NORTHWESTERN
UNIVERSITY

Computer Science Department

Technical Report
Number: NU-CS-2024-09

April 2024

A Computational Analysis and Exploration of Linguistic Borrowings in French Rap Lyrics

Lucas Zurbuchen

Abstract

In France, linguistic borrowing in the French language can occur in several ways due to the frequent contact that French has with different linguistic groups in the country. However, people inside France view this borrowing with widely mixed perspectives. This thesis examines French rap lyrics -- one of the country's largest proponents of borrowed terms and slang -- and the usage of linguistic borrowings that appear in them. By doing this, new knowledge about French rap and its implications on the French language (and about linguistic borrowings in general) through computational analysis can be contributed.

Collecting over 8000 lyrics (extracted from Spotify and Genius) from 1991 to 2024 and over 700 words and phrases borrowed into the language, four variables about the borrowings (length, origin, semantics, and part of speech) can be analyzed to view both how they shape the number of times a borrowing is used in lyrics and how its usage can change over time, using regression models and temporal analysis respectively. The thesis finds that all four factors can contribute to a borrowing's popularity in lyrics under several circumstances, with subcategories of these factors experiencing both increases and decreases in popularity over time. It also finds that linguistic borrowing usage in French rap has been increasing over time. Further research will include examining the effect that factors outside the music and its words (such as a song's popularity) have on the usage trends of a linguistic borrowing as well as using machine learning methods to predict future trends of a linguistic borrowing's popularity.

Keywords

Linguistic Borrowing, France, Rap Lyrics, Wordlist and Lyric Collection

A Computational Analysis and Exploration of Linguistic Borrowings in French Rap Lyrics

Northwestern University

Department of Computer Science

An undergraduate senior thesis submitted in partial fulfillment of the honors requirements for the degree of BS in Computer Science at Northwestern University.

Lucas Zurbuchen

Advisor: Dr. Rob Voigt

April 2024

A Computational Analysis and Exploration of Linguistic Borrowings in French Rap Lyrics

Lucas Zurbuchen
Northwestern University / 633 Clark Street; Evanston, IL 60208
lucaszurbuchen2024@u.northwestern.edu

Abstract

In France, linguistic borrowing in the French language can occur in several ways due to the frequent contact that French has with different linguistic groups in the country. However, people inside France view this borrowing with widely mixed perspectives. This thesis examines French rap lyrics – one of the country's largest proponents of borrowed terms and slang – and the usage of linguistic borrowings that appear in them. By doing this, new knowledge about French rap and its implications on the French language (and about linguistic borrowings in general) through computational analysis can be contributed.

Collecting over 8000 lyrics (extracted from Spotify and Genius) from 1991 to 2024 and over 700 words and phrases borrowed into the language, four variables about the borrowings (length, origin, semantics, and part of speech) can be analyzed to view both how they shape the number of times a borrowing is used in lyrics and how its usage can change over time, using regression models and temporal analysis respectively. The thesis finds that all four factors can contribute to a borrowing's popularity in lyrics under several circumstances, with subcategories of these factors experiencing both increases and decreases in popularity over time. It also finds that linguistic borrowing usage in French rap has been increasing over time. Further research will include examining the effect that factors outside the music and its words (such as a song's popularity) have on the usage trends of a linguistic borrowing as well as using machine learning methods to predict future trends of a linguistic borrowing's popularity.

1 Introduction and Background

1.1 Introduction

Ever since its origin, the musical genre of rap has changed the languages it has touched. In America, its birthplace, it placed several African American Vernacular English (AAVE) terms like the adjective "woke" and the verb "ghost" into the lexicons of English speakers far beyond that linguistic group (Lewis, 2023). This linguistic borrowing has led to interesting debates about its morality, like whether it should be considered linguistic evolution or cultural appropriation. The popularity of rap music has allowed the genre to go far beyond the United States and into countries in Latin America, Europe, Africa, Asia and more. For these countries, the opportunity for linguistic borrowing to occur from the linguistic minorities that choose to partake in its rap community is present.

Linguistic borrowing in European rap can be especially intriguing. Generally, borrowings in Europe tend to come from three different sources: Anglicisms from the United States, linguistic minorities from countries affected by European colonialism, and European linguistic communities being in close proximity to each other. The Western European country of France is a country that is uniquely vulnerable to all three of these sources of borrowings. As one of history's largest colonizing nations, it lies across the ocean from the United States and borders eight other countries.

The usage of linguistic borrowings in the French language is a socially complex issue. Some people and organizations in France are especially known for resisting this change. One of the biggest symbols of French linguistic preservation is the *Académie Française* (AF), which is a forty-member organization having close ties to the French government (Estival and Pennycook, 2011). They try to act as a "guide" for French speakers, and they often resist the introduction of borrowings outside of French, like how they created the alternative term "ordinateur" for the English "computer" – even though other Romance languages like Spanish simply borrowed the English term (Estival and Pennycook, 2011).

Rap music in France poses challenges for these types of organizations. Borrowings from more

languages just than English, Spanish, or Arabic regularly occur. For example, linguist and public speaking teacher Julien Barret explains that many borrowings have also occurred from Romani (often inside prisons), which have been output through rap songs and artists (Rhrissi, 2021). He also explains that linguistic borrowings through French rap have become so widespread that students he works with sometimes don't know if they learned about a borrowed word in a rap song or from their neighborhood (Rhrissi, 2021).

French rap artists also helped influence the popularity of a new type of *argot*, or slang, called Verlan, which is composed of inverting syllables of French words to create new unusual-sounding ones ("Verlan" itself derives from French "l'envers", meaning "the inverse") (Hassa and Tekourafi, 2010). People are motivated to use Verlan due to the excitement of making a common word understandable to only a select few people and because it creates a sense of identity and power for those who can flip ordinary words to go against the status quo (Westphal, 2013). An example of this can be seen in an interview with old-school French rapper Zoxea, where he and his friend group changed the names one of the major places that they would meet in the city of Boulogne, Pont de Sèvres (Sèvres Bridge), to "Pont de Vreuss" – inverting the syllables of bridge's name and removing the last syllable of the inverted word to simplify it (Westphal, 2013). Nevertheless, the popularity of new words, whether Verlan or linguistic borrowings, in French rap is something peculiar to be studied since their usage can reflect the personal, social, and political outlook of the French rap community. In fact, as of 20 years ago, 92 percent of rappers in French rap were immigrants, many of whom used their music to highlight social problems and postcolonialism (Hassa and Tekourafi, 2010).

As a result, what would make a linguistic borrowing in French popular compared to others? It might be valuable to see which types of borrowings get propelled forward by the French rap community while also which types of borrowings can be held back by organizations like the AF for the sake of linguistic preservation. Additionally, analyzing not just if but when some types of borrowings were more popular than others could be equally useful. The thesis will explore both topics, asking what the largest determining factors of a linguistic borrowing in French rap are for determining a word's popularity both overall and over time.

1.2 Background

There exists current literature on sections of this topic. What this thesis hopes to do is bridge the gap between the computational literature on temporal word popularity, more social research on the causes and impact of linguistic borrowing in rap songs, and specific research of this topic done with France in mind.

Computationally, there is work being done with both identifying borrowed words and factors contributing to their relative popularity. Regarding lexical borrowing identification, still a major challenge in computational linguistics, much of the current methods revolve around pre-processing the text in such a way that one can computationally search for deviations in sound patterns (often with classifier-based machine learning methods). Examples include Tsvetkov et al. (2015) converting Swahili text to the International Phonetic Alphabet to find derived words from Arabic, or analyzing phonotactic patterns in the Siberian language of Sakha to uncover borrowings from Russian (Mæhlum and Ivanova, 2023). Lexical borrowing detection is quite reliant on data collection to verify the deviations from sound patterns, like how Miller and List (2023) needed a wordlist for every language when searching for Spanish borrowings in indigenous Central and South American languages. Therefore, examining linguistic borrowings from as many types of languages as possible using these above techniques can quickly become a difficult data collection and annotation task.

There are alternatives like computational slang detection that can sometimes expose commonly used linguistic borrowings (even though in the context of the thesis, it might also return non-linguistic borrowings like Verlan). Approaches to this have been relatively effective, like BiLSTM (Bidirectional Long Short-Term Memory) construction to identify English slang, though it was shown that identifying the presence of slang in a sentence is more accurate than specifying the word in a sentence that is slang (Pei et al., 2019).

The other computational research of interest centers on analyzing the popularity of lexical borrowings in languages. First, an important thing to consider is whether a word, regardless as its status of a lexical borrowing, will remain well-used in a language. Studies such as WordWars have been

written on this topic, describing a natural selection approach to two different words with the same meaning, finding that there are some important factors such as word length but that the more important changes are those to the word's morphology (Mohammad, 2020). The popularity of English slang words over time (as related to non-slang words) has been studied as well, but the main finding was that simply the word's status as a slang word was the largest determining factor of its popularity (Keidar et al., 2022). When it comes to the popularity of specific linguistic borrowings, much of the work outside of English is done on Anglicisms. In Spanish, for instance, there exist corpora of Anglicisms in Spanish newspaper headlines (Alvarez-Mellado, 2020), and studies of Anglicisms in Spanish tweets whose popularity and use seems to depend on social factors and contexts, though newspapers tend to use more Anglicisms than other social media users (Stewart et al., 2021). Studies about linguistic borrowings and their popularity also exist in French, but an interesting finding that Chesley and Baayen (2010) have is that the dispersion of a French term (a measure of a word's spread in text and not simply its quantity) is a better indicator of its foothold in French than frequency, though both are important factors. They also mentioned that languages other than English were less likely to be borrowed (Chesley and Baayen, 2010), but it would be interesting to analyze that in more detail in this thesis.

Socially, France isn't the only Western European country that has rap artists eager to spread social awareness about certain linguistic groups. For example, Arabic linguistic borrowings have a foothold in German rap, though much of these Arabic borrowings often have as much to do with Muslim religious identity as with the Arabic linguistic community itself (Hebblethwaite, 2018). It's debatable if this is equally true in French rap, because there are many Arabic lexical borrowings in French for secular terms, such as money ("flouze"), what's up ("wesh"), or party ("nouba"). In Spain, the rap group Dios Ke Te Crew is the only one that raps in Galician, frequently about local or global issues (Loureiro-Rodríguez, 2013). Depending on the locality of the message they want to send, they switch to Spanish or English instead of Galician (Loureiro-Rodríguez, 2013). Lastly, identity and anti-imperialism are a large topic of discourse in Portugal's Kriolu rap (a Portuguese-based creole from Cape Verde) (Pardue, 2012).

All these additions to the literature help demonstrate what the current progress, limitations, and gaps are in this thesis topic, and they serve as a great path of action for collecting and analyzing data. Moreover, they are a tool to evaluate what exactly what this thesis will contribute:

- Bridging the gap between related technical research, social research, and existing research on French language change through rap.
- Leveraging computation to generate corpora of rap songs and linguistic borrowings which can be used both in and beyond this thesis.
- Exploring trends in the overall usage and temporal usage of linguistic borrowings, providing direction on what future research could address.

2 Data Collection

2.1 Introduction

To obtain sufficient information to answer the question of interest, the data that was sought out was a set of temporal graphs, each showing the frequency of a borrowed word in French rap songs (relative to all words in the lyrics) over time. The main steps of this process were building a corpus of French rap songs, gathering a data sheet of linguistic borrowings in the French language, and building the temporal graphs.

2.2 Corpus Collection

Spotify's API[1] and Genius's API[2] were used in tandem to build a large set of French rap songs that could be analyzed for linguistic borrowings. There exist corpora on French rap lyrics like RapCor[3], but what it offers in annotations and detail it lacks in both number of songs and currency, as its last update was over a year ago. As a result, using Spotify and Genius meant that the most recent songs and lyrics could be collected (as well as many older ones) to analyze. Searching for songs in the Spotify API started with searching for a select number of songs within the genre of French rap. Because Spotify's API has a limit of songs one can receive in a single API request, a recursive search was used where the artists returned

[1] https://developer.spotify.com/documentation/web-api

[2] https://docs.genius.com/#/getting-started-h1

[3] https://is.muni.cz/do/phil/Pracoviste/URJL/rapcor/index_en.html

Wesh, cette bitch veut mon corps, pourtant je sue comme un sumo
"Yo [Arabic borrowing], this bitch [English borrowing] wants my body, yet I'm sweating like a sumo [Japanese borrowing]"
From "Intro (Introduced by Caspi)" by JMK$ & Beamer

À 14 ans dans la tess, igo c'était gore
"In the streets [Verlan] at 14 years old, dude [Spanish borrowing] that was gory"
From "Des Nomes" by Fresh laDouille

C'est la crise au mic, fait la bise au mac
"It's a crisis at the mic [English borrowing], give the pimp [Argot] a kiss"
From "Oh mama oh" by Le Classico Organisé

Figure 1: Three examples of lexical borrowings and other *argot* from the collected corpus.

1. Referring to a certain identity of people
2. Person – occupational
3. Food/drinks/drugs
4. Other inanimate material/product
5. Places
6. Events/materials related to conflict/crime
7. Sex/sexual connotations
8. Common exclamations/expressions
9. Common usage/grammatical function/other
10. Related to music/other arts

Figure 2: List of 10 semantic categories borrowed terms were assigned to.

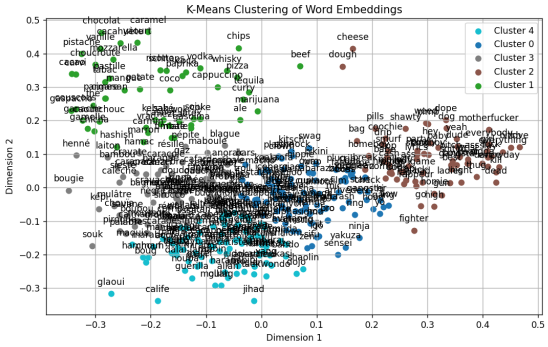

Figure 3: A sample of collected borrowings embedded with Facebook MUSE embeddings and clustered with K-Means (with 52 words not found).

from the first request were queried on (still with the constraint of the French rap genre) until a certain depth in search was reached. Something important to consider was that Spotify's API tended to oversample newer songs over older ones (given the increase of the popularity both of French rap since the late 20th century and of Spotify in general), but this was attempted to be mitigated by trying to sample as many songs as possible with this approach. Nonetheless, for each song, the song name, artist(s), and release date were sampled. For each song received from Spotify's API, the Genius API was queried to find its lyrics, filtering out noise accordingly. After the process was completed, a total of 8222 French rap lyrics from 1991 to 2024 were ready to analyze.

Examples of borrowings were prevalent all over the corpus, with three examples in Figure 1 to provide some context. Even though the research question remains only about linguistic borrowings, it's useful to examine that these aren't the only linguistic innovations in French rap – Verlan and other types of *argot* exist as well.

2.3 Linguistic Borrowing Collection

The next step into collecting the desired data was finding a selection of words that were borrowed into the French language. This collection was done manually (for the sake of verifying the words' etymologies and appearance in lyrics), though rough automatic techniques like language detection libraries or looking for words that weren't in French dictionaries acted as helpful tools. The most useful source was Wiktionary[4], maintaining a list of words borrowed into the French language from others. For each word, the origin of the borrowing (a singular value, also known as the donor language), its part(s) of speech, and its semantic meaning(s) were all recorded. Wiktionary helped with determining both the word's origin and part of speech. After some deliberation, the words' semantics were manually collected with a hand-made classification system (see Figure 2).

Substantial time was spent on finding a set of word embeddings that one could cluster to calculate semantic categories, but it was deduced that several word embedding methods heavily biased words based on their origin. As an example, Figures 3 and 4 show K-Means clustering assignments (projected onto 2 dimensions) for multilingual Facebook MUSE embeddings[5] (aligned in the same space) and embeddings generated from Urban Dictionary (Wilson et al., 2020) respectively. Figure 3 has categories on the top-left related to food and drugs, but many of the terms on the right derive from AAVE and are placed there regardless of their

[4] https://en.wiktionary.org/wiki/Category:French_terms_derived_from_other_languages

[5] https://github.com/facebookresearch/MUSE

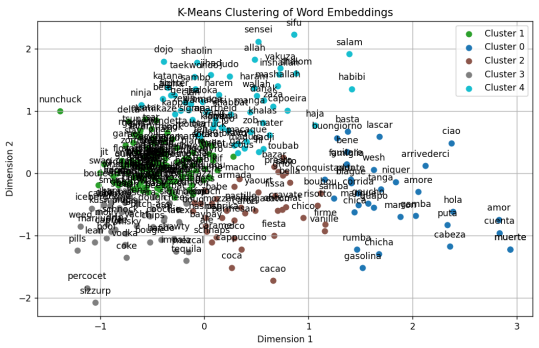

Figure 4: A sample of collected borrowings embedded with Urban Dictionary embeddings and clustered with K-Means (with 130 words not found).

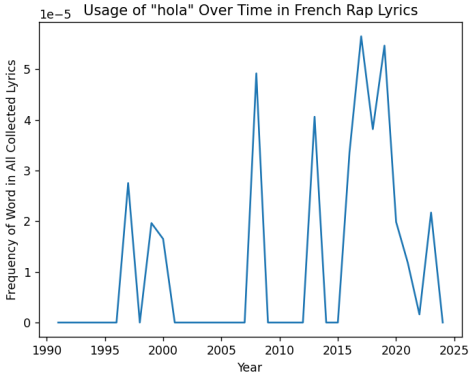

Figure 5: A plot showing the frequency at which the Spanish borrowing "hola" appears in the collected lyrics over time.

meaning. Likewise, Figure 4 shows borrowings derived from Arabic, Asian languages, and Italian clustered in their own areas. Luckily, this bias could still be taken advantage of while searching for new borrowed words and inflections because of its tendency to have higher cosine similarities for words of similar origins.

A total of 741 instances of linguistic borrowing in French were collected, which could be used along with the generated corpus of French rap lyrics to obtain the graphs.

2.4 Graph Generation

Generating a graph of the frequency of a word in French over time was essentially an instance of string matching, going through the entire corpus to find all instances of the word, dividing by the number of words in the text. If there was a string match in a line of text, it was important to check that the borrowing was its own individual token (or set of tokens). Additionally, to account for the variations in spelling that occur when transcribing lyrics, all popular inflections of a word were also recorded as a word. The release date of each song was recorded with a granularity of a year to make it easier to record broader temporal trends. At the end of the process, one gathers a graph such as Figure 5 for all the collected instances of linguistic borrowing. Furthermore, in the data collection process, there were 4563577 total tokens in the lyrics, with 98740 borrowings counted. Table 1 shows basic statistics about the distribution of number of occurrences of recorded borrowings. Lastly, all code and data are available for open-source use on the GitHub repository for this project[6].

[6] https://github.com/ljz112/CLResearch

Table 1: Basic statistics of the number of times borrowings appear in the collected corpus (STD is Standard Deviation and IQR is Interquartile Range).

Maximum	Minimum	Mean	Median	STD	IQR
10362	0	133.3	28	525.3	71

3 Data Analysis

3.1 Introduction and Methods

Analysis of the data was completed in two parts: finding correlations between the overall usage of a borrowed word and four independent variables and searching for temporal trends in the usage of different types of linguistic borrowings in French rap.

To do the first, an RLM (Robust Linear Model) regression model was run on the collected data. Using this model specifically instead of standard linear regression was beneficial mainly because of the data's potential sensitivity to outliers (see Figure 5 for an example of noise that often occurs in the data), which can be difficult to prevent. For each recorded linguistic borrowing, the four parameters that served as independent variables for the regression were its length (in characters), origin of borrowing, semantic cluster(s), and part(s) of speech, with all except the first variable being dummy-encoded due to their categorical nature. The output was the total number of times that the word appeared in the collected lyrics. To reduce the number of potentially noisy dimensions in the model, parameters corresponding to language groups with an insufficient number of samples (less than 5 in this case) were removed before running the model. Lastly, to be a statistically significant

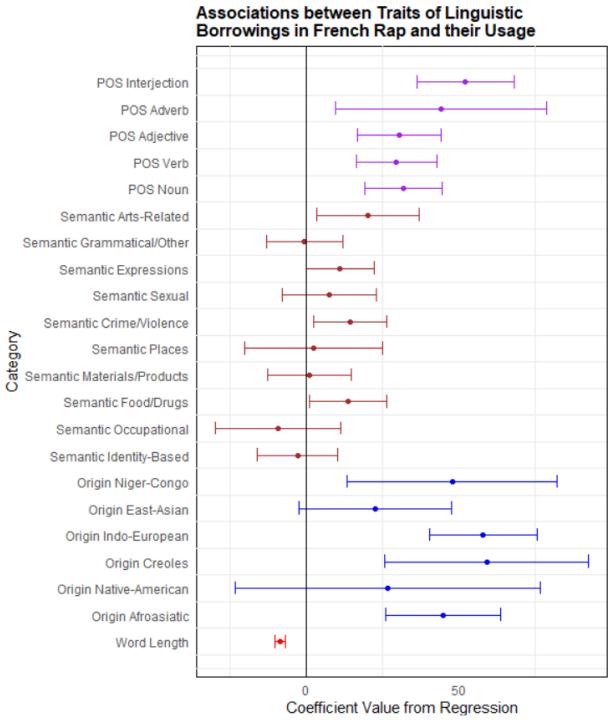

Figure 6: A plot of all categories and their coefficients from the RLM regression run on all collected borrowings (with a 95% confidence interval).

correlation, the p-value for a category (with the null hypothesis that the category's coefficient in the regression was 0) had to be less than $\alpha = 0.05$.

The second method examined trends among the graphs of a given category of linguistic borrowings over time, searching for plausible lines of best fit. Specifically, the main dependent variable of interest was the proportion of borrowings of the said category relative to all borrowings from the collected data. Both a Linear Model and LOESS (LOcally Estimated Scatterplot Smoothing) were used to find fits (depending on the linearity of the observed relationship).

3.2 Regression Analysis

Performing the regressions returned several interesting correlations (see Figure 6 for a forest plot on the regression performed on all the data). Perhaps the clearest one was with word length, with there being a negative relationship between the length of the word and the number of times it appears in French rap songs. This is intuitive since people are inherently more likely to use shorter words than longer ones.

On the broadest level of origin, the four language families with statistically significant correlations (all positive) were Afroasiatic, Creoles, Indo-European, and Niger-Congo – none of which were significantly higher than each other. Out of these four language groups, Indo-European languages were the only ones to have statistically significant results on borrowed languages when the same RLM regression was run only on Indo-European borrowings. Germanic and Romance languages both had positive correlations to word frequency (see Table 2, showing the output coefficient, its standard error, and a 95% confidence interval for these groups), making sense as these contain the most popular borrowed languages, like English in Germanic and Spanish in Romance. Inside the regression run on only Germanic languages, English was the only language to have a significant (positive) correlation, emphasizing the popularity of Anglicisms in French rap lyrics (see Table 3). In a regression run only on English borrowings, the usage of words from AAVE versus those not from AAVE were analyzed, and it was found that both had significant positive relationships (see Table 4). Inside the regression run on only Romance borrowings, both Spanish and Portuguese had statistically significant correlations, but Spanish was positive while Portuguese was negative, emphasizing the popularity

Table 2: Statistically significant Indo-European language groups from regression run on Indo-European linguistic borrowings.

Borrowing Origin	Coefficient in Regression	Standard Error	95% CI
Romance	18.2	5.47	(7.48 , 28.9)
Germanic	26.8	4.94	(17.2 , 36.5)

Table 3: Statistically significant Germanic languages from regression run on Germanic linguistic borrowings.

Borrowing Origin	Coefficient in Regression	Standard Error	95% CI
English	20.2	4.14	(12.1 , 28.3)

Table 4: Statistically significant results from regression run to compare AAVE English borrowings and non-AAVE English borrowings in French rap

Borrowing Origin	Coefficient in Regression	Standard Error	95% CI
AAVE	24.6	6.83	(11.2 , 37.9)
Non-AAVE	18.6	4.46	(9.84 , 27.3)

Table 5: Statistically significant Romance languages from regression run on Romance linguistic borrowings.

Borrowing Origin	Coefficient in Regression	Standard Error	95% CI
Spanish	10.9	5.34	(0.414 , 21.3)
Portuguese	-33.9	15.8	(-64.9 , -2.87)

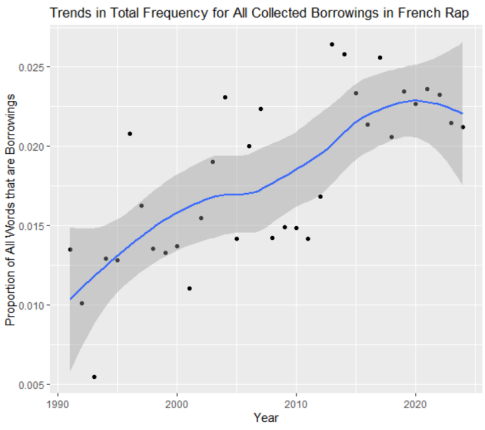

Figure 7: Fitted curve of the frequency of every collected borrowing relative to all words in lyrics over time.

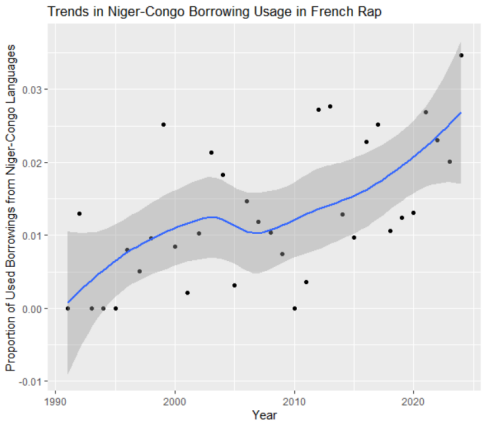

Figure 8: Fitted curve of the proportion of collected borrowings from Niger-Congo languages over time.

Spanish has over other Romance languages as potential borrowings (see Table 5).

On the regression run on all borrowings (in Figure 6), three semantic categories turned out to have significant positive correlations over others: terms related to food and drugs, crime and violence, and music and the arts. This is interesting to see because it suggests that all three have been quite common topics to mention in French rap songs. Furthermore, all five examined parts of speech had significantly positive correlations, though all the confidence intervals overlap with each other, so the prevalence of borrowings of one part of speech over another is inconclusive with this regression model.

3.3 Temporal Analysis

Quite a few insights on a linguistic borrowing's origin, semantic categories, and parts of speech in French rap could be found through examining their respective plots.

Before addressing specific categories of borrowings, examining the proportion of all collected linguistic borrowings relative to every word in the lyrics provided some interesting findings (see Figure 7). It demonstrates that the overall usage of these borrowings has been increasing over time – essentially doubling since the 1990s to 2 percent of all words in the lyrics. This also acts as a guide for the rest of the data analysis since it suggests that it's more productive to analyze the usage of a certain category relative to the usage of all collected borrowings over time (as a proportion) because it prevents any false interpretation of results that is simply this overall trend.

Examining languages of origin this way, Niger-Congo and Afroasiatic languages exhibit interesting trends. Niger-Congo languages were shown to increase the most rapidly, getting gradually more popular up to this day (see Figure 8). A possible explanation for this could that Sub-Saharan Africa regions are front-runners in population growth (Uni, 2019), which includes regions that France has colonized like the Democratic Republic of the Congo. On the other hand, borrowings from Afroasiatic languages have stayed both substantial (at around 10% of total borrowings) and consistent over time (see Figure 9), suggesting that Afroasiatic languages, many of which are from Arabic, have been a staple in linguistic borrowing usage in French rap since the beginning. Inside Indo-European languages, Romance languages have decreased in popularity (see Figure 10), with the largest decrease happening in the 1990s, possibly because some of the collected borrowings were already engrained in the French language before rap gained traction (like "armada" from Spanish or "paparazzi" from Italian).

With semantic categories, the largest finding was that arts-related terms have had a linear decrease (see Figure 11) since the start while common expressions have had a linear increase (see Figure 12). This potentially indicates a major shift in song topics or even style since then.

Something else that could indicate a stylistic change is in the interesting finding with part of speech over time, which is that the proportion of borrowings that are interjections has been increasing rapidly after 2010 (see Figure 13) while the inverse has been happening to nouns (see Figure 14). This could be because many interjections in rap

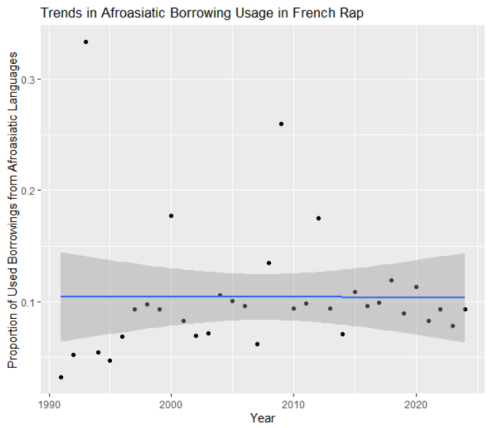

Figure 9: Fitted line of the proportion of collected borrowings from Afroasiatic languages over time.

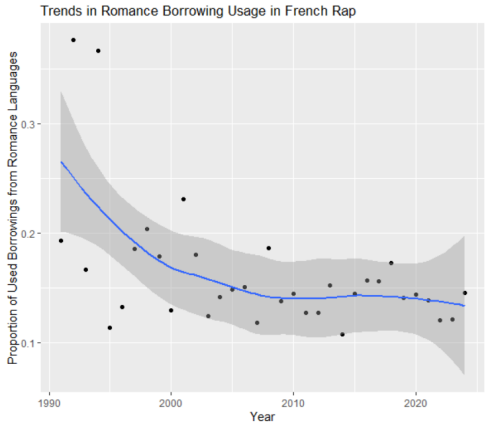

Figure 10: Fitted curve of the proportion of collected borrowings from Romance languages over time.

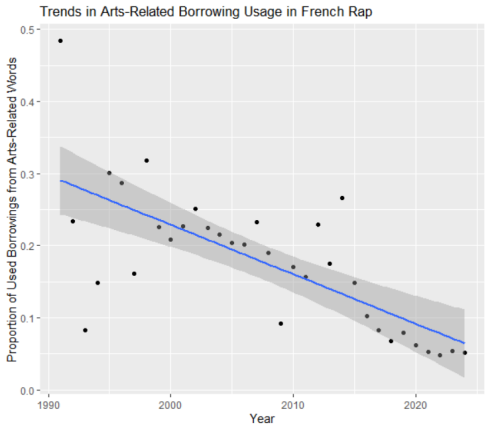

Figure 11: Fitted line of the proportion of collected borrowings that are arts-related over time.

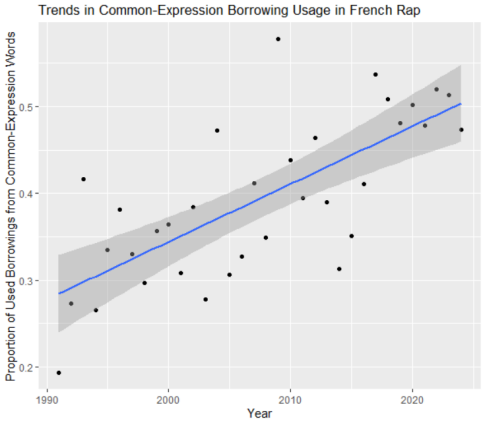

Figure 12: Fitted line of the proportion of collected borrowings that are common expressions over time.

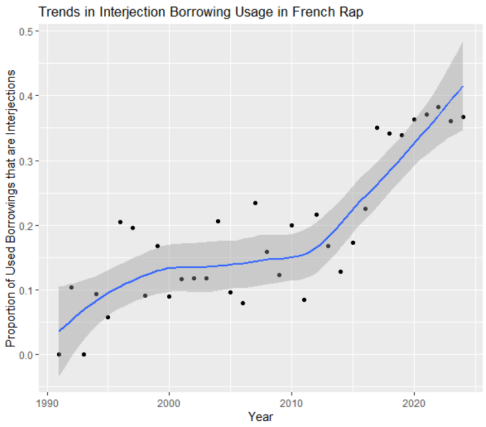

Figure 13: Fitted line of the proportion of collected borrowings that are interjections over time.

songs, at least in the United States, act as ad-libs, whose role has become steadily more important since 2010 for new subgenres of rap like mumble rap (Waugh, 2020), so it wouldn't be surprising if this stylistic trend moved over to France as well.

4 Conclusion

4.1 Discussion

This thesis finds that a linguistic borrowing's length is a likely determining factor in its popularity while the origin, semantic category, and part of speech of a borrowing can all influence its overall usage under certain circumstances. Furthermore, certain categories of borrowings have experienced both ups and downs in their popularity over the years relative to each other hinting at possible trends involving demographics, musical style, or more. Most importantly, the number of borrowings has been increasing over time, showing the influence that linguistic borrowings are having on the French

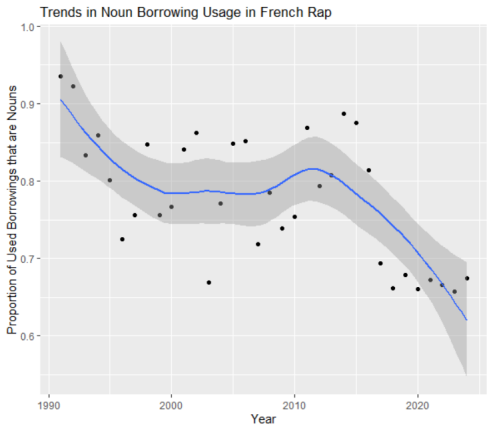

Figure 14: Fitted line of the proportion of collected borrowings that are nouns over time.

language.

These results were deduced by collecting over 8000 songs with Spotify and Genius and analyzing the trends of the usage of 700 words over time with regression models and temporal analysis. These findings are compelling for several reasons. Firstly, it explains some of the dynamic linguistic environment that occurs in French rap, exposing both the successes of linguistic minorities and organizations like the AF to influence the French language as it is. It also studies linguistic borrowing with a large breadth and depth on the donor languages that make their way into French rap lyrics. Lastly, it provides a small window into the large role that rap music can serve to change the languages it encounters, whether in the United States, France, Korea, or anywhere else.

The ethics of this type of linguistic borrowing are under constant debate. On one hand, it allows linguistic minorities to assert their own identity and pridefully distinguish themselves from others, often in environments where doing so can be looked down upon. Languages have also constantly changed using linguistic borrowing, like how English and French share around 40% of words due to the Norman invasion of England almost a millennium ago (Pervan, 2021). However, the effect that linguistic borrowing can have on fostering a linguistic group's sense of identity can diminish when people outside of it start using those borrowed words while ignorant of the words' origins. When the origin is known, it can sometimes even run the risk of cultural appropriation when people outside of the linguistic group use it. Even though this can be a difficult debate, this data was collected and analyzed with the hopes of possibly exploring this phenomenon in France. A possible next step could be to gather details about the origins of artists and analyze their usage of borrowed words to see how it relates to artists of that specific linguistic group. Nevertheless, this thesis has potential ramifications like these and others that will be mentioned later that could be conducted as next steps.

4.2 Limitations

An inherent number of limitations exist within this research. First, the manual collection of words limited the sample of words that could be analyzed. However, computational lexical borrowing detection methods weren't in a place where it could be used without the presence of well collected, annotated, and transliterated data on several languages and dialects. To add on, determining the singular origin of a borrowing was sometimes difficult due to complex etymologies, so accounting for that in further research may be helpful for accuracy and generating more insightful findings. The data collection was done by only one person, making it easier for biases to enter the labeling (even with efforts to have consistent and objective labelling protocols). They only speak French at an intermediate level, so they don't understand the context of words and phrases (without external aids) like a native speaker can. Moreover, they were quicker with identifying Anglicisms and newer words than other borrowings since English is their first language and since they were born after 2000.

Because of rap's growth in France, less songs were available for older dates than for newer ones. This led to results where the data was noisier in the earlier years than the last (preventing the granularity of temporal graphs from being less than a year). Accounting for this error present in earlier years during data analysis may be beneficial to address in further research. Another variable causing this is Spotify's and Genius's API, which are imperfect measures for obtaining as many French rap songs as possible. The collection of a song's lyrics in this thesis occurs under the assumption that a song is recorded both on Spotify and Genius, which removes any songs that aren't on the streaming platform (likely disproportionately affecting older songs) or that got taken away from it.

4.3 Future Improvements

To build on this research in the future, many directions could be taken. One possible direction would be to analyze word usage over time relative to ex-

ternal information about the French rap songs in which they are used. For example, it would be interesting to deduce if the popularity of a song affects a word's usage over time, like if a popular song with a borrowed word triggers its increased popularity. Examining if French linguistic borrowings act similarly in rap songs in other European francophone countries, like Switzerland or Belgium, could potentially be valuable as well. Additionally, one could look at the external feature of how well-integrated a borrowing is in its home language compared to French. Though this would require a larger data collection process, this could provide insight on if there exist borrowings used more in French than its actual language, or vice versa.

Another direction in which this research could go is in the direction of machine learning, like using a neural network or transformer-based model to predict the future popularity trends of a borrowed word in French rap songs over time given some current information. There is some research working on similar things, like predicting the popularity of an online petition given the headlining text and image (Kitayama et al., 2020), or the popularity of newspaper headlines, with interesting insights provided about the types of models that can be used and what types of headlines gather more popularity than others (Lamprinidis et al., 2018). Inside music, work has been done to evaluate the sentiment of music with Large Language Models using social media comments (Donnelly and Beery, 2022), which could be an interesting gateway for evaluating relationships between variables like sentiments and popularity using machine learning techniques. Furthermore, being able to predict a graph of a borrowing's usage over time could provide insight into both short- and long-term trends that could occur with a borrowing's usage and not just a singular quantity. The implications of this work could be many, like researchers having a heightened understanding of which borrowed words thrive in a linguistic environment. However, it runs the risk of being taken advantage of by companies wanting to use it to capitalize on these trends and by governments who would like to foster or prevent changes in their citizens' speaking patterns. Nonetheless, if used properly, it could be a very intriguing direction in which to go.

5 Acknowledgements

First, I would like to thank Professor Voigt for all his guidance as my thesis advisor. Through our meetings, I have gained understanding of what it means to do research, learning about its difficulties and rewards. For someone who decided to start studying computational linguistics in his senior year, I thank Professor Voigt for his faith that someone could learn the basics and generate a thesis in one academic year.

I'd also like to thank Professor Worsley for welcoming me into the tiilt lab, getting me into the BLINC project which sparked much of my interest in computational linguistics, and for helping me revise my thesis. Professor Worsley cultivates such a positive environment in his lab that he is impossible not to thank.

As a musician, I'd like to give my thanks to Professor Raciti, my bass teacher, for being there both inside and outside of my bass playing -– understanding what it means when I tell him I'm having a "hell week" during midterms. I'm grateful for my friends and family who supported me through this process. Though it's been a while, I'm also indebted to my high school French teacher, Monsieur Click, who gave me invaluable background on both the French language and society.

Lastly, I'd like to acknowledge Northwestern's Computer Science department and the McCormick School of Engineering for giving me the opportunity to write this thesis.

References

2019. 9.7 billion on earth by 2050, but growth rate slowing, says new un population report.

Elena Alvarez-Mellado. 2020. An annotated corpus of emerging anglicisms in Spanish newspaper headlines. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 1–8, Marseille, France. European Language Resources Association.

Paula Chesley and R. Harald Baayen. 2010. Predicting new words from newer words: Lexical borrowings in french. *Linguistics*, 48(6).

Patrick Donnelly and Aidan Beery. 2022. Evaluating large-language models for dimensional music emotion prediction from social media discourse. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 242–250, Trento, Italy. Association for Computational Linguistics.

Dominique Estival and Alastair Pennycook. 2011. L'académie française and anglophone language ideologies. *Language Policy*, 10(4):325–341.

Samira Hassa and Marina Tekourafi. 2010. *Kiff My Zikmu: Symbolic Dimensions of Arabic, English and Verlan in French Rap Texts*, page 44–66. Continuum.

Benjamin Hebblethwaite. 2018. Arabic lexical borrowings in german rap lyrics: Religious, standard and slang lexical semantic fields. *Delos*, 31:113–125.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Kotaro Kitayama, Shivashankar Subramanian, and Timothy Baldwin. 2020. Popularity prediction of online petitions using a multimodal DeepRegression model. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 110–114, Virtual Workshop. Australasian Language Technology Association.

Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy. 2018. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 659–664, Brussels, Belgium. Association for Computational Linguistics.

Miles Marshall Lewis. 2023. How hip-hop changed the english language forever.

Verónica Loureiro-Rodríguez. 2013. "if we only speak our language by the fireside, it won't survive": The cultural and linguistic indigenization of hip hop in galicia. *Popular Music and Society*, 36(5):659–676.

Petter Mæhlum and Sardana Ivanova. 2023. Phonotactics as an aid in low resource loan word detection and morphological analysis in sakha. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 111–120, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

John E. Miller and Johann-Mattis List. 2023. Detecting lexical borrowings from dominant languages in multilingual wordlists. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2599–2605, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif M. Mohammad. 2020. WordWars: A dataset to examine the natural selection of words. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3087–3095, Marseille, France. European Language Resources Association.

Derek Pardue. 2012. Cape verdean creole and the politics of scene-making in lisbon, portugal. *Journal of Linguistic Anthropology*, 22(2).

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

Zoran Pervan. 2021. Social elements in the norman french influence on english. *Hum*, (24):163–179.

Lina Rhrissi. 2021. D'aya nakamura à pnl: Comment les artistes musicaux·les transforment la langue française.

Ian Stewart, Diyi Yang, and Jacob Eisenstein. 2021. Tuiteamos o pongamos un tuit? investigating the social constraints of loanword integration in Spanish social media. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 286–297, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-based models of lexical borrowing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 598–608, Denver, Colorado. Association for Computational Linguistics.

Michael Waugh. 2020. "every time i dress myself, it go motherfuckin" viral': Post-verbal flows and memetic hype in young thug's mumble rap: Popular music.

Kelsey Quinn Westphal. 2013. Teuf love: Verlan in french rap and beyond.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4764–4773, Marseille, France. European Language Resources Association.