sciendo
DOI 10.2478/jazcas-2022-0017

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 »

ALENA PODHORNÁ-POLICKÁ[1] – ANNE-CAROLINE FIÉVET[2]
[1]Université Masaryk de Brno, Brno, Tchéquie
[2] L'École des hautes études en sciences sociales, Paris, France

PODHORNÁ-POLICKÁ, Alena – FIÉVET, Anne-Caroline: How different types of linguistic corpora shed light (or not) on various categories of substandard lexicon: contrastive analysis of vocabulary in the comedy "Les Kaïra" [Porn in the hood], a typical example of the hood film genre. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 3, pp. 927 – 941.

Abstract: The arrival of WaC corpora, including Aranea family corpora, with its "close-to-spoken language" writings from different non-formal web pages brought the new options to researchers of sociolects, mainly to those who were previously obliged to observe youth collectives in its spontaneous discourses with its consequent time-consuming transcripts. Non-spontaneous spoken language from rap songs or youth film dialogues also help researchers to describe the level of societal diffusion of some typical features of youth slang. In this paper, we focus on demonstration of these crossed approaches in order to describe three types of verbs, used in a successful comedy about Parisian peri-urban post-adolescents *Les Kaïra* (2012), representing different types of substandard lexicon.

Key words: substandard verbs, French, neology, film dialogues, corpus linguistics, hood films

0. INTRODUCTION

La dynamique langagière est traitée abondamment par les linguistes, surtout les lexicologues, qui s'appuient sur des corpus textuels de plus en plus larges et de plus en plus accessibles. Jusqu'à récemment, ces corpus textuels avaient pour inconvénient de présenter essentiellement les néologismes des journalistes (cf. Sablayrolles, 2013 ; Cartier, 2019) et de faire l'impasse sur la néologie créée, promue et partagée notamment par la jeune génération. En effet, pour ce type de néologisme, seule l'observation sur le terrain - aussi bien des réseaux concrets (Popovičová Sedláčková, 2012 pour le slovaque, à titre d'exemple) que des réseaux virtuels (Chovancová, 2009 pour le français, entre autres) - permettait d'observer les tendances de cette dynamique inhérente à chaque langue vivante.

Cette dichotomie était encore plus prononcée dans les pratiques lexicographiques. En effet, cette deuxième catégorie de néologismes identitaires à un moment donné pour la jeune génération en entier ou pour une partie d'entre elle ne rentre pas facilement dans les dictionnaires généraux, elle est au contraire retraçable à partir des différents dictionnaires spécialisés de jeunes amateurs de langue. Une récupération dictionnairique officielle, si elle a lieu, s'opère avec un décalage temporel qui pourrait être expliqué, parmi d'autres facteurs, par l'instabilité sémantique de nombreux nouveaux items, plus particulièrement de ceux qui comportent une forte valeur expressive et identitaire pour les jeunes. Malgré le caractère instable de tous les aspects variationnistes (Gadet, 2003), d'où l'impression de « futilité » de toute entreprise lexicographique sur le sujet, nous partageons l'idée de Gaétane Dostie que « s'il faut choisir [...], il est préférable d'avoir une vue figée de ce qui bouge, que de n'avoir aucune vue sur le sujet » (Dostie, 2004, p. 192). C'est dans cette perspective qu'Alena Podhorná-Polická a lancé, en 2009, le projet d'un corpus de chansons de rap francophone, le *RapCor*. Ce dernier permet de compléter les résultats de recherches qualitatives et quantitatives sur le terrain et aide à reconstituer l'histoire des néologismes identitaires pour les jeunes à l'époque de leur promotion via le rap (voir Podhorná-Polická – Fiévet, 2013). La version la plus actuelle qui comprend les textes de 1288 chansons de rap est disponible sur la plateforme SketchEngine (pour une description plus détaillée de cette version intitulée *RapCor 1288*, voir Podhorná-Polická, 2020).

La situation a considérablement évolué depuis l'arrivée des corpus « big data », gratuitement accessibles et compilables de manière répétitive et à faible coût à partir des pages web (web as corpus ou WaC). L'apparition des corpus de la famille WaCky (Baroni et al., 2010), TenTen (cf. Jakubicek et al., 2013) et d'Aranea (Benko, 2014) a ouvert et démocratisé l'accès à des productions du « parlécrit »[1] français : le *frWaC* de 2010 a été rendu disponible en 2013 avec une taille de plus d'un milliard trois cent millions de mots. La même année arrive aussi la version française de la famille TenTen : le *FrTenTen12*, compilé du web francophone en 2012 avec presque 10 milliards de mots ; suivi par la version compilée en 2017, qui fait la moitié de la taille de la version de 2012 mais s'avère intéressante du point de vue de la synchronie dynamique. Et, enfin, en 2014, arrive la famille de corpus Aranea (Benko, 2014) qui a l'avantage de montrer les contrastes, sous *Araneum Francogallicum*, entre les différentes parties de la francophonie grâce aux cinq sous-corpus (*Gallicum, Belgicum, Canadiense, Helveticum* et *Africanum*). En effet, leur taille inégalable (pour la dernière version d'*Araneum Franco-*

[1] Étudiant les spécificités de la communication sur Minitel à l'époque, Anne-Marie Jeay a proposé un mot-valise néologique qui n'a rien perdu de son ergonomie et adéquation quant à la description de ce qui se passe sur les réseaux sociaux trente ans plus tard : « **Sur les messageries télématiques les individus sont censés dialoguer, mais en fait ils se 'parlent' par écrit.** Le langage y est donc un 'parlécrit' » (Jeay, 1991, p. 31). C'est l'auteure qui souligne.

gallicum Maximum, compilée en mai 2020, par exemple, il s'agit de 10,9 G, soit presque 11 milliards de positions (tokens) / 9,3 milliards de mots) et leur constitution extrêmement rapide en comparaison avec des corpus partant de bases de données de textes soumis aux droits d'auteurs (cf. Cvrček et al., 2020) fait de ce type d'outil une méthode puissante de vérification des données de terrain, désormais incontournable pour les chercheurs variationnistes. Si nous n'avons trouvé aucune attestation pour les argotismes plutôt « jeunes » tels que, à titre d'exemple, *keum* (verlan de *mec*, au sens de « garçon, (jeune) homme ») dans le corpus littéraire qui était un des rares disponibles gratuitement entre 2008 et 2013, à savoir le sous-corpus français du corpus parallèle *InterCorp* (créé dans le cadre du Corpus national tchèque), la situation s'améliore un peu quantitativement mais pas qualitativement à partir de sa version 7, lancée en 2014, où nous avons pu découvrir les 11 premières occurrences de *keum*, ceci dans une collection *Open subtitles*, formée de traductions de films par des amateurs inconnus que le Corpus national tchèque a décidé d'abriter dans le souci de diversifier les textes parallèles et d'augmenter leur taille (Nádvorníková – Vavřín, 2014). Dans le *frWac*, qui est abrité par le même Corpus national tchèque dès 2013 mais qui n'a attiré notre attention qu'en 2015, 53 occurrences de *keum* sont désormais disponibles. Mais il faut attendre l'intégration d'une des versions du corpus *Aranea Francogallicum* (AF) sous le même « toit » tchéco-slovaque, à savoir le *AF Maius* (de mars 2015 ; 1,2 G) pour que la véritable puissance des corpus de type WaC devienne évidente (avec 400 occurrences de *keum*), ouvrant ainsi la porte à la « big-data argotologie ».

Notre attention dans ce travail portera sur un film racontant l'histoire de trois *keums* d'une banlieue sud de Paris, *Les Kaïra*, sorti en 2012, année où les informaticiens de la Faculté d'informatique de l'Université Masaryk ont pu compiler le susmentionné *FrTenTen12* qui comporte 4 538 occurrences du lexème *keum*. Les nouvelles versions du corpus *TenTen* et d'*AF Maximum* (respectivement de 2017 et de 2020) montrent bien, si l'on observe la chute des chiffres de fréquence relative (i.p.m.), que cet argotisme circule actuellement moins dans le parlécrit aspiré sur la toile et serait donc moins répandu qu'en 2012.

1. VERS UN CORPUS PARALLÈLE DE DIALOGUES ET DE SOUS-TITRES

En parallèle du RapCor, de manière plus ponctuelle mais depuis plus longtemps (2007), les deux auteures constituent une base de données portant sur les films qui mettent en scène des jeunes, vivant ou non en banlieue. Nous préférons le terme anglais de « hood films » (cf. Mével, 2008) à ses équivalents français « cinéma de banlieue » (Grodner, 2020) ou encore « cinéma sur les jeunes de la banlieue » car ces derniers ont pour inconvénient de véhiculer des connotations négatives autour des termes « banlieue » et « jeune de banlieue ».

Notre but scientifique est d'observer la dynamique de diffusion et la variation diachronique du lexique substandard dans les dialogues de plusieurs « hood films » (Podhorná-Polická – Fiévet, 2008 ; Fiévet – Podhorná-Polická, 2020). Il est également nécessaire de prendre en compte les défis traductologiques et épistémologiques qui émergent quand on compare les dialogues avec les sous-titres intra-lingues (pour les sourds et malentendants, SM) et inter-lingues, ceci avec une question primordiale : les traducteurs vers les langues dites mineures telles que le tchèque, traduisent-ils à partir des dialogues ou se simplifient-ils la tâche en s'appuyant sur des sous-titres SM et, surtout, anglais, qui sont souvent disponibles avec la copie du film ?

C'est pour cette raison que notre base de données interne, comportant des transcriptions incomplètes de dialogues, a été récemment revue et retravaillée avec l'objectif de rendre disponible un corpus parallèle, formé : 1) de dialogues de « hood films » (un corpus similaire à *The Movie corpus* (Davies, 2019–) étant indisponible pour le français jusqu'à présent, nous nous inspirons du corpus de Dekhissi (2013), formé de transcriptions sélectives de 38 « films de banlieue » français parus entre 1984 et 2011) et 2) de sous-titres officiels de différents types (faisant ainsi écho au corpus de Open subtitles[2] qui a l'inconvénient de ne pas fournir les métadonnées sur les protagonistes du film et sur les sous-titreurs – amateurs de fan-subbing). Le résultat de ce travail sera disponible sous le nom de *HoodFilmCor* sur la plateforme Sketch Engine courant 2022. Afin de présenter ce nouveau corpus parallèle de plus près, nous montrons ci-dessous (voir Tableaux 1 et 2) les extraits de deux films qui vont y apparaître prioritairement, *Les Kaïra* (2012) et *Bande de filles* (2014). Il s'agit de deux tableaux, pour l'instant sous Excel, parce que la mise en page dans un gestionnaire de corpus avec balisage n'est pas encore achevée (en partenariat avec Pavel Rychlý, co-auteur de Sketch Engine, de la Faculté d'informatique de l'Université Masaryk). Le balisage consiste à donner plusieurs informations : sur quelle ligne apparaît le texte en question, qui énonce la réplique, de quelle couleur est le sous-titre et comment il est positionné (la couleur et la position jouent un rôle important dans les sous-titres SM surtout). Le corpus apporte également des informations sur le minutage des sous-titres (colonne de gauche), sur les métadonnées, sur les protagonistes du film (tranche d'âge, sexe, apparence, etc.) et, au niveau textuel, apporte des informations sur l'alignement des répliques (pour la transcription fidèle des énoncés) et au niveau des images pour les différentes versions des sous-titres disponibles (SM, anglais et d'autres langues qui nous intéressent, notamment le tchèque, le slovaque et l'allemand). Par exemple, pour le cas de *Bande de filles*, deux versions de sous-titres tchèques du même film, créés pour deux festivals distincts par deux traducteurs professionnels, nous permettent de faire l'analyse des procédés de traduction audiovisuels détaillés.

[2] Corpus disponible dans le cadre du corpus parallèle *InterCorp* dès sa version 7 (publiée le 19 décembre 2014).

Tab. 1. Les Kaïra (extrait)

Minutage	Sous-titres EN	Sous-titres SM
		<\V><1>Musique classique</1></\V>
		<g><\V><1>...</1></\V></g>
		<R><1>Le vinyle grésille.</1></R>
		<\V><1>Musique hip hop</1></\V>
		<g><\V><1>...</1></\V></g>
0:00:28	<1>Listen, imagine</1><2>that if you could bang</2>	<J><1>- Regarde.</1><2>Imagine que si tu peux ken</2></J>
0:00:31	<1>Beyonce or Shakira,</1>	
0:00:33	<1>who's a better bang?</1>	<J><1>ou Beyoncé ou Shakira, laquelle</1><2>des deux tu préfères le plus la ken ?</2></J>
0:00:35	<1>Stop with the lame questions.</1><2>Jackass.</2>	<J><1>-Vas-y, dégage,</1><2>avec tes questions en carton.</2></J>
0:00:38	<1>Bone any babes lately?</1>	<J><1>- Depuis quand t'as pas niqué ?</1></J>
0:00:42	<1>You serious?</1><2>You think I don't fuck?</2>	<g><B><1>-T'es sérieux,</1><2>tu crois que je ken pas ?</2></B></g>
0:00:44	<1>I fuck every babe I want.</1><2>I know tons.</2>	<g><B><1>Moi je ken toutes les meufs</1><2>que je veux.</2></B></g>
0:00:47	<1>It's easy. They dig me.</1>	<g><B><1>J'ai pas de problème,</1><2>elles me kiffent. C'est toi qui ken pas.</2></B></g>
0:00:49	<1>You can't get laid, loser.</1>	
0:00:51	<1>Why are you lying, bro?</1><2></2>	<B><Abd><1>Il est là le doss.</1></Abd><Mou><2>-Pourquoi tu mens ?</2></Mou></B>
0:00:53	<1>We're always together.</1><2>If you fucked, I'd	<B><1>On est tous les jours ensemble,</1><2>je t'aurais cramé !</2></B>

Sous-titres SM	Qui parle FR	Transcription fidèle (par répliques)
<\V><1>Musique classique</1></\V>		
<g><\V><1>...</1></\V></g>		
<R><1>Le vinyle grésille.</1></R>		
<\V><1>Musique hip hop</1></\V>		
<g><\V><1>...</1></\V></g>		
<J><1>- Regarde.</1><2>Imagine que si tu peux ken</2></J>	Mousten	Tiens, regarde, imagine que si tu peux ken ou Beyoncé ou Shakira, laquelle des deux tu préfères le plus la ken ?
<J><1>ou Beyoncé ou Shakira, laquelle</1><2>des deux tu préfères le plus la ken ?</2></J>	Mousten	
<J><1>-Vas-y, dégage,</1><2>avec tes questions en carton.</2></J>	Abdelkrim	Vas-y, dégage avec tes questions en carton là. Super chelou, c'keum.
<J><1>- Depuis quand t'as pas niqué ?</1></J>	Mousten	Genre toi, ça fait combien d'temps qu't'as pas niqué une meuf, toi ?
<g><B><1>-T'es sérieux,</1><2>tu crois que je ken pas ?</2></B></g>	Abdelkrim	Eh, t'es sérieux là. Et mais, regarde-moi bien, tu crois qu'j'ken pas, moi ? Moi, j'ken toutes les meufs que j'veux. J'en connais trop des meufs. Moi j'ai pas de problème avec ça, les meufs, elles m'kiffent, mon gars. C'est toi l'chien d'la casse, qui ken pas. Il est là, l'doss.
<g><B><1>Moi je ken toutes les meufs</1><2>que je veux.</2></B></g>	Abdelkrim	
<g><B><1>J'ai pas de problème,</1><2>elles me kiffent. C'est toi qui ken pas.</2></B></g>	Abdelkrim	
<B><Abd><1>Il est là le doss.</1></Abd><Mou><2>-Pourquoi tu mens ?</2></Mou></B>	Abdelkrim, puis Mousten	Et mais, pourquoi tu mens, mon frère ? Eh, on est tous les jours ensemble. Si tu niquais des meufs, j't'aurais d'jà cramé, sale mytho.
<B><1>On est tous les jours ensemble,</1><2>je t'aurais cramé !</2></B>	Mousten	

Tab. 2. Bande de filles (extrait)

Minutage	Sous-titres EN	Sous-titres CZ1(JM)	Sous-titres CZ2
0:15:26	<B><1><Lad>- Always with your crew.<Lad></1><2><Abd>- You too.</Abd></2></B>	partě?</Abd></1><2><Lad>- Ty taky.</Lad></2></B>	<B><1><Abd>- Vždycky s partou.</Abd></1><2><Lad>- Ty taky.</Lad></2></B>
0:15:29	<B><1><Adi>- Right.</Adi></1><2><Lad>- Always together.</Lad></2></B>	<B><1><Lad>- Jo.</Lad></1><2><Adi>- Pořád spolu.</Adi></2></B>	<B><1><Lad>- Jasně.</Lad></1><2><Adi>- Vždycky spolu.</Adi></2></B>
0:15:32	<1><B>I heard you clashed with some girls.</B></1>	<B><1>Prej ste se srazily</1><2>s ňákejma holkama.</2></B>	<B><1>Slyšel jsem, že ste se pohádaly.</1><2>s nějakejma holkama.</2></B>
0:15:34	<1><B>Bullshit!</B></1>	<1><B>Kecy!</B></1>	<1><B>Kecy!</B></1>
0:15:37	<1><B>That's what they said.</B></1>	<1><B>Říkaly to.</B></1>	<B><1><Abd>- Říkaly to.</Abd></1><2><Lad>- Počkej.</Lad></2></B>
0:15:39	<1><B>Hold on a second.</B></1>	<1><B>Tak počkej.</B></1>	
0:15:40	<1><B>What did they say exactly?</B></1>	<1><B>Co přesně říkaly?</B></1>	<1><B>Co přesně říkaly?</B></1>
0:15:42	<1><B>That you guys chickened out.</B></1>	<1><B>Že jste zdrhly</B></1>	<1><B>Že jste zbabělci.</B></1>
0:15:44	<1><B>Chickened out?</B></1>	<1><B>Zdrhly?</B></1>	<1><B>Zbabělci?</B></1>
0:15:45	<1><B>We chickened out?!</B></1>	<1><B>My a zdrhnout?</B></1>	<1><B>My a zbabělci?</B></1>
0:15:48	<1><B>We'll waste them any time.</B></1>	<1><B>Rozmáznem je, kdykoli se nám zachce.</B></1>	<1><B>Zničíme je, kdykoli se nám zachce.</B></1>
0:15:51	<B><1>You got that from the chick.</1><2>always doing selfies?</2></B>	<B><1>To máš od tý kozy,</1><2>co furt fotí selfíčka?</2></B>	<B><1>To máš od tý holky,</1><2>co si pořád fotí selfie?</2></B>
0:15:54	<1><B>Doing duck-face in her bathroom.</B></1>	<B><1>Co špulí pusu v koupelně</1><2>jak topmodelka?</2></B>	<B><1>Jak na sebe neustále</1><2>špulí pusu v koupelně?</2></B>

Sous-titres CZ2	Sous-titres SM	Qui parle FR	Transcription fidèle (par répliques)
<B><1><Abd>- Vždycky s partou.</Abd></1><2><Lad>- Ty taky.</Lad></2></B>	<1><d><J>-Toujours en équipe.</J></d></1>	filles	Toujours en équipe, ce aue je vois.
	<1><ddc><B>-Vous aussi.</B></ddc></1>	Abdel	
<B><1><Lad>- Jasně.</Lad></1><2><Adi>- Vždycky spolu.</Adi></2></B>	<B><Lad><1><gcc><h>-On est là.</h></gcc></1></Lad><Adi><2><d>-Toujours ensemble.</d></2></Adi></B>	Lady, puis Adiatou	
<B><1>Slyšel jsem, že ste se pohádaly.</1><2>s nějakejma holkama.</2></B>	<1><g><J>-Ça a parlé sur vous.</J></g></1>	Abdel	Ça a parlé sur vous là si vous êtes embrouillé avec des filles.
	<1><Abd><dcc><h><B>-Ons embrouillé avec des filles.</B></h></dcc></Abd></1><2><Lad><d><J>-N'importe quoi !</J></d></Lad></2>		
<1><B>Kecy!</B></1>			
<B><1><Abd>- Říkaly to.</Abd></1><2><Lad>- Počkej.</Lad></2></B>	<J><g><1><Abd><h>-Ils ont pas dit ça ?</h></Abd></1><2><Lad>-Attends.</Lad></2></g></J>	Abdel, puis Lady	Wallah, ils ont pas dit ça ? Attends, attends
<1><B>Co přesně říkaly?</B></1>	<1><gcc><B>Elles ont dit quoi ?</B></gcc></1>	Lady	Elles ont dit quoi éxactement?
<1><B>Že jste zbabělci.</B></1>	<1><g><J>-Elles ont dit que</J></g></1> <1><gc><B>-vous avez peur d'elles.</B></gc></1>	Abdel	Moi j'étais pas là, mais elles ont dit que vous étiez des lâches et vous avez peur d'elles.
<1><B>Zbabělci?</B></1>	<B><1><Lad><dcc><h>-Genre, on a peur d'elles.</h></dcc></Lad></1><2><Adi><d>-	Lady, puis	Comment ça, des lâches ?
<1><B>My a zbabělci?</B></1>	Des lâches ?</d></Adi></2></B>	Adiatou	Genre. Genre, on a peur d'elles.
<1><B>Zničíme je, kdykoli se nám zachce.</B></1>	<1><B>-On va les exploser.</B></1>		Mais on les prend quand elles veulent, on va les exploser.
<B><1>To máš od tý holky,</1><2>co si pořád fotí selfie?</2></B>	<g><J><1>C'est pas celle qui se prend en photo</1><2>dans sa salle de bains ?</2></J></g>		Ça c'est pas un coup de la meuf là qui se prend toujours en photo dans sa salle de bains là ?
<B><1>Jak na sebe neustále</1><2>špulí pusu v koupelně?</2></B>	<1><g><J>Avec sa bouche de canard, là.</J></g></1>	Lady	Genre top-modèle avec la bouche de canard là

2. MÉTHODES CROISÉES POUR MIEUX CIRCONSCRIRE LA DIFFUSION DU LEXIQUE EXPRESSIF POUR LES JEUNES

Depuis les débuts de nos travaux sur les corpus filmiques, nous avons appliqué la méthode des filtres successifs qui consiste à rechercher les lexèmes dans des dictionnaires, du plus standard au plus argotique. Cette méthodologie a été utilisée dans le cadre de tous nos travaux mentionnés *supra*. De plus, depuis l'arrivée des WaC, les mots sont également recherchés dans les différents corpus en ligne, ce qui permet de valider ou d'invalider certaines hypothèses sur leur circulation. Il s'agira ici d'approfondir cette méthodologie que nous avons pour la première fois appliquée sur le corpus du film de Franck Gastambide sorti en 2012, *Les Kaïra* (Fiévet – Podhorná-Polická, 2020). *Kaïra* est le verlan de *racaille* (« personne peu recommandable » (PR), aujourd'hui plutôt avec le sens en usage de « délinquant juvénile » (*Dictionnaire de la zone*, DZ)), mais tandis que *racaille* est négativement connoté, *kaïra* (orthographié aussi *kaillera* ou *caillera*) apporte plutôt un effet laudatif, voire comique. Le film a connu un grand succès au cinéma lors de sa sortie en salles en 2012 avec 1 million d'entrées. *Les Kaïra* raconte l'histoire de trois amis, jeunes adultes d'une cité de la banlieue parisienne, plus exactement Melun en Seine-et-Marne. Dans l'espoir de faire des conquêtes féminines, ils vont essayer de percer dans le milieu du cinéma pornographique. Une grande partie du film retrace leurs tentatives de se procurer une sex-tape démo pour un producteur de films X, tentatives qui finissent la plupart du temps par des échecs.

Puisque le film *Les Kaïra* parle très ouvertement des relations entre filles et garçons, ceci nous a amenées, dans cette recherche précédente (Fiévet – Podhorná-Polická, 2020), à analyser ce qu'on peut appeler la dragolalie et la pornolalie : en effet, plus de 50 mots et expressions différents, pour plus de 150 occurrences, ont été relevés concernant les thématiques de la drague et de la sexualité. Nous appuyant sur *Le Petit Robert* (PR) comme dictionnaire d'exclusion (considérant que les mots qui y sont répertoriés avant la sortie du film sont déjà très connus, nous avons plus exactement exclu les mots présents dans le PR jusqu'en 2011, considérée comme l'année de tournage du film), nous avons analysé en détail les 81 occurrences restantes (31 lexèmes). Parmi eux, six expressions ont été ajoutées au PR après 2011 (comme *mettre/se prendre un vent* (PR2012), *pécho* (PR2015) ou *être en chien* (PR2017) pour les dragolaliques et *avoir la gaule* (PR2012), *film de boule* (PR2015) ou *défoncer* (PR2019) pour les pornolaliques). La plupart des lexèmes ont été trouvés dans les dictionnaires d'argot (19 sur 31, surtout dans le DZ) mais les six restants n'ont été trouvés nulle part (*dragon de Komodo, chacaler, surcheum* et *taper un blocage sur qqn* pour les dragolaliques et *anaconda* et *poutre de Bamako* pour les pornolaliques).

Ainsi, les résultats ont pu mettre au jour plusieurs niveaux de circulation, du plus évident (le lexème relevé est présent dans tous les dictionnaires d'argot des

jeunes consultés) au plus énigmatique (le lexème relevé dans le film n'est présent nulle part). Afin d'étudier notre hypothèse que le recours aux grands corpus web peut nous donner des indications supplémentaires sur ces différents niveaux de circulation ou encore spécifier les nuances sémantiques que les dialogues sous-entendent mais que les dictionnaires ne notent pas, nous avons décidé, pour cet article, de sélectionner trois verbes dont la circulation est *a priori* différente :

- un lexème à faible circulation, qu'on trouve dans aucun dictionnaire d'argot (pas d'indication sémantique, pas d'indication sur sa circulation et sur la période) : ***chacaler***
- deux lexèmes à large circulation, qui comportent plusieurs notations graphiques avec une faible dictionnairisation officielle: ***ken*** pour les pornolaliques et ***choper/pécho*** pour les dragolaliques (le verbe *choper* et sa verlanisation *pécho*, sachant qu'il existe des nuances sémantiques entre les deux).

2.1 Le lexème *chacaler* : hapax ou argotisme à définir ?

Dans le corpus de dialogues du film *Les Kaïra*, nous avons pu relever deux occurrences du verbe *chacaler*. La première est prononcée dans une phrase énoncée par un des trois principaux personnages qui s'appelle Mousten et qui est incarné par le scénariste Franck Gastambide (« et qu'y'avait qu'une seule meuf que tout l'monde a chacalée », 16.49) et la deuxième est prononcée par son acolyte, Abdelkrim (« J'ai vu quand t'es allé la chacaler à côté des chiottes » (36.04). Le lexème *chacaler* est intéressant à observer de plus près puisqu'il n'est présent dans aucun des dictionnaires consultés (PR, AFP, CTT, DZ, BOQ – voir la liste *infra*). On trouve seulement le lexème *chacal* dans le dictionnaire papier d'argot commun des jeunes, *Lexik des cités* (2007, p. 99, LC ; « 1) radin. Synonyme : crevard, pince 2) avide. Synonyme : crevard ») et dans le dictionnaire en ligne collaboratif « Wiktionnaire »[3] (« personne sournoise et opportuniste »).

Quant au lemme *chacaler* dans les différents corpus disponibles cités *supra*, il n'est présent ni dans les corpus WaC, ni dans le corpus RapCor1288. En revanche, une recherche plus complexe dans l'AF Maximum (AFM) grâce à la requête [lemma="chacal.*"&tag="V.*"] permet d'obtenir un seul résultat correspondant au sémantisme relevé dans le film : « Moi je refuse de penser que ce sont des gens frustrés et en manque de sensations sexuelles qui la ***chacalent*** » (saisi en 2015 à partir du site https://ntrjack.mondoblog.org/2013/08/29/les-camerounaises-sont-belles/). Cet exemple unique apporte pourtant des informations intéressantes : d'une part que le mot est bien annoté syntaxiquement mais mal lemmatisé (lemme *chacalent*) puisque les concepteurs de corpus WaC en général, y compris Benko (auteur d'AFM), préfèrent prendre les formes des mots pour lemmes si ces formes manquent dans le dictionnaire interne du *Tree tagger* (logiciel gratuit largement utilisé pour

[3] https://fr.wiktionary.org/wiki/chacal

l'annotation morphosyntaxique et la lemmatisation du français, cf. Stein et Schmid, 1995) à la place du « none » (absence de lemme) que propose *TreeTagger* dans de pareils cas. D'autre part, du fait que nous n'avions jamais entendu ce verbe en dehors de ce contexte filmique dans des discussions entre jeunes, cet exemple nous permet de répondre négativement à notre questionnement : est-ce que, en 2012, *chacaler* était un lexème identitaire pour le groupe de pairs autour du scénariste (voire sa création idiolectale), que ce dernier aurait essayé de faire connaître plus largement en le faisant répéter dans les dialogues ?

Dans les autres corpus WaC interrogés, à savoir *frWaC, FrTenTen12* et *FrTenTen17*, il est possible de trouver une série dérivationnelle basée sur ce verbe : *chacaliser, chacalisoter*, ce qui montre la vivacité de la métaphore animalière, importée fort probablement de l'habitat traditionnel des chacals au Maghreb, dans son acception verbale. Les informations recueillies autour de ce lexème nous permettent ainsi de proposer une définition de *chacaler* : « draguer avec avidité, de façon opportuniste (en approchant sa « proie », sans y mettre les formes) ».

2.2 Le lexème *choper* et sa verlanisation *pécho* : changements formels et sémantiques

Le verbe *pécho* a été choisi à partir de la liste des dragolaliques comme un exemple prototypique d'une circulation large dans le français hexagonal d'aujourd'hui mais aussi d'une dictionnarisation problématique qui se reflète dans le traitement de ces lemmes dans les grands corpus. Résultat d'une métathèse régulière sur le mot bisyllabique *choper* (absent des dialogues du film étudié), *pécho* a été énoncé trois fois dans *Les Kaïra*, chaque fois par Mousten dont deux fois dans la locution *pécho des meufs* (« on va pécho les meufs », 24.48 et « où est-ce qu'on va pouvoir pécho des meufs », 59.27) et la dernière fois, juste après cette dernière scène, dans le sens de « se procurer ; voler » (« tu lui demandes qu'il pécho l'enveloppe », 59.57). La polysémie de *pécho* reflétant la polysémie de son « verbe-miroir » *choper*, c'est paradoxalement seulement le sens de « draguer, séduire » qui est répertorié dans l'ouvrage lexicographique de référence, le PR. Au fait, il faut le chercher sous l'entrée *choper* qui comporte la marque lexicographique FAM. (« familier ») pour toutes ces acceptions dont les trois premières, anciennes : 1) Voler (vieilli) ; 2) Arrêter, prendre (qqn) ; 3) Attraper. La quatrième, « Parvenir à séduire qqn », toujours marquée comme FAM., n'est ajoutée qu'à partir de l'édition 2015, et c'est justement sous cette acception du champ sémantique de la dragolalie qu'apparaît *pécho* sous une entrée cachée, avec une limitation (erronée comme en témoignent les occurrences du film ainsi que de nombreux exemples dans les corpus WaC) de son emploi uniquement en tant que participe passé[4]. Absent du dictionnaire de référence en matière de français substandard, l'Argot &

[4] « Parvenir à séduire (qqn). ABSOLT *Il a chopé !* - VERLAN au p.p. *pécho*. 'le fils d'un chanteur très connu que j'ai pécho' (L. Pille) ».

français populaire (AFP), *pécho* figure, en revanche, dans le DZ, dictionnaire de référence pour l'argot commun des jeunes, en ligne, comme entrée distincte et autonome par rapport à *choper*. Il est à noter qu'ici, la suite d'acceptions légèrement différente témoigne d'une spécialisation de *pécho* par rapport à *choper.* Pour ce dernier, le lien paronymique avec *chiper* (« voler »[5]) est plus marquant mais, à la différence de nos observations sur l'usage du verbe *pécho* dans les corpus WaC, la prédominance du sens « draguer, séduire » pour le verbe verlanisé n'est pas pris en compte dans l'ordre des acceptions. De la même manière que pour le rapport *racaille – kaïra* cité *supra*, on assiste ici aussi à un effet d'allégement du poids référentiel que constatait Goudaillier en 2002 déjà : « le verlan est une pratique langagière qui vise à établir une distanciation effective par rapport à la dure réalité du quotidien […]. Le lien au référent serait plus lâche et la prégnance de celui-ci moins forte, lorsque le signifiant est inversé, verlanisé » (2002, p. 18).

Il est intéressant de noter que le dictionnaire *Comment tu tchatches !* de Goudaillier (CTT) a introduit, dès sa première édition parue en 1997, la variante phonétique *peucho* [pøʃo] juste en dessous de l'entrée *pécho* [peʃo]. Dans le DZ, en revanche, l'entrée *pécho* renvoie encore à sa variante où la fin du mot subissait une reverlanisation monosyllabique [peoʃ], orthographiée *péauche* et *péoch.* Les corpus du type WaC peuvent venir sur ce point éclairer le niveau de circulation de ces variantes phonétiques (et graphiques). Comme en témoignent les résultats du Tableau 3, les trois variantes n'ont été trouvées nulle part, ce qui renforce notre hypothèse qu'il s'agit de variantes créées et diffusées de manière locale qui tendent à disparaître avec la perte d'expressivité du verbe *pécho*.

La variation graphique reste néanmoins une caractéristique typique des créations *a priori* orales qui surgissent à partir des interactions dans un groupe en quête d'identité (générationnelle, socio-ethno-géographique ou autre). Quant à *pécho*, sa graphie a été longtemps variable (les mots verlanisés étant transcrits soit avec un tiret, soit sans, en gardant le digramme *-er* de l'infinitif ou non, avec une hésitation pour l'accent au-dessus du *e*, etc.), mais comme pour d'autres verlanisations à haute circulation (*tess* pour *cité*, par exemple ou *ken*, voir *infra*), elle tend à se stabiliser dans le parlécrit des usagers des réseaux sociaux, tout en respectant les règles de l'économie. Le fait de faire entrer la graphie *pécho* dans le PR ne fera certainement qu'accélérer cette stabilisation. Pour faire ressortir l'oscillation graphique telle que nous avons pu la voir dans nos questionnaires auprès de jeunes dans nos travaux précédents, une série d'autres graphies (*pecho, per-cho, pé-cho, pêcho*) a été évaluée dans le Tableau n°3.

Même pour le verbe *choper*, dont l'orthographe semble être stabilisée, à en croire le PR, l'oscillation entre un ou deux P (*choper* ou *chopper*) est lisible à partir des citations données dans l'AFP, par exemple. Ce phénomène, ainsi que l'homony-

[5] Cf. article *choper* dans le PR et celui d'AFP, où il serait dérivé de *coper* (« faire un faux pas »).

mie du verbe *chopper* avec le déonyme *chopper* (anglicisme, « moto de sport avec les guidons très haut », prononcé et souvent aussi orthographié comme *choppeur*) est reflété également dans le Tableau 3.

	frWaC (2010)	Araneum Francogallicum III – Maximum (2013-2019) AFMaxi	French Web 2012 (FrTenTen12)	French Web 2017 (FrTenTen17)	RapCor1288 (2020)
DOCs	2 268 304	17 767 539	20 400 411	14 088 683	1288
TOKENS	1 613 814 684	10 881 222 203	11 444 973 582	6 845 630 573	767 483
MOTS	5 911 017	9 327 453 482	9 889 689 889	5 752 261 039	709 057
choper	2 243	25 893	41 490	11 858	11
chopper	1 245 (dont 1075 pour verbe)	17 336	30 512	5 861	3
pécho	128	2 013	1 697	1 203	6
pecho	95	547	310	185	0
pêcho	6	100	0	52	0
pé-cho	1	14	0	10	0
per-cho	0	8	2	3	0
peucho, péoch, péauche	0	0	0	0	0

Tab. 3. Résultats des requêtes pour lemmes *choper, pécho* et leurs variantes graphiques dans le *frWaC*, *AFM, FrTenTen12, FrTenTen17* et dans le *RapCor*1288.

Le Tableau 3 apporte un témoignage intéressant de la distribution de différentes variantes graphiques pour les verbes *choper* et *pécho* dans différents corpus WaC (frWaC de 2010, AFM de 2013-2019, les deux FrTenTen de 2012 et de 2017) et dans un corpus annoté manuellement, le RapCor. La requête simple permet de regrouper la flexion verbale pour *choper* mais le taggeur automatique n'arrive pas à gérer sans faute l'homonymie entre *chopper* verbe et *chopper* nom, ce qui est sous-entendu pour les corpus de cette taille. Or, pour *pécho*, invariable au niveau des désinences mais variable au niveau de ses éléments vocaliques, la lemmatisation automatique est extrêmement erronée. Si l'on prend pour exemple les 128 occurrences du frWaC pour *pécho* (requête simple, sans sensitivité à la majuscule), 5 d'entre eux ont pour lemme « Pécho » puisque la majuscule a fait que le TreeTagger l'a interprété comme un nom propre. Or, lorsqu'on regarde la distribution de 128 occurrences de *pécho* en catégories grammaticales, on a affaire à seulement 51 verbes (soit un taux de 39,8 % d'annotation correcte), les autres cas ont été interprétés comme NOM (nom commun ; 45,3 %), ADJ (adjectif ; 2,3 %) ou encore comme NAM (nom propre ; d'autres

11 cas ayant la minuscule, soit 12,5 %). Les défauts d'annotation automatique par TreeTagger se font encore plus sentir sur les 2 013 résultats pour *pécho* dans l'AFM où les tags comportent 0 verbe, 1 650 noms, 302 adjectifs et 57 noms propres – le taux d'annotation correcte est alors de 0 %. Quant au FrTenTen17, annoté par un autre outil, le FreeLing, la totalité des occurrences de *pécho* est annotée comme nom (propre ou commun, en fonction de l'initiale).

Le *RapCor* a été ajouté aux côtés de ces grands corpus web pour deux raisons : d'une part pour montrer que la lemmatisation y est semi-automatique et les variantes graphiques sont prises en considération (le résultat de TreeTagger est revu chanson par chanson et les variantes graphiques sont intégrées dans notre dictionnaire interne sous le même lemme fédérateur afin de simplifier les requêtes dans un avenir proche). D'autre part, si l'on regarde la fréquence relative de *pécho* par rapport à la taille du corpus, on constate une spécialisation du RapCor par rapport au français substandard (frWaC – 0,08 ; AFM – 0,22 ; FrTenTen12 – 0,17, FrTenTen17 – 0,21 et RapCor – 7,96). Pour *choper*, ce rapport (les argotismes y sont 7 fois plus fréquents) se confirme.

De plus, toujours en chiffres relatifs (i.p.m.), il est intéressant du point de vue de la synchronie dynamique de comparer FrTenTen12 (2012 : année de sortie du film) d'un côté et AFM et FrTenTen17 (postérieur au film) de l'autre, pour constater que la circulation est en baisse pour *choper* et en légère hausse pour *pécho*.

2.3 Le lexème *ken* : un usage ergonomique mais une lexicographie complexe

Notre troisième exemple tiré du film *Les Kaïra* est le verbe le plus fréquent de la catégorie des pornolaliques. La verlanisation apocopée formée à partir de l'arabisme entièrement intégré au français, *niquer* [nike] > *[keni] > [kɛn], *ken* est apparue 25 fois dans les dialogues du film, dans le sens de « posséder sexuellement ».

La recherche dans les différents corpus nous sera utile, ici aussi, pour mieux circonscrire : 1) la distribution réelle des variantes graphiques qui ont été notées dans les différents dictionnaires d'argot et 2) le problème de la lemmatisation défaillante des verbes invariables dans les corpus annotés automatiquement.

De même que pour l'homographie entre le verbe et le nom *chopper*, la graphie la plus fréquente pour la suite phonique [kɛn], à savoir *ken*, apporte une homographie avec le nom propre d'origine celte, assez fréquent, à savoir Ken. Cette forme abrégée du prénom masculin Kenneth est connue surtout comme compagnon de la poupée Barbie. Dans le discours (entre filles notamment), il peut fonctionner également comme déonyme pour désigner un synonyme de *beau gosse* (« bel homme/garçon »), où l'initiale peut s'orthographier avec une minuscule. La négligence du style, typique pour le parlécrit des sms d'abord, puis pour le parlécrit sur les réseaux sociaux, et très présente dans les corpus web, complique les requêtes. Par exemple, parmi les 32 706 occurrences de *ken* (requête simple) dans l'AFM, on peut trouver, en dehors des emplois nominaux susmentionnés, à la fois les formes phonétisantes de *qu'en = ken*, les noms de produits divers importés du Japon (*ken* étant un homonyme très fréquent en japo-

nais, ayant une trentaine de sens différents dont 7 extrêmement fréquents, p. ex. le fameux manga *Hotaru no ken*, le nom du département + *ken* « département », etc.). Avec la lemmatisation erronée (0 % de verbes, comme *supra* pour *pécho*), il est difficile de trouver des occurrences de *ken* avec le sens de « posséder sexuellement » ou « abîmer, détruire ». En utilisant la requête CQL complexe : [tag="PRO.*"][word="ken"] qui donne 83 occurrences, on enlève à la fois les prénoms Ken mais on limite la recherche uniquement aux pronoms dont certains renvoient à la personne qui fait l'action (*je ken, j'ken, on ken*) et d'autres renvoient toujours au Ken (déonyme ou prénom) : *ce ken*, *moi ken*. D'autres contextes verbaux peuvent être retrouvés avec la requête [lemma="avoir|faire"] [word="ken"] – 71 occurrences, notamment les participiales (*je l'ai ken, on a ken*) et les pronominaux passifs (*se faire ken*).

Le chercheur peut alors accéder (même si c'est de façon peu aisée) aux attestations de circulation fréquente de ce verbe et apprendre, par une requête rapide, que *ken* est la variante graphique privilégiée par les scripteurs sur la toile : dans l'AFM, *kène* renvoie au verbe étudié moins de 30 fois parmi les nombreux africanismes orthographiés ainsi, *kèn* une seule fois, et aucune preuve n'est trouvée de l'utilisation des orthographes *quène*, *kéne* ou *kén*. Or, si un étudiant en FLE ou un traducteur sont à la recherche de ce verbe dans les dictionnaires, ils ne trouveront *ken* ni dans le PR, ni dans l'AFP. Pour le CTT, ils devront aller le chercher sous la lettre Q (entrée *quène*). En ce qui concerne d'autres dictionnaires d'argot, seul *Bien ou quoi* (BOQ) privilégie la graphie *ken*. Quant au DZ en ligne, l'entrée principale y est bizarrement *kéner*, avec la remarque « s'emploie aussi sous sa forme invariable *kène* ». L'adverbe « aussi » est inapproprié parce que nous avons observé l'emploi exclusif de cette forme invariable, aussi bien dans les corpus web que dans le RapCor (23 occurrences dont une seule avec la graphie originelle du rappeur Ol'Kainry « elles kennent toutes », chanson *En attendant* de 2001, qui ne prouve pourtant pas l'existence de la forme régularisée en *-er* à l'infinitif). Pour être précis, l'AFM apporte une occurrence de *kéner* (taggué en tant que NOM) dans un contexte métalinguistique : « ce qui le rend ultra 'kénable' (du verbe 'kéner'). ».

Cet exemple du verbe *ken* renforce alors notre conviction qu'il y a matière à creuser aussi bien du côté des lexicographes traditionnels (ajustement des graphies des entrées principales en fonction des usages réels) que du côté des chercheurs en linguistique de corpus (élargissement des dictionnaires intégrés aux taggers par le lexique substandard).

3. CONCLUSION

Nos trois exemples ont permis de confirmer notre hypothèse de travail. En effet, les corpus linguistiques du type WaC (Web as Corpus) apportent des informations précieuses sur le niveau de diffusion des argotismes (souvent des néologismes) qui sont identitaires pour les jeunes à une époque donnée, ils servent de **banques d'attestation des usages dans le passé proche**. La riche palette de **notations gra-**

phiques qui les accompagne (notamment s'il s'agit d'emprunts ou de verlanisations) complique les requêtes. Et ce sera donc un défi pour celui qui aura comme mission de revoir les lemmes de façon manuelle (ex.: *ken*/ *kène*/*quène,* verlan de *niquer*, « posséder sexuellement »). Les WaC sont un outil intéressant pour ceux qui s'intéressent au français substandard mais l'accès à certaines formes, en particulier aux formes verbales invariables, reste difficile face à la **complexité des variantes formelles** et encore plus aux **nuances sémantiques**, faute d'annotation morphosyntaxique et de lemmatisation automatiques sans révision manuelle.

On arrive alors à l'époque où les pratiques orales de jeunes qui n'ont eu que peu d'impact dans les corpus écrits lors de la décennie précédente deviennent scripturalisées grâce à l'essor des réseaux sociaux et peuvent être étudiées par les chercheurs grâce à l'arrivée des corpus WaC, ce qui ouvre une brèche pour la « big-data argotologie ».

Acknowledgements: The research has been supported by the Masaryk University Development Fund (project MUNI/FR/1366/2019).

Bibliographie

BARONI, Marco – BERNARDINI, Silvia – FERRARESI, Adriano ZANCHETTA, Eros : The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. In : Language Resources and Evaluation, 2009, Vol. 43, No 3, pp. 209–226. Disponible sur: https://doi.org/10.1007/s10579-009-9081-4

BENKO, Vladimír : Aranea: Yet Another Family of (Comparable) Web Corpora. In : Text, Speech and Dialogue. Eds. P. Sojka *et al*. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, Springer International Publishing Switzerland, pp. 257–264.

CARTIER, Emmanuel : Néoveille, plateforme de repérage et de suivi des néologismes en corpus dynamique. In : Neologica, 2019, No 13, pp. 23–54.

CHOVANCOVÁ, Katarína : Pour une pragmatique de l'écriture interactive en ligne : le statut de l'énoncé dans le chat ». In : La langue en contexte, Helsinki : Université d'Helsinki 2009, pp. 199–211.

CVRČEK, Václav – KOMRSKOVÁ, Zuzana – LUKEŠ, David – POUKAROVÁ, Petra – ŘEHOŘKOVÁ, Anna – ZASINA, Adrian Jan – BENKO, Vladimír : Comparing web-crawled and traditional corpora. In : Language Resources and Evaluation, 2020, No 54, pp. 713–745. Disponible sur : https://doi.org/10.1007/s10579-020-09487-4

DAVIES, Marc : The Movie Corpus. (2019–). Disponible sur : https://www.english-corpora.org/movies/

DEKHISSI, Laurie : Variation syntaxique dans le français multiculturel du cinéma de banlieue. Thèse de doctorat sous la direction d'Aidan Coveney et Zoë Boughton, Exeter : Université 2013.

DOSTIE, Gaétane : Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique. Bruxelles : De Boeck/Duculot 2004.

FERRARESI, Adriano – BERNARDINI, Silvia – PICCI, Giovanni – BARONI, Marco : frWaC. Ústav Českého národního korpusu FF UK 2013, Praha. Disponible sur : http://www.korpus.cz

FIÉVET, Anne-Caroline – PODHORNÁ-POLICKÁ, Alena : La variation du lexique substandard dans le cinéma sur la banlieue : analyse argotologique du champ lexical des relations garçons-filles dans le film « Les Kaïra ». In : Diversité et variations de la langue française au XXIe siècle. Eds. R. Mudrochová – B. Courbon. Plzeň : Nakladatelsví Nava 2020, pp. 183–224.

GADET, Françoise : La variation sociale en français. Paris : Ophrys 2003.

GOUDAILLIER, Jean-Pierre : De l'argot traditionnel au français contemporain des cités. In : La linguistique, 2002, Vol. 38, No 1, pp. 5–23.

GRODNER, Manon : Le « cinéma de banlieue » : représentation des quartiers populaires ? Enjeu d'un cinéma entre réalité et fantasme. Paris : L'Harmattan 2020.

JAKUBÍČEK, Miloš – KILGARRIFF, Adam – KOVÁŘ, Vojtěch – RYCHLÝ, Pavel – SUCHOMEL, Vít : The tenten corpus family. In : 7th International Corpus Linguistics Conference CL, 2013, pp. 125–127.

JEAY, Anne Marie : Les messageries télématiques. Une communication paradoxale. Paris : Eyrolles 1991.

MÉVEL, Pierre-Alexis : Traduire La haine : banlieues et sous-titrage. In : Glottopol, revue sociolinguistique en ligne, 2008, No 12, pp. 161–181.

NÁDVORNÍKOVÁ, Olga – VAVŘÍN, Martin : Korpus InterCorp – francouzština, verze 7 z 19. 12. 2014. Ústav Českého národního korpusu FF UK 2014, Praha. Disponible sur : http://www.korpus.cz

PADRÓ, Lluís – STANILOVSKY, Evgeny : FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul 2012. Disponible sur : http://nlp.lsi.upc.edu/publications/papers/padro12.pdf.

PODHORNÁ-POLICKÁ, Alena – FIÉVET Anne-Caroline : Argot commun des jeunes et français contemporain des cités dans le cinéma français depuis 1995 : entre pratiques des jeunes et reprises cinématographiques. In : Glottopol, revue sociolinguistique en ligne, 2008, No 12, pp. 212–240.

PODHORNÁ-POLICKÁ, Alena – FIÉVET Anne-Caroline : Le rap en tant que vecteur des innovations lexicales : circulation médiatique et comportement des locuteurs. In : Écarts et apports des médias francophones. Eds. M. Abecassis – G. Ledegen. Oxford, Bern, Berlin, Bruxelles, Frankfurt : Peter Lang 2013, pp. 113–139.

PODHORNÁ-POLICKÁ, Alena : RapCor, Francophone Rap Songs Text Corpus. In : Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2020. Eds. A. Horák et al. Brno : Tribun EU 2020, pp. 95–102.

POPOVIČOVÁ SEDLÁČKOVÁ, Zuzana : Slang v mládežnickom diskurze. Bratislava : Univerzita Komenského v Bratislave 2013.

RYCHLÝ, Pavel : Manatee/Bonito – A Modular Corpus Manager. In : 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University 2007, pp. 65–70.

SABLAYROLLES, Jean-François : D'où viennent les mots nouveaux ?. In : Sciences humaines, Le langage en 12 questions, 2013, Vol. 3, No 246, p. 14.

STEIN, Achim – SCHMID, Helmut : Étiquetage morphologique de textes français avec un arbre de décisions. In : Traitement automatique des langues, 1995, Vol. 36, No 1-2, pp. 23–35.

Dictionnaires (avec les abréviations utilisées) :

DZ : Cobra le Cynique [Abdelkarim Tengour] (2000-2020). Le Dictionnaire de la Zone. Disponible sur : https://www.dictionnairedelazone.fr

AFP : COLIN, Jean-Paul – MÉVEL, Jean-Pierre – LECLÈRE, Christian : Argot et français populaire. (1ère édition sous le titre « Dictionnaire de l'argot », 1990), Paris : Éditions Larousse 1990-2008.

LC : Collectif Permis de vivre la ville : Lexik des cités. Paris : Fleuve Noir 2007.

CTT : GOUDAILLIER, Jean-Pierre : Comment tu tchatches! Dictionnaire du français contemporain des cités. Paris : Maisonneuve & Larose (4ème éd. 2019 ; 1ère éd. 1997).

BOQ : LAFFITTE, Roland – YOUNSI, Karima : Bien ou quoi? La langue des jeunes à Ivry et Vitry-sur-Seine. Paris : SELEFA 2004.

PR : *Le Petit Robert*. Paris : Dictionnaires Le Robert 1997–2021.