Revue d'Études Françaises
N° 25 (2021)
DOI : 10.37587/ref.2021.1.14

ALENA PODHORNÁ-POLICKÁ – LAURENT CANAL – ADÉLAÏDE EVREINOFF – ANNE-CAROLINE FIÉVET

Autour de l'enseignement de la dia-variation en FLE universitaire. Le cas de la constitution d'un dictionnaire thématique à partir du corpus de rap francophone *RapCor*

Since 2006, texts of rap songs have been used in seminars of sociolexicology and translation at the Institute of Romance Languages and Literatures of Masaryk University, in order to teach dia-variation in authentic, current and attractive contexts for students. In parallel to these annual seminars, which aim at a qualitative approach, a linguistic corpus, the RapCor, has been built from the texts available on the booklets of French-speaking hip-hop artists, since 2009, aiming then at a rather quantitative approach. Since 2018, this corpus is available online, as an open-source linguistic corpus throughout the query platform Sketch Engine (actual size from October 2021 is 1292 songs, i.e. 769k of tokens). In this article, we explain the phases of interaction between professors and students for the elaboration of a specialized dictionary serving to train existing lemmatizers for French by improving the annotation of substandard items. We focus on two different levels: on the one hand, the teaching of dia-variation to students of French as a foreign language and, on the other hand, the debates on the categorization of the substandard lexicon identified by the students within the laboratory bringing together researchers and PhD students, specialists in sociolects and/or in intercultural sociodidactics. Some examples from the categories of proper names, polysemic verbs and loanwords will be presented in more detail.

1. Introduction

Les paroles des chansons de rap représentent une source précieuse pour attester des innovations lexicales qui sont diffusées auprès des jeunes générations comme en témoignent les travaux des sociologues (*cf.* Hammou,

2012) ou des littéraires (*cf.* Ghio, 2012), entre autres. Concernant l'argotologie, des exemplifications extraites de chansons de rap figurent dans les divers dictionnaires d'argot contemporain (*cf.* Tengour, 2013 ; Goudaillier, 2019) et parmi la multitude de genres de chansons, le rap est le plus souvent au cœur de l'intérêt des chercheurs depuis les pionniers dans ce domaine (Calvet, 1999 ; Trimaille, 1999). À partir de 2006, les textes de chansons de rap sont utilisés dans les séminaires de sociolexicologie et de traduction à l'Institut des Langues et Littératures romanes de l'Université Masaryk, afin d'enseigner la dia-variation dans des contextes authentiques, actuels et attirants pour les étudiants (*cf.* Fiévet, Podhorná-Polická, 2010). Parallèlement à ces séminaires annuels qui visent une approche qualitative, depuis 2009, un corpus linguistique a été initié à partir des textes disponibles sur les livrets des albums de rap francophones, visant alors une approche plutôt quantitative. Dès 2018, ce corpus est en ligne, en open source[1], sur la plateforme d'interrogation Sketch Engine (co-créée à l'Université Masaryk ; Kilgarriff, Rychlý, Smrž, Tugwell, 2004). Dans cet article, notre objectif sera d'éclairer les phases d'interaction entre les professeurs et les étudiants pour l'élaboration d'un dictionnaire servant à entraîner le lemmatiseur actuellement utilisé, notamment en ce qui concerne l'amélioration de l'annotation des lexèmes substandard. Nous nous focaliserons sur deux niveaux différents : d'un côté, sur les étapes d'enseignement du traitement textuel et de la dia-variation auprès des étudiants de français langue étrangère (FLE) et de l'autre, sur les débats autour de la catégorisation du lexique substandard relevé par les étudiants au sein du laboratoire regroupant les chercheurs et les doctorants, tous spécialistes des sociolectes et/ou de la sociodidactique de l'interculturel.

2. L'apprentissage grâce à la création d'un corpus de textes de rap : enjeux et écueils

Dans les cours de FLE universitaire, l'enseignement s'oriente vers le registre standard qui, étant donné son caractère prestigieux et relativement stable, est le plus propice à la maîtrise de la langue à des fins professionnelles, c'est-à-dire à une communication efficace lors de situations formelles. Or, pour

[1] Pour l'instant, l'accès libre est assuré jusqu'en mars 2022 pour le public académique. Actuellement, la version *RapCor1292* (c'est-à-dire comportant 1292 textes de chansons du rap francophone est disponible, d'une taille de 769, 649 tokens) d'octobre 2021 est disponible.

la communication de tous les jours, en dehors du cadre professionnel mais aussi pour satisfaire à des tâches traductologiques diverses, il est également important d'utiliser ou, du moins, de reconnaître d'autres registres utilisés dans la francophonie, avec ses riches variétés diatopiques, diaphasiques et diachroniques. L'efficacité de l'enseignement du registre familier via les textes de rap a été évoquée dans de nombreux travaux (entre autres, voir Gendron, 2017 ; Sakano Fernandes, 2013). C'est pourquoi nous passerons directement aux aspects pratiques de l'expérience pédagogique telle que nous avons pu la vivre grâce au travail effectué lors de cours semestriels qui avaient pour but de travailler sur le projet de la création du corpus de textes de rap francophone *RapCor*.

Ainsi, depuis l'automne 2020, les étudiants du Département de langue et littérature françaises de l'Université Masaryk de Brno, de niveaux L2, L3 et M1, ont la possibilité de découvrir le langage substandard via ce cours optionnel qui leur permet de travailler pendant un ou plusieurs semestres, selon leur motivation, sur les données numériques. Le travail se divise en plusieurs phases :

- dans un premier temps, chaque étudiant choisit ou se voit attribuer un rappeur ou un groupe et son album (ou ses albums) dont les livrets avec les paroles ont été scannés ou que l'étudiant scannera lui-même. Après avoir recherché les métadonnées de l'artiste ou du groupe (nom, pseudonyme artistique/nom de scène, information sur la naissance, origine, banlieue à laquelle l'artiste se réfère, etc.) ainsi qu'après avoir complété la discographie et le tracklisting (liste des titres) chronologiquement (y compris la vérification de la disponibilité sur Internet des transcriptions des paroles qui n'ont pas été intégrées aux booklets (livrets des disques ou pochettes des albums)), s'ensuit l'étape de la numérisation des textes. Si le livret contient des paroles, chaque texte de rap est découpé un par un à l'aide du logiciel d'édition GIMP et ensuite océrisé avec l'application Abby FineReader. L'étape suivante consiste à baliser les parties du texte sur le plan structurel (comment la chanson est segmentée en couplets et en refrains, qui chante quel couplet, etc.), et marquer, tout en réécoutant la chanson et en comparant les paroles écrites avec les paroles réellement chantées, toutes les différences entre les deux versions (documents appelés « version pochette » et « version son »). Les enjeux de ce travail minutieux sont multiples (voir Podhorná-Polická, 2020), l'un des plus importants pour les

recherches en argotologie et pour l'application aux corpus du type WaC (web as corpus ; *cf.* Jakubíček et al., 2020), est le repérage des variantes graphiques des lexèmes substandard fréquents à l'oral mais à graphie flottante (pour l'exemplification, voir Podhorná-Polická, Fiévet, 2022) ;
- dans un deuxième temps, lors de la phase à caractère plus « interactif », l'étudiant déclenche l'opération automatique de segmentation du texte travaillé en tokens (positions dans le texte). Il arrive ainsi à une organisation de texte où chaque mot occupe une ligne d'un tableau, sauvegardé sous le format Libre office .ods, qu'on nommera alors ODS. Le tableau contient, en outre la colonne A (comportant ligne par ligne le texte tokenisé), également une annotation automatique en catégories morphosyntaxiques (en colonne B) et une lemmatisation (en C), voire d'autres colonnes comme expliqué *infra*, tout cela grâce à un seul clic après avoir copié-collé le texte sur l'interface web qui comporte un outil adapté spécifiquement à nos besoins par nos collègues informaticiens de l'Université de Brno. Cet outil est basé sur le célèbre logiciel *TreeTagger* (désormais TT, Stein, Schmid, 1995) mais en plus, il est enrichi par les lemmes inconnus (la ligne comportant un lexème inconnu par TT est coloriée en jaune). On y fait ressortir également la nécessité de désambiguïsation (la ligne comportant une hésitation du TT dans le choix parmi deux lemmes possibles est coloriée en rouge), certaines marques de l'oralité (p.ex. si l'élision est effectuée devant les mots comportant la consonne à l'initiale, l'annotation dans la colonne D, consacrée entre autres à la morphosyntaxe va comporter automatiquement une balise ELIS). En ce qui concerne certains lexèmes substandard annotés dans les versions précédentes du corpus, la colonne D sera coloriée en bleu turquoise et va comporter différentes autres balises de l'attribut ***ms*** (s'il s'agit d'un procédé morphosyntaxique particulier, p.ex. APO pour apocope, APH pour aphérèse, etc. ; disponible en utilisant les requêtes CQL), de l'attribut ***emprunt*** (s'il s'agit d'un emprunt quelconque) ou de l'attribut ***lexik*** (s'il s'agit d'un lexème marqué par les dictionnaires de référence[2]). Comme le dictionnaire du TT est très faiblement fourni concernant le français substandard et que notre vieux dictionnaire interne devient suranné, nécessitant un travail de désambiguïsation approfondie, notre mission commune depuis janvier 2021 est de proposer un

[2] Ces derniers nous servent également en tant que dictionnaires d'exclusion pour décider ce qui est ou n'est pas substandard. Il s'agit des éditions les plus récentes du *Petit Larousse* et du *Petit Robert.*

nouveau dictionnaire. Ce dernier a pour objectif de regrouper sous un lemme toutes les variantes graphiques du lexème et d'expliquer la lexicogénèse et le registre supposé par les dictionnaires de référence[3] ainsi que le sens (colonne E nourrissant l'attribut ***sens***) des lexèmes substandard connus ou méconnus de ces derniers (à l'aide d'une recherche lexicographique sur la toile). Ce travail est à destination de tous ceux qui vont se servir de cet outil dans le futur (par exemple lors de la recherche des séries synonymiques argotogènes par thèmes) mais surtout de tous les étudiants de FLE qui obtiennent ainsi une information rapide sur le(s) sens possible(s) d'un lexème substandard lors du traitement des chansons. Il s'agit en quelque sorte d'un entraînement du TT (tel qu'évoqué dans Poudat, 2009 ou Dupuis, Kapitan, Daoust, 2010, entre autres) mais plus particulièrement ciblé vers des objectifs traductologiques qui visent la pédagogie de la variation et des allusions interculturelles.

Les étudiants ont alors pour tâche de travailler avec le tableau ODS (parcourir le texte segmenté à la verticale et, si nécessaire, regrouper les défigements des expressions substandard figées, corriger les fautes dans les annotations morpho-syntaxiques au niveau des catégories grammaticales et proposer des corrections des lemmes selon les règles internes – basées sur la fréquence ou la convention de simplification orthographique maximale en cas de variation graphique non-dictionnarisée). Ce tableau est partagé entre l'étudiant et les enseignants-chercheurs du laboratoire qui contrôlent les différentes étapes du travail de l'étudiant et lui proposent des améliorations. Les étudiants créent également un autre document Excel partagé nommé *Dictionnaire RapCor ODS_nom et prénom*. Dans ce petit dictionnaire individuel qui alimente et est alimenté par le grand dictionnaire appelé *Dictionnaire RapCor Validé* (voir *infra*), ils recopient et travaillent toutes les lignes coloriées par notre tagger (c'est-à-dire non reconnues par le TT ou identifiées comme substandard ou ambiguës) et proposent des lignes à désambiguiser, s'il s'agit d'un néosémantisme ou du sens substandard d'un mot polysémique. Cette dernière tâche s'avère la plus problématique et nécessite le suivi des professeurs puisqu'il s'agit ici de locuteurs non-natifs assez peu expérimentés pour la détection du français substandard. Un travail où ils se

[3] Tout en prenant en compte l'évolution des marques du substandard et la variation diatopique (incluant également le dictionnaire canadien USITO parmi les sources de référence pour le volet franco-canadien du corpus).

sentent plus autonomes consiste en une proposition d'étiquettes indiquant les informations sur la lexicogenèse et sur le registre – si cette information est présente dans les dictionnaires de référence –, et en une recherche des définitions des sociolectismes (souvent des néologismes) absents desdits dictionnaires, ainsi qu'en la traduction vers le français standard. Les tokens ainsi enrichis permettront non seulement une compréhension plus rapide et plus efficace dans le futur grâce au remplacement du vieux dictionnaire par un nouveau, mais également une recherche plus fiable car plus homogène par catégories morpho-syntaxiques, par exemple l'étiquette VERL regroupera de manière plus fiable qu'à présent tous les lexèmes verlanisés (créés par métathèse).

Parmi les écueils rencontrés les plus problématiques, au niveau de la didactique du FLE, nous pouvons citer la vitesse de travail (motivée par la volonté de terminer la dernière phase de traitement de la chanson aussi rapidement que possible), la trop grande difficulté d'une chanson (passages inaudibles, lexèmes peu diffusés en dehors d'un micro-argot du rappeur, etc.) ou le faible niveau de français malgré une bonne implication dans la tâche demandée. Il est donc important d'enseigner aux étudiants une certaine rigueur et de vérifier minutieusement leur travail s'ils « tombent » sur une chanson extrêmement longue ou truffée de sociolectismes, voire leur réexpliquer certaines bases linguistiques (notamment en syntaxe et en morphologie). Un exemple typique des négligences rencontrées est que le logiciel peut dire qu'un lemme est un nom propre alors que c'est un nom commun parce qu'il y a une majuscule en première lettre et pas de point avant (comme c'est souvent le cas des textes présentés à l'instar des poèmes sur les livrets), ce qui implique que l'étudiant doit revenir jusqu'à la phase de comparaison des versions et réajuster plusieurs colonnes de son tableau. Il en est de même pour les virgules qui permettent au TT de désambiguïser les différentes catégories grammaticales plus correctement.

3. Catégoriser le substandard et expliquer les allusions socio-culturelles

Lors des séminaires hebdomadaires de recherche qui regroupent les quatre auteurs de cet article – deux chercheuses et deux doctorants – ainsi qu'un

troisième doctorant qui s'est joint au projet plus tardivement[4], les cas problématiques sont débattus et les lexèmes substandard vus par les étudiants dans les différents textes de rap sont intégrés dans le *Dictionnaire RapCor Validé* susmentionné qui sera très prochainement utilisé pour relemmatiser et réannoter le corpus entier. Dans ce fichier Excel, plusieurs onglets ont été créés, à savoir : noms, gentilés, noms propres, toponymes, anthroponymes, locutions nominales, adjectifs, locutions adjectivales, pronoms, verbes invariables, verbes, verbes pronominaux, locutions verbales, adverbes, locutions adverbiales, prépositions et interjections. Parmi ces catégories, on retrouve les catégories traditionnelles telles que les noms, les adjectifs ou les verbes mais également les sous-catégories des noms tels que les gentilés, les toponymes ou les anthroponymes. Le plus souvent, ce sont ces deux dernières catégories qui soulèvent des allusions socio-culturelles pouvant poser des problèmes d'identification pour les étudiants si ce ne sont pas des références globalement connues.

Lors de la correction des travaux d'étudiants, nous avons pris conscience qu'il était nécessaire d'adopter des approches spécifiques en fonction des types de lexèmes. Parmi toutes les problématiques qui se sont présentées lors du séminaire, nous avons choisi, pour cet article, trois catégories intéressantes du point de vue des interactions pédagogiques, à savoir le traitement des noms propres (y compris les déonomastiques), des polysèmes et des emprunts à la langue la plus exposée, l'anglais.

3.1 Noms propres et déonomastiques

Pour décider si un nom propre doit rentrer dans le *Dictionnaire RapCor Validé* ou non, il est tout d'abord recherché dans la deuxième partie (encyclopédique) du *Petit Larousse* ainsi que sur Internet, afin de circonscrire au mieux les allusions socioculturelles adhérentes qui sont facilement identifiables pour les non-natifs, ce qui nous amène à prendre une décision sur la nécessité éventuelle de les expliciter. Par exemple, nous avons longtemps débattu sur *Prozac/prozac*. En effet, le terme *prozac* renvoie au sens généralisé d'« antidépresseur » mais c'est aussi une « marque de médicament », s'il est

[4] Il s'agit de Simon Naumann, spécialiste en plurilinguisme dans le rap francophone (voir Naumann, 2020).

utilisé avec une majuscule à l'initiale. Or, cette distinction est très peu respectée dans le « parlécrit » des rappeurs. Pour trancher entre les catégories *nom propre* et *nom commun*, les deux variantes (avec et sans majuscule) ont été ajoutées au *Dictionnaire RapCor Validé* en tant que catégorie DEON, nom propre en train de glisser vers les noms communs, c'est-à-dire la catégorie des déonomastiques. Dans de nombreuses chansons de rap, nous avons également pu trouver des noms propres à référent humain, des anthroponymes, comme le terme *King Kong* que nous avons défini comme « gorille géant, personnage du film éponyme » mais également en tant que locution déonomastique comme « nom générique pour quelqu'un ou quelque chose de grand et puissant » afin que la désambiguïsation puisse être opérée la prochaine fois que cet item sera retrouvé, peu importe comment il sera orthographié, avec une minuscule ou une majuscule, avec un tiret ou en deux mots – cet exemple, même s'il est relativement simple et à la limite des allusions globalement connues (mais jusqu'à quelle génération ?), montre la complexité de la tâche et la nécessité d'aborder plusieurs aspects à la fois. En ce qui concerne les noms propres à référents géographiques, les toponymes, la tâche s'avère complexe dans le cas de références à des endroits micro-toponymiques évoqués souvent par des rappeurs débutants n'ayant que des ambitions locales mais aussi dans le cas des macro-toponymes ayant deux graphies possibles comme *Irak* ou *Iraq*, par exemple. *Le Petit Larousse* l'écrit avec un *q* final mais dans l'orthographe « populaire » des internautes dont témoignent les corpus du type WaC (tels qu'*Araneum Francogallicum maximum*, etc.), il apparaît le plus souvent avec un *k* à la fin. Cela sera traité par un regroupement sous le même lemme où la priorité est donnée à la variante normée par le dictionnaire de référence et non à la fréquence, ce qui est le critère prédominant pour les cas où aucun repère n'est disponible dans ces mêmes dictionnaires. Enfin, le problème le plus fréquent avec les noms propres, que ce soit pour les étudiants ou pour nous, c'est de ne pas identifier la référence. On peut trouver le cas d'une personnalité importante pour le rappeur comme p.ex. *King Zaman* (empereur d'Afghanistan du XIX^e^ siècle[5]) ou encore, la référence mentionnée est connue seulement dans le milieu du rap restreint, comme *Panama* (qui renvoie ici au collectif de rap parisien *Panama Bende* autour du rappeur PLK[6]). Il peut s'agir également

[5] https://en.wikipedia.org/wiki/Zaman_Shah_Durrani (consulté le 10/12/2021.)
[6] https://fr.wikipedia.org/wiki/Panama_Bende (consulté le 10/12/2021.)

d'une référence comprise par une génération comme *double effet Kiss Cool* qui se rapporte à une publicité pour une pastille rafraîchissante (*Kiss Cool*) diffusée dans les années 1980–1990[7], ayant pour sens « effet supplémentaire à celui normalement attendu[8] ». Il est à noter que la concentration de ce type de noms propres dans certaines chansons ralentit considérablement le travail sur d'autres sociolectismes.

3.2. Mots polysémiques et locutions

La polysémie qu'un mot peut revêtir ou sa présence dans une locution qu'il faut nécessairement prendre en compte complexifient le processus de recherche de consensus entre les membres de l'équipe lors de l'étiquetage. Ces tropes, souvent (mais pas exclusivement) définis comme argotiques, sont généralement nombreux dans les langues et le sont d'autant plus dans le rap. Dans l'argot, la polysémie est l'un des principes d'élaboration sémantique majeurs, et l'utilisation des locutions, parfois excessive, souvent sujettes, entre autres, à des jeux de mots[9] (paronymie, faux proverbe, etc.), ou à des modifications actualisantes.

Les lexèmes de différents types sont recherchés dans des dictionnaires papier et en ligne en suivant la méthode des filtres successifs (Fiévet, Podhorná-Polická, 2013). *Le Petit Robert* (PR) et *Le Petit Larousse* (PL), qui font généralement autorité, sont tout d'abord consultés (premier filtre). Puis le *Dictionnaire de l'argot* (Colin, 1992) (DA), plus spécialisé et *Comment tu tchatches ! Dictionnaire du français contemporain des cités* (Goudaillier, 2019) (DC) plus actuel, y compris les dictionnaires en ligne tels que le *Dictionnaire de la Zone* (DZ) ou le *Dictionnaire d'argot avec Bob, l'autre trésor de la langue* sur le site *ABC de la langue française* (AB) (deuxième filtre). Si nécessaire, cette prospection se poursuit dans d'autres dictionnaires en ligne, parfois étrangers pour les emprunts à d'autres langues. Dans le cas où aucune définition n'est trouvée, nous proposons notre propre définition mais seulement

[7] https://www.youtube.com/watch?v=TntufaVUQxs (consulté le 10/12/2021.)

[8] https://www.laculturegenerale.com/double-effet-kiss-cool-deuxieme-effet-definition-origine-signification/ (consulté le 12/12/2021.)

[9] *A priori*, l'objectif de l'auteur est de joindre à son discours une fonction esthétique : la rime, même si certaines compositions semblent suivre un processus inverse.

si cette dernière recueille le consensus de la totalité des personnes du groupe de recherche.

Dans la phrase suivante : « Le prof tapait trop de prozac » (RimK *feat.* Awa Imani, Bac-5, album *Chef de famille*, 2012), le verbe *taper* illustre bien cette problématique multiple. En effet, ce verbe, transitif, intransitif ou pronominal, ainsi que ces expressions associées (car *taper* peut se verlaniser sous la forme invariable *péta*) recouvre un grand nombre de nuances sémantiques peu documentées dans les dictionnaires de tous types. L'objectif de l'étiquetage est donc d'intégrer au corpus l'acception correspondant au verbe dans son contexte, en l'occurrence « absorber », ou « consommer » (des médicaments).

Le PR, qui comporte 14 acceptions (v. tr., v. intr., et v. pron.) et 20 locutions concernant le verbe *taper* ne donne aucune définition satisfaisante, seuls des synonymes de *manger* et *boire* : « s'enfiler » et « s'envoyer » s'en rapprochent mais ils dépendent uniquement de la forme pronominale. Le PL (9 acceptions et 4 locutions), le DC (2 acc. de *péta* et 4 loc. avec *taper*) et le DZ (12 acc.) sont dépourvus de toute acception appropriée. Le DA (12 acc. et 10 loc.) propose les deux verbes correspondants : « consommer » et « absorber », mais ici aussi, ces acceptions dépendent de la forme pronominale plus usuelle. L'AB (23 acc. et 100 loc., + 8 acc. de *péta*), quant à lui, propose « consommer » pour le verbe *taper*. C'est donc cette dernière référence qui sera prise en compte dans le *Dictionnaire RapCor Validé*. Cette acception de *taper*, plus connue à la forme pronominale, est le résultat d'une modification (*se taper* > *taper*) à considérer. Une recherche de *taper* sur le corpus *RapCor* nous permet d'inventorier de nombreuses acceptions qu'il est nécessaire d'étiqueter afin de les rendre accessibles et compréhensibles de tous dans le contexte qui leur est donné.

Dans cette même recherche se trouvent aussi de nombreuses locutions telles que *taper la fuite* : « s'enfuir » (AB), *se taper des barres* (de rire) : « rire » (AB), ou *taper la causette* : « parler », syntagmes qu'il est nécessaire de regrouper sur une ligne. Les locutions *taper la discussion* ou *la discute* (DC/AB) coexistent dans les dictionnaires mentionnés ci-dessus mais pas *taper la causette*. Comme mentionné *supra*, en l'absence de références dictionnairiques, notre propre définition, sous la balise ND, est apposée.

Après une éventuelle déduction du sens dans le contexte donné (quand celui-ci n'est pas évident), et après une recherche dictionnairique approfondie, un verbe tel que *taper*, très utilisé dans le rap, peut donc se comprendre rapidement selon l'acception qu'il recouvre réellement nonobstant les méandres

de sa polysémie. Le corpus *RapCor* constitue ainsi un apport considérable pour les apprenants et un outil de recherche particulièrement précis.

3.3. Emprunts et anglicismes

Les alternances codiques sont annotées différemment des emprunts (par exemple, la balise AC : angl indique que la phrase entière est prononcée en anglais, avec une prononciation sans adaptation phonotactique, tandis que sous la balise ANGL, on retrouve différents emprunts à l'anglais ; pour les autres langues, cette division sera identique, AC : ital *vs* ITAL, par exemple). Dans les discussions au sein du groupe de recherche, la catégorisation des anglicismes a donné lieu à de nombreux débats et la réflexion se doit d'être poursuivie. D'une part, ce qui est indiscutable, c'est que nous devons classer comme ANGL (anglicisme) les lexèmes qui sont clairement notés comme tels dans le PR ou le PL (marque *anglicisme*), par exemple le substantif *flash* (« vision de très courte durée » dans le PR 2022). D'autre part, en tant que spécialistes du substandard, nous avons également décidé de considérer comme anglicismes deux autres cas : 1) Quand le mot est noté comme standard dans le PR et le PL mais marqué comme « Anglic. » ou « de l'anglais » dans les dictionnaires d'argot. Par exemple, *squatter* (dans le DZ : « occuper un endroit, rester à un endroit », de l'anglais *to squat* « s'accroupir »). Et 2) Quand le mot n'est présent ni dans le PR, ni dans le PL et que, dans l'imaginaire linguistique d'au moins 2 sur 3 (actuellement 3 sur 4) des natifs francophones du groupe, cela reste un anglicisme évident. Par exemple, *rap* avec la prononciation adaptée [ʁap] ou *featuring* [fityʁiŋ]. En revanche, quand le PL ou le PR disent que le lexème vient « de l'anglais », comme le lexème *footballeur*, nous considérons qu'il est déjà assez intégré donc que ce n'est plus un anglicisme et nous le notons en tant que ANG : itg (anglicisme intégré). Enfin, nous avons créé une troisième catégorie, AGC (anglicisation), pour les cas où, pour des raisons paronymiques, on trouve une graphie anglaise pour la notation graphique d'un mot français. On pourra citer *beef,* employé à la place de *biff,* où on constate une confusion entre le « beefsteak » (variante graphique : *bifteck* ; avec l'expression *gagner son bifteck*, « gagner sa vie[10] » et le *biff*, (apocope de *biffeton* ou *bifton*) pour « argent, billet de banque » (DZ). Une catégorisation plus fine que celle que nous

[10] https://fr.wiktionary.org/wiki/gagner_son_bifteck (consulté le 10/12/2021.)

venons de présenter est à la fois séduisante mais dangereuse car entrent en jeu les facteurs diatopiques (vision de la francophonie européenne *vs* vision de la francophonie canadienne) et diachronique (intégration rapide des emprunts à l'anglais).

4. En guise de conclusion

Ce projet, à la fois d'enseignement et de recherche, permet de travailler en équipe sur les différents niveaux du substandard. Le fait que le travail soit réalisé en amont par des étudiants dont la langue maternelle n'est pas le français a certes l'inconvénient de générer plus d'erreurs que pour un natif mais cet écueil est largement compensé par l'apport des découvertes dans le domaine de la socio-didactique du FLE (quels sont les endroits qui, justement, posent problème pour un non-natif).

Pour l'instant, le traitement d'une chanson prend environ 12 heures (si l'on compte le travail des étudiants puis de l'équipe de recherche) mais, une fois que le « noyau dur » du familier-argotique sera traité, nous pensons que les lexèmes qui circulent activement seront de moins en moins nombreux (un seuil de saturation du lexique va être atteint) et que nous trouverons de plus en plus de néologismes formels et sémantiques. Plus nous arriverons à étoffer le nouveau dictionnaire, plus le tagger pourra être entraîné avec nos données et pourra être applicable sur d'autres corpus du type WaC.

Bibliographie

COLIN et al. (1992), Dictionnaire de l'argot, Paris, Larousse

DEBOV Valéry (2012), *Diko des rimes en verlan dans le rap français,* Paris, La Maison du dictionnaire.

DUPUIS Fernande, KAPITAN Robert, DAOUST François (2010), « Expérience d'entraînement de TreeTagger et d'intégration à l'interface Web de SATO », JADT 2010, 10èmes Journées internationales d'Analyse statistique des Données Textuelles, Lexicometrica, Université Paris 3, p. 1013-1020.

FIÉVET Anne-Caroline, PODHORNÁ-POLICKÁ Alena (2010), « Argot des jeunes et français contemporain des cités en didactique du FLE/S : motivations des jeunes et limites des dictionnaires pour une étude des divergences socioculturelles », in : *Les Voix des Français* (M. Abecassis, G. Ledegen éds.), Volume 1, Bern, Peter Lang, p. 159-174.

GENDRON Catherine (2017), « Rap et slam : réflexions sur des pratiques urbaines et des jeunes », *Fictions, école et société,* MUNAÉ, Déc. 2017, Rouen, France.

GHIO Bettina (2012), *Le rap français: désirs et effets d'inscription littéraire*. Thèse sous la direction de Bruno Blanckeman, Paris, Université Sorbonne Nouvelle – Paris 3.

GOUDAILLIER Jean-Pierre (1997, 1998, 2001, 2019), *Comment tu tchatches ! Dictionnaire du français contemporain des cités,* Paris, Maisonneuve & Larose.

HAMMOU Karim (2012), *Une histoire du rap en France*, Paris, La Découverte.

JAKUBÍČEK Miloš, KOVÁŘ Vojtěch, RYCHLÝ Pavel et Vít SUCHOMEL (2020), « Current Challenges in Web Corpus Building », in : *Proceedings of the 12th Web as Corpus Workshop* (A. Barbaresi, F. Bildhauer, R. Schafer, E. Stemle éds.). Marseille, France, European Language Resources Association, p. 1-4.

NAUMANN Simon (2020), *Plurilinguisme dans le rap francophone. Passerelle mutuelle entre rap francophone et langue française*, Mémoire de master sous la direction de Claire Lesacher et Gudrun Ledegen. Rennes, Université de Rennes.

PODHORNÁ-POLICKÁ Alena, FIÉVET Anne-Caroline (2013), « Le rap en tant que vecteur des innovations lexicales : circulation médiatique et comportement des locuteurs », in : *Écarts et apports des médias francophones, Lexique et grammaire* (M. Abecassis, G. Ledegen éds.), Bern, Peter Lang, p. 113-139.

PODHORNÁ-POLICKÁ Alena, FIÉVET Anne-Caroline (2022), « Comment les différents types de corpus linguistiques éclairent (ou non) les différents types du lexique substandard : analyse contrastive à partir du vocabulaire de la comédie « Les Kaïra », exemple typique du genre filmique dit « de banlieue » », *Jazykovedný časopis*, sous presse.

PODHORNÁ-POLICKÁ Alena (2020), « RapCor, Francophone Rap Songs Text Corpus », in : *Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2020,* Brno, Tribun EU, p. 95-102.

POUDAT Céline (2009), « Etiqueteurs morphosyntaxiques. Présentation détaillée de quatre taggers et de leur fonction d'entraînement », *Texto!*, volume XIV-2 [en ligne]. Disponible sur : http://www.revue-texto.net/index.php?id=2293.

SAKANO FERNANDES Helena Yuriko (2013), « Hip Hop au cours de FLE ? En quoi le rap peut-il intéresser l'enseignement-apprentissage des langues-cultures », *Synergies Brésil,* 11, p. 141-149.

STEIN Achim, SCHMID Helmut (1995), « Étiquetage morphologique de textes français avec un arbre de décisions », *Traitement automatique des langues*, 36, 1-2 (Traitements probabilistes et corpus), p. 23-35.

TENGOUR Abdelkarim (2013), *Tout l'argot des banlieues. Le dictionnaire de la zone en 2 600 définitions,* Paris, Les éditions de l'Opportun.

TRIMAILLE Cyril (1999), Le rap français ou la différence mise en langues. *LIDIL*. Grenoble, Lidilem, n°19, p. 80–98.

https://is.muni.cz/do/phil/Pracoviste/URJL/rapcor/index.html,

https://www.sketchengine.eu/

GIMP. GNU Image Manipulation Program. https://www.gimp.org/

ALENA PODHORNÁ-POLICKÁ

Université Masaryk de Brno
Courriel : podhorna@phil.muni.cz

LAURENT CANAL

Université Masaryk de Brno
Courriel : 233716@muni.cz

ADÉLAÏDE EVREINOFF

Université Masaryk de Brno
Courriel : evreinoff.adelaide@gmail.com

ANNE-CAROLINE FIÉVET

École des hautes études en sciences sociales, Paris
Courriel : acfievet@gmail.com