Univerzita Karlova

Filozofická fakulta

Ústav informačních studií a knihovnictví

Diplomová práce

BcA. Julie Klimentová

Texty frankofonního hip hopu z pohledu digital humanities

Francophone Hip Hop Lyrics from the Perspective of Digital Humanities

Praha 2022

Vedoucí práce: Mgr. Josef Šlerka, Ph.D.

Tímto bych ráda poděkovala Mgr. Josefu Šlerkovi, Ph.D. za odborné vedení. Poděkování patří také doc. PhDr. Aleně Polické, Ph.D. za poskytnutí konzultace, přístupu k rapovému corpusu RapCor a především za inspiraci.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 11. května 2022

Julie Klimentová

Abstrakt (česky)

Tato práce obsahuje explorační analýzu textů rapových písní ve francouzském jazyce. Hlavním záměrem této práce je odpovědět na tři základní otázky: Jaká je slovní zásoba rapových písní ve francouzském jazyce? Jaká témata tyto písně reflektují? Jsou v těchto písních média a technologie důležitá témata?
Metody použité za účelem zodpovězení těchto otázek jsou následující: přehled dostupné literatury, sběr dat (sběr francouzských rapových textů), statistická textová analýza a algoritmické modelování témat.
Výsledky tohoto výzkumu potvrzují závěry nalezené v přehledu literatury. Podporují tvrzení, že výskyt nestandardních slov ve francouzských rapových textech není tak vysoký. Co se týče modelování témat, výsledky potvrzují, že francouzské rapové texty obsahují mnoho různých témat, například protisystémové postoje, existenční potíže, cenzuru a kritiku nepravdivých informací prezentovaných autoritami.

Abstract (in English)

This paper contains an explorative analysis of French rap lyrics. The focus of the work is to answer three questions: What is the vocabulary used in the French rap lyrics? What are the topics present in the French rap lyrics? Are media and technology a significant topic in French rap lyrics?
The methods used to obtain the answers to the research questions can be listed as follows: literature research, data collection of French rap lyrics, statistical textual analysis, and algorithmic topic modelling.
The findings confirm previous research on the subject by supporting claims that the level of non-standard language in French rap lyrics is not as high as myths suggested. In terms of the topics, the topic modelling confirms that there is a variety of themes present in French rap lyrics, including anti-systemic sentiments, struggle, censorship, and false information presented by the authorities.

Klíčová slova (česky)

Francouzský rap, analýza písňových textů, modelování témat, LDA, statistická textová analýza, média, technologie, lingvistika

Keywords (in English)

French rap, lyrics analysis, topic modelling, LDA, statistic textual analysis, media, technology, linguistics

Table of Contents

1 Introduction

This paper contains an explorative analysis of French rap lyrics. The focus of the work is to answer three questions: What is the vocabulary used in French rap lyrics? What are the topics present in French rap lyrics? Are media and technology a significant topic in French rap lyrics?

This interest in media related words and topics comes from the perspective of new media studies. It is due to the observation of topics related to the media and technology currently present in musical genres – both conscious use of the subject in a critical way and natural appropriation of the related vocabulary.

Our research covers a brief review of the history of rap in France as well as a literature review summarising so far conducted research about themes and vocabulary in French rap. The tools used to answer the questions are coming mostly from statistical textual analysis and are further described in the methodology.

2 Terminology

Genius.com

Online database of musical lyrics. It provides a reliable API exposing songs and artists metadata, but not song lyrics.

GitHub.com

A site where code repositories are stored and shared using the git version control system.

Rap Corpus of Masaryk University, a.k.a. RapCor

A project that maintains a high-quality French rap lyrics corpus. Founded and led by Czech linguist doc. PhDr. Alena Polická, Ph.D.

Sketch Engine

A text analysis tool which allows institutional login.

Jupyter Notebook

An open-source application which allows sharing executed code and visualisations.

3 Methodology

In order to answer the research questions, several steps needed to be carried out: the performers names collection, the lyrics collection, and the statistical textual analysis together with topic modelling.

First, a selection of the most known French rap / hip hop performers and bands had to be collected. In order to acquire such information, a combination of non-academic sources was used for the purpose of obtaining the widest range of performers. For this purpose, Wikipedia API was used together with the curated list by one of the subscribers from the website SensCritique (Artzgild, 2020).

For the purpose of this research, statistical textual analysis was applied on gathered lyrics from the online lyrics database Genius.com. Genius.com was chosen because it is one of the most popular lyrics sources and it provides a reliable API to query artists and songs' urls. The full lyrics were parsed from the HTML content provided by the site.

In the interest of getting a solid idea of what the statistical analysis of the text would comprise, extensive research has been conducted and different approaches investigated. In the end, the R programming language was used, mainly the libraries *tidytext* (Robinson and Silge, 2021) and *udpipe* (Wijffels, Jan et al., 2021), as well as Python programming language and the package *gensim* (Řehůřek and Sojka, 2021).

The content of the analysis can be briefly described as follows:

A subcorpus of songs containing media related words (further referred to as **the media words subcorpus** or **subcorpus**) was created from the lyrics corpus (further referred to as the **main corpus** or **corpus**) and the ratio of the media words subcorpus to the main corpus was calculated.

The media words subcorpus was created by subsetting songs, which contain media related words. The list of media related words used was carefully selected based on extensive lyrics research and investigating media specific words throughout the history.

Word frequencies were calculated for the corpus and subcorpus. For the main corpus, a comparative analysis was performed with French Web Corpus 2017 (Sketch Engine, 2017).

For both the corpus and subcorpus, topic modelling was performed with the LDA algorithm.

4 Evolution of rap in France

Considering the subject, a brief introduction into the context and history of rap in France is deemed necessary. However, the topic will not be examined in great detail as there have been works written on this subject, which are more apt for giving a profound historical insight, such as *Une histoire du rap en France* by Karim Hammou (2014) or *Regarde ta jeunesse dans les yeux* by Vincent Piolet (2017), or others.

American rap beginnings can be linked to the birth of hip hop culture in 1970s in the Bronx, New York, whereas in France, it started emerging only in the 1980s (Devilla, 2011, p. 76). It was developing on two fronts: by the already established artists in the media and by the amateurs who appropriated the funk heritage. It was diffused mainly by the radio (specifically Radio Nova) and in the music clubs. Until 1984, when the H.I.P H.O.P. TV programme started, it was mostly ignored by the TV. The situation changed with the 1990s when rap became the symptom of social problems reframed by the media as a practice employed by minorities. From that moment, rap became associated with the suburbs (Hammou, 2014 as cited in Mayaud, 2015).

In terms of the first track/rap album released, there are some discrepancies among the claims by the researchers (e.g. Verbeke, Devilla, Vicherat). However, when overlooking individual opinions, the names such as DJ Dee Nasty, IAM, Suprême NTM, and MC Solaar are associated with the rap beginnings in the late 1980s/early 1990s.

In the 1990s, due to the success of rappers in the media, the music labels invested into rap artists. However, they still do encounter rejections by the radio and have only limited commercial success. They are also still largely associated with the suburbs and violence. This stigma is enhanced by the state institutions and their policies and more new agents influencing the definition of rap come into place. Independent labels and collectives emerge in opposition to the rappers promoted by the major labels and manage to instil a new rap scene (Hammou, 2014 as cited in Mayaud, 2015).

In the last period studied by Hammou, the 2000s, rap already constitutes a segment recognized by the music industry and even participates in the emergence of another new genre, R&B. Nevertheless, the association with street life and underground culture is still present. Despite the widespread popularity and acceptance of the genre by larger public, rap is still often slandered by the state representatives when being part of various

judiciary battles and open discussions. According to Hammou, that is a positive phenomenon as the public behind rap music suddenly receives a voice to speak about its realities which was not possible before (Hammou, 2012 as cited in Tamagne, 2014). As we have entered the 2020s, new reflections on the past decade in rap need to come. However, from the current point of view, the general tendency seems to be continuing the trend set in the 2000s as defined by Hammou (2014). The rap genre is already an established genre with a large audience and new performers are continuously emerging and gaining more influence.

5 Literature review

5.1 Research questions

In order to have a clear overview of already conducted research on the subject of topics and vocabulary in French Rap lyrics, the following research questions were central to the literature review:

- What is the vocabulary used in French rap lyrics?
- What are the topics in French rap lyrics?

However, the number of literature and other sources reviewed also contained other subjects important for this paper, such as French rap history, statistical textual analysis, data collection, or topic modelling techniques. For the purpose of clarity and focus, these subjects are omitted in this section and the respective sources are mainly used in the corresponding sections of this paper - Evolution of rap in France, Data collection, Analysis tools, and Topic modelling.

The answers to these questions were researched with respect to the methodology described as follows.

5.2 Methodology

For the literature review, two main literature sources were selected. The most useful sources proved to be the bibliography and related works of the Rap Corpus of Masaryk University. The complementary source was the Charles University search service of accessible academic sources, ukaz.cz. It allows easy access to a large number of academic texts from various sources and thus it is an ideal source for creating an overview of published literature.

5.3 Inclusion criteria

As the literature on the specific subject of French rap lyrics textual analysis is not extensive, the inclusion criteria needed to be flexible enough to be able to work with enough sources, but not too loose in order to be sufficiently relevant to the research.

Nevertheless, the inclusion criteria are also delimited by the geographical location and language skills of the researcher.

Considering these specifics, the inclusion criteria were defined as follows:

- Relevancy: The source must be contributing to answering the research questions or provide some important context to the research.
- Quality: The source must be of high quality and trustworthy sources can be included.
- Academic level: It needs to be conducted at a minimum level of a Master thesis.
- Availability: The source must be accessible, i.e. it must be either in the scientific databases available at Charles University, or available online, or affordable.
- Language: The source must be in English, Czech, or French language.

5.4 Overview of literature

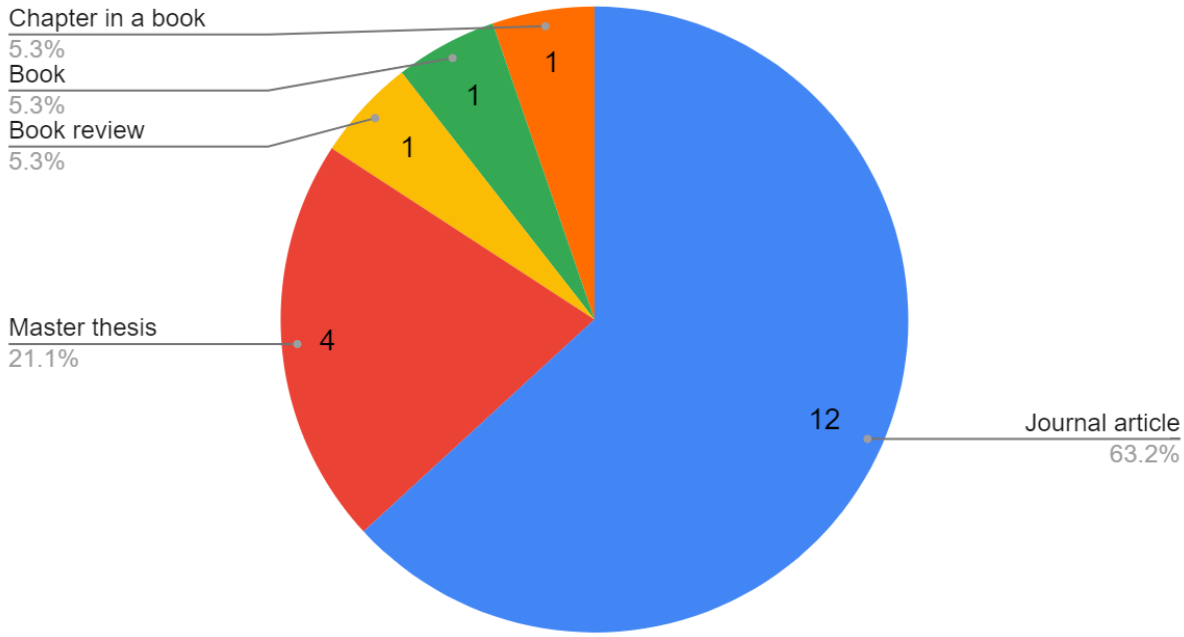

Figure 1: Types of reviewed literature

Selection of nineteen works was chosen for the literature review considering the aforementioned inclusion criteria. The majority of the works are journal articles – twelve / 63.2%, a smaller cluster, which will be further on analysed separately, are master

theses – four / 21.1 %. The following types – book chapter, book, and book review – scored each a count of one / 5.3%. This makes the journal articles the most impactful group for our research (see Figure 1).

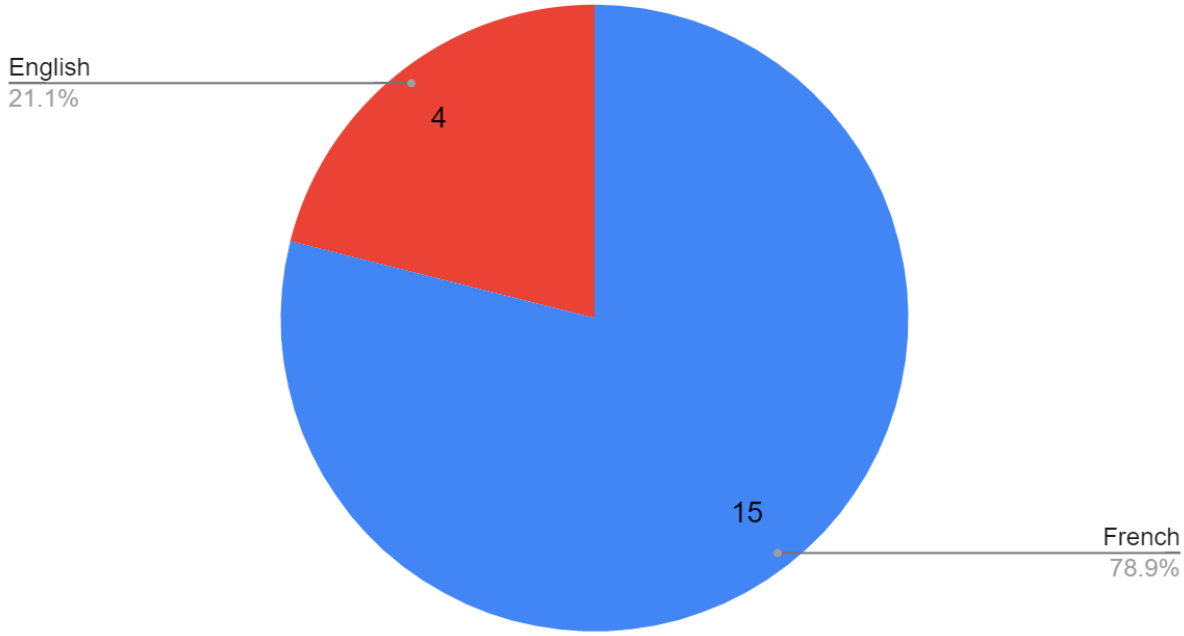

Figure 2: Languages of reviewed literature

Looking at the languages present in the book review, French language is understandably prevalent – fifteen / 78.9% with a small portion of texts being in English – four / 21.1% (see Figure 2).

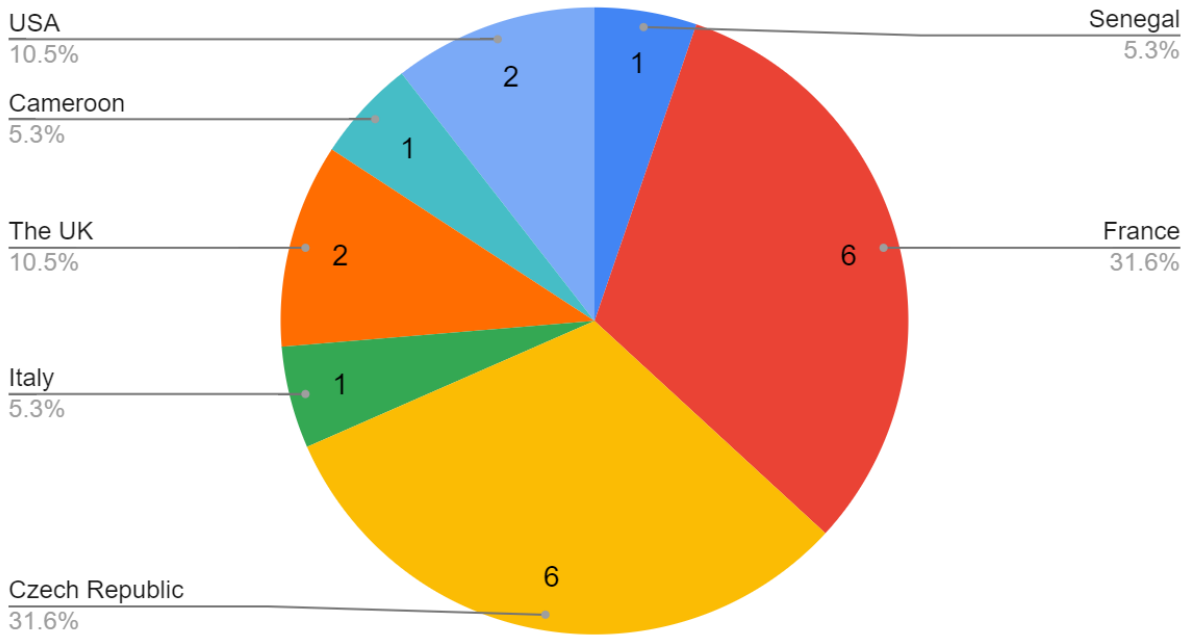

Figure 3: Countries of residence

In terms of the country of residence of the researchers, the selection is quite diverse and comprises three continents, Europe, America and Africa. France is understandably prominent – 6 occurrences / 31.6% together with the Czech Republic – the country of the researcher had a significant impact on access to resources – 6 / 31.6%. The English-speaking countries, the UK and the USA having both a count of two/ 10.5% came as third (see Figure 3). The rest of the countries scored an occurrence each, all of them having some relation to the French language – the African countries being partially francophone and Italy being the neighbour of France and Italian belonging to the same language group.

5.5 Content analysis

The deeper content analysis focuses on different types of works than master theses. The master theses will be analysed further on separately. The aptitude for the research was measured by the level of impact on the answering of the research questions – "What is the vocabulary used in French rap lyrics?" and "What are the topics in French rap lyrics?" with the scale from Low through Medium to High.

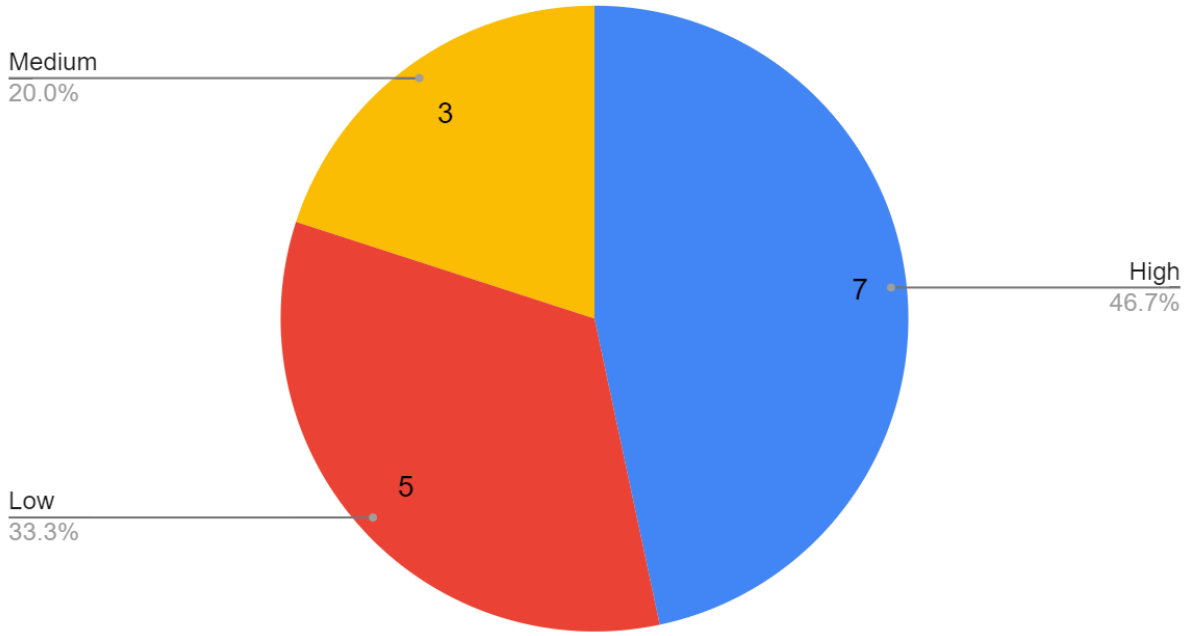

Figure 4: Impact on vocabulary research answer

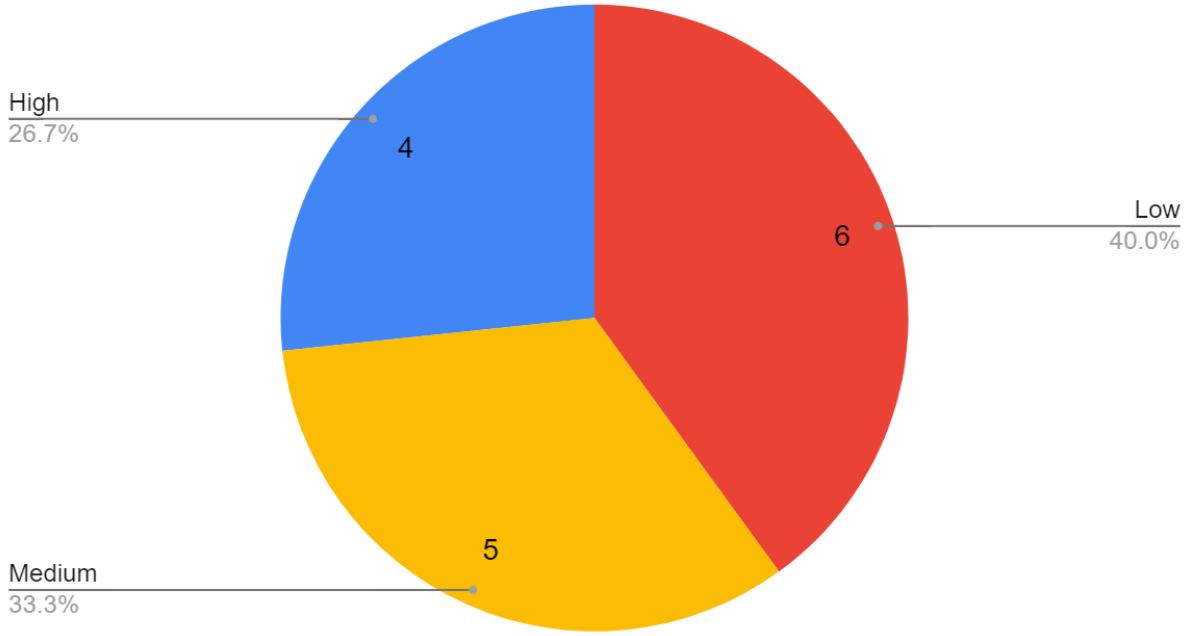

Figure 5: Impact on topics research answer

Overall, more texts were having a high impact on answering the vocabulary related question than on the topic related question (see Figures 4 and 5 for details). In order to maintain the literature review highly focused, only texts which scored High or Medium score in either of the categories were selected for the deeper content analysis. This selective approach leaves us with thirteen texts with medium to high impact on answering our research questions.

5.6 What is the vocabulary used in French rap?

Perhaps the most straightforward answer to this question is proposed by Skye Paine, an associate professor of French at State University of New York. In his article *The Quadrilingual Vocabulary of French Rap,* he asserts that the language used in French rap lyrics consists of four categories: standard French, French slang (*argot*), American slang and the various *langues de bled.* By *langues de bled* he means vocabulary coming from various languages currently used in France ("bled" – the word coming from Arabic word "balad" – "a village") (Paine, 2012).
One of the conclusions interesting for our research is that according to Paine the non-standard language takes a relatively small part of the corpus:

"When one thinks rap one necessarily thinks of non-standard language and much of the corpus supports this presupposition. The data shows that in the 72 songs, there are 1445 separate utterances that could be considered non-standard (calculated as a combination of French slang, American slang, and words du bled). Yet, this must not be overstated as it only represents 4% of all words in the songs." (Paine, 2012, p. 53)

He also suggests that the French language employed is often grammatically correct forms instead of using more colloquial forms, the speakers tend to use more correct syntax, e.g., proper using of "ne" in forming a negative sentence or usage of subjunctive, a specific form of verb conjugation usually used to express a level of subjectivity of the statement or bound by specific conjunctions (Paine, 2012).

Paine (2012) also describes the various *argot* used with general examples, such as "niquer", the French equivalent of the "f-word", and more region-specific examples such as *verlan,* the reversal of syllables in a word to create a new one, or Provençal. Lastly, he analyses the usage of English slang together with its possible significations and regional specifics:

"The data shows that American slang makes up 26% of all colloquial speech (the combination of French Argot, langue du bled, and English). This percentage is higher significantly in Paris (33%) than in Marseille (17%)." (Paine, 2012, p. 62)

Paine also deliberates on the purpose of using English slang. He designates two ways: the useful and the ornamental. The useful is used when expressing a word which does not have a French equivalent, whereas the ornamental is used for effect. He is also stressing that using English is not necessarily pro-American, but more of an anti-French attitude (2012, pp. 63-64).

In his work, Paine references Hassa's text (2010) – *Kiff my zikmu: Symbolic Dimensions of Arabic, English and Verlan in French rap texts,* a chapter in Languages of Global Hip Hop. She discusses the role of Arabic, English and *verlan* and in which context it is used. According to her, Arabic is mostly used in reference to the cultures of countries such as Morocco, Algeria, and Tunisia. It is also used in relation to Islam in different ways. One example would be an indication to moral values as part of conscious rap, another example would be the discourse about stereotypical viewing of Muslims and countering "Islamophobia". Hassa also mentions the words "khay" – "brother" and "khti" – "sister", which are used when establishing social contact and expressing solidarity among group members. Lastly, she refers to using Arabic when evoking nostalgia for their country of origin and difficulties of being torn between two cultures (and identities), the culture of origin and the French culture (Hassa, 2010, pp. 52–53).

When describing the context of using English in French rap lyrics, she explains that the image of American culture in France is portrayed mainly by the media and thus English words are sometimes linked to this portrayal, e.g. "Hollywood", "Starsky". The other group of words coming from English would be the terminology of the original American hip hop movement such as "beat", "vibe", "flow", and "crew". Furthermore, in the corpus analysed by Hassa, the English language is often used when degrading women using sexual references and in the context of violence. The examples of such

usage would be: “bitch”, “fuck”, “dead”, “gang”, “shoot” (verbs are also often conjugated as a French verb – e.g. “shootait”). This violence perceptible in the lyrics might be a result of the “gangsta” image presented by the American rappers as well as a need to reproduce the atmosphere of violence associated with the “banlieue” – “suburban” lifestyle. Hassa gives an example of code-switching when mentioning the Marseillais group IAM – according to her the English vocabulary present in one of the extracts is very explicit whereas the French text is softer and more poetic. Hassa also highlights the possible transcendent nature of violent English language as it might be easier to express themes linked to taboos in a different language than one’s own (Hassa, 2010, p. 57).

Finally, Hassa describes the context of using *verlan*. She describes it mainly as a means of “insiders talk” - a way for group members to recognize one another and on the contrary exclude those who do not belong. It was created out of immigrant experience in Parisian suburbs and initially was used as a playful encryption and identity indicator. It is thus more often used by Parisian rappers than elsewhere in the country (in the USA, the rap rivalry is between East and West Coast, in France it is North vs. South, specifically Paris vs. Marseille). In terms of context, she mentions that *verlan* is mostly used when referring to the socio-economic background of the suburbs and the difficulties of being an immigrant descendant facing assimilation. According to Hassa, rappers turn to *verlan* when they formulate the intricate reality of the suburbs related to among other subjects to delinquency and racial tensions. The example of such words would be: “ivé” – “vie” - “life”, “du-per” – “perdu” – “lost”, “car-pla” – “placard’ – “jail”, or “beu-her” – “herbe” – “marijuana”. Also, there are words serving as group identity markers such as “refré” – “frére” - “brother”, or “roeus” ´– “sœur”. At last, Hassa also mentions usage of *verlan* as a means of fostering the suburban identity (2010, p. 61).

Table 1. Percentage of total word count.

NSL categories	1990/1991	2001	2011
Total NSL	3.12	7.11	12.38
Abbreviations	0.21	0.65	1.47
Slang	0.18	1.34	2.32
Colloquial words	1.15	2.58	3.51
Foreign words	1.24	1.91	4.47
Verlan	0.03	0.65	1.13
Vulgar words	0.53	0.65	1.05
Combinations	0.21	0.62	1.53

Figure 6: Percentage of total word count (Verbeke, 2017)

Table 2. Percentage of total borrowings.

Foreign borrowings	1990/1991	2001	2011
English	100%	87.50	81.65
Arabic	0	6.25	13.29
Spanish	0	1.25	1.26
Romani	0	2.50	1.26
German	0	1.25	0
Bambara	0	1.25	1.26
Russian	0	0	1.26

Figure 7: Percentage of total borrowings (Verbeke, 2017)

Verbeke (2017) builds on the work of the previous two authors when presenting his diachronic analysis of non-standard language used by French rappers. The analysis focuses on describing the potential influences on the number of non-standard language words during the years as well as bringing further understanding of which non-standard vocabulary is present in the given corpus. For the analysis, Verbeke chose lyrics of rappers from the Île-de-France region.

For our research, the most important findings are: across the corpus, the usage of non-standard vocabulary increased with time and one of the accelerators of such increase is the usage of the internet and platforms such as YouTube (see Figure 6 for overview of non-standard vocabulary for given years). Also, the borrowings' origins developed from solely English origin in the earlier years to more diverse language origins later on (see Figure 7 for precise percentages).

Verbeke describes this process of continuous non-standard language development as such:

"This much faster spread of NSL thanks to the Internet means that such words lose their cryptic dimension even faster. This might be seen as a problem for its original users who will then attempt to replace these terms with other NSL words, therefore contributing to the
appearance of new and often more complex NSL. This phenomenon stems from a desire to keep ownership of the language." (2017, p. 289)

He further describes other determinants behind the development of the non-standard vocabulary usage. He mentions factors such as fear of acceptance by the wider audience at the beginning of rap in France leading to less non-standard words and on the contrary, the need to stay relevant and interesting by creative usage of non-standard language leading to the increase.

Devilla (2011) brings us a reflection on how the multilingual properties of French rap lyrics represent the multiple identities and roots of the artists. His contextual analysis of the given corpus describes the usage of borrowed words coming from various languages – English, Arabic, Italian, Spanish, Provençal, and Occitan. His findings in terms of associated topics will be further analysed when answering the question about topics present in French rap lyrics.

Specific view into the French rap vocabulary is provided by Dramé and Ndiaye (2012) when comparing French rap to poetry and proposing using French rap as a tool to learn French. Their analysis is influenced mainly by the Senegalese rap scene and evokes its development from a niche genre to a widely accepted art form. The specific language they mention apart from standard French includes: loan words from various languages such as English, Wolof, or Arabic, plays with numbers and letters, spelled out words, acronyms, abbreviations, or metaplasms (pp. 121–147).

Particularly, the recount of metaplasms is interesting in terms of non-standard language presenting words created by suffixation:

- Dolécratie': 'dolé' – French root ('doler' – flatten with a sharp object) + 'cratie' – French suffix (way of government) = government by using force.
- 'Politichien': 'politique' – French root (politic] + 'chien': French suffix (dog) = Politics who behave like dogs (Dramé and Ndiaye, 2012, p. 135).

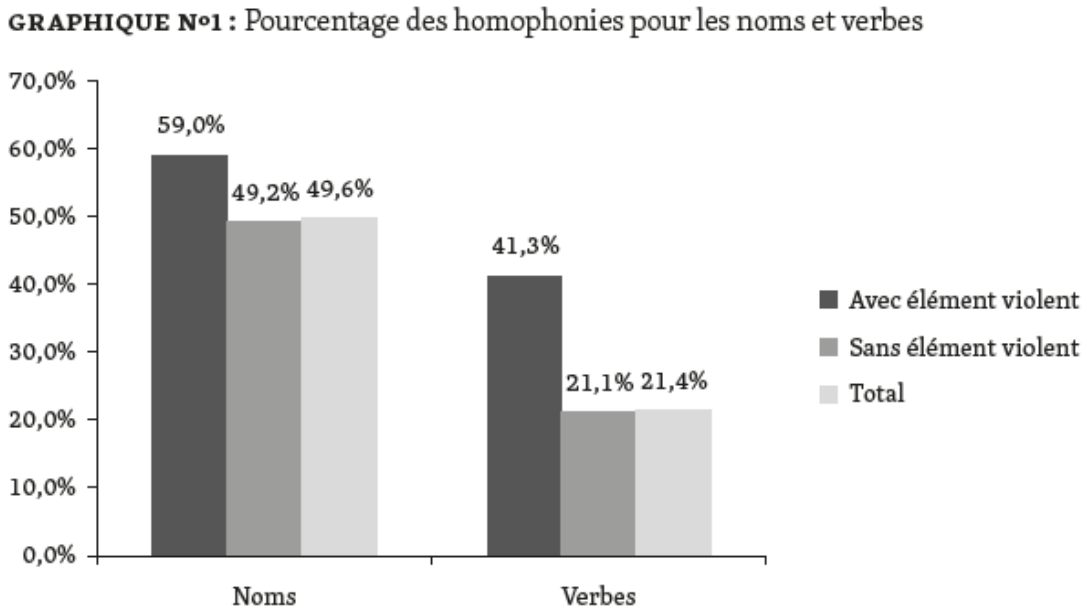

Figure 8: Percentage of homophonies for nouns and verbs (Zubčeková, 2015)

A quantitative approach to violent language present in French rap lyrics was taken by Helena Zubčeková (2015) who analysed a corpus of 97 songs by Diam's, MC Solaar, and Rohff. She argues that violent words are often used in relation to homophony – rhyme, assonance, alliteration – and serve an artistic intention / purpose. In her data set, the violent words were more often used in homophony than other words (see Figure 8 for details). She bases her argument on four premises: Firstly, the rapper supports the rhythm by homophony. Secondly, the insults have phonetic, rhythmic, and articulative value when forming rhymes. Thirdly, the usage of obscene language focuses the attention of the audience. Finally, the violence of the words supports the artistic performance by aiming to stir the emotions of the listener. (pp. 297–300)

Table 1. Non-standard language in the corpus (% of total word count)

NSL categories – % of total Word count	*Knowledge (4,528 words)*	*Jazz/poetic (2,863 words)*	*Ego trip (2,877 words)*
Total NSL	2.32	3.18	10.22
Abbreviations	0.37	0.24	1.11
Slang words	0.17	0.28	1.81
Colloquial words	1.19	1.95	3.47
Foreign borrowings	0.42	0.63	3.3
Verlan	0.06	0	0.83
Vulgar words	0.18	0.1	0.94
Combinations	0.09	0.03	1.25

Figure 9: Non-standard language in the corpus (% of total word count) (Verbeke, 2019)

Table 2. Foreign borrowings in the corpus (% of total borrowings)

Foreign borrowings – % of total borrowings	*Knowledge (19 borr.)*	*Jazz/poetic (18 borr.)*	*Ego trip (95 borr.)*
English	84.21	50	91.58
Arabic	0	0	6.31
Romani	0	0	1.05
Latin	15.79	50	0
Unknown	0	0	1.05

Figure 10: Foreign borrowings in the corpus (% of total borrowings) (Verbeke, 2019)

In order to further develop his research, Verbeke (2019) focused on how different rap genres influence the number of nonstandard words. He analysed a corpus of seven tracks per genre. The year of song release was set to 2005 in order to avoid time related determinants. For the specific results, please see Figure 9 and Figure 10.

Table 5. Themes in the selected lyrics (number of occurrences per 100 words)

Occurrences of themes per 100 words	MC Solaar (*jazz/poetic*)		La Fouine (*ego trip*)	
	'Carpe Diem' 4.42% NSL	'Le Davinci Claude' 2.62% NSL	'J'arrive en balle' 9.09% NSL	'Jalousie' 11.11% NSL
Materialism	0	0	0.45	1.66
Misogyny	0	0	0.68	0.83
Violence/crime	0.67	0.78	3.18	2.77
Political awareness	0.67	1.57	0	0
Expression of culture	0	0	2.27	1.94
Dissatisfaction with mainstream society	1.12	0	0.68	0.27
Group unity	0	0	0.45	0

Figure 11: Themes in the selected lyrics (number of occurrences per 100 words) (Verbeke, 2019)

Furthermore, he conducted a qualitative analysis of four songs by two artists – MC Solaar for "jazz/poetic" songs and La Fouine for "ego trip" songs (please see Figure 11 for details) and interviewed a number of francophone rappers to complement his research.

Verbeke suggests that different themes lead to a different language employed, which is confirmed when interviewing rapper Akro: "[In this quotation,] we can see very clearly that the artist is making a conscious decision to employ a specific language register depending on the theme and the objective of the track." (2019)

Among his conclusions is that the "ego trip" genre can be linked to higher levels of non-standard vocabulary and he suggests that one factor in particular could justify that

– the link between ego trip and violence. He explains that by ego trip genre being linked to competing over others and thus covering hostile themes and speaking to a specific target group:

"As the goal of many ego trip artists is to attack other rappers and convince the audience of their superiority, their tracks often contain adversarial themes and narratives to criticise the competitors and FCC elements to appeal to specific audiences, which translate into higher occurrences of NSL words that are associated with violence, materialism, misogyny and life in the banlieues." (2019, p. 62)

In alignment with these findings, he revisits his previously mentioned research from 2017 and claims that most of the tracks analysed as latest were of the ego trip genre. That, he claims, makes sense as the increased popularity of rap also brought higher competition between the rappers and thus increased need for non-standard language. However, he stresses that further studies on larger corpora are needed to confirm or deny his suggestions.

In order to conclude, these findings can be outlined from the literature research: despite the myth observed by Pecqueux that French rappers use only slang, *verlan* and vulgar words (Verbeke, 2019, p. 44), the French rap lyrics consist in majority of standard French language. However, there is a large non-standard vocabulary such as foreign borrowings, vulgar words, abbreviations, *verlan* and *argot*. Additionally, it seems to expand with time, vary from artist to artist, and is often related to code-switching and specific subjects directly linked to the group and individual identity.

5.7 What are the topics in French rap lyrics?

Following up on Verbeke's research from 2019, we can distinguish some topics according to genres. He also asserts that there are myths about French rap which have been already debunked by researchers such as "that French rap should be regarded as commercial music only, that all rap artists come from the *banlieues*, or that rap artists talk about hatred and themselves only." (p. 44)

Verbeke uses for his analysis three major genres - jazz/poetic, knowledge, and ego trip as he finds them the most distinctive. However, he states that the distinction is sometimes not trivial as the genres often overlap for artists, albums or even tracks. Also,

there are many other genres which can be defined in relation to the major genres and also be extensions of them in some way (2019, p. 51).

For clarity, he briefly defines the three genres as such:

- *Ego trip* – an egocentric display of abilities – in this genre, rappers practically boast about everything they can – possessions, skills, deeds, to name a few. It is also close to *gangsta* rap and rarely has any narrative.
- *Knowledge* – as opposed to the *ego trip,* it usually focuses on conveying some value by the lyrics content and often focuses on critique of societal issues, such as racism or segregation. It is also known as *message* rap in English or *rap engagé* in French.
- *Jazz/poetic* – harder to distinguish as its characteristics are less clear, named after *jazz* as it borrows from its code; *poetic* not because of poetic nature of the whole rap genre, but in this context because of the artist's desire to express feelings and beauty (p. 52)

In order to delimit the themes and be able to measure their presence, Verbeke takes over the defined themes from 'Controversial Rap Themes, Gender Portrayals and Skin Tone Distortion: Distortion: A Content Analysis of Rap Music Videos' by Conrad et al. (2009) and adapts it to the French scene:

- **materialism**: the display of wealth or consumption
- **misogyny**: sexualizing or objectification of women and/or dominance of men over women
- **violence**: threats, physical force, displaying or firing weapons, and criminality
- **political awareness**: advocating a political position or raising awareness of political or societal problems
- **expression of culture**: displaying symbols of hip hop or banlieue culture
- **disaffection with mainstream society**: showing contempt for dominant beliefs or societal pillars
- **group unity**: groups of people gathered together (2019, p. 55)

In order to identify the themes, Verbeke relied on music videos and further on established a measure of topics per one hundred words.

Violence, or perceived violence, seems to be an important subject of discussion by the researchers. Pecqueux (2004) discusses the cathartic element of violence in rap: "À partir de là on peut légitimement soutenir que la violence et plus généralement le rap servent d'exutoire, ou katharsis, à la violence réelle qui caractérise l'environnement social des rappeurs et de leurs auditeurs." – "From this, we can legitimately argue that violence and more generally rap can serve as a relief or catharsis of the real violence which characterises the social environment of rappers and their listeners." (p. 59)

Zubčeková (2015) gives an overview of researchers' view on violence. She summarises that there are three perceptions of violence in rap. Firstly, that in fact, rap is not violent at all. This idea is supported by Pecqueux who claims that the perception of violence is caused by the violent language used. Furthermore, Marti considers violence in rap as purely symbolic whereas Martínez finds the expression of violence in rap fictional. Secondly, that the violence in rap is exogenous – fabricated by the media in order to sell better. According to Trimaille, it is the scandal which sells. (p. 295) Finally, the idea, which Zubčeková supports, that violence in rap is endogenous – it is its inherent quality. For this theory, the claims by researchers can be summarised as such:

- The violence in rap is a heritage of blues and jazz (Laurent Béru, 2006).
- The violence in rap is a reaction to the real violence of the environment (Paul Yonnet, 2000; André Prévos, 2003).
- The violence in rap is an invitation to the audience to take the violence onto the city (Alain Milon, 2004; Manuel Boucher, 1998).
- The violence in rap is a contestation against the rules (Yves et Émilie Morhain, 2011).
- The violence in rap serves as a catharsis for the rapper (Véronique Petetin, 2009; Manuel Boucher, 1998).
- The violence in rap is the symbol of real violence in the streets (Cyril Trimaille, 1999).
- The violence in rap is reflecting the oppressive element of life (Morgan Jouvenet, 2006).

- The violence in rap is a means of artistic expression and technique of performance (Karim Hammou, 2005; Christian Béthune, 2003; Richard Shusterman, 2003; Anthony Pecqueux, 2007) (2015, p. 296).

Devilla (2011) recounts the themes present in French rap with the primary focus on the multiple roots the French rappers often have and their common affiliations to cité or banlieue. He recounts themes related to the places of origins or roots, themes related to life in difficult social reality, but also links to everyday life as simple as food. He also highlights the critique of society which is often aimed against the French Republic, but at the same time evoking their idealistic values citing the famous "Liberté, Égalité, Fraternité" often in a reproachful or ironic manner. Devilla (2011) puts the themes of social justice in rap into context of the French chanson seeing rap as continuation of its tradition. (p. 62)

Looking at the political themes, Sonnette (2014) analyses the opposition of pronouns "nous" – "us" and "eux"– "them" in French rap songs. She elaborates on the way French "conscious" rappers oppose themselves against the postcolonial power and successively class dominance:

"En cherchant à rassembler autour de leurs propos, ils invitent à la confrontation – symbolique – de leur camp contre celui d'en face. Précisément, ceux d'en face sont les groupes ou les personnes qui exercent une domination critiquable et à combattre. Alors que les thématiques les plus largement brassées sont celles de l'immigration, du racisme, de l'histoire coloniale et des guerres actuelles, il semble pertinent de les inclure sous l'idée plus générale de domination postcoloniale."

–

"Trying to unite themselves around their ideas, they invite to a confrontation – symbolic - of their camp against the other. To be precise, the others are the groups of persons who are exercising domination which should be fought against. While the themes in majority touched upon are those of immigration, racism, colonial history and current

wars, they seem to include them under the more general idea of postcolonial domination."
(Sonnette, 2014, p. 176)

Sonnette bases her research on songs from the 2000s and which have already been labelled as "rappeurs conscients" – "conscious rappers", "rappeurs engagés" – "engaged rappers", "rappeurs politiques" – "political rappers".
Following the political themes, Ngmaleu (2021) focuses on analysis of four texts of four rappers where the texts are politically engaged. Two among the chosen rappers, Booba and La Fouine, are from the French rap scene, both coming from immigration. Their texts describe the difficult situations the life of a low class of immigration encounters. The two Cameroon rappers Général Valsero and Maalhox le Viber also construct the analysed songs on the sociopolitical and sociomoral context.
Vicherat (2001) takes a more general approach to themes in rap. Also, he defends rap as a genre against its critics who degrade it as a lower form of art.
He focuses on several areas of analysis, bringing the themes in rap closer to the reader with specific examples. For the sake of brevity, the themes, standing out as different from those which have been already largely covered in this paper, can be denoted as: Rappers' relation to time, rappers' relation to territory, rappers' relation to religion, rappers' relation to integrity, and rappers' relation to morals.
Vicherat describes the relation to time in rap as paradoxical because, despite being innovative as a genre, it contains some sort of nostalgic or reactionary relationship to the past. Seeing the world as degrading itself with new problems such as AIDS or unemployment, they look to "before" in a nostalgic manner. There is also a reference to the beginnings of rap being idealist in contrast with the increasing competition in rap. Also, there is a perspective that life in the (sub)urban areas degrades to violence which was not there so heavily present before. To conclude, according to Vicherat, the nostalgia of the past expresses the fear of the future as well as a profound state of sadness linked to their current situation.
In terms of the relation to territory, Vicherat states that it is particularly pronounced – it is very hard to find a rapper who would not be talking about their neighbourhood both in positive and negative manner. There is pride when talking about their quarters, but

there are also themes of segregation and violence, referring to their neighbourhoods as "ghettos", similarly to the American rap. What Vicherat highlights is the fact that the rappers refer to their quarters almost as to a living entity possessing features and characteristics. Despite the negative references, the view of the rappers' place of living seems to be rather positive, referring to the negativity as "hardness" which reinforces its inhabitants. Generally speaking, the positive image of their quarter or town seems to stem from the sense of belonging and is stronger than any other affiliation including hip hop culture. The other territorial theme present in rap is belonging to "somewhere else" referring to the rappers' roots.

When discussing the theme of religion in French rap, particularly Catholicism and Islam, Vicherat describes it as a possible way of taming the violence by following specific role models. Secondly, it can be a way for the rappers to connect with their cultural heritage. He also draws a parallel between rappers and preachers, suggesting a formal similarity between mystic litanies and rap songs.

When referring to rappers' relation to integrity, Vicherat describes the importance of a set of rules and authenticity. For example, MC Solaar was condemned for his abandonment of authentic content by his preference to verbal artistry. According to Vicherat, preserving authentic content of the songs, particularly in terms of social justice, is vital for rappers. Another situation where rappers can be particularly judged is when they create purely for commercial success.

According to Vicherat, rappers have a particular relation to morality. For many, rap seems to be a means of getting out of life's difficulties (in contrast with American rap), evolving to almost a form of moral advice. Some of the older rappers feel especially "qualified" for giving advice on how to get out of a violent lifestyle as they have been through similar tracks. Similar attitudes can be observed towards drugs where "soft drugs" are generally accepted whereas "hard drugs" are condemned. Other themes include protection against AIDS or recommendations on work ethic – how to become a rapper and make a breakthrough.

To summarise, Vicherat provides a qualitative analysis of themes which serves as a good complement to the sociological reflections and quantitatively oriented studies

Finally, Diallo (2009) tries to debunk a myth that rap is just a form of resistance. He insists that rap contains a plethora of topics and cannot be reduced to only a theme of

resistance. He quotes Costello and Wallace to support his argument of how many different topics are present in rap: 'How bad / cool /fresh /def the rapper and his lyrics are; how equally un-all-these his musical rivals are; how troublesome, vacuous, and acquisitive women are; how wonderful it is to be « paid in full » for rapping instead of stealing or dealing; how gangs are really families and in particular, how sex and violence and yuppie toys represent perfectly the urban life drive' (Costello and Wallace, 1990, p. 24) He affirms that viewing rap only as a form of resistance to oppression comes largely from researchers' ideological bias and that the originality and expression of rap cannot be so simply reduced.

To conclude, the answer to our research question by the literature can be defined as such: The French rap lyrics contain a variety of themes where some are more prominent than others. The most commonly discussed seem to be:

- "ego trip" – the boasting of self, focus on material satisfaction, women and money, and rivalry with others
- "the conscious or political rap" – focusing on themes of social justice and protest against the ruling class
- "violence" – both in describing real-life events and as a way of emotional catharsis
- "territory" – themes of quarter/town pride, reference to roots, or hip hop affiliation
- "ethical" – followed by the themes of religion, ethical teachings
- "everyday life" – references to the daily life in all its forms

However, it needs to be stressed that the list is not exhaustive and the themes are also continuously evolving.

5.8 Master Theses Cluster

The master theses on the given subject were mostly collected from the publications using Rap Corpus from the Masaryk University, Brno (for more information about RapCor, please see the section Terminology or Data collection), but also from the Charles University. They are all written in French, which suggests a high engagement in

the matter on the Czech academic scene. For the purpose of this paper, the master theses were mostly used as an inspiration for the methodology and to make the research more informed in terms of recently published works of the same academic level on the same subject. However, studying the master theses cluster was not aimed at drawing any conclusions for the research.

The master thesis of Pavla Kudličková, 2009 – *Français contemporain des cités dans les chansons de rap – Contemporary French of the Housing Estates in Rap Songs* – describes the richness of the vocabulary of housing estates of large French cities. She puts it into the context of immigration, foreign languages influences, hard life conditions, and collective identity. Her main goal is to describe the ways of how the new vocabulary is created in this ambiance and if there is any means which is more productive than the others. She concludes that in comparison with the traditional argot, there is a higher number of borrowed words from other languages. Nevertheless, it is not significantly higher. Thus the author believes that there is no way of neologisms creation favoured nor particularly more productive.

The work of Pavlína Závodská, 2010 – *Vulgarismes dans un corpus de chansons de rap : étude lexicométrique en synchronie dynamique – Vulgarisms in rap songs: a lexicometric study in dynamic synchrony* – studies vulgarisms in French rap songs from the lexicometric point of view. She asserts that vulgarisms in French rap do not reflect low levels of cultivation. On the contrary, they are part of a strategy of how to shock the listener and draw attention to social issues. She bases her study on comparison of corpora from years 1990–1994 and from years 2005–2009. Besides findings on types of vulgarisms, ways of forming them, and the usage of borrowed words, the main result of the comparative study is that the later corpus contains more vulgar words (which, from our point of view, would be in alignment with Verbeke's findings about increasing usage of non-standard language with time).

In her work, Anna Zelenková, 2013, *Arabismes dans les chansons de rap français : traitement lexicographique, adaptation phonique et rôle de l'origine des rappeurs – Arabisms in French rap songs: lexicographic analysis, phonic adaptation, and the role of the rappers* – examines the role of Arabic language in French rap songs. In formulating her hypotheses, she focuses mainly on the difference of employing Arabic terms by rappers of Arabic origins in comparison with rappers of other origins. She

concludes that a higher percentage of Arabic borrowed words in the corpus comes from the rappers who are of Arab origin. She also affirms low levels of dictionarisation of Arabic borrowed words in French dictionaries.

The paper of Dana Ondrušková, 2014 – *Lexique de la drogue dans le corpus des chansons de rap: analyse sémantique en synchronie dynamique – The drug vocabulary in rap songs: semantic analysis in dynamic synchrony* – focuses on the vocabulary related to drug use in rap songs. It was mostly inspirational in the lexicometric part where the author analyzes drug related words frequencies. Otherwise, the thesis focuses on the creative potential of the necessity of ciphering the drug related vocabulary by its users. Thus, it explores further the non-standard vocabulary usage, its dictionarisation and its development in French rap. It also explores and confirms several hypotheses. Firstly, she affirms that the dictionarisation of non-standard vocabulary related to drug use is relatively low in the time of her research. Secondly, she claims that the dictionarisation depends on the word's origin and on the way the word was created. Thirdly, based on her research, the two genres – "gangsta rap" and "conscious rap" have comparable crypto-ludic character in terms of the drug related vocabulary, but the difference is thematic and contextual.

5.9 Discussion

The biggest limitation of the conducted literature review is the geographical bias of the researcher. The subject is analysed externally through the prism of a different culture, and thus to some extent limits the research also by restricted access to the academic resources issued in France. However, it can also serve positively when evaluating the findings with more distance. What also needs to be considered is that this literature review is, metaphorically speaking, "a tip of the iceberg". It does not aim to be exhaustive in terms of all the possible sources, but to specifically address the questions asked. It includes a vast amount of research and references, on which the analysed works were previously built, but those references were not explicitly included in the review. To name a few out of the many influential researchers, the works of Manuel Boucher, Christian Béthune, Debov Valéry, Hugues Bazin, Cyril Trimaille (or Alena Podhorná-Polická from the Czech academic context), are important to the research as

they have been cited frequently or present in the reviewed work, but were not necessarily included in the literature review as individual entries.

6 Data collection

6.1 Description of the used solution

6.1.1 Preliminary research and chosen lyrics source

First of all, a research of potential lyrics sources and collection methods was conducted. There were several options considered such as manual lyrics search through the search engine google.com, checking lyrics sites for APIs which would provide the songs' data, or ready-to-be-used lyrics corpora.
The manual search was quickly dismissed as it would be too time-consuming and would not allow for collecting large amounts of data repeatedly. Thus, only automatic lyrics sites scraping and already established corpora were further examined as a feasible approach.

Genius.com choice

Based on initial search, several popular lyrics sites were evaluated, such as lyrics.com, azlyrics.com, lyricsfreak.com, metrolyrics.com. None of them were considered as usable since they either did not have an API to exploit or they would not contain sufficient amounts of French rap lyrics. After further research of possible solutions, a study of French rap lyrics was discovered – R.A.P. – Rap Analysis Project (2015) which was using Genius.com as its source.
Based on this study, several other sources confirming Genius.com were found on github.com. For example, one code snippet (Ng, 2015) shows a solution of searching the Genius.com API for songs, and the article *Getting Song Lyrics from Genius's API + Scraping* by Jack Schultz (2016) describes a way to search the lyrics, scrape and parse the HTML content. Because of these successful examples, Genius.com was chosen as the preferred lyrics source.

Already established corpora

Two other possible sources were considered before finally deciding for the Genius.com option.

Rapresearchlab.com

Rap Research Lab is a community based project using rap lyrics as an analytical tool to be used in fields such as education or art. It contains tools such as the Rap Almanac Database which contains more than 500 000 rap lyrics and their metadata or Mapper's Delight, a visualisation tool to represent semantic relationships. All the tools look excellent, but the focus is not on French rap and it was not easily accessible (Rap Research Lab, 2020).

RapCor

RapCor is a French Rap Corpus project started by docent Alena Polická from Masaryk University in Brno, the Czech Republic. It is available through Sketch Engine and it is an excellent source for French rap songs lyrics analysis since all the songs are manually collected either directly from the CD booklets or verified internet sources. It was used for the preliminary method research to try out calculation of word frequencies and work with concordances. However, in the end, the research needs of this paper did not fully align with using it as a primary source for the analysis.

6.1.2 Selection of performers and bands

The artists and bands were not chosen from academic sources in order to guarantee a wide selection of artists. The first step was automatic collection of the names by genre from Wikipedia.org. The Wikipedia search was done through Wikipedia API through a client written in JavaScript. The search included the following categories:

- Rappeur français
- Rappeur belge
- Groupe de hip hop français
- Groupe de hip hop belge
- Rappeuse belge
- Rappeuse française

The results were reviewed for duplicates and artists from an additional source, a curated list by one of the subscribers from the website SensCritique (Artzgild, 2020) were added. Followingly, more manual editing had to take place in order to be able to search

the Genius.com API. During the trial searches, not all Genius artist IDs were collected due to the difference between Wikipedia names and Genius.com names. Some of the performers have several artist names and not always the correct results were found, but where possible, a Genius.com artist name was added to compensate for the differences.

6.1.3 Collection of Genius.com data

Genius.com provides a reliable API which offers several functions. The most important feature used for this research is the search through keywords. The querying of the API was again performed by a JavaScript client and the results were collected in JSON format. First of all, the artists Genius.com IDs were collected. The queries for song IDs were performed by looping over the artist IDs and requesting artist data from the API. This was again collected in JSON format and used for further querying of songs' content.

6.1.4 Collection and parsing of the lyrics

As the Genius.com API provides the hyperlinks to the songs, but not the lyrics, a different mechanism had to be devised in order to collect the songs texts. Looping over the songs IDs, the songs data were collected and with another loop over the songs data, harvesting of the songs hyperlinks was performed and used for an HTTP request collecting the HTML content with songs lyrics. In order to parse the HTML content and extract the lyrics, a JavaScript library was used. The process required several adjustments and code conditions as the HTML content provided by the Genius.com is not consistent and alters regularly in order to prevent machine collection and copyright breaches. To view the whole Genius.com API documentation for more details, please visit https://docs.genius.com.

The resulting JSON data were converted into the CSV format in order to be easily processed in Excel and converted into R and Python data frames. See Appendix 1 to review the resulting list of songs which contains Genius.com artist ID, artist name, artist corpus ID, Genius.com song ID, Genius.com URL, song lyrics, and if available release date. To inspect the code or list of JavaScript dependencies, please visit the code repository on GitHub, https://github.com/julieklimentova/le_rap_francophone.

6.1.5 Media words list collection

The media related words list was created based on brainstorming sessions, studying internet resources, and representative lyrics. The list of words is definitely not exhaustive, and it was intentionally limited to enable higher focus. The list of words contains spans from words such as 'livre' – 'book' and 'journal' – 'newspaper' to words such as 'Facebook' and 'Internet' (please see Appendix 2).

For better orientation, the list was divided into these categories:

- Social Media
- Video & Cinema
- Chat & Communication
- Devices
- Sex & Porn
- Internet
- Virtual Assistants
- Technology
- Computing
- Image
- Music & Music making
- Shopping
- Information
- Gaming
- Smartphone
- Personalities
- Groups & Projects

7 Data analysis

7.1 Analysis tools

At the beginning of the process, research was conducted to identify correct technologies to obtain the research goals. Based on analysing several code samples and related libraries, two possible approaches were identified – writing the analysis scripts in R programming language or in Python programming language. The libraries which were considered for the analysis were: Tidytext by Julia Silge and David Robinson, UDpipe developed at the Institute of Formal and Applied Linguistics, Charles University, Gensim by Radim Řehůřek and Petr Sojka, and LDAvis by Carson Sievert.

Tidytext is a R library which uses tidy data principles as presented by Wickham (2014). The tidy data format can be defined as a table with one token per row (Robinson and Silge, 2017). Robinson and Silge further define a token: "A token is a meaningful unit of text, such as a word, that we are interested in using for analysis, and tokenization is the process of splitting text into tokens." The tidy data is in contrast with other ways of storing data for analysis, such as strings, corpus, or document-term matrix (Robinson and Silge, 2017). The package allows a large amount of analysis approaches – measuring word frequencies, calculating tf_idf (inverse document frequency), sentiment analysis, and topic modelling.

"UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files." (Charles University, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, 2017) It was developed by Milan Straka at the Institute of Formal and Applied Linguistics, Charles University. It is available as a binary, as a library for Python, R, and several more programming languages. The first necessary step is the annotation using a trained UDpipe model for a given language. There is a ready selection of open source UDpipe models for both living and ancient languages, or one's own model can be trained with the help of the UDpipe library. The resulting annotation is in CoNLL-U format. The documentation of the R package is easy to understand with practical steps to perform tasks on your corpus, such as parts of speech statistics, word frequencies based on parts of speech tag, topic modelling, or keywords extraction.

"Gensim is a free open-source Python library for representing documents as semantic vectors, as efficiently (computer-wise) and painlessly (human-wise) as possible." (Řehůřek, 2021) It is designed to process plain text using machine learning algorithms such Word2Vec, FastText, LsiModel, and LdaModel. It was designed and developed by Radim Řehůřek and other contributors such as Gordon Mohr, or Misha Penkov (Řehůřek, 2021).

LDAvis is an R package designed for interactive visualisation of topic models. It can be used to explore the topics created by the LDA algorithm through an interesting interactive form while exploring the word frequencies within the topic and the whole corpus. It also uses the "saliency" introduced in the paper *Termite: Visualization Techniques for Assessing Textual Topic Models* (Chuang et al., 2012) which estimates the word relevancy for the given topic and thus creating more or less distinctive topics based on the scale from 0 to 1. A detailed demo analysis on movie reviews can be found on the package's website (Sievert, 2018).

As the aim of this thesis is an introductory insight into the topic, the tools to be chosen needed to be comprehensive and offer a steep learning curve. After several exploratory coding exercises, the R programming language and libraries Tidytext, UDpipe and LDAvis were chosen as the main tools for the analysis for their ease of access and at the same time expected quality. Other R packages such as ggplot2 or Lattice were used. Gensim Python library was also used mainly for the topic modelling section. A full list of dependencies can be inspected in the code repository on GitHub, https://github.com/julieklimentova/le_rap_francophone.

7.2 French rap main corpus analysis

7.2.1 Word frequencies

For the word frequencies analysis a comparative analysis with French Web Corpus 2017 (Sketch Engine, 2017) was conducted to establish a clear distinction between the French rap main corpus language and more commonly used language.

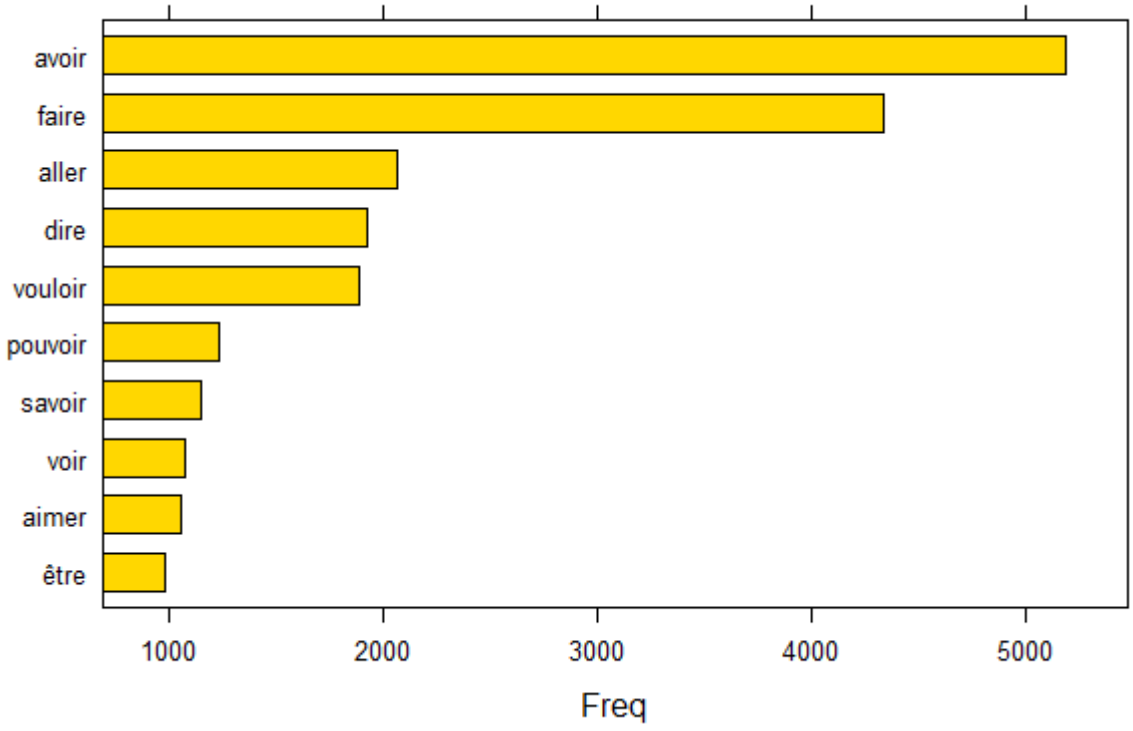

Figure 12: Most frequent verbs in the main corpus

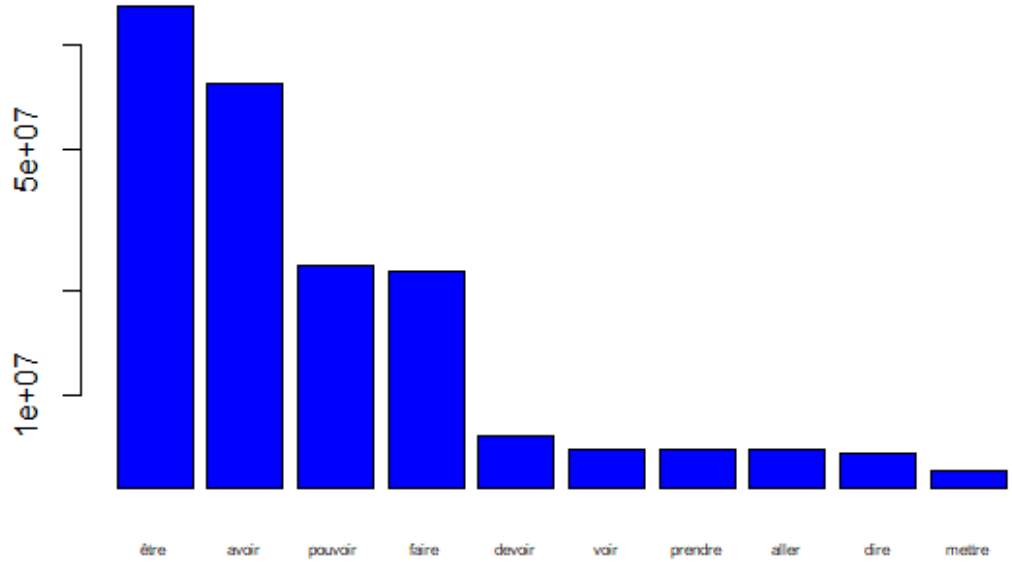

Figure 13: Most frequent verbs in the French web corpus 2017

Looking at the top ten verbs occurring in both corpora (Figure 12 and Figure 13), we can directly see some differences in the significance of certain words. The verb 'avoir' – 'to have', is more common in the main corpus, whereas the verb 'être' is less common in the main corpus than in the French web corpus 2017. From that, we can possibly insinuate that the main corpus texts might have more focus on defining oneself by what can be owned than what one is considering oneself to be. Another interesting difference is the frequency of the word 'dire' – 'to tell', which again has more weight in the French rap corpus implying that the verbal expression is an important topic of the corpus. Finally, the words 'aimer' – 'to love', and 'savoir' – 'to know' are present in the top ten main corpus verbs, but not at all present in the French web corpus. In the case of 'aimer', we can again imply that love is one of the most common topics in the French rap main corpus, in a similar way as in pop music. In the case of 'savoir' there might be focus on transmitting information as well as defining self by gaining some sort of knowledge or wisdom.

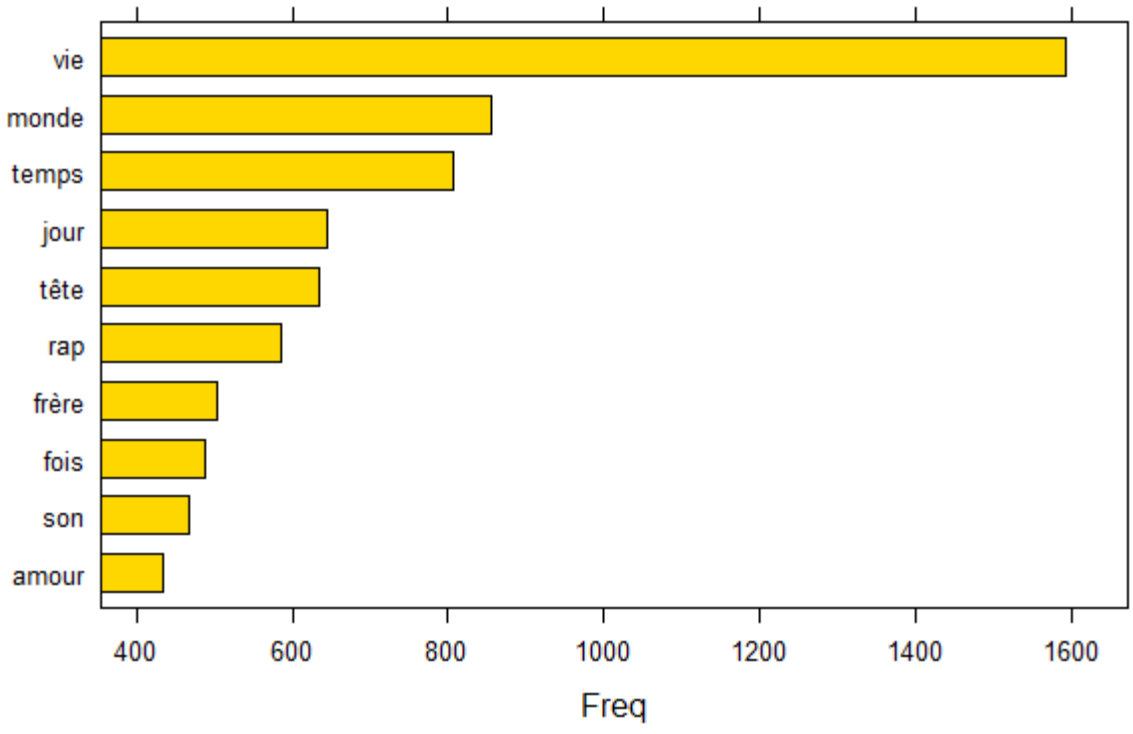

Figure 14: Most frequent nouns in the main corpus

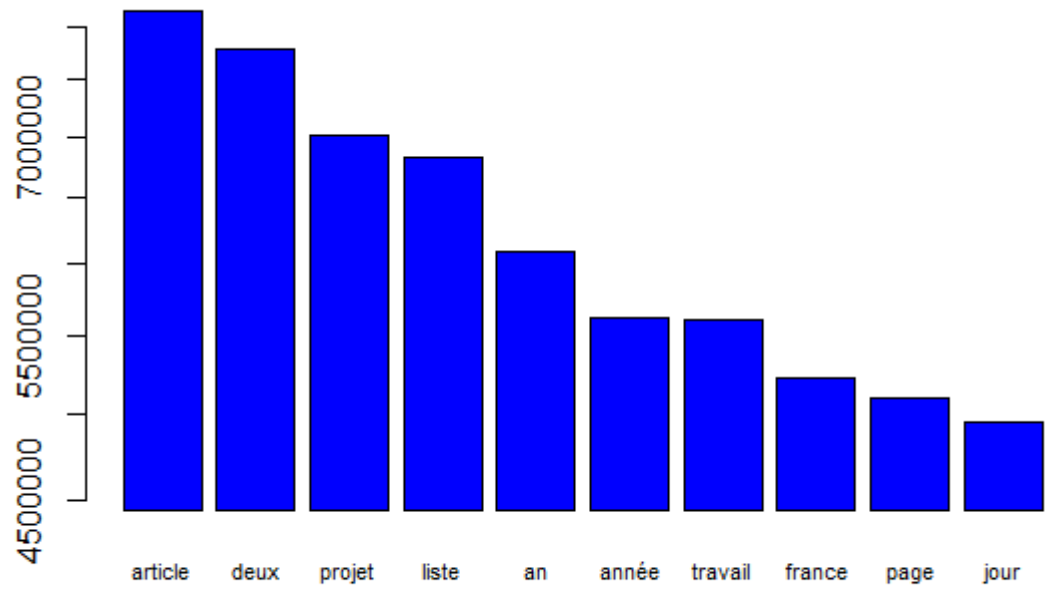

Figure 15: Most frequent nouns in the French web corpus 2017

Repeating the same process for nouns (see Figure 14 and Figure 15), we can again clearly distinguish the difference in topics between the corpora. This time the difference is even clearer since there is only one overlap between the word sets, the word 'jour' – 'day'. From a high level, we can evaluate that the main corpus includes again reference to 'love' – 'amour' and then words which belong more to the subject of reflection about life, world, and others (e.g. 'vie' – 'life', 'monde' – 'world', but also in an expression 'tout le monde' – 'everybody', 'temps' – 'times), to the subject of rap itself (e.g. 'rap', 'son' – 'sound'), a dialogue (e.g. 'frère' – 'brother', used when talking to someone familiar). On the contrary, the French web corpus 2017 collection of top ten nouns suggests focus on more formal subjects such as work (e.g. 'travail' – 'work', 'projet' – 'project'), state affairs (e.g. 'France').

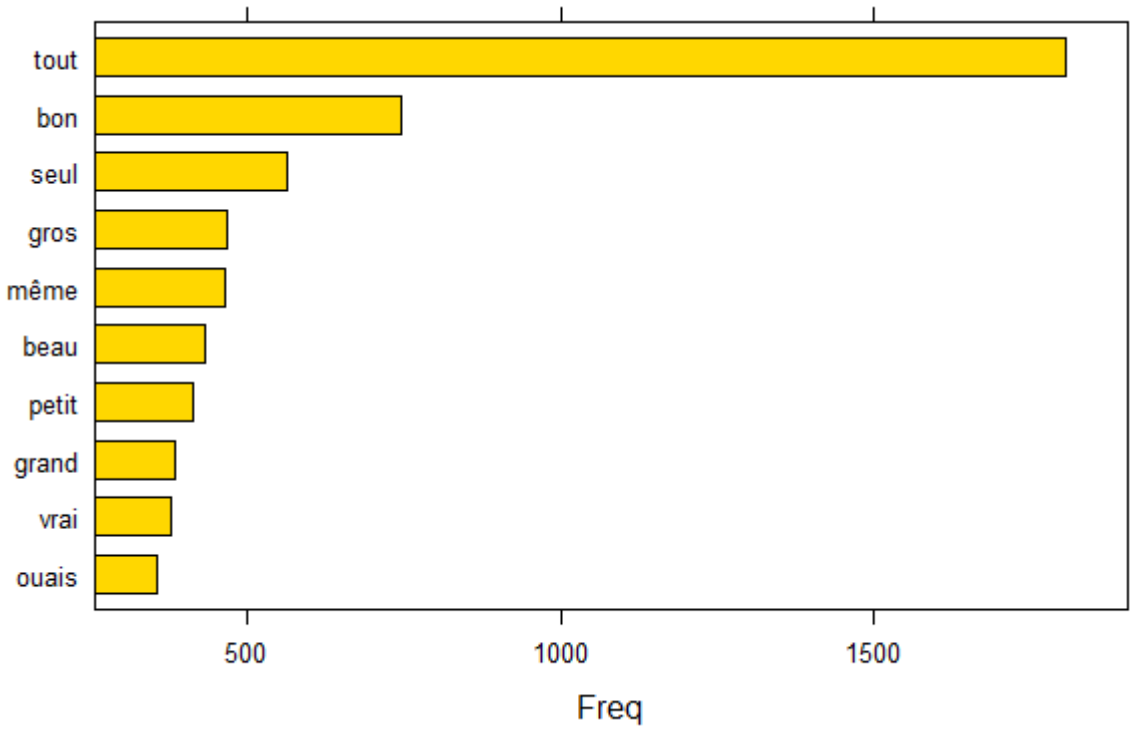

Figure 16: Most frequent adjectives in the main corpus

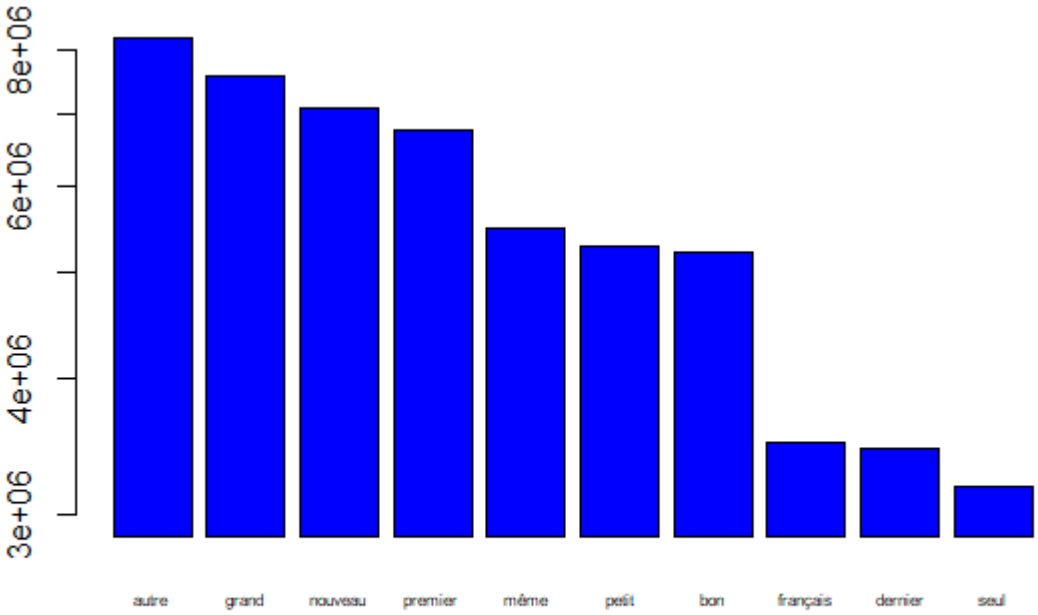

Figure 17: Most frequent adjectives in the French web corpus 2017

Finally, the comparison of adjectives marks the least significant differences (see Figure 16 and Figure 17). The composition of top ten words are rather similar to one another, moreover the words are of rather generic character and not clearly distinguishing any topics (e.g. 'petit' – 'small', 'grand' – 'big', 'bon' – 'good', 'seul' – 'alone'). Also, the word 'ouais', a colloquial version of 'oui' – 'yes', has been wrongly tagged as an adjective by the udpipe library parts of speech tagging.

7.2.2 Bigram frequencies

The first step to analysing bigram frequencies was to remove bigrams containing stopwords in both French rap corpus bigrams and French web corpus 2017. Please see the resulting tables as Appendix 3 and Appendix 4. The composition of words gives a clear distinction between the two corpora as the French rap corpus top ten results contain mainly interjections, whereas the French web corpus 2017 top ten results consist primarily of verbs and nouns specific for more formal communication.

7.2.3 Conclusion

The French rap corpus is different from the French web corpus 2017 mainly in the selection of words reflecting its topic as well as the parts of speech composition, which was significantly visible in the bigrams analysis, where interjections were prevalent.

7.3 Subcorpus analysis

A subset of the main corpus was created by identifying the songs which contain media related words.

	Main corpus	Media words subcorpus - Higher ambiguity	Media words subcorpus - Lower ambiguity
Number of songs	5061	885	817

Figure 18: Table of songs count for the corpora

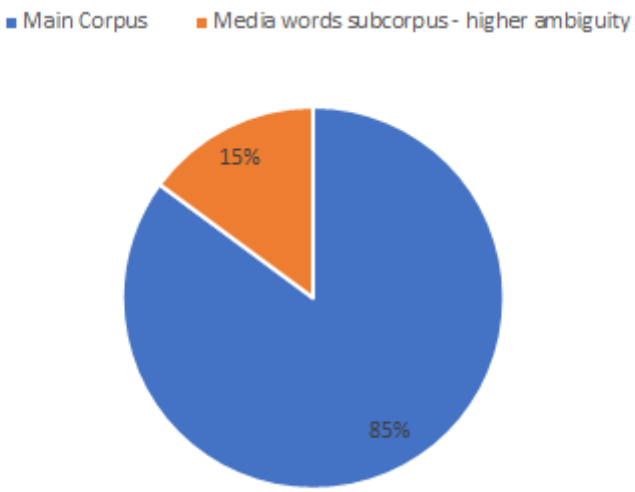

Figure 19: Main corpus – media words subcorpus ratio – higher ambiguity

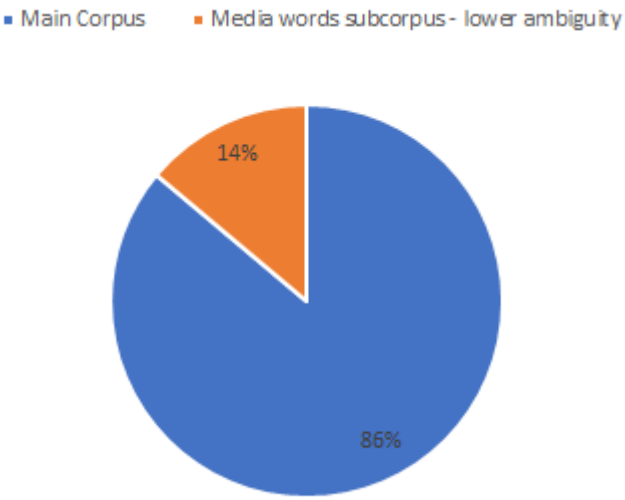

Figure 20: Main corpus – media words subcorpus ratio – lower ambiguity

There are two subsets differing by higher / lower proportion of words which can potentially fall into a different topic than only media (e.g. 'ami' – 'friend', 'système' – 'system'). Those words will be further referred to as 'ambiguous'.

In both cases, the subsets containing media related words comprise a significant part of the main corpus – approximately one sixth of the songs (please see Figure 18 for precise numbers and Figure 19 and Figure 20 for percentages).
For subsetting, a list of media related words was chosen based on research of the lyrics and considering the cultural context as described earlier. However, the subset was based on the words which already had some frequency in the corpus. Please see Appendix 5 for media words frequencies with higher ambiguity and Appendix 6 media words frequencies with lower ambiguity.

7.3.1 Media words subcorpus – lower ambiguity

For the main analysis, the media words subcorpus with less ambiguous words was chosen (further on referred to as media words subcorpus) for being more representative.

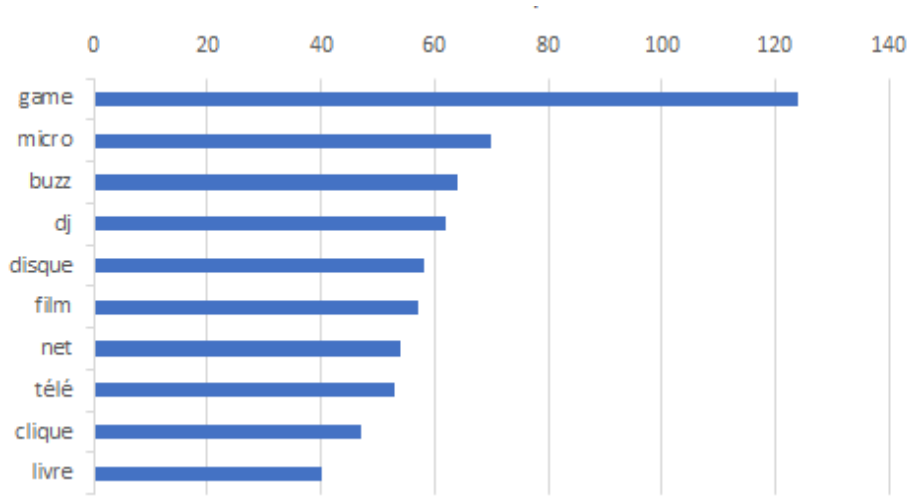

Figure 21: Media words frequencies

Figure 21 shows ten most frequent media words in the subcorpus. As rap and hip hop have their origins in the United States, it is natural that the language contains borrowed words. Even more so for media related words since a lot of them are used in their English original form (e.g. smartphone). To understand better why the word 'game' has the most occurrences, the context analysis was necessary. Looking at several examples,

the word game is the most often used as its French equivalent 'jeu' and it is notably used as a metaphor when referring to the rap environment, specifically rap battles and the rivalry between the rappers, e.g. 'rap game'.

»T'es la caricature de toi-même, la copie de la copie d'untel. Au final, qu'est-ce que t'amènes ? Une pierre de plus à ce fichu rap game. »

–

"You are your own caricature, a copy of someone else. At the end, what are you bringing? Just one more stone in this damn rap game." (Guizmo, 2013)

Another notable usage of the word game is in a bigram 'game over' which is a name of two albums *Game Over* and *Game Over 2* by 50k Editions.

Since the music industry is closely linked with technological development, the top ten words contain references such as 'micro' – a shortcut for 'microphone', 'disque' – 'disc', and 'dj'. Other topics present are TV and film making – 'télé' – 'TV', 'film'; and internet – 'net', 'clique' – 'click'.

'Clique' can however refer to 'a bag' or 'a gang' as well and 'clic clac' means a sofa bed, so it is in this case quite ambiguous. In the texts, it is however also used in the context of the internet and it can be used as a play of words. The example would be a video / statement by Morsay, *Message à Internet* (2008). *"Envoyez-moi des commentaires, faites moi de la Pub. Cliquez sur ses clips bande de fils de pute!" – "Send me comments, advertise me. Click on my clips, you bunch of sons of a bitch."* Here the word 'cliquez' is used in its regular context. Later on in the same video, the rapper plays with the same sound of the words: *'Je la mets dans mon clic–clac.' – 'I put her on my sofa'*. The last two words in the top ten list fall into different topics – 'buzz' refers mainly to the effect of talking loudly about something (in the media, but also in person) or to a buzzing sounds, whereas the word 'livre' – 'book' is tenth and represents the old media on the list.

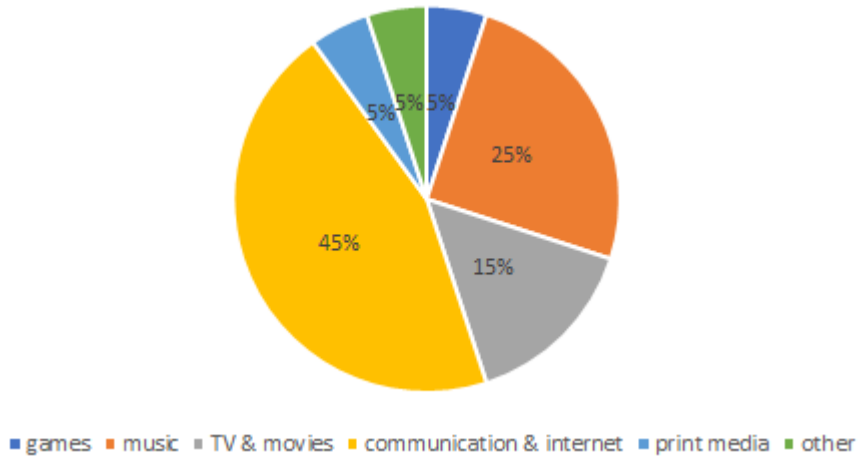

Figure 22: Topics in top twenty media words

To further analyse the topics represented in the most common media words, the top twenty media words were divided into distinct topics (see Figure 22 for the graph representation). Even though the first internet & communication word appears only 7th, this topic is the most significant. Even the music category, which would seem more probable to be the largest, comprises only 25% of the top twenty media words.

Song Title	Musician	Song corpus ID	Media words count
Message à Internet	Morsay	MORSAY11	37
Bad Buzz	Liza Monet	LIZA_MONET5	20
Clique Dance	Morsay	MORSAY4	17
Legal Geneva	Liza Monet	LIZA_MONET20	16
Les Nouvelles Synthèses	Spoke Orkestra	SPOKE_ORKESTRA16	15
Follow Me	Kobo	KOBO18	12
L'arme secrète de laser	L'Armée Des 12	L'ARMÉE_DES_128	11
Freestyle baise le game	Ladea	LADEA18	9
Change leur game	Tito Prince	TITO_PRINCE14	9
Baby Mama	Small-X ft. Di-Meh	DI-MEH8	9
Panorama	Vincent Delerm (Ft. Alain Souchon, Aloïse Sauvage & François Truffaut)	ALOÏSE_SAUVAGE11	9

Figure 23: Top eleven songs with the highest media words count

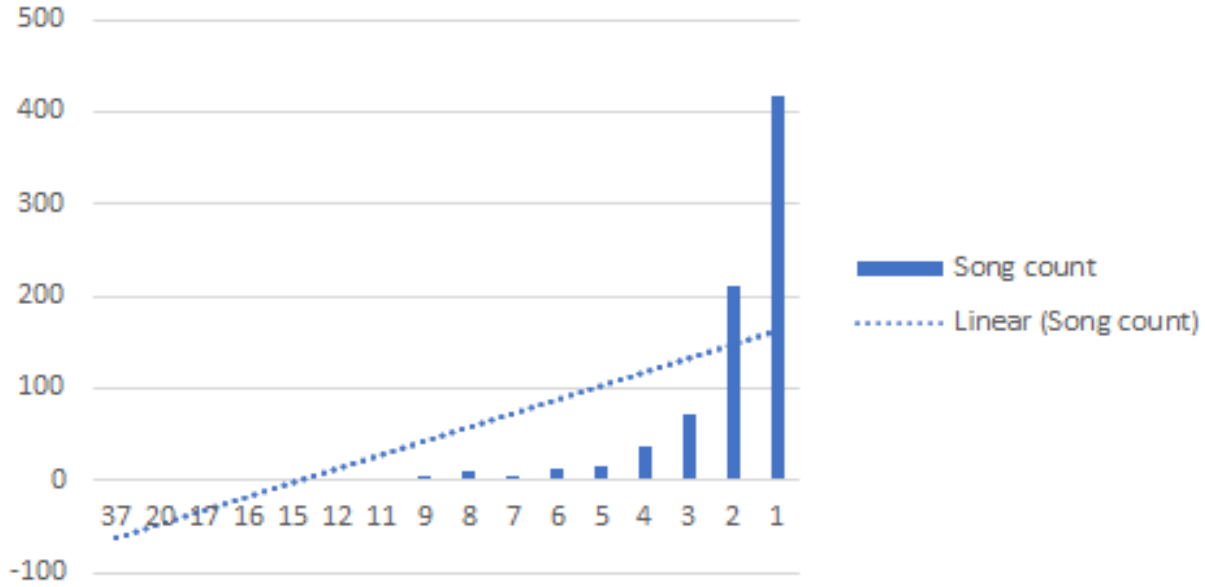

Figure 24: Song count – media words count relationship

Media words count	Song count
37	1
20	1
17	1
16	1
15	1
12	1
11	1
9	4
8	11
7	5
6	13
5	15
4	37
3	71
2	211
1	416

Figure 25: Table of media words count per song count

To analyse how the media words are distributed throughout the subcorpus and in which context, a table with a number of media words occurrences was created. Above, the top ten word counts can be inspected (please see Figure 23). Please also see Figure 24 for a graph depicting the relationship between song count and number of media words and Figure 25 for a table of media words count per song count). You can also refer to Appendix 7 for precise data.

Aforementioned *Message à Internet* by Morsay has the highest count of 37 media related words. To be precise, in this case it is not a song. It is an approximately 10 minutes long

statement commenting on and criticising (insulting) the people commenting on his videos on the internet and the rappers on French mainstream rap radio network Skyrock. However, it is still included as an original text in the Genius.com database. Considering that, it can be safely concluded that the activities on the internet played an important role in the rapper's life in 2008.

Morsay's song *Clique Dance* is also present on the list of songs with the most media related words. The word 'clique' here refers to the mechanical action of clicking together with the stereotypical actions in a dance club.

Another musician with a high frequency of media words is Liza Monet. Her common word is 'buzz' referring to the bad gossip spreading throughout the rap community and media (Liza Monet, 2020). Another interesting point in her song *Legal Geneva* is making a metaphor when comparing providing (online) sexual services and performing music (Liza Monet, 2020).

An obscure story telling takes place in the song *Les Nouvelles Synthèse* by Spoke Orchestra. It contains 16 media related words and that is mainly due to portraying the feeling of excitement when viewing events through camera.

Despite the title, in the song *Follow me* by Kobo it is not entirely clear why this English bigram was chosen, and thus it falls into the category of ambiguous meaning. Same goes for the song *Ailleurs Higher* by Aloïse Sauvage, 2017 where the word 'follow' is definitely not used in the media related context. The text *L'arme secrète de laser* by L'Armée Des 12 was not able to be fully identified since there is not sufficient information to be found online. However, it contains several media words such as 'laptop', 'programme', 'vidéo'; or 'livre'. In the song *Freestyle baise le game* by Ladea, the artist uses the word 'game' as a metaphor for the rap competitive environment and same goes for the usage in *Change leur game* by Tito Prince. In fact, the bigram "rap game" is a relatively common bigram as its count is 20 in the subcorpus, and as already mentioned in the section Media words frequencies, the word 'game' is the most common media word in the corpus. When referring to the rap game/game, the rappers mean the competitive spirit among the rappers where everyone wants to be the best, sell the most. There are multiple events in the rap culture which demonstrate this principle, such as battles or freestyles (Culturap, 2021). To investigate the usage of the metaphor further, the characteristics of the game elements of the rap culture can be analysed. If we look at the formal definition of a

game, there are several options to choose from. For the purpose of this analysis, the definition of a Danish game researcher Jesper Juul was chosen. "*A game is 1) a rule-based formal system with 2) variable and quantifiable outcome, where 3) different outcomes are assigned different values, 4) the player exerts effort in order to influence the outcome, 5) the player feels attached to the outcome, and 6) the consequences of the activity are optional and negotiable*." (Juul, 2003) We can compare the individual points in order to check if they qualify.

1. *A game is a rule-based formal system* – Each rap battle has its specific rules by the organisers, there are even entire battle leagues such as King of the Dot or Ultimate Rap League (Rap Wiki, 2022).
2. *With variable and quantifiable outcome* – The participants compete with the most innovative lyrics and technical ability while improvising. There is either a judge appointed to decide the winner or the reaction of the audience indicates the win/loss.
3. *Where different outcomes are assigned different values* – The win is a positive outcome and the loss is negative (Game studies Wiki, 2022).
4. *The player exerts effort in order to influence the outcome* – That completely applies as the rappers have to exert an enormous effort to win the battle.
5. *The player feels attached to the outcome* – The rapper is attached to the outcome as it can influence his status among the rap community.
6. *The consequences of the activity are optional and negotiable* – Even though often rap battles have real life consequences, they are primarily a source of amusement. "A specific playing of a game may have assigned consequences, but a game is a game because the consequences are *optionally* assignable on a per-play basis." (Juul, 2003)

Visibly, there are strong game elements present in the rap community rites which explain the frequent usage of the word game. At the same time, this phenomenon demonstrates the interconnectedness between real world and media related terms.
In the song *Baby Mama* by Small-X and Di-Meh, the used word is 'call' which further highlights the high usage of borrowed words. The last song Panorama by Vincent Delerm (Ft. Alain Souchon, Aloïse Sauvage & François Truffaut) can be questioned in terms of genre as it was largely selected for the corpus due to the featuring by Aloïse

Sauvage. However, for the completeness of information, the word 'films' is the used word.

The relationship between the media words count and song count is close to inverse – the less media words the higher song count.

Word frequencies

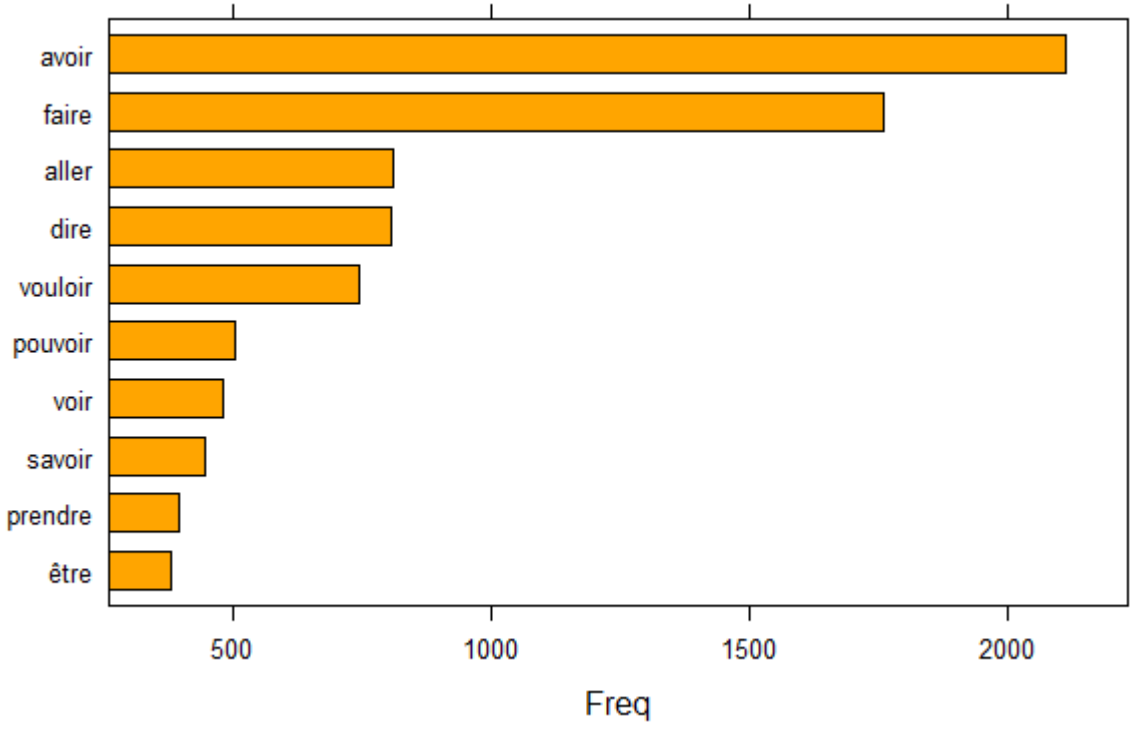

Figure 26: Most frequent verbs in the subcorpus

The list of most frequent verbs is close to identical with the main corpus. The only notable difference is the swapped rank of words 'savoir' – 'to know' and 'voir' – 'to see' and the fact that the word 'aimer' – 'to love' in the main corpus is exchanged with the word 'prendre' – 'to take' in the subcorpus (please see Figure 26 above).

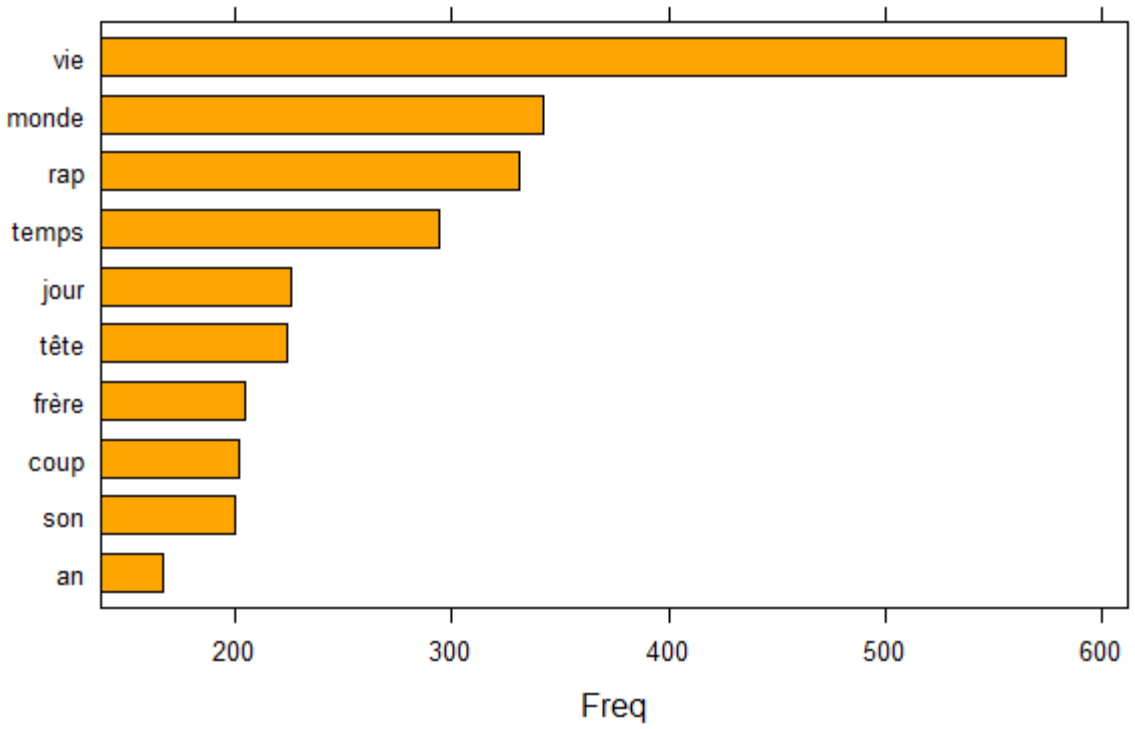

Figure 27: Most frequent nouns in the subcorpus

Similarly, the most frequent nouns in subcorpus largely correlate with the ones in the main corpus (please see Figure 27). Notably, the word 'rap' has a higher rank in the subcorpus, where it is 3rd, than in the main corpus, where it is 6th.

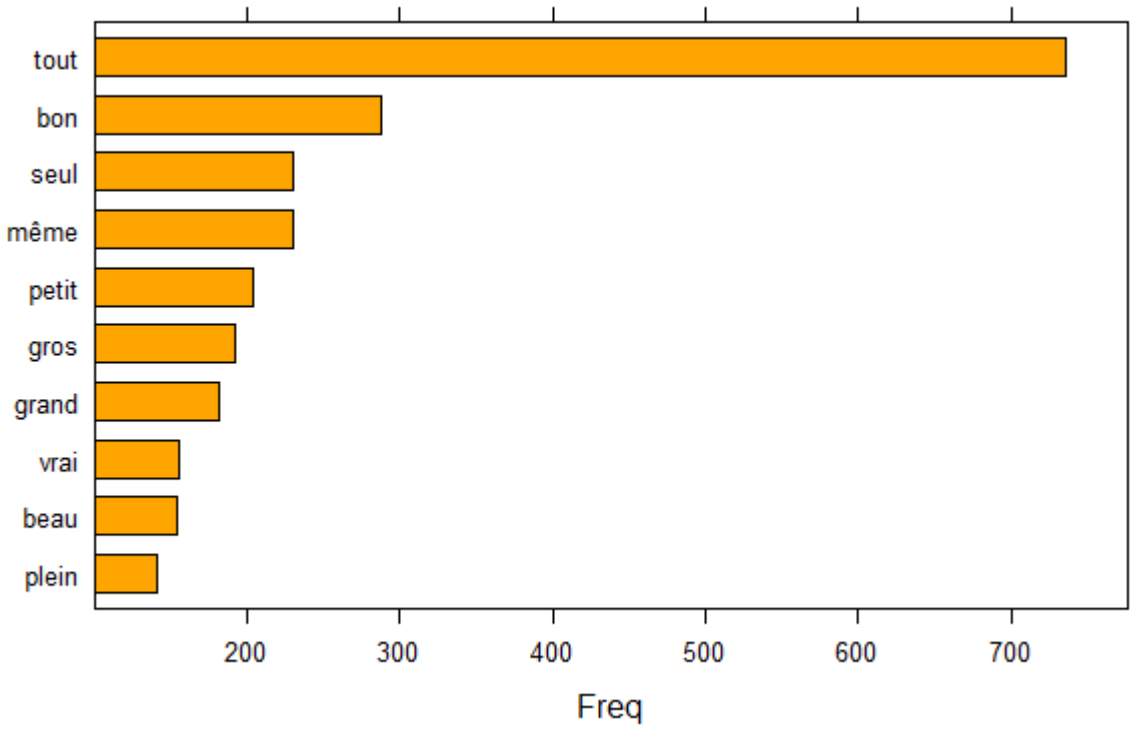

Figure 28: Most frequent adjectives in the subcorpus

Same goes for the adjectives which are mostly similarly distributed in the top ten without differences worth attention (please see Figure 28).

Bigram analysis

The subcorpus bigrams (see Appendix 8) have understandably lower frequencies than the main corpus bigrams (see Appendix 3). That puts the bigram "hip hop" to the first place with 60 occurrences in the subcorpus in comparison with the bigram 'ouais ouais' which is first in the main corpus with 285 occurrences. The bigram 'hip hop' has the rank 5 with 110 occurrences in the main corpus and the bigram 'ouais ouais' has the rank 10 with only 29 occurrences in the subcorpus. ' That suggests that in our media words subcorpus the genre has even more weight than in the main corpus and the most frequent bigrams in the subcorpus are slightly less generic than in the main corpus which suggests richer vocabulary use. Also, the focus on the media words subject puts the rank of bigrams 'bad buzz' and 'rap game' higher. 'Bad buzz' (aforementioned when

analysing frequencies in the song *Bad Buzz* by Liza Monet) has 26 occurrences in both corpora, which puts it in the rank range 140–161 in the main corpus whereas in the subcorpus it scores 15th position. 'Rap game' has 20 occurrences in both corpora and scores the rank range 142–165 in the main corpus, whereas in the subcorpus it scores as high as 24.

Songs containing media related words comprise a substantial part of the corpus. The most frequent media word is 'game' which is at the same time part of one of the important bigrams 'rap game'. The metaphor is important across the corpus as it is often used to denote the game aspect of the rap community rites. The composition of words in top ten word frequencies is similar between the corpus and the subcorpus. However, bigram counts comparison between corpus and subcorpus suggests richer vocabulary in the subcorpus.

7.4 Topic modelling

In order to choose the best analysis method, a selection of several most popular algorithms was compared, namely LSA, PLSA, and LDA. Eventually, the LDA algorithm was chosen. The main reasons for the choice were its popularity and easy accessibility through currently largely employed programming libraries. Based on that, it was considered as the most appropriate for an explorative study such as this paper. However, it should be taken into consideration that it is not completely reliable, one example of possible issues can be inability to distinguish the importance of word order in phrases. (Wallach, H., 2006) Another issue related to LDA can be the lower suitability for short documents. Moreover, improvements and new methods are still being developed in order to perfectionate current topic modelling techniques.

7.4.1 LDA

As previously mentioned, topic modelling was performed with the LDA algorithm - Latent Dirichlet Allocation. LDA is one of the most popular topic modelling algorithms. The LDA is a technique originally developed by David Blei, Andrew Ng, and Michael Jordan (Nguyen, 2014). Blei et al. (2003) describe LDA as "generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-

level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities".

As the mathematical description of this algorithm is beyond of the scope of this work, we can work with the simplified explanation provided by Robinson and Silge (2017), who claim that LDA treats each document as a mixture of topics – each document contains several topics in proportions, e.g. document contains 80% of topic A and 20% of topic B, and at the same time, each topic is a mixture of words, e.g. when the number of topics would be 2 for American news, we could label them as "politics" and "entertainment", then the words "government", "minister", and "president" would surely belong to the "politics" topic, whereas words like "cinema", "actor", "comedian" would be part of the "entertainment" topic. The topics can also overlap – the terms can be shared among the topics, e.g. a word "budget" could be identified as belonging to both of the topics.

7.4.2 Hyperparameters setting

The LDA algorithm accepts hyperparameters α, η, K (the parameters can be differently named based on the package implementation, for example η is very often referred to as β). The hyperparameter α influences the document-per-topic distribution and the hyperparameter η influences the word-per-topic distribution. The hyperparameter K defines a number of topics to work with. Overall, the hyperparameters setting significantly influences the quality of the inferred topics. To work with the hyperparameters, these two sources were used as an inspiration for the coding part of the analysis: Shashank Kapadia's *Evaluate Topic Model in Python: Latent Dirichlet Allocation* (LDA) (2020) and *Topic Modeling* with Gensim (Python) by Selva Prabhakaran (2018).

Hyperparameters α and η

The α hyperparameter influences the topic distribution over a document - "Also, in the special case where $\alpha = (\alpha, \ldots, \alpha)$, so that DirK (α) is a symmetric Dirichlet indexed by the single parameter α, when α is large, each document tends to involve many different topics; on the other hand, in the limiting case where $\alpha \to 0$, each document involves a

single topic, and this topic is randomly chosen from the set of all topics." (George and Doss, 2018). In case of the η, the hyperparameter influences the words per topic distribution – "For example, when η is large, the topics tend to be probability vectors which spread their mass evenly among many words in the vocabulary, whereas when η is small, the topics tend to put most of their mass on only a few words." (George and Doss, 2018) In both of those cases, we are referring to the symmetric distribution where the hyperparameter value stays the same for all the topics. However, asymmetric distribution of the α hyperparameter (where the hyperparameter value is different for each topic) is supposed to be more efficient in terms of quality of inferred topics, while asymmetric distribution of the η hyperparameter is not helpful (Wallach et al., 2009).

Hyperparameter K

The LDA algorithm takes a defined number of topics as a hyperparameter, thus it needs to be determined by the researcher. Several ways to achieve the most harmonic result were investigated. If the number of the topics is too high, the results can be too granular to interpret. If the number is too low, the results can be too broad. The optimal results can be achieved either manually/intuitively – by iterating over the data with different numbers of topics set and analysing the results – estimating the topic coherence by a human reader (if the topics created and the words inside are fitting together), or by taking a quantitative approach (Oleinikov, 2022).

Quantitative indicators

Perplexity

When evaluating a topic model, we often read about perplexity. "Perplexity is a statistical measure of how well a probability model predicts a sample." (Soltoff, 2021) Perplexity can be calculated for a given model and a low value is considered a good result. The best approach is to calculate perplexity for a number of models with different K values and choose the K number with the lowest perplexity. However, perplexity is just one indicator and it should not be the only decision making factor when evaluating if the topic number serves well the analysis purposes. (Soltoff, 2021) This claim is supported by research of Chang et al. (2009) where a topic score involving human evaluation was introduced and did not always correlate with the perplexity

measures. The research involved human judges which were supposed to indicate "intruder words" in a topic. The word intrusion based scores were subsequently compared to perplexity scores. (Korenčic et al., 2018)

Topic coherence measures

Since perplexity is not a reliable indicator of topic quality, the need for a different evaluation score stimulated further research on topic coherence. Topic coherence is supposed to evaluate how much the combination of words in a topic is semantically consistent. Following the research of Chang et al. (2009), most of the new approaches aimed to automate the topic coherence score calculations (Korenčic et al., 2018). There is a number of research papers suggesting different ways of calculating topic coherence. Korenčic et al. (2018) provide a comprehensive overview of the word-based methods presented in the years 2010–2017 which can be briefly reiterated for informative purposes as follows:

- Newman et al. (2010) – coherence calculation using WordNet based and Wikipedia based word similarity over the top topic words pairs together with pointwise mutual information and a method using search engine queries
- Mimno et al. (2011) – coherence calculation based on averaging pairwise log–conditional probability of top topic words
- Musat et al. (2011) – coherence calculation based on mapping top topic words to WordNet concepts
- Aletras and Stevenson (2013) – coherence calculation by averaging similarity of distributional vectors using normalised pointwise mutual information
- Nikolenko (2015) – coherence calculation by calculating tf-idf based similarity of top topic words
- O'Callaghan et al. (2015) – coherence calculation based on averaging cosine similarity of word embedding vectors over pairs of top topic words
- Rosner et al. (2013) – dividing the top topic words set into subsets and averaging a coherence measure based on conditional probabilities over the subset pairs
- Röder et al. (2015) – proposing a framework for topic coherence calculation

- Ramrakhiyani et al. (2017) – coherence measure based on clustering word embeddings

Korenčic et al. in their paper propose a document-term based coherence measure (which does not try to estimate the coherence based on word-topic distribution as above, but focuses on the document-topic distribution). They compare their work to AlSumait et al. (2009), and Ramirez et al. (2012). Arguing that some type of data (in this case, from news media) might be more suitable for document-term based topic coherence as the topic coherence might be better assessed in terms of documents associated with the topic.

The purpose of this paper is not to fully explore or understand the aforementioned coherence measure methods. However, seeing the number of research focused on evaluating topic coherence (and the above list is not exhaustive) gives us a clear idea about the importance, complexity of the coherence calculation, and the underlying growing need for a reliable topic quality indicator. It is also necessary to emphasise that the currently available tools and libraries do not necessarily offer the coherence measures calculation as built-in and readily available features and if they do, they can be computationally heavy.

Choosing the most efficient hyperparameters

Fortunately, the Gensim library provides a perplexity calculation feature together with some support for coherence measures calculation. It also allows for manual, symmetric, and asymmetric setting of the alpha parameter and eta hyperparameter. (For more details, please view Gensim documentation.)

In order to evaluate, which hyperparameters are the best for our corpus, we ran an analysis inspired by Shashank Kapadia's Jupyter notebook (2019) with the addition of the perplexity calculation. The analysis is based on a loop which creates a Gensim LDA model with two corpora – on the original corpus and on the original corpus reduced to 75% of the size. Each iteration creates a model with different hyperparameters and evaluates its perplexity and topic coherence. Gensim provides these options for topic coherence measures:

- *u_mass*
- *c_v*
- *c_uci*
- *c_npmi*

Those were described by Röder et al. (2015) (mentioned in the overview above) in their paper *Exploring the Space of Topic Coherence Measures* referencing the previously conducted research and coming up with a unifying framework. According to the results of their research, *c_v* coherence measure correlates the most with human scoring. Based on that, it was also used for the hyperparameters evaluation for this paper.

The whole analysis was counting also with asymmetric and symmetric values as a possibility for alpha and eta. The operation was very computationally expensive as it lasted 28 hours on a Mac Pro laptop with a M1 chip. For the below provided results, only the values for the full corpus were included.

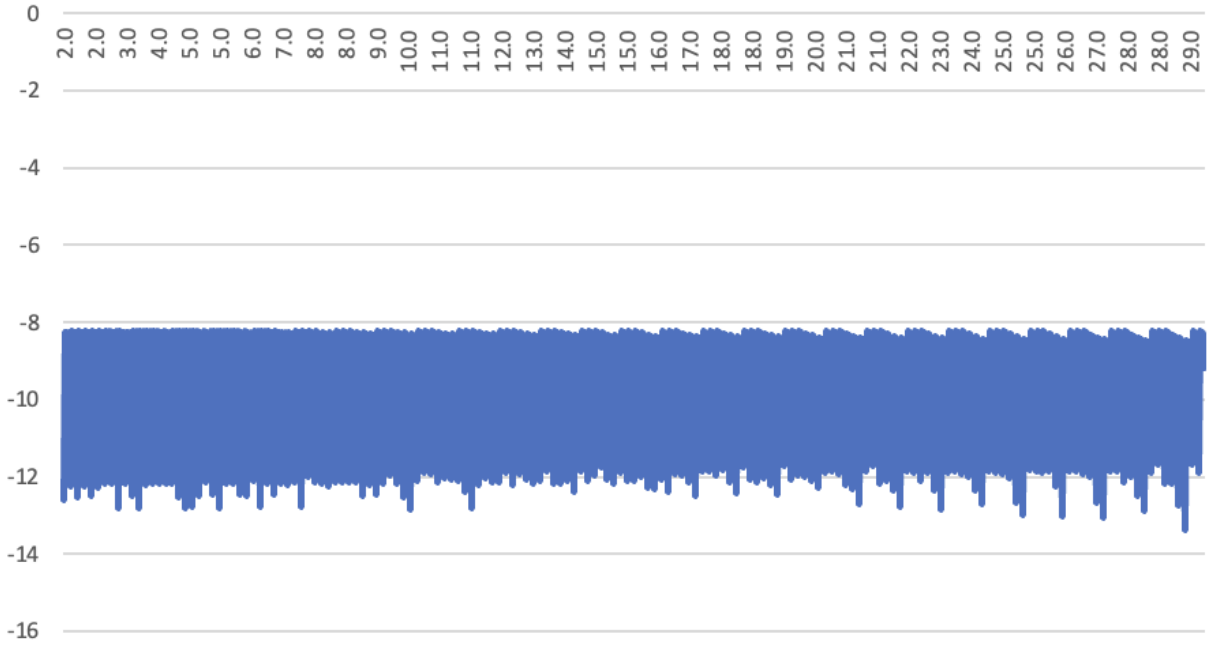

Figure 29: Perplexity levels per number of topics for the main corpus

Looking at the graph of perplexity per K (please see Figure 29), number of topics, the lowest perplexity is at 29 topics with certain combinations of a relatively high alpha and a very low eta. That would overall suggest more topics per document and less words in a topic. However, the coherence score in this case is quite low. Since the tendency of the perplexity value is looking mostly descending with the number of topics, there is a possibility that we could obtain even a lower score with a higher number of topics if given specific Alpha and Eta values.

Number of topics	Alpha	Eta	Coherence	Perplexity
29	0.91	0.01	0.26897764	-13.359099

Figure 30: Values for 29 topics – Alpha, Eta, Coherence, and Perplexity

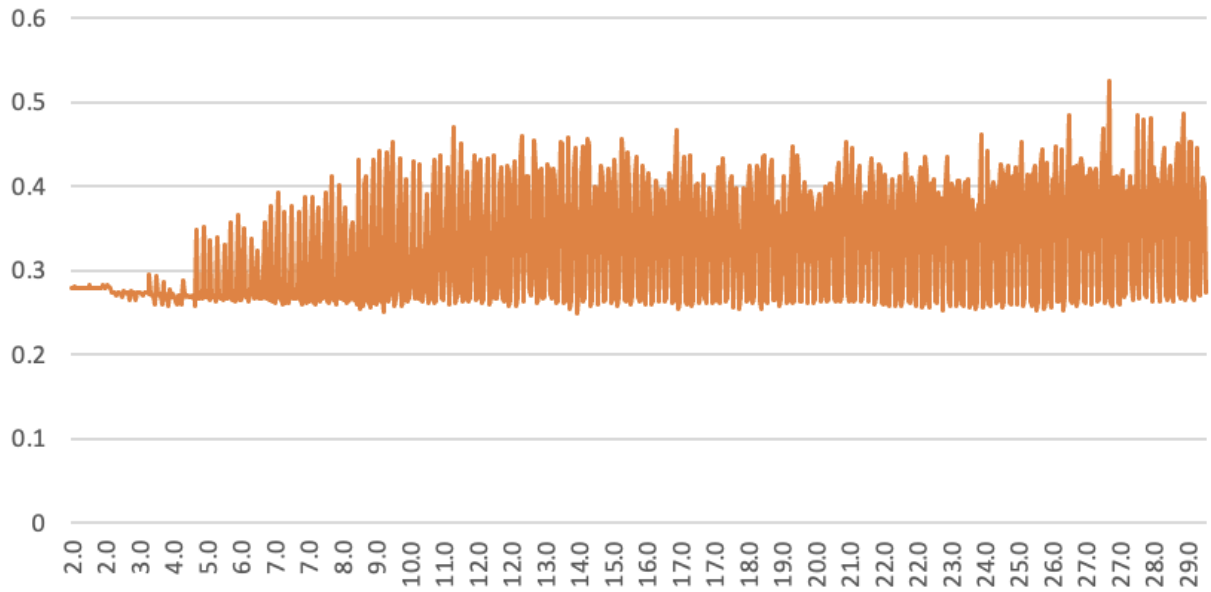

Figure 31: Coherence score per number of topics for the main corpus

When studying the coherence score per K graph (please see Figure 31), we can see that the highest coherence score is achieved at number 27 with a high alpha and eta suggesting higher number of topics per document and more words in a topic. However, perplexity is very high in comparison with the lowest perplexity. That again confirms that coherence and perplexity do not have to correlate.

Number of topics	Alpha	Eta	Coherence	Perplexity
27	0.91	0.91	0.52488081	–8.5316083

Figure 32: Values for 27 topics – Alpha, Eta, Coherence, and Perplexity

7.4.3 Corpus topic modelling

In order to gain more insight into how the hyperparameters choice influences the final topic creation, 3 variations of the LDA model were created. The first model was based on the lowest perplexity value, the second on the highest coherence value, and the third model on searching for a compromise between highest possible coherence and lowest possible perplexity. To provide the most complete analysis, the most prominent 5 topics of each model were examined when creating the model for the full corpus.

Model	Number of topics	Alpha	Eta	Coherence	Perplexity
Lowest perplexity	29	0.91	0.01	0.26897764	–13.359099
Highest Coherence	27	0.91	0.91	0.52488081	–8.5316083
Compromise between lowest perplexity and highest coherence	28	0.31	symmetric	0.290924108	-9.279195784

Figure 33: Overview of values for chosen models – Number of topics, Alpha, Eta, Coherence, and Perplexity

The library used for visualisation of the topic modelling results was LDAvis and saliency (as described in the section Data analysis tools) was set to 1 to for all the visualisations in order to provide the highest relevance of the words for the topics. In addition to the terms frequencies visualisation per topic (visible on the right side of Figure 34), it provides an intertopic distance map represented as a two-dimensional graph using Jensen-Shannon divergence (in simple words, the topics are close or distant based on how many words they share). The library also allows visualising relevance of each of the topics based on a specific word (visible on the left side of Figure 34) (Sievert, 2014).

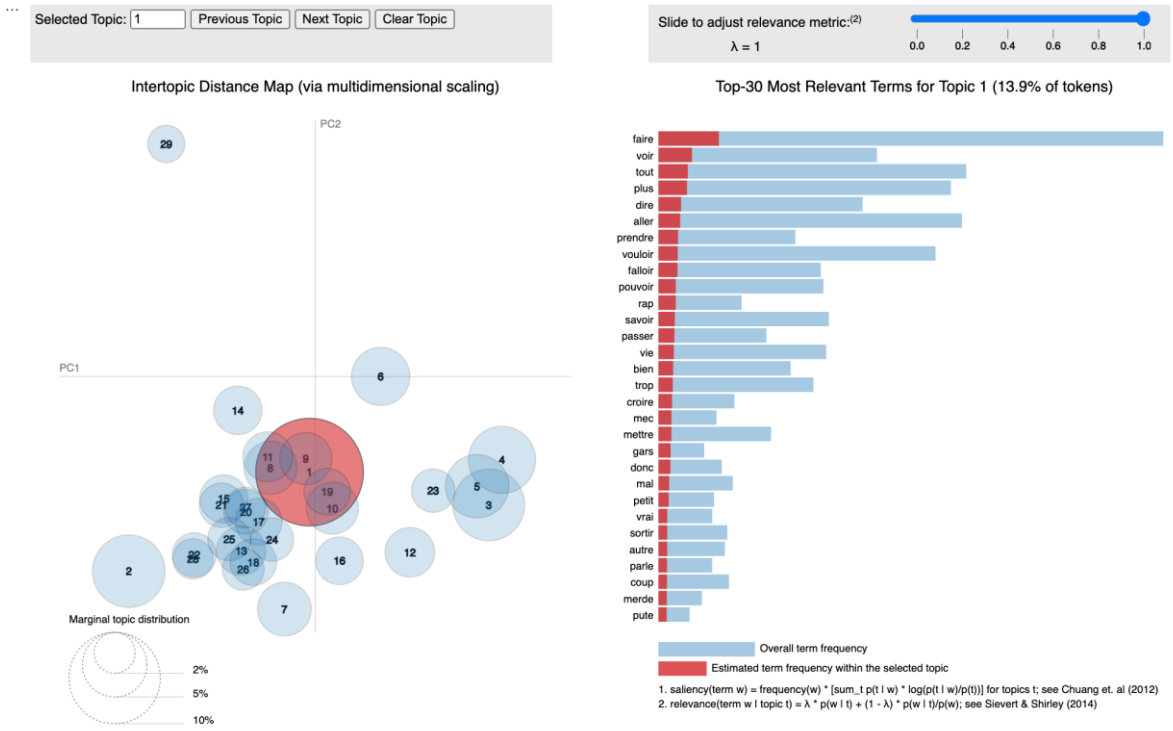

Figure 34: The lowest perplexity model

The model based on lowest perplexity shows mostly a proximity between the topics as they are closely concentrated in one area apart from the topic 29 which is placed far from the remaining group (see Figure 34).

Topic 1 - General	Topic 2 – Street life	Topic 3 – Love and relationships	Topic 4 – Death and depression	Topic 5 – Life purpose
faire	faire	plus	vie	jour
voir	ouai	vie	voir	tout
tout	vouloir	amour	tout	pouvoir
plus	aller	cœur	plus	savoir
dire	tout	dire	fort	dire
aller	gros	faire	pouvoir	faire
prendre	plus	vouloir	même	vie

vouloir	voir	nuit	sang	voir
falloir	falloir	aimer	autre	plus
pouvoir	mettre	temps	dire	vouloir
rap	fume	mal	combien	fois
savoir	passer	pouvoir	mourir	prendre
passer	ville	apprendre	larme	mieux
vie	sale	encore	monde	demain
bien	pute	savoir	faire	choisir
trop	savoir	peur	homme	non
croire	prendre	voir	passer	vivre
mec	dire	aime	jamais	fond
mettre	trop	tout	reste	temps
gars	ouer	alors	jour	seul
donc	gang	seul	seul	monde
mal	donne	moins	trop	dieu
petit	sortir	jamais	ici	chemin
vrai	fou	ciel	devenir	autre
sortir	couille	beaucoup	devoir	mal
autre	monnaie	trop	temps	falloir
parle	zone	bien	mal	donne
coup	connaître	petit	haine	souvent
merde	fait	jour	paix	envie
pute	vie	peu	peu	bien

Figure 35: The lowest perplexity model – top five topics

The 5 most prominent topics (please see Figure 35) could be labelled as: Topic 1 – Rapper's life in general, Topic 2 – Street life, Topic 3 – Love and relationships, Topic 4 – Death and depression, Topic 5 – Life purpose. Overall, the topics in this model feel quite general with larger scale of words which are not always directly indicating a clear topic. However, while focusing on the most specific words, we can distinguish topics more easily.

The first topic, Rapper's life in general, is the most polysemous and we could say it somehow encompasses life in general with mundane words like "faire" – "to do", "autre" – "other", "pouvoir" – "to be able", "sortir" – "go out". Nevertheless, it is still quite clear that the topic is rap specific as it contains words like "rap" and "pute" – "whore".

Second topic, Street life, also contains a large proportion of general words, but comprises keywords such as "monnaie" – "coin", "gang", "fume" – "smoke", "fou" – "crazy", "ville" – "city" which suggest a focus on life on the street, group identity, and money.

Third topic, Love and relationships, follows the others with a proportion of generic words, but is more clearly distinguished than the previous two. Terms like "amour" – "love", "cœur" – "heart", "nuit" – "night", "peur" – "fear", "seul" – "alone", "aime" – "love" (1st person singular conjugation, verb "aimer"), "ciel" – "sky/heaven" distinctly suggest that the topic focus is mainly on love, sentimental life and relationships.

Fourth topic, Death and depression, is not an exception in terms of number of generic words, but offers understandable collection of keywords such as "vie"– "life", "sang" – "blood", "mourir" – "to die", "larme" – "teardrop", "haine" – "hatred", and "paix" – "peace". The existential sentiment of the topic is clear in a similar way to the previous topic's focus on love life.

Fifth topic, Life purpose, confirms the previously stated observation that the topics in this model feel slightly generic with a large proportion of common words, and as the low coherence score suggests, they do not feel very coherent as units. In spite of that, even the fifth topic offers words like "choisir" – "to choose", "chemin" – "the way", "dieu" – "the God", "vivre" – "to live", "demain" – "tomorrow". The underlying meaning could be reflecting on life purpose and how to continue living one's life.

The terms "seul" – "alone", and "vie" – "life" appeared several times across the topics. This higher frequency could point at the feeling of adversity and having to deal with one's life alone as those words are less generic than for example very frequent "faire" – "to do" which is present in all the five topics.

The highest coherence score model

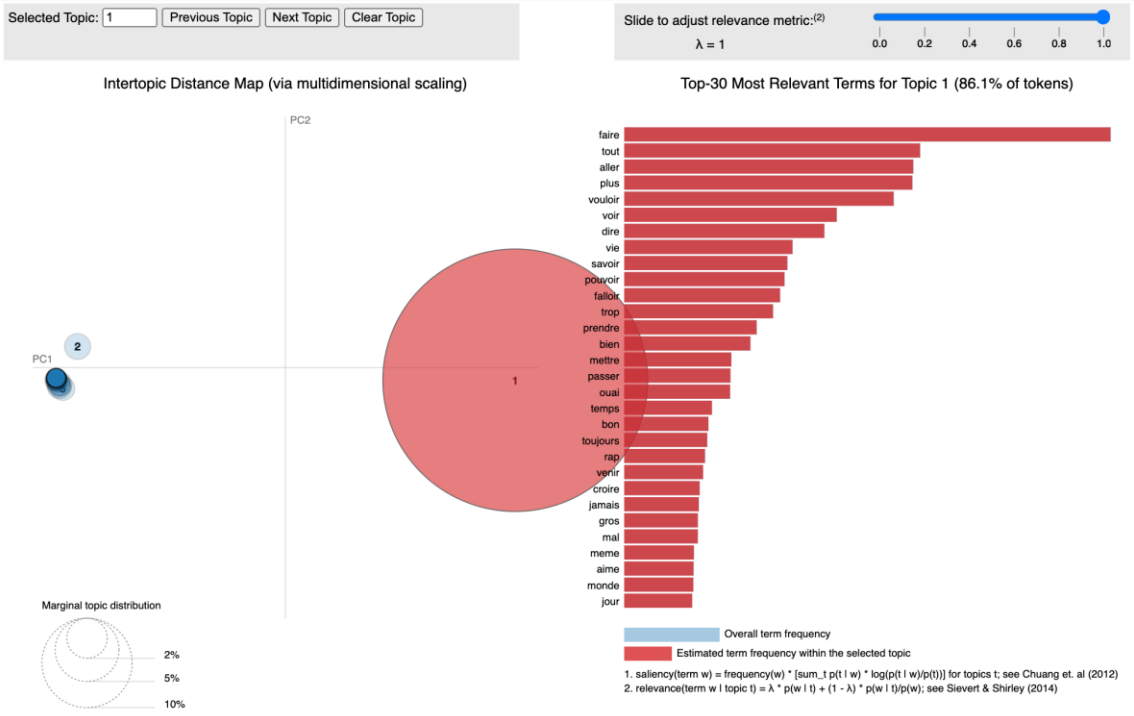

Figure 36: The highest coherence score model

What immediately strikes us when analysing the highest coherence score model visualisation (see Figure 36), is the fact that practically all the most frequent words (and thus the most generic ones) are organised into one major topic. This allows the other less frequent words to be divided more coherently, with higher focus, and the model provides very specific results with overall less common (less frequent) words in the topics. The major topic is also very distant from the rest of the topics.

Topic 1 – General	Topic 2 – Crime and violence	Topic 3 – Anarchy and against the system	Topic 4 – State censorship	Topic 5 – Gang, African heritage, and love life
faire	allume	peuple	censure	moula
tout	alphabet	anarchie	geant_geant	gang_gang
aller	laser	deterr	aucune_censur	santiag
plus	brandir	bon_grigri	censur	africain
vouloir	hisser	hache	tais	touss
voir	savoir	engager	cercle_rouge	darwah
dire	flingu	chiraquie	etat	botte
vie	ruelle	tandem	lutece	buvance_apres
savoir	alien	detient_arsenal	cr	beaufs_coups
pouvoir	milli	armes_egalesa	zique	grr_grr
falloir	lav	glisse	censurer	skur
trop	domestique	cœur_vailler	baillon	benefice_benefice
prendre	briquet	poing	furet	delire
bien	deal	debrouille	susurre	boom_boom
mettre	rap	combat	siete_cinco	vire
passer	abuser	coma_aquo	aucune_censure	bisous_tendre
ouai	lettre	chiraqui	white_spirit	wet
temps	glisser	gribouille	baudi	gang
bon	methode	memes_magouille	irrite	escrocs_prennent
toujours	charcler	memes_patrouille	section	bamboula_bamboula

rap	top	vadrouillent_vandale	resorbe	maine_weed
venir	crim	vecu_accusateur	delter	bis_bis
croire	chargeur	ca	stoppe	cheveux_crepus
jamais	plomb	balltrap	absorber	gominer
gros	propre	guerre	chatouille	santiag_niquer
mal	doux	vendeur	maure	tumeur_tumeur
meme	compter	lalala_lalala	ganja_deliquance	clic_clac
aime	police	bledi	juvenile_il	chatouilles_partout
monde	foule	repic	transistor	inquieter_maman
jour	claque	faire	quage_brer	bise

Figure 37: The highest coherence score model – top five topics

This major topic is comparable with the first topic of the lowest perplexity model, Rapper's life in general, and thus was named in the same way.

As the intertopic distance suggests, the second topic could not be more different from the first one. The vocabulary used is much more specific and less common in general speech. Even though the vocabulary involved largely differs from the second topic of the lowest perplexity model, it aligns on the underlying meaning. The second topic in this model is named Crime and violence, because it contains terms such as "brandir" – "brandish", "flingu"– misspelt "flingue" – a slang word for a gun, "deal" – a borrowed word from English, "abuser" – "to abuse", "charcler" – "A verb with a Catalan origin which means to fight with violence. It's a term usually used in the language spoken on the street." (Linternaute.fr, 2021), "crim" – misspelt "crime", "chargeur" – could be ambiguous, but can mean a gun magazine, "plomb" – can also be ambiguous, but can

mean a bullet, and finally "police". This selection of words clearly indicates a topic full of life on the street actions linked with crime and violence.

Topic 3 was named Anarchy and against the system as it contains many references to the fight against the system and rebellious tendencies. The most prominent terms indicating such theme are: 'anarchie' – 'anarchy', 'chiraquie' – a colloquial term for the period of Jacques Chirac presidency, 'detient_arsenal' – 'keeps_arsenal', 'combat', 'memes_magouille' – 'same_shady_business', 'vadrouillent_vandale' – 'wandering_vandal', 'guerre' – 'war'. Even though the topic also contains words which are not coherent with the rest, those words suggest strong emotions and identification with being in a fight.

Topic 4 was called State censorship simply because it contains the word 'censure' – 'censorship' several times. The sentiment of speaking freely (and not being able to) seems to be clearly present in this topic, even though it also contains terms which are not directly linked. To support that statement, the list of the related terms is: 'censure' – 'censorship', misspelt 'aucun_censur' – 'no_censorship' and 'censur', 'censurer' – 'to censor', 'tais' – 'being silent', misspelt 'etat' (correctly 'état') – 'state', misspelt 'baillon' (correctly 'bâillon') – 'gag'.

Topic 5 seems to contain several smaller subtopics than being one coherent whole. Thus, it was named Gang, African heritage, and love life.

- The terms for subtopic Gang:
    - 'gang', 'gang_gang', 'maine_weed'.
- The terms for subtopic African heritage:
    - 'africain' – 'african', 'cheveux_crepus' (misspelled cheveux crépus – 'frizzy hair'
- The terms for Love life:
    - 'bisous_tendre' – 'soft_kisses', 'bise' – colloquial term for a kiss

Overall, this topic model contains many references to activities outside of ordinary life. Topic 2, Topic 3, Topic 4, and Topic 5 are semantically close as they could be arranged into a superordinate topic Anti-system activities (please inspect Figure 37 above for words highlighted in light red as they are all members of the aforementioned

superordinate topic). Despite the high coherence score, there are still many word occurrences which break the topic coherence of all the 5 topics. The most obvious examples would be: 'doux' – 'soft' in Topic 2, 'faire' – 'to do' in Topic 3, 'chatouille' – 'tickle' or 'lalala_lalala' in Topic 4, or 'tumeur_tumeur' – 'tumour' in Topic 5. We can however observe the distinct division between Topic 1 and the rest of the topics which is straight away remarkable through the intertopic distance visualisation. The division could be named as Ordinary life vs. Life in a marginalised community and describes very well the topic content of the corpus.

Compromise between highest possible coherence and lowest possible perplexity

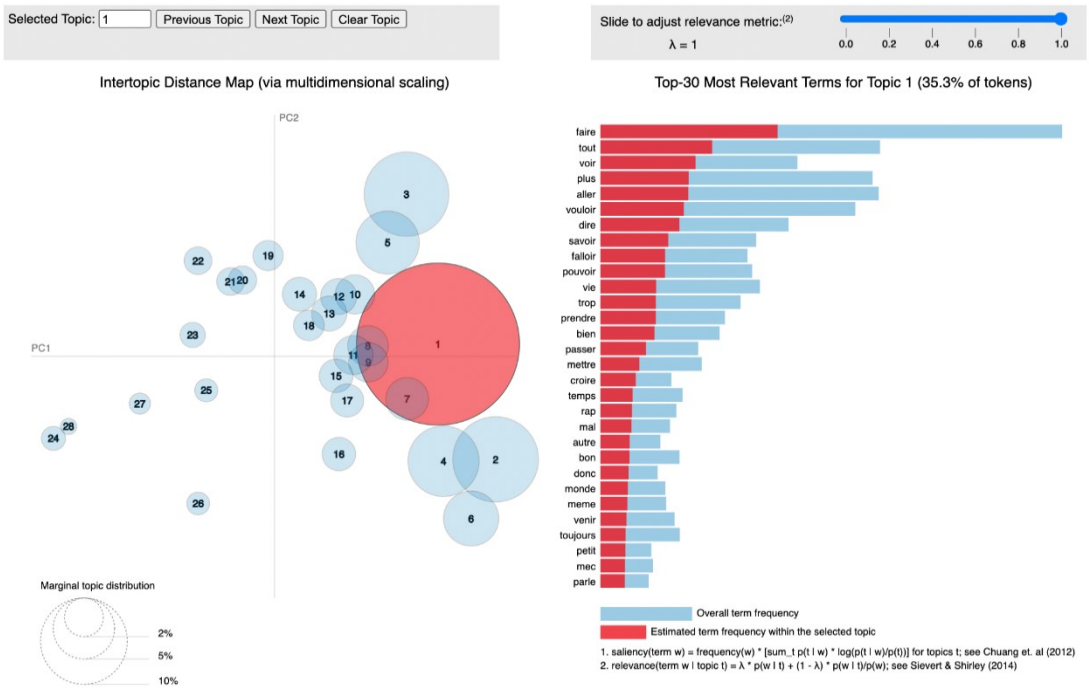

Figure 38: The compromise model

In order to have a comparison reference point to the coherence and perplexity-based model, one more model was chosen to be analysed (see Figure 38). The hyperparameters values were based on the intersection of lowest possible perplexity and highest possible coherence. Predictably, the results are the least distinctive and the topics seem very similar to one another. They contain a larger number of common words and a lower number of words clearly specific to a certain subject. However, there are some words which can suggest a certain theme even though less strongly.

Topic 1 – General	Topic 2 – Love and emotions	Topic 3 – General / gang	Topic 4 – Fight	Topic 5 – General / rap
faire	plus	faire	faire	faire
tout	vie	ouai	tout	vouloir
voir	vouloir	aller	plus	aller
plus	dire	vouloir	guerre	tout
aller	faire	tout	pouvoir	coup
vouloir	aime	gros	monde	non
dire	amour	falloir	arme	rap
savoir	cœur	voir	sang	hein
falloir	savoir	mettre	ici	putain
pouvoir	tout	plus	terre	guele
vie	jour	prendre	jeune	gros
trop	pouvoir	savoir	droit	plus
prendre	temps	trop	prendre	rime
bien	voir	fume	face	trop
passer	nuit	dire	histoire	dire
mettre	trop	sortir	mort	mettre
croire	aimer	passer	noir	voir

temps	jamais	sale	vouloir	pouvoir
rap	mal	ville	voir	merde
mal	encore	pute	vie	savoir
autre	toujours	donne	mourir	couille
bon	seul	chaud	combat	meuf
donc	bien	venir	homme	venir
monde	perdre	connaître	venir	pote
meme	falloir	vie	coup	mec
venir	juste	toujours	pays	toujours
toujours	peur	fait	bas	prendre
petit	croire	gang	enfant	beuh
mec	peu	couille	seul	fou
parle	non	niqu	trop	ici

Figure 39: The compromise model – top five topics

This vagueness makes it more difficult to name the topics clearly (please see Figure 39 to inspect the topics). Despite the general sentiment being possible to be identified as similar to the previous two analyses, the lower number of distinctive words caused naming most of the topics as general or general with a suggestion (e.g. General / rap) based on only few words which could indicate a specific subject (e.g. 'rap' or 'rime' – 'rhyme'). When revisiting the lowest perplexity model, these two models seem to be very similar and thus suggesting that perplexity might not have much significance when evaluating topic model quality. The most easily identifiable topics in this analysis would be Topic 2 – Love and emotions and Topic 4 – Fight.

The most representative words for Topic 2 – Love and emotions are:

- 'aime' – a conjugation of 'aimer' – 'to love'
- 'amour' – 'love'

- ‘cœur’ – ‘heart’
- ‘peur’ – ‘fear’

The most representative words for Topic 4 – Fight:

- ‘guerre’ – ‘war’
- ‘arme’ – ‘weapon’
- ‘sang’ – ‘blood’
- ‘mort’ – ‘death’
- ‘mourir’ – ‘to die’
- ‘combat’ – ‘fight’

Topics frequent words analysis

To understand better the relationship between the overall term frequency and the frequency in the given topic, it is possible to investigate the blue and red bars in the visualisation (e.g. Figure 38). For the full visualisation please visit https://julieklimentova.github.io/le_rap_francophone.
In order to investigate further how the overall frequency relates to the three models, the top ten frequent verbs, adjectives, and nouns in the corpus were counted in the top five topics of each of the models.

The most frequent verbs	
avoir	0
faire	5
aller	2
dire	5
vouloir	4
pouvoir	4
savoir	5

voir	5
aimer	1
être	0
	31
The most frequent nouns	
vie	4
monde	2
temps	3
jour	2
tête	0
rap	1
frère	0
fois	1
son	0
amour	1
	14
The most frequent adjectives	
tout	5
bon	0
seul	3
gros	1
même	1
beau	0

petit	2
grand	0
vrai	1
	13
Total count	58

Figure 40: The lowest perplexity model – an overview of frequent words

The most frequent verbs	
avoir	0
faire	1
aller	1
dire	1
vouloir	1
pouvoir	1
savoir	2
voir	1
aimer	0
être	0
	8
The most frequent nouns	
vie	1
monde	1

temps	1
jour	0
tête	0
rap	2
frère	0
fois	0
son	0
amour	0
	5
The most frequent adjectives	
tout	1
bon	1
seul	0
gros	1
même	1
beau	0
petit	0
grand	0
vrai	0
	3
Total count	17

Figure 41: The highest coherence model – an overview of frequent words

The most frequent verbs	
avoir	0
faire	5
aller	3
dire	3
vouloir	5
pouvoir	4
savoir	4
voir	5
aimer	1
être	0
	30
The most frequent nouns	
vie	4
monde	2
temps	2
jour	1
tête	0
rap	2
frère	0
fois	0
son	0

amour	1
	12
The most frequent adjectives	
tout	5
bon	1
seul	2
gros	2
même	1
beau	0
petit	1
grand	0
vrai	0
	11
Total count	54

Figure 42: The compromise model – an overview of frequent words

The top five topics of the highest coherence model contain the least amount of the most frequent words in the corpus (see Figure 41). On the contrary, the lowest perplexity model has the highest number of the most frequent words in the corpus (see Figure 40), followed closely by the compromise model (see Figure 42).

Conclusion

Analysing the results of the main corpus topic modelling, we can conclude that the highest coherence model has the most significant value for this research paper. Based

on the comparison of numbers of the most common words, it contains more specific (and less common) words and thus defines the topics the most clearly, establishing a visible division between the first topic containing more general words and the rest. It also contains the most distinctly divided top 5 topics even though some of the topics could still be split into several subtopics (e.g. Topic 5 – Gang, African heritage, and love life) as they do not necessarily feel completely coherent to a human reader.
However, the second two models, the lowest perplexity and compromise models, confirm a similar sentiment despite containing too many general words to form the topics more distinctly.
At the same time, they provide a valuable reference point to the highest coherence model as they offer a different perspective on the same corpus.

7.4.4 Subcorpus topic modelling

As for the word frequencies analysis, the subcorpus with lower ambiguity was chosen to provide more focused results. In order to provide enough data for comparative analysis, the same process as for the main corpus for evaluating ideal hyperparameters setting was repeated.

Finding the lowest perplexity

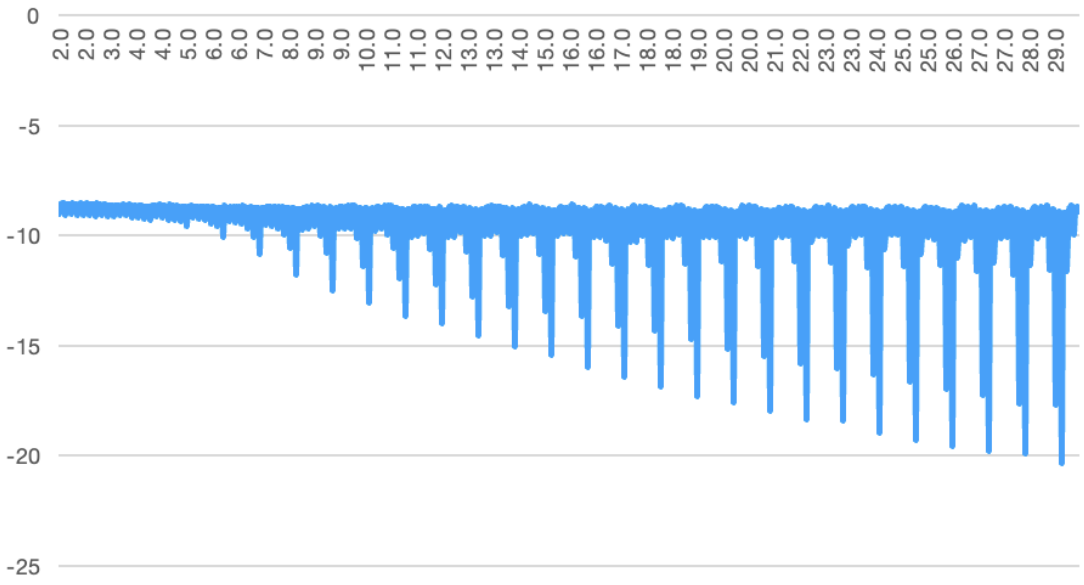

Figure 43: Perplexity levels per number of topics for the subcorpus

Number of topics	Alpha	Eta	Coherence	Perplexity
29	0.91	0.01	0.32522	–20.342

Figure 44: Values for 29 topics – Alpha, Eta, Coherence, and Perplexity

In comparison to the main corpus lowest perplexity value, –13.359099, the lowest possible value for the subcorpus is substantially lower. Interestingly, the K, Alpha, and Eta values are the same as for the main corpus (see Figure 44 for the lowest perplexity model subcorpus values and Figure 30 for the main corpus values). Otherwise, the overall trend is the same as for the main corpus, but it is more clearly distinguished. The K variable and the Perplexity variable are inversely proportional given certain Alpa and Eta values. That could suggest that the size of the subcorpus is more suited for this number of topics than the main corpus.

Finding the highest coherence

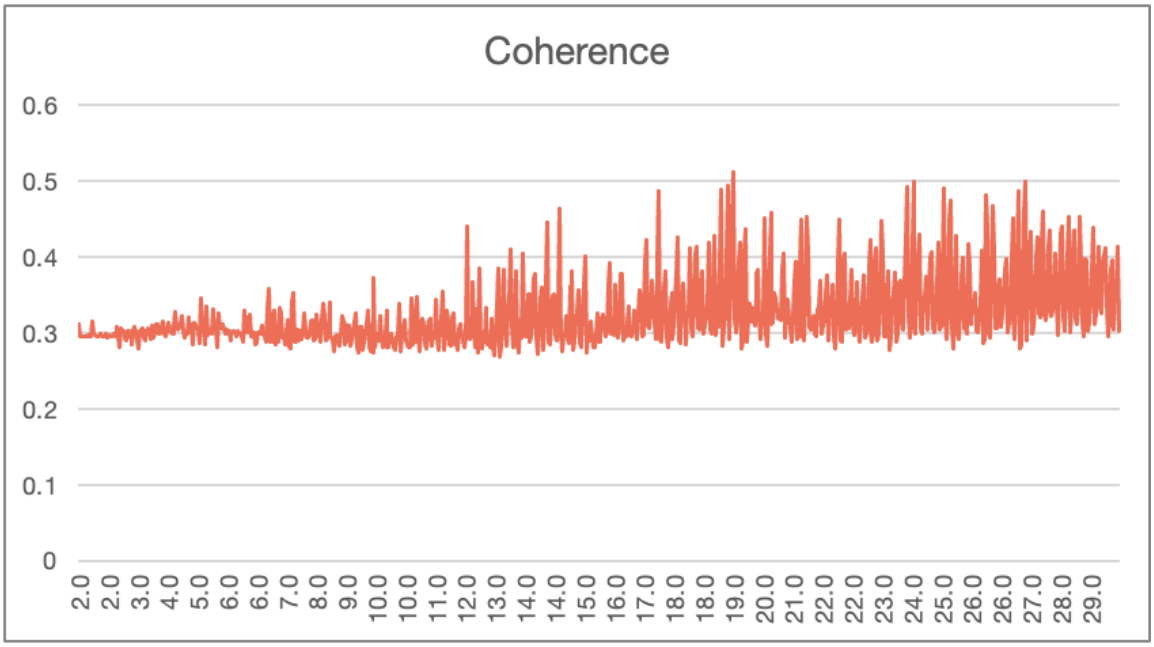

Figure 45: Coherence levels per number of topics for the subcorpus

Number of topics	Alpha	Eta	Coherence	Perplexity
19	0.91	0.91	0.51121	–8.8726

Figure 46: Values for 19 topics – Alpha, Eta, Coherence, and Perplexity

The subcorpus highest coherence model also shares Alpha and Eta values with the main corpus highest coherence topic model (see Figure 46 for the subcorpus highest coherence model values and Figure 32 for main corpus values). However, the K value is much lower than in the case of the main corpus which correlates with the subcorpus' smaller size. Also, the coherence score is slightly lower than the highest coherence, 0.52488081, for the main corpus.

Number of topics	Alpha	Eta	Coherence	Perplexity
16	0.61	0.01	0.3114	–13.666

Figure 47: Values for 16 topics – Alpha, Eta, Coherence, and Perplexity

As for the main corpus, a reference model was chosen for comparison with the two others. The approach was the same – finding a compromise between lowest possible perplexity and a highest possible coherence (see Figure 47 for precise values).

Topic modelling

	Number of topics	Alpha	Eta	Coherence	Perplexity
Lowest perplexity	29	0.91	0.01	0.32522	–20.342
Highest coherence	19	0.91	0.91	0.51121	–8.8726
Compromise between lowest perplexity and highest Coherence	16	0.61	0.01	0.3114	–13.666

Figure 48: Overview of value for the subcorpus models – Alpha, Eta, Coherence, and Perplexity

For the subcorpus topic modelling, we will use the values from the three chosen models (see Figure 48 for overview of values). Further on, the top five topics will be analysed for each model and we will summarise the results and compare them with the main corpus modelling.

The lowest perplexity model

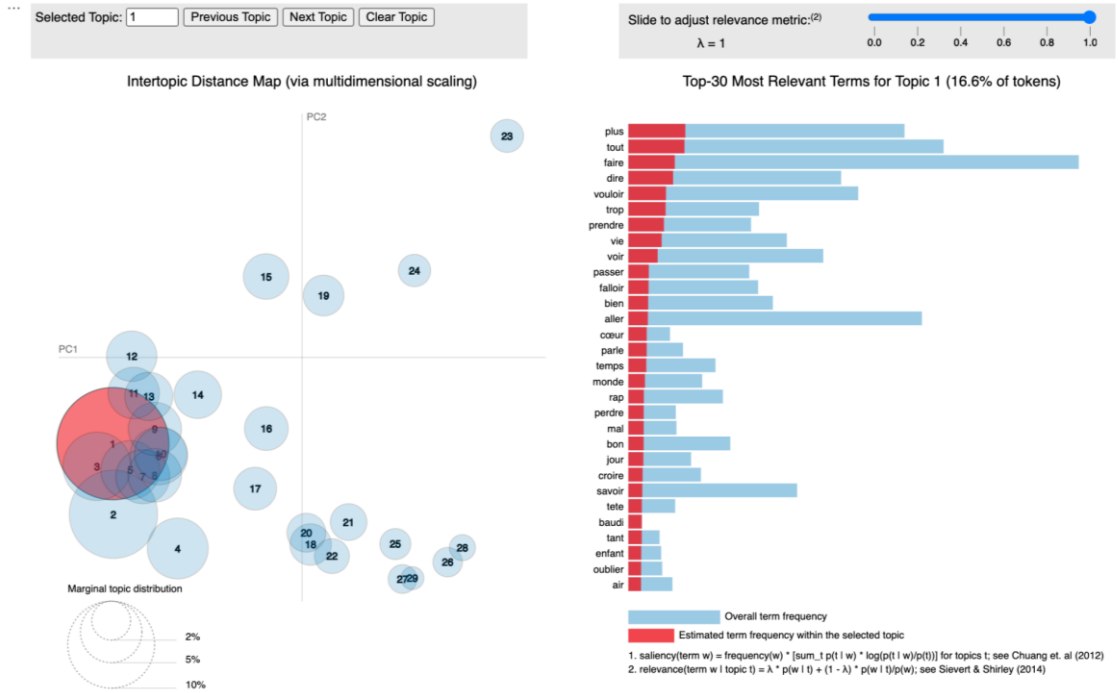

Figure 49: The lowest perplexity model – subcorpus

Topic 1 – General	Topic 2 – General 2 / Colloquial speech	Topic 3 – General 3	Topic 4 – Suburban identity	Topic 5 – General 4
plus	faire	tout	faire	aller
tout	tout	jour	ghetto	toujours
faire	aime	aller	vie	petit
dire	plus	plus	tout	faire
vouloir	aller	faire	savoir	voir
trop	vouloir	soir	aller	tout
prendre	mec	laisse	plus	trouver
vie	dire	passer	sale	vie
voir	pute	vrai	falloir	oublier
passer	rap	croire	vouloir	jamais

falloir	trop	meme	bien	encore
bien	pouvoir	an	temps	cote
aller	voir	bien	voir	comment
coeur	parler	putain	venir	pouvoir
parle	monde	savoir	mieux	dire
temps	fond	dire	bon	injoignable
monde	savoir	joue	banlieue	vibe
rap	bande	vie	vien	moins
perdre	vie	reste	aime	chemin
mal	niqu	bon	mec	venir
bon	reproche	dur	dire	temps
jour	regarde	toujours	donc	battre
croire	toujours	sortir	putain	meme
savoir	temps	yo	devoir	savoir
tete	mettre	ouai	parle	dessou
baudi	fils	attendre	jeune	heure
tant	cher	falloir	peu	mentalite
enfant	bien	seul	souci	prendre
oublier	meuf	sale	taire	vouloir
air	baise	pouvoir	zesau	trouve

Figure 50: The lowest perplexity model – subcorpus – top five topics

The subcorpus lowest perplexity model in the comparison with the main corpus Lowest Perplexity model leaves more space between the topics (see Figure 49). However, similarly, the top five topics are closely grouped together. Even though the topics are feeling quite coherent to the human reader, the number of common words is too high to determine specific sentiments for each of the topics (to inspect the topics, please see

Figure 50). However, there is still a large amount of rap specific words distinctive to the corpus. The most easily distinguishable topic is Topic 4 – Suburban identity. This is mainly due to the presence of two words – 'ghetto' and 'banlieue' and the name of Zesau, the French rapper from Malassis, Vitry-sur-Seine. It also contains the word 'souci' – 'worry' which could be associated with harder life conditions and 'taire' – 'shut up' which could suggest the impossibility of expressing oneself freely. The other more easily distinguishable topic is Topic 2 – Colloquial speech. As the name suggests, this topic contains words which are often used in colloquial context. The most representative examples would be 'mec' – 'guy', 'meuf' – a *verlan* word for 'a woman', followed by 'pute' – 'whore' or 'baise' from 'baiser' – 'to fuck'.

The highest coherence

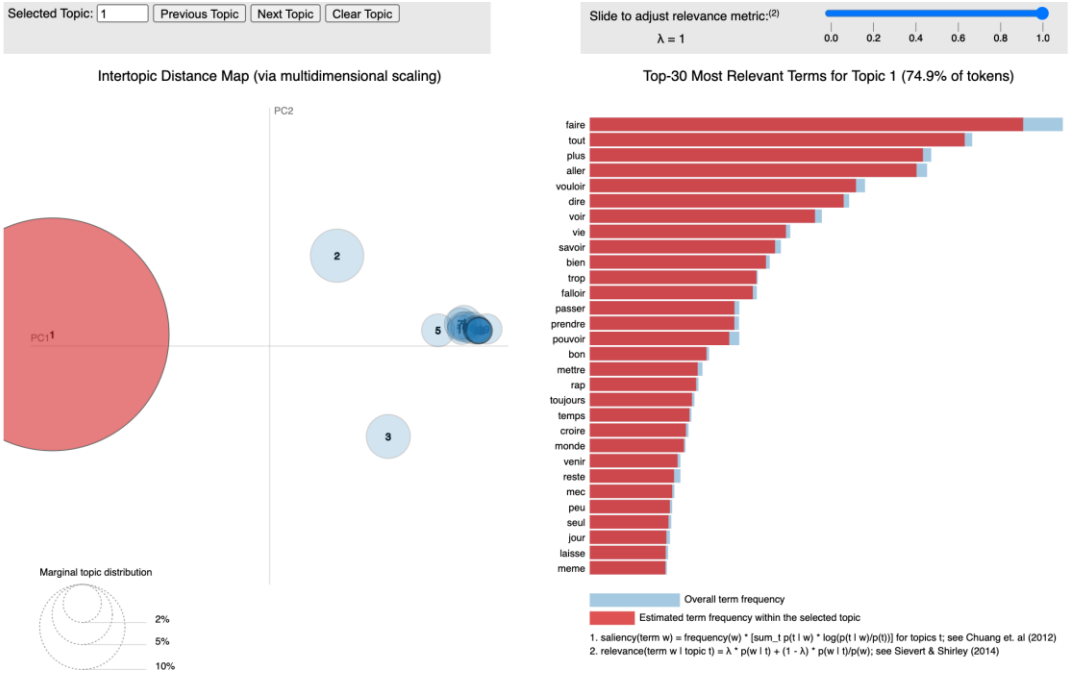

Figure 51: The highest coherence model – subcorpus

Topic 1 – General	Topic 2 – General / War	Topic 3 – Food and drink / Unidentified	Topic 4 – Technology and power	Topic 5 – Propaganda
faire	faire	faire	laser	non
tout	pouvoir	pastis	arme	vouloir

plus	aller	oeuf	faisceau	ment
aller	aime	ghetto	defense	oui
vouloir	vouloir	petit	armes_laser	aller
dire	reste	gout	air	toujours
voir	plus	ome_om	energie	voir
vie	foutre	plus	pouvoir	parler
savoir	tout	savoir	dire	tenir
bien	avancer	aller	optique	chanter
trop	voir	blanc	utilise	regarder
falloir	seule_chose	tout	americain	dire
passer	petit	bien	tres	monsieur
prendre	tourner	alors	developpe	mentir
pouvoir	passer	battez_battez	programme	bien
bon	kader	omelette	similaire	aussi
mettre	prohibition	amaiam	energi	payer
rap	prendre	casser	chimique	parle
toujours	coup	fai	jeu	voiler
temps	reproche	voir	dollar	falloir
croire	force	paradis	puissance	plus
mode	savoir	manger	etat	maintenant
venir	ton	couille	peut_etre	continuer
reste	vie	attendre	sortir	emission
mec	mettre	falloir	portable	entreprise
peu	dire	eau	nouveau	contredire
seul	bombe	guerre	technologie	poli
jour	guerre	vouloir	cible	boniment

laisse	jour	changer	grand	repondre
meme	age	cauchemar	fibr	rate

Figure 52: The highest coherence model – subcorpus – top five topics

The highest coherence model is similar to the main corpus highest coherence model in the way that it contains one large topic containing more common words opposed to the rest of the topics (see Figure 51 for details). However, the main difference is that the topics 2 and 3 are closer to topic 1 than in the main corpus. The topic 5 is also closer but on a smaller scale. This difference in intertopic distance is visible also in the topics' word composition as they share words together. On the level of topic distinction, that makes the topics more similar to one another with fewer specific words and more common words (to inspect the topics, please see Figure 52). Overall, that makes the topics less easily distinguishable as with the previous model and the Topic 4 and Topic 5 more specific and distinctive than the previous three. As many times before, the first topic is composed of the most common words. The second topic was identified as containing many general words, but with several ones suggesting a war / fight subject. The example words would be 'coup' – 'shot' or 'blow', 'force', 'bombe' – 'bomb', or 'guerre' – 'war'.

The third topic is the hardest to evaluate among all the topics as it contains a mixture of words with the majority linked to food and drink such as 'pastis', 'eau' – 'water', 'manger' – 'to eat', 'omelette', 'gout' – 'taste'. However, there are several words which break the coherence. The most disturbing examples would be 'guerre' – 'war' or 'cauchemar' – 'nightmare'.

At first sight, the Topic 4 seems the most relevant to our research as it was named 'Technology and power'. It contains words related to the technologies, but also words related to money, power, and war.

The terms possibly related to technology are:

- 'laser'
- 'Armes_laser' – 'laser weapons'
- 'energie' – 'energy'

- ‘optique’ – ‘optic’
- ‘developpe’ – from ‘développer’ (with removed accents) – ‘to develop’
- ‘programme’ – ‘program’
- ‘jeu’ – ‘game’
- ‘portable – ‘cell phone’ (as one of the possible meanings)
- ‘nouveau’ – ‘new’
- ‘technologie’ – ‘technology'

However, the actual relevance to our research is limited by the fact that the aforementioned words have very low frequency and thus it is questionable how important they really are for the corpus.

The words related to power, fight, and money are:

- ‘arme’ – ‘weapon’
- ‘defense’
- ‘armes_laser’
- ‘pouvoir’ – ‘power’ (as one of the meanings)
- ‘dollar’
- ‘puissance’ – ‘power’
- ‘etat’ – ‘état’ – ‘state’
- ‘cible’ – ‘target’

The Topic 5 was rather decisively identified as ‘Propaganda’ due to terms like ‘ment’, ‘mentir’ – ‘to lie’, ‘parler’ – ‘to talk’, ‘dire’ – ‘to say’, ‘contredire’ – ‘to contradict’, ‘boniment’ – ‘sweet talk’, ‘payer’ – ‘to pay’, ‘voiler’ – ‘to cover’ in the context with the words ‘enterprise’, ‘monsieur’ – ‘mister’, ‘poli’, and ‘emission’ – ‘broadcast’ (as one of the meanings). This combination of words indicates a sentiment of mistrust towards what is being told by the institutions and those in power. The presence of this topic confirms the general anti-systemic mood which was highlighted in the main corpus highest coherence model.

The compromise model

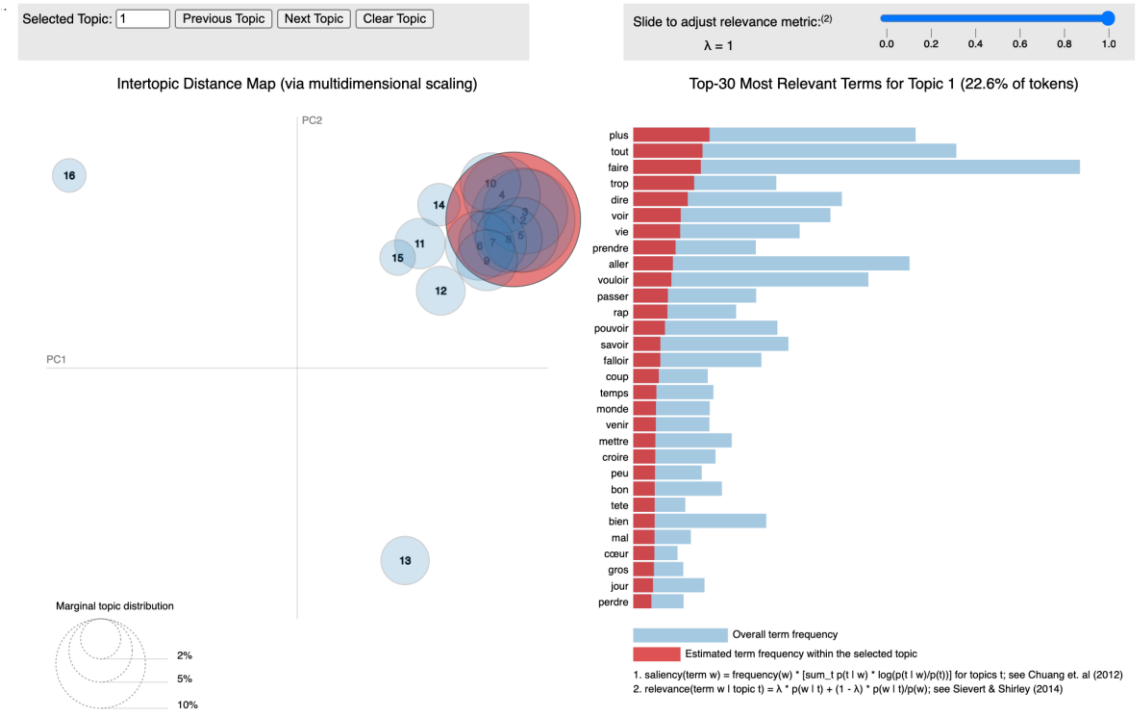

Figure 53: The compromise model – subcorpus

Topic 1 – General 1	Topic 2 – General / Colloquial speech	Topic 3 – General 2	Topic 4 – General / Food and drink	Topic 5 – General 3
plus	faire	aller	faire	dire
tout	tout	tout	aller	plus
faire	aime	faire	plus	vouloir
trop	aller	plus	vie	tout
dire	vouloir	ouai	vouloir	trop
voir	rap	jour	tout	vie
vie	mec	savoir	bien	comprendre
prendre	savoir	laisse	ghetto	pourquoi
aller	plus	bien	savoir	voir

vouloir	dire	soir	falloir	faire
passer	mettre	passer	bon	bien
rap	voir	pouvoir	sale	falloir
pouvoir	pouvoir	dire	pastis	appeler
savoir	niqu	an	gout	savoir
falloir	vie	bon	voir	toujours
coup	trop	meme	passer	prendre
temps	pute	toujours	peu	man
monde	croire	croire	alors	meme
venir	regarde	vie	mettre	etat
mettre	bien	putain	oeuf	style
croire	coup	seul	temps	gros
peu	baise	attendre	venir	star
bon	gars	voir	ome_om	alors
tete	falloir	reste	aime	mettre
bien	fond	joue	donc	rap
mal	cher	falloir	manger	aujourd_hui
coeur	grand	vouloir	fois	bas
gros	reproche	trop	gros	ouai
jour	prendre	autre	valoir	autant
perdre	mal	juste	meme	reste

Figure 54: The compromise model – subcorpus – top five topics

The subcorpus compromise model differs from the main corpus compromise model mainly by having more favourable coherence and perplexity values. It is also largely concentrated around the Topic 1 and the top five topics are closely linked together

(please see Figure 53 for visualisation and Figure 54 for the table of topics). Only two smaller topics are exceptions to this formation. That makes the topics share more words and the topics are even more similar to one another than in the subcorpus lowest perplexity model. Overall, it is the least distinctive model. This observation suggests that the best performing models are those based solely on coherence score, but further testing would be needed to confirm such hypothesis and that is beyond the scope of this paper. The two notable topics are Topic 2 and Topic 4. Topic 2 – Colloquial speech resembles the subcorpus lowest perplexity model Topic 2 of the same name. It contains colloquial terms such as 'mec', 'pute','baise', 'gars' – another word for a 'guy'. The Topic 4 – General/Food and drink shares some characteristics with the subcorpus highest coherence model Topic 3 – Food and drink/Unidentified. It contains words such as 'manger', 'pastis', 'gros' – 'fat', or 'oeuf' – 'egg'. The reason for this topic's presence suggests the link to the overall importance of food culture in France and portrayals of daily life in rap songs.

Topics frequent words analysis

	Lowest Perplexity model	Highest Coherence model	Compromise model
Corpus	58	17	54
Subcorpus	60	25	62

Figure 55: The overview of frequent words across corpora models

The most frequent verbs	
avoir	0
faire	5
aller	5
dire	5

vouloir	4
pouvoir	3
savoir	5
voir	4
aimer	0
être	0
	31
The most frequent nouns	
vie	5
monde	2
temps	4
jour	2
tête	1
rap	2
frère	0
fois	0
son	0
amour	0
	16
The most frequent adjectives	
tout	5
bon	3
seul	1
gros	0
même	2
beau	0

petit	1
grand	0
vrai	1
	13
Total count	60

Figure 56: The lowest perplexity model – an overview of frequent words

The most frequent verbs	
avoir	0
faire	3
aller	4
dire	4
vouloir	4
pouvoir	3
savoir	3
voir	4
aimer	0
être	0
	25
The most frequent nouns	
vie	2
monde	0
temps	1

jour	2
tête	0
rap	1
frère	0
fois	0
son	0
amour	0
	6
The most frequent adjectives	
tout	3
bon	1
seul	1
gros	0
même	1
beau	0
petit	2
grand	1
vrai	0
	9
Total count	40

Figure 57: The highest coherence model – an overview of frequent words

The most frequent verbs	
avoir	0
faire	5
aller	4
dire	4
vouloir	5
pouvoir	3
savoir	5
voir	5
aimer	0
être	0
	31
The most frequent nouns	
vie	5
monde	1
temps	2
jour	2
tête	1
rap	3
frère	0
fois	1
son	0
amour	0

	15
The most frequent adjectives	
tout	5
bon	3
seul	1
Gros	3
même	3
beau	0
petit	0
grand	1
vrai	0
	16
Total count	62

Figure 58: The compromise model – an overview of frequent words

As for the main corpus, the topics corpus frequent words analysis was conducted on the subcorpus. Overall, the subcorpus topics contain more frequent words than the topic in the main corpus (please see Figure 55 for precise numbers). The subcorpus lowest perplexity model exceeds the corpus lowest perplexity frequent words number with 2 words, the corpus highest coherence model frequent words number is exceeded by the subcorpus highest coherence model with 8 words, and the number of frequent words in the subcorpus compromise model is also 8 words higher than in the corpus compromise model (please view Figure 56, Figure 57, and Figure 58 for precise numbers). This statistics confirms the observation that the subcorpus models' top five

topics contained a higher number of more common words which made it more difficult to identify the distinctive topics. Overall, it can mean that the subcorpus based on media words presence gravitates towards more general topics, discussing the reality of everyday life. Vice versa, thanks to its broadness, the main corpus encompasses more specific and distinctive topics. However, both the highest coherence models (which also contain less high frequency words) contain, as a strong element, the topics of censorship and propaganda. This correlates with the media being the main communication channel of state and private sectors. Another sentiment present in both the highest Coherence models is the theme of fight and war. This correlates with the anti-systemic mood and feeling of struggle in life often present in French rap music.

8 Conclusion

To conclude, our findings mostly align with the findings from the literature review. The top ten frequent nouns and verbs reflect the topics differences between the main corpus and the reference French web corpus, but mainly the top ten adjectives confirm the fact that, despite the large non-standard vocabulary, statistically, the language is actually not that different from more standard French. However, the bigram analysis shows that interjections are more common in the French rap corpus which makes sense due to the spoken nature of the lyrics in comparison with web corpus.

In terms of the importance of media words – the subsets containing media related words comprise a significant part of the main corpus – approximately one sixth of the songs. Also, the analysis of the top 20 highest frequency media words shows that the most significant topic in the media related words is communication and internet. The most significant media term is "game" which highlights both the ludic aspect of rap rites and the interconnectedness with the media terms as well as the usage of English borrowings.

To summarise the most important outcomes from the Topic modelling section, the corpus highest coherence model confirms two statements previously found through the literature review – the largest topic seems to be surprisingly rather banal together with very common words inside. Otherwise, the next four significant topics confirm the importance of anti-systemic sentiment, the theme of censorship and freedom of speech, and multiple identities of rappers.

Furthermore, the subcorpus highest coherence model confirms the anti-systemic mood and marks the topic of propaganda / mistrust towards what authorities claim. In the results, there is also a separate topic of technology and power, but its legitimacy would need to be further investigated.

To answer the initial research questions, we can support the claim that the language used in French rap lyrics is mostly standard French mixed with a smaller percentage of rich non-standard vocabulary. In terms of the topics, we can support the claim that the topics in French rap are not primarily about resistance as the largest topic is rather banal. However, the rest of the topics suggest an anti-systemic sentiment as well as topics of every-day struggle. In terms of importance of media and technology, according

to the word frequencies analysis and topic modelling, we can suggest that it is a specific subject in French rap lyrics but cannot be called as particularly significant in comparison with the other matters. However, its importance is linked to the aforementioned topics of state censorship and false claims in the media by the authorities.

9 Discussion

The limits of this study define the possible following research. Further research, which wants to develop on findings from this paper, should focus on eliminating these three limitations.

Firstly, in order to get more refined results, a more curated corpus needs to be developed or found. Rap Corpus from Masaryk University is a very good candidate for the next research due to their focus on quality and maintaining the corpus with deep knowledge of the domain.

Secondly, the technological limitations of topic modelling need to be further examined and faulty mechanisms fixed. Deeper understanding of LDA algorithm and topic coherence as well as further research in the domain could help to sharpen the results.

And finally, more qualitative study could bring further understanding into how the topics get formed and their legitimacy.

10 List of references

ALETRAS, Nikolaos et al. Evaluating Topic Coherence Using Distributional Semantics. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* [online]. 2013. Available at: https://aclanthology.org/W13-0102/

ALSUMAIT, Loulwah et al. Topic Significance Ranking of LDA Generative Models. In: BUNTINE, Wray et al. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science*. Berlin: Springer. 2009, vol 5781.

ARTZGILD. Les meilleurs rappeurs français. In: *Sens Critique* [online]. 2020[cit. 2020-09-09]. Available at: https://www.senscritique.com/liste/Les_meilleurs_rappeurs_francais/1266999

Battle Rap. In: *Rap Wiki* [online]. 2022 [accessed 2021-10-02]. Available at: https://rap.fandom.com/wiki/Battle_Rap

BÉRU, Laurent. Popularisation et récuperation d'un marginalisme artistique :Le rap, une liberté d'expression mort-née ou mort vivante ? *Questions de communication* [online]. 2006, no. 9 [cit. 2021-10-12]. Available at: https://www.cairn.info/revue-questions-de-communication-2006-1-page-251.htm

BÉTHUNE, Christian. Le Rap: Une esthétique hors la loi. Autrement, 2003. ISBN: 978-2746703841.

BLEI, David et al. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* [online]. 2003, vol. 3, pp 993–1022 [cit. 2021-10-07]. Available at: https://dl.acm.org/doi/10.5555/944919.944937

BOUCHER, Manuel. Rap, expression des lascars : Significations et enjeux du Rap dans la société française. Paris: L'Harmattan, 1998, p. 50. ISBN : 2-7384-7380-6.

CCM BENCHMARK. Charcler. In: *Linternaute* [online]. 2021 [cit. 2021-10-12]. Available at: https://www.linternaute.fr/dictionnaire/fr/definition/charcler/

CHENG, Xueqi et al. A biterm topic model for short texts. *In: Proceedings of the 22nd international conference on World Wide Web* [online]. 2013 [cit. 2021-10-13]. Available at: https://www.researchgate.net/publication/262244963_A_biterm_topic_model_for_short_texts

CHANG, Jonathan et al. Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems [online]. 2009 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/221618226_Reading_Tea_Leaves_How_Humans_Interpret_Topic_Models

CHARLES UNIVERSITY. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics. Udpipe description. In: *GitHub.com* [code repository]. 2017 [cit. 2021-10-0] Available at: https://github.com/ufal/udpipe

CHUANG, Jason, Jeffrey HEER, and Christopher D. MANNING. Termite: Visualization Techniques for Assessing Textual Topic Models. In: *Advanced Visual Interfaces* [online]. 2012 [cit. 2021-10-02]. Available at: http://vis.stanford.edu/files/2012-Termite-AVI.pdf

CONRAD, Kate et al. *Controversial Rap Themes, Gender Portrayals and Skin Tone Distortion: A Content Analysis of Rap Music Videos*. Journal of Broadcasting & Electronic Media [online]. 2009, vol. 53, no. 1, pp. 134–156 [cit. 2021-09-25]. Available at: https://www.researchgate.net/publication/232932971_Controversial_Rap_Themes_Gender_Portrayals_and_Skin_Tone_Distortion_A_Content_Analysis_of_Rap_Music_Videos

Corpus of the French Web. In: *Sketch Engine* [online]. 2017 [accessed 2021-10-02]. Available at: https://www.sketchengine.eu/frtenten-french-corpus/

DEVILLA, Lorenzo. C'est pas ma France à moi... »: identités plurielles dans le rap français. *Synergies Italie* [online]. 2011, vol. 7, pp. 75–84 [cit. 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_proquest_journals_2062085785

DIALLO, David. La musique rap comme forme de résistance ? *Revue de recherche en civilisation américaine* [online]. 2009, vol. 1 [cit. 2021-09-25]. Available at: https://journals.openedition.org/rrca/80

DRAME, Mamadou and Assane NDIAYE. Le français employé dans le rap : menace ou chance ? Comparaison avec la poésie. *Anadiss* [online]. 1st June 2012, vol.1, no. 13, pp. 121–147 [cit. 2021-09-25] Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_doaj_primary_oai_doaj_org_article_58cb71abc9744e5db6eafc0ea66c84f9

Game. In: *Culturap* [online]. 2021 [cit. 2021-10-07]. Available at: https://culturap.fr/glossaire/cest-quoi-le-game/

Game. In: *Game studies Wiki* [online]. 2022 [cit. 2021-10-07]. Available at: https://game-studies.fandom.com/wiki/Game

GEORGE, Clint P. and Hani DOSS. Principled selection of hyperparameters in the latent dirichlet allocation model. *The Journal of Machine Learning Research* [online]. 2017, vol. 18, no. 1, pp 5937–5974 [cit. 2021-10-09]. Available at: https://dl.acm.org/doi/abs/10.5555/3122009.3242019

Genius.com API documentation. In: *Genius.com* [online]. 2020 [cit. 2021-10-02]. Available at: https://docs.genius.com

GUIZMO. Bonnet d'âne. In: *Genius.com* [online]. 2013 [cit. 2021-20-05]. Available at: https://genius.com/Guizmo-bonnet-dane-lyrics

HAMMOU, Karim. La vérité au risque de la violence. Remarques sur la stylistique du rap en français. In: MOÏSE, Claudine. *De l'impolitesse à la violence verbale*. Avignon, 2005. pp. 203–222.

HAMMOU, Karim. *Une histoire du rap en France*. Paris: La Découverte, 2012. 1st edition. 978–2707171375.

HAMMOU, Karim. *Une histoire du rap en France*. Paris: La Découverte, 7th May 2014. 2nd edition. 978–2707181985.

HASSA, Samira. Kiff my zikmu: Symbolic Dimensions of Arabic, English and Verlan in French rap texts. In: TEKOURAFI, Marina. *The Languages of Global Hip Hop*. Bloomsbury Publishing, 2010, pp. 44–66. ISBN 9780826431608.

JOUVENET, Morgan. Rap, techno, électro... Le musicien entre travail artistique et critique sociale. Paris : Éditions de la Maison des sciences de l'homme, 2006, p. 106.

JUUL, Jesper. The Game, the Player, the World: Looking for a Heart of Gameness" In: COPIER, Marinka and Joost RAESSENS. *Level Up: Digital Games Research Conference Proceedings*. Utrecht: Utrecht University, 2003, pp. 30–45.

KAPADIA, Shashank. Evaluate Topic Model in Python: Latent Dirichlet Allocation (LDA). In: *GitHub.com* [code repository]. 2020 [cit. 2021-10-08]. Available at: https://github.com/kapadias/mediumposts/blob/master/natural_language_processing/topic_modeling/notebooks/Evaluate%20Topic%20Models.ipynb

KORENČIC, Damir et al. Document-based Topic Coherence Measures for News Media Text. *Expert Systems with Applications* [online]. 2018 [cit. 2021-10-10]. Available at: https://bib.irb.hr/datoteka/950542.Document-based_Topic_Coherence_Measures_for_News_Media_Text_2018_preprint.pdf

KUDLIČKOVÁ, Pavla. *Français contemporain des cités dans les chansons de rap*. Praha, 2009. Master thesis. Charles University. Faculty of Education, Department of French Language and Literature.

MAYAUD, Isabelle. Karim Hammou, Une histoire du rap en France. *Transposition* [online]. 2015, no. 5 [cit. 2021-09-24]. Available at: http://journals.openedition.org/transposition/1294

MIMMO, David et al. Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* [online]. 2011 [cit. 2021-10-10]. Available at: https://aclanthology.org/D11-1024/

MONET, Liza. Bad Buzz. In: *Genius.com* [online]. 2020 [cit. 2021-10-07]. Available at: https://genius.com/Liza-monet-bad-buzz-lyrics

MONET, Liza. Legal Geneva. In: *Genius.com* [online]. 2020 [cit. 2021-10-07]. Available at: https://genius.com/Liza-monet-legal-geneva-lyrics

MORALES, Joseph et al. R.A.P. Rap Analysis Project. In: *Berkeley School of Information* [online]. 2015 [cit. 2020-09-25]. Available at: https://www.ischool.berkeley.edu/projects/2015/rap-rap-analysis-project

MORHAIN, Yves and Émilie MORHAIN. La création adolescente. *Adolescence* [online]. 2011, vol. 29, no 1), pp. 87–111 [cit. 2021-10-13]. Available at: https://www.cairn.info/revue-adolescence-2011-1-page-87.htm

MORSAY. Message à Internet. In: Genius.com [online]. 2008 [cit. 2021-10-07]. Available at: https://genius.com/Morsay-message-a-internet-annotated

MUSAT, Claudiu et al. Improving topic evaluation using conceptual knowledge. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* [online]. 2011 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/220815876_Improving_Topic_Evaluation_Using_Conceptual_Knowledge

NEWMAN, David et al. Automatic Evaluation of Topic Coherence. In: Human Language Technologies: Conference of the North American Chapter of the Association

of Computational Linguistics [online]. 2010 [cit. 2021-10-10]. Available at: https://aclanthology.org/N10-1012.pdf

NG, Jason. Genius-lyrics-search. In: *GitHub.com* [code repository]. 2015 [accessed 2021-09-27]. Available at: https://github.com/jasonqng/genius-lyrics-search

NGAMALEU, Jovensel. Poésie et discours social dans le rap français et camerounais : Booba, La Fouine, Valsero et Maalhox le Viber. *Itinéraires* [online]. 10th December 2021, vol.2020, no. 3 [cit. 2021-12-20]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_crossref_primary_10_4000_itineraires_9022

NGUYEN, Eric. Text Mining and Network Analysis of Digital Libraries in R. In: ZHAO, Yanchang and Yonghua CEN. *Data Mining Applications with R* [online]. 2014, pp. 95–115 [cit. 2021-10-07]. Available at: https://www.sciencedirect.com/science/article/pii/B9780124115118000049

NIKOLENKO, Sergey. Topic Quality Metrics Based on Distributed Word Representations. In: *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information* [online]. 2016, pp. 1029–1032 [cit. 2021-10-10]. Available at: https://dl.acm.org/doi/10.1145/2911451.2914720

O'CALLAGHAN, Derek et al. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* [online]. 2015, vol. 42, no. 13, pp. 5645–5657 [cit. 2021-10-10]. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417415001633

OLEINIKOV, Pavel. FInding the best number of topics. In: *Datacamp* [online]. 2022 [cit. 2021-10-10]. Available at: https://campus.datacamp.com/courses/topic-modeling-in-r/how-many-topics-is-enough?ex=1

ONDRUŠKOVÁ, Dana. *Lexique de la drogue dans le corpus des chansons de rap: analyse sémantique en synchronie dynamique*. Brno, 2014. Master thesis. Masaryk University, Faculty of Arts, Department of Romance Languages and Literature.

PAINE, Skye. The Quadrilingual Vocabulary of French Rap. *Revue Kinephanos: Multilingualism in Popular Arts* [online]. 2012, vol. 3, no. 1 [cit. 2021-09-24]. Available at: https://www.kinephanos.ca/Revue_files/2012-paine.pdf

PECQUEUX, Anthony. La violence du rap comme katharsis : vers une interprétation politique. *Volume ! La revue des musiques populaires* [online]. 2004, vol. 3, no. 2 [cit. 2021-09-24]. Available at: https://journals.openedition.org/volume/1959

PECQUEUX, Anthony. VOIX DU RAP: Essai de sociologie de l'action musicale. L'Harmattan, 2007. ISBN : 978-2-296-04463-0.

PETETIN, Véronique. Slam, rap et « mondialité ». *Études* [online]. 2009, pp. 797–808 [cit. 2021-10-13]. Available at: https://www.cairn.info/revue-etudes-2009-6-page-797.htm

PIOLET, Vincent. *Regarde ta jeunesse dans les yeux : Naissance du hip-hop français*, 1980-1990. Le mot et le reste, 20th April 2017. ISBN 978-2360542901.

PRABHAKARAN, Selva. Topic Modeling with Gensim (Python). In: *Machine Learning +* [online]. 2018 [cit. 2021-10-08]. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#14computemodelperplexityandcoherencescore

PRÉVOS, André J. M. "In It for the Money": Rap and Business Cultures in France. *Popular Music and Society* [online]. 2003, vol. 26, no. 3 [cit. 2021-10-12]. Available at: https://www.tandfonline.com/doi/abs/10.1080/03007760320001449l3?journalCode=rpms20

RAMIREZ, Eduardo. Topic model validation. *Neurocomputing* [online]. 2012, vol. 76, no. 1 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/220549982_Topic_model_validation

RAMRAKHIYANI, Nitin et al. Measuring topic coherence through optimal word buckets. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* [online]. 2017 [cit. 2021-10-10]. Available at: https://aclanthology.org/E17-2070/

Rap Corpus. [online]. 2021 [accessed 2021-09-10]. Available at: https://is.muni.cz/do/phil/Pracoviste/URJL/rapcor/index.html

Rap Research Lab. [online]. 2020 [cit. 2021-09-28]. Available at: https://rapresearchlab.com

ROBERTS, David and Julia SILGE. *Text Mining with R: A Tidy Approach*. [online] O'Reilly Media. 1st edition. 2017 [cit. 2020-09-10] Available at: https://www.tidytextmining.com/

ROBERTS, David and Julia SILGE. *Tidytext 0.3.2* [software].[accessed 2021-10-01]. Available at: https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html

RÖDER, Michael, et al. Exploring the Space of Topic Coherence Measures. In: *WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* [online]. 2015, pp. 399–408 [cit. 2021-10-10]. Available at: https://dl.acm.org/doi/10.1145/2684822.2685324

ROSNER, Frank et al. Evaluating topic coherence measures. In: *Conference: Neural Information Processing Systems Foundation (NIPS 2013) – Topic Models Workshop* [online]. 2013 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/261101181_Evaluating_topic_coherence_measures

ŘEHŮŘEK, Radim. People behind Gensim. In: *radimrehurek.com* [online]. [cit. 2021-10-01]. Available at: https://radimrehurek.com/gensim/people.html

ŘEHŮŘEK, Radim. What is Gensim? In: *radimrehurek.com* [online]. [cit. 2021-10-01]. Available at: https://radimrehurek.com/gensim/intro.html#what-is-gensim

ŘEHŮŘEK, Radim and Petr SOJKA. *Gensim 4.1.2* [software].[accessed 2021-10-02]. Available at: https://pypi.org/project/gensim/

SCHULTZ, Jack. Getting Song Lyrics from Genius's API + Scraping. In: *Big-Ish Data* [online]. 2016 [cit. 2021-09-27]. Available at: https://bigishdata.com/2016/09/27/getting-song-lyrics-from-geniuss-api-scraping/

SHUSTERMAN, Richard. Pragmatisme, art et violence : le cas du rap. *Mouvements* [online]. 2003, vol. 2, no. 26, pp. 116–122.

SIEVERT, Carson. A topic model for movie reviews. In: *Carson Paul Sievert* [online]. 2018 [cit. 2021-10-02]. Available at: https://ldavis.cpsievert.me/reviews/reviews.html

SIEVERT, Carson. LDAvis. In: *GitHub.com* [code repository]. 2018 [accessed 2021-09-27]. Available at: https://github.com/cpsievert/LDAvis

SIEVERT, Carson and Kenneth SHIRLEY. LDAvis: A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* [online]. 2014 [cit. 2021-10-12]. Available at: https://aclanthology.org/W14-3110/

SOLTOFF, Benjamin. Topic modeling. In: *Computing for the social sciences* [online]. 2021 [cit. 2021-10-10]. Available at: https://cfss.uchicago.edu/notes/topic-modeling

SONNETTE, Marie. Des mises en scène du « nous » contre le « eux » dans le rap français: De la critique de la domination postcoloniale à une possible critique de la domination de classe. *Sociologie de l'Art* [online]. 2015, no. 1–2, pp. 153–177 [cit. 2021-09-24]. Available at: https://www.cairn.info/revue-sociologie-de-l-art-2015-1-page-153.htm

TAMAGNE, Florence. Karim Hammou, Une histoire du rap en France, Paris, La Découverte, 2012, 302 P., ISBN 978-2707171375. *Revue d'histoire moderne & contemporaine* [online]. 2014, no. 61-1 [cit. 2021-09-24]. Available at: https://www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2014-1-page-190.htm

TRIMAILLE, Cyril. *De la planète Mars... Codes, langages, identités : étude sociolinguistique de textes de rap marseillais*. Grenoble, 1999. Master thesis. Université Stendhal.

Ukaz.cuni.cz [online].[accessed 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/discovery/search?vid=420CKIS_INST:UKAZ&lang=cs

VERBEKE, Martin. Rapping through time: an analysis of non-standard language use in French rap. *Modern & contemporary France* [online]. 3rd July 2017, vol.25, no. 3, pp. 281–298 [cit. 2021-09-20]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_crossref_primary_10_1080_09639489_2017_1304903

VERBEKE, Martin. French Rap Genres And Language: An Analysis Of The Impact Of Genres

On Non-Standard Language Use. *Nottingham French studies* [online]. 2019, vol.58, no. 1, pp. 44–63 [cit. 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_edinburghupress_primary_10_3366_nfs_2019_0235

VICHERAT, Mathias. *Pour une analyse textuelle du rap français*. L'Harmattan, 2001. 2-7475-1089-1.

WALLACH, Hanna M. Topic Modeling: Beyond Bag-of-Words. In: *Proceedings of the 23rd International Conference on Machine Learning* [online]. 2006, pp. 977–984 [cit. 2021-10-07]. Available at: http://dirichlet.net/pdf/wallach06topic.pdf

WALLACH, Hanna M. et al. Rethinking LDA: Why priors matter. In: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems* [online]. 2009 [cit. 2021-10-10]. Available at: https://papers.nips.cc/paper/2009/hash/0d0871f0806eae32d30983b62252da50-Abstract.html

WICKHAM, Hadley. Tidy data. *The Journal of Statistical Software.* 2014, vol. 59. Available at: https://vita.had.co.nz/papers/tidy-data.html

WIJFFELS, Jan et al. *Udpipe 0.8.6* [software]. [accessed 2021-10-02]. Available at: https://CRAN.R-project.org/package=udpipe

YONNET, Paul. Rap: Musique, langage, violence, sexe. *Le Débat* [online]. 2000, vol. 5, no. 112, pp. 124–127 [cit. 2021-10-12]. Available at: https://www.cairn.info/revue-le-debat-2000-5-page-124.htm

ZÁVODSKÁ, Pavlína. *Vulgarismes dans un corpus de chansons de rap : étude lexicométrique en synchronie dynamique.* Brno, 2009. Master thesis. Masaryk University, Faculty of Arts, Department of Romance Languages and Literature.

ZELENKOVÁ, Anna. *Arabismes dans les chansons de rap français : traitement lexicographique, adaptation phonique et rôle de l'origine des rappeurs*. Brno, 2013. Master thesis. Masaryk University, Faculty of Arts, Department of Romance Languages and Literature.

ZUBČEKOVÁ, Helena. Langage violent dans le rap français : caractéristique ou cliché? *Svět literatury* [online]. 1st December 2015, vol.25 [cit. 2021-09-25]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_doaj_primary_oai_doaj_org_article_c1d22dd24a834defa92bc25ec7632af3

11 Bibliography

ALETRAS, Nikolaos et al. Evaluating Topic Coherence Using Distributional Semantics. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* [online]. 2013. Available at: https://aclanthology.org/W13-0102/

ALSUMAIT, Loulwah et al. Topic Significance Ranking of LDA Generative Models. In: BUNTINE, Wray et al. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science*. Berlin: Springer. 2009, vol 5781.

ARTZGILD. Les meilleurs rappeurs français. In: *Sens Critique* [online]. 2020[cit. 2020-09-09]. Available at: https://www.senscritique.com/liste/Les_meilleurs_rappeurs_francais/1266999

Battle Rap. In: *Rap Wiki* [online]. 2022 [accessed 2021-10-02]. Available at: https://rap.fandom.com/wiki/Battle_Rap

BLEI, David et al. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* [online]. 2003, vol. 3, pp 993–1022 [cit. 2021-10-07]. Available at: https://dl.acm.org/doi/10.5555/944919.944937

CCM BENCHMARK. Charcler. In: *Linternaute* [online]. 2021 [cit. 2021-10-12]. Available at: https://www.linternaute.fr/dictionnaire/fr/definition/charcler/

CHENG, Xueqi et al. A biterm topic model for short texts. *In: Proceedings of the 22nd international conference on World Wide Web* [online]. 2013 [cit. 2021-10-13]. Available at: https://www.researchgate.net/publication/262244963_A_biterm_topic_model_for_short_texts

CHANG, Jonathan et al. Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems [online]. 2009 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/221618226_Reading_Tea_Leaves_How_Humans_Interpret_Topic_Models

CHARLES UNIVERSITY. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics. Udpipe description. In: *GitHub.com* [code repository]. 2017 [cit. 2021-10-0] Available at: https://github.com/ufal/udpipe

CHUANG, Jason, Jeffrey HEER, and Christopher D. MANNING. Termite: Visualization Techniques for Assessing Textual Topic Models. In: *Advanced Visual Interfaces* [online]. 2012 [cit. 2021-10-02]. Available at: http://vis.stanford.edu/files/2012-Termite-AVI.pdf

CONRAD, Kate et al. *Controversial Rap Themes, Gender Portrayals and Skin Tone Distortion: A Content Analysis of Rap Music Videos*. Journal of Broadcasting & Electronic Media [online]. 2009, vol. 53, no. 1, pp. 134–156 [cit. 2021-09-25]. Available at: https://www.researchgate.net/publication/232932971_Controversial_Rap_Themes_Gender_Portrayals_and_Skin_Tone_Distortion_A_Content_Analysis_of_Rap_Music_Videos

Corpus of the French Web. In: *Sketch Engine* [online]. 2017 [accessed 2021-10-02]. Available at: https://www.sketchengine.eu/frtenten-french-corpus/

DEVILLA, Lorenzo. C'est pas ma France à moi... »: identités plurielles dans le rap français. *Synergies Italie* [online]. 2011, vol. 7, pp. 75–84 [cit. 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_proquest_journals_2062085785

DIALLO, David. La musique rap comme forme de résistance ? *Revue de recherche en civilisation américaine* [online]. 2009, vol. 1 [cit. 2021-09-25]. Available at: https://journals.openedition.org/rrca/80

DRAME, Mamadou and Assane NDIAYE. Le français employé dans le rap : menace ou chance ? Comparaison avec la poésie. *Anadiss* [online]. 1st June 2012, vol.1, no. 13, pp. 121–147 [cit. 2021-09-25] Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_doaj_primary_oai_doaj_org_article_58cb71abc9744e5db6eafc0ea66c84f9

Game. In: *Culturap* [online]. 2021 [cit. 2021-10-07]. Available at: https://culturap.fr/glossaire/cest-quoi-le-game/

Game. In: *Game studies Wiki* [online]. 2022 [cit. 2021-10-07]. Available at: https://game-studies.fandom.com/wiki/Game

GEORGE, Clint P. and Hani DOSS. Principled selection of hyperparameters in the latent dirichlet allocation model. *The Journal of Machine Learning Research* [online]. 2017, vol. 18, no. 1, pp 5937–5974 [cit. 2021-10-09]. Available at: https://dl.acm.org/doi/abs/10.5555/3122009.3242019

Genius.com API documentation. In: *Genius.com* [online]. 2020 [cit. 2021-10-02]. Available at: https://docs.genius.com

GUIZMO. Bonnet d'âne. In: *Genius.com* [online]. 2013 [cit. 2021-20-05]. Available at: https://genius.com/Guizmo-bonnet-dane-lyrics

HAMMOU, Karim. *Une histoire du rap en France*. Paris: La Découverte, 2012. 1st edition. 978-2707171375.

HAMMOU, Karim. *Une histoire du rap en France*. Paris: La Découverte, 7th May 2014. 2nd edition. 978-2707181985.

HASSA, Samira. Kiff my zikmu: Symbolic Dimensions of Arabic, English and Verlan in French rap texts. In: TEKOURAFI, Marina. *The Languages of Global Hip Hop*. Bloomsbury Publishing. 2010, pp. 44–66. 9780826431608.

JUUL, Jesper. The Game, the Player, the World: Looking for a Heart of Gameness" In: COPIER, Marinka and Joost RAESSENS. *Level Up: Digital Games Research Conference Proceedings*. Utrecht: Utrecht University. 2003, pp. 30–45.

KAPADIA, Shashank. Evaluate Topic Model in Python: Latent Dirichlet Allocation (LDA). In: *GitHub.com* [code repository]. 2020 [cit. 2021-10-08]. Available at: https://github.com/kapadias/mediumposts/blob/master/natural_language_processing/topic_modeling/notebooks/Evaluate%20Topic%20Models.ipynb

KORENČIC, Damir et al. Document-based Topic Coherence Measures for News Media Text. *Expert Systems with Applications* [online]. 2018 [cit. 2021-10-10]. Available at: https://bib.irb.hr/datoteka/950542.Document-based_Topic_Coherence_Measures_for_News_Media_Text_2018_preprint.pdf

KUDLIČKOVÁ, Pavla. *Français contemporain des cités dans les chansons de rap*. Praha, 2009. Master thesis. Charles University. Faculty of Education, Department of French Language and Literature.

MAYAUD, Isabelle. Karim Hammou, Une histoire du rap en France. *Transposition* [online]. 2015, no. 5 [cit. 2021-09-24]. Available at: http://journals.openedition.org/transposition/1294

MILON, Alan. Pourquoi le rappeur chante ?: Le rap comme expression de la relégation urbaine. Cités [online]. 2004, vol. 19, no. 3, p. 71 [cit. 2021-10-13]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_hal_shs_oai_HAL_halshs_00153361v1

MIMMO, David et al. Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* [online]. 2011 [cit. 2021-10-10]. Available at: https://aclanthology.org/D11-1024/

MONET, Liza. Bad Buzz. In: *Genius.com* [online]. 2020 [cit. 2021-10-07]. Available at: https://genius.com/Liza-monet-bad-buzz-lyrics

MONET, Liza. Legal Geneva. In: *Genius.com* [online]. 2020 [cit. 2021-10-07]. Available at: https://genius.com/Liza-monet-legal-geneva-lyrics

MORALES, Joseph et al. R.A.P. - Rap Analysis Project. In: *Berkeley School of Information* [online]. 2015 [cit. 2020-09-25]. Available at: https://www.ischool.berkeley.edu/projects/2015/rap-rap-analysis-project

MORSAY. Message à Internet. In: Genius.com [online]. 2008 [cit. 2021-10-07]. Available at: https://genius.com/Morsay-message-a-internet-annotated

MUSAT, Claudiu et al. Improving topic evaluation using conceptual knowledge. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* [online]. 2011 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/220815876_Improving_Topic_Evaluation_Using_Conceptual_Knowledge

NEWMAN, David et al. Automatic Evaluation of Topic Coherence. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics [online]. 2010 [cit. 2021-10-10]. Available at: https://aclanthology.org/N10-1012.pdf

NG, Jason. Genius-lyrics-search. In: *GitHub.com* [code repository]. 2015 [accessed 2021-09-27]. Available at: https://github.com/jasonqng/genius-lyrics-search

NGAMALEU, Jovensel. Poésie et discours social dans le rap français et camerounais : Booba, La Fouine, Valsero et Maalhox le Viber. *Itinéraires* [online]. 10th December 2021, vol.2020, no. 3 [cit. 2021-12-20]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_crossref_primary_10_4000_itineraires_9022

NGUYEN, Eric. Text Mining and Network Analysis of Digital Libraries in R. In: ZHAO, Yanchang and Yonghua CEN. *Data Mining Applications with R* [online]. 2014, pp. 95-115 [cit. 2021-10-07]. Available at: https://www.sciencedirect.com/science/article/pii/B9780124115118000049

NIKOLENKO, Sergey. Topic Quality Metrics Based on Distributed Word Representations. In: *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information* [online]. 2016, pp. 1029–1032 [cit. 2021-10-10]. Available at: https://dl.acm.org/doi/10.1145/2911451.2914720

O'CALLAGHAN, Derek et al. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* [online]. 2015, vol. 42, no. 13, pp. 5645–5657 [cit. 2021-10-10]. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417415001633

OLEINIKOV, Pavel. FInding the best number of topics. In: *Datacamp* [online]. 2022 [cit. 2021-10-10]. Available at: https://campus.datacamp.com/courses/topic-modeling-in-r/how-many-topics-is-enough?ex=1

ONDRUŠKOVÁ, Dana. *Lexique de la drogue dans le corpus des chansons de rap: analyse sémantique en synchronie dynamique*. Brno, 2014. Master thesis. Masaryk University, Faculty of Arts, Department of Romance Languages and Literature.

PAINE, Skye. The Quadrilingual Vocabulary of French Rap. *Revue Kinephanos: Multilingualism in Popular Arts* [online]. 2012, vol. 3, no. 1 [cit. 2021-09-24]. Available at: https://www.kinephanos.ca/Revue_files/2012-paine.pdf

PECQUEUX, Anthony. La violence du rap comme katharsis : vers une interprétation politique. *Volume ! La revue des musiques populaires* [online]. 2004, vol. 3, no. 2 [cit. 2021-09-24]. Available at: https://journals.openedition.org/volume/1959

PECQUEUX, Anthony. Un témoignage adressé. Parole du rap et parole collective. *Les cahiers de psychologie politique*. 2005. Available at: https://hal.archives-ouvertes.fr/hal-00349605

PIOLET, Vincent. *Regarde ta jeunesse dans les yeux : Naissance du hip-hop français*, 1980–1990. Le mot et le reste, 20th April 2017. 978-2360542901.

PODHORNÁ-POLICKÁ, Alena. Debov Valéry, Diko des rimes en verlan dans le rap français. Paris: La Maison du dictionnaire, 2012, 315 pp. 978 28 560 8290 4 (broché). *Journal of French language studies* [online]. 2016, vol.26, no.3, pp. 381–382 [cit. 2021-10-12]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_crossref_primary_10_1017_S0959269515000484

PRABHAKARAN, Selva. Topic Modeling with Gensim (Python). In: *Machine Learning + [online]*. 2018 [cit. 2021-10-08]. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#14computemodelperplexityandcoherencescore

RAMIREZ, Eduardo. Topic model validation. *Neurocomputing* [online]. 2012, vol. 76, no. 1 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/220549982_Topic_model_validation

RAMRAKHIYANI, Nitin et al. Measuring topic coherence through optimal word buckets. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* [online]. 2017 [cit. 2021-10-10]. Available at: https://aclanthology.org/E17-2070/

Rap Corpus. [online]. 2021 [accessed 2021-09-10]. Available at: https://is.muni.cz/do/phil/Pracoviste/URJL/rapcor/index.html

Rap Research Lab. [online]. 2020 [cit. 2021-09-28]. Available at: https://rapresearchlab.com

ROBERTS, David and Julia SILGE. *Text Mining with R: A Tidy Approach*. [online] O'Reilly Media. 1st edition. 2017 [cit. 2020-09-10] Available at: https://www.tidytextmining.com/

ROBERTS, David and Julia SILGE. *Tidytext 0.3.2* [software].[accessed 2021-10-01]. Available at: https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html

RÖDER, Michael, et al. Exploring the Space of Topic Coherence Measures. In: *WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* [online]. 2015, pp. 399–408 [cit. 2021-10-10]. Available at: https://dl.acm.org/doi/10.1145/2684822.2685324

ROSNER, Frank et al. Evaluating topic coherence measures. In: *Conference: Neural Information Processing Systems Foundation (NIPS 2013) – Topic Models Workshop* [online]. 2013 [cit. 2021-10-10]. Available at: https://www.researchgate.net/publication/261101181_Evaluating_topic_coherence_measures

ŘEHŮŘEK, Radim. People behind Gensim. In: *radimrehurek.com* [online]. [cit. 2021-10-01]. Available at: https://radimrehurek.com/gensim/people.html

ŘEHŮŘEK, Radim. What is Gensim? In: *radimrehurek.com* [online]. [cit. 2021-10-01]. Available at: https://radimrehurek.com/gensim/intro.html#what-is-gensim

ŘEHŮŘEK, Radim and Petr SOJKA. *Gensim 4.1.2* [software].[accessed 2021-10-02]. Available at: https://pypi.org/project/gensim/

SCHULTZ, Jack. Getting Song Lyrics from Genius's API + Scraping. In: *Big-Ish Data* [online]. 2016 [cit. 2021-09-27]. Available at: https://bigishdata.com/2016/09/27/getting-song-lyrics-from-geniuss-api-scraping/

SIEVERT, Carson. A topic model for movie reviews. In: *Carson Paul Sievert* [online]. 2018 [cit. 2021-10-02]. Available at: https://ldavis.cpsievert.me/reviews/reviews.html

SIEVERT, Carson. LDAvis. In: *GitHub.com* [code repository]. 2018 [accessed 2021-09-27]. Available at: https://github.com/cpsievert/LDAvis

SIEVERT, Carson and Kenneth SHIRLEY. LDAvis: A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* [online]. 2014 [cit. 2021-10-12]. Available at: https://aclanthology.org/W14-3110/

SOLTOFF, Benjamin. Topic modeling. In: *Computing for the social sciences* [online]. 2021 [cit. 2021-10-10]. Available at: https://cfss.uchicago.edu/notes/topic-modeling

SONNETTE, Marie. Des mises en scène du « nous » contre le « eux » dans le rap français: De la critique de la domination postcoloniale à une possible critique de la domination de classe. *Sociologie de l'Art* [online]. 2015, no. 1–2, pp. 153–177 [cit. 2021-09-24]. Available at: https://www.cairn.info/revue-sociologie-de-l-art-2015-1-page-153.htm

TAMAGNE, Florence. Karim Hammou, Une histoire du rap en France, Paris, La Découverte, 2012, 302 P., ISBN 978-2707171375. *Revue d'histoire moderne & contemporaine* [online]. 2014, no. 61–1 [cit. 2021-09-24]. Available at: https://www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2014-1-page-190.htm

Ukaz.cuni.cz [online].[accessed 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/discovery/search?vid=420CKIS_INST:UKAZ&lang=cs

VERBEKE, Martin. Rapping through time: an analysis of non-standard language use in French rap. *Modern & contemporary France* [online]. 3rd July 2017, vol.25, no. 3, pp. 281–298 [cit. 2021-09-20]. Available at:

https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_crossref_primary_10_1080_09639489_2017_1304903

VERBEKE, Martin. French Rap Genres And Language: An Analysis Of The Impact Of Genres

On Non-Standard Language Use. *Nottingham French studies* [online]. 2019, vol.58, no. 1, pp. 44–63 [cit. 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_edinburghupress_primary_10_3366_nfs_2019_0235

VICHERAT, Mathias. *Pour une analyse textuelle du rap français*. L'Harmattan, 2001. 2-7475-1089-1.

WALLACH, Hanna M. Topic Modeling: Beyond Bag-of-Words. In: *Proceedings of the 23rd International Conference on Machine Learning* [online]. 2006, pp. 977–984 [cit. 2021-10-07]. Available at: http://dirichlet.net/pdf/wallach06topic.pdf

WALLACH, Hanna M. et al. Rethinking LDA: Why priors matter. In: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems* [online]. 2009 [cit. 2021-10-10]. Available at: https://papers.nips.cc/paper/2009/hash/0d0871f0806eae32d30983b62252da50-Abstract.html

WICKHAM, Hadley. Tidy data. *The Journal of Statistical Software.* 2014, vol. 59. Available at: https://vita.had.co.nz/papers/tidy-data.html

WIJFFELS, Jan et al. *Udpipe 0.8.6* [software]. [accessed 2021-10-02]. Available at: https://CRAN.R-project.org/package=udpipe

ZÁVODSKÁ, Pavlína. *Vulgarismes dans un corpus de chansons de rap : étude lexicométrique en synchronie dynamique.* Brno, 2009. Master thesis. Masaryk University, Faculty of Arts, Department of Romance Languages and Literature.

ZELENKOVÁ, Anna. *Arabismes dans les chansons de rap français : traitement lexicographique, adaptation phonique et rôle de l'origine des rappeurs*. Brno, 2013. Master thesis. Masaryk University, Faculty of Arts, Department of Romance Languages and Literature.

ZUBČEKOVÁ, Helena. Langage violent dans le rap français : caractéristique ou cliché? *Svět literatury* [online]. 1st December 2015, vol.25 [cit. 2021-09-25]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_doaj_primary_oai_doaj_org_article_c1d22dd24a834defa92bc25ec7632af3

12 List of figures

https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_edinburghupress_primary_10_3366_nfs_2019_0235
Figure 10: VERBEKE, Martin. *Foreign borrowings in the corpus (% of total borrowings)* [table, 2019]. In: VERBEKE, Martin. French Rap Genres And Language: An Analysis Of The Impact Of Genres
On Non-Standard Language Use. *Nottingham French studies* [online]. 2019, vol.58, no. 1, pp. 44–63 [cit. 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_edinburghupress_primary_10_3366_nfs_2019_0235
Figure 11: VERBEKE, Martin. *Themes in the selected lyrics (number of occurrences per 100 words)* [table, 2019]. In: VERBEKE, Martin. French Rap Genres And Language: An Analysis Of The Impact Of Genres
On Non-Standard Language Use. *Nottingham French studies* [online]. 2019, vol.58, no. 1, pp. 44–63 [cit. 2021-09-24]. Available at: https://cuni.primo.exlibrisgroup.com/permalink/420CKIS_INST/1pop0hq/cdi_edinburghupress_primary_10_3366_nfs_2019_0235
Figure 12: KLIMENTOVÁ, Julie. *Most frequent verbs in the main corpus* [bar chart, 2021].
Figure 13: KLIMENTOVÁ, Julie. *Most frequent verbs in the French web corpus 2017* [bar chart, 2021].
Figure 14: KLIMENTOVÁ, Julie. *Most frequent nouns in the main corpus* [bar chart, 2021].
Figure 15: KLIMENTOVÁ, Julie. *Most frequent nouns in the French web corpus 2017* [bar chart, 2021].
Figure 16: KLIMENTOVÁ, Julie. *Most frequent adjectives in the main corpus* [bar chart, 2021].
Figure 17: KLIMENTOVÁ, Julie. *Most frequent adjectives in the French web corpus 2017* [bar chart, 2021].
Figure 18: KLIMENTOVÁ, Julie. *Table of songs count for the corpora* [table, 2021].
Figure 19: KLIMENTOVÁ, Julie. *Main corpus – media words subcorpus ratio – higher ambiguity* [pie chart, 2021].
Figure 20: KLIMENTOVÁ, Julie. *Main corpus – media words subcorpus ratio – lower ambiguity* [pie chart, 2021].
Figure 21: KLIMENTOVÁ, Julie. *Media words frequencies* [bar chart, 2021].

Figure 58: KLIMENTOVÁ, Julie. *The compromise model – an overview of frequent words* [table, 2021].

13 Appendices

Appendix 1: Corpus [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/FinalCorpus/metadata/songsMetadata.csv

Appendix 2: Media words table [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/FinalCorpus/mediaWords.csv

Appendix 3: Main corpus bigrams [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/TextAnalysis/Word%20Frequencies/csvs/main%20corpus/bigrams_filtered.csv

Appendix 4: French web corpus bigrams [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/TextAnalysis/Word%20Frequencies/csvs/french%20web/bigrams_filtered_wb.csv

Appendix 5: Media words frequencies – more ambiguous words [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/TextAnalysis/Word%20Frequencies/nwmFrequencies/newMediaWordsFrequencies.csv

Appendix 6: Media words frequencies – less ambiguous words [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/TextAnalysis/Word%20Frequencies/nwmFrequencies/newMediaWordsFrequencies_noSoft.csv

Appendix 7: Number of media words per song counts [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/TextAnalysis/Word%20Frequencies/csvs/media%20words%20subcorpus/count_songs.csv

Appendix 8: Media words subcorpus bigrams [CSV table]. Available at: https://github.com/julieklimentova/le_rap_francophone/blob/master/TextAnalysis/Word%20Frequencies/csvs/media%20words%20subcorpus/bigrams_subcorpus.csv