"Balance Ton Rap": Quantifying gender stereotypes in French rap using word embeddings

Matteo Larrode

Abstract

This study quantifies gender bias in French rap lyrics using word embedding association tests (WEAT). By training static embeddings on a corpus of 8,208 French rap songs and comparing them with off-the-shelf embeddings trained on general French text, I examine differences in stereotypical gender associations across four conceptual categories. Results reveal significant gender bias in French rap lyrics, particularly in intelligence-appearance associations, where intelligence-related terms are more strongly associated with male attributes and appearance-related terms with female attributes. Interestingly, general French text exhibits more pervasive gender bias across all examined categories, with web text showing significant bias in all four dimensions. Although limited by corpus size differences, these findings provide empirical evidence of gender stereotypes in French rap while challenging simplistic narratives that single out rap music as uniquely problematic. The study contributes to our understanding of how cultural products reflect and potentially reinforce gender biases, with implications for the development of more equitable representation in music, and the development of embedding debiasing methods.

1 Introduction

In her 2019 feminist anthem *Balance ton quoi*, which inspired the title of this paper, Belgian singer Angèle sings "J'ai vu que le rap est à la mode, et qu'il marche mieux quand il est sale" (*I've seen that rap is trending, and that it works better when it's dirty*). As she points out, while rap music has been playing an major role in the empowerment of underrepresented communities (Lewis, 2024; Pégram, 2019), this male dominated music genre has also often been criticised for its misogynistic lyrics (Brown & Turner, 2020; Cobb & Boettcher III, 2007). Conversely, these claims often lack empirical evidence, and are sometimes instrumentalised in a form of 'moral panic' – a socio-legal overreaction to rap as a seemingly new threat to society – observed by critical race theory scholars (Khan, 2022).

NLP techniques have shown great promises in identifying discriminatory language and bias in text corpora (Betti et al., 2023). Word embeddings, which represent words in a high-dimensional vector space (Bengio et al., 2000), have been shown to capture linguistic

biases from text they have been trained on (Caliskan et al., 2017). Basing their work on this observation, scholars have demonstrated the efficiency of leveraging word embeddings to measure bias and stereotypes towards women and ethnic minorities in text (Bolukbasi et al., 2016; Chaloner & Maldonado, 2019; Garg et al., 2018). In keeping with this literature, the present study quantifies gender bias in French rap lyrics for the first time. Specifically, I consider the main research questions:

- **RQ 1.** Is there evidence of gender bias in French rap lyrics?
- **RQ 2.** What are the differences in the extent and/or nature of stereotypical gender associations between French rap lyrics and general text?

After providing an assessment of the presence and implications of sexism in music, I describe related studies using word embeddings to assess gender bias in text and lyrics corpora. Then, using embeddings trained on a corpus consisting of lyrics of more than 8,000 French rap songs (Zurbuchen & Voigt, 2024), and importing off-the-shelf embeddings trained on large corpora of general French text, I employ association tests to detect gender language bias in both the rap lyrics corpus, and general French text (Betti et al., 2023; Caliskan et al., 2017; Charlesworth et al., 2021). I show the significant presence of gender-biased word associations in both corpora, with subtle differences in the stereotype categories in which bias reveals itself.

The implications of this research are two-fold. First, I highlight the importance of evidence-based assessments of gender bias in rap music, contrasting with stereotypical claims in the context of a pushback against the genre, which hurt fights against both racism and sexism (Khan, 2022; Reyna et al., 2009). Second, with NLP models trained on increasingly large and diverse corpora, the overwhelming evidence of gender and ethnic biases in text, including in song lyrics, (Betti et al., 2023; Casanovas-Buliart et al., 2024; Hansen, 2022) ought to be a source of concern.

2 Literature review

2.1 Sexism in music and rap, and on the importance of nuance

Music is among society's most influential cultural products, reflecting and shaping collective attitudes (Bennett, 2001; Casanovas-Buliart et al., 2024). Song lyrics function as both mirror and moulder of cultural narratives (Davis, 1985), making them ideal for examining how gender biases manifest in culture. From feminist and postmodern perspectives, music lyrics serve as archives where language reflects power dynamics (van Maanen, 1995). They embody cultural conditions that either reinforce positive identity formation or perpetuate harmful stereotypes, particularly against women and marginalized communities[1] (Dhoest

[1]Throughout this study, the term 'sexism' specifically addresses prejudice against women.

et al., 2015; Eze, 2020; van Bohemen et al., 2017). Sexism in this context manifests as patterns subordinating women through lyrical content and visual representation, operating through intersecting axes of gender, race, ethnicity, and class (Casanovas-Buliart et al., 2024).

Despite its industry dominance, both in the US and in France, it faces intense scrutiny, with feminist critiques often focusing on perceived misogyny (Griffin et al., 2024; Khan, 2022). Studies suggest misogynistic rap can activate stereotype priming (Cobb & Boettcher III, 2007), while content analyses reveal unbalanced gender dynamics in discussions of topics like sexual health (Griffin & Fournet, 2020). However, critical approaches to rap require nuance. Khan (2022) argues that the disproportionate scrutiny of rap reflects racial bias in cultural evaluation – a moral panic (Garland, 2008) targeting Black cultural expression. This is particularly evident in criticism of female rap artists creating sexually explicit content, despite research showing their music creates empowering spaces for marginalized communities (Griffin et al., 2024; Keyes, 2000), and support listeners' development of sexual agency (Carney et al., 2016).

French rap presents a particularly interesting yet understudied context for analysing gender representation. While French rap has gained tremendous cultural significance, analyses of gender bias in its content remain scarce, and qualitative in its nature (Pégram, 2019). The present study is the first computational analysis of a corpus of French rap lyrics.

2.2 Related works: analysis of gender stereotypes using NLP

Natural Language Processing (NLP) techniques have demonstrated great potential for identifying discriminatory language and biases in large text corpora. The analysis of gender stereotypes and sexism expressed in language with NLP methods has been a growing area of research in recent years (Stanczak & Augenstein, 2021). Word embeddings – which encode words in a high-dimensional space based on their co-occurrence patterns, positioning semantically related words near each other (Bengio et al., 2000; Goldberg, 2017) – form the foundation of many contemporary approaches to measuring linguistic bias. Initially developed for improving NLP task performance, researchers discovered these embeddings also capture patterns of semantic evolution (Hamilton et al., 2016), as well as cultural associations and biases present in their training data.

Inspired by this observation, Bolukbasi et al. (2016) pioneered the study of gender bias in word embeddings. After establishing a gender direction defined by the subspace containing gendered word pairs like 'she'/'he' and 'woman'/'man', they projected words that should ideally be gender-neutral (such as occupation terms) onto this gender subspace. Their analysis revealed that word embeddings trained on Google News articles contained concerning correlations between gender and various occupations.

Building on this foundation, Caliskan et al. (2017) developed the Word Embedding

Association Test (WEAT), adapting the Implicit Association Test (IAT) from Psychology (Greenwald et al., 1998), to measure gender bias in word embeddings. This technique, exploited in this study (more details in Sect. 3.4), was later successfully used by Garg et al. (2018), and refined by Charlesworth et al. (2021) and Bianchi et al. (2021), responsible for the creation of the Single-Category WEAT (SC-WEAT) and Sliced Word Embedding Association Test (SWEAT) respectively.

The application of word embedding techniques to music lyrics represents a relatively recent development. Barman et al. (2019) examined lyrics spanning five decades, finding that biases in lyrics correspond to human biases more broadly. Betti et al. (2023), combined multiple bias measures to examine gender bias in English songs from 1960-2010. However, French rap lyrics remain unexplored through this computational lens.

3 Methods

3.1 Dataset

The corpus derives from Zurbuchen and Voigt (2024)'s dataset of French rap lyrics, collected via Spotify and Genius APIs (1991-2024). After removing duplicates, invalid lyrics, and non-French songs using the Lingua[2] language detection library, the final dataset comprises 8,208 songs, comparable to Zurbuchen and Voigt's 8,222.

The preprocessing pipeline was designed to preserve distinctive linguistic features of French rap (including borrowings, slang, *verlan*) while creating a dataset suitable for embedding training. I removed structural annotations in square brackets (e.g., [Intro], [Chorus]), and used Gensim's[3] `simple_preprocess` for tokenization, retaining accent marks to maintain linguistic nuances. This preprocessing step included the removal of punctuation and conversion of all text to lowercase. I avoided lemmatisation as it would nullify the ability to perform analogy tasks, and standard French lemmatisers would likely eliminate many distinctive linguistic properties of French rap, including *verlan* and slang variations. This cleaning process resulted in a corpus containing more than 3.8 million tokens, and a vocabulary size of 26,149.

3.2 Word embeddings choice and training

While attention-based pretrained models like FlaubBERT (Le et al., 2020), CamemBERT (Martin et al., 2020), and BARThez (Kamal Eddine et al., 2021) have demonstrated superior performance on many natural language understanding tasks, static word embeddings were chosen over their contextual counterparts. This choice was motivated by their computational efficiency, and ubiquity in previous studies utilising the word association hypothesis

[2] https://github.com/pemistahl/lingua-py
[3] https://radimrehurek.com/gensim/

testing strategy laid out in subsection 3.4 (Caliskan et al., 2017; Chaloner & Maldonado, 2019; Charlesworth et al., 2021). To improve comparability with general French text embeddings, I employed Word2Vec (Mikolov et al., 2013) to generate word embeddings from the French rap corpus. For robustness, I report results for four distinct configurations using both CBOW and Skip-gram methods at 100 and 200 dimensions, with a minimum word frequency threshold of 5 and context window size of 5, based on results from the empirical validation process described in subsection 3.3.

Comparing French rap to general text entailed a careful consideration during hyperparameter optimisation, to maximise comparability between my trained embeddings and available off-the-shelf embeddings, while accommodating corpus size differences. Doing so would allow to attribute observed gender bias discrepancies to content of the corpora rather than to training contexts more validly. For comparative analysis, I selected French word embeddings developed by Fauconnier (2015). These include embeddings trained on FrWaC, a 1.6 billion word corpus constructed from web content within the .fr domain (Baroni et al., 2009), and a French Wikipedia dump. Two limitations affect the comparative analysis: corpus size differences (affecting embedding quality) and temporal disparity (Fauconnier's embeddings only include data to 2015, while the rap corpus extends to 2024). Access to more recent embeddings by Abdine et al. (2022) was requested but not granted, and limiting the rap corpus to pre-2015 would have further reduced an already modest corpus size.

Table 1 gives the details of each word embeddings used in this study.

Embeddings	Vocab	Tokens	Method	Corpus	Dimension	Min. word freq.
`frRap_100_cbow`	26K	4M	CBOW	French rap lyrics	100	5
`frRap_200_cbow`	26K	4M	CBOW	French rap lyrics	200	5
`frRap_100_skip`	26K	4M	Skip-gram	French rap lyrics	100	5
`frRap_200_skip`	26K	4M	Skip-gram	French rap lyrics	200	5
`frWac_200_cbow` (Fauconnier, 2015)	155K	12B	CBOW	FrWac	200	100
`frWac_200_skip` (Fauconnier, 2015)	155K	12B	Skip-gram	FrWac	200	100
`frWiki_1000_cbow` (Fauconnier, 2015)	66K	2B	CBOW	Wikipedia dump	1000	100
`frWiki_1000_skip` (Fauconnier, 2015)	66K	2B	Skip-gram	Wikipedia dump	1000	100

Table 1. Summary of the models used in this study.

3.3 Word embeddings validation

A three-part intrinsic validation framework was implemented to evaluate embedding quality. First, k-means clustering, with t-SNE dimensionality reduction for the visualisation, examined whether semantically related terms formed coherent clusters. Second, I assessed semantic similarity by comparing cosine distances in the embedding space to human judgments – as determined by crowd-assessed similarity averages. Due to the lack of a gold-standard French dataset, I translated Finkelstein et al. (2001)'s WordSim353, removing

3 pairs whose associations lacked French equivalents (e.g. 'soap' and 'opera'). Finally, I implemented Grave et al. (2018)'s French analogy task to assess whether semantic relationships – such as "king is to queen what man is to woman" – and syntactic patterns were accurately captured in the vector space, providing comparative metrics across embedding models (Mikolov et al., 2013).

3.4 WEAT hypothesis testing

To test for gender bias in the three corpora, I apply the Word Embedding Association Test (WEAT) methodology developed by Caliskan et al. (2017). This approach tests whether there exists a statistically significant difference in the strength of association between two sets of target words with respect to two sets of attribute words. In the context of this gender bias analysis, let X and Y be two sets of **target words** representing different conceptual categories (e.g. career vs. family), and let M and F be two sets of **attribute words** representing unambiguously male and female concepts, respectively.

For each word w in the target sets, represented by a word embedding vector $\vec{w}$, I compute an association measure $s(w, M, F)$ that quantifies its relative association with male versus female attributes using cosine similarity:

$$s(w, M, F) = \frac{1}{|M|} \sum_{m \in M} \cos(\vec{w}, \vec{m}) - \frac{1}{|F|} \sum_{f \in F} \cos(\vec{w}, \vec{f}) \tag{1}$$

A higher value of $s(w, M, F)$ indicates that word w is more strongly associated with male attributes than female attributes, while a lower or negative value indicates stronger association with female attributes. The test statistic for the entire target set comparison is defined as:

$$S(X, Y, M, F) = \sum_{x \in X} s(x, M, F) - \sum_{y \in Y} s(y, M, F) \tag{2}$$

The null hypothesis (H_0) states that there is no difference between the two sets of target words (X and Y) in terms of their relative association with the attribute sets (M and F). To test it, I perform a permutation test. Let $Z = X \cup Y$ be the union of the two target sets. The permutation test creates all possible partitions (X_i, Y_i) of Z such that $|X_i| = |X|$ and $|Y_i| = |Y|$. For each partition, I compute the test statistic $S(X_i, Y_i, M, F)$. The one-sided p-value is calculated as:

$$p = \Pr_i[S(X_i, Y_i, M, F) > S(X, Y, M, F)]. \tag{3}$$

This represents the probability of observing a test statistic more extreme than the one computed for the original sets, assuming the null hypothesis is true. Following Betti et al. (2023), I perform 1,000 random permutations instead of exhaustively testing all possible

partitions, which is sufficient for statistical inference.

To measure the magnitude of the bias, I compute a standardized **effect size** using Cohen's d (Cohen, 1988):

$$d = \frac{\text{mean}_{x \in X} s(x, M, F) - \text{mean}_{y \in Y} s(y, M, F)}{SD_{\text{pooled}_{w \in X \cup Y}} s(w, M, F)} \tag{4}$$

Based on this formulation, a positive effect size would indicate that the words in X are more similar to the words in M than in F, and the words in Y are more similar to the words in F than in M (Betti et al., 2023).

Charlesworth et al. (2021) note that WEAT effect sizes often appear statistically non-significant at the .05 level, due to relatively large standard errors. The latter are influenced by several corpus-related factors, which include the number of stimulus words used to define targets or attributes, the frequency with which these words appear in the corpus, and the number of co-occurrences between stimulus words. Studies continue to examine the various factors that influence the reliability and sensitivity of WEAT outcomes (Ethayarajh et al., 2019). Despite nonsignificance at conventional thresholds, I join Charlesworth et al. (2021) in arguing that large effect sizes (>.5) remain descriptively meaningful, especially when interpreted in the context of broader meta-analytic findings.

3.5 Selection of words and bias categories

The final selection of stimulus words is displayed in Table A.1 in Appendix A, along with a detailed discussion on its curation. Mostly, words were translated from Betti et al. (2023), and complemented with words recurrent in French rap. In terms of target words, the experiment includes five distinct categories, which represent concept pairs that tend to exhibit gender bias. These were identified from the works of Caliskan et al. (2017), Chaloner and Maldonado (2019), and Betti et al. (2023):

(B1) Career vs. family

(B2) Maths / science vs. arts

(B3) Differences on personal descriptions in terms of intelligence vs. appearance

(B4) Physical or emotional strength vs. weakness

An additional category on gender-based differences in sexuality was considered but ultimately excluded. This category could have captured how women are typically portrayed as being centred on romance and often sexually objectified, while men are expected to prioritise sexual activity (Smiler et al., 2017). However, the rising importance of sexual agency in female rappers' discourse would have complicated this analysis. Indeed, research has documented how female rappers create spaces for reclaiming sexual narratives, empowering listeners to develop independent, sex-positive attitudes (Lewis, 2024).

4 Results

4.1 Quality of embeddings

4.1.1 Clustering

An initial evaluation of the quality of the word embeddings trained on French rap lyrics was conducted through k-means clustering, visualised after t-SNE dimensionality reduction. Figure 1 reveals semantically coherent groupings in the vector space.

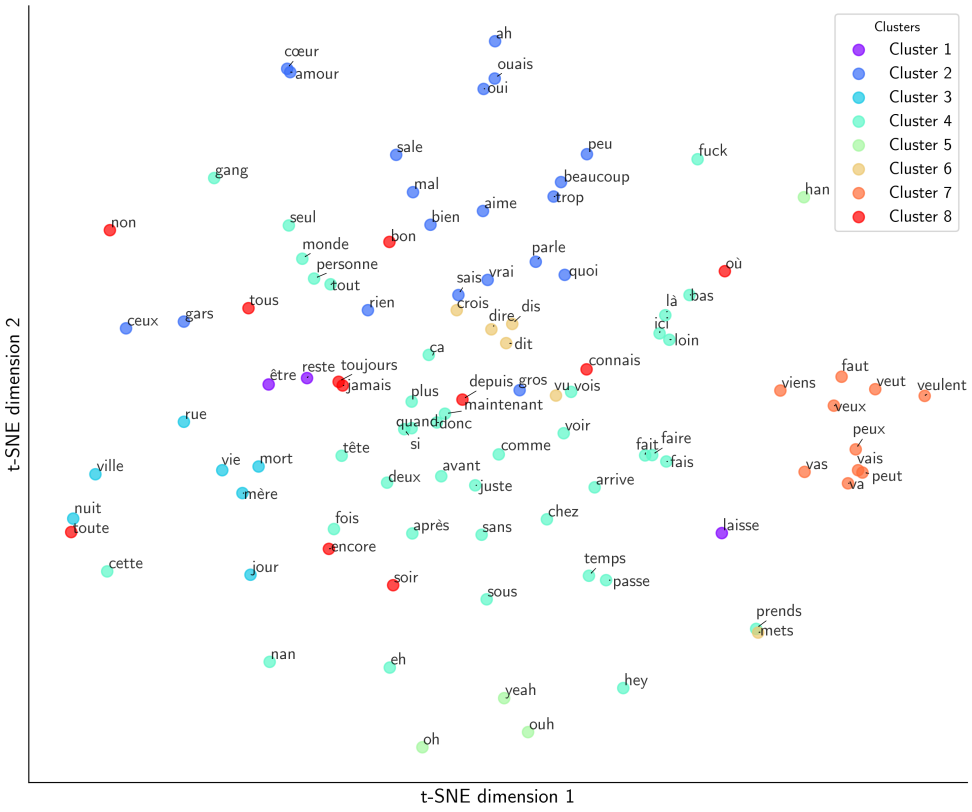

Figure 1. t-SNE visualization of word clusters in the embedding space. The plot displays 8 distinct groupings of word after k-means clustering of word embeddings produced by `frRap_100_skip`.

Notable clusters include: cluster 7 containing modal verbs (e.g., “veux,” “peux,” “faut”); cluster 2 grouping quantity terms (“beaucoup,” “peu,” “trop”) alongside emotional concepts (“coeur,” “amour”); and cluster 3 containing existential terms like “vie” (life) and “mort” (*death*). These observations hint at a relative success in the encoding of syntactic and semantic relationships in the trained embeddings, despite the relatively modest corpus size.

4.1.2 Word similarities

Embeddings were evaluated using the translated WordSim353 dataset, comparing cosine similarities between word pairs against human judgments. As shown in Table B.1, the

rap corpus embeddings achieved moderate performance with Pearson's coefficients ranging from 0.20 to 0.23.

Among rap embeddings, 100-dimensional models slightly outperformed their 200-dimensional counterparts, suggesting that for this relatively small corpus, additional dimensionality introduced noise rather than captured meaning. When compared to off-the-shelf embeddings, the rap corpus models demonstrated substantially lower correlations, with FrWac embeddings achieving coefficients exceeding 0.50. This performance gap, illustrated in Fig. B.1 likely reflects the considerable difference in corpus size.

4.1.3 Analogies

The final validation assessment employed word analogy tasks to measure the embeddings' ability to capture semantic and syntactic relationships. Table B.2 shows the rap corpus embeddings achieved notably lower overall accuracy compared to off-the-shelf embeddings (best rap model: 8.2% vs. best reference model: 37.8%). CBOW consistently outperformed skip-gram counterparts, while poor performance on some syntactic analogies likely reflects distinctive linguistic features of rap or corpus size limitations. Accuracy in analogy tasks has however sometimes been described as a poor metric for the quality of embeddings (Abdine et al., 2022; Khalife et al., 2019), with extrinsic evaluation deemed more reliable. However, no such validation was conducted in this study.

To conclude on this validation process, the French rap embeddings arguably demonstrated sufficient quality for the gender bias analysis to follow. While the performance gap with reference embeddings somewhat hinders the validity of comparative analysis, this evaluation provides confidence that gender associations identified through the trained embeddings reflect authentic patterns in the rap lyrics corpus. Importantly however, the relatively lower quality of rap embeddings may lead to underestimation of gender bias, as poorer semantic understanding could obscure subtle but real gender associations in the lyrics.

4.2 Gender bias in French rap and general French text

To begin analysing gender bias in French rap lyrics, I report results of the WEAT across the five categories of gender associations. Figure 2 presents results of the permutation tests for the `frRap_100_skip` model, illustrating the distribution of test statistics – calculated using Eq. 2 – for each category under the null hypothesis.

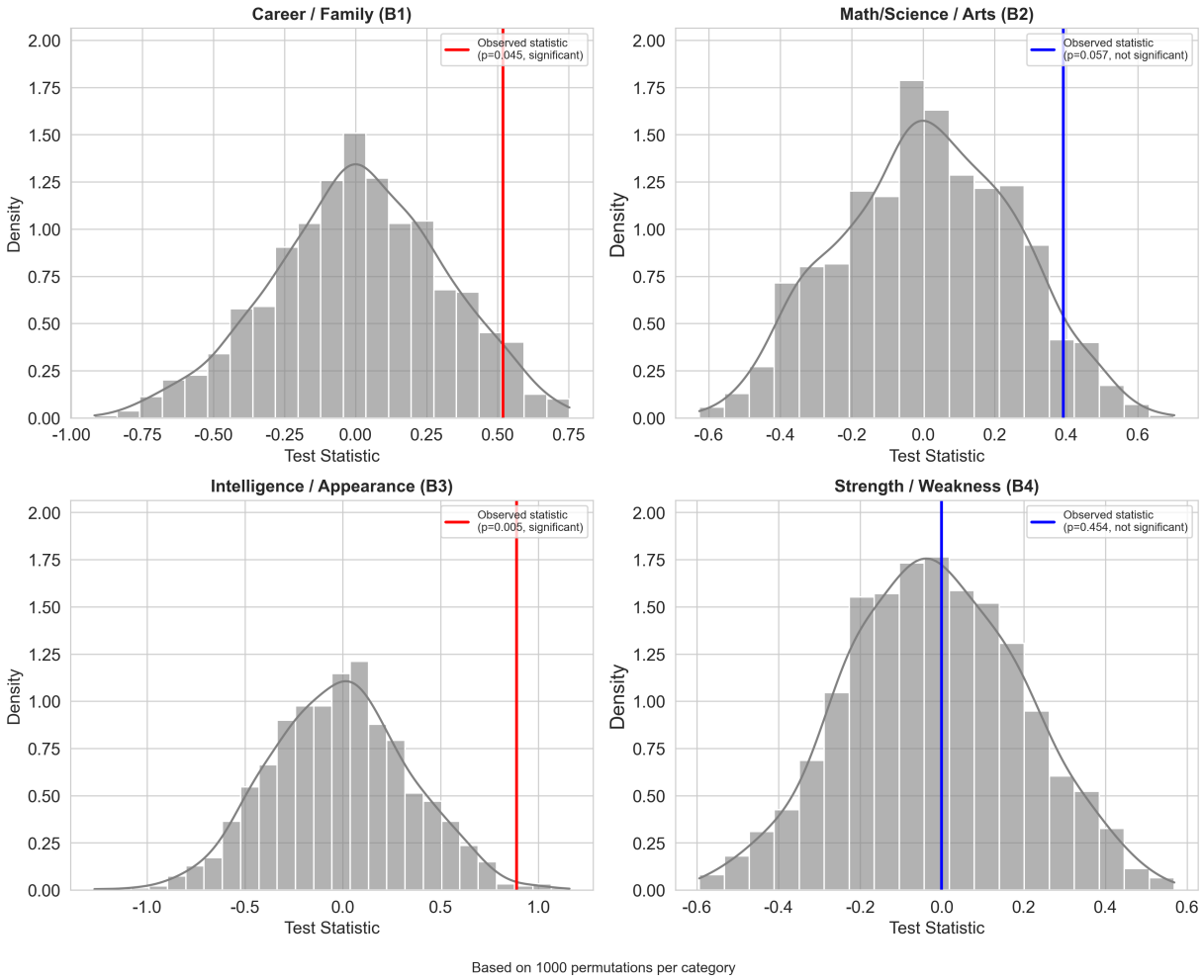

Figure 2. Distribution of test statistics from 1,000 random permutations for each bias category for the `frRap_100_skip` embeddings. Vertical lines represent observed test statistics for each category, with blue indicating non-significant results and red indicating statistically significant results ($p < 0.05$).

As shown in Figure 2, the permutation tests reveal that both the B1 (Career vs. Family) and B3 (Intelligence vs. Appearance) categories demonstrate statistically significant gender bias in the `frRap_100_skip` model. This suggests words related to career and intelligence are more strongly associated with male attributes, while words related to family and appearance show stronger association with female attributes. The remaining categories do not exhibit statistically significant gender bias in this model, although B2 (Maths & Science vs. Arts) approaches significance (p=0.06).

These initial findings from a single embedding model provide a first glimpse into gender associations in French rap lyrics. To establish more robust and comparative conclusions, we can now examine results across all embedding models for both rap lyrics and general French text. Building on the initial permutation test analysis, Table 2 presents the effect sizes (ES) and p-values from WEAT analyses across all models and bias categories.

Corpus	Model	Career-Family	Math/Science-Arts	Intelligence-Appearance	Strength-Weakness
frRap	frRap_200_cbow	ES=0.281, p=0.288	ES=0.821, p=0.080	**ES=0.934, p=0.020**	ES=-0.107, p=0.562
	frRap_100_cbow	ES=0.218, p=0.325	ES=0.880, p=0.070	**ES=1.080, p=0.009**	ES=0.006, p=0.498
	frRap_200_skip	ES=0.679, p=0.097	ES=0.521, p=0.178	**ES=1.221, p=0.007**	ES=-0.091, p=0.560
	frRap_100_skip	**ES=0.866, p=0.045**	ES=0.940, p=0.057	**ES=1.354, p=0.005**	ES=0.010, p=0.454
frWac	frWac_200_cbow	**ES=1.307, p=0.007**	**ES=2.173, p=0.001**	**ES=0.832, p=0.029**	**ES=1.093, p=0.027**
	frWac_200_skip	**ES=1.778, p=0.002**	**ES=1.445, p=0.008**	**ES=1.196, p=0.007**	**ES=1.526, p=0.006**
frWiki	frWiki_1000_cbow	**ES=1.043, p=0.028**	ES=0.655, p=0.136	ES=-0.150, p=0.629	ES=0.403, p=0.252
	frWiki_1000_skip	**ES=1.817, p=0.001**	**ES=2.001, p=0.001**	ES=0.328, p=0.227	ES=0.030, p=0.450

Table 2. Gender bias WEAT results across embedding models. Each cell reports the effect size (ES) and p-value (p) for a WEAT association test. Higher ES indicates a stronger association between target and attribute concepts. Bold values indicate statistically significant associations ($p < 0.05$). Models are grouped by corpus.

The Intelligence-Appearance category (B3) shows strong gender bias in the rap corpus, with significant effects across all rap embedding models (ES ranging from 0.934 to 1.354, p < 0.05). This robust tendency to associate intelligence with male attributes and appearance with female attributes also appears in both frWac models, suggesting consistency across different French language domains.

For the Math/Science-Arts category (B2), significant gender bias appears in both frWac models and the frWiki skip-gram model, but not in any rap embedding model, though some approach significance (p=0.057 to p=0.178). This suggests that associations linking Math/Science with male attributes and Arts with female attributes may be more pronounced in general French text than in rap lyrics specifically. This finding might be explained by the fact that, rapping being an art form, male rappers push their own artistry to the foreground and do not necessarily associate it more to women.

The Career-Family category (B1) exhibits a more complex pattern. Significant gender bias appears in three models: `frRap_100_skip` (ES=0.866), `frWac_200_cbow` (ES=1.307), `frWac_200_skip` (ES=1.778), and both frWiki models (ES=1.043 and ES=1.817). The strongest effects appear in the frWiki and frWac corpora, suggesting that encyclopedic and web French text more strongly associate career-related terms with male attributes and family-related terms with female attributes compared to most rap embedding configurations. This is a rather surprising finding, given the importance of monetary success and the 'hustle' culture in rap music.

Interestingly, the Strength-Weakness category (B4) shows significant gender bias only in the frWac models, while being notably absent in all rap and Wiki models. This suggests that associations between strength and male attributes or weakness and female attributes may be more specific to general web French text rather than rap lyrics or encyclopedic content.

When comparing across corpus types, frRap embeddings demonstrate consistent gender

bias primarily in intelligence-appearance associations, while frWiki models show stronger biases in career-family and math/science domains. The frWac models exhibit the most pervasive bias, showing significant effects across all four categories in at least one configuration. Figure 3 illustrates these findings.

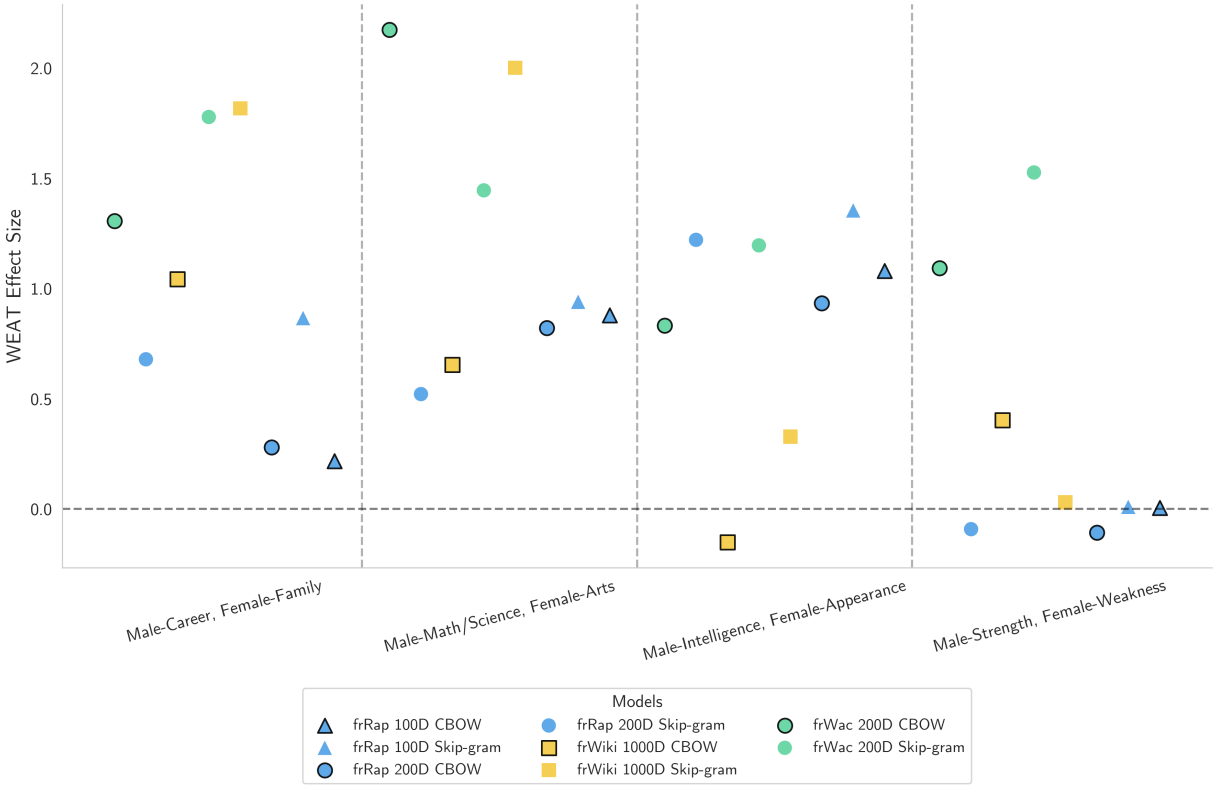

Figure 3. WEAT effect sizes across different bias categories and embedding models. Positive effect sizes indicate associations between male attributes and the first concept in each pair (Career, Math/Science, Intelligence, Strength) and between female attributes and the second concept (Family, Arts, Appearance, Weakness), as computed in Eq 4. The horizontal line at zero represents no bias.

5 Discussion

5.1 Main results and implications

This study provides the first computational analysis of gender bias in French rap lyrics. Results reveal significant gender bias in rap, particularly in intelligence-appearance associations, where intelligence terms link strongly with male attributes and appearance terms with female attributes. General French text embeddings demonstrate more pervasive gender bias than rap lyrics, with web text (frWac) showing significant bias across all four categories, while French Wikipedia exhibits strong bias in career-family and maths/science dimensions. Any comparative interpretation is however strongly limited by differences in corpus size, and resulting performance disparities explored in the validation process.

These findings contribute to the literature by providing empirical evidence for debates about sexism in rap music, with several implications. First, they support calls for more balanced gender representation in French rap, as advocated by organizations like Madame Rap, as increasing female artist visibility could help challenge stereotypical representations (Pégram, 2019). Second, they complicate simplistic critiques of rap as exceptionally misogynistic, supporting, Khan (2022)'s argument that moral panic surrounding rap often reflects racial bias in cultural evaluation. Finally, gender bias identified across all corpora raises concerns for large language models and embeddings trained on these texts, which risk internalising and amplifying stereotypical associations. This underscores the importance of developing effective debiasing techniques to prevent perpetuating harmful stereotypes (Bolukbasi et al., 2016).

5.2 Limitations and future research

Several limitations warrant acknowledgement. First, while WEAT methodology effectively captures stereotypical word associations, it detects only one dimension of sexism in rap lyrics. Critical forms of misogyny – including sexual objectification, explicit insults, and promotion of rape culture – remain undetected. Future research should employ complementary methods to provide a more comprehensive picture of gender representation in rap.

Second, the limited quality of embeddings trained on the rap corpus, evidenced by modest validation scores, is likely to lead to an underestimation of bias. Expanding the corpus and improving validation datasets and processes for French text would improve future studies. Besides improvement in understanding of French, enlarging the rap corpus would allow to conduct diachronic analyses examining how gender bias has evolved over time, potentially tracking changes as female representation in French rap has increased.

Third, this analysis lacks intersectionality – a critical shortcoming given how gender interacts with race and sexuality in music (Casanovas-Buliart et al., 2024). As intersectional bias measures are currently being developed (Charlesworth et al., 2024), future studies should work towards examining how gender representations shift across these identity dimensions.

6 Conclusion

This study provided the first computational analysis of gender bias in French rap lyrics, revealing significant bias in intelligence-appearance associations while showing that general French text also exhibits pervasive bias across multiple dimensions. Besides challenging simplistic critiques that single out rap as uniquely problematic, these findings also emphasise the urgent need for effective debiasing techniques as NLP systems trained on biased corpora risk perpetuating harmful stereotypes.

References

Abdine, H., Xypolopoulos, C., Eddine, M. K., & Vazirgiannis, M. (2022). Evaluation Of Word Embeddings From Large-Scale French Web Content. *arXiv preprint arXiv:2105.01990*.

Barman, M. P., Awekar, A., & Kothari, S. (2019). Decoding The Style And Bias of Song Lyrics. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1165–1168.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2000). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*(Feb), 1137–1155.

Bennett, A. (2001). *Cultures Of Popular Music*. McGraw-Hill Education.

Betti, L., Abrate, C., & Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, *12*(1), 10.

Bianchi, F., Marelli, M., Nicoli, P., & Palmonari, M. (2021). SWEAT: Scoring Polarization of Topics across Different Corpora. *arXiv preprint arXiv:2109.07231*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, *29*.

Brown, T., & Turner, J. (2020). What the WAP: Part 1-Black Feminist Scholars on Black Women's Popular Culture. *Black Feminisms*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Carney, C., Hernandez, J., & Wallace, A. M. (2016). Sexual knowledge and practiced feminisms: On moral panic, black girlhoods, and hip hop. *Journal of Popular Music Studies*, *28*(4), 412–426.

Casanovas-Buliart, L., Alvarez-Cueva, P., & Castillo, C. (2024). Evolution over 62 years: An analysis of sexism in the lyrics of the most-listened-to songs in Spain. *Cogent Arts & Humanities*, *11*(1), 2436723.

Chaloner, K., & Maldonado, A. (2019). Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 25–32). Association for Computational Linguistics.

Charlesworth, T. E. S., Ghate, K., Caliskan, A., & Banaji, M. R. (2024). Extracting intersectional stereotypes from embeddings: Developing and validating the Flexible Intersectional Stereotype Extraction procedure. *PNAS Nexus*, *3*(3).

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, *32*(2), 218–240.

Cobb, M. D., & Boettcher III, W. A. (2007). Ambivalent Sexism and Misogynistic Rap Music: Does Exposure to Eminem Increase Sexism? *Journal of Applied Social Psychology*, *37*(12), 3025–3042.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.

Davis, S. (1985). Pop Lyrics: A Mirror and a Molder of Society. *Et Cetera*, *42*, 167.

Dhoest, A., Herreman, R., & Wasserbauer, M. (2015). Into the Groove - Exploring lesbian and gay musical preferences and 'LGB music' in Flanders. *Observatorio*, *9*(2).

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding Undesirable Word Embedding Associations. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1696–1705). Association for Computational Linguistics.

Eze, S. U. (2020). Sexism and Power Play in the Nigerian Contemporary Hip Hop Culture: The Music of Wizkid. *Contemporary Music Review*, *39*(1), 167–185.

Fauconnier, J.-P. (2015). French Word Embeddings. http://fauconnier.github.io

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web*, 406–414.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Garland, D. (2008). On the concept of moral panic. *Crime, Media, Culture*, *4*(1), 9–30.

Goldberg, Y. (2017). *Neural network methods in natural language processing*. Morgan & Claypool Publishers.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Griffin, M., & Fournet, A. (2020). F**k B*tches Raw on the Kitchen Floor: A Feminist Examination of Condom Messages in Hip Hop and Rap Music, 1991–2017. *Sexuality & Culture*, *24*(1), 291–304.

Griffin, M., Fournet, A., Zhai, A., & Mascary, D. (2024). There's Some Whores in this House: An Examination of Female Sexuality in R&B/Hip Hop and Pop Music, 1991–2021. *Sexuality & Culture*, *28*(2), 610–631.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Association for Computational Linguistics.

Hansen, K. A. (2022). *Pop Masculinities: The Politics of Gender in Twenty-first Century Popular Music*. Oxford University Press.

Kamal Eddine, M., Tixier, A., & Vazirgiannis, M. (2021). BARThez: A Skilled Pretrained French Sequence-to-Sequence Model. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 9369–9390). Association for Computational Linguistics.

Keyes, C. L. (2000). Empowering self, making choices, creating spaces: Black female identity via rap music performance. *Journal of American Folklore*, *113*(449), 255–269.

Khalife, S., Liberti, L., & Vazirgiannis, M. (2019). Geometry and Analogies: A Study and Propagation Method for Word Representations. In C. Martín-Vide, M. Purver, & S.

Pollak (Eds.), *Statistical Language and Speech Processing* (pp. 100–111). Springer International Publishing.

Khan, U. (2022). A guilty pleasure: The legal, social scientific and feminist verdict against rap. *Theoretical Criminology*, *26*(2), 245–263.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2479–2490). European Language Resources Association.

Lewis, A. A. (2024). Rapping about pleasure: The role of Black women's rap music in shaping Black women's sexual attitudes. *Culture, Health & Sexuality*, 1–15.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Clergerie, É. V. d. l., Seddah, D., & Sagot, B. (2020). CamemBERT: A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Pégram, S. (2019). Respectez-nous as We Feminize the Rapped Rhyme: Women Rappers and Gender Empowerment in French Hip-Hop. *Popular Culture Review*, *30*(1), 53–94.

Reyna, C., Brandt, M., & Tendayi Viki, G. (2009). Blame It on Hip-Hop: Anti-Rap Attitudes as a Proxy for Prejudice. *Group Processes & Intergroup Relations*, *12*(3), 361–380.

Smiler, A. P., Shewmaker, J. W., & Hearon, B. (2017). From "I Want To Hold Your Hand" to "Promiscuous": Sexual Stereotypes in Popular Music Lyrics, 1960–2008. *Sexuality & Culture*, *21*(4), 1083–1105.

Stanczak, K., & Augenstein, I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

van Bohemen, S., den Hertog, L., & van Zoonen, L. (2017). Music as a resource for the sexual self: An exploration of how young people in the Netherlands use music for good sex. *66*, 19–29.

van Maanen, J. (Ed.). (1995). *Representation in Ethnography* (First Edition). SAGE Publications, Inc.

Zurbuchen, L., & Voigt, R. (2024). A Computational Analysis and Exploration of Linguistic Borrowings in French Rap Lyrics. In X. Fu & E. Fleisig (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 200–208). Association for Computational Linguistics.

Appendix A. Additional material on the choice of words used for the word embedding association tests.

Table A.1 displays the sets of words used in the WEAT experiment, translated from Betti et al. (2023), Caliskan et al. (2017), and Chaloner and Maldonado (2019), and complemented with French rap-specific vocabulary. Words that occurred less than 5 times in the French rap corpus were removed, as well as words absent from the vocabulary of the off-the-shelf embeddings of general text.

Category	Word set name	Words
Attribute	Male attributes (M)	homme, mec, gars, frère, il, père, papa, oncle, grand-père, bro, cousin
	Female attributes (F)	femme, meuf, fille, sœur, elle, mère, maman, tante, grand-mère, miss, cousine
Target	Career (B1-X)	business, patron, patronne, argent, money, travail, boss, cash, hustle, bureau, carrière
	Family (B1-Y)	foyer, parents, maison, enfant, enfants, famille, mariage, marier
	Math/Sci (B2-X)	calcul, logique, science, chiffres, physique, maths, chimie
	Arts (B2-Y)	poésie, art, danse, littérature, chanson, peinture
	Intelligence (B3-X)	brillant, brillante, intelligent, intelligente, stratège, cerveau, sage, lucide, génie
	Appearance (B3-Y)	beau, belle, mince, moche, laid, laide, joli, jolie, maigre, gros, grosse, corps
	Strength (B4-X)	confiant, confiante, confiance, puissant, puissante, puissance, fort, forte, force, dominant, dominante, dominance
	Weakness (B4-Y)	faible, fragile, timide, doux, douce, sensible, soumis, soumise, peur, vulnérable

Table A.1. Word sets used for word embedding association tests. Most attribute and target words were translated from Caliskan et al. (2017), Chaloner and Maldonado (2019), and (Betti et al., 2023).

Addressing grammatical gender in target words

A key methodological challenge in this analysis stems from French grammatical gender, particularly as the corpus was not lemmatised. French adjectives and many nouns have distinct masculine and feminine forms (e.g., "intelligent"/"intelligente"), which can split semantically equivalent concepts across different word vectors in the embedding space. This might introduce artificial bias or reduce statistical power.

To assess the potential impact of grammatical gender on the analysis, cosine similarities were measured between the masculine and feminine version of gendered word pairs across all models. This complementary analysis revealed substantial variation, both across models (Fig. A.1) with average similarities ranging from 0.692 to 0.035, and across word pairs (Fig. A.2). This indicates that most models represent gender variants as distinct concepts rather than mere morphological variations.

To address this issue, the word choice approach prioritised gender-invariant terms where possible. For essential words with distinct gender forms, both variants were included. To verify the robustness of the findings, additional WEAT analyses were conducted using word sets that excluded gendered pairs, retaining only gender-invariant terms. These analyses yielded consistent effects and similar significance patterns across all bias categories, confirming that the results are not artifacts of the grammatical gender representation.

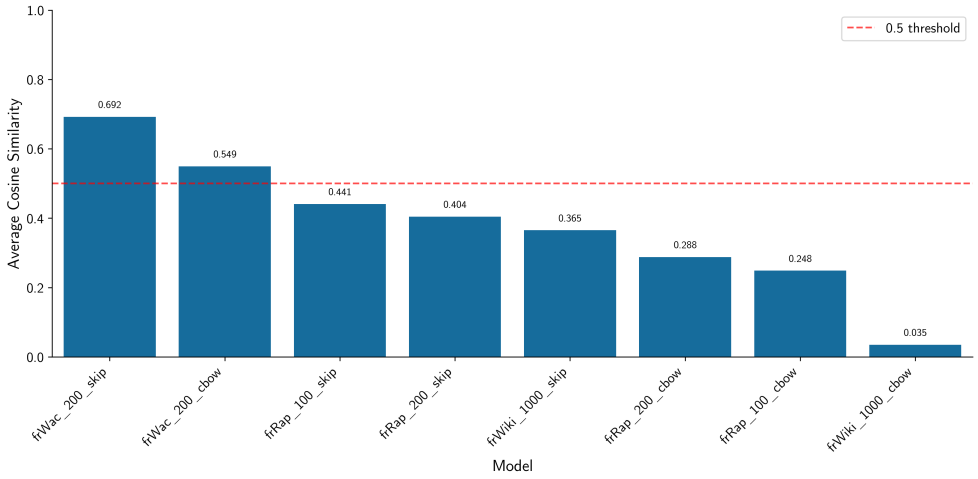

Figure A.1. Average cosine similarity between gendered word pairs across models. Mean cosine similarity between masculine and feminine word forms across different embedding models. Higher values indicate that gendered variants are represented as more semantically similar. Only frWac models exceed the 0.5 similarity threshold.

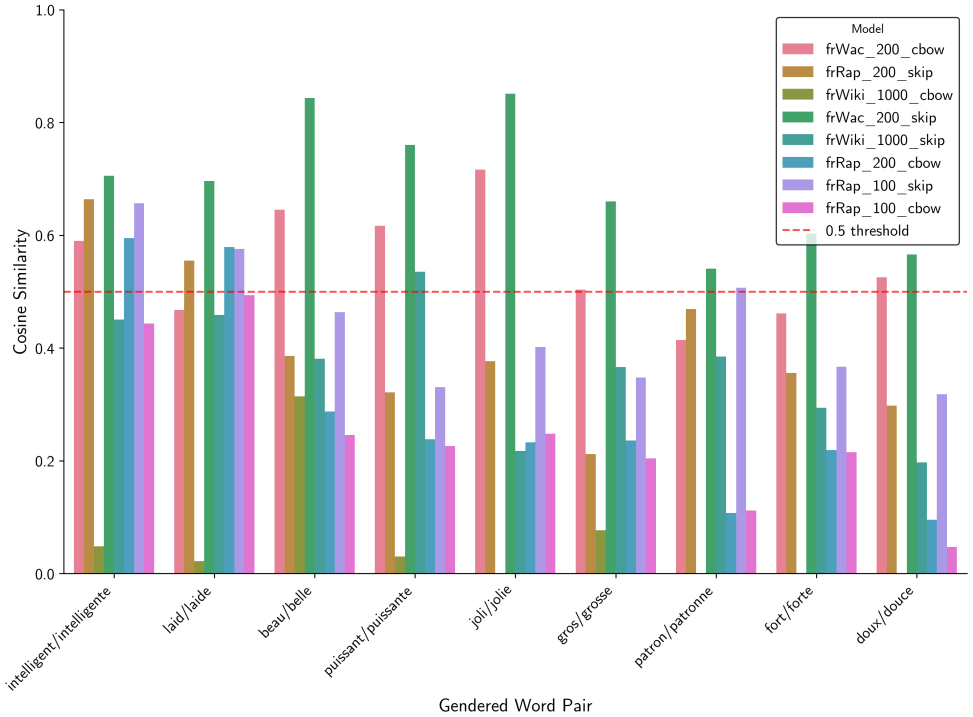

Figure A.2. Cosine similarity by gendered word pair across models. Note the substantial variation both between word pairs and across models.

Appendix B. Additional material on word embeddings evaluation.

Model	Pearson's r (p-value)	Spearman's r (p-value)	Pairs (coverage)
`frWac_200_cbow`	0.56 (1.13e-27)	0.57 (1.35e-28)	320/346 (92.5%)
`frWac_200_skip`	0.53 (1.48e-24)	0.53 (8.97e-25)	320/346 (92.5%)
`frWiki_1000_skip`	0.52 (2.63e-23)	0.54 (2.11e-25)	318/346 (91.9%)
`frRap_100_cbow`	0.23 (1.73e-04)	0.23 (1.62e-04)	256/346 (74.0%)
`frRap_100_skip`	0.23 (1.91e-04)	0.23 (1.71e-04)	256/346 (74.0%)
`frRap_200_cbow`	0.22 (3.96e-04)	0.21 (5.39e-04)	256/346 (74.0%)
`frRap_200_skip`	0.20 (1.44e-03)	0.19 (1.95e-03)	256/346 (74.0%)
`frWiki_1000_cbow`	0.17 (1.92e-03)	0.14 (1.09e-02)	318/346 (91.9%)

Table B.1. Correlation between embedding-based cosine similarities and human similarity judgments on the translated WordSim353 dataset.

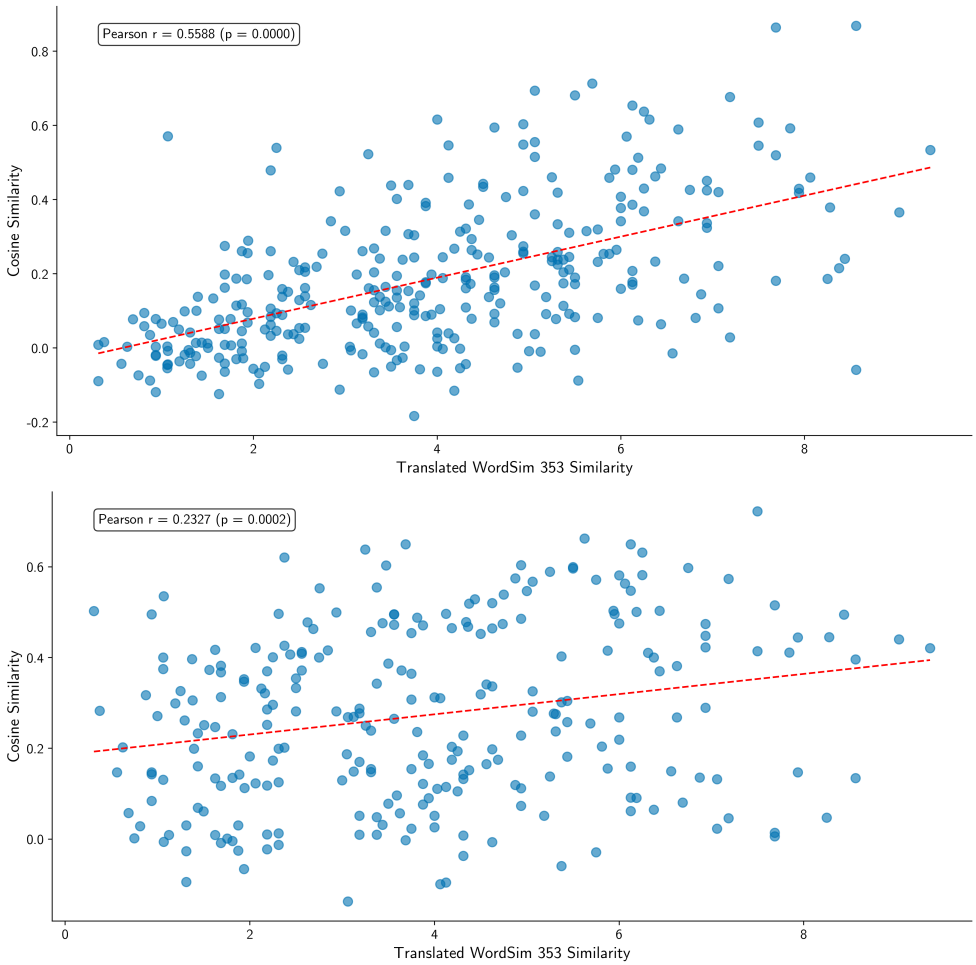

Figure B.1. Scatterplots of word pairs' cosine similarities against human judgments on the translated WordSim353 dataset. Top: `frWac_200_cbow` model. Bottom: `frRap_100_cbow` model. The difference in slope illustrates the performance gap between general French text embeddings and rap lyrics embeddings.

Model	Overall	Antonyms	Family	Gram1	Gram2	Gram3	Gram4	Gram5	Gram7	Gram8
frWiki_1000_skip	**37.8%**	**27.8%**	51.4%	N/A	20.8%	**47.4%**	30.3%	54.8%	13.3%	15.4%
frWac_200_cbow	37.3%	16.8%	68.8%	18.1%	22.1%	35.2%	**32.7%**	**67.9%**	**34.9%**	**40.2%**
frWac_200_skip	36.1%	9.3%	**72.9%**	**21.5%**	**22.4%**	45.4%	25.4%	62.8%	33.9%	32.5%
frRap_100_cbow	8.2%	2.9%	8.2%	0.0%	0.0%	7.6%	13.6%	3.6%	16.1%	7.1%
frRap_200_cbow	8.1%	2.9%	8.7%	0.0%	1.6%	7.1%	13.0%	4.5%	15.2%	1.8%
frRap_100_skip	3.7%	2.2%	5.3%	0.0%	0.0%	0.4%	7.3%	0.8%	6.0%	3.6%
frRap_200_skip	2.6%	0.4%	5.8%	0.0%	0.5%	0.4%	4.8%	0.5%	5.8%	0.0%
frWiki_1000_cbow	0.2%	0.0%	4.3%	N/A	0.0%	0.0%	0.3%	0.1%	0.0%	0.0%

Table B.2. Accuracy on word analogy tasks across different categories. Gram 1: adjective-to-adverb; Gram2: opposite; Gram3: present-to-participle; Gram4: infinitive-to-past principle; Gram5: plural; Gram7: conjugate to past tense; Gram8: verbs to plural form. Bold numbers indicate best performance across models for each category. The analogy dataset is available at the fastText documentation.

Appendix C. Code

For transparency and replicability, all code used for data pre-processing, visualisation, and analysis has been joined to this submission as zip file. It is structured as in a Github repository, with a `README.md` giving all necessary explanations to reproduce the findings.

Find this repository at: https://github.com/MatteoLarrode/NLP-SexismRapFr