LRFAF : une exploration numérique du rap français depuis les années 1990

Benoît de Courson *

2024-02-05

Le temps passe et passe et passe, et beaucoup de choses ont changé
Qui aurait pu s'imaginer que le temps serait si vite écoulé ?
(On fait l'bilan)
Calmement
En s'remémorant chaque instant

(Nèg' Marrons, 2000)

Abstract

Cet article introduit le grand corpus ouvert LRFAF, qui contient 37307 textes de rap français et leurs métadonnées. En utilisant des méthodes de lexicométrie et de traitement automatique des langues, il se propose une première exploration numérique du corpus. Il combine (i) une analyse en diachronie courte, à la recherche des tendances lourdes connues par le rap français pendant ces trente dernières années, et (ii) une étude transversale, qui compare les rappeurs, à la recherche des lignes de force qui décrivent leur style. L'analyse diachronique révèle deux phases. Les textes deviennent plus sombres, plus agressifs et plus vulgaires entre 1990 et 2014, dans une sorte d'élargissement de la fenêtre d'Overton, avant de devenir plus légers, polis et moins politiques depuis, dans un mouvement de « variétisation ». L'analyse transversale aboutit à une analyse factorielle interprétable, qui propose de résumer le profil lexical des rappeurs par deux axes, que nous avons nommés littérarité et hardcorité.

1 Introduction

« Ma qualité de lyrics est quantifiable », affirmait le rappeur Médine dans *Bangerang* (Médine 2018). Simple provocation peut-être, et il convient de toute façon de se garder de tout jugement de valeur en sciences sociales - en particulier lorsque l'on touche aux goûts et aux couleurs. Cependant, les lyrics - le mot étant lexicalisé dans le milieu du rap, on se permettra cet anglicisme - constituent des donneés textuelles. A ce titre, elles peuvent se prêter à la lexicométrie de la même façon que des corpus de presse ou de littérature.

Il n'existe à ce jour, et à notre connaissance, pas de large corpus ouvert de rap français. Depuis 2009 - avant la création de Genius -, des chercheurs de l'Université Masaryk de Brno mènent le projet RapCor et rassemblent des textes de rap français (Podhorná-Polická 2020). Mais le corpus n'est ni ouvert [1], ni massif [2]. En 2024, il contient 1300 titres, soit environ l'équivalent de la production totale du seul rappeur JuL. Ce choix de privilégier la qualité à la quantité convient à l'ambition sociolinguistique du corpus et à l'étude des néologismes, mais pas à des études statistiques. Dans le champ universitaire, les premières études quantitatives ont, de même, reposé sur des petites bases de données, courageusement renseignées à la main (Hammou 2008 ; Hammou 2009).

Pendant ce temps, des petites mains transcrivent massivement les textes de leurs idoles sur le site collaboratif Genius. Les erreurs des uns sont corrigées par les autres, et le résultat s'avère, au bout de 15 ans, très satisfaisant. Bien qu'étant une

*Max Planck Institute for the Study of Crime, Security and Law, Freiburg im Breisgau, Germany

1. Le corpus est accessible sur demande pour les chercheurs mais, par crainte du flou juridique, n'a pas été rendu disponible directement.
2. Le projet RapCor a par contre réalisé un archivage assez massif, et contient les scans de plus de 1500 de disques de rap.

entreprise à but lucratif revendiquant une propriété intellectuelle sur les textes, Genius offre un accès aisé à ses données à travers son API à qui sait programmer. Celles-ci ont été utilisées dans de nombreux projets hors de France (par exemple : Meinecke, Hakimi, and Jänicke 2021; Baltazar and Västfjäll 2020; Kryva and Dilai 2019). Concernant le rap français, quelques chercheurs se constituent manifestement des corpus massifs Klimentová (2022), tout comme les blogueurs spécialisés (voir par exemple l'ambitieuse infographie de RapMinerz (2023)). Mais ces corpus dorment dans les disques durs de leurs collectionneurs, et, on l'espère, sont passés sous le manteau à l'occasion. En cause, probablement, la crainte d'une violation du droit d'auteur.

En juin 2023, Genius a pourtant perdu son procès contre Google, qu'il accusait de piller ses données pour les afficher sur son moteur de recherche. La Cour suprême des Etats-Unis a conclu que ces données appartenaient aux artistes eux-mêmes et non à Genius. D'un point de vue éthique, ces données sont produites par des petites mains bénévoles. Il semble donc naturel de les rendre réutilisables pour un usage non commercial - comme le sont les données de Wikipédia, produites de la même façon. Pour le dire de façon très caricaturale, ces données viennent doublement du peuple : des rappeurs d'une part, qui ont « pour pierre angulaire la pratique d'un art populaire » (Médine 2012), et d'autre part d'une multitude de passionnés, qui ont offert leur temps libre et leur oreille affutée pour transcrire ces textes. Il nous a donc semblé naturel de les « rendre au peuple » en publiant ce corpus. Parce que l'auteur trouve la formule de Médine (2022) [3] désarmante, nous l'avons baptisé « LRFAF », pour « Le rap français aux Français » (ou aux francophones, si l'on veut sacrifier la rime pour plus d'inclusivité). L'auteur en profite pour s'excuser d'avance pour la monochromie de ses exemples, où l'on devinera ses tropismes musicaux.

Concrètement, nous mettons à disposition trois sources et outils :

— Le corpus à proprement parler, soit 37307 textes en format csv, avec pour métadonnées l'artiste, l'album (à ajouter), l'URL Genius et deux mesures de la popularité du titre : le nombre de vues de la page (renseigné à partir de 5 000) et le nombre de contributeurs. A l'aide de l'outil Bunka, les titres ont enfin été classés en 6 grands *topics*, ce sur quoi nous reviendrons. Il est disponible à cette addresse : https://huggingface.co/datasets/regicid/LRFAF, et téléchargeable directement à cette adresse : https://huggingface.co/datasets/regicid/LRFAF/resolve/main/corpus.csv?download=true. A noter que les vues sont certainement une mesure biaisée de la popularité. D'abord, elles mesurent vraisemblablement mieux le succès d'estime que le succès auprès du grand public - en témoigne les scores très élevés de Freeze Corleone. Ensuite, Genius est un site de consultation des paroles, et d'explication. Les textes les plus hérmétiques ou les moins audibles ont plus de chances de mener l'auditeur à consulter la page. Médine (2012) déclarait ainsi rapper pour « [donner du taf au webmaster du site Rap Genius](https ://genius.com/859813/Medine-biopic/Jcontinuerai-ddonner-du-taf-au-webmaster-du-site-rap-genius) », et dans une pastille vidéo pour expliquer sa punchline, confiait mesurer la qualité de ses textes au nombre d'annotations sur le site, signe d'un texte doté de multiples niveaux de sens.
— Une batterie de mesures par chanson - que nous détaillerons dans cet article, pour chaque titre. En les agrégeant, on peut calculer ces mesures par rappeur ou par année. C'est ce que fait cet article, écrit en R Markdown. Le code qui produit les analyses entre les lignes est disponible ici : https://huggingface.co/datasets/regicid/LRFAF/blob/main/article/rap.Rmd. Pour explorer les mesures par rappeurs, nous avons aussi codé un explorateur des donneés à cette adresse : https://shiny.ens-paris-saclay.fr/app/rappeurs_explorateur.
— Les fréquences annuelles d'occurrences de chaque mot et groupe de mots (jusqu'à trois mots) sont explorables sur l'application interactive Gallicagram, à cette adresse : https://shiny.ens-paris-saclay.fr/app/rap, en sélectionnant le corpus « Rap français ». En cliquant sur les points du graphique, l'application affiche les titres dont sont issues les occurrences. Pour plus de détails sur le projet, voir Azoulay and Courson (2021) et Courson et al. (2023).

Cet article se propose un premier défrichage du corpus, à une échelle macroscopique. Il insistera sur l'approche diachronique, et sur les tendances lourdes d'évolution du rap français depuis le milieu des années 1990. Il se place dans une approche ouvertement *data-driven* : plutôt que de chercher à répondre à une question préexistante à l'étude, l'auteur a farfouillé dans les données et cherché ce que l'on peut y mesurer, dans l'espoir d'obtenir des résultats éclairants. De fait, la moisson fut bonne.

2 La collecte des données : croiser Wikipedia, Wikidata et Genius

Décrivons d'abord brièvement le processus de constitution du corpus. Genius n'a malheureusement pas d'arborescence, qui aurait pu nous permettre de recenser en amont les rappeurs français. Ce qui s'en approche le plus est la page des comptes vérifiés de rappeurs français(https://genius.com/Genius-france-liste-des-comptes-verifies-de-rappeurs-francophones-lyrics),

3. À la veille du second tour de l'élection présidentielle française de 2022, Médine a publié un titre intitulé « La France au rap français », retournant le slogan d'extrême droite « La France aux Français »

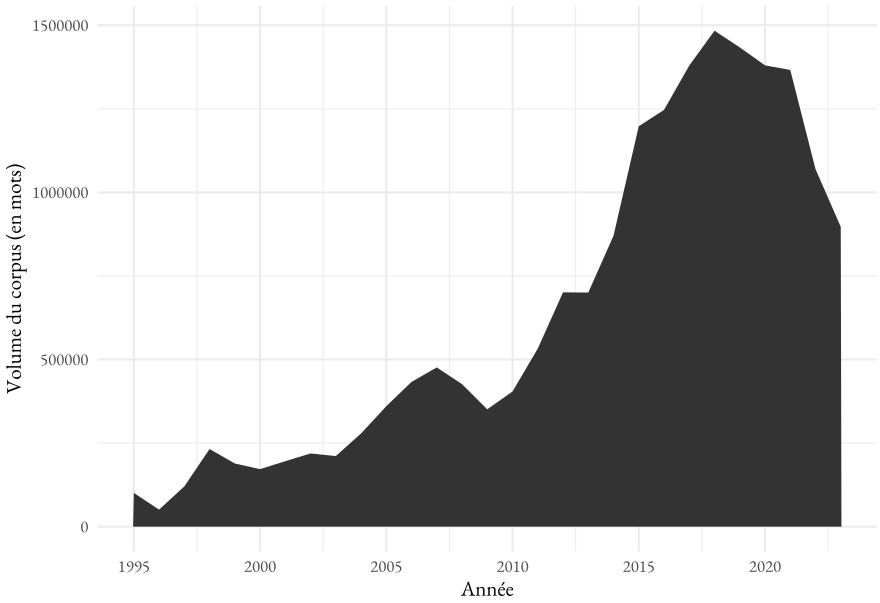

Figure 1 – Volume de données par an, entre 1990 et 2023

qui recense ceux qui sont actifs sur Genius. Cette page n'est pas exhaustive (il y manque par example IAM et PNL). Elle contient aussi les rappeurs belges Damso, Hamza, Roméo Elvis, Scylla et Caballero que nous avons gardé de bon cœur. Nous l'avons donc croisé avec les catégories Wikipedia suivantes : « rappeuse française », « rappeur français » et « groupe français de hip hop » (le rap étant ici subsumé dans le hip hop). Nous avons aussi employé l'outil query de Wikidata, pour lui demander les mêmes catégories. Après quelques corrections et ajouts manuels, par exemple pour préciser qu'on s'intéresse au groupe français Assassin et non à l'un de ses multiples homonymes, on arrive à 822 rappeurs. Certains sont aux frontières du rap, voire au-delà de cette frontière. C'est le cas du groupe Thérapie Taxi, classé en « groupe de hip-hop » par Wikipédia. Nous avons choisi de garder, puisque ses résultats extrêmes sur nos métriques permettent d'entrevoir le contraste entre le rap et les autres genres lusicaux.

Nous avons ensuite injecté cette liste dans le formidable package python LyricsGenius, codé par John W. Miller. Après avoir filtré les textes en langue étrangère, il nous reste 37307 textes, issus de 611 artistes. Certains rappeurs répertoriés sur Wikipedia n'ont pas de textes dans Genius, d'autres sont manqués par le moteur de recherche Genius. A l'œil nu, nous n'avons remarqué l'absence d'aucun nom majeur du rap français, à part le groupe ATK, présent sur Genius mais introuvable par l'API. Nous avons aussi remarqué l'absence du titre « Ma Benz », malencontreusement répertorié comme étant en langue anglaise. Ce genre d'erreur est inévitable avec notre approche, et il vaut mieux manquer quelques titres que polluer le corpus avec des textes en langue étrangère.

Notons que la méthode d'extraction serait aisément déclinable à d'autres genres : il existe aussi des catégories de « chanteurs français » sur Wikipedia, et l'on trouve aussi les textes de Georges Brassens sur Genius. Cependant, il nous a semblé qu'il faut savoir arrêter sa gourmandise. Le code qui a servi à l'extraction est bien sûr à la disposition de quiconque aimerait appliquer le procédé sur un autre genre.

2.1 Limites du corpus (à écrire)

-> Un corpus évidemement non exhaustif et non représentatif, en particulier avant 2010

-> Partir de Genius opère fatalement à une « textualisation » du rap, ce qui appauvrit l'analyse (Carinos and Hammou 2017).

3 Analyse

3.1 Que mesurer ?

3.1.1 Mesurer la fréquence de lexiques

Nous voici face à un petit océan textuel, d'un peu plus de 21 millions de mots. Une première idée vient à l'esprit : y compter les occurrences de n'importe quel mot. « Statistique(s) » y apparaît ainsi 131 fois. C'est peu, et l'on peut déplorer un certain dédain des rappeurs pour les approches quantitatives - « J'suis qu'une statistique pour la place Beauvau », rappait Médine (2015) dans Speaker Corner. Mais avouons-le : ce n'est pas très intéressant. Il est déjà plus instructif de chercher d'où proviennent ces occurrences. En quelques lignes de code, on apprend que 9 d'entre elles viennent du duo Bigflo & Oli, suivi par IAM et Sopico avec 4 utilisations chacun. Une série temporelle de l'utilisation par année donne un résultat peu interprétable, que nous n'osons reproduire ici. On touche ici à un premier problème : bien qu'étant aussi large que possible, notre corpus n'est pas assez gros pour ce genre d'approche. On risque à chaque fois de n'avoir qu'une poignée de résultats, et ne pas pouvoir séparer le signal du bruit. A titre de comparaison, le corpus formé par les archives du journal *Le Monde* contient 1,5 milliard de mots au total, soit en moyenne 20 millions de mots par an. C'est environ le volume nécessaire pour obtenir des résultats interprétables en diachronie. Il va nous falloir ruser, et choisir soit des mots fréquents (« je »), soit de larges lexiques récupérés ailleurs, de sorte à mesurer par exemple la fréquence de mots en verlan, ou la fréquence d'adverbes. Cela nous permet aussi d'éviter l'écueil du choix arbitraire du mot étudié, et le risque d'en essayer plusieurs avant de choisir celui qui reflète le mieux les préconceptions du chercheur. Pour cette étude, nous avons choisi les lexiques suivants :

- Les pronoms « je » (et sa variante « j' », particulièrement utilisée dans le rap).
- Un grand lexique de mots en français, aussi exhaustif que possible, pour mesurer la proximité du texte avec un français standard Nous avons choisi Morphalou3, développé par l'Atilf, car sa taille dépasse de loin celui d'autres bases (comme Lexique.org). De l'aveu de ses concepteurs, il recense des mots en français standard. On y trouvera par exemple « mec », mais pas « schneck » ou « pookie ».
- Trois lexiques du verlan [4], de l'argot [5] et de la sexualité [6], récupérés sur le Wiktionnaire.
- Une liste d'insultes trouvée sur Github, initialement développée pour modérer les contenus haineux sur internet, elle-même basée sur le Wiktionnaire [7]. Nous y avons adjoint les mot anglais « fuck », « nigga » et « bitch », très utilisés dans le rap français.
- De larges lexiques de mots associés aux émotions positives et négatives, développés pour l'analyse de sentiments dans le project LIWC (Piolat et al. 2011). Ces lexiques, souvent utilisés en fouille de textes, regroupent des mots sémantiquement liés, annotés par des humains. On trouve par exemple dans le premier lexique les mots « gloire », « honoré » et « gracieux », et dans le second, les mots « flipper », « enculeront » et « plaindra ». Ici, ils ont pour but de mesurer la tonalité, sombre ou joyeuse, des lyrics.

3.1.2 Mesurer la complexité et la diversité du discours

Sans faire appel à des lexiques extérieurs, nous avons aussi cherché à mesurer la complexité et la diversité du discours, à travers deux *proxys*. D'abord, la longueur moyenne des mots employés, souvent utilisée en traitement automatique du langage comme une mesure de la complexité du discours (Biran, Brody, and Elhadad 2011; Gala et al. 2014), car elle semble empiriquement refléter leur complexité conceptuelle perçue par les receveurs (Lewis and Frank 2016; Bonvin and Lambelet 2019; Piantadosi, Tily, and Gibson 2011).

Ensuite, nous avons cherché à mesurer la richesse du vocabulaire employé. Le plus naturel ici est de compter le nombre de mots différents [8]. Le problème immédiat est que cette mesure est fort sensible au volume de textes étudiés : plus un rappeur écrit de textes, plus il découvrira de nouveaux mots. Or, les variations sont considérables, et le sous-corpus formé par les textes de JuL est par exemple 13 fois plus gros que celui de Lino. Si l'on divise par le nombre total de mots, pour calculer le taux de mots uniques dans les lyrics - une mesure appelée le *type-token ratio* - alors le problème s'inverse : plus un rappeur rappe, et

4. https://fr.wiktionary.org/wiki/Catégorie:Verlan
5. https://fr.wiktionary.org/wiki/Annexe:Liste_de_termes_argotiques_en_français
6. https://fr.wiktionary.org/w/index.php?title=Catégorie:Lexique_en_français_de_la_sexualité
7. https://github.com/MauriceButler/badwords
8. Ici, nous avons considéré l'apostrophe comme un espace, afin d'éviter que « j'suis » soit considéré comme un nouveau mot par rapport à « je suis ». Le mot « aujourd'hui » est donc scindé en deux. Les apostrophes au milieu du mot étant rares en français, ce n'est pas un drame. Le même problème se pose pour les élisions comme « p'tit » ou « v'là », qui sont parfois très fréquentes dans les transcriptions, pour rendre compte du rythme. Il semble heureusement que l'usage de ces élisions soit relativement stable dans le temps à partir de 2000, suffisamment pour ne pas craindre que cela biaise nos analyses diachroniques.

plus il a de chances de se répéter, ce qui le pénalise sur cette mesure. Ces problèmes sont bien connus des psycholinguistes (Daller, Hout, and Treffers-Daller 2003; Bonvin and Lambelet 2019). Une rustine couramment utilisée est de ne prendre en compte qu'un nombre fixé de mots par rappeur, ce qui contraint tous les sous-corpus à avoir la même taille. C'est le choix qu'à fait RapMinerz (2023) dans son infographie, en se limitant aux 7000 derniers mots de chaque artiste. Cela pose cependant des choix cornéliens : 7000 mots, c'est fort peu pour une mesure comme le nombre de mots uniques. Ci-dessous, la Figure 2 présente le résultat de cette opération répétée 1000 fois sur Roff et sur Médine, en prenant chaque fois des textes différents.

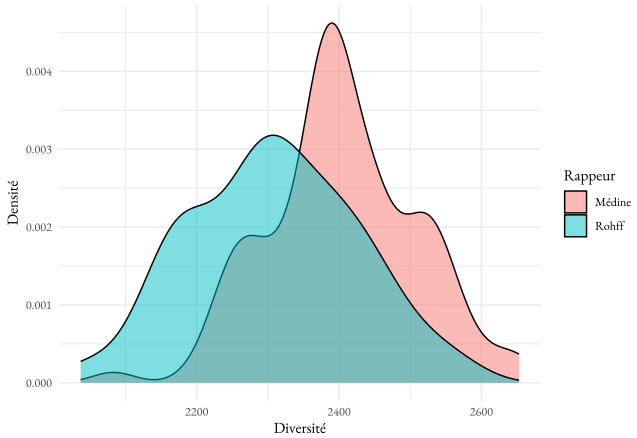

FIGURE 2 – Graphique de densité de l'estimation de la diversité lexicale de Médine et Rohff en prenant 7000 mots issus de textes choisis aléatoirement, et en répétant l'opération 1000 fois

Il est clair qu'en moyenne, Médine a une diversité lexicale supérieure à Rohff, d'environ 10%. Mais l'algorithme arrive au résultat inverse dans 26% des cas. En augmentant fortement le nombre de mots pris en compte (disons 25 000), on tombe tout juste sous les 5% d'erreurs. Mais on est alors contraint de laisser de côté des rappeurs comme Gaël Faye, qui n'a que 20465 mots dans son escarcelle. Se pose en plus question du choix de l'échantillon : si RapMinerz (2023) a choisi de prendre les derniers textes publiés, nous verrons plus tard que la diversité lexicale n'est pas stable au fil de la carrière d'un rappeur. Bref, cette méthode suppose des choix cornéliens, ainsi que de jeter, dans le cas du seuil à 25 000 mots, 64% des mots du corpus aux oubliettes, au prix d'un plus grand bruit statistique.

Revenons au corpus complet, sans échantillonner. De fait, si l'on affiche dans un nuage de points le nombre de mots différents par rappeur selon le nombre de mots qu'ils ont prononcé, (Figure 3A), on observe comme prévu une relation concave : la croissance du vocabulaire décélère à mesure que l'artiste produit, et se répète fatalement. Cependant, si l'on place ces deux variables sur une échelle logarithmique, une relation manifestement linéaire apparaît. Ce n'est pas une vraie surprise : la loi empirique de Heap, connue depuis les années 1960, assure que la diversité vocabulaire employé croît de façon polynomiale avec la taille du corpus en mots, c'est à dire en Kn^{α}, où K est une constante, n le nombre total de mots et α un exposant, communément situé entre 0,4 et 0,6 - ici estimé à 0.7, ce qui témoigne de la grande inventivité lexicale du rap. Cette linéarité sur l'échelle logarithmique est une nouvelle formidable : elle signifie qu'on pourrait en théorie « contrôler » pour le volume du corpus avec une banale régression linéaire. Formellement, on pourrait prendre les « résidus » du modèle linéaire représenté par la ligne bleue de la Figure 3B - la distance verticale à cette ligne - comme mesure de la richesse lexicale. Nous avons aussi effectué cette mesure, cette fois pour tous les rappeurs. La corrélation avec la mesure sur corpus de taille fixe est très forte ($r = 0.87$), mais on présume cette dernière plus précise. On remarque cependant en haut à droite de la Figure 3B un point qui décroche de cette ligne. On devine le rappeur JuL, ses 37 albums et son freestyle de 43 minutes - un temps le record de France - intégralement retranscrit dans Genius. Cet écart à la valeur prédite ne semble pas authentique : JuL se classe 359ème avec la première méthode, et 608 avec la seconde, soit proche de la dernière place. Il semble que la loi de Heap soit violée par le volume considérable de sa production. La loi de Heap est connue pour ne pas fonctionner sur des corpus trop gros : elle prédirait que le nombre de mots unique tende vers l'infini comme tout bon polynôme de degré positif, ce qui est impossible. Il semble que JuL soit pénalisé car il a, à force de rapper, atteint le plafond de tous les mots qu'il pouvait prononcer. Le même phénomène semble affecter, dans une moindre mesure, les autres rappeurs très productifs : Guizmo, la Fouine et les groupes Sexion d'Assaut et surtout IAM, qui chute de la 8ème place à la 95ème . Plus généralement, on devine une légère concavité dans la Figure 3B. Il est possible de la prendre en compte en appliquant une régression non pas linéaire, mais quadratique. Graphiquement, le résultat est presque le même. Mais elle semble mieux rendre compte de la relation entre nombre de mots

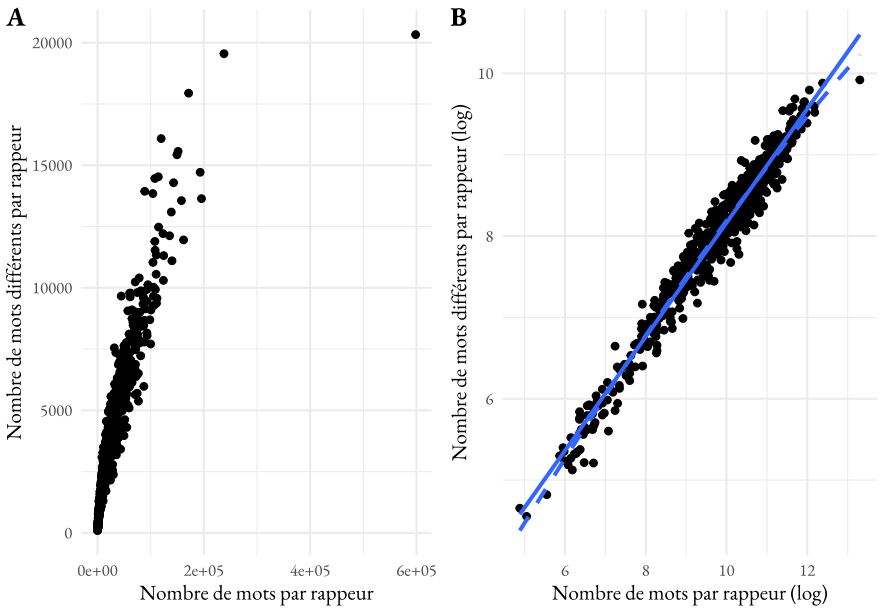

FIGURE 3 – Nombre de mots uniques par rappeur selon le nombre de mots prononcés, en échelle linéaire (A) et logarithmique (B)

et nombre de mots uniques : dans un test du rapport de vraisemblance, le modèle quadratique domine clairement le modèle linéaire ($p < 10^{-4}$). La corrélation avec la première mesure s'en trouve améliorée ($r = 0.89$) et justice est mieux rendue à JuL et IAM, qui remontent respectivement à la 581ème et à la 37ème place. Nous garderons donc cette mesure dans la suite, sous le nom de « diversité lexicale », car elle nous semble la plus raisonnable. Cette mesure n'est cependant pas parfaite, et nous invitons le lecteur à garder en tête le risque de biais pour les rappeurs les plus productifs.

Pour vérifier que nos mesures sont raisonnables, on peut visualiser les résultats par rappeur. La Figure 4 présente le score sur nos deux dimensions des 200 rappeurs les plus vus, avec la taille dépendant du nombre de vues. Ce graphique étant quelque peu surchargé, une version interactive, zoomable et explorable, est disponible en ligne à cette adresse : https://regicid.github.io/complexity_rappers.

La forte corrélation entre nos deux métriques ($r = 0.77$) est fort rassurante : nos deux mesures semblent bien saisir une qualité commune. Qualitativement, il semble que l'on retrouve bien en haut à droite les rappeurs à texte, ou « lyricistes ». On pourrait aussi analyser le déséquilibre entre les deux dimensions : Stupeflip semble employer des mots complexes (les « Argémiones », « Pétaouchnok ») (Stupeflip 2011b) mais relativement répétitifs. Cela renvoie au côté obsessionnel du groupe, résumé dans La Menuiserie : « Prendre des petits bouts de trucs/Prendre, prendre/Prendre des petits bouts de trucs et puis me les assembler ensemble » (Stupeflip 2011a). Inversement, Hugo TSR, Seth Gueko, Népal ou encore Nakk Mendosa utilisent des mots relativement courts, compte tenu de l'étendue de leur vocabulaire. Ce couplet d'@hugotsr2012 illustre sa méthode, en employant des mots relativement courts, mais rares :

> Bicrave d'aya, des cavales, schlags, Boulevard Ney y a plus de putes qu'à Pattaya
> Les petits graffent le reureu ou bicrave le teuteu
> Du mal à faire le mois quand t'as la petite femme et le teum-teum
> Alors dites pas que j'suis pas déter', certes un passé terne
> Mais à cette heure j'veux juste un taf pépère, calculez pas mes cernes

Nous avons aussi mesuré la diversité lexicale par année plutôt que par rappeur. Dans ce cas, nous ne disposons que de 30 points de données, ce qui est insuffisant pour calibrer finement une régression. Mais on dispose aussi de bien davantage de mots par points de données. A partir de 1995, chaque année (sauf 1996) contient plus de 100 000 mots, ce qui, on l'a vu, permet une estimation assez fiable. On utilisera donc la première méthode ici, en prenant 100 000 mots issus de textes aléatoirement choisis dans l'année.

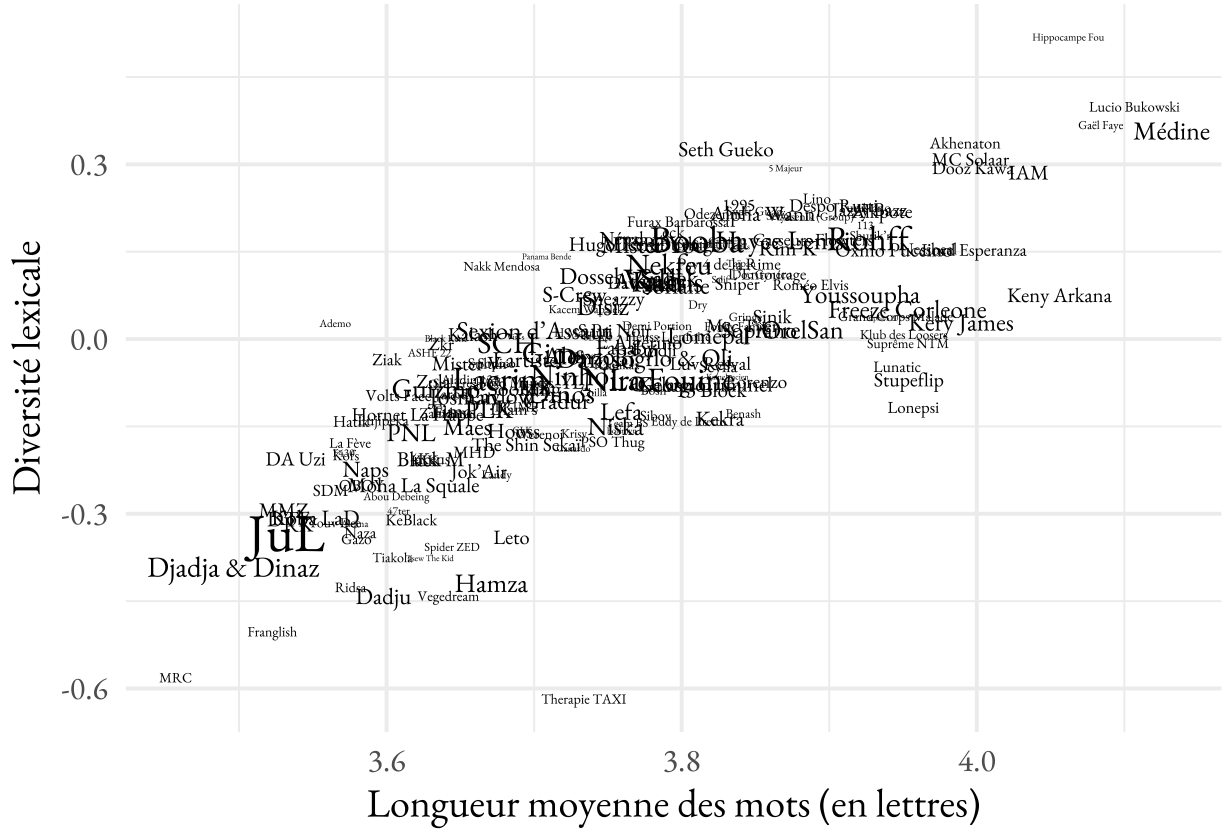

FIGURE 4 – Profil de complexité pour les 200 rappeurs les plus vus. La taille des noms est proportionnelle au nombre de vues des rappeurs sur Genius. NB : ceci n'est pas une mesure de qualité, mais d'orientation artistique.

3.1.3 Reconstituer les genres de rap : le *topic modeling*

Enfin, nous avons appliqué un *topic modeling* au corpus, une méthode de machine learning capable de rassembler les documents en plusieurs *clusters* de par leur ressemblance textuelle. Nous avons d'abord effectué un « plongement lexical » avec le modèle Solon-large, développé par OrdalieTech. Nous avons ensuite appliqué l'outil Bunka, développé par Charles de Dampierre. Celui-ci a l'avantage de fournir une liste de mots qui caractérisent le corpus, ce qui permet de les comprendre. Les *topics* sont affichés dans la Table 1, muni du nom que nous lui avons attribué [9] et d'une chanson emblématique, choisie parmi les 20 textes les plus centraux du corpus« C'est-à-dire les plus proches du « centroïde » du *topic*, au sens de la distance dans l'espace vectoriel des plongements lexicaux. ». On a omis deux *topics* inclassables, représentant moins de 0,5% du corpus.

Table 1 : Topics du corpus

Mots caractérisant le topic	%	Nom (selon l'auteur)	Titre emblématique	Artiste représentatif
beuh, sacoche, binks, billets, poto, monnaie, Audi, moula, zone, condés, euros, pétard, cli, mets, frappe	19.96	Gangsta rap	Ninho : Mama No Cry	Timal
rêves, ciel, espoir, bonheur, vie, larmes, lumière, souvenirs, jour, destin, âme, vide, monde, silence, soleil	16.37	Rap poétique	Grand Corps Malade : Jour de doute	Dooz Kawa
peuple, pays, système, justice, liberté, jeunes, violence, médias, familles, politique, État, droits, racisme, peuples, pauvres	15.47	Rap conscient	Akhenaton et al. : 11'30 contre les lois racistes	Keny Arkana
rap, mic, flow, MC, rimes, beat, style, hip, micro, rime, rappeurs, phases, lyrics, textes, rappeur	14.91	Egotrip/Méta-rap	Nekfeu : Enorme freestyle de Nekfeu en live dans Planète Rap!	Ärsenik
baby, soir, might, chérie, yeah, bébé, nuit, miss, night, soirée, copines, danse, fille, mama, belle	13.43	Rap coquin	Fatal Bazooka : Mauvaise foi nocturne	JuL
amour, cœur, yeux, might, larmes, sentiments, fois, nuit, fille, souvenirs, temps, vie, soir, mal, relation	10.35	Chansons d'amour	La Fouine & Zaho : Ma meilleure	Dadju
négro, négros, ekip, bitch, LDO, Fuck, nigga, pétasse, Dakar, chatte, game, bite, secte, flow, gang	8.92	Rap hardcore/Drill	Freeze Corleone : Ekip	Freeze Corleone

Comme toute méthode automatique, cette classification est faillible. Cependant, elle a la qualité de se faire « par en bas », par les proximités langagières entre les textes, et non par le choix possiblement arbitraire d'un chercheur. Elle semble de plus convaincante : on reconnaît clairement les grands thèmes du rap dans les quatre premières catégories. La dernière (« Rap hardcore/Drill ») fut plus délicate à nommer. Elle est caractérisée par des termes très vulgaires, agressifs et sexistes. On y trouve tout d'abord Freeze Corleone et son groupe 667 (85% de leurs titres), ainsi qu'un bon nombre de chansons d'Ateyaba, Gazo ou Gradur. Le terme de drill renvoie habituellement à un rythme plus qu'à un vocabulaire, mais puisqu'il est accollé à ces quatre rappeurs, on a choisi de l'utiliser ici. On y trouve aussi une cinquantaine de titres de Booba, comme A.C. Milan, Tombé pour elle et Wesh Morray. On n'y trouve par contre peu de titres d'Ärsenik, Médine ou Kery James, parfois associés au rap hardcore – probablement davantage du fait de leur timbre et leur flow que de leur vocabulaire.

A noter également que la catégorie « Egotrip/Méta-rap » n'est pas forcément caractérisée par la mégalomanie, mais plutôt par l'évocation du rap. On y trouvera donc Grand Médine et grand nombre de freestyles, mais aussi des morceaux programmatiques comme *Boxe avec les mots* d'Ärsenik, ou encore les morceaux d'hommage de Médine : *Lecture aléatoire*, *Ali X*, *Bataclan* et *Le jour où j'ai arrêté le rap*.

9. Ces noms sont choisis *a posteriori* par l'auteur, au vu des mots retournés par l'algorithme. Nous avons choisi là où c'était possible de reprendre les « catégories indigènes » (Maillochon 2021) comme le « rap conscient », d'une part parce qu'elles semblaient coller à merveille, et d'autre part parce que nous n'avons pas connaissance de « catégories savantes » en la matière. Le « rap coquin » fait exception, parce qu'il n'existait ici pas de catégorie pertinente.

On peut voir ce traitement comme une façon d'enrichir les métadonnées : dans le corpus que nous mettons en ligne, on a ajouté ce classement thématique, ce que nous espérons utile à de futures recherches.

3.1.4 Stratégie d'analyse

Nos données sont structurées par deux variables principales : le rappeur et l'année de publication. Nous les grouperons donc par ces deux variables, et présenterons nos variables de façon chronologique et de façon transversale. Nous tenterons brièvement de croiser les deux et d'étudier les rappeurs de façon longitudinale, afin d'appréhender l'effet de l'avancement dans sa carrière. Nous aurions aimé prendre en compte l'effet du genre de l'artiste. Malheureusement, les femmes sont trop rares dans notre corpus : parmi les 200 rappeurs les plus « vu(e)s », on ne trouve que deux femmes (Diam's et Keny Arkana). C'est d'ailleurs un résultat en soi : ces deux rappeuses ayant passé leur pic de notoriété depuis plus de 10 ans, le rap de Genius est aujourd'hui quasi exclusivement une « affaire de mec » (Dole and Strausz 2010, 12), du moins derrière le micro (Lesacher 2013).

.

3.2 L'évolution du rap français depuis 1990

Notre corpus commence en 1984 avec Paname City Rappin de Dee Nasty. Cependant, le volume de textes semble insuffisant avant 1990 pour faire une étude statistique. Avant 1995 - année charnière du genre, où Suprême NTM, Assassin, Akhenaton et Les Sages Poètes de la rue publient chacun un album - certaines de nos courbes sont très bruitées, la faute à un corpus rachitique. Pour cette raison, nous nous permettrons d'amputer certains de nos graphes des premières années, où le bruit statistique peut nuire au rendu visuel.

3.2.1 Le rap se personnalise et s'éloigne du Français formel

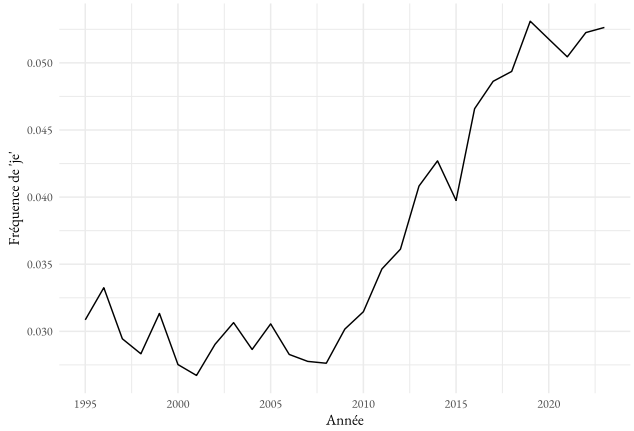

Figure 5 – Usage de la première personne du singulier, 1990-2023

Il ressort de nos données deux tendances continues depuis 1990. D'abord, le rap se personnalise de façon plus explicite [10] : la fréquence de la première personne du masculin, quasi stable entre 1995 et 2010, a depuis doublé (Figure 5). On mesure ici la place considérable de la première personne dans le rap : sa fréquence dans un corpus comme les archives du Monde est inférieure à 0,3%, soit près de 20 fois moins que le rap en 2023. Au sein du rap, on trouve aussi des différences frappantes : on trouve chez IAM seulement 1.8% de "je", contre 6.5% chez PNL. On trouve d'ailleurs une forte relation négative (mesuré par le coefficient de corrélation) entre l'utilisation du « je » et nos deux mesures de complexité du discours : la richesse du vocabulaire employé (r = -0.52) et la longueur moyenne des mots (r = -0.59).

La fréquence de la première personne dépend aussi fortement du *topic*. Le « je » est sous-représenté parmi le rap conscient, mais aussi, de façon contre-intuitive, dans le *topic* « Egotrip-Méta-rap ». De fait, si un titre comme Jimmy Punchline d'Orelsan contient une forte proportion de « je » (6.8%), ce ne semble pas être une condition nécessaire à l'égotrip. Un

10. Consulté, Anthony Pecqueux considère que le « je » est souvent implicite, par exemple dans *Boxe avec les mots*, et que la tendance que nous observons relève davantage d'une explicitation que d'une augmentation de la personnalisation, qui est de toute façon une règle constitutive du genre. Les « je » implicites n'étant pas mesurables, nous suspendons notre jugement.

morceau mégalomane comme Grand Médine contient 4.6% de « je », soit à peine plus que dans le corpus complet (4.2%). Boxe avec les mots d'Ärsenik en contient seulement3.7%, et Qui sème le vent récole le tempo de MC Colaarà peine 1.8%. Il est cependant possible que les morceaux qui soient davantage « méta-rap » qu'égotrip - c'est-à-dire qui parlent de rap sans célébrer le rappeur - tirent la moyenne vers le bas. Lecture aléatoire, un morceau d'hommage aux rappeurs qui ont marqué Médine, en contient la proportion dérisoire de 0.3%).

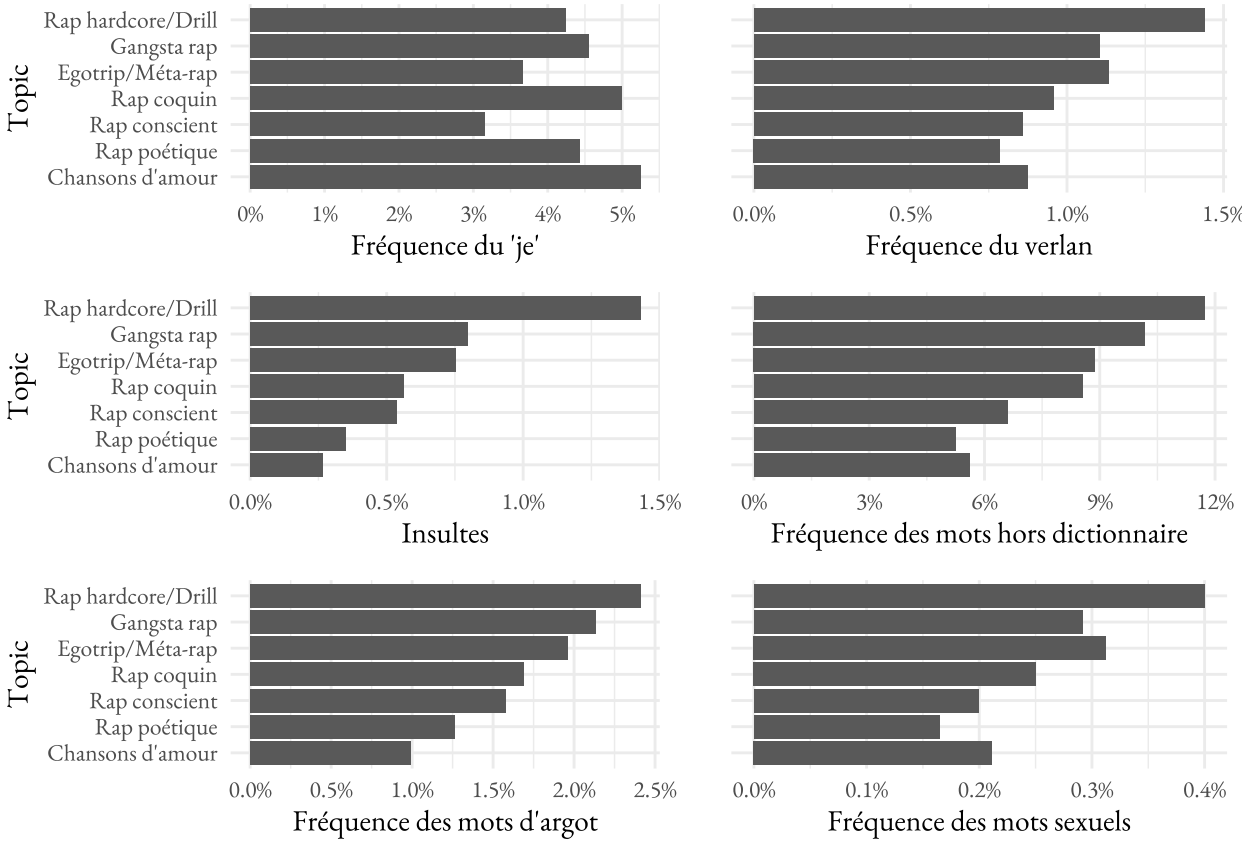

Figure 6 – Fréquence des lexiques dans les différents topics du corpus

Ensuite, les textes semblent s'éloigner du français standard : le pourcentage de mots reconnaissables est en chute continue, de près de 98% en 1990 à 92% aujourd'hui (Figure 8). Cette chute ne semble pas, ou peu, due à l'adoption du verlan : après une apparition fulgurante en 1994 (due à MC Solaar et au Ministère A.M.E.R), son usage est presque stable depuis (Figure 4B). On pourrait plutôt l'attribuer notamment à l'usage des onomatopées, qui explosent (Figure 10), et probablement à l'augmentation du *code-switching* entre français, anglais, arabe ou lingala.

Au sein des *topics*, le rap hardcore/drill et le gangsta rap sont, de façon prévisible, les genres les moins fidèles au dictionnaire (Figure 6). Nos estimations sont probablement affectées par la prégnance des « gimmicks » dans ces deux genres, comme le « izi » de Booba. Le verlan, l'argot et les mots sexuels suivent la même répartition entre les *topics* (Figure 6).

3.2.2 Le rap, la « nouvelle variété » ?

Pour nos autres mesures, une rupture claire se présente au début des années 2010. D'abord dans l'évolution des sous-genres. La Figure 7 présente l'évolution de la part de chaque *topic* depuis 1993. La composition semble assez stable jusqu'en 2010, avec une domination de l'Egotrip/Méta-rap, du rap conscient et du rap poétique - qui contient une bonne partie des textes de IAM, Akhenaton, Oxmo Puccino et Keny Arkana. Vers 2012, le gangsta rap, les chansons d'amour et le rap coquin mordent soudainement sur la part du rap conscient et de l'égotrip, tous deux en voie de disparition en 2023. La disparition du rap conscient est une tendance bien repérée par les rappeurs eux-mêmes. Si en 1998, Calbo pouvait proclamer comme une évidence « Qui peut prétendre faire du rap sans prendre position » (Ärsenik 1998) , Kery James (2012) déplore déjà que « les autres font du rap d'inconscients ». Quelques années plus tard, Youssoupha (2018) considère que la messe est dite : « Et, très

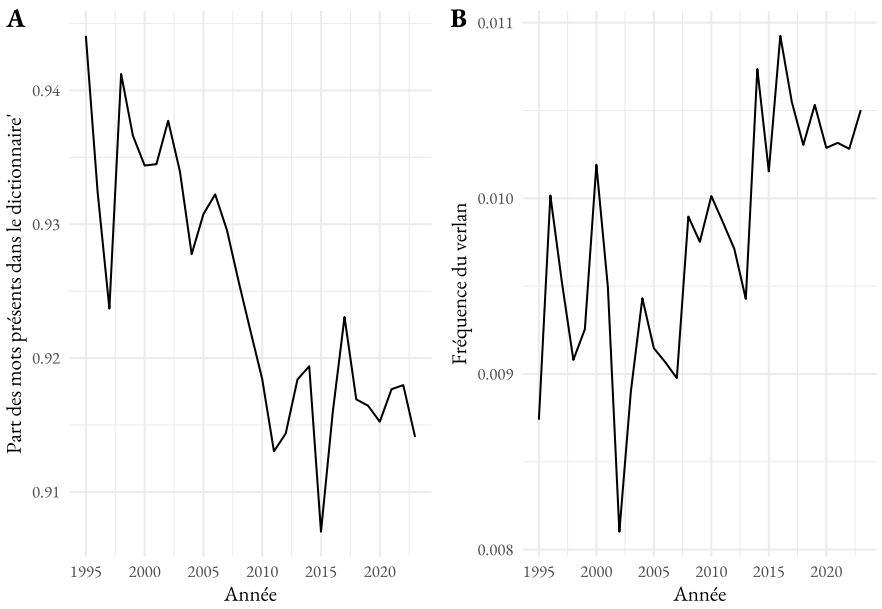

FIGURE 7 – Usage de mots présents dans le dictionnaire et du verlan, 1990-2023

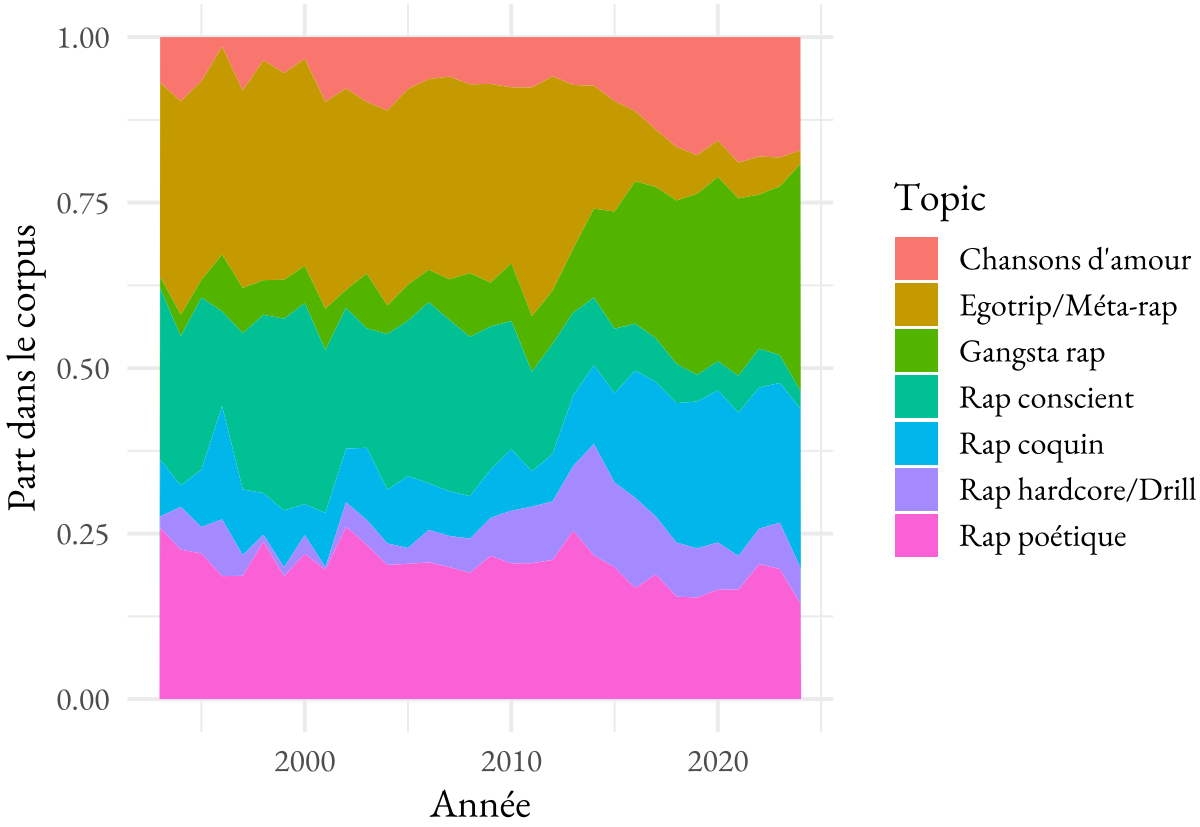

FIGURE 8 – Evolution de la part de chaque topic dans le nombre de titres du corpus, 1993-2023

vite, j'ai senti la douille, j'suis pas complètement inculte/Tout est parti en couilles quand rap conscient est devenu une insulte ». Dans *Le flingue à Renaud*, Lino (2015) suggère lui que cela participe d'une tendance plus générale dans l'industrie musicale : « J'ai retrouvé le flingue à Renaud dans les égouts du top albums/Coincé entre la rage de Trust et la guitare à Brassens/Parce que l'insoumission fait du lap-dance à plein temps ».

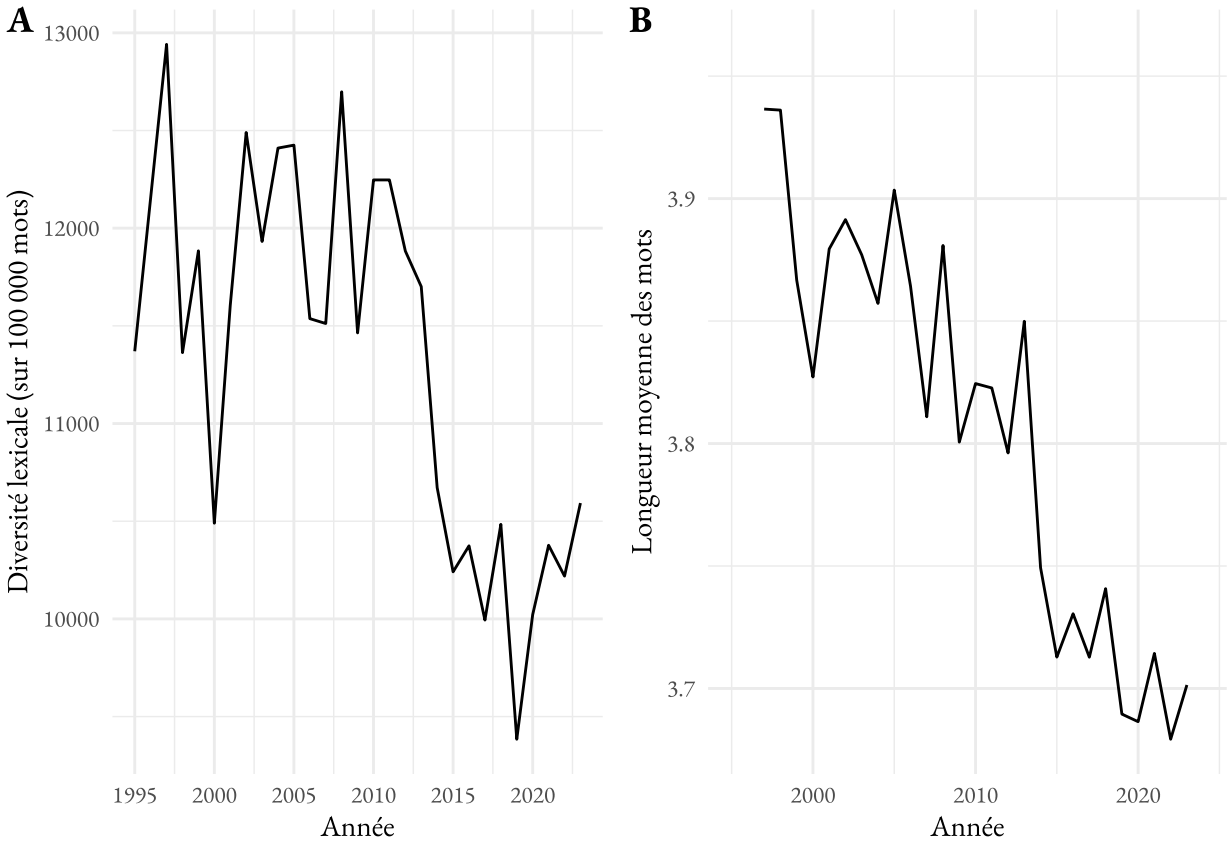

Figure 9 – Mesures de complexité linguistique, 1990-2023

Cette tendance a fait l'objet d'un débat dans le champ journalistique et universitaire. En 2018, la revue Mouvements consacrait une section entière d'un numéro à la question « Le rap doit-il être politique ?» (voir en particulier Jesu (2018)). Plus récemment, Olivier Cachin constatait sur France Inter que la remarque de Kaaris « Marion Maréchal elle est bonne » avait remplacé le morceau engagé *Marine* de Diam's (Cachin 2023). Certains journalistes (Levaché 2023) proposent cependant que le genre n'ait rien perdu en politisation, mais ait seulement changé de moyen d'expression. Notre mesure par topic modeling classe cependant en rap conscient non pas le rap qui se proclame conscient, mais celui qui en a le vocabulaire (peuple, pays, système, liberté, justice..., voir Table 1). Si tel est le cas, la politisation a donc su se faire indétectable. On pourrait enfin imaginer que la politisation se soit faite plus diffuse, que le rap conscient ait disparu en tant que genre à part, mais que les thèmes soient demeurés. En réalité, tous les mots qui caractérisent le *topic* sont en décroissance dans le corpus. La décrue est particulièrement marquée pour le syntagme « Le Pen », ce qui dénote tant sa banalisation que la raréfaction de l'opposition frontale à l'extrême droite dans le rap français.

La disparition de l'égotrip/méta-rap est une tendance moins documentée. Cette moindre évocation du rap dans les textes de rap semble aller de pair avec une moindre attention aux textes, révélée par la chute brutale de nos deux mesures de complexité en 2014 (Figure 9). D'ailleurs, ce graphique bat en brèche l'idée reçue d'un « âge d'or » du rap dans les années 1990, où les textes auraient été plus ciselés que par la suite. D'un point de vue quantitatif, c'est plutôt la période récente qui sort de l'ordinaire [11].

11. Sur la Figure 9, nous avons décidé de tronquer la période 1989-1994. Si on l'inclut, on observe une longueur des mots exceptionnelle, supérieure à 4 lettres. Mais il semble que seuls les rappeurs les plus lettrés aient été conservés dans Genius, ce qui explique ce résultat. En 1991, deux tiers des titres proviennent par exemple de IAM ou de MC Solaar. A partir de 1995 (où l'on trouve 27 artistes différents), il nous a semblé qu'on pouvait faire une étude sur une population, et non une quasi-monographie.

Enfin, le ratio de positivité - rapport entre le nombre de mots positifs et le nombre de mots négatifs - s'envole depuis 2014 et retrouve son niveau du début des années 1990 (Figure 10). La fréquence des insultes - après une hausse continue depuis l'époque des Sages poètes de la rue et du non moins sage MC Solaar - recule depuis 2014, de même que le lexique de la sexualité et l'argot. Le rap semble bien être devenu, comme le proclamait Gims, « la nouvelle variété » : moins sombre, moins agressif, moins tourné vers les textes. En passant, on mesure aussi ici à quel point le rap du début des années 1990 - ou du moins celui qui est parvenu jusqu'à Genius - a pu être sage, avant l'arrivée de groupes comme Ministère A.M.E.R. et Lunatic, dont le titre « Le crime paie » (Lunatic 1996) est souvent associé à la naissance d'un rap plus sombre. On peut supposer que pour se faire une place dans l'industrie musicale, il a d'abord fallu montrer patte blanche (Hammou (2014) parle d'un « espoir de professionnalisation ») et que les textes qui ont survécu soient ceux qui ont su arrondir les angles pour passer à la radio. Le tournant vers un rap plus agressif à partir du milieu des années 1990 est certainement lié au désintérêt des *majors* pour le rap à partir de 1993 et à l'irruption de labels indépendants comme le Secteur Ä (Hammou 2014). On peut aussi présumer qu'il y ait là un phénomène analogue à la fenêtre d'Overton : les discours de Suprême NTM suffisaient à choquer dans les années 1990. Vingt ans plus tard, Kaaris et Gradur ont dû aller bien plus loin pour choquer à nouveau, Lunatic, Ärsenik et Seth Gueko étant notamment passés par là.

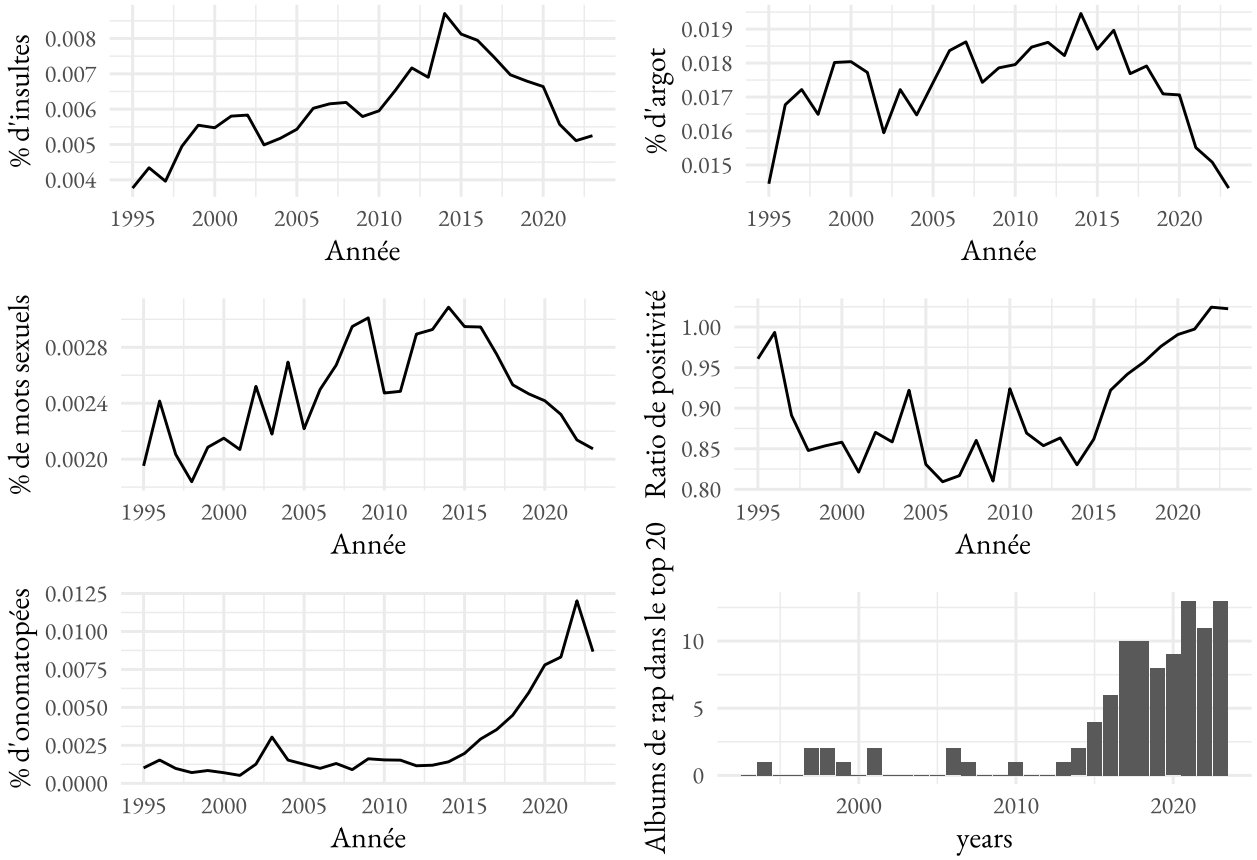

Figure 10 – Le tournant de 2014 : au moment où les albums de rap envahissent le top 20 annuel du SNEP, les insultes, l'argot et les mots sexuels chutent, le ratio entre mots positifs et mots négatifs augmente, tout comme les onomatopées.

Si le rap se mue en « variété », c'est aussi que la frontière avec les autres genre se brouille. Encore une fois, Gims illustre cette évolution. Rappeur à ses débuts dans la Sexion d'Assaut - même aux débuts de sa carrière solo, des morceaux comme *Meurtre par strangulation* appartiennent encore clairement au rap -, il devient exclusivement chanteur au fil de sa carrière. Autrement dit, il se peut que ces tendances soient dues non à l'évolution des rappeurs, mais à la contamination du corpus par des textes plus légers, et dont l'appartenance au rap et donc à ce corpus est devenue discutable.

Certaines branches du rap semblent d'ailleurs refuser ce tournant. Dans la Figure 11, on a affiché l'évolution des mesures

de complexité par *topic*. Si le raccourcissement des mots affecte tous les genres [12], la perte de richesse lexicale est à peine perceptible parmi le rap conscient et l'égotrip/méta-rap [13], deux genres autrefois dominants et manifestement attachés à l'ancienne façon d'écrire. Certains combattent d'ailleurs ouvertement ce tournant du rap : « J'm'étais dit que le rap n'accoucherait pas de chanteur de variét' », peut-on entendre dans *Grand Médine* (Médine 2014). Celui-ci fait d'ailleurs un lien immédiat avec la marchandisation du genre : « Main qui donne est main qui dirige, coupe la de toutes les manières », tout comme Hugo TSR (2012) (« dis qu'tu fais d'la variét' pour tenter les concerts »).

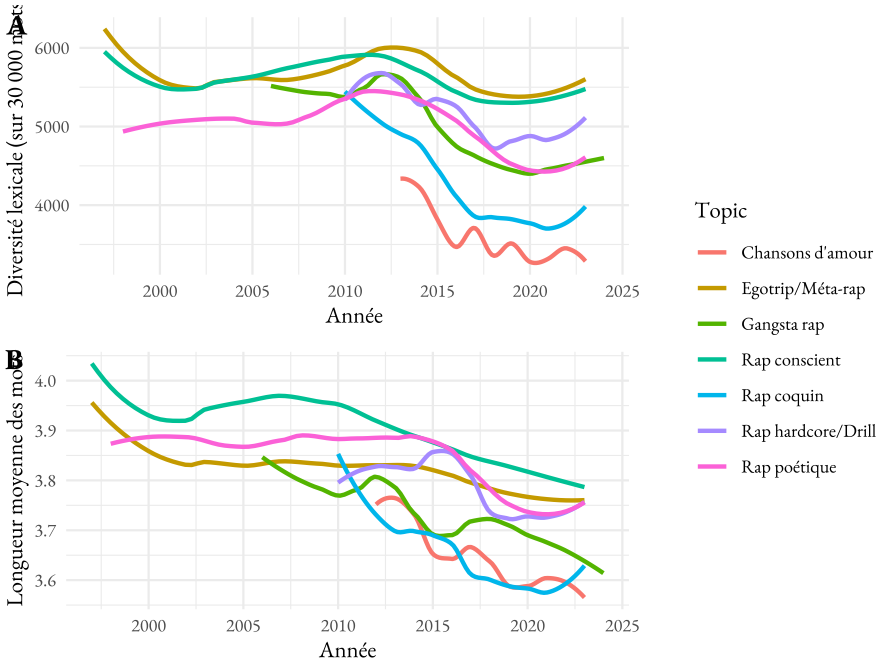

FIGURE 11 – Mesures de complexité par topic et par année, 1995-2023. Pour plus de lisibilité, on a ici appliqué un lissage loess.

Notons enfin que ce tournant est à peu près synchrone avec la prise de pouvoir du rap sur l'industrie musicale. Dans la Figure 10, nous avons aussi dénombré le nombre d'albums de rap parmi le top 20 annuel publié par le SNEP. L'explosion a lieu en 2015, moment où la transition semble s'engager dans les textes. On retrouve cette explosion dans l'enquête INSEE « Pratiques culturelles », qui a mesuré que 30% des amateurs de musique en France écoutaient du rap en 2018, contre 14% en 2008 (Hammou and Molinero 2022). Le rap semble pénétrer l'ensemble des secteurs de la société française : l'augmentation est plus marquée chez les femmes, les plus de 25 ans et les cadres, catégories auparavant rétives (Hammou and Molinero 2022).

On peut supposer soit que le tournant du rap ait permis la conquête du grand public, soit que ce succès ait renforcé les incitations à se faire « variété ».

Concluons. Médine (2008b) posait la question rhétorique « Sais-tu vraiment ce qu'est le rap français ? Pas une machine à sous, mais une machine à penser ». Depuis, le rap français est indéniablement devenu une « machine à sous ». Il est impossible de mesurer s'il a pour autant cessé d'être une « machine à penser », mais il est clair qu'il a en moyenne beaucoup perdu en politisation, en radicalité, en agressivité et en complexité lexicale. Ce faisant, il a aussi considérablement élargi son audience.

3.3 Mesurer les discours haineux et sexistes

Jusqu'ici, nous avons traité nos textes commes des « sac de mots », et avons construit nos mesures en se contentant de compter les mots qui figuraient dans nos lexiques. Cette classe de méthode par fréquence d'occurrences, aujourd'hui assez *old-school* dans le traitement automatique des langues, offre deux avantages. D'abord, un temps de calcul réduit, de l'ordre de la seconde. Ensuite, une grande transparence et interprétabilité : le lecteur pourrait en théorie faire cette mesure lui-même à la main, s'il avait quelques centaines d'heures devant lui. À son passif, on peut reprocher à ces methodes qu'elles sont sujettes au

12. Notons que le raccourcissement des mots peut aussi venir de la prolifération des onomatopées.

13. Ce graphique livre un autre enseignement : le rap conscient emploie des mots plus longs, mais moins divers que l'égotrip/méta-rap. Il semble que le second soit plus proche de l'exercice de style, là où le premier est plus conceptuel.

choix des mots par le chercheur, et donc à son arbitraire. Ensuite, ces méthodes détachent les mots de la la syntaxe. Il est par exemple impossible de détecter avec des fréquences d'occurrences la nuance entre « je te hais » et « je ne te hais point ».

Une approche alternative est d'utiliser les plongements lexicaux évoqués plus haut et d'autres méthodes de *machine learning*. On peut par exemple lire manuellement une poignée de textes, isoler ceux qui remplissent à nos yeux un certain critères (par exemple, les textes politiques), et laisser ensuite la machine généraliser avec des calculs qui nous dépassent (pour plus de détails sur cette classe de méthode, on recommande la lecture de Do, Ollion, and Shen (2022)). Dès lors, les calculs prennent des nuits entières et sont ininterprétables. Mais on peut espérer avoir gagné en profondeur et en fiabilité.

Cette méthode est notamment appliquée pour repérer automatiquement les contenus problématiques sur les réseaux sociaux, et en particulier les contenus haineux ou sexistes [définir haine et sexisme ?]. Il existe des modèles aptes à détecter ces discours, pré-entraînés sur des corpus de posts Twitter (Chiril et al. 2020; Chang, May, and Lerman 2023), média dont la langue orale ne s'éloigne pas tant celle utilisée en rap. Ces algorithmes ont été développés par des experts du champ, et il nous a semblé tentant de les appliquer à nos textes de rap. Nous avons dont appliqué deux modèles, qui détectent respectivement les discours haineux (Chang, May, and Lerman 2023) [14] et sexistes [chiril2020] [15]. Après des jours à faire chauffer notre malheureuse CPU, ces modèles nous ont renvoyé pour chaque ligne une probabilité que cette ligne soit haineuse ou sexiste. Ces modèles permettent en théorie de détecter plus finement l'agressivité et le sexisme que nos lexiques d'insultes et de mots liés à la sexualité. De fait, « les femmes doivent rester à la cuisine » et « espèce d'islamo-gauchiste » sont classés comme respectivement sexiste et haineux, sans pourtant contenir d'insulte ou de mot sexualisé.

Nous n'en étions pas convaincus en nous lançant, mais nos mesures semblent avoir du sens. Dans la Figure 12, nous avons représenté les rappeurs les plus connus dans le plan haine-sexisme. Les deux caractéristiques, issues pourtant de modèles indépendants, sont fortement corrélées (à l'échelle des rappeurs, $r = 0.69$). De plus, nos modèles sont en mesure de repérer qu'un rappeur comme Médine est nettement plus haineux que sexiste, ce qui nous semble très juste : « Pour respecter nos soeurs, on n'attendra pas le 8 mars » clame-t-il dans *Don't Panik* (Médine 2008a), avant de consacrer une chanson à l'excision (Médine 2017). De même pour Freeze Corleone qui, s'il utilise l'insulte sexiste « pétasse » comme gimmick, voue plutôt sa haine aux Etats-Unis et à Israël qu'aux femmes. A l'inverse, Moha la Squale est détecté comme beaucoup plus sexiste que haineux, ce qui fait écho aux terribles violences sexistes dont il est accusé.

C'est donc un pari osé, mais pas dément, d'utiliser cette mesure pour détecter l'évolution du sexisme et de la haine dans nos textes de rap [16]. La Figure 13A présente l'évolution du degré de haine et de sexisme, mesuré par nos modèles. La haine semble suivre la tendance observée jusqu'ici, avec un effondrement depuis 2014. Le sexisme, lui, suit un chemin différent : il connaît un second pic en 2019-2020, alors que la haine est déjà retombée aux niveaux de 2005. la « variétisation » du rap semble avoir, dans un premier temps, coopté le sexisme. Il chute fortement depuis 2021, probablement rattrapé dans un second temps par le mouvement Me Too. On remarque d'ailleurs, sur l'application Gallicagram, une chute spectaculaire du mot « pute » entre 2020 et 2021, de 734 à 473 occurrences, pour un volume de texte presque inchangé. Il demeure que pour quelques années, le sexisme a pu être à un niveau historiquement haut dans le rap français, alors même que la haine avait déjà fortement reculé.

Ce découplage entre haine et sexisme ne se manifeste pas seulement en diachronie, mais aussi de façon transversale. Si l'on effectue chaque année« On a ici exclu les chansons qui comportaient moins de dix lignes, pour lesquelles l'estimation est trop bruitée. » une régression entre haine et sexisme au niveau de la chanson, on remarque que le coefficient [17] (qui mesure à quel point une augmentation de la haine de la chanson est associée à une augmentation du sexisme) s'effondre à partir de 2011, et chute de plus d'un tiers jusqu'en 2019 (Figure 13B). Autrement dit, un texte peut désormais plus aisément être sexiste sans être reconnu comme haineux. Si l'on se plonge dans les données et cherche les chansons qui soient sexistes sans être haineuses, on trouve en haut lieu *Réseaux* de Niska (2017). Son refrain est, de fait, fortement sexiste, mais ne manifeste pas d'agressivité particulière :

J'fais repérage de femmes sur les réseaux

14. Le modèle est disponible ici : https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-french

15. https://huggingface.co/annahaz/xlm-roberta-base-misogyny-sexism-indomain-mix-bal

16. Ici, on court le risque que les modèles soient meilleurs pour détecter la haine et le sexisme sur la période où ils ont été entraînés, soit 2017-9 dans les deux cas. Ce risque était d'ailleurs aussi présent pour nos autres méthodes : il est possible que la liste d'insultes que nous avons utilisé ne contienne pas certaines insultes des années 1990, tombées en désuétude depuis. Il ne nous semble pas possible de s'en prémunir : les catégories elles-mêmes font au mieux, mais nous sommes contraints de faire comme si elles étaient fixes pour faire des études diachroniques.

17. On a préféré cette mesure au coefficient de corrélation, qui ne quantifie pas l'effet d'une variable sur une autre, mais leur tendance à varier ensemble. Un effet très faible mais systématique de l'une sur l'autre pourra donner lieu à une corrélation de 1 si la variable dépendante (ici le sexisme) n'a pas d'autre source de variation. Mais dans ce cas, on obtiendra bien un faible coefficient de régression - formellement, celui-ci ne dépend pas de l'écart-type de la variable dépendante, contrairement au coefficient de corrélation.

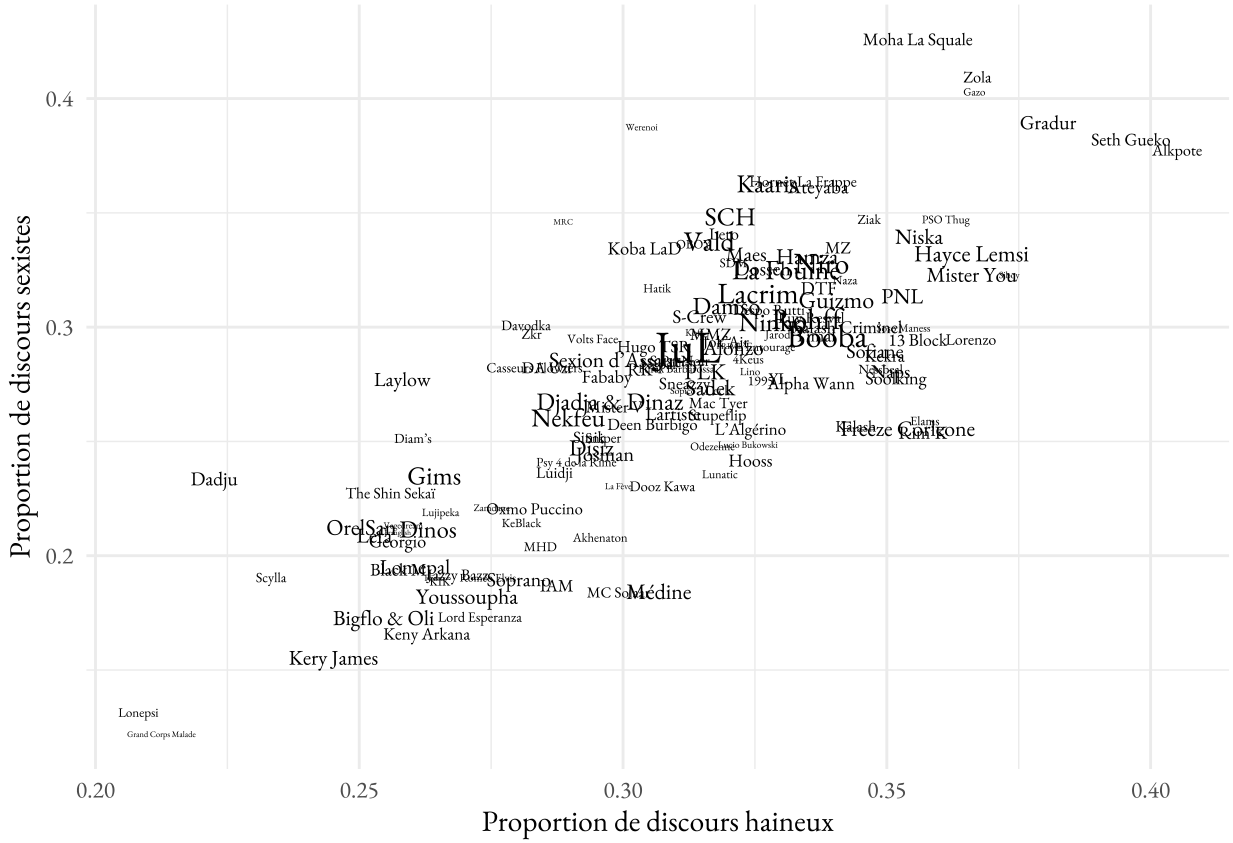

FIGURE 12 – Proportion de discours sexistes et haineux pour les rappeurs les plus vus

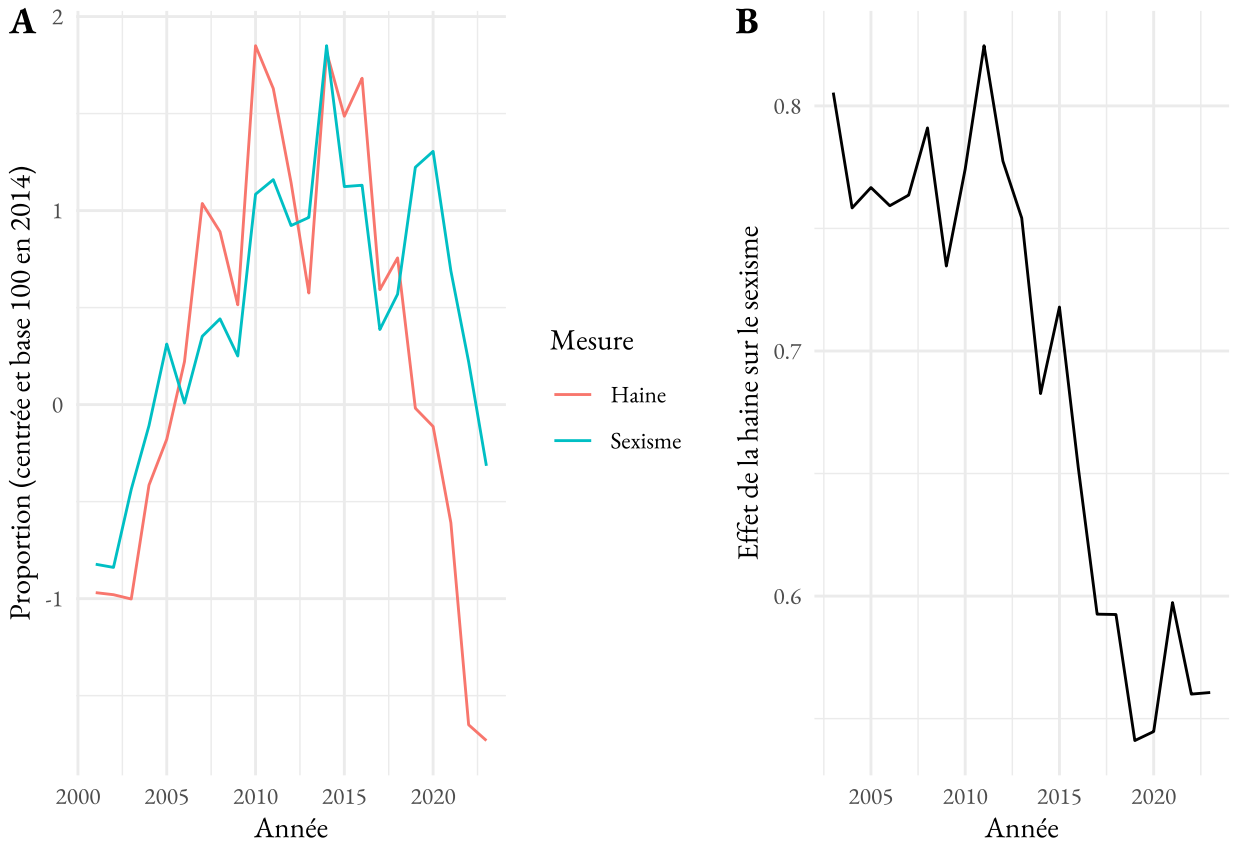

FIGURE 13 – Les discours sexistes et haineux dans le rap français depuis 2001 : leur évolution (A) et leur association au niveau de la chanson (B)

J'ai vu ses lolos
Elle m'a pas followback quand je l'ai follow
Elle fait la go qui connaît pas Charo
Elle m'a vu à la télé', elle a dansé ma choré'

Ce titre de Niska semble donc représentatif d'un sexisme peu haineux, qui se serait diffusé dans le rap français autour de 2015 - on peut considérer que le sexisme véhicule de toute façon une forme de haine, mais nous entendons par là une forme de sexisme qui se donne pour non-agressive, et qui est mieux acceptée socialement que les provocations d'un Booba ou du jeune OrelSan, quand bien même elle dépendrait d'un même substrat. Ce phénomène participe vraisemblablement de la variétisation du rap : ce « sexisme pop » nous semble avoir été présent de longue date dans la variété française, par exemple chez Charles Aznavour et Michel Sardou. La différence fondamentale est que le sexisme n'y est pas utilisé pour choquer ou provoquer l'auditeur. On peut illustrer cette différence en opposant deux chansons. Dans *Cathatsis*, Dooz Kawa (2013) l'utilisait dans une démarche de *clash* envers ses auditeurs : « Bah ouais bien sûr que je nique vos meufs, en douce/Le produit de l'année c'est mon zboub, déclaré en label rouge ». Ce titre est bell et bien détecté par nos modèles comme davantage haineux que sexiste. L'année suivante, Black M (2014), pionnier du rap apaisé et pourtant fort sexiste, écrivait *Je ne dirai rien*, mesuré comme fort sexiste, mais peu haineux :

T'aimes te faire belle, oui, t'aimes briller la night
T'aimes les éloges, t'aimes quand les hommes te remarquent
T'aimes que l'on pense haut et fort que t'es la plus... Oh !
Je ne dirai rien... [...] À moitié dénudée, t'es prête à tout pour plaire
T'aimes pas mon son, mais tu veux ton pass backstage
T'aimes pas les canards, mais t'enchaînes les duckfaces

Dans ce titre, ce n'est plus le rappeur qui passe pour déviant mais au contraire la femme - comme chez Niska, d'ailleurs. Les mots qui pourraient choquer sont tus, ou sous-entendus derrière le « Je ne dirai rien ». En ce sens, même si ce nouveau sexisme est moins obscène et moins agressif, il n'en est pas moins problématique. Il ne peut d'ailleurs pas se justifier par une esthétique « rabelaisienne », comme ont pu le faire Seth Gueko explicitement (Meliholo 2019), et Alkpote implicitement (Berardis 2021).

3.4 Renouvellement générationnel, ou évolution des rappeurs ?

Sont-ce les rappeurs qui ont changé, ou ont-ils été submergés par une nouvelle génération moins sombre, plus égocentrique et moins « lyriciste » ? Cette question est analogue à un problème fréquent en sociologie. Lorsqu'un trait culturel évolue au cours du temps, on peut l'attribuer à trois mécanismes distincts (Ryder 1965) :

— L'effet de période, où le *zeitgeist* évolue avec le temps et s'impose à tous
— L'effet de génération, où le trait dépend du moment où l'individu est venu au monde, puis se fixe - on pourrait ici imaginer que chaque rappeur adopte le style propre à sa génération, puis le conserve tout au long de sa carrière
— L'effet d'âge, où le vieillissement influence le trait - on pourrait imaginer que les rappeurs s'assagissent au fur et à mesure qu'ils vieillissent, ou bien qu'ils se reposent sur leurs lauriers

L'auteur s'est, avouons-le tout de go, cassé les dents sur une question qu'il croyait simple. Pour cause, il est mathématiquement impossible de séparer ces trois facteurs en même temps, du fait d'une redondance d'information (Bell and Jones 2013). En effet, si l'on connaît à la fois la date d'une chanson et l'âge du rappeur au moment où il la chante, alors on peut déduire par soustraction sa date de naissance. Cette redondance viole l'« identifiabilité » du modèle statistique : il existe toujours une infinité de manières possibles, et également vraisemblables, de partager l'effet observé entre ces trois effets. Pour contourner ce problème, il est possible de diviser le temps en tranches grossières (Mason et al. 1973), mais la brièveté de la période ici étudiée rend cette rustine impraticable.

Table 2 : Différence entre le début et la fin de carrière, sur 25 chansons prises au début et à la fin. Ici, seuls sont pris en compte les 216 artistes ayant au moins 30 000 mots, étalés sur une période d'amplitude d'au moins 5 ans. La différence standardisée est la différence divisée par l'écart-type de la variable concernée. La fréquence de chute est la part des rappeurs qui ont une valeur plus faible sur cet indicateur à la fin de leur carrière qu'au début. La p-value est le résultat d'un t-test, une différence étant par convention considérée comme significative lorsqu'elle est inférieure à 5%.

Variable	Différence	Différence standardisée	Fréquence de chute	p-value
Diversité lexicale sur 10 000 mots	-9 %	-0.54	76 %	<0.001
Longueur moyenne des mots	-1 %	-0.28	65 %	0.0033
Fréquences des insultes	-28 %	-0.50	73 %	<0.001
Fréquences des mots négatifs	-11 %	-0.57	74 %	<0.001
Fréquences des mots positifs	6 %	0.29	39 %	0.0023
Fréquence du verlan	2 %	0.08	49 %	0.43
Fréquence du "je"	12 %	0.37	36 %	<0.001
Part des mots dans le dictionnaire	0 %	0.04	51 %	0.65
Fréquence des mots sexuels	-22 %	-0.40	66 %	<0.001
Fréquence des discours sexistes	-7 %	-0.50	73 %	<0.001
Fréquence des discours haineux	-9 %	-0.41	70 %	<0.001

Il est cependant possible de faire plusieurs observations. D'abord, qu'il s'agisse d'un effet d'âge ou de période, il est clair que les rappeurs changent avec le temps. Dans la Table 2, nous avons isolé les rappeurs au long cours, et comparé leurs 25 premières chansons avec leurs 25 dernières. D'impressionnantes différences apparaissent : à la fin de leur carrière, les rappeurs utilisent un vocabulaire 9% moins divers (Figure 14), 28% moins d'insultes, 12% plus de premières personnes, 11% moins de mots négatifs et 6% de plus de mots positifs.

Il est notable que l'effet soit moins net pour la longueur des mots que pour la diversité lexicale (Table 2). Une partie de cette différence s'explique par le fait que la longueur des mots varie beaucoup moins dans la population que la diversité lexicale Lorsqu'on standardise par l'écart-type, la différence s'atténue, mais persiste. Cela suggère cependant que plutôt que de changer de vocabulaire, les rappeurs le restreignent et se concentrent sur certains mots. Certains revendiquent d'ailleurs un virage vers un style plus épuré. En interview, Gaël Faye l'exposait ainsi « Pendant longtemps, j'ai ressenti un complexe de supériorité [...] je trouvais presque cela jouissif [...] d'être un peu hermétique. Puis j'ai appris à écrire différemment. Car une chanson peut aussi être quelque chose de très simple. » (Faye 2020). Dans le hit de l'album qu'il évoque ici [18], le refrain consiste en un seul mot, très long, répété six fois : « Chalouper ». Dans le séminaire La Plume et le Bitume, organisé par Emmanuelle Carinos et Benoît Dufau, Lino constate le même phénomène chez lui, et l'attribue explicitement au vieillissement : « Au début, je rappais pour impressionner les rappeurs, je rappais avec soixante-dix milliards de mots et c'était de la découpe [...] puis mon flow a évolé, j'épure, je mets moins de mots, je vais droit au but. [...] C'est l'âge. »

Notons cependant que les évolutions constatées dans la Table 2 sont globalement congruentes avec l'évolution récente du rap français depuis 2014, vers des textes moins complexes, moins noirs, moins injurieux et davantage à la première personne. Cela suggère un effet de période. On peut par exemple noter que le « rap zumba », initié par Gims et d'abord décrié, a ensuite été partiellement adopté par des rappeurs âgés et installés comme Booba (Validée) et Kaaris (Tchoin) (Thesaurap 2019).

Concernant la part des mots dans le dictionnaire, les rappeurs semblent toutefois ne pas évoluer, à rebours de la tendance générale (Figure 10). Cela suggère un effet de génération : la chute observée en diachronie est due à l'irruption de nouveaux rappeurs au langage moins formel.

Pour étayer cette hypothèse, on peut reproduire l'analyse précédente en tronquant le corpus de tous les textes postérieurs à 2014, date que nous avons identifié comme la rupture chronologique des évolutions susmentionnées. Dans ce cas, les différences de complexité s'annulent toutes deux. Pour le reste, les différences s'amoindrissent, mais persistent : les rappeurs emploient 28% moins d'insultes, 11% moins de mots négatifs et 6% plus de mots positifs. Mais ici, ces trois différences se font

18. *Chalouper* est la chanson la plus vue sur Genius de l'album Lundi Méchant

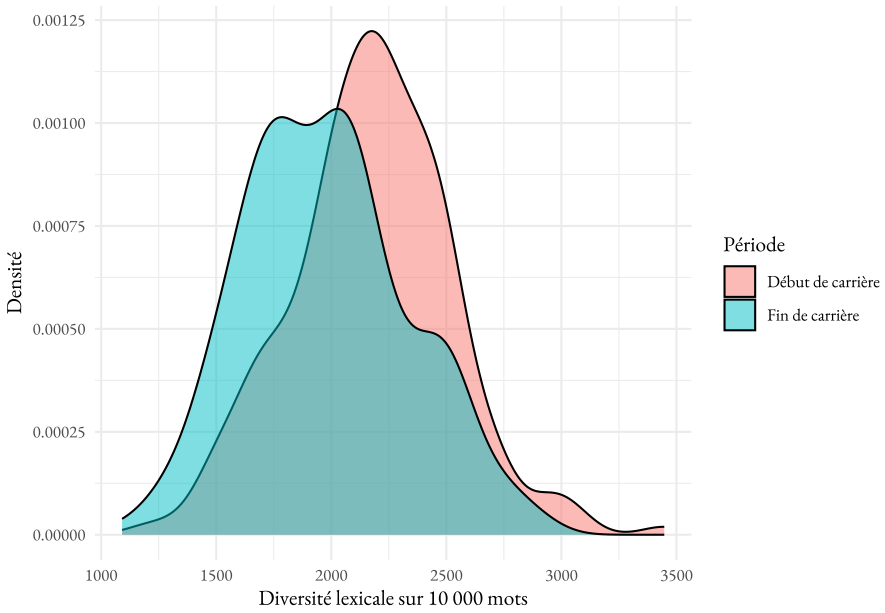

FIGURE 14 – La perte de vocabulaire au fil de la carrière des rappeurs. Ici, nous n'avons gardé que les rappeurs ayant ayant plus de 30 000 mots, étalés sur une période d'au moins 5 ans.

au rebours de l'évolution générale. On l'a vu, entre ses débuts et 2014, le rap français tend à employer de plus en plus d'insultes et à être plus négatif. Il semble donc que l'on ait isolé deux effets d'âge : en vieillissant, les rappeurs deviennent plus polis, et moins noirs, ou moins agressifs. En psychologie de la personnalité, l'âge est d'ailleurs connu pour augmenter l'« agréabilité » et diminuer le « neuroticisme » (Soto et al. 2011), ce qui ne peut pas être neutre dans la façon d'écrire.

Cette analyse se veut exploratoire, et n'épuise pas les explications possibles. Dans le corpus, nous avons renseigné la date de naissance du rappeur et son âge lors de la publication du morceau. Un chercheur plus compétent que l'auteur sur le sujet pourrait mener une analyse plus poussée, par exemple à la façon de Underwood et al. (2022).

3.5 Une cartographie des rappeurs ?

Au fil de l'article, nous avons collecté une quinzaine de variables qui peuvent chacune décrire un rappeur : son usage des insultes, sa diversité lexicale, etc.. On pourrait donc décrire chaque rappeur par un point dans un espace en 15 dimensions - soit 13 de trop pour un écran d'ordinateur. Cependant, il existe une forte redondance entre nos variables. On l'a vu, les rappeurs qui utilisent des mots longs ont aussi tendance à employer des mots divers, et à éviter le pronom « je ». Autrement dit, il existe des structures sous-jacentes. On peut ainsi ébaucher une cartographie des rappeurs, qui met au jour ces structures.

Pour ce faire, on peut recourir aux analyses factorielles, popularisées par Bourdieu (1979) et son analyse des goûts culturels. Sous le capot, l'algorithme (ici l'analyse en composantes principales) cherche deux axes, qui combinent linéairement nos variables pour résumer au mieux les données, c'est-à-dire les trahir le moins possible. Dans la Figure 15, on a représenté le résultat de notre analyse factorielle. Ce graphique étant touffu, une version interactive et zoomable est disponible en ligne : https://regicid.github.io/pca_rappers. Ici, les axes expliquent respectivement 35% et 25% de la variance, ce qui est fort convenable. L'auteur s'est permis de nommer les deux axes, car leur signification lui sautait aux yeux.

En abcisse, on devine le degré auquel l'auteur parle en langage soutenu, avec des mots longs, divers, sans insultes et sans verlan. La pratique du rap conscient, de l'égotrip/méta-rap et du rap poétique semble aller de pair. On a choisi de nommer cet axe la « littérarité », à défaut d'un adjectif substantivé plus neutre. Nous ne voulons pas suggérer qu'un Freeze Corleone est moins « littéraire » qu'un Grand Corps Malade, mais il semble clair que sa langue s'éloigne davantage du français soutenu, que l'on rencontre en général dans la littérature.

En ordonnée, on devine le degré auquel un rappeur attente aux bonnes mœurs, ce qu'on peut appeler son « hardcorité ». Ici pèsent principalement la fréquence des insultes, de l'argot, des mots sexuels et l'absence de mots positifs, et dans une moindre mesure le verlan et une faible proportion de mots issus du dictionnaire. On trouve au sommet de la hiérarchie

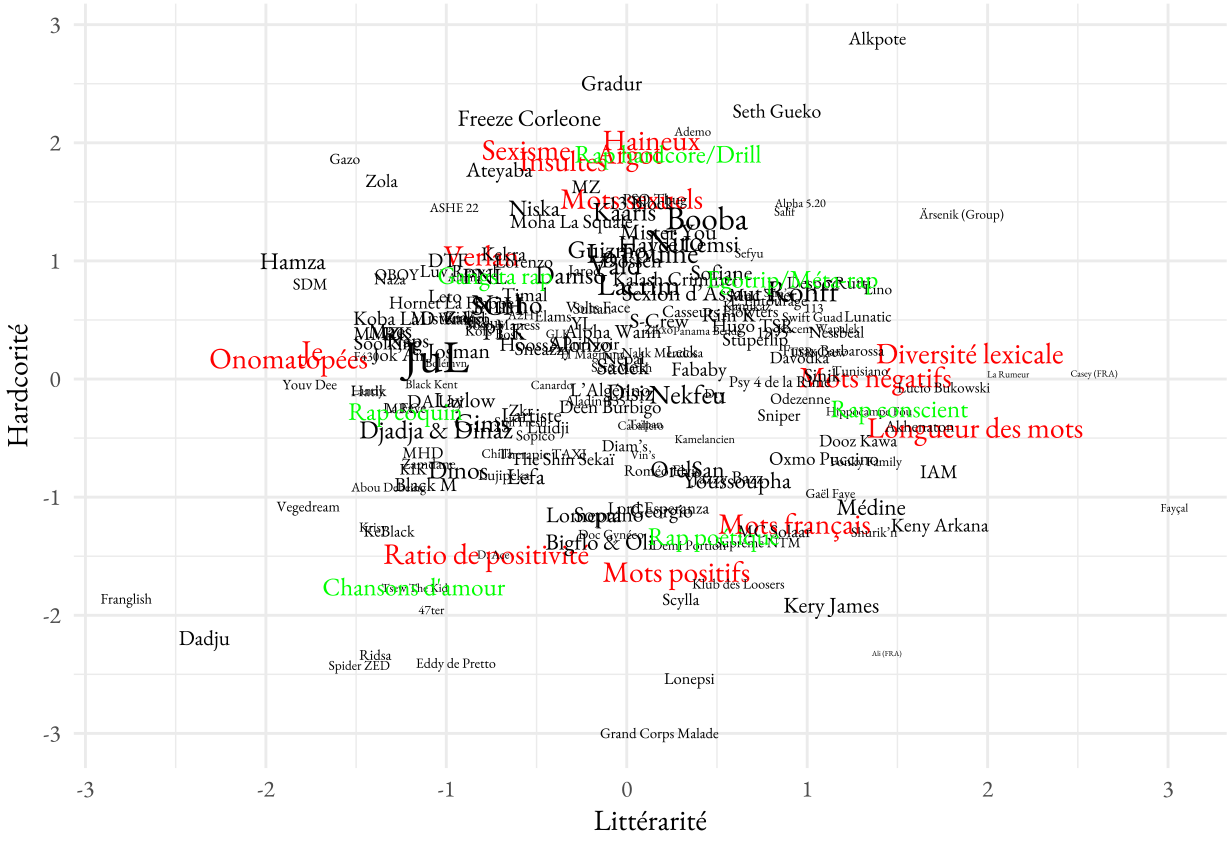

FIGURE 15 – Analyse factorielle de 17 variables calculées par rappeur. Seuls les 200 rappeurs les plus consultés sont affichés. Les axes - des combinaisons linéaires des différentes variables produits de façon endogène par l'analyse factorielle - ont été nommés a posteriori par l'auteur. Les rappeurs sont affichés en noir, les variables en rouge et les topics en vert.

Freeze Corleone, Gradur, Seth Gueko et Alkpote. La diversité lexicale est légèrement associée à cet axe, à l'inverse la longueur des mots. Notons enfin que cet axe ne représente en rien la façon de rapper : le flow, le timbre du rappeur et le beat ne sont pas mesurables avec nos données textuelles. Si Kery James, Médine ou Keny Arkana ont des scores piteux sur cet axe, c'est vraisemblablement parce que leurs textes sont moins hardcores que la façon dont ils les délivrent.

Ces deux axes sont, par construction, orthogonaux, c'est-à-dire qu'ils séparent les deux catactéristiques pour chaque rappeur. Un rappeur peut donc être à la fois littéraire et hardcore. C'est le cas de Lino, de son groupe Ärsenik, d'Alkpote et de Lunatic. Ces rappeurs associent un soin de la langue avec une certaine vulgarité (« La vulgarité ne me pose pas de problème », déclarait d'ailleurs Lino à l'ENS). On pourrait qualifier ce pôle de « rap célinien » : Lino (2014) revendique faire de la « sale littérature », « lisse et trop classe comme du Céline ». Ces rappeurs céliniens sont cependant rares. Pour cause, peut-être, une injonction de l'industrie à choisir un camp, dont Hugo TSR (2008) dit chercher à s'extraire :

> J'ai un stylo c'est pour écrire pas pour signer des contrats
> Dans la musique parait qu'faut être hardcore ou faut être beau et tendre
> J'lai prouvé sur c't'album j'ai l'cul fermé et un cerveau étanche

Au sud-est du graphique, on trouve un pôle que l'on pourrait qualifier de « rap hugolien ». Il se caractérise par une littérarité forte, une grande négativité, mais une faible vulgarité. S'y trouve le rap conscient, qui dénonce une réalité dure, et ses principaux représentants. Parmi eux, Médine, chez qui Victor Hugo est une référence constante « C'est maître Victor Hugo qui disait qu'être contesté, c'est être constaté », rappelle-t-il dans *Voltaire* (Médine 2020), après avoir mêlé son visage au sien sur la pochette de l'album Prose Élite. Ce rap pratique une grande ascèse : peu d'insultes, d'onomatopées, de « je », de lexique sexuel, de sexisme et de discours haineux. Ce relatif puritanisme va de pair avec une forte religiosité (Médine, Kery James, Akhenaton et Ali [19] ont un fort attachement à l'islam, et Shurik'n au taoïsme). Keny Arkana est la seule athée du groupe, mais c'est aussi l'une des deux seules femmes représentées, d'où peut-être une faible appétance pour les discours graveleux.

On l'a dit, Médine, Kery James et Keny Arkana sont parfois associés au rap *hardcore*, du fait d'un *flow* agressif. De fait, on réalise en passant de l'écoute à la lecture que ces textes sont souvent moins agressifs que le *flow* de ces trois rappeurs. D'une certaine façon, ces rappeurs débitent avec rage des discours humanistes. Citons par exemple ce couplet de Keny Arkana (2011), issu d'un morceau intitulé *La Rage* :

> La rage car c'est l'homme qui a créé chaque mur
> S'est barricadé de béton, aurait-il peur de la nature ?
> La rage car il a oublié qu'il en faisait partie
> Disharmonie profonde mais dans quel monde la colombe est partie ?

Plein sud, on trouve le rap modérément littéraire, et résolument anti-*hardcore*. Grand Corps Malade et Lonepsi sont les idéaux-types de ce rap lyrique et sentimental, qu'on pourrait qualifier en caricaturant de « lamartinien » (« il n'a jamais pissé que de l'eau claire », en disait Flaubert). Moins extrêmes, on retrouve Bigflo & Oli et Demi Portion, qui rappent tous deux la bienveillance et l'amour du prochain. À mi-chemin entre rap célinien et lamartinien, on remarque Dooz Kawa, qui se revendique de Lautréamont et « veu[t] les chants de Maldoror comme seule oraison funèbre » (Dooz Kawa 2016).

Au sud-ouest, on trouve les rappeurs qui ne sont ni littéraires ni hardcores. On reconnaît des rappeurs festifs comme Vegedream et Franglish - dont les textes sont plutôt construits pour le dancefloor - ainsi que Dadju, dont le style tend vers la pop. On trouve enfin au nord les rappeurs proprement hardcores, et en particulier Freeze Corleone, Gradur et Seth Gueko. L'auteur s'excuse de n'avoir pas grand chose à dire sur l'ouest du graphique, du fait de son incompétence sur ces sous-genres du rap français.

4 Prolongements (avis aux personnes de bonne volonté)

L'auteur aimerait pour finir suggérer des pistes qu'il aurait aimé suivre, « Mais [s]on temps est restreint [il a] besoin de plus de temps » (Médine 2005).

— On pourrait appliquer la cartographie élaborée pour comparer les années entre elles. Une tentative exploratoire a révélé un mouvement en croissant, du sud-est au sud-ouest, en passant par le nord.

19. Nous avons ajouté le rappeur Ali au graphique, même s'il ne remplissait pas les conditions de popularité. Son positionnement extrême met en valeur la distance qui le sépare de Booba, comme l'ombre et la lumière. Leur groupe Lunatic est d'ailleurs presque au barycentre des deux rappeurs.

— On pourrait de même comparer les différents albums d'un même artiste à ce prisme. Il serait par exemple intéressant de visualiser la trajectoire de Booba sur ces deux axes, au cours de ses 30 ans de carrière.
— Il serait formidable de lemmatiser le corpus, voire d'étiqueter les fonctions grammaticales des mots. La taille du corpus ne poserait ici pas de problème, mais la chose est délicate lorsqu'il est constitué de « français substandard ». Contactée, la linguiste Alena Němcová Polická affirme y travailler
— On pourrait en faire une étude des néologismes (par exemple « faire du sale »).
— On pourrait aussi étudier de la mort des expression. Sur l'application, on a pu observer que « du lourd » s'effondre, remplacé par « du sale » qui commence à son tour à disparaître.
— On pourrait construire une mesure de la radicalité du style. Pour ce faire, votre serviteur a tenté de prendre toutes les variables, de les standardiser et de calculer une distance de Mahalanobis, pour mesurer à quel point les rappeurs sont loin de la norme. Freeze Corleone apparaît comme le rappeur le plus radical, ce qui semble raisonnable. On retrouve ensuite Serane, Fayçal et Ali. Parmi les moins radicaux, Roméo Elvis et les Casseurs Flowters. La distance de Mahalanobis permet de prendre ou non en compte les covariances, et donc de repérer soit les rappeurs qui sont radicaux dans l'absolu, soit ceux qui sont surprenants compte tenu des associations attendues entre variables (par exemple très sexistes mais peu haineux).
— On pourrait croiser cette mesure de radicalité - et plus généralement, nos mesures lexicales - avec le nombre de vues. Il semble, à l'œil nu, que les rappeurs à succès prennent souvent peu de risques lexicaux, à l'exception de Freeze Corleone. *Aurea mediocritas* ? L'exception semble être le sexisme, qui a la plus forte corrélation avec le succès.
— On pourrait chercher à détecter les avant-gardes. Une possibilité serait d'utiliser des plongements lexicaux et de comparer chaque œuvre à celles qui le précèdent, à la façon de Barré and Poibeau (2023).

5 Conclusion

En réalisant ce projet, l'auteur a réalisé l'étendue de son inculture en rap français. Dans le dernier graphique, tous les rappeurs qu'il affectionnait sont situés dans le même quadrant. L'écriture de cet article fut l'occasion de combler certaines lacunes - et en particulier de découvrir Ärsenik. Mais il se sent toujours incapable de produire une analyse fine comme le ferait un linguiste ou un sociologue. Ce n'est pas son métier. Sans affirmer qu'elle suffise, il préfère s'en tenir à la « lecture distante », à la fois par goût et parce qu'il lui manque l'érudition nécessaire à la « lecture proche ». Il espère cependant avoir ouvert des portes - certaines l'étant probablement déjà avant lui. Les données et le code de cet article sont accessibles, et une division du travail pourrait se mettre en place. Il n'y a plus qu'à.

6 Remerciements

Niels Laloé, Clara de Courson, Gaspard de Courson, Benjamin Azoulay, Côme Jaulin, Alexandra Brouillet, Gudrun Ledegen, Anthony Pecqueux, Alena Podhorná-Polická

Bibliographie

Ärsenik. 1998. "Ärsenik (Group) – Boxe Avec Les Mots." https://genius.com/Arsenik-group-boxe-avec-les-mots-lyrics.
Azoulay, Benjamin, and Benoît de Courson. 2021. "Gallicagram : Un Outil de Lexicométrie Pour La Recherche." https://doi.org/10.31235/osf.io/84bf3.
Baltazar, Margarida, and Daniel Västfjäll. 2020. "Songs Perceived as Relaxing : Musical Features, Lyrics, and Contributing Mechanisms." In. Faculty of Music, University of Arts in Belgrade.
Barré, Jean, and Thierry Poibeau. 2023. "Beyond Canonicity." In.
Bell, Andrew, and Kelvyn Jones. 2013. "The Impossibility of Separating Age, Period and Cohort Effects." *Social Science [?] Medicine* 93 (May) : 163–65. https://doi.org/10.1016/j.socscimed.2013.04.029.
Berardis, Chloé. 2021. "La Misogynie Dans Le Rap Français Du XXIe Siècle : Pour Une Lecture Carnavalesque."
Biran, Or, Samuel Brody, and Noémie Elhadad. 2011. "Putting It Simply : A Context-Aware Approach to Lexical Simplification." In, 496–501.
Black M. 2014. "Je Ne Dirai Rien." https://genius.com/Black-m-je-ne-dirai-rien-lyrics.
Bonvin, Audrey, and Amelia Lambelet. 2019. "Exploration Empirique de La Richesse Lexicale : La Perception Humaine." *Linguistik Online*.
Bourdieu, Pierre. 1979. *La Distinction : Critique Sociale Du Jugement*. Paris : Les Editions de Minuit.

Cachin. 2023. “Faire l’histoire du rap.” https://www.radiofrance.fr/franceculture/podcasts/le-book-club/faire-l-histoire-du-rap-6457985.
Carinos, Emmanuelle, and Karim Hammou. 2017. “Approches du rap en français comme forme poétique.” In, edited by Stéphane Hirschi, Corinne Legoy, Serge Linarès, Alexandra Saemmer, and Alain Vaillant, 269–84. Orbis litterarum. Nanterre : Presses universitaires de Paris Nanterre. https://books.openedition.org/pupo/10358.
Chang, Rong-Ching, Jonathan May, and Kristina Lerman. 2023. “Feedback Loops and Complex Dynamics of Harmful Speech in Online Discussions.” In, edited by Robert Thomson, Samer Al-khateeb, Annetta Burger, Patrick Park, and Aryn A. Pyke, 14161 :85–94. Cham : Springer Nature Switzerland. https://link.springer.com/10.1007/978-3-031-43129-6_9.
Chiril, Patricia, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. “An Annotated Corpus for Sexism Detection in French Tweets.” In, 1397–1403.
Courson, Benoît de, Benjamin Azoulay, Clara de Courson, Laurent Vanni, and Étienne Brunet. 2023. “Gallicagram : les archives de presse sous les rotatives de la statistique textuelle.” *Corpus*, no. 24 (January). https://doi.org/10.4000/corpus.7944.
Daller, Helmut, Roeland van Hout, and Jeanine Treffers-Daller. 2003. “Lexical Richness in the Spontaneous Speech of Bilinguals.” *Applied Linguistics* 24 (2) : 197–222. https://doi.org/10.1093/applin/24.2.197.
Do, Salomé, Étienne Ollion, and Rubing Shen. 2022. “The Augmented Social Scientist : Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy.” *Sociological Methods & Research*, December, 00491241221134526. https://doi.org/10.1177/00491241221134526.
Dole, Antoine, and Sté Strausz. 2010. *Fly Girls : Histoire(s) du hip-hop féminin en France : oooo*. Vauvert : Au Diable Vauvert.
Dooz Kawa. 2013. “Catharsis.” https://genius.com/Dooz-kawa-catharsis-lyrics.
———. 2016. “Dooz Kawa – Si Les Anges n’ont Pas de Sexe.” https://genius.com/Dooz-kawa-si-les-anges-nont-pas-de-sexe-lyrics.
Faye, Gaël. 2020. “Gaël Faye : “Beaucoup de mes rimes m’agacent mais je les garde car je ne trouve pas mieux”.” https://www.numero.com/fr/musique/gael-faye-lundi-mechant-album-petit-pays-christiane-taubira-alexis-thibault-interview-sofiane-pamart-rap-zanzibar-roman-zanzibar-paroles.
Gala, Núria, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. “Un Modèle Pour Prédire La Complexité Lexicale Et Graduer Les Mots.” In, 91–102.
Hammou, Karim. 2008. “Le Disque Comme Document : Une Analyse Quantitative de l’usage Du Refrain Dans Les Albums de Rap En Français (1990-2004).” In, 177–93. L’Harmattan.
———. 2009. “Des raps en français au « rap français ».” *Histoire & mesure* XXIV (1) : 73–108. https://doi.org/10.4000/histoiremesure.3889.
———. 2014. “6. Une nouvelle génération de rappeurs en France.” In, 142–68. Poche / Essais. Paris : La Découverte. https://www.cairn.info/une-histoire-du-rap-en-france--9782707181985-p-142.htm.
Hammou, Karim, and Stéphanie Molinero. 2022. “Chapitre IV. Rap et RnB dans les pratiques culturelles en France.” In, 117–48. Ministère de la Culture - DEPS. https://www.cairn.info/40-ans-de-musiques-hip-hop-en-france--9782724638905-page-117.htm.
Hugo TSR. 2008. “Hugo TSR – 2 Minutes Pour Conclure.” https://genius.com/Hugo-tsr-2-minutes-pour-conclure-lyrics.
———. 2012. “Hugo TSR – Alors Dites Pas.” https://genius.com/Hugo-tsr-alors-dites-pas-lyrics.
Jesu, Louis. 2018. “De la subversion sociale et politique dans le rap français contemporain.” *Mouvements* 96 (4) : 43–53. https://doi.org/10.3917/mouv.096.0043.
Keny Arkana. 2011. “Keny Arkana – La Rage.” https://genius.com/Keny-arkana-la-rage-lyrics.
Kery James. 2012. “Kery James – Dernier MC.” https://genius.com/Kery-james-dernier-mc-lyrics.
Klimentová, Julie. 2022. “Francophone Hip Hop Lyrics from the Perspective of Digital Humanities.”
Kryva, Uliana, and Marianna Dilai. 2019. “Automatic Detection of Sentiment and Theme of English and Ukrainian Song Lyrics.” In, 3 :20–23. IEEE.
Lesacher, Claire. 2013. ““Le Rap Est Sexiste”, Ou Quand Les Représentations Sur Le Rap En France Engagent Une Réflexion à Partir de l’intrication Et de La Coproduction Des Rapports de Pouvoir.” *Genre Et Migrations Postcoloniales : Lectures Croisées de La Norme. Rennes, Presses Universitaires de Rennes*, 155–70.
Levaché, Tim. 2023. “Le rap français se mobilise-t-il autant qu’avant ?” https://www.radiofrance.fr/mouv/le-rap-francais-se-mobilise-t-il-autant-qu-avant-9907455.
Lewis, Molly L., and Michael C. Frank. 2016. “The Length of Words Reflects Their Conceptual Complexity.” *Cognition* 153 (August) : 182–95. https://doi.org/10.1016/j.cognition.2016.04.003.
Lino. 2014. “Lino – 12ème Lettre.” https://genius.com/Lino-12eme-lettre-lyrics.
———. 2015. “Le Flingue à Renaud.” https://genius.com/Lino-le-flingue-a-renaud-lyrics.

Lunatic. 1996. “Lunatic – Le Crime Paie.” https://genius.com/Lunatic-le-crime-paie-lyrics.
Maillochon, Florence. 2021. “Classification.” *Sociologie*, September. https://journals.openedition.org/sociologie/9488.
Mason, Karen Oppenheim, William M. Mason, Halliman H. Winsborough, and W. Kenneth Poole. 1973. “Some Methodological Issues in Cohort Analysis of Archival Data.” *American Sociological Review*, 242–58.
Médine. 2005. “Médine – Besoin de Résolution.” https://genius.com/Medine-besoin-de-resolution-lyrics.
———. 2008a. *Don’t Panik*. https://genius.com/Medine-dont-panik-lyrics.
———. 2008b. “Médine – Lecture Aléatoire.” https://genius.com/Medine-lecture-aleatoire-lyrics.
———. 2012. “Médine (Ft. Kayna Samet) – Biopic.” https://genius.com/Medine-biopic-lyrics.
———. 2014. “Médine – Grand Médine.” https://genius.com/Medine-grand-medine-lyrics.
———. 2015. “Médine – Speaker Corner.” https://genius.com/Medine-speaker-corner-lyrics.
———. 2017. *L’homme Qui Répare Les Femmes*. https://genius.com/Medine-lhomme-qui-repare-les-femmes-lyrics.
———. 2018. *Médine – Bangerang*. https://genius.com/Medine-bangerang-lyrics.
———. 2020. “Voltaire.” https://genius.com/Medine-voltaire-lyrics.
———. 2022. “Médine – La France Au Rap Français.” https://genius.com/Medine-la-france-au-rap-francais-lyrics.
Meinecke, Christofer, Ahmad Dawar Hakimi, and Stefan Jänicke. 2021. “Explorative Visual Analysis of Rap Music.” *Information* 13 (1) : 10.
Meliholo. 2019. “Cinq à Seth Gueko - Interview.” https://www.abcdrduson.com/interviews/cinq-a-seth-gueko/.
Niska. 2017. “Réseaux.” https://genius.com/Niska-reseaux-lyrics.
Piantadosi, Steven T., Harry Tily, and Edward Gibson. 2011. “Word Lengths Are Optimized for Efficient Communication.” *Proceedings of the National Academy of Sciences* 108 (9) : 3526–29. https://doi.org/10.1073/pnas.1012551108.
Piolat, A., R. J. Booth, C. K. Chung, M. Davids, and J. W. Pennebaker. 2011. “La version française du dictionnaire pour le LIWC : modalités de construction et exemples d’utilisation.” *Psychologie Française* 56 (3) : 145–59. https://doi.org/10.1016/j.psfr.2011.07.002.
Podhorná-Polická, Alena. 2020. “RapCor, Francophone Rap Songs Text Corpus,” December.
RapMinerz. 2023. “Diversité & Originalité du Rap FR.” https://www.rapminerz.io/articles/diversite-originalite-du-rap-fr.
Roquebert, Corentin. 2020. “Le Capital Social Des Rappeurs : Les Featurings Entre Gains de Légitimités Et Démarche d’authentification Professionnelle.” *Volume !. La Revue Des Musiques Populaires*, no. 17 : 2 : 61–81.
Ryder, Norman. 1965. “The Cohort as a Concept in theStudy of Social Change,[la Cohorte Concepto En El Estudio Del Cambio Social] American SociologicalReview 30 (6), 843-861.” *Recuperado de Http ://Personal. Psc. Isr. Umich. Edu/Yuxie-Web/Files/Soc543-2004/Ryder1965. Pdf*.
Soto, Christopher J., Oliver P. John, Samuel D. Gosling, and Jeff Potter. 2011. “Age Differences in Personality Traits from 10 to 65 : Big Five Domains and Facets in a Large Cross-Sectional Sample.” *Journal of Personality and Social Psychology* 100 (2) : 330.
Stupeflip. 2011a. *La Menuiserie*. https://genius.com/Stupeflip-la-menuiserie-lyrics.
———. 2011b. “Stupeflip – Stupeflip Vite !!!” https://genius.com/Stupeflip-stupeflip-vite-lyrics.
Thesaurap. 2019. “Le rap zumba : d’infréquentable à incontournable.” https://thesaurap.fr/articles/le-rap-zumba-dinfrequentable-a-incontournable/.
Underwood, Ted, Kevin Kiley, Wenyi Shang, and Stephen Vaisey. 2022. “Cohort Succession Explains Most Change in Literary Culture.” *Sociological Science* 9 (May) : 184–205. https://doi.org/10.15195/v9.a8.
Youssoupha. 2018. “Youssoupha – Le Jour Où j’ai Arrêté Le Rap.” https://genius.com/Youssoupha-le-jour-ou-jai-arrete-le-rap-lyrics.