Analyse Exploratoire d'un Corpus Poétique
Étude de la Variable *Anticipation*

Étudiant.e

STT-7335 Méthodes d'analyse de données

March 2, 2025

Contents

1 Description du jeu de données

Le présent rapport s'appuie sur un ensemble de 3443 vers, chacun annoté par un ensemble de variables descriptives concernant l'auteur ou l'autrice ainsi que plusieurs scores émotionnels.

1.1 Détails des variables du jeu de données

Les noms de colonnes dans le fichier source sont listés ci-dessous :

- **poet_id** : Identifiant unique associé à chaque poète.
- **poet** : Nom du poète ou de la poétesse.

- **suicidal** : Indicateur booléen précisant si l'auteur ou l'autrice s'est suicidé(e). La valeur peut être `TRUE` ou `FALSE`.
- **period** : Période d'écriture (par exemple, *Early* ou *Modern*).
- **sex** : Sexe de la personne à l'origine du poème (`Male` ou `Female`).
- **heterosexual** : Indicateur booléen spécifiant l'orientation sexuelle, `TRUE` pour hétérosexuel(le), `FALSE` dans les autres cas.
- **date_of_birth** : Date de naissance de l'auteur ou de l'autrice.
- **date_of_death** : Date de décès de l'auteur ou de l'autrice, si connue.
- **country_of_birth** : Pays de naissance.
- **poem_title** : Titre du poème auquel le vers appartient.
- **n_words** : Nombre de mots contenus dans le vers.
- **anger** : Score reflétant la présence de la colère dans le vers (valeur numérique).
- **anticipation** : Score reflétant l'anticipation (ou l'attente) véhiculée dans le vers (variable réponse principale de cette étude).
- **disgust** : Score reflétant la présence du dégoût dans le vers.
- **fear** : Score reflétant la présence de la peur dans le vers.
- **joy** : Score reflétant la présence de la joie dans le vers.
- **sadness** : Score reflétant la présence de la tristesse dans le vers.
- **surprise** : Score reflétant la présence de la surprise dans le vers.
- **trust** : Score reflétant la confiance véhiculée par le vers.
- **negative** : Score global regroupant des émotions à valence négative.
- **positive** : Score global regroupant des émotions à valence positive.
- **verse** : Position normalisée du vers dans le poème (`0` : début, `1` : fin).

Après nettoyage et vérification de la cohérence des données, le corpus final comporte 3443 vers annotés. Le nombre de données manquantes est faible et n'impacte pas significativement les analyses présentées ci-après.

Table 1: Variables numériques : statistiques récapitulatives (en français)

Variable	Min	1er quartile	Médiane	Moyenne	3e quartile	Max
Anticipation	0	0.00	0.0	0.15	0.00	3
Colère	0	0.00	0.0	0.11	0.00	4
Confiance	0	0.00	0.0	0.13	0.00	3
Dégoût	0	0.00	0.0	0.12	0.00	2
Joie	0	0.00	0.0	0.14	0.00	3
Nombre de mots	1	5.00	7.0	6.76	8.00	40
Négatif	0	0.00	0.0	0.33	1.00	4
Peur	0	0.00	0.0	0.17	0.00	4
Positif	0	0.00	0.0	0.27	0.00	4
Surprise	0	0.00	0.0	0.08	0.00	2
Tristesse	0	0.00	0.0	0.19	0.00	3
Vers	0	0.24	0.5	0.49	0.74	1

2 Présentation des variables

2.1 Statistiques descriptives

Le tableau ci-dessous présente un résumé statistique (minimum, quartiles, médiane, moyenne, maximum) des variables numériques, dont **Anticipation**, **Colère**, **Peur**, etc.

2.2 Distribution de la variable Anticipation

La Figure 1 montre l'histogramme de la variable **Anticipation**. On constate une forte concentration à 0 et quelques valeurs plus élevées, quoique peu fréquentes dans le corpus.

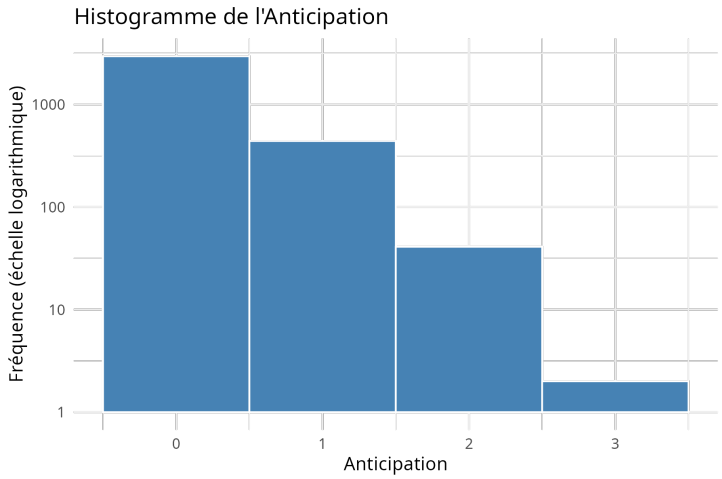

Figure 1: Histogramme de la variable **Anticipation**

2.3 Corrélations avec les autres variables émotionnelles

Les Figures 2 et 3 illustrent les corrélations entre **Colère**, **Anticipation**, **Dégoût**, **Peur**, **Joie**, **Tristesse**, **Surprise**, **Confiance**, **Négatif** et **Positif**.

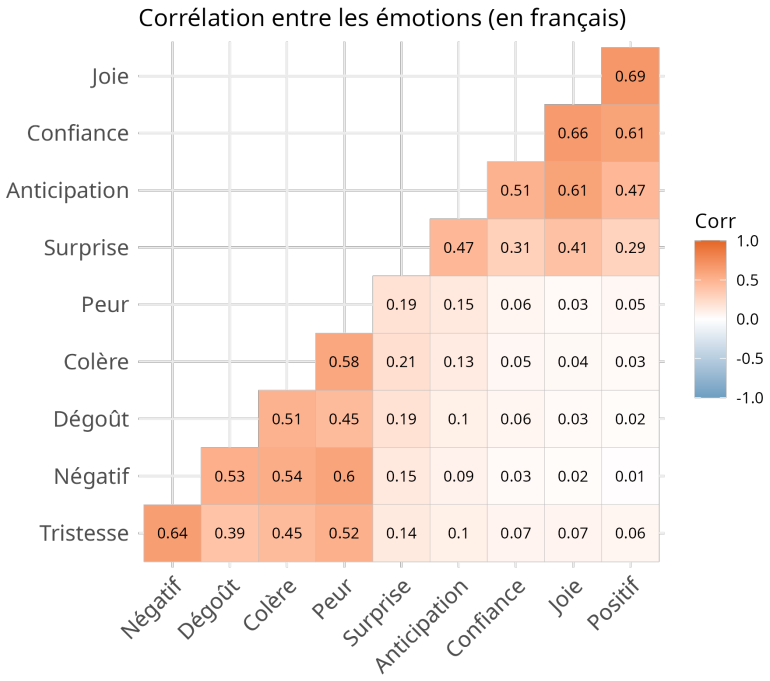

Figure 2: Matrice de corrélation (méthode `ggcorrplot`)

Cette première représentation indique la valeur des coefficients de corrélation, avec un réordonnancement par clustering hiérarchique.

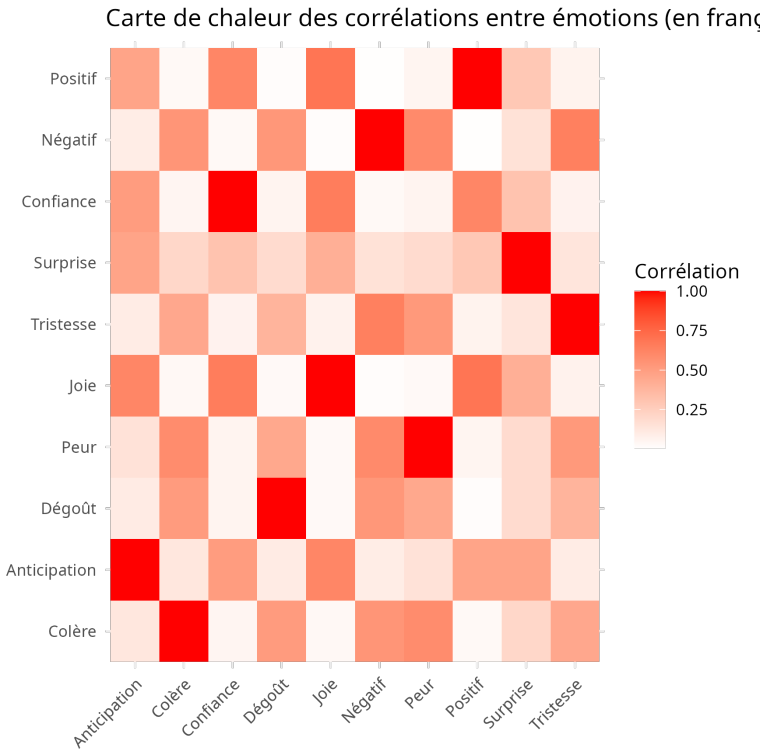

Figure 3: Carte de chaleur de la corrélation entre les émotions

Cette carte permet de repérer facilement les relations positives et négatives entre les différentes émotions. Le bleu reflète des corrélations négatives, le rouge des corrélations positives, et le blanc une corrélation nulle.

3 Analyse avec réduction de la dimension

3.1 ACP et éboulis des valeurs propres

Afin de mieux cerner la structure multidimensionnelle de ces variables émotionnelles, nous appliquons une analyse en composantes principales (ACP). La Figure 4 présente l'éboulis des valeurs propres, qui illustre la part de variance expliquée par chaque axe principal.

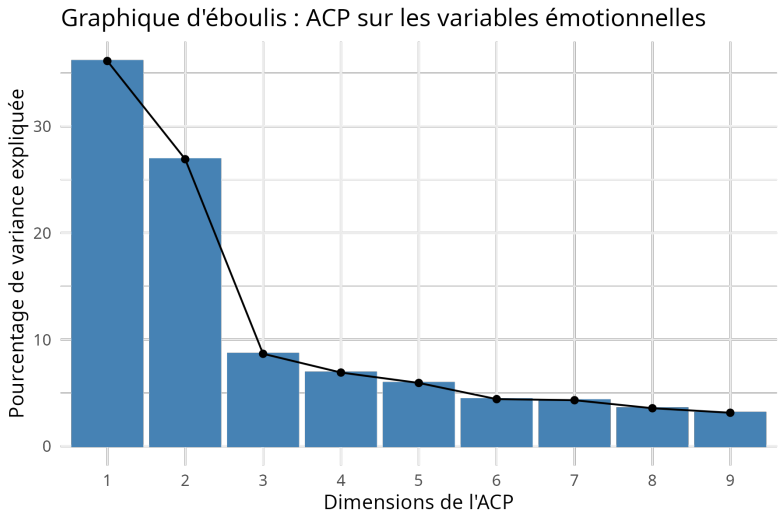

Figure 4: Éboulis (Scree Plot) des composantes principales

3.2 Biplot de l'ACP

La Figure 5 montre un biplot où les axes principaux sont utilisés pour projeter à la fois les variables (flèches) et les observations (points). Les observations sont colorées en fonction de **Anticipation** afin de situer cette variable dans l'espace factoriel.

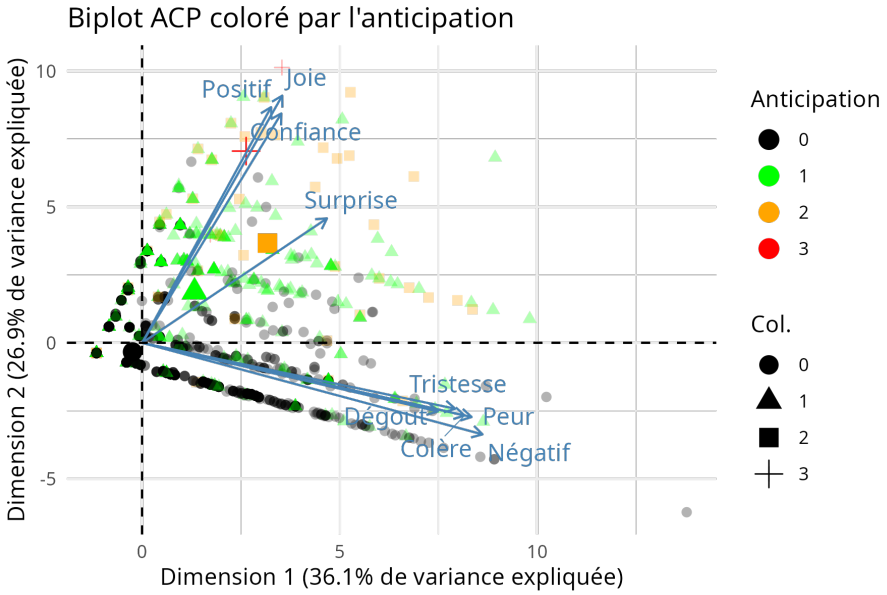

Figure 5: Biplot de l'ACP, coloré selon la variable Anticipation

On observe que **Colère**, **Peur** et **Négatif** s'opposent fortement à **Joie**, **Confiance** et **Positif**. La variable **Anticipation** se retrouve plutôt du côté des émotions positives.

3.3 Conclusion et perspectives

- L'analyse révèle une distribution faiblement étalée de la plupart des scores émotionnels, **Anticipation** comprise.
- Les corrélations confirment la présence de deux grands ensembles : **Colère**, **Peur**, **Dégoût**, **Négatif** face à **Joie**, **Confiance**, **Positif**, où **Anticipation** semble s'associer davantage à ces dernières.
- L'ACP confirme cette dichotomie et met en évidence une forte opposition sur le premier axe. Les deux premiers axes expliquent la majorité de la variance, ce qui facilite l'interprétation graphique.

Les travaux futurs pourraient inclure un modèle prédictif approfondi de la variable **Anticipation**, en intégrant davantage de caractéristiques (par exemple des indicateurs biographiques plus détaillés ou des mesures stylistiques). Il serait également intéressant d'évaluer d'autres méthodes de réduction de dimension, comme `t-SNE` ou `MDS`, afin de comparer leurs performances et d'explorer d'éventuelles structures non linéaires dans les données.