la richesse lexicale relative d'une œuvre par rapport à l'ensemble qui sert de norme peut être un indice de l'unité du vocabulaire employé ou bien de sa variété.

Sur l'histogramme fondé sur le décompte des lemmes - ce sont les « bâtons » les plus courts qui signalent les œuvres au vocabulaire le plus varié[5] - se dessine une courbe qui accorde la plus grande variété lexicale au premier roman du Cycle et la moindre au dernier, ce qui signale un resserrement de la thématique.

La courbe des hapax

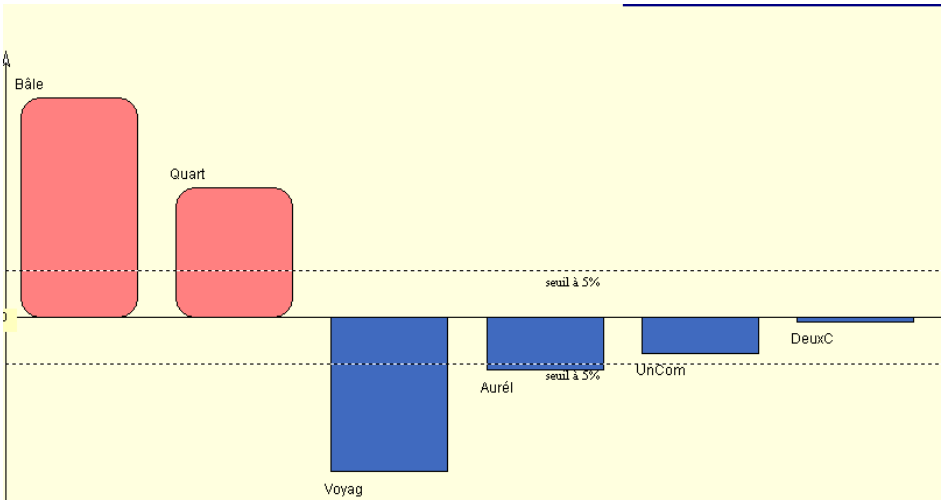

Figure 11 – les hapax

Le graphique des hapax sous la forme des lemmes, autrement dit des mots qui n'apparaissent qu'une seule fois dans le corpus et qui témoignent de l'hétérogénéité du vocabulaire, établit sensiblement la même hiérarchie entre les textes. Les deux premiers romans restent en tête tandis que *Les Voyageurs* accusent un déficit en hapax plus grand que *Les Communistes*. De même, les mots de fréquence rare se rapprochent des premiers romans. *Les Cloches de Bâle* se singularise à la fois par la plus forte représentativité des hapax et par la richesse lexicale relative la plus grande. Ces observations peuvent trouver une justification dans la construction dite baroque du roman qui repose sur la juxtaposition de personnages dans des chapitres successifs et disparates. « La multiplicité des centres d'intérêt » (*La suite dans les idées*, Gallimard, 1965, p. 9) va de pair avec une variété du vocabulaire employé.

[5] L'axe horizontal sépare la partie haute du graphique (zone positive) de la partie basse (zone négative). Les bâtons correspondent aux écarts réduits affectés à chacune des œuvres. En raison des répétitions thématiques, la richesse lexicale réelle d'une œuvre est presque toujours inférieure à ce qu'elle pourrait être dans l'absolu, c'est-à-dire à la richesse mathématiquement attendue représentée par le niveau zéro de la ligne horizontale. A cause du calcul de l'écart réduit qui repose en partie sur une différence entre la fréquence réelle et la fréquence théorique, les écarts sont tous négatifs. Les bâtons les plus courts correspondent aux œuvres dont la richesse lexicale réelle s'éloigne le moins de la richesse théorique, donc qui ont la richesse lexicale la plus grande relativement aux autres.

En observant la liste du vocabulaire en progression ou en régression au fil des œuvres, on constate l'emploi croissant de deux ensembles d'unités : les unes relèvent du registre familier voire argotique comme « bouffer, boulot, ça, engueuler, flanquer, popote, truc », d'autres dénotent l'interaction verbale tels le paradigme du verbe « dire », les interjections « eh », « hein », le mot-phrase « oui », l'expression interrogative renforcée de la tournure directe « est-ce que », des formes verbales de la personne cinq (« imaginez », « croyez ») ou encore les pronoms « tu » et « te ». Cette évolution de l'emploi du vocabulaire témoigne de l'orientation manifeste du corpus vers la structure dialogale au sens où la trame narrative insère de plus en plus de scènes dialoguées, qui s'accompagnent d'un registre de langue plus familier.

Conclusion

Pour un tel corpus, il faut garder à l'esprit les perturbations que peuvent introduire les réécritures successives des œuvres, une fois que « les yeux ont changé » (*C'est là que tout à commencé*, Paris, Gallimard, 1972, p. 39) ; la genèse devrait être prise en compte si on voulait inférer de ces observations statistiques des conclusions sur l'évolution de l'écriture d'Aragon. Ces tests statistiques ont pu simplement mettre en évidence des affinités entre les œuvres et quelques particularités d'écriture intrinsèques.

Le roman *Les Communistes* s'est clairement distingué de l'ensemble par des particularités d'écriture qui affichent le privilège accordé à la dynamique verbale associée à des indices de structure dialogale dominante. Le système qui lie la deuxième personne du singulier et du présent de l'indicatif se présente comme armature énonciative distinctive, alliée aux interjections, à la classe du verbe, à une moindre variété lexicale. Le privilège accordé encore au code masculin renvoie au titre même de l'œuvre et peut-être à un mouvement de concrétisation. Une continuité thématique unit en revanche les deux premiers romans du Cycle tandis qu'un rapprochement s'affirme entre *Les Cloches de Bâle* et *Aurélien*. A plusieurs reprises cependant, *Aurélien* garde une position quelque peu isolée de l'ensemble. Si la classe du verbe est privilégiée par *Les Communistes*, c'est celle du nom qui est mise en valeur dans les trois premiers romans, associée à l'adverbe comparatif, le seul qui se désolidarise des autres adverbes et de la classe du verbe. Quelques particularités s'esquissent encore au sein même de ce groupe : la valorisation des adjectifs dans *Les Voyageurs*, la

variété et l'hétérogénéité lexicales les plus grandes qui affectent *Les Cloches de Bâle*, de même que le privilège du code féminin dans ce premier roman, à interpréter comme marquage du personnage féminin ou comme une tendance plus grande à l'abstraction.