Distance intertextuelle et connexion lexicale : outils de catégorisation générique ou stylistique ? Approche expérimentale d'un corpus inédit : le corpus aragonien

Véronique Magri

Distance intertextuelle et connexion lexicale : outils de catégorisation générique ou stylistique ? Approche expérimentale d'un corpus inédit : le corpus aragonien

Véronique Magri-Mourgues

BCL, Université de Nice Sophia-Antipolis, CNRS ;

MSH de Nice, 98 bd E. Herriot, 06200 Nice

Résumé

L'enjeu de ce travail est de tester et de comparer différentes méthodes utilisées par les analyses statistiques des textes littéraires pour mesurer la plus ou moins grande proximité lexicale ou grammaticale de deux corpus. Seront ainsi expérimentées les fonctions implémentées dans le logiciel Hyperbase, la mesure de la « distance intertextuelle » qui repose sur les indices de Jaccard et la méthode de Labbé mais aussi l'évaluation de la « connexion lexicale » que l'on doit à Ch. Muller. L'analyse arborée (X. Luong) qui met en valeur la hiérarchie classificatoire des textes sera utilisée notamment pour présenter la distribution des hautes et des basses fréquences comme possible critère de différenciation générique. Le corpus d'expérimentation choisi est un corpus inédit qui regroupe l'œuvre poétique d'Aragon écrite entre 1917 et 1952 et des œuvres narratives. Il a le double avantage de croiser deux ensembles génériques et de s'étendre sur une tranche chronologique importante (1917 à 1967). Le paramètre générique et la variation chronologique seront ainsi éprouvés comme possibles critères de différenciation linguistique et stylistique du corpus lemmatisé.

Abstract

The purpose of this paper is to test and to compare various methods used by the statistical analyses of the literary texts to measure the more or less big lexical or grammatical nearness of two corpora. The functions included in Hyperbase, the measure of the "intertextual distance" which is based on the indications of Jaccard and the method of Labbé will so be experimented but also the evaluation of the "lexical connection" which we owe to Ch. Muller. The tree diagrams (X. Luong) which emphasize the classificatory hierarchy of texts will be used in particular to present the distribution of the high and the low frequencies as possible criterion of generic differentiation. The chosen corpus of experiment is an original corpus which groups together the poetic work of Aragon written between 1917 and 1952 and narrative works. It has the double advantage to cross two generic sets and to extend over an important chronological slice. The generic parameter and the chronological variation will so be felt as possible criteria of linguistic and stylistic differentiation of the lemmatized corpus.

Keywords: semantics, linguistics, intertextual distance, lexical connection, text classification, text corpora, literary genre

1. Introduction [1]

Le corpus d'étude a été constitué en vue de proposer un champ d'étude propice aux recherches sur la typologie textuelle et à l'exercice de l'outil statistique à ces fins, croisant différenciations

1 Mes remerciements vont à E. Brunet qui a apporté son soutien technique à la réalisation de la base de données ainsi qu'à H. Béhar qui m'a aimablement prêté ce corpus poétique pour mes recherches.

génériques et variable chronologique. Il s'étend sur une période de 50 ans et associe œuvres poétiques et narratives [2]. Il présente de surcroît l'avantage d'être un corpus inédit pour ce type d'expérimentation et de relever d'un écrivain, Aragon, qui affirmait ne voir aucune différence entre vers et prose. L'enjeu est alors de tester divers outils statistiques de la mesure de la distance intertextuelle tels qu'ils sont implémentés dans le logiciel d'analyse hypertextuelle, Hyperbase, sur un corpus a priori réfractaire au partage générique.

Pour l'exploitation statistique, les œuvres ont été classées dans l'ordre chronologique, afin de conserver la possibilité d'une observation sur l'évolution éventuelle de l'écriture d'Aragon, indépendante du partage générique, d'autant plus que textes poétiques et romanesques s'entrelacent au cours du temps. Vingt-trois fragments composent ce corpus – sachant que chaque texte noté *poésie* regroupe plusieurs recueils, conformément à l'édition originale et que *Les Communistes*, récit plus volumineux, a dû être divisé en deux segments.

2. La connexion lexicale

La nouvelle version d'Hyperbase intègre désormais trois types de calculs pour mesurer la distance entre les textes. Je choisis de commencer par le calcul de la connexion lexicale, initié par Ch. Muller mais implémenté dans les versions les plus récentes seulement d'Hyperbase.

Ch. Muller applique les calculs de la loi binomiale et table ses observations sur l'appréciation de l'écart entre un effectif attendu et un effectif réel. Grâce à ce calcul, la dernière version du logiciel Hyperbase permet en particulier une observation sur la distribution des hautes et des basses fréquences, la ligne de partage entre l'un et l'autre ensemble étant fixée à cinquante occurrences d'un même item.

2.1. Hautes fréquences

Le segment plus long de l'arbre de la figure 1 sépare nettement deux ensembles de textes qui ne correspondent pas exactement aux deux groupes génériques établis *a priori* puisque *Le Fou d'Elsa* et le *Roman inachevé* se rapprochent plus des textes narratifs que du corpus poétique. L'hésitation constitutive de ces deux recueils entre vers et prose peut sans doute expliquer ce partage et assurer la transition entre deux grands ensembles tout de même clairement constitués aux deux extrémités de l'arbre, récits d'une part, recueils poétiques d'autre part. Les « bouquets » portés par les ramifications de l'arbre paraissent dépendre d'un autre paramètre, celui de la variation chronologique. Ainsi, du côté des textes narratifs, *Les Cloches de Bâle* et *Les beaux Quartiers*, écrits respectivement en 1934 et 1936, se rejoignent en s'éloignant à la fois des autres romans du *Cycle du Monde réel*, *Les Voyageurs de l'Impériale* (1942), *Aurélien*

[2] Aragon a rassemblé autour de 1970 son œuvre poétique assortie de textes introductifs en un ensemble qui compte 15 volumes dans l'édition du Libre Club Diderot (1974). Ces textes ont été numérisés par H. Béhar pour les écrits qui vont jusqu'en 1952, exclusion faite des textes dont les amis d'Aragon étaient auteurs. Je prolonge cet ensemble avec deux autres recueils poétiques, *Le Roman inachevé* (1956) et *Le Fou d'Elsa* (1963), et je l'associe, dans un souci contrastif, à des œuvres romanesques, celles du *Cycle du Monde réel*, *Servitude et grandeur des Français, La Mise à mort et Blanche ou l'oubli* (Voir le détail du corpus en annexe). Le corpus poétique doit être accepté avec l'unité que lui a donnée Aragon et cela en dépit de l'hétérogénéité générique qui peut être perçue par tout lecteur. *Le Paysan de Paris* est par exemple envisagé plutôt comme un roman par la critique ou comme un « essai » par l'auteur lui-même. *Le Trésor des Jésuites* est une pièce de théâtre écrite à deux mains, avec A. Breton. Aragon mêle ainsi des œuvres écrites en prose, d'autres écrites en vers et mélange de surcroît vers et prose dans une même œuvre comme *Le Roman inachevé* qui compte ce qu'il est convenu d'appeler *Les trois Proses* ou encore dans *Le Fou d'Elsa*.

(1944) et *Les Communistes* (1951) et des derniers romans du corpus, *La Mise à mort* (1965) et *Blanche ou l'oubli* (1967). Du côté des textes poétiques, les bouquets coïncident avec quatre tranches chronologiques, correspondant au découpage par livres souhaité par Aragon : 1917-1929 : depuis *Feu de joie* (Livre I) jusqu'à *Snark* (Livre IV) ; 1930-1935 : Livres V et VI (*Peinture au défi* et *Hourra l'Oural*) ; 1936-1942 : depuis *La poésie soviétique à Paris* (Livre VII) jusqu' à *En français dans le texte* (Livre IXbis) ; 1943 à 1952 : Livres X et XI (*Musée Grévin* et *Chroniques du Bel Canto*).

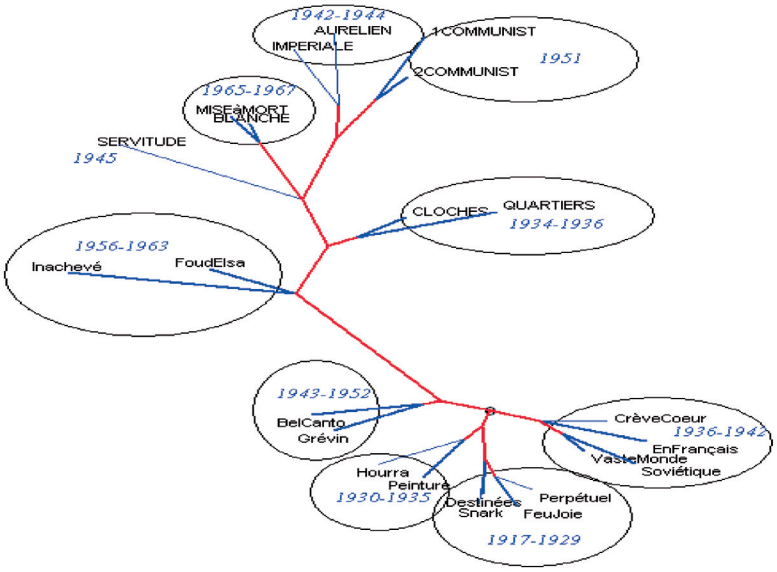

Figure 1 : Distribution des hautes fréquences

Les deux premiers groupes sont par ailleurs portés par une même branche du graphique, ce qui pourrait être expliqué par un élément biographique d'Aragon. En effet, on considère que deux époques se succèdent : la première dite période surréaliste, depuis 1917 jusqu'à la rupture officielle avec le groupe surréaliste en 1932, à laquelle succède la période dite réaliste. Béhar (in press) a montré le rôle charnière que joue le volume V dans l'architecture des recueils poétiques. Dans le graphique 1, le volume V se trouve néanmoins associé au volume suivant, proche chronologiquement.

Quant au binôme formé par *Le Roman inachevé* et *Le Fou d'Elsa*, il est difficile dans cette perspective de se prononcer en faveur de l'argument chronologique ou générique pour expliquer son isolement puisque ces deux œuvres cumulent deux originalités ; elles sont en effet particulières sur le plan formel et occupent la dernière position sur l'axe chronologique.

L'observation – grâce à l'analyse factorielle des correspondances – de la distribution des formes de plus de 1000 occurrences confirme encore le partage générique sans qu'une distribution selon les tranches chronologiques soit très claire ; un décalage est notable toutefois entre les deux derniers romans du corpus écrits le plus tardivement, *La Mise à mort et Blanche ou l'Oubli* qui paraissent présenter des affinités avec la négation (« ne, n', rien ») et la condition. Mais il est vrai aussi que les formes de plus de 1000 occurrences sont majoritairement des morphèmes grammaticaux – hormis les deux substantifs, « monde » et « vie » qui se placent du côté des recueils poétiques – dont la répartition donnera des indications sur la structure grammaticale des textes davantage que sur une caractérisation thématique. L'évolution chronologique éventuelle d'une œuvre ne peut se lire au travers de certains mots-outils que si elle est liée à une

variation par exemple de l'appareil énonciatif qui se concrétise au travers du jeu du système personnel, des tiroirs verbaux ou des marqueurs modaux. Dans le cas de ce corpus, les mots-outils ne paraissent pas *a priori* jouer un rôle discriminant quant à l'hypothèse d'une évolution de l'écriture.

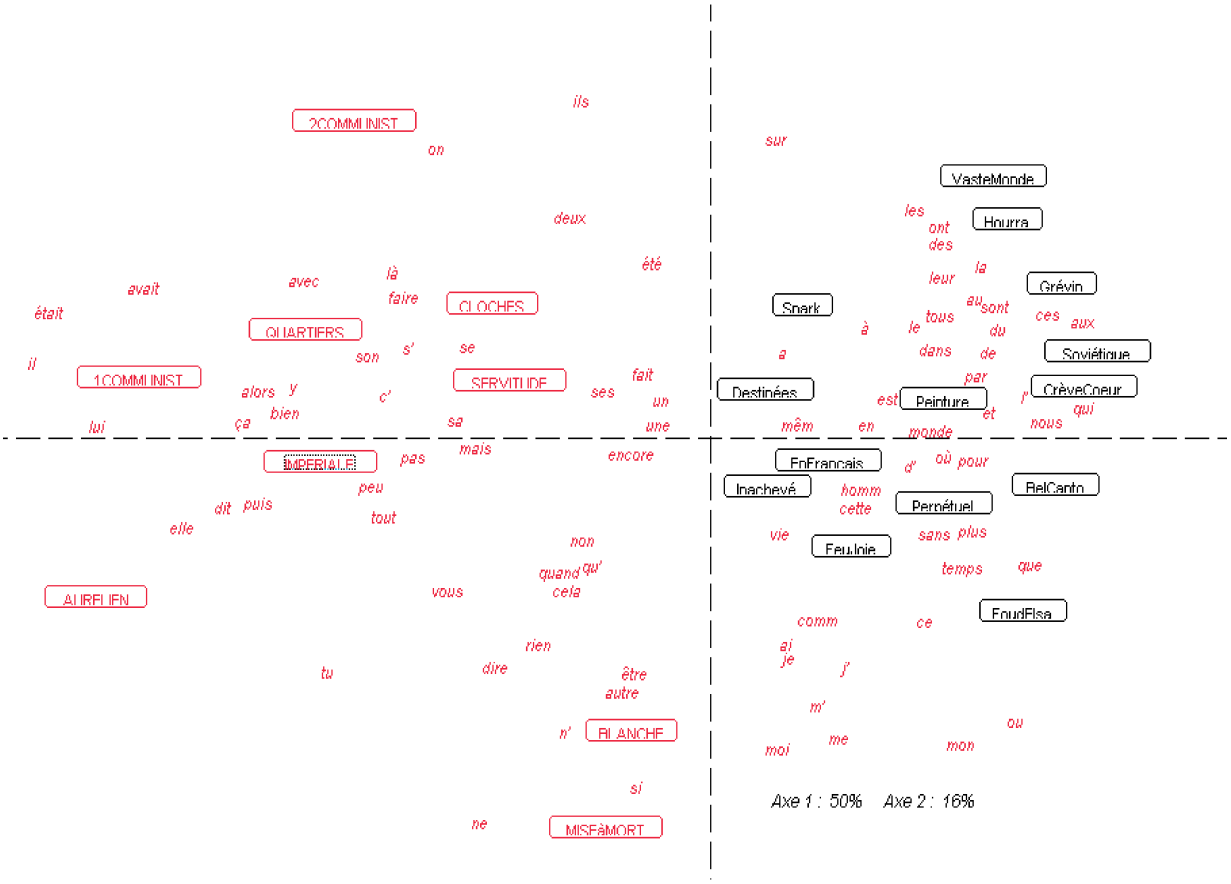

Figure 2 : Les formes de plus de 1000 occurrences

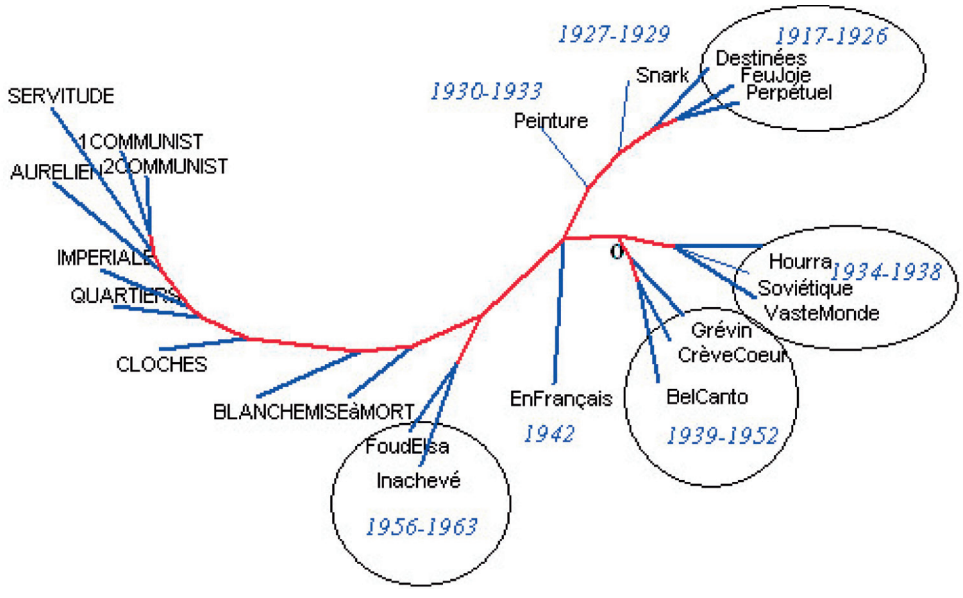

Figure 3 : Distribution des basses fréquences

2.2. Basses fréquences

Un test similaire a été mené à partir cette fois du calcul des basses fréquences – mots dont la fréquence est inférieure à 50.

Graphique 3 paraît d'emblée offrir des résultats moins tranchés. Si on retrouve globalement, de part et d'autre de l'arbre, les deux ensembles génériques avec une hésitation similaire du *Fou d'Elsa* et du *Roman inachevé* qui assurent le passage de l'un à l'autre, un échelonnement est plus manifeste sur la branche principale. Les œuvres narratives paraissent davantage s'individualiser sans regroupements en bouquets (hormis pour les deux parties des *Communistes*), tandis que *La Chasse au Snark* (Livre IV, 1927-1929), *En français dans le texte* (Livre IXbis, 1942) et *La Peinture au défi* (Livre V, 1930-1933) s'isolent en quittant les groupes chronologiques auxquels les hautes fréquences les rattachent dans le graphique précédent. Le volume V, incluant *La Peinture au défi*, paraît cette fois mieux marquer la transition entre les deux époques surréaliste et réaliste, se trouvant à mi-chemin entre deux périodes chronologiques.

3. Les calculs de la distance intertextuelle

La distance intertextuelle permet d'obtenir une image de la proximité des vingt-trois textes en comparant la distribution de tous les mots et en établissant quatre ensembles de mots [3].

Comme ce travail utilise la version d'Hyperbase qui permet d'analyser les corpus lemmatisés, il sera évidemment toujours possible d'effectuer les calculs sur les occurrences comme sur les lemmes, qui sont complémentaires. Les premiers feront ressortir les contrastes d'ordre grammatical, les seconds ceux qui sont d'ordre lexical. Plus l'intersection des textes est grande, plus la distance entre eux sera dite faible. Le logiciel Hyperbase propose deux modes de calculs de cette distance intertextuelle, l'une repose sur les indices de Jaccard, l'autre est la « méthode Labbé » [4].

Pour mettre en évidence les divergences éventuelles qui ressortissent à l'emploi du lexique, le choix s'est porté sur la distribution des lemmes – qui gomment donc les variations d'ordre grammatical – en suivant la méthode de Jaccard mettant davantage l'accent sur les aspérités du lexique. On pourrait espérer faire ressortir les particularités de chaque œuvre au travers des sphères lexicales développées. L'analyse arborée de la figure 4 établit cependant un partage générique, quoiqu'un échelonnement soit perceptible du côté des textes narratifs qui ne correspond pas toujours à l'évolution chronologique : si *Blanche ou l'oubli* et *La Mise à mort*, les deux romans les plus tardifs, sont regroupés, il n'en va pas de même du récit *Les Beaux Quartiers* qui présente plus d'affinités lexicales avec *Aurélien* et *Les Voyageurs de l'Impériale* qu'avec *Les Cloches de Bâle* dont il est cependant plus près chronologiquement. La répartition chronologique est plus claire du côté des textes poétiques. Graphique 4 est

[3] L'étendue du vocabulaire (lemmes et occurrences) du texte 1 et du texte 2 ; l'étendue du vocabulaire cumulée présent dans les deux textes ; l'étendue du vocabulaire commun aux deux textes et exclusif pour chacun d'eux ; l'étendue du vocabulaire absent des deux textes.

[4] La première ne prend en compte que la présence ou l'absence d'un mot donné, sans que sa fréquence n'intervienne. Dès lors, un biais d'analyse est introduit puisque seront privilégiés les mots rares, voire les hapax, qui donnent l'avantage aux variations lexicales et thématiques en mettant en sourdine les divergences d'emploi qui pourraient affecter les mots les plus usuels, forcément présents dans n'importe quel texte. La somme des rapports établis entre le vocabulaire exclusif du texte et le vocabulaire total du texte correspond à la distance entre les textes et oscille entre 0 et 2. La seconde observe les fréquences de chaque item et permet de porter ainsi une attention plus grande aux catégories grammaticales. La plus ou moins grande fréquence d'une catégorie grammaticale peut être significative.

quasiment superposable au Graphique 3 qui propose la distribution des basses fréquences : on retrouve des regroupements qui correspondent aux tranches chronologiques avec les mêmes particularités : par exemple *Le Crève-cœur* qui quitte l'ensemble (1936-1942) et rejoint une période plus tardive (1939-1952). Le livre V (*La Peinture au défi*) se trouve pareillement isolé, dénotant des particularités lexicales qui en font une œuvre d'exception dans l'ensemble. Seul *La Chasse au Snark* n'adopte pas la même position isolée que dans Graphique 3 puisque dans l'analyse arborée de Fig. 4, ce recueil se rapproche plus fortement de sa tranche chronologique.

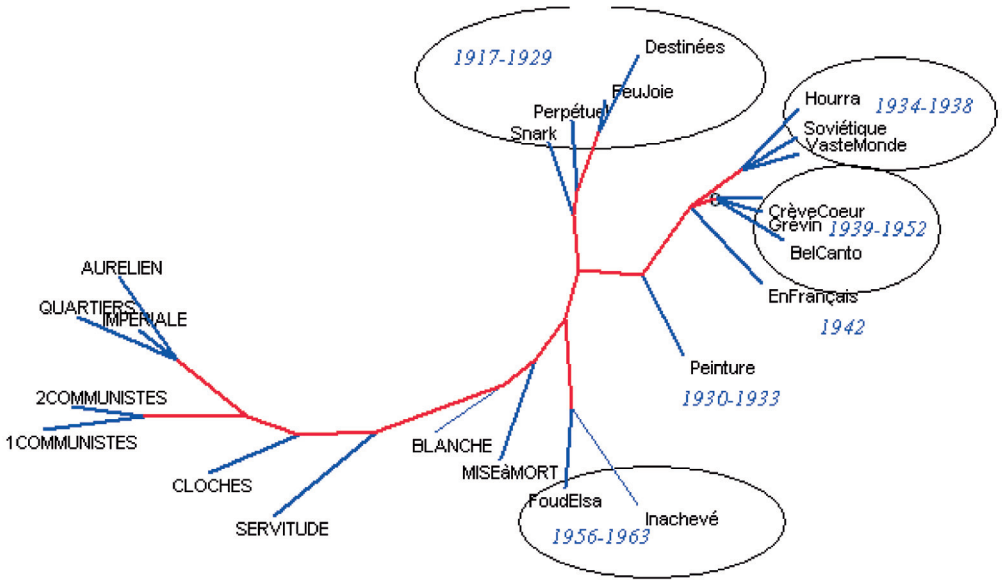

Figure 4 : Distribution des lemmes (indice de Jaccard ; critère de l'absence / présence)

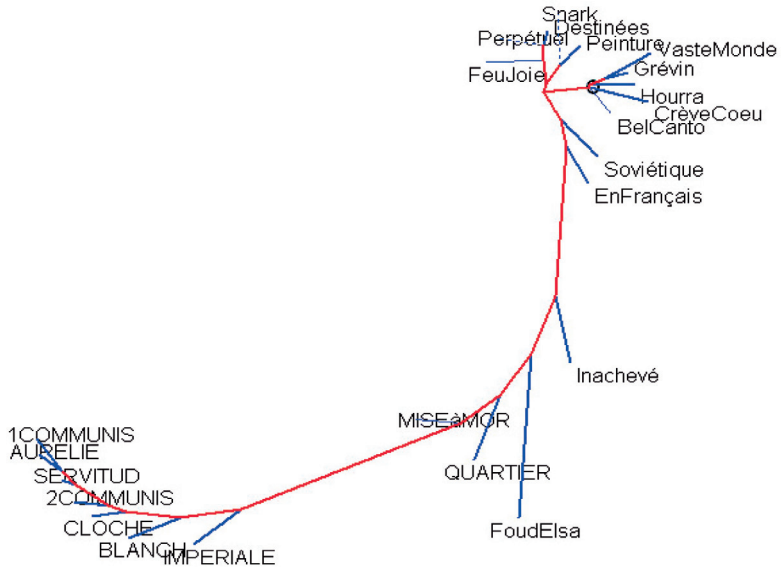

Figure 5 : Distribution des codes grammaticaux (méthode Labbé ; critère de la fréquence)

En contrepoint à l'analyse précédente, les calculs à l'origine de Fig. 5 tentent d'établir une typologie textuelle à partir cette fois des codes grammaticaux ; le calcul porte forcément dans ce cas sur la fréquence de chacun de ces codes correspondant au codage implémenté dans Hyperbase par le biais de l'étiqueteur Cordial. Les observations prennent alors comme champ d'étude les variations observables dans la fréquence d'emploi des parties de discours,

indépendamment du lexique. La bipartition générique subsiste avec toutefois quelques nuances. Le point remarquable est que le Livre V n'affirme plus son caractère hétérogène ; il rejoint au contraire le champ générique de la poésie et se rattache plus étroitement aux premières œuvres poétiques. La distribution des codes grammaticaux ne permet pas de redessiner clairement des bouquets chronologiques qui pourraient témoigner de l'évolution de l'écriture d'Aragon. Au contraire, des récits qu'une trentaine d'années séparent, *Les beaux Quartiers* et *La Mise à mort*, se trouvent rapprochés.

Afin d'articuler les observations précédentes qui ont porté sur la distribution des lemmes, des codes et sur l'ensemble des items regroupés selon leur seule fréquence, avec le contenu même du discours, une incursion dans le domaine de la sémantique, qui « nous introduit au domaine de la langue en emploi et en action » (Benveniste, 1974 : 24) a été tentée mais la place manque pour livrer ici les graphiques afférents. La conclusion rend compte succinctement des principaux résultats.

4. Conclusion

Hautes et basses fréquences se sont avérées sensibles à la différenciation générique. Aucun test statistique n'a permis au critère chronologique de prévaloir sur le critère générique, même si la constitution de ce corpus qui croise les deux paramètres pouvait encourager ce résultat. Cependant, les analyses arborées qui mettent en valeur la distribution des œuvres selon l'emploi qu'elles font des hautes et des basses fréquences ne sont pas superposables ; leurs divergences peuvent être commentées. Les hautes fréquences paraissent plus sensibles à la variation chronologique, autrement dit, peuvent être postulées comme un paramètre valide pour suivre l'évolution d'une écriture au cours du temps ; les œuvres se regroupent en sous-corpus réunis par un point commun, celui d'appartenir à la même tranche chronologique. Le genre établit une classification première qui peut se subdiviser en sous-classifications selon les bouquets chronologiques. Les basses fréquences ont tendance à mieux marquer les caractères plus individualisés des œuvres : les moments de rupture d'une part (le Livre V s'individualise comme tournant qui coïncide avec la rupture qui a amené Aragon à s'éloigner du mouvement surréaliste), d'autre part les spécificités de l'écriture qui s'approprie certains mots peu répétés mais essentiels – sans doute glisse-t-on alors davantage vers ce qu'il est convenu d'appeler le stylistique, envisagé comme une appropriation de la langue par le discours d'un individu, ou encore comme « processus de singularisation » (Herschberg-Pierrot, 2006 : 31) d'une œuvre.

Toutefois, en élevant le seuil de la fréquence pour observer la distribution des formes de plus de mille occurrences, on perd la distinction chronologique pour ne plus conserver que la répartition générique. La structure grammaticale – qu'on peut voir matérialisée par ces très hautes fréquences – n'est pas *a priori* pertinente pour ce corpus d'étude pour en tester l'éventuelle variation au cours du temps. C'est ce que confirme par ailleurs l'analyse arborée qui porte uniquement sur les codes grammaticaux tout en fournissant la surprise du rapprochement de deux œuvres qu'une trentaine d'années séparent.

La distance intertextuelle, fondée sur les indices de Jaccard et portant sur les lemmes, i.e. mettant en évidence le profil lexical du corpus, donne un graphique superposable dans son ensemble à l'analyse arborée des basses fréquences. L'amorce de caractérisation sémantique n'a fait que confirmer le poids des différences de genres ; toutefois, si la répartition des substantifs en grands groupes lexicaux sommaires ne fait pas apparaître de contrastes évidents entre les ensembles chronologiques, la distribution plus fine des champs lexicaux s'est avérée plus sensible au facteur-temps.

La différenciation générique paraît une catégorie transhistorique qui fournit des critères d'évaluation stables et dominants reposant essentiellement sur l'emploi des classes grammaticales et des mots grammaticaux ; la dynamique temporelle qui suit les mouvements d'une pensée et les variations socio-historiques se laisse davantage lire au travers des réseaux lexicaux qui esquissent des isotopies mobiles.

Références

Béhar H. (in press). Le lexique surréaliste d'Aragon. In *La Langue d'Aragon*, Dijon, 11-12 mars.

Benveniste E. (1974) *Problèmes de linguistique générale*, 2. Paris : Gallimard.

Brunet E. (1988). Une mesure de la distance intertextuelle. *Revue informatique et statistique dans les Sciences humaines*, vol. 1-4 : 81-116.

Brunet E. (2009). *Où l'on mesure la distance entre les textes*. http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html.

Corpus, n° 2, *La Distance intertextuelle*.

Herschberg-Pierrot A. (2006) Style, corpus et genèse. *Corpus*, 5 : 19-36.

Muller Ch. (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette.

Annexe : Le corpus

1. 1917-1920 : *Feu de joie (Écritures automatiques, Les Aventures de Télémaque)*
2. 1921-1925 : *Le Mouvement perpétuel, Une vague de rêves*
3. 1926 : *Les destinées de la poésie, Le Paysan de Paris*
4. 1927-1929 : *La Chasse au snark, La défense de l'infini, La Grande Gaité, Le Trésor des Jésuites*
5. 1930-1933 : *La Peinture au défi, Front rouge, Persécuté persécuteur, La Littérature internationale*
6. 1934-1935 : *Hourra l'Oural, Commune*
7. 1934 : LES CLOCHES DE BALE *(roman)*
8. 1936-1937 : *La poésie soviétique à Paris, Réalisme socialiste et réalisme français*
9. 1936 : LES BEAUX QUARTIERS *(roman)*
10. 1938 : *Nouvelles du vaste monde*
11. 1939-1941 : *Le Crève-cœur, Les Yeux d'Elsa*, *La Leçon de Ribérac*
12. 1942 : *En français dans le texte*, *Brocéliande*
13. 1942 : LES VOYAGEURS DE L'IMPÉRIALE *(roman)*
14. 1943-1945 : *Musée Grévin*, *Diane française, L'enseigne de Gersaint*
15. 1944 : AURÉLIEN *(roman)*
16. 1945 : SERVITUDE ET GRANDEUR DES FRANÇAIS *(roman)*
17. 1946-1952 : *Chroniques du Bel canto*, *Nouveau crève-cœur*, *Mes caravanes*, *La Patrie en danger*
18. 1951 : LES COMMUNISTES 1 *(roman)*
19. 1951 : LES COMMUNISTES 2 *(roman)*
20. 1956 : *Le Roman inachevé*
21. 1963 : *Le Fou d'Elsa*
22. 1965 : LA MISE À MORT *(roman)*
23. 1967 : BLANCHE OU L'OUBLI *(roman)*

Dans un souci de clarté, les titres des œuvres narratives sont notés en majuscules ; les recueils poétiques portent le titre du premier texte qui apparaît dans le volume, en minuscules.