Semaine Data-SHS (décembre 2021)
"Traiter et analyser des données en sciences humaines et sociales"

Plateforme Universitaire de Données (TIR-PROGEDO)
Université de Grenoble-Alpes
Maison des sciences de l'homme

Cyril Labbé
Laboratoire informatique de Grenoble – Université Grenoble-Alpes
cyril.labbe@imag.fr

Dominique Labbé
Université Grenoble-Alpes (r)
dominique.labbe@umrpacte.fr

Pierre Hubert †
Association Internationale des Sciences Hydrologiques

Les polars dans la Bibliothèque électronique du français moderne

10 décembre 2021

La Bibliothèque Electronique du Français Moderne (BEFM en ligne sur le site l'Université de Grenoble) comporte des milliers d'oeuvres littéraires et notamment une collection de romans policiers contemporains. Après une présentation de la BEFM et de son utilisation, les caractéristiques particulières des "polars" sont étudiées au niveau du vocabulaire et du style. Ce "sous-ensemble" littéraire met en scène des personnages singuliers. Le récit s'organise autour d'une énigme criminelle. Il est soumis à une forte contrainte temporelle. L'action domine, les phrases sont plus courtes et plus simples que dans le roman traditionnel. Enfin, la BEFM peut être utile pour l'étude de la littérature et, au-delà, pour la connaissance et l'enseignement du français.

A la mémoire de Pierre Hubert (1943-2020)

Pendant plus de cinquante ans (1969-2020), il a participé à nos travaux de statistique appliquée au langage.

Pierre Hubert a joué un rôle important dans le développement de l'hydrologie, au sein de l'Ecole des Mines de Paris et surtout à la tête de l'Association Internationale des Sciences Hydrologiques (IAHS) dont il a été :
- vice-président puis président de la section française (1987-2000)
- secrétaire général de l'association internationale (2000-2011)

En 2015, il a reçu la médaille Volker (le "Nobel" des hydrologues délivré par l'UNESCO et l'Organisation Mondiale de la Météorologie) pour l'ensemble de son œuvre et pour son leadership exceptionnel dans le développement des sciences hydrologiques.

Le roman policier - ou « polar » - est un succès commercial : d'après Collowald et Neveu (2013), plus d'un roman sur quatre vendus en France serait un policier. Sans doute en partie à cause de cet engouement du public, il a acquis ses lettres de noblesse dans la littérature française. En témoignent les nombreuses études universitaires qui lui sont consacrées depuis un demi-siècle. L'élan a été donné par l'étude de Boileau et Narcejac aux Presses Universitaires de France (1975), la traduction en français de l'étude de Kracaeur (1981) puis celle de Lacassin en 1993. Les analyses savantes se sont multipliées sur quelques auteurs mais aussi à propos de ce que certains pensent être un nouveau genre littéraire. Il y a même, à notre connaissance, deux dictionnaires : Mesplède (2003) et Tulard (2005).

Cependant, à la lecture de ces études, il est difficile d'identifier quelles sont les caractéristiques lexicales et stylistiques de ces œuvres et de comprendre en quoi elles se distinguent du reste de la fiction romanesque.

La statistique lexicale peut apporter quelques éléments de réponse à ces questions. L'étude des polars français a été possible grâce à l'aide de Pierre Hubert qui, peu avant de disparaître, nous a aidés à constituer une section particulière de la Bibliothèque Électronique du Français Moderne (BEFM) (http://lexicometrie.univ-grenoble-alpes.fr/)

Cette bibliothèque a été présentée, ici même, lors de la semaine Data-SHS organisée l'année dernière par la Maison des Sciences de l'Homme Alpes (Urien, Labbé et Labbé, 2020). Elle utilise les logiciels CQPweb (Evert 2009) et TXM (Heiden et Al. 2010).

Après avoir décrit les méthodes et les corpus utilisés pour cette expérience, on analysera les caractéristiques particulières du vocabulaire et du style de la littérature policière.

I. Corpus et méthodes

Les corpus utilisés pour cette étude sont présentés en annexe. Après avoir décrit ces deux ensembles, les méthodes de traitement seront exposées.

Les polars

Le corpus "Romans policiers" correspond aux premiers objectifs que nous nous étions fixés avec P. Hubert avant sa mort pour atteindre une certaine représentativité : amplitude de près d'un siècle - de 1931 (apparition de Maigret) à 2020 - au moins une soixantaine d'œuvres par une vingtaine d'auteurs différents dont aucun ne devait peser plus de 10% du corpus. Des œuvres réparties à peu près uniformément sur l'ensemble de la période. Nous avions convenu que, si cette première exploration se révélait intéressante, le corpus serait augmenté pour une étude plus approfondie.

On remarque que, dans le corpus actuel, deux auteurs pèsent à peu près 10% chacun (Simenon et Thilliez) ; que deux autres auteurs n'ont pour l'instant qu'une seule œuvre (Djian et Manchette) alors que nous avions convenu qu'il en faudrait au moins deux par auteur. Il manque aussi un

certain nombre d'écrivains auxquels tenaient P. Hubert – notamment Daenincks, Fajardie, Izzo, Jonquet, Villiers… - qui seront ajoutés dans les mois à venir.

P. Hubert avait souhaité faire figurer beaucoup d'œuvres ayant marqué le genre. Cependant, il a fallu composer avec certaines difficultés techniques, notamment l'impossibilité d'utiliser l'OCR sur certaines éditions anciennes pour lesquelles n'existe aucune édition électronique. En effet, nous avons pu récupérer environ la moitié des fichiers électroniques, notamment à partir des ebooks et l'autre moitié des textes ont été scannés par P. Hubert puis par D. Labbé. Les mois à venir seront consacrés à ce travail ainsi qu'à une analyse plus approfondie des "polars".

Un point de comparaison : les romans contemporains

Pour déterminer la singularité des "polars", on tire de la BEFM un corpus de comparaison présentant les mêmes caractéristiques : 22 auteurs, un peu plus de 70 romans parus durant le dernier siècle (liste en annexe). Cependant, ici deux auteurs dépassent les 10% (Gary et Le Clézio).

Sous cette réserve, la comparaison permettra d'apporter une première réponse à quelques questions concernant les polars, particulièrement l'existence d'un "genre" spécifique. A ce propos, on remarquera que quatre auteurs sont présents dans les deux corpus (Dard, Lemaître, Pennac, Vian), ce qui permettra de voir si l'écrivain change significativement son vocabulaire et sa manière d'écrire quand il passe des romans traditionnels aux polars.

La norme de dépouillement

Avant d'entrer dans la bibliothèque électronique, chaque texte est soumis à une série de traitements préalables – correction et standardisation orthographique, balisage du para-texte, découpage en mots et lemmatisation. Cette dernière opération consiste à attacher à chacun des mots du texte une étiquette comportant le vocable correspondant, c'est-à-dire son entrée de dictionnaire : mot vedette et catégorie grammaticale. Par exemple, "est" reçoit l'étiquette "*être*, *verbe au présent*" ou "*est substantif masculin"* (point cardinal) ; "suis" peut être rattaché aux verbes *être* ou *suivre*, "le" est un article ou un pronom, etc. L'homographie touche plus du tiers des mots de tout texte écrit en français et, comme le suggèrent les trois exemples ci-dessus, elle concerne des mots très fréquents.

La norme de dépouillement utilisée pour découper les mots et les rattacher à leurs vocables s'inscrit dans la ligne des travaux de Muller (1963 et 1967). Elle se calque sur les principes de la lexicographie française et sur la pratique des usagers du français (Labbé 1990).

Des outils informatiques permettent de réaliser cet étiquetage de manière automatisée pour la plus grande partie du traitement (Pibarot, Picard et Labbé 1995), avec le souci de limiter au maximum le nombre d'erreurs (par rapport à la norme de dépouillement). Pour les textes mis en ligne, ce taux est inférieur à 0,3% et nous travaillons à le réduire).

En suivant ces normes de dépouillement, à l'heure où ces lignes sont écrites, les deux corpus utilisés pour l'expérience comportent 10 467 381 mots – sans compter les signes de ponctuation.

Ils sont rattachés à 64 531 vocables. Encore faut-il préciser qu'il y a, parmi ces vocables, 20 296 noms propres, 2 880 mots étrangers et 325 expressions et interjections, ce qui réduit à environ 40 000 le nombre de vocables français présents dans ces 145 ouvrages.

L'ensemble des textes annotés constitue la BEFM, soit actuellement un corpus de plus de 68 millions de mots dont une cinquantaine de millions sont librement consultables en ligne. Les outils de consultation (CQP et TXM) permettent d'obtenir les concordances d'un mot/vocable ou d'un groupe de mots, des listes de fréquences, des index indiquant par exemple, le poids des catégories grammaticales dans un corpus (le tableau 6 ci-dessous en donne un exemple).

Les renseignements fournis par la BEFM éclairent les questions posées en introduction de l'exposé et, en premier lieu, celle du genre.

II. Existe-t-il un « genre » policier distinct du reste de la fiction romanesque ?

Une réponse objective à cette question ne laisse pas de place à l'intuition du chercheur et doit reposer uniquement sur des données vérifiables et reproductibles. La classification automatique répond à ces impératifs. Elle opère des regroupements dans de vastes populations sans aucune intervention de l'observateur. La meilleure classification possible sera celle qui minimise les distances entre les individus classés ensemble et qui maximise les distances entre les différents groupes ainsi créés (Sneath & Sokal 1973).

Nous utilisons la distance intertextuelle combinée à la "classification arborée". Cette classification est classique en génétique, en biologie (Holmes 1999, Felsenstein 2004) ou en linguistique historique (Embleton 1986). Elle repose sur la propriété suivante : si les distances séparant les individus étudiés présentent les propriétés requises d'une distance, il existe un "arbre" qui représente, le mieux possible sur un plan, les positions respectives de ces individus les uns par rapport aux autres et les meilleurs groupements possibles entre eux. L'algorithme a été présenté pour la première fois dans X. Luong (1988). Pour une présentation complète de la méthode suivie ici : Labbé et Labbé (2011).

Le « genre » est une collection de règles qui s'imposent aux auteurs. Par exemple, la poésie (Labbé et Labbé 2009), la correspondance (Labbé et Labbé 2009), le théâtre, ou le roman (Labbé 2011).

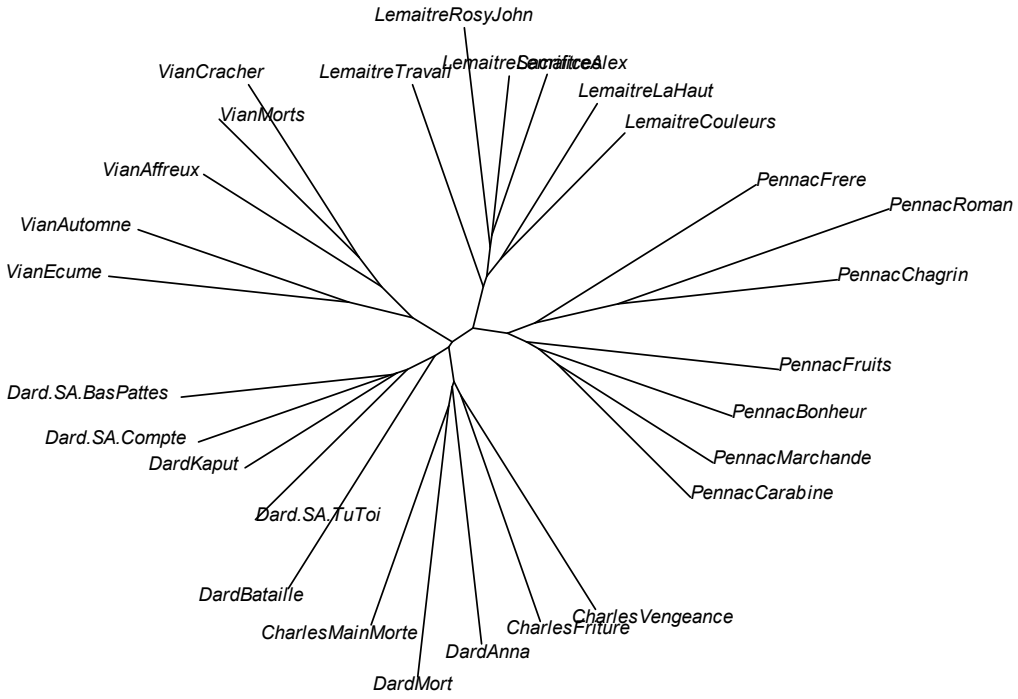
Si le "polar" est un genre distinct du roman traditionnel, la classification rangera les textes en deux ensembles disjoints (P et R comme ils sont présentés dans les annexes). A l'inverse, si la fiction romanesque est un genre unique incluant le polar, les textes seront rangés en 38 ensembles (autant qu'il y a d'auteurs).

Menons cette expérience sur les quatre auteurs qui sont dans les deux corpus (Dard, Lemaitre, Pennac et Vian). La figure 1 présente l'arbre obtenu sur les 28 textes de ces auteurs.

Dans cette figure, les feuilles terminales figurent les textes ; les nœuds intermédiaires donnent les meilleurs groupements possibles, c'est-à-dire ceux pour lesquels les distances entre les éléments qui composent le groupe sont les plus faibles et les distances les séparant des autres groupes les

plus grandes possibles. La distance entre deux points est figurée par le chemin les unissant et la longueur de ce chemin est proportionnelle à la distance originelle correspondante. Il ne faut pas attacher d'importance au placement des branches rattachées à un même nœud, spécialement pour les branches terminales.

Figure 1. Classification arborée sur les romans (policiers et classiques) de Dard, Lemaître, Pennac et Vian.



Graphique réalisé en novembre 2021avec R (version 3.5.1 – librairie APE) (R Core Team 2021 ; Becue 2017 ; Savoy 2020)

Dans le sens des aiguilles d'une montre à partir du haut.

Lemaître : *Travail soigné*,

Policiers : *Rosy & John*, *Sacrifices*, *Alex* ;

Romans : *Au Revoir là-haut*, *Les Couleurs de l'incendie*.

Pennac Romans : *Mon Frère, Comme un roman, Chagrin d'école* ;

Pennac Policiers : *Aux Fruits de la passion, Au Bonheur des ogres, La Petite marchande de prose La Fée Carabine*

Dard & Charles : *Vengeance*, *La Grande friture*, *Anna*, *Quand la mort vient*, *La Main morte*.

Dard roman : *Batailles sur la route*

Dard (Kaput et San Antonio) : *A Tue et à toi*, *Kaput*, *Réglez-lui son compte*, *Bas les pattes*.

Vian Romans : *L'Ecume des jours, L'Automne à Pékin*

Vian Policiers : *On Tuera tous les affreux, Les Morts ont tous la même peau, J'irai cracher sur vos tombes* ;

En comparant les distances originales et les chemins correspondants sur l'arbre, on peut mesurer la fidélité de la représentation. Ici tous les groupes sont agrégés à partir de nœuds dont l'indice de qualité est supérieur à 95%. Autrement dit, il y a moins de 5% de chances de se tromper en affirmant que la figure représente fidèlement les proximités relatives entre les textes et que les 4

groupes principaux et les 8 sous-groupes isolés par cette classification correspondent aux meilleurs classements possibles respectivement d'ordres 4 et 8.

Pour comprendre la complexité de l'opération et le caractère non-trivial des résultats, il faut se souvenir qu'avec 28 individus, on peut former 378 couples différents, 6 552 trios, etc. Cette combinatoire étant immense, le fait de parvenir – sans intervention de l'observateur - à un classement qui semble "naturel" n'a rien d'évident.

Cette classification amène deux conclusions.

Premièrement, tous les textes sont assignés à leurs auteurs. On remarque aussi que se trouve confirmée la paternité de Dard sur les romans signés Charles, Kaput et San Antonio (Dard les a reconnus de son vivant). L'outil peut donc être utilisé pour résoudre certaines énigmes posées par les pseudos dans la littérature policière. Un sondage dans Mesplède et Schleret (1995) montre que les auteurs de plus du dixième des ouvrages parus dans la *Série noire* se dissimulent derrière des pseudonymes, sans compter tous ceux pour lesquels on ne dispose pas d'élément biographique fiable. Ainsi Dard se serait dissimulé derrière au moins une cinquantaine de pseudos (Clément 2009 ; Rivière 1999 et le site "toutdard.fr"). Le polar français mérite donc de se voir appliquer des outils d'attribution d'auteur comme ceux qui ont permis d'identifier l'écrivain qui se cache derrière Elena Ferrante (Savoy 2018 ; Tuzzi et Cortelazzo 2018).

Deuxièmement, les textes sont également assignés à leur catégorie respective (P ou R) à l'exception du *Travail soigné* de Lemaître qui semble à peu près équidistant des autres romans de l'auteur qu'ils appartiennent à P ou à R. Pour Dard, *Bataille sur la route* - roman non-policier - se place à part des deux groupes qui correspondent à un classement chronologique et thématique dans les policiers publiés par Dard sous divers noms (outre le sien) : Charles, Kaput et San Antonio (SA).

Cette expérience infirme donc la théorie selon laquelle le "roman policier" serait un genre à part entière distinct du roman traditionnel. Cependant, il existe bien des règles particulières, puisque, au sein des groupes principaux (les auteurs), les polars se placent à part des autres romans. Par conséquent, les policiers appartiennent bien au genre de la fiction romanesque, mais ils semblent ne pas être écrits de la même manière que les autres romans.

F. Dard illustre la porosité de la frontière entre les deux ensembles. En 1954, alors qu'il commençait à rencontrer le succès avec ses policiers, il a publié chez son éditeur lyonnais, deux romans qu'il voulait "sérieux" : *Anna Soleil* et *Quand la mort vient*. Contre l'avis de Dard, l'éditeur a placé le second dans sa collection "Spécial Police" (raison pour laquelle, nous avons placé *a priori* ce roman dans le corpus des policiers). La classification donne raison à l'éditeur et ne fait pas grâce non plus à *Anna Soleil* qui semble avoir été en quelque sorte "contaminé" par les polars publiés à la même époque chez le même éditeur (sous les noms de Dard et Charles).

Ces constats permettent de formuler l'hypothèse selon laquelle le polar serait une sorte de sous-ensemble dans la fiction romanesque, tout comme la tragédie, la tragi-comédie ou la comédie se distinguent nettement au sein du genre théâtral. Mais il faut ajouter que, dans le cas du roman policier, la frontière semble parfois poreuse.

La lexicométrie permet de donner consistance à ces hypothèses.

III. Comparaison des deux corpus

Il s'agit de comparer deux corpus de longueurs approximativement égales et d'identifier les vocables caractéristiques de chacun. La méthode statistique employée ici est présentée dans Monière et Labbé (2019).

Prenons en exemple le substantif le plus utilisé dans les polars (le nom féminin "heure" au singulier et au pluriel). Grâce à la BEFM, on dispose de trois informations à son sujet :

- le rang : premier substantif le plus employé dans P (les polars), il figure au neuvième rang dans R (les romans contemporains) ;

- les effectifs (E_i). Il y en a 7 664 dans P et seulement 5 164 dans R. Comme les deux corpus n'ont pas exactement la même longueur, on utilise une troisième information.

- les fréquences (F_i). Dans P, *heure* apparaît en moyenne 1,53 fois tous les mille mots et 0,94 dans R. Autrement dit, la densité de ce mot est de 61% plus élevée dans P par rapport à R.

Remarque : si la convention "pour mille mots" s'est imposée à la place des pourcentages, cela tient à ce que ces densités sont très faibles. Ici pour le substantif le plus utilisé, cette fréquence est seulement de 0,153 %.

La différence de densité entre les deux corpus est donc considérable. Peut-on pour autant considérer que "heure" est une caractéristique des polars ? Un test statistique simple permet de répondre à cette question. Ce test diffère légèrement de la procédure standard qui consiste à considérer le corpus étudié comme un échantillon tiré d'un très vaste ensemble de référence que l'on nomme "population parente". En effet, nous ne disposons pas d'un corpus de référence très grand par rapport à l'échantillon mais simplement de deux corpus de taille équivalente.

Deuxième difficulté. Comme on le verra plus bas, la densité de la catégorie grammaticale "substantif" n'est pas la même dans les deux corpus. Pour surmonter cette difficulté, on considère que l'apparition, dans un texte, du mot "heure" résulte de deux choix successifs, le premier consiste à placer un substantif à cet endroit précis et le second choix se fait en faveur de "heure". Dès lors, pour juger de la singularité de ce vocable, il faut neutraliser l'influence du premier choix (en faveur du substantif).

Sachant que l'effectif total des substantifs (Ns) est :

- dans P : Nsp = 876 150

- dans R : Nsr = 968 338.

Si les corpus P et R sont issus de tirages aléatoires dans la même population parente, la fréquence de "heure" dans cette population parente inconnue ($Ftheo_i$) est estimée à :

$$Ftheo_i = \frac{E_{ip} + E_{ir}}{N_{sp} + N_{sr}} = \frac{7\,664 + 5\,164}{876\,150 + 968\,338} = 0{,}06955$$

D'où l'on tire l'effectif théorique (du vocable étudié) attendu dans P (sous l'hypothèse d'échantillons tirés dans le réservoir des substantifs d'une population parente unique) :

$Etheo_{ip} = Nsp * Ftheo_i =$ 876 150* 0,06955 = 6 093

Ce qui donne un écart entre cette prévision et la valeur observée de :

$Ecart_{ip} = E_{ip} * Ftheo_{ip} = 7\ 664 - 6\ 093 = +\ 1\ 571$

Le même calcul donne l'effectif théorique du vocable dans R :

$Etheor_{ir} = 968\ 338 * 0{,}06955 = 6\ 735$

Pour les romans, l'écart entre la valeur observée et la valeur attendue est de -1 571 (puisque les écarts se compensent nécessairement).

On associe à ces valeurs une déviation standard ou écart type.

Sous l'hypothèse d'un échantillon tiré d'une population parente unique, la probabilité (*p*) qu'un substantif employé dans les polars soit le vocable *heure* est de :

$$p = \frac{Etheo_{ip}}{Ns_p} = \frac{6\ 093}{876\ 150} = 0{,}0070$$

La probabilité (*q*) de l'événement contraire (un autre substantif)

$q = 1 - p = 0{,}9930$

D'où l'on tire un écart type théorique (ou déviation standard) autour de la valeur attendue :

$$\sigma = \sqrt{Nsp.p.q} = \sqrt{876\ 150 * 0{,}007 * 0{,}993} = 77{,}8\ mots$$

On obtient un « écart réduit » (z-score) en divisant l'écart observé ($Ecart_{ip}$) par cette déviation standard :

$$z = \frac{Ecart_{ip}}{\sigma} = \frac{1\ 571}{77{,}8} = 20{,}2$$

Autrement dit, l'écart entre la valeur observée et la valeur attendue est égal à vingt fois l'écart type. Dans le cas de deux échantillons tirés dans une population parente unique, 66% des observations seraient comprises entre la valeur attendue ± un écart-type ; 95% dans un intervalle de ± 2σ ; 99% dans un intervalle de ± 2,6σ et 99,9% dans un intervalle de ±3,3σ (courbe en cloche ou de Laplace-Gauss). De ces propriétés, on tire un risque d'erreur (accepter ou rejeter à tort l'hypothèse d'un usage semblable dans les deux échantillons). Par exemple, pour les écarts positifs, le risque sera inférieur à 5% avec {z ≥ 2 et z < 2,6} ; inférieur à 1% avec {z ≥ 2,6 et z < 3,3} ; inférieur à 1‰ pour {z > 3,3}, etc.

Dans la suite de cette communication, ces probabilités sont exprimées grâce à un indice unique :

++ : emploi significativement supérieur avec un risque d'erreur inférieur à 1 chance sur mille,

+ : emploi significativement supérieur avec un risque d'erreur inférieur à 1 chance sur cent,

≈ risque d'erreur supérieur à 0,01 et inférieur à 0,99. On ne peut rejeter l'hypothèse selon laquelle il n'y a pas de différence significative entre les emplois du vocable dans les deux corpus

- : emploi significativement inférieur avec un risque d'erreur inférieur à 1 chance sur cent,

-- : emploi significativement inférieur avec un risque d'erreur inférieur à 1 chance sur mille.

Généralisation : Si l'effectif d'un vocable dans un texte s'écarte significativement de l'effectif attendu, le vocable est une caractéristique positive ou négative de l'un des corpus par rapport à l'autre. A l'inverse, si l'effectif de ce vocable ne diffère pas significativement dans A et B, le vocable est considéré comme "non-caractéristique" (on le suppose "commun" aux deux corpus). Ajoutons,

que lorsque les deux corpus sont de longueurs approximativement égales, comme ici, les caractéristiques positives de l'un sont les négatives de l'autre.

Ce calcul est simple à mettre en œuvre dans un tableur. Il s'agit d'une approximation fiable pour les effectifs élevés (au moins une dizaine). Le calcul exact (loi hypergéométrique) est présenté dans Monière et Labbé (2019). Nous espérons pouvoir un jour mettre ces logiciels à disposition des chercheurs.

On a donc beaucoup moins d'une chance sur mille de se tromper en affirmant que la préférence pour le mot "heure" est une caractéristique du vocabulaire des polars par rapport à la littérature générale. Une fois cette singularité identifiée, il faut se demander si le vocable n'est pas inséré dans un réseau sémantique (que l'on pourrait intituler la "durée") caractéristique du roman policier. Le même calcul montre que sont également caractéristiques de P : le *temps* (*avoir* ou *ne pas* avoir le *temps* sont deux des syntagmes favoris dans ces romans). Dans le vocabulaire caractéristique, on trouve également *minute*, *seconde*, *an* et *année* (mais pas "jour" car ce mot signifie aussi "clarté" et "lumière" qui sont plutôt bannies du polar). Autrement dit, la contrainte du temps est une dimension essentielle du polar. Elle est en tout cas nettement plus présente que dans les romans contemporains. Elle fait peser sur les principaux personnages une tension particulière qui est censée rejaillir sur le lecteur.

Ce calcul permet donc de reconstituer le vocabulaire caractéristique des polars.

IV. Le vocabulaire des polars

Dans un premier temps, on examine les vocables les plus fréquents, comme on vient de le faire pour *heure*. Les tableaux ci-dessous donnent les principaux vocables, classés par catégories grammaticales. Les fréquences sont calculées sur l'ensemble des mots afin de faciliter les comparaisons entre catégories.

Le groupe des verbes

Les verbes sont habituellement associés à des pronoms et à des adverbes. On examine donc ces trois catégories ensemble.

Jusqu'au 12e rang, polars et romans partagent les mêmes verbes usuels. Il en est généralement ainsi dans tous les grands corpus, du moins pour les trois ou quatre premiers verbes qui se classent toujours dans cet ordre : *être*, *avoir*, *faire*, *dire*). Autrement dit, ces outils appartiennent au cœur de la langue et aucun locuteur ne peut s'en passer, tout juste peut-il faire varier les densités. Par exemple, la préférence apparente pour "avoir" s'explique avant tout par une utilisation plus importante du passé composé des verbes (voir tableau 6 ci-dessous). De même pour "aller" à cause de son rôle de "pseudo-auxiliaire" qui permet de placer une action dans le futur tout en utilisant le présent (il va faire). Or, comme on va le voir plus bas, le présent est plus utilisé dans les polars que dans les autres romans.

Tableau 1. Les verbes les plus usuels des romans policiers rangés par fréquence décroissante (en ‰ mots), comparés aux romans.

Polars Rang	Vocable	Fréquence	Romans Rang	Fréquence	Indice (P/R)
1	être	25,69	1	25,10	≈
2	avoir	24,06	2	20,83	++
3	faire	5,80	3	5,26	++
4	dire	4,59	4	4,90	-
5	aller	3,57	6	2,95	++
6	pouvoir	3,00	5	2,97	--
7	savoir	2,39	8	2,12	+
8	voir	2,37	7	2,44	-
9	vouloir	1,94	9	1,77	++
10	venir	1,64	10	1,65	-
11	prendre	1,62	11	1,42	++
12	devoir	1,58	12	1,30	++
13	passer	1,44	16	1,17	++
14	demander	1,37	22	0,93	++
15	trouver	1,21	19	0,98	++
16	falloir	1,19	13	1,27	--
17	mettre	1,17	18	1,04	--
18	parler	1,07	15	1,18	--
19	croire	1,07	17	1,05	≈
20	laisser	1,02	27	0,79	++

Ces principaux verbes indiquent les dimensions essentielles du polar (outre la contrainte temporelle).

Premièrement, la détention de l'information (*savoir*), la *demande* de celle-ci et sa découverte (*trouver*) jouent un rôle clef dans le récit, toujours organisé autour d'une énigme. En effet, parmi les verbes les plus caractéristiques des polars, on trouve également, plus bas dans le classement, par ordre de caractéristique : *questionner*, *planquer*, *vérifier*, *découvrir*, *piger*, *expliquer*, *comprendre*, *annoncer*… Les substantifs correspondants apparaissent également dans les caractéristiques positives principales des polars (*question*, *planque*, *vérification*, *explication*, etc.). Avec la contrainte temporelle, l'énigme et sa recherche sont donc la seconde dimension caractéristique du polar.

Ensuite, on trouve la volonté (*vouloir*), la captation (*prendre*) ou l'action (*prendre* une chose, une attitude, une décision, un coup) les modalités de l'obligation morale ou matérielle (*devoir*), le déplacement (*passer*), l'abandon (*laisser*). En examinant les mots avec lesquels se combinent *laisser*, on trouve comme sens principal *abandonner* (un lieu, une action, une personne, une chose), *laisser tomber*.

Les pronoms, spécialement personnels, accompagnent les verbes.

Tableau 2. Les pronoms les plus usuels des romans policiers rangés par fréquence décroissante (en ‰ mots), comparés aux romans.

Polars Rang	Vocable	Fréquence	Romans Rangs	Fréquence	Indice P/R
1	il	28,01	1	27,91	-
2	je	25,32	2	21,99	++
3	se	12,14	3	11,66	≈
4	ce	10,23	4	9,48	++
5	le	8,25	6	7,69	+
6	vous	7,21	9	4,69	++
7	qui	7,14	5	8,20	--
8	que	6,09	7	6,07	≈
9	tu	5,49	14	3,32	++
10	on	5,10	8	5,02	+

La préférence va au "je" qui indique à la fois un récit à la première personne et une forte présence du dialogue (avec "tu" et "vous"), couplé au fait de montrer ("ce"). Enfin, le complément d'objet le plus simple (le) est également privilégié.

A l'inverse, le polar utilise peu les relatifs : *qui*, *que* (et plus bas dans les listes, *dont*, *lequel*, etc.). Cette caractéristique s'explique avant tout par une construction des phrases plus simples dans le polar que dans les romans contemporains (construction que nous examinons plus bas)

Tableau 3. Les adverbes les plus usuels des romans policiers rangés par fréquence décroissante (en ‰ mots), comparés aux romans.

Polars Rang	Vocable	Fréquence	Romans Rang	Fréquence	Indice (P/R)
1	ne	14,01	1	13,91	+
2	pas	11,06	2	8,95	++
3	plus	4,44	3	5,64	--
4	bien	2,50	4	2,81	--
5	où	1,64	5	1,96	--
6	encore	1,47	6	1,79	--
7	là	1,47	10	1,50	-
8	non	1,43	17	1,08	++
9	peu	1,41	7	1,80	--
10	tout	1,39	8	1,59	--

Les principaux adverbes sont plutôt plus rares dans les polars à l'exception de la contradiction (surtout "pas" et "non"). Plus bas dans la liste, le "oui" est très sous-employé par rapport aux romans. Ces constats sont à mettre en relation avec la prédominance des rapports "je-tu" et "je-vous". On sait donc maintenant que ces rapports sont dominés par la tension et la négation (ou la contradiction). Dans la suite de la liste, les adverbes suivants sont caractéristiques des polars : *très*,

hier, *demain*, *rapidement*, *brutalement*. On retrouve ainsi le temps et la violence comme dimensions essentielles du polar.

Le groupe du nom

Ce groupe est composé des noms communs (tableau 4), des adjectifs (tableau 5) et des déterminants.

Tableau 4. Les substantifs les plus usuels des romans policiers rangés par fréquence décroissante (en ‰ mots), comparés aux romans.

Polars Rang	Vocable	Fréquence	Romans Rang	Fréquence	Indice (P/R)
1	heure	1,53	9	0,95	++
2	main	1,43	5	1,22	++
3	homme	1,42	1	1,68	--
4	oeil	1,39	2	1,27	++
5	temps	1,21	7	1,12	++
6	chose	1,16	4	1,25	-
7	fois	1,13	6	1,17	≈
8	coup	1,12	14	0,80	++
9	tête	1,12	12	0,88	++
10	femme	1,12	8	0,99	++
11	jour	0,92	3	1,26	--
12	porte	0,92	24	0,59	++
13	monsieur	0,71	10	0,92	--
14	moment	0,71	15	0,79	-
15	fille	0,70	29	0,53	++
16	an	0,67	22	0,62	++
17	nuit	0.67	16	0.79	--
18	air	0.65	19	0.72[22222]	--
19	rue	0.64	35	0.49	++
20	vie	0.63	11	0.91	--

Ici les contrastes entre les deux corpus sont plus nets tant au niveau des rangs que des fréquences. Outre la contrainte temporelle (dans le tableau : *heure*, *temps*, *an*), on note l'importance de la *main*, de la *tête* (dont un des synonymes est *visage*), notamment l'*œil*, du *coup*, de la *femme* – alors que "homme" est avec "monsieur" des caractéristiques négatives du polar - et des *filles*. Les principaux réseaux sémantiques attachés à ces vocables sont décrits plus bas.

Tableau 5. Les adjectifs les plus usuels des romans policiers rangés par fréquence décroissante (en ‰ mots), comparés aux romans.

Polars Rang	Vocable	Fréquence	Romans Rang	Fréquence	Indice (P/R)
1	petit	1,74	1	1,62	++
2	seul	0,91	3	1,10	--
3	bon	0,90	4	0,72	++
4	grand	0,84	2	1,35	--
5	nouveau	0,63	6	0,63	=
6	jeune	0,62	5	0,68	-
7	dernier	0,54	10	0,47	+
8	vieux	0,50	7	0,62	--
9	beau	0,41	8	0,55	--
10	sûr	0,41	15	0,36	++

A nouveau, on constate l'existence de couples de mots. Par exemple, *bon* et *mauvais* (en 18e position) sont deux caractéristiques principales des polars. De même *petit* qui est associé à *gros* et non pas à *grand*, car la fonction principale de "petit" est d'indiquer une familiarité plus qu'une caractéristique physique. Enfin, *sûr* est couplé à *certain* qui sont à relier à la dimension de la connaissance. A l'inverse, *jeune* et *vieux* sont tous deux des caractéristiques négatives (le polar est affaire d'adultes).

Dans les déterminants, les polars se caractérisent essentiellement par une sur-utilisation des chiffres, dates, nombres à mettre en relation avec la contrainte de la durée (l'heure, la minute, la seconde sont généralement accompagnées d'un nombre).

Personnages et actions

Dans les tableaux ci-dessus, les écarts les plus significatifs (à la fois rang et fréquence) s'observent dans la catégorie des substantifs. Le tableau en annexe 2 détaille les principaux substantifs caractéristiques des polars au-delà des vingt premiers déjà décrits dans le tableau 4 ci-dessus. Ces informations apportent une réponse synthétique aux questions suivantes : quels sont les personnages, les lieux, les actions les plus caractéristiques du polar ?

Les personnages se classent en quelques catégories.

- En premier lieu, le tableau vérifie la justesse du nom générique de "roman policier". En effet, les principaux personnages sont d'abord les *flics* et plus précisément, *commissaire*, *inspecteur*, *lieutenant*, *commandant*, secondairement "policier" ou "gendarme". Ils conduisent des *enquêtes*, recherchent des *indices* et *interrogent* des *suspects* et les placent en *garde à vue*…

- en face des flics il y a des "types" (quasi-synonyme de "homme"). Ils sont souvent en *bande*. Ce sont des *gars*, des *copains* mais aussi souvent des *braqueurs*, *tueurs* ou *assassins*.

- avec les *types*, des "femmes", mais surtout des "filles". Alors que dans les romans, le vocable "fille" a un champ sémantique proche de celui de "fils" avec le sens de "enfant de...", ici les synonymes sont plutôt à rechercher du côté de *putain*, *môme*, *souris* et *poupée*… Non seulement, ces vocables sont beaucoup plus employés dans les polars que dans les romans, mais de plus, ils n'ont pas le même sens. Dans les romans, les "mômes" sont des garçons, ou des enfants désignés par ce masculin familier quel que soit leur sexe (la majorité des emplois se rencontrent chez Céline et Gary), les "souris" sont des rongeurs et les "poupées", des jouets.

Certains *types* (parfois des *filles*) sont "morts". En effet, le *cadavre* (ou *macchabée*) et la *victime* sont des personnages importants du polar. A ce propos, on remarque que les substantifs "mort" au masculin et surtout au féminin sont moins employés dans les polars que dans la littérature générale. En effet, à partir de l'entre-deux guerres, *la mort* est devenue un thème essentiel de la littérature générale. Dans le polar, il s'agit d'un *assassiné* dont il faut rechercher l'*assassin*.

Outre la recherche de l'information, les activités principales des personnages semblent être de distribuer les *coups* (de *tête*, d'*œil*, de *main*, de *feu*, de *poing*, de *pied*, d'*épaule*…). Ils *agissent* – et *réagissent* - toujours *vite* et *brusquement*. Ils *courent*, *entrent* et *sortent*, *ouvrent*, *ferment* (*claquent*) des *portes* (aussi "lourdes"). Ils se déplacent en *voiture* – ou en *tire* - pour *aller* dans les *bars* et les *restaurants*, *boire* des *verres* et fumer des *cigarettes* (aussi *clopes*), voir les *filles* et monter dans leurs *chambres*. A tout propos, ils sortent les *armes* : *flingue*, *revolver*, *couteau*, *lame*, et *tirent (lancent)* des *balles*, *pruneaux*, *valdas*…

A l'opposé, les mots les plus caractéristiques des romans comparés aux polars dessinent en creux les absences relatives. Très peu de *monsieur* et *madame*, de *famille*, d'*enfant*, *frère*, *sœur*... L'emploi d'*amour* est deux fois plus faible dans les polars (idem pour le verbe *aimer*). On y rencontre peu d'*amis*, de *camarades*, d'*amitié,* de *sentiment* ni de *bonheur*. Au fond, qu'ils soient policiers ou mauvais garçons, les personnages principaux des polars sont fondamentalement solitaires, incapables d'avoir une vie sociale normale et, en premier lieu, ce sont des sans-familles (ou avec une famille pour le moins atypique comme chez Pennac !). D'autres absences relatives sont moins prévisibles. Par exemple, la *nature* - *mer*, *rivière*, *forêt*, *arbre*... - n'appartient pas au décor usuel du polar (contrairement à la *rue*). De même, le polar est généralement incolore, le *noir* et le *blanc* sont sous-employés – tant les adjectifs que les substantifs - mais surtout toutes les couleurs y compris le *rouge* (alors que le *sang* coule à flots !). C'est également le cas de toutes les saisons.

Pour compléter ce rapide tableau des manques, voici la suite de la liste des principaux substantifs – non cités ci-dessus – qui sont significativement moins employés dans les polars comparés aux romans (classés par indice décroissant) :

ciel, vent, soleil, terre, sable, colline, pierre, lumière, guerre, nuage, bruit, peuple, ville, dune, fleuve, âme, jour, vie, horizon, vallée, troupeau, professeur, navire, montagne, plaine, paysan, joie, herbe, désir, pays, poussière, parti, vague, être, force, jeunesse, monde, dieu, orgueil, oiseau, rivage, étoile, révolution, village, rocher, eau, langage, plage, bête, tente, révolte, élève, ombre

Plus fondamentalement, l'analyse révèle également le style propre au polar. En premier lieu, celui-ci apparaît particulièrement tourné vers l'action.

De l'action !

Le tableau 6 présente les densités relatives des principales catégories grammaticales dans les policiers (P)° comparées à celles des romans contemporains (R).

Tableau 6. Densités relatives des catégories grammaticales dans les romans policiers (P) comparés aux autres romans contemporains (R) (en ‰ pour mots).

Catégories	Proportion P	Proportion R	Indice (P/R)
Verbes	181,7	168,7	++
Futurs	2,4	2,3	≈
Conditionnels	3,9	3,4	+
Présent	62,8	52,3	++
Imparfaits	34,2	41,1	--
Passés simple	18,4	15,5	++
Participes passés	24,9	21,5	++
Participes présents	4,2	3,8	++
Infinitifs	31,0	28,9	++
Noms propres	30,8	24,3	++
Noms communs	174,5	177,8	-
Adjectifs	48,3	55,4	--
Adjectifs du participe passé	9,8	10,4	-
Pronoms	149,4	142,4	++
Pronoms personnels	96,5	90,1	++
Déterminants	157,2	159,2	-
Articles	110,3	114,4	-
Nombres	11,4	8,5	++
Possessifs	20,4	20,2	≈
Démonstratifs	6,9	7,0	≈
Indéfinis	8,2	9,0	--
Adverbes	75,7	78,4	--
Prépositions	133,5	140,6	--
Coordinations	24,0	28,0	--
Subordinations	21,6	22,7	--

La préférence pour le verbe est frappante. Elle s'accompagne d'une réticence équivalente envers les noms communs, les adjectifs et les déterminants (sauf les nombres).

Du point de vue linguistique, le verbe a une double fonction : la fonction "cohésive" qui organise "en une structure complète les éléments de l'énoncé" et la fonction "assertive" qui "dote l'énoncé d'un prédicat de réalité" car l'élément verbal implique une référence à un ordre qui n'est plus simplement celui du discours mais aussi celui de la réalité (Benveniste 1966, 1, p. 154). Sans doute

les histoires racontées dans les polars sont-elles particulièrement extraordinaires pour qu'il soit nécessaire d'injecter un supplément de réalité et d'action dans le récit.

Les temps des verbes sont très différents dans les deux compartiments. Les polars contiennent 20% de présents et de passés simples en plus que les romans. Ici le présent est clairement une technique narrative visant à faire vivre l'événement au lecteur. En français, le passé simple situe l'événement dans un instant précis, clairement délimité dans le temps du récit, raison pour laquelle il est préféré à l'imparfait et plus encore le passé composé – encore appelé "parfait" – qui donnent une durée à cet événement passé et abolissent, au moins partiellement, la frontière entre passé et présent.

Il n'y a apparemment pas de différence quant au futur mais l'emploi du "pseudo-auxiliaire" *aller* - nettement plus présent dans les polars - permet de dilater l'instant présent en y incluant un futur proche.

La réticence envers le nom n'est pas totale. On remarque qu'il s'agit des noms communs, pas des noms propres qui sont plus utilisés dans les polars. Ceci est à mettre en relation avec l'utilisation des pronoms personnels et des nombres (également fortement employés). En effet, patronymes et toponymes assurent l'ancrage spatial et social du récit, de même que les dates et les chiffres, les horaires, les distances, les hauteurs. On retrouve ainsi la contrainte temporelle qui est une dimension essentielle du polar et le souci de rendre plus "réel" le récit grâce à ces ancrages multiples.

Plus au fond, on remarque que les verbes, les pronoms, les adverbes et les conjonctions de subordination sont liés. Pour un niveau de complexité comparable, la préférence pour le verbe entraîne également un fort emploi des trois autres catégories. A l'inverse, noms communs, adjectifs, déterminants, prépositions évoluent ensemble. Cette caractéristique permet de former deux groupes et d'observer les variations de leurs densités (Tableau 7)

Tableau 7. Densités des groupes du nom et du verbe dans les polars (P) comparés aux romans (R) (en ‰ pour mots).

Groupe	Proportion P	Proportion R	Indice (P/R)
Verbes	428,7	412,1	++
Noms	568,2	585,0	--

On observe donc, dans les polars, une préférence marquée pour le verbe (avec les pronoms et les adverbes) et une réticence envers le nom et ses satellites (adjectifs et déterminants) La différence peut paraître faible (+8% dans les policiers par rapport à la littérature générale) mais elle est statistiquement très significative et la marge de choix est relativement faible par rapport aux contraintes de la langue et surtout aux habitudes de chacun des auteurs. Mais au sein de cette marge, les auteurs font effectivement des choix. Les corpus offrent un moyen simple de le vérifier.

Vérification

En effet, chez les quatre auteurs présents dans les deux ensembles (Dard, Lemaître, Pennac, Vian), on observe les mêmes tendances (tableau 8).

Tableau 8. Densités des groupes du nom et du verbe dans les polars (P) comparés aux romans (R) chez Dard, Lemaître, Pennac et Vian (en ‰ pour mots).

Auteur	Groupe	Proportion P	Proportion R	P/R %
Dard	Verbes	458,6	426,9	+7,4
	Noms	536,1	568,0	-5,6
Lemaitre	Verbes	434,6	418,5	+3,9
	Noms	563,3	578,9	-2,7
Pennac	Verbes	395,8	390,1	+1,5
	Noms	601,9	607,2	-0,9
Vian	Verbes	486,1	442,7	+9,8
	Noms	510,4	552,7	-7,7

Quand ils passent d'un registre à l'autre, ces quatre auteurs changent la proportion relative des catégories grammaticales, privilégiant celles appartenant au groupe du verbe quand ils écrivent un "polar". Les variations peuvent sembler faibles mais elles se manifestent sur des milliers de pages alors que ces propensions sont des choses assez stables chez un auteur donné, du moins quand il reste dans un même genre. De ce point de vue, Dard et Vian sont intéressants puisque, dans leurs polars, le mouvement est plus ample que chez les deux autres alors que l'écriture de ces oeuvres est exactement contemporaine et que, comme l'a déclaré Vian, ils voulaient faire de "vrais" polars.

Il y aurait donc bien un "style" polar que les auteurs essaient de respecter quand ils écrivent un roman destiné à ce marché. C'est ce que confirment les longueurs de phrases.

Les phrases

Les longueurs de phrases sont analysées à l'aide d'un certain nombre d'indices. Pour décrire un caractère très dispersé comme ces longueurs, la moyenne est peu représentative. On utilise plutôt d'autres valeurs centrales : mode (la longueur la plus fréquente), médiane (longueur de la phrase du milieu de la distribution), médiale (ou seconde médiane qui partage le texte en deux parts égales). Il existe aussi des indices de dispersion des longueurs : étendue, écart-type de la moyenne (et coefficient de variation), rapports entre valeurs centrales, comme la médiane et la médiale, et indice de concentration du caractère sur les individus les plus grands (Labbé et Labbé 2018). La médiane est particulièrement intéressante car elle donne approximativement la valeur correspondante à la moitié du temps de lecture.

Tableau 9. Comparaison des longueurs de phrases (en mots) dans les policiers et les romans contemporains de Dard, Lemaître, Pennac et Vian.

Auteur	Catégorie	Mode	Médiane	Médiale
Dard	Policier	6	9	14
	Roman	6	11	17
Lemaitre	Policier	4	10	19
	Roman	7	12	23
Pennac	Policier	5	9	18
	Roman	4	11	26
Vian	Policier	6	8	14
	Roman	6	7	15
Ensemble	Policier	6	9	16
	Roman	6	12	23

Ainsi, dans les polars de Dard comme dans ses autres romans, les phrases les plus nombreuses (mode) ont 6 mots. Dans ses policiers, la moitié des phrases ont moins de 9 mots (et l'autre moitié plus de 9) alors que cette médiane est de 11 mots dans les autres romans de Dard. Enfin, on lit pendant la moitié du temps des phrases de 14 mots et plus quand il s'agit d'un polar alors que, dans un roman du même auteur, cette médiale est de 17 mots (soit une augmentation d'un cinquième). Pour Lemaître, l'augmentation de la médiale est également d'un cinquième, ce qui est loin d'être négligeable. On remarque que, chez Dard, Pennac et Vian, la dimension la plus fréquente (mode) ne bouge guère d'un registre à l'autre.

Les deux dernières lignes donnent ces valeurs pour les deux corpus entiers. Les phrases les plus nombreuses ont la même longueur dans les deux corpus (6). En revanche, médiane et médiale augmentent de façon importante quand on passe de P à R. C'est donc l'étendue de la distribution qui augmente, ou encore : les phrases longues sont plus nombreuses et plus longues dans les fictions autres que policières. Toutefois, dans pratiquement tous ces textes, les phrases restent relativement brèves. A titre de comparaison, la médiale est de 50 mots chez Proust (Ragonneau, 2021) et elle est proche des 35 mots dans les romans de la BEFM pour la période allant du milieu du XIXe siècle à 1920.

Les dialogues se caractérisent par des phrases courtes. Il en est de même pour les énoncés rapportés et pour les scènes d'action. A l'opposé, les phrases longues servent aux descriptions – décors et caractères - ou à des réflexions plus philosophiques. La règle implicite semble donc que ces choses-là doivent occuper moins de place dans un polar que dans un roman ordinaire. On remarquera que, dans les romans depuis les années 1920, ces longueurs ne cessent de baisser, ce qui indique une prépondérance grandissante de l'oralité et de l'action dans les ouvrages de fiction.

Là encore, il ne s'agit que de tendances et non d'un automatisme. Vian, et Dard dans une certaine mesure, en témoignent qui restent fidèles à une phrase courte et grammaticalement simple dans leurs romans comme dans leurs polars.

Conclusions

Il n'existe pas de « genre policier ». Les ouvrages parus sous cette étiquette appartiennent à un compartiment particulier de la littérature romanesque ou de fiction dont ils respectent les principales règles. Cependant, on peut parler d'un "sous-ensemble" qui présente beaucoup de caractéristiques particulières par rapport aux autres romans contemporains, notamment un vocabulaire moins diversifié, une thématique centrée sur une énigme criminelle enfermée dans une forte contrainte temporelle, un rythme plus soutenu, une violence verbale et physique. Les personnages principaux vivent une certaine solitude, voire une a-socialité. L'action, les dialogues, ou les énoncés rapportés dominent, ce qui implique également une phrase courte et grammaticalement simple.

Résumées ainsi, les principales caractéristiques du roman policier peuvent sembler évidentes. Mais grâce à la statistique appliquée au langage, ces conclusions changent de statut. D'intuitions toujours contestables, elles deviennent des certitudes raisonnables. On voit également que l'outil semble efficace pour abstraire les trames fondamentales de plusieurs centaines de milliers de pages...

Ajoutons que cette communication ne présente qu'une petite partie des résultats obtenus et que ces résultats sont provisoires, les corpus devant être complétés prochainement. Nous reviendrons aussi sur un certain nombre d'aspects qui n'ont pu être abordés, notamment l'évolution des polars depuis une quarantaine d'années environ. La trame essentielle demeure la même, mais quelques personnages semblent gagner en épaisseur, spécialement les femmes qui tiennent maintenant des rôles principaux tout en ayant parfois un ersatz de famille avec des *mômes* (au sens familier de "enfants" et non plus dans celui de *filles* aux mœurs légères).

Des travaux ultérieurs permettront également d'affiner les portraits lexicaux et stylistiques des principaux auteurs qui ont dominé la littérature française moderne. On examinera aussi la répartition des mots (et des thèmes), le vocabulaire de chaque auteur et ses singularités par rapport aux contemporains, le sens spécifique qu'il donne à ses mots favoris, les principales combinaisons de mots ; on approfondira les thèmes, les affinités, les ruptures et les continuités ; on mesurera la richesse des vocabulaires, etc. En quelque sorte, il s'agit d'une véritable stylométrie (Savoy 2020) qui contient la lexicométrie classique mais la dépasse singulièrement et dont la plus grande partie reste à inventer.

En attendant, nous espérons avoir montré combien la bibliothèque électronique du français moderne, couplée avec les outils de la statistique lexicale, peut apporter des informations utiles aux chercheurs et aux enseignants en sciences humaines et sociales.

Remerciements et crédit.

Nous remercions les organisateurs de cette semaine DATA-SHS et les responsables de la PUD Grenoble-Alpes.

La BEFM est hébergée par le Laboratoire Informatique de Grenoble (Université de Grenoble-Alpes). Elle est interrogeable grâce à CQP-web et TXM.

Frédéric Urien a aidés à placer en ligne cette bibliothèque.

Origine des fichiers électroniques suivants :

Alexandre Clément : les deux Audiard, les trois F. Charles ainsi que les trois romans "sérieux" de F. Dard (*Batailles sur la route*, *Anna Soleil*, *Quand la mort vient*).

Margarete Katsberg : les romans de J.-M. Le Clézio.

Maurizio Lana : R. Gary : *Au-delà de cette limite*, *votre ticket n'est plus valable, L'Homme à la colombe, les Racines du ciel, Les Têtes de Stéphanie*.

Edward Arnold, Guy Bensimon, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Gaétan Péaquin, Jacques Picard, André Pibarot, Mathieu Ruhlman et Jacques Savoy ont collaboré à la mise au point des outils de lexicométrie.

Les logiciels d'analyse des corpus lemmatisés sont disponibles auprès de Cyril et Dominique Labbé.

Bibliographie

Toutes nos études citées dans ce texte sont consultables en ligne sur Research Gate et sur les Archives ouvertes du CNRS (HAL)

Benveniste Emile (1966 & 1970). *Problèmes de linguistique générale*. Paris: Gallimard (rééd. 1980).

Becue Monique (2018). *Analyse textuelle avec R*. Rennes : Presses universitaires de Rennes.

Boileau Pierre et Narcejac Thomas (1975). *Le roman policier*. Paris : PUF.

Clément Alexandre (2009). *Frédéric Dard, San-Antonio et la littérature d'épouvante*. Les polarophiles tranquilles.

Collovald Annie et Neveu Erik (2013). *Lire le noir*. Rennes : Presses universitaires de Rennes

Embleton Sheila M. (1986). *Statistics in Historical Linguistics*. Bochum : Brokmeyer.

Evert Stefan (2009). *The CQP Query Language Tutorial*.
https://cwb.sourceforge.io/temp/CQPTutorial.pdf

Felsenstein Joseph (2004). *Inferring Phylogenies*. Sunderland : Sinauer Ass.

Heiden Serge, Magué Jean-Philippe, Pincemin Bénédicte (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010) (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma.

Holmes Susan (1999). Phylogenies: an overview. In Halloran Elizabeth, Geisser Seymour (Eds.). Statistics and Genetics. New York : Springer-Verlag.

Kracauer Siegfried (1981). *Le roman policier : un traité philosophique*. Paris : Payot.

Labbé Cyril, Labbé Dominique (2007). Baudelaire, Rimbaud et Verlaine. *Communication aux VIIIe Journées de l'ERLA*. Brest : 16-17 novembre 2007. Publié dans Banks David (Ed). *Aspects linguistiques du texte poétique*. Paris, l'Harmattan, 2011, p. 17-45

Labbé Cyril, Labbé Dominique (2009). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. *Communication aux Xe Journées de l'ERLA*. Brest : 20-21 novembre 2009. Publié dans : Banks David (Ed). *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, 2013, p. 53-85.

Labbé Cyril, Labbé Dominique (2011). La classification des textes. Comment trouver le meilleur classement possible au sein d'une collection de textes ? Images des mathématiques. *La recherche mathématique en mots et en images*. (http://images.math.cnrs.fr/La-classification-des-textes.html). 28 mars 2011.

Labbé Cyril, Labbé Dominique (2018). Les phrases de Marcel Proust. In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 2018, p. 400-410.

Labbé Dominique (1990). Normes de saisie et de dépouillement des textes politiques. Grenoble : *Cahiers du CERAT*.

Labbé Dominique (2011). Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). Conférence invitée au séminaire *L'œuvre et son auteur : problèmes d'attribution*. Lille : Université de Lille-Nord de la France, 21 mai 2014.

Lacassin Françis (1993). *Mythologie du roman policier*. Paris : Christian Bourgois.

Luong Xuan (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat ès sciences. Paris : Université de Paris V.

Mesplède Claude (Dir.) (2003). *Dictionnaire des littératures policières*. Nantes : Joseph K.

Mesplède Claude et Schleret Jean-Jacques (1996). *Les auteurs de la série noire (1945-1995)*. Nantes : Joseph K.

Monière Denis, Labbé Dominique (2019). Analyse comparée du discours gouvernemental au Canada et au Québec. *Document numérique*. Volume 22, 1/2019, p. 85-105

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

Muller Charles (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Pibarot André, Picard Jacques & Labbé Dominique (1995). Un outil de statistique textuelle : le lemmatiseur. *Travaux scientifique du Service de Santé des Armées*. XVI, p. 305-307.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ragonneau Nicolas (2021). *Le Proustographe*. Paris : Denoël.

Rivière François (1999). *Frédéric Dard ou la vie privée de San-Antonio*. Paris : Fleuve noir.

Savoy Jacques (2020). *Machine Learning Methods for Stylometry*. Cham : Springer.

Savoy Jacques (2018). Elena Ferrante Unmasked. In Tuzzi Arjuna et Cortelazzo Michele (eds). *Drawing Elena Ferrante's Profile*. Padova University Press, p. 123-141.

Tulard Jean (2005). *Dictionnaire du roman policier*. Paris : Fayard.

Urien Frédéric, Labbé Cyril, Labbé Dominique (2020). La bibliothèque en ligne du français moderne. Le vocabulaire de V. Hugo. Semaine Data-SHS. *Traiter et analyser des données en sciences humaines et sociales*. Plateforme Universitaire de Données (TIR-PROGEDO). Université de Grenoble-Alpes. 7-11 décembre 2020.

Sneath Peter & Sokal Robert (1973). *Numerical Taxonomy*. San Francisco: Freeman, 1973.

Annexe 1
Les deux corpus

Romans policiers

Auteurs	Titre	Date	Mots	Vocables
Audiard Michel	Priez pour elle	1950	56 339	6 424
	Massacre en dentelles	1952	61 215	6 578
			117 554	9 569
Benacquista Tonino	La Commedia des ratés	1991	60 219	5 285
	Malavita	2004	85 545	7 737
	La Maldonne des sleepings	1989	63 125	5 185
			208 889	11 339
Boileau Pierre et	Celle qui n'était plus (les Diaboliques)	1952	46 766	4 467
Narcejac Thomas	Les Pistolets de Sans-Atout	1973	36 845	3 804
	D'Entre les morts (Sueurs froides)	1954	49 500	4 507
			133 111	7 467
Bussi Michel (1965-)	Un Avion sans elle	2012	133 379	8 122
	Le Temps est assassin	2017	142 570	8 547
			275 949	11 837
Charles Frédéric	La Grande friture	1953	33 736	4 026
	La Main morte	1953	29 031	3 447
	Vengeance !	1953	29 334	3 698
			92 101	6 810
Dard Frédéric	Quand la mort vient	1954	25 973	2 971
San Antonio	Réglez lui son compte	1949	67 605	6 235
	Bas les pattes	1954	37 976	4 630
	A tue et à toi	1956	38 261	4 797
	L'Archipel des Malotrus	1967	54 300	7 664
Kaput	La Foire aux asticots (Un tueur 1)	1955	42 161	4 942
	La Dragée haute (Un tueur 2)	1955	38 981	4 696
	Pas tant de salades (Un tueur 3)	1956	34 616	4 485
	Mise à mort (Un tueur 4)	1956	37 563	4 659
			351 463	15 698
Djian Philippe	Ca, c'est un baiser	1999	113 533	7 062
Giebel Karine	Les Morsures de l'ombre	2007	64 370	5 046
	Jusqu'à ce que la mort nous unisse	2009	140 504	7 017
	Terminus Elicius	2004	59 324	4 305
			264 198	9 445
Grangé Jean-Christophe	Kaïken	2012	123 464	9 702
	L'Empire des loups	2002	123 439	9 458
	Les Rivières pourpres	1998	107 863	7 694
			354 766	15 068
Héléna André	Les Salauds ont la vie dure	1948	129 481	6 654
	Le Festival des macchabées	1951	110 139	6 273
	Le Demi-sel	1952	421 64	3 766
			281 784	9 464
Lemaître Pierre	Alex	2011	103 749	6 652
	Rosy et John	2012	34 012	3 933
	Sacrifices	2012	94 142	6 536
	Travail soigné	2010	104 725	6 936
			336 628	12 003
Malet Léo	120 rue de la gare	1943	58 198	5 763
	Nestor Burma contre CQFD	1945	57 263	5 580
	Brouillard au Pont de Tolbiac	1956	50 384	5 327
	Du Rébecca rue des Rosiers	1958	56 943	5 481
			222 788	11 335
Manchette Jean-Patrick	Nada	1975	42 304	4 802
Musso Guillaume	La Fille de Brooklyn		100 629	8 714
	La Fille de Paris		95 496	8 989

			196 125	12 649
Pennac Daniel (1944-)	Au Bonheur des ogres	1985	62 498	6 420
	La Fée Carabine	1987	73 371	6 728
	Aux Fruits de la passion	1999	40 770	4 791
	La Petite marchande de prose	1990	93 078	7 876
			269 717	13 619
Simenon Georges	Les Dossiers de l'Agence "O" (nouvelles)	1943	162 616	7 059
	Maigret et l'homme au banc	1953	44 270	3 180
	Le Charretier de La Providence	1931	35 734	3 506
	Le Petit Docteur (nouvelles)	1943	143 514	6 791
	Pietr-le-Letton	1931	40 638	4 366
	Maigret et le voleur paresseux	1961	37 888	3 385
	Maigret a peur	1953	43 265	3 318
			507 925	11 985
Simonin Albert	Touchez pas au grisbi !	1953	65 758	5 827
	Le Cave se rebiffe	1954	58 513	5 360
	Grisbi or not grisbi	1955	57 091	5 388
			181 362	9 500
Thilliez Franck	AtomKa	2012	152 137	8 322
	La Chambre des morts	2006	82 243	7 902
	L'Anneau de Moebius	2008	129 884	7 500
	Pandemia	2015	159 039	8 170
			523 303	15 220
Vargas Fred	L'Armée furieuse	2011	111 595	7 035
	L'Homme aux cercles bleus	1991	66767	4 771
	Sans feu ni lieu	1997	71 731	4 892
	Les Jeux de l'amour et de la mort	1988	54 940	4 105
	Pars vite et reviens tard	2001	98 147	6 820
			403 180	12 563
Vian Boris	J'irai cracher sur vos tombes	1946	35 904	2 977
	Les Morts ont tous la même peau	1947	32733	2 779
	On tuera tous les affreux	1948	51 067	4 125
			119704	5 957
Total	20 auteurs et 68 livres		5 022 345	43 742

Corpus romans de littérature générale 1920-2020

Auteur	Titre	Date	Longueur (mots)	Vocabulaire
Ajar Emile (Gary Romain)	Gros Câlin	1974	26 380	2 903
	La Vie devant soi	1975	37 635	2 368
	Pseudo	1976	44 735	4 137
	L'Angoisse du roi Salomon	1979	81 663	4 515
	Total Ajar		190 413	7 845
Bazin Hervé	Vipère au poing	1948	57 810	6 891
	La mort du petit cheval	1950	72 892	6 891
	MadameEx	1974	86 019	7 755
	Total Bazin		216 721	13 180
Bernanos Georges	Sous le soleil de Satan	1926	94 727	6 381
	L'Imposture	1927	87 667	6 393
	Journal d'un curé de Campagne	1936	98 504	6 234
	Total Bernanos		280 898	10 428
Camus Albert	L'étranger	1942	35 441	2 835
	La Peste	1947	89 182	5 712
	La Chute	1956	31 267	3 638
	Total Camus		155 890	7 466
Céline Louis Ferdinand	Voyage au bout de la nuit	1932	196 145	9 829
	Mort à crédit	1936	234 665	12 575

	Total Céline		430 810	15 945
Dard Frédéric	Batailles sur la route	1949	40 871	4 752
	Anna Soleil	1954	36 377	4 343
			77 248	6 713
Gary Romain	Racines du ciel	1956	178 901	8 460
	Homme à la colombe	1958	28 129	3 429
	Promesse de l'aube	1960	31 109	4 147
	Chien blanc	1970	63 547	6 255
	Au-delà de cette limite	1974	58 876	5 242
	Têtes de Stéphanie	1974	101 249	7 495
	Charge d'âme	1977	76 666	6 631
	Clair de femme	1977	33 788	3 480
	Cerfs-Volants	1980	103 227	6 848
	Total Gary		675 492	18 237
Gide André	Les Faux-monnayeurs	1925	121 356	6 447
Giono Jean	Un Roi sans divertissement	1947	62 954	4 817
	Le Hussard sur le toit	1951	143 527	7 736
	L'Iris de Suse	1970	70 080	6 053
	Total Giono		276 561	11 116
Giraudoux Jean	Adorable Clio	1920	38 401	4 926
	La Disparition de Jérôme Bardini	1930	12 713	2 395
	Siegfried et le Limousin	1922	68 576	7 948
	Total Giraudoux		119 690	10 284
Houellebecq Michel	Extension du domaine de la lutte	1994	38 815	4 996
	Les Particules élémentaires	1998	99 751	8 678
	Plateforme	2001	96 465	7 993
	Possibilité d'une île	2005	127 494	9 351
	Total Houellebecq		362 525	15 958
Le Clézio Jean-Marie Gustave	Procès verbal	1963	78 773	6 714
	Fièvre	1965	85 891	6 441
	Mydriase	1973	8 586	1 441
	Voyages	1975	98 852	5 360
	Mondo et autres histoires	1978	85 938	3 900
	Désert	1980	132 577	5 164
	Ronde	1982	72 500	4 106
	Rodrigues	1986	33 493	3 611
	Printemps	1989	60 340	4 153
	Pawana	1992	9 386	1 393
	Quarantaine	1995	144 474	7 275
	Hasard	1999	59 595	5 049
	Total Le Clézio		870 405	17 993
Lemaître Pierre	Au revoir là-haut	2013	150 927	8 681
	Les couleurs de l'incendie	2018	130 983	8 669
	Total Lemaître		281 910	12 087
Malraux André	Les Conquérants	1928	63 526	5 292
	La voie royale	1930	49 830	4 366
	La condition humaine	1933	95 062	5 929
	L'espoir	1937	153 100	7 590
	Total Malraux		361 518	11 580
Martin du Gard Roger	Le Cahier gris	1922	34 653	4 076
	Le Pénitentier	1922	52 584	4 719
	La Belle saison	1923	88 639	6 835
	Les Thibault (3 tomes)		175 876	9 194
Mauriac François	Le désert de l'amour	1926	50 593	4 500
	Thérèse Desqueyroux	1927	35 683	3 688
	Le Nœud de vipères	1932	60 027	4 877
	Total Mauriac		146 303	7 512
Modiano Patrick	Villa triste	1975	44 808	4 579
	Rue des Boutiques obscures	1978	47 303	3 998
	Total Modiano		92 111	6 240
Pennac Daniel	Comme un roman	1992	27 761	3 982

	Chagrin d'école	2007	58 945	6 115
	Mon frère	2018	21 044	3 242
	Total Pennac		107 750	8 651
Queneau Raymond	Exercices de style	1947	12 816	2 222
Radiguet Raymond	Le Diable au corps	1923	32 687	3 310
	Le Bal du comte d'Orgel	1924	34 157	3 428
	Total Radiguet		66 844	4 863
Saint-Exupéry Antoine de	Courrier sud	1928	26 527	3 258
	Terre des hommes	1939	46 608	4 465
	Vol de nuit	1931	19 258	2 428
	Total Saint-Ex		92 393	6 009
Vian Boris	L'Ecume des jours	1947	54 013	4 850
	L'Automne à Pékin	1947	72 071	5 454
	Total Vian		126 084	7 531
Yourcenar Marguerite	Les mémoires d'Hadrien	1951	93 437	8 081
	L'œuvre au noir	1968	110 072	9 482
	Total Yourcenar		203 509	12 713
22 auteurs	76 œuvres		5 445 036	49 792

Annexe 2. Les substantifs les plus caractéristiques des romans policiers (rangés par fréquence décroissante) comparés avec leur emploi dans la littérature générale contemporaine*.

Polars Rang	Vocable	Fréquence	Littérature Rang	Fréquence
22	type	0,620	200	0,146
27	flic	0,530	1144	0,029
30	voiture	0,500	120	0,216
47	minute	0,390	97	0,260
51	police	0,390	323	0,100
52	commissaire	0,380	899	0,038
76	sang	0,313	142	0,200
78	verre	0,307	206	0,144
79	peur	0,302	55	0,373
80	mètre	0,295	169	0,167
92	café	0,272	202	0,145
110	enquête	0,241	1331	0,024
114	hôtel	0,236	176	0,164
119	gars	0,222	751	0,045
120	victime	0,220	645	0 ,053
157	arme	0,181	256	0,122
161	cigarette	0,176	231	0,131
164	inspecteur	0,175	1566	0,020
166	cadavre	0,173	1170	0,070
172	policier	0,169	1102	0,030
175	tueur	0,167	3679	0,006
193	meurtre	0,155	1527	0,020
203	assassin	0,151	1578	0,020
218	bar	0,141	522	0,065
220	merde	0,137	468	0,072
229	balle	0,133	413	0,079
233	juge	0,132	721	0,047
260	poing	0,122	388	0,084
262	copain	0,121	669	0,051
271	kilomètre	0,117	410	0,080
320	revolver	0,098	838	0,041
333	lieutenant	0,095	997	0,034
338	commandant	0,094	1120	0,030
344	empreinte	0,093	2492	0,011
379	gendarme	0,083	1033	0,032
382	bande	0,083	574	0,060
402	putain	0,080	1617	0,019
404	couteau	0,080	668	0,051
442	môme	0,072	965	0,035
447	prison	0,072	556	0,062
466	restaurant	0,070	698	0,048
485	alcool	0,067	1065	0,031
593	lame	0,060	1091	0,031
605	poupée	0,056	2245	0,012

649	pétard	0,052	2778	0,009
699	taule	0,049	-	0,001
709	indice	0,049	2778	0,009
714	uniforme	0,048	1826	0,004
734	interrogatoire	0,047	2632	0,010
740	flingue	0,047	-	0,001
760	bordel	0,046	1782	0,017
827	souris	0,042	1313	0,025
856	pistolet	0,041	1091	0,031

* Pour tous ces vocables, l'indice de caractéristique est ++ (moins de une chance sur mille de se tromper en affirmant que ce vocable est très sur-employé dans les polars par rapport aux romans contemporains).