RapCor, Francophone Rap Songs Text Corpus

Alena Podhorná-Polická

Faculty of Arts, Masaryk University
Arna Nováka 1, 602 00 Brno, Czechia
podhorna@phil.muni.cz

Abstract. The paper introduces the RapCor corpus, which is a specific text corpus for French, based on francophone rap songs' texts from the last three decades when rap music became one of most popular music genres. An overview of more than ten years of rap corpora building presents our motivations, text processing methods, annotation decisions, as well as achievements and problematic issues. The published part of rap corpora, available in Sketch Engine manager for interdisciplinary research, the RapCor 1288, consists of 709,057 words of 1288 francophone rappers' texts. It had been used mainly for the detection and longitudinal observation of so-called "identitary neologisms", i.e. expressions emerging from communication between peers, motivated by search for group belonging, playfulness and expressivity. Rappers' language is also a valuable resource for investigating metaphors and idioms that have been formed by assigning a new meaning to existing language items. The main goal of this largely substandard linguistic corpora is to uncover the phonemic and semantic innovations and trends in modern French.

Keywords: French corpus; text processing; rap music; hip hop; lyrics; substandard language; neology; written orality; corpus building

1 Introduction

Text corpora are useful resources for investigation into a broad range of linguistic issues, including the study of neologisms in a short-term diachronic perspective. This approach became more accessible recently thanks to the arrival of big data corpora (e.g. [12]). Together with gradual perfection of neologism detection tools based on typical co-occurrence analysis (so called "discriminants", see [8]), or, more usually, based on exclusion lists or dictionaries (e.g. European project *Néoveille*, cf. [1]), the automatic extraction of neologism candidates seems to be able to facilitate the detection of ongoing changes in the lexicon (for an overview in Czech, see [17]).

However, our previous collaboration with the Lingea publishing house on the second edition of the substandard French-Czech dictionary *Pas de blème !* [15] and our ongoing participation in Neoveille's project confirms our conviction that there is still a lot to do as regards the detection of so-called "identitary neologisms", i.e. expressions emerging from communication between peers, motivated by search for group belonging, playfulness and expressivity [11]. It

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 95–102, 2020.

is due to: a) its low frequency in newspaper monitor corpora which are usually used to neo-crawling, b) difficulties in automatic detection of neosemantisms (changes of meaning of existing word forms, as is very frequent in the formation of new words in slang) and c) gaps in integrated exclusion dictionaries that are rarely based on spoken language vocabularies. Thus, in order to fill these gaps and being motivated by pedagogical and translation considerations, we set out to compile a specific slang corpus for French with a focus on rap music lyrics, well known for a high frequency of slang words [4, 20], as well as a variety of identitary neologisms in early stages of their societal diffusion.

2 Corpus Construction Motivations

More than ten years ago, the idea of a rap corpus emerged in the Department of Romance Languages and Literatures of Masaryk University, motivated by the then research goal to track identitary neologisms (e.g. *bolos* [10]) as well as by the students' constant interest[1] in discovering substandard French via rap lyrics, mainly through translation or linguistic analysis in their scholar theses.

At that time, French rap was growing exponentially until it became the second largest market for the production and consumption of this genre after the American scene [18]. Nowadays, rap music, as one of hip-hop culture pillars (together with break dance, DJing, beat-box and graffittis), dominates French music industry despite several controversies that have even led to lawsuits [5].

Over the past four decades, rappers have been reflecting historical, political, and economic circumstances, focusing the listeners' attention to their own origins, usually but not exclusively those emblematic districts and suburbs (called *cités* in French), which are ethnically mixed and geographically more or less segregated. But rap lyrics also describe everyday life, personal problems, dreams and visions along with other features, characteristic for youth culture, such as the use of new buzz words, old slang expressions and idiolectal formations created on the spot in order to meet aesthetic-artistic expectations [3] as well as intrinsic metric restrictions of the genre of rap music [6].

3 Corpus Priorities and Choices

The creation of a linguistic corpus of francophone rap songs began in spring 2009, when the first one hundred rap songs were included in the Bonito corpus manager[2] [14]. That is also when the corpus obtained its name and logo[3]. The original idea was to create a corpus based on rap music production in

[1] The first data compiling and text processing were accomplished thanks to the enthusiastic work of our student Jiří Marek.

[2] Thanks to the web page description at `https://nlp.fi.muni.cz/projects/bonito/` (still active).

[3] For additionnal presentation, see: `https://is.muni.cz/do/phil/Pracoviste/URJL/rapcor/index_en.html`.

France but students quickly became interested in extending the text compilation by involving other francophone hip-hop scenes as well (i.e., French-speaking rappers from Belgium, Switzerland, Canada or even from Senegal including French overseas territories and departments, e.g. Reunion Island in the Indian Ocean). Furthermore, another one hundred texts from Czech and Slovak rap albums were prepared for tokenization but were not published yet.

So far, we have relied on transcriptions of lyrics that the authors themselves insert into album booklets. Until today, we have been listing in our Google document database 2,424 francophone rap albums (EP, LP and mixtapes) published between 1984 and 2020, with 904 (37 % of whole database) physically verified against their CD or vinyl versions. In sum, almost 53 % (479) of all verified booklets contain at least one text transcription, in 16 % of them (78) all texts are transcribed. The advantage of this method (as opposed to obtaining text transcripts from fan pages on the internet) is that it deals with authors' licensed texts, thus enabling us to identify their own grammar and typing errors, as well as any disparities between the written- and the sung-versions (labelled as P and S versions in RapCor's database). Such differences are often intentional, particularly where the sung-version (S) is way too explicit. Another advantage of this approach is that it helps us to obtain a huge variability of transcriptions of oral neologisms that do not have a fixed graphic form yet (due to the phonemic orthography of written French, in comparison to Czech, for example). Moreover, this method allows us to compare P and S versions in parallel-corpora, using *MK-align* software[4] (see the scholar work of Vaňková, 2014, listed in the next link) and to point out the most frequent categories of those P and S disparities (consisting mainly in non-transcription of introductive or final rappers' "gimmicks", territorial references and echoed voices).

The newest version of RapCor, released on 21st October 2020 and available in the Sketch Engine corpus manager, is *RapCor 1288*, i.e. having the lyrics of 1288 songs. Relying on authors' transcripts still remains the most original feature of RapCor's text processing, even though this corpus version does contain 133 texts of songs by well-known rappers that were gathered from internet's fan pages because the rappers never transcribed their lyrics into album booklets. Their inclusion is the result of our students' decision, either motivated by seeking a representative nature of their sub-corpora in several theses[5] or by their personal choice for seminar projects on lyrics translation into Czech. In the future, we expect that the number of unauthorized lyrics retrieved from the internet (or authorized by interprets directly on the internet) will increase because the importance of publishing in physical format is decreasing.

As the building of the RapCor corpus is strictly based on students' work, and as our students are mostly non-native speakers of French, it is not possible to ask them for their own transcriptions of the voice on the music track. Rappers often fill their "punchlines" with substandard and trendy words, including

[4] Freely available at: `http://www.tal.univ-paris3.fr/mkAlign/#p1` (see [2]).

[5] Listed at: `https://is.muni.cz/do/phil/Pracoviste/URJL/rapcor/kvalifikacni_MU_en.html`.

borrowings from non-European languages that reflect the dynamics of lexical innovations in multicultural suburbs. The frequency of the "XXX" symbol (i.e. a word or words that is impossible to hear or understand) in online published version of corpus (301 times) serves as evidence of how difficult it is when one tries to accurately transcribe the rappers' rappings.

The frequent gaps in lyrics transcriptions on the internet clearly show that speech recognition is sometimes hard even for native speakers and experts in local hip-hop scenes. That is one of the reasons why the digital media company *Genius.com* was so successful in 2014 with their song annotation system. The company changed their earlier "explanations" of hip-hop lyrics on their original website *RapGenius.com* into a collaborative database, based on the contributions of "annotators" of lyrics and opened itself to other music genres.

Online publication of rap lyrics in French started on RapGenius just in 2010, but until the re-launch of Genius, our interest in searching for non-authorized text transcripts has led us to explore several other web sites. Thanks to our collaboration with The Natural Language Processing Centre at the Faculty of Informatics, a more convenient extraction of texts from various specialized websites (so-called *RapCor Text Crawler*[6]), was launched in 2016. With this tool, online versions of lyrics, if available, can be obtained in several (from one to six) different versions, which helps our students to clarify content ambiguities in case of doubt over the accuracy of the fan's transcription of the analyzed song. Together with booklet lyrics, there are currently (November 2020) 4,588 transcriptions of francophone rap songs in different stages of corpus treatment.

4 Text Processing Methods

The preparation of texts to include in the linguistic corpus is much more time consuming in case of authentic lyrics' taken from album booklets than lyrics taken from internet text crawlers. For that reason, the expansion of RapCor has been relatively slow in comparison with other text corpora, mainly those crawled from web pages with automatic annotations[7].

Firstly, students are completing structural metadata about songs (singer, featuring, album, publication year, song's length) and, more recently but not exhaustively, open personal metadata about artists (date and place of birth, geo-localisation, parents' origins, etc.).

Prior to tokenisation, there are several text processing phases:

1. the scanning of texts from booklets (from 2015 until now, 684 albums were (re)scanned in high resolution);

[6] Thanks to Pavel Rychlý's student Lukáš Banič, available at: `https://nlp.fi.muni.cz/projekty/rapcor/`.

[7] By writing an open-source script with R program in order to crawl texts from Genius.com, Corentin Roquebert [13] offered in 2015 a comprehensive to double the size of RapCor in a quite short time (see `https://nycthemere.hypotheses.org/541`). For various discourse analysis, this quantitative method is sufficient but it doesn't offer further qualitative features as our text processing methodology.

2. after encoding of available transcriptions of songs, the texts are cut and vertically pasted (presently 2,427 documents);
3. the result of automatic text resolution (OCR) is then checked and saved as text recognized image (decision made just in 2015; until now 664 texts were saved as double-layer pdf files in our internal storage);
4. the machine-recognized and word-by-word controlled text is tagged by structural information tags (full title, interpret, featuring, album, year of publication, length of music track) and two text files, P (written-) and S (sung-) versions, are created. The authorized booklet text's transcription (P version) is checked against audio recordings (S version) and any differences found in lyrics are colour-coded according to predefined disparity categories (both in P and S version). As of now, 2,323 P versions have been prepared by controlling OCR recognition and by adding structural tagging. Because of the time-consuming overhearing of audios, the mirror S version has already been completed for only 1,332 of them. There are actually seven categories of disparities to be coloured:

   – correction of typographical or grammatical error;
   – deliberately missed correction if it was author's intention (mainly for jokes or special adaptations of loanwords);
   – replacing of text by another one (frequent for vulgarisms);
   – addition of text (especially opening and closing sections called "intro" and "outro" are sung but not transcribed in booklets), including punctuation in order to increase the quality of morphosyntactic annotations (lyrics are often written with upper case in the beginning of each punchline and without any finishing mark of sentences);
   – omission of text (less frequent but annotated as empty token);
   – different position of stanza or chorus in booklet transcription in comparison of what is really sung;
   – pronunciation mismatch (frequent for graphically non-adapted loanwords; annotation include phonetic transcription in international phonetic alphabet);

   and the a special category with colours belongs to the aforementioned "XXX" in S versions of all types:

   – incomprehensible passage (i.e. waiting for overhearing by native speakers).

   Only after creating of S version of two types (a) colour-coded files in case of traditional text processing as explained before or b) just tagged by structural information tags in case of fans transcriptions from the internet – other 142 texts), we can proceed to the final stage of pre-annotation text processing;
5. the creation of a plain text without any tags that can be submitted for automatic tokenization and semi-automatic annotation (1,474 texts are completed until this stage).

5 Remarks on Lemmatisation and Annotation

When the first texts for corpus were prepared, there was only one tool available as open-source, the famous *TreeTagger* tool[8]. For this reason, all segmentation and annotation was carried out by the TreeTagger, even if the result had to be reviewed carefully because of its poor dictionary in the matter of oral/substandard French. About 2012, we also tested some RapCor texts on a licensed tool *Cordial Analyseur*[9] with more successful syntactic annotation results but its wider use became unsustainable in our collaborative project for financial and technical reasons.

Until 2019, students were adding words that TreeTagger lemmatized as "unknown" to the shared Google table file. This table was designed as an internal dictionary with a special focus on loanwords, inversed slang words (well-known as "verlan"), regionalisms (mainly from well-represented Canadian French) and socio-cultural references invoked by usage of proper names (mainly toponyms and anthroponyms). Thanks to close collaboration with Pavel Rychlý during this year, students don't have to download and learn to handle TreeTagger tool in their computers and then search for unknown lemmas in our dictionary anymore. By simple pasting of the plain text into a so called French tagger web page window, together with the code of the song[10], they can download encoded Open office table file with tokenized and automatically pre-annotated text. The special feature is adding to TreeTagger's lemmatization result lemmas from our dictionary. Furthermore, tokens lines containing unguessed ambiguities, disparities between both dictionaries and unknown words are coloured in three different colours in order to attract students' attention. This will help us to decrease the number of errors in a decision process for POS and lemmas. The advent of big web corpora enables us to verify frequencies in writing down of orality.

As the corpus building started a long time before that, the question of unique lemma's choice in case of several graphic variants was solved by applying of so called *method of lexicographic filters* [9:341-353]. For example, [tɛs] is an identitary neologism which knew a quick diffusion in common youth slang and it is semantic equivalent of the aforementioned word *cité* (created by truncation of bissyllabic (verlan) inversion: *cité* [site] > *téci* [tesi] > [tɛs]. In booklets, rappers transcribed this word as *tess, tece, téc, té-ce(s), tèc* and *tèce*. As it is absent from reference dictionaries such as *Le Petit Robert* or *Le Petit Larousse*, the search of lemma passed through the first lexicographic filter and, as explained in [16:154], the lemma *téc* was fixed for our corpus, according to the graphic form chosen as entry in unique slang dictionary with academic background that recorded [tɛs] at that time (Goudaillier's 3rd edition of *Comment tu tchatches!*, published in

[8] Freely available at: `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` (for detail explanation in French, see [19]).

[9] Called also Cordial Universités, this product of French company Synapse development was widely used in lexicometric studies [7].

[10] Available at: `https://nlp.fi.muni.cz/projekty/rapcor/tag.cgi`.

2001)[11]. However, the aging of the then trendy words brings usually an increase of graphic match that can be seen on frequencies and year stamp of each form: *tece* (once in a song released in 2004), *téc* (once in 2009), *tè-ce* and *tè-ces* (both once by the same singer in 2005), *tèce* (twice in 2004 and 2006) and *tèc* (twice in RapCor, in 1998 and 2009), in comparison of longitudinal use of *tess* (48 times in RapCor1288). This form dominates also in very large web corpora for French (both in *FrTenTen* 2012 and 2017 or in *Araneum Francogallicum*) where it is mixed with homonymic female name *Tess* (written often without upper case). While occurrences in those big corpora exceed easily several hundreds, our small numbers are more relevant statistically (e.g. relative frequency of *tess* in RapCor is 62.54 i.p.m, but only 0.71 i.p.m. in FrTenTen12, which is still more than 0.5 i.p.m. in newer FrTenTen17). This example shows the dynamics of slang words in recent diachrony. It is also a typical case of our "drifting in the current", when corpus building includes permanent integrating of new NLP tools on old and new lyrics and (re)considering lemmatization for old and new slang words.

6 Conclusion and Future Work

The RapCor corpus is to our knowledge the only existing comprehensive linguistic corpus of francophone rap songs, which were released between 1984 and 2020. In its current published version, RapCor incorporates a sample of 1,288 individual songs from 169 albums and sung by 570 artists. However, this is still a relatively small part from the universe of the French rap production and a small portion of rappers' population. In order to be able to benefit from the full potential of the database (another 3,300 songs are under treatment and thousands of others can be incorporated already), RapCor needs further expansion and equilibration. Together with displaying of P and S versions disparities in the next version, our short-term objective is to rebuild our dictionary and to test other taggers than TreeTagger, especially UDPipe and FreeLing.

Acknowledgments. This work was supported by the project of specific research "Románské jazyky a románské literatury 2020" (Romance languages and romance literatures 2020), project no. MUNI/A/1262/2019.

References

1. Cartier, E. (2017). *Neoveille, a Web Platform for Neologism Tracking.* Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017. `https://www.aclweb.org/anthology/E/E17/E17-3024.pdf`.

[11] This choice of *téc* is logic if one wants to refer to original word *téci* but improper form phonological point of view (*é* refers always to closed-mid vowel [e] and not to open-mid [ɛ]). This confirms the Tengour's choice of the entry *tèce* (with sub-entry *tess*) for its online peri-urban slang dictionary (`http://www.dictionnairedelazone.fr`; printed version was published in 2013).

2. Fleury, S. (2012). *MkAlign (version 2.0) : Manuel d'utilisation.* Paris: Université Sorbonne Nouvelle Paris 3.
3. Ghio, B. (2012). *Le rap français: désirs et effets d'inscription littéraire. Disertation Thesis.* Paris: Université Sorbonne Nouvelle – Paris 3.
4. Goudaillier, J.-P. (2019). *Comment tu tchatches! Dictionnaire du français contemporain des cités.* 4th edition. Paris: Maisonneuve & Larose, (1st ed. 1997).
5. Hammou, K. (2012). *Une histoire du rap en France.* Paris: La Découverte.
6. Chodakova, P. (2014). *Metrická inovace ve francouzském a českém rapu.* Lingvistika Praha. http://lingvistikapraha.ff.cuni.cz/node/199.
7. Lebart, L., Pincemin, B. and Poudat, C. (2019). *Analyse des données textuelles.* Québec: Presses de l'Université du Québec.
8. Paryzek, P. (2008). *Comparison of selected methods for the retrieval of neologisms.* Investigationes linguisticae, Vol. XVI, pp. 163–181.
9. Podhorná-Polická, A. (2009). *Universaux argotiques des jeunes.* Brno : Munipress.
10. Podhorná-Polická, A. and Fiévet, A.-C. (2009). *À la recherche de la circulation d'un néologisme identitaire: le cas de bolos.* In Kacprzak, A. and Goudaillier, J.-P. (eds.). Standard et périphéries de la langue. Łódź: Oficyna Wydawnicza Leksem, pp. 207–223.
11. Polická, A. (2018). *Lexikální inovace. Dynamika šíření identitárních neologismů.* Brno: Masarykova univerzita. `https://www.muni.cz/inet-doc/1129409HAB_SPIS_Policka_2018_final.pdf`.
12. Renouf, A. (2016). *Big data and its consequences for neology.* Neologica, 10, pp. 15–38.
13. Roquebert, C. (2015). *Tutoriel: Récupérer des paroles de rap du site Rapgenius* [online]. Nycthémères. Mesures du rap. Academic blog Hypotheses.org. `https://nycthemere.hypotheses.org/533`.
14. Rychlý, P. (2007). *Manatee/Bonito – A Modular Corpus Manager.* 1st Workshop on Recent Advances in Slavonic Natural Language Processing, pp. 65–70.
15. s.a. (2012). *Pas de blème ! Slovník slangu a hovorové francouzštiny.* Brno: Lingea (1st edition 2009).
16. Sekaninová, T. (2012). *Stéréotypes liés au verlan : variation diatopique dans le rap français* [master's thesis]. Brno: Masaryk University. `http://is.muni.cz/th/263203/ff_m/`.
17. Sláma, J. (2017). *K (polo)automatické excerpci neologismů*. Jazykovědné aktuality, LIV, 3–4, pp. 34–46.
18. Spady, J., Meghelli, S. and Alim, H. S. (2006). *Tha global cipha: Hip Hop culture and consciousness.* Philadelphia: Black History Museum Press.
19. Stein, A. and Schmid, H. (1995). *Étiquetage morphologique de textes français avec un arbre de décisions.* Traitement automatique des langues, 36, 1-2 (Traitements probabilistes et corpus), pp. 23 –35.
20. Tengour, A. (2013). *Tout l'argot des banlieues. Le dictionnaire de la zone en 2 600 définitions.* Paris: Les éditions de l'Opportun.