

Text Mining Freud's Complete Works: A Lexical-Affective Cartography across Volumes

Alana Farias

NeuroIn

Kiyoshi Fukutani

ferreirafk@gmail.com

Universidade Estadual de Feira de Santana

Research Article

Keywords: Freud, text mining, digital humanities, psychoanalysis, sentiment analysis, NRC Emotion Lexicon, lexical-affective mapping

Posted Date: July 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10195435/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Sigmund Freud's work constitutes one of the most influential theoretical corpora in the humanities and health sciences, yet its scope and conceptual density make it difficult to identify large-scale patterns distributed across the texts. This study presents an exploratory text-mining analysis of the Portuguese edition of Freud's Complete Works, published by Editora Imago, comprising 23 editorial volumes spanning the period from 1886 to 1939. After textual curation, the final corpus retained 12,524,088 characters and 4,145 pages and was analyzed through lexical frequency, n-grams, directed bigram networks, lexical polarity, NRC Emotion Lexicon categories, correlations between psychoanalytic keywords and emotion categories, contextual enrichment analysis, and principal component analysis. The results show that the computational approach recovered recognizable conceptual nuclei of Freudian theory, including dream, sexuality, symptom, neurosis, hysteria, unconscious, desire, pleasure, reality, and childhood. The longitudinal distribution of these terms indicates that key concepts do not appear homogeneously across the collection, but vary across editorial moments, suggesting shifts in lexical emphasis consistent with the historical and theoretical development of psychoanalysis. Lexicon-based affective analyses revealed oscillations in polarity and multidimensional emotional profiles involving positivity, negativity, trust, anticipation, fear, sadness, anger, disgust, joy, and surprise. Psychoanalytic keywords were organized into distinct affective-semantic fields, separating clinical-symptomatic nuclei from dream-interpretive and metapsychological fields. PCA further indicated that different parts of the collection share comparable affective signatures. Taken together, these findings suggest that Freud's work exhibits a measurable longitudinal lexical-affective organization. This exploratory computational cartography complements traditional theoretical interpretation and opens new possibilities for digital humanities approaches to psychoanalysis.

Introduction

Sigmund Freud's work constitutes one of the most influential and complex theoretical corpora in the humanities and health sciences (Ricoeur, 1970; Laplanche & Pontalis, 1973). Since the late nineteenth century, his formulations on the unconscious, dreams, sexuality, neurotic symptoms, and unconscious psychic processes have profoundly transformed understandings of human subjectivity, influencing not only psychology, psychiatry, and psychoanalysis, but also philosophy, anthropology, literature, sociology, and cultural studies (Freud, 1900/1953; Freud, 1905/1953; Ricoeur, 1970; Lear, 2005). Across more than four decades of intellectual production, Freud developed a broad conceptual framework that continually reworked questions concerning the functioning of the human mind (Mezan, 2014).

Traditionally, Freud's work has been investigated through hermeneutic, historical, and clinical approaches, often centered on close readings of specific texts or particular periods of his production (Assoun, 1983). These approaches remain indispensable for understanding psychoanalysis. However, they face important limitations when the goal is to examine global patterns distributed across thousands of pages. The increasing availability of computational resources and the development of digital humanities methodologies have enabled new ways of investigating large textual corpora, making it

possible to explore quantitative dimensions of intellectual production that are difficult to identify through traditional reading alone. This macrotextual perspective converges with Moretti's concept of distant reading, which proposes the use of computational and quantitative tools to track large-scale structures, patterns, and movements within extensive literary and cultural archives (Moretti, 2013).

In this context, text mining emerges as a promising approach for the systematic study of classical works (Jockers, 2013). By combining natural language processing, statistical analysis, and data visualization, text mining allows researchers to identify lexical patterns, association networks, conceptual relations, and thematic trajectories across large document collections. Rather than replacing theoretical interpretation, computational analysis expands the possibilities of inquiry by offering a complementary perspective capable of detecting regularities, tendencies, and associations that remain dispersed throughout the textual corpus. As Meunier (2015) argues, computational treatment of text does not compete with traditional hermeneutics, but functions as an additional analytical lens capable of revealing the material and statistical structure of thought as it is inscribed in large-scale textual corpora.

Recent studies have also shown that the relationship between psychoanalysis and data mining can be explored not only as a technical issue, but also as a theoretical problem concerning subjectivity, identification, and symbolic formation in contemporary digital culture. Aydin and Possati (2025), for example, draw on Lacanian theory to examine how data mining may transform processes of self-formation, arguing that profiling technologies do not merely describe subjects, but also participate in the production of new forms of identification. Unlike this predominantly theoretical and philosophical approach, the present study shifts the question to the empirical and textual level. It applies text-mining methods to a classical psychoanalytic corpus in order to investigate the lexical, conceptual, and affective organization of Freud's Complete Works.

Applying these methods to Freud's work is particularly relevant because of the dynamic nature of his theoretical construction. Fundamental concepts such as dream, desire, repression, the unconscious, sexuality, symptom, anxiety, and fantasy do not appear as static elements across the texts. Rather, they undergo conceptual displacements that accompany the historical development of psychoanalytic theory. This dynamism, together with the continuous reconfiguration of central metapsychological terms, justifies a systematic and longitudinal tracing of their textual trajectories (Garcia-Roza, 2008). Understanding how these terms are distributed, associated, and reorganized across different periods of Freud's production may provide new evidence about the internal architecture of his work and the historical trajectory of its central concepts.

Beyond the conceptual dimension, computational analysis also makes it possible to investigate affective aspects of the language used in Freud's texts. Although psychoanalysis emerged as a theory deeply concerned with emotional life, few studies have systematically mapped the affective profiles that emerge across Freud's writings as a whole. Quantitative mapping of affective language is aligned with recent developments in affective computing and lexicon-based emotion analysis applied to large historical and literary corpora, highlighting the potential of computational methods to trace affective

valence and emotional categories in lexical choices (Mohammad, 2016). The use of emotion lexicons and sentiment analysis therefore makes it possible to examine how different emotions, polarities, and affective tonalities are distributed throughout Freud's work, contributing to a broader understanding of the relationship between conceptual elaboration and affective expression in psychoanalytic discourse.

This study aims to map the lexical, conceptual, and affective organization of Sigmund Freud's Complete Works through text mining. Specifically, it investigates the distribution of central psychoanalytic concepts across Freud's production and examines their association with different emotional profiles and affective-semantic fields.

Starting from the hypothesis that Freud's work contains large-scale structural patterns that can be computationally identified, this study asks whether central psychoanalytic concepts occupy distinct emotional regions, whether thematic reorganizations occur across the editorial sequence, and whether different moments of Freud's production can be characterized by specific affective signatures. By establishing an interface between psychoanalysis, data science, and digital humanities, the study seeks to contribute to the development of new methodological strategies for textual inquiry and to expand the possibilities for analyzing one of the most important intellectual legacies of the twentieth century.

Methods

Corpus construction, curation, and textual preprocessing

An exploratory text-mining analysis was conducted on the Portuguese edition of Sigmund Freud's Complete Works, published by Editora Imago. The corpus comprised 23 editorial volumes corresponding to the chronological period from 1886 to 1939. Each volume was treated as an independent documentary unit for longitudinal, lexical, and multivariate analyses.

The PDF files were converted into plain text in R and submitted to a systematic textual curation process. This step included the removal of editorial paratexts, headers, footers, commercial elements, publication metadata, non-analytical pages, and noise derived from PDF conversion. Quality control was performed separately for each volume by recording the total number of characters before cleaning, the number of characters after cleaning, the absolute reduction, the percentage reduction, and the final number of pages retained for analysis (Table 1). Text extraction from PDF files was performed using the `pdftools` package, and only files with directly extractable textual layers were processed without optical character recognition (Ooms, 2023).

After extraction, the texts were organized in tabular format and processed according to tidy text principles. This structure allows text mining to be integrated with filtering, grouping, summarization, and visualization procedures within the tidyverse ecosystem (Silge & Robinson, 2016; Silge & Robinson, 2017). Preprocessing included conversion to lowercase, punctuation removal, removal of isolated numbers, elimination of non-alphabetic characters, and ASCII normalization during computational

processing. Stopwords, publisher-related terms, ISBN information, copyright notices, tables of contents, commercial calls, and other elements unrelated to the theoretical content of the corpus were removed.

When necessary, related lexical forms were manually harmonized to reduce conceptual fragmentation. Thus, variants such as “sonho” and “sonhos”, “desejo” and “desejos”, “sintoma” and “sintomas”, and “inconsciente” and “inconscientes” were treated as expressions of the same conceptual axis. This harmonization was used only to support concept-level analyses and did not replace the broader lexical structure of the corpus.

Lexical frequency analysis and conceptual keywords

Absolute word frequencies were calculated for the full corpus and separately for each volume. This step aimed to identify the dominant lexical terms in Freud’s work and to provide an initial quantitative characterization of the corpus. The most frequent words were visualized using bar plots (Silge & Robinson, 2017).

Bigrams and trigrams were extracted to identify recurrent lexical associations between consecutive words. Bigrams were defined as adjacent two-word sequences, whereas trigrams were defined as adjacent three-word sequences. This procedure made it possible to capture compound expressions and recurrent semantic units that are not detectable when words are analyzed in isolation (Silge & Robinson, 2017).

A directed lexical network was constructed from the most frequent bigrams. In this network, nodes represent words and edges represent transitions between consecutive terms. The direction of each edge reflects the textual order of the terms, and edge thickness is proportional to bigram frequency. This type of lexical network makes it possible to visualize local clusters of terms and recurrent relations between lexical units, complementing simple frequency-based analysis.

In addition to global lexical frequency, a matrix of conceptual keywords was constructed. These keywords were manually selected according to their theoretical relevance to Freudian psychoanalysis. The selected terms were related to dream, desire, unconscious, sexuality, symptom, neurosis, hysteria, anxiety, censorship, repression, pleasure, unpleasure, fantasy, reality, childhood, father, mother, and memory. For each keyword, relative frequency was calculated across the 23 volumes.

Lexical polarity and emotion analysis

Lexical polarity was assessed using SentiLex, a sentiment lexicon developed for opinion and sentiment extraction in Portuguese texts (Silva et al., 2010; Silva et al., 2012). SentiLex contains lemmas and inflected forms annotated with positive, negative, or neutral polarity, as well as grammatical information (Silva et al., 2010; Carvalho & Silva, 2015). Words from the corpus were matched against SentiLex, and each recognized term was classified according to the polarity value assigned by the lexicon.

Absolute and relative frequencies were calculated for each polarity class in the full corpus and by volume. For the longitudinal analysis, a polarity score was calculated for each volume as the difference between the number of positive and negative terms. Values above zero indicate a relative predominance of positive terms, values below zero indicate a relative predominance of negative terms, and values close to zero indicate a balance between the two classes. This metric was used as an exploratory indicator of lexical valence variation across Freud's work.

Discrete emotion analysis was performed using the NRC Emotion Lexicon, also known as EmoLex. This resource associates words with eight basic emotion categories, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, as well as two polarity categories, positive and negative (Mohammad & Turney, 2010; Mohammad & Turney, 2013). The NRC lexicon was accessed in R through the *syuzhet* package, which provides functions for sentiment analysis and retrieval of emotion dictionaries in multiple languages (Jockers, 2015).

Because the same word may be associated with more than one emotion category, NRC annotations were treated as non-mutually exclusive relations between lexical terms and affective categories. Global frequencies of NRC categories were calculated for the full corpus, followed by volume-level frequencies. For the longitudinal analysis, emotional proportions by volume were normalized using z-scores within each emotion category. This normalization was adopted to reduce the effect of globally abundant categories and to highlight volumes in which a given emotion was relatively more or less represented in comparison with the mean of that same emotion across the entire corpus.

Correlation between psychoanalytic keywords and emotion categories

To evaluate the association between psychoanalytic keywords and NRC emotion categories, two matrices were constructed: a volume-by-keyword matrix and a volume-by-emotion matrix. In the first matrix, each cell represented the relative frequency of a given keyword in a given volume. In the second matrix, each cell represented the proportion of a given NRC category in the same volume.

Associations between keywords and emotion categories were estimated using Spearman's rho correlation. This non-parametric method was chosen because textual data are often asymmetric, sparse, and potentially non-linear. Positive coefficients indicate that the relative frequency of a keyword tends to increase in volumes where a given emotion category also shows higher proportional representation. Negative coefficients indicate an inverse association. Correlation coefficients were represented in a heatmap using a divergent scale to distinguish positive, negative, and near-zero associations.

Contextual enrichment analysis of psychoanalytic keywords

To investigate whether the lexical contexts surrounding psychoanalytic keywords were enriched for specific emotion categories, a contextual enrichment analysis was performed based on the principle of over-representation analysis. This approach is commonly used to assess whether a given category appears more frequently than expected within a set of interest (Boyle et al., 2004; Huang et al., 2009).

In this adaptation for text mining, each psychoanalytic keyword was treated as a central term. Words located within a fixed window before and after each occurrence of that keyword were extracted across the volumes. These contextual words were then matched against the NRC Emotion Lexicon to identify which terms were associated with emotion categories such as fear, sadness, trust, joy, anger, disgust, surprise, anticipation, positive, and negative.

The background universe was defined as all corpus words recognized by the NRC Emotion Lexicon. For each emotion category, the observed frequency in the context of a given keyword was compared with the expected frequency in the full set of NRC-annotated words in the corpus. When an emotion category appeared proportionally more often in the context of a keyword than in the background universe, that category was considered enriched in that lexical context.

The statistical significance of enrichment was assessed using the hypergeometric test, and p values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure to control the false discovery rate (Benjamini & Hochberg, 1995). Results were visualized using bubble plots, in which the y-axis represents psychoanalytic keywords, the x-axis represents NRC emotion categories, point size represents the number of emotion-associated terms detected in the keyword context, and color intensity represents $-\log_{10}(\text{FDR})$. This visualization integrates, in a single plot, the enriched emotion category, the associated keyword, the magnitude of the observed signal, and its corrected statistical significance.

Principal component analysis of keywords and volumes in the affective space

Principal component analysis (PCA) was used to represent the multivariate organization of psychoanalytic concepts in an affective-lexical space. The first PCA was performed using a keyword-by-emotion matrix derived from the emotional profile associated with the lexical context of each psychoanalytic keyword. In this representation, each point corresponds to a conceptual keyword, while vectors represent NRC emotion categories. The proximity between keywords indicates similarity in contextual affective profile, whereas the direction and length of the vectors indicate the relative contribution of emotion categories to the reduced multivariate space. This approach allowed us to evaluate whether central Freudian concepts cluster according to distinct affective fields.

A second PCA was performed at the volume level using only the normalized proportions of NRC emotion categories as input variables. In this analysis, each point represents an editorial volume, and its position in the reduced space indicates similarities or displacements associated with specific affective profiles. Because the unit of analysis corresponds to editorial volumes rather than individual essays or chapters, this PCA was interpreted as an exploratory cartography of the affective organization of the collection.

Visualization and statistical analysis

All analyses were conducted in R. Visualizations were generated using ggplot2. Bar plots were used to represent lexical frequencies, polarity classes, and emotion categories. Heatmaps were used to represent longitudinal matrices, normalized affective profiles, and keyword-emotion correlations. Bubble

plots were used for contextual enrichment analyses, integrating observed signal magnitude and corrected statistical significance. Directed lexical networks were used to represent bigrams and directional relations between consecutive terms.

All statistical analyses were treated as exploratory. Correlations were estimated using Spearman's rho. Enrichment tests were based on the hypergeometric distribution. When multiple hypotheses were tested simultaneously, p values were adjusted using the Benjamini-Hochberg false discovery rate procedure (Benjamini & Hochberg, 1995). Associations with $FDR < 0.05$ were considered statistically significant in the enrichment analyses, provided that they were also accompanied by inspection of the words supporting each association.

Results

Textual curation of the Freudian corpus

The analyzed corpus consisted of 23 volumes of Freud's Complete Works, published in Portuguese by Editora Imago and covering the period from 1886 to 1939. The textual curation step preserved most of the extracted content while removing predominantly editorial paratexts, conversion noise, publication metadata, and excerpts not belonging to the analytical body of the work.

Overall, the volumes contained 12,824,424 characters before cleaning and 12,524,088 characters after curation. This corresponded to the removal of 300,336 characters, equivalent to approximately 2.34% of the initial corpus. At the end of preprocessing, 4,145 textual pages were retained for analysis.

The percentage reduction ranged from 0.39% to 7.21% across volumes, indicating that the cleaning procedure was conservative and did not substantially reduce the textual content of the corpus. Volume 1 showed the highest relative reduction, possibly because it contained a larger amount of introductory, editorial, or paratextual material, whereas the remaining volumes showed more limited reductions (Table 1).

Table 1

Summary of textual curation and character reduction across the volumes of Freud's Complete Works used in the corpus.

Work	Publisher	Total Characters (N)	Total Characters After Cleaning (n)	Character Reduction (N - n)	Percentage Reduction (%)	Pages Retained (n)
Freud.Vol.01 (1886–1889)	Imago	790382	733410	56972	7.21	292
Freud.Vol.02 (1893–1895)	Imago	660005	652202	7803	1.18	204
Freud.Vol.03 (1893–1899)	Imago	591753	559914	31839	5.38	199
Freud.Vol.04 (1900–1900)	Imago	708846	697352	11494	1.62	218
Freud.Vol.05 (1900–1901)	Imago	722538	709182	13356	1.85	227
Freud.Vol.06 (1901–1901)	Imago	541934	533056	8878	1.64	174
Freud.Vol.07 (1901–1905)	Imago	545969	534150	11819	2.16	190
Freud.Vol.08 (1905–1905)	Imago	473694	466938	6756	1.43	149
Freud.Vol.09 (1906–1908)	Imago	401276	392684	8592	2.14	133
Freud.Vol.10 (1909–1909)	Imago	554298	541032	13266	2.39	182
Freud.Vol.11 (1910–1910)	Imago	421095	411097	9998	2.37	138
Freud.Vol.12 (1911–1913)	Imago	594660	577650	17010	2.86	204
Freud.Vol.13 (1913–1914)	Imago	494663	487801	6862	1.39	160
Freud.Vol.14 (1914–1916)	Imago	638751	623425	15326	2.40	201
Freud.Vol.15 (1915–1916)	Imago	500555	497013	3542	0.71	151
Freud.Vol.16 (1915–1916)	Imago	494220	492316	1904	0.39	145

Work	Publisher	Total Characters (N)	Total Characters After Cleaning (n)	Character Reduction (N - n)	Percentage Reduction (%)	Pages Retained (n)
Freud.Vol.17 (1917–1918)	Imago	524915	513580	11335	2.16	168
Freud.Vol.18 (1920–1922)	Imago	543754	534785	8969	1.65	172
Freud.Vol.19 (1923–1925)	Imago	481184	463372	17812	3.70	163
Freud.Vol.20 (1925–1926)	Imago	559004	548684	10320	1.85	173
Freud.Vol.21 (1927–1931)	Imago	482252	472321	9931	2.06	154
Freud.Vol.22 (1932–1936)	Imago	511403	503230	8173	1.60	161
Freud.Vol.23 (1937–1939)	Imago	587273	578894	8379	1.43	187

Freud's work shows a lexical architecture organized around recurrent conceptual nuclei

An integrated view of the global lexical structure of Freud's work showed that the most frequent terms were associated with both general argumentative vocabulary and central psychoanalytic nuclei, including "dream", "ego", "sexuality", "life", and "fact" (Fig. 1A). Bigram analysis refined this interpretation by recovering recurrent lexical associations such as "sexual life", "dream thoughts", "dream interpretation", "external world", and "mental life" (Fig. 1B). Complementarily, the most frequent trigrams included expressions such as "latent dream thoughts", "internal understanding insight", "manifest dream content", "early years of", and "hysteria and obsessive" (Fig. 1C).

The directed bigram network revealed an organized lexical architecture in which terms such as "dream", "desire", "interpretation", "content", "neurosis", "symptom", "sexuality", "pleasure", "anxiety", "unconscious", and "reality" appeared articulated in chains of textual association (Fig. 1D). The thickness of the connections indicates that some pairs of terms function as recurrent units of conceptual construction, particularly within the fields of dream theory, neurotic clinical experience, sexual life, and psychic processes.

The longitudinal trajectory of conceptual keywords showed that terms such as "dream", "sexuality", "symptom", "repression", "reality", "pleasure", "father", "neurosis", "memory", "fear", "mother", "childhood", "unconscious", "hysteria", "fantasy", "unpleasure", "desire", "censorship", and "anxiety" displayed distinct relative intensities across the volumes (Fig. 1E). This distribution indicates that Freud's work does not exhibit a homogeneous conceptual organization across the collection, but rather alternates thematic nuclei throughout the editorial sequence.

(A) Most frequent words in the complete corpus after textual curation, removal of editorial paratexts, and lexical filtering. (B) Most frequent bigrams identified across the corpus, representing recurrent lexical associations between pairs of consecutive terms. (C) Most frequent trigrams, highlighting compound expressions and recurrent semantic units in the corpus. (D) Directed network of the most frequent bigrams, in which nodes represent words and edges indicate the order of occurrence between consecutive terms; edge thickness is proportional to bigram frequency. (E) Relative trajectory of conceptual keywords across the volumes after grouping lexical variants by concept. Each cell represents the relative intensity of a keyword in a given volume, scaled from 0 to 1 within each concept, allowing the identification of thematic concentration peaks throughout the work. The analysis highlights the global predominance of terms related to dreams, sexuality, psychic life, symptoms, neurosis, and interpretive processes, while also revealing longitudinal oscillations in the distribution of the main psychoanalytic concepts.

Freud's work exhibits affective oscillations and multidimensional emotional profiles

The global and longitudinal affective characterization of the corpus using SentiLex revealed a broad presence of terms classified as positive and negative, with a smaller proportion of neutral terms (Fig. 2A). This result indicates that the Freudian corpus contains substantial evaluative lexical material, consistent with the recurrent presence of terms related to conflict, suffering, pleasure, sexuality, symptoms, mental life, and psychic elaboration.

Analysis with the NRC Emotion Lexicon further expanded this characterization by decomposing the vocabulary into specific affective categories. In addition to the positive and negative classes, occurrences associated with trust, anticipation, fear, sadness, anger, joy, disgust, and surprise were identified. The global frequency profile showed a strong presence of positive, negative, trust, and anticipation categories, as well as relevant contributions from fear and sadness (Fig. 2B).

The longitudinal analysis of lexical polarity showed oscillations across volumes, with some segments presenting a relative predominance of positive polarity and others showing a predominance of negative polarity (Fig. 2C). This pattern suggests that the lexical tonality of the work is neither stable nor uniform, but varies according to the thematic and clinical focus of the volumes.

The z-score-normalized emotional profile revealed additional volume-specific variations. This result indicates that Freud's work exhibits a dynamic affective landscape, marked by alternating patterns of trust, anticipation, fear, sadness, negativity, anger, disgust, joy, and surprise across the editorial sequence (Fig. 2D).

(A) Global distribution of lexical polarity according to SentiLex, classified as positive, negative, and neutral terms. (B) Global frequency of NRC Emotion Lexicon categories in the complete corpus. (C) Longitudinal variation in lexical polarity across the volumes, calculated as the difference between the number of positive and negative terms in each volume. The dashed horizontal line indicates balance between positive and negative polarity; values above zero indicate positive predominance, whereas

values below zero indicate negative predominance. (D) NRC emotional profile across the volumes, normalized by z-score within each emotion category. Positive values indicate volumes with a higher relative proportion of a given emotion compared with the mean of that same emotion across the corpus, whereas negative values indicate lower relative representation. This normalization reduces the effect of globally abundant categories and highlights volume-specific variation in each affective category across the editorial sequence. The results indicate that Freud's work does not show a uniform affective predominance, but rather alternating patterns of positivity, trust, anticipation, fear, sadness, anger, negativity, and other NRC categories across the corpus.

Clinical and metapsychological concepts occupy distinct emotional regions

To explore the relationship between selected psychoanalytic concepts and NRC emotion categories, we constructed a keyword-by-emotion correlation matrix. This analysis showed that specific concepts were differentially associated with particular affective dimensions (Fig. 3A). Clinical terms such as "symptom", "neurosis", "hysteria", and "anxiety" showed greater proximity to negative categories, including fear, sadness, and negative polarity. In contrast, terms such as "dream", "desire", "pleasure", "unconscious", "reality", and "censorship" occupied regions closer to dimensions such as trust, positivity, surprise, and joy. These statistically significant associations do not imply that such concepts are simply positive or negative. Rather, they suggest that their lexical contexts are more closely related to processes of elaboration, interpretation, symbolization, and metapsychological organization than to directly clinical-negative markers.

The keyword-by-emotion enrichment analysis reinforced this interpretation by identifying specific contextual associations (Fig. 3B). Terms such as "symptom", "repression", "hysteria", and "anxiety" showed enrichment in negative affective categories, including fear, sadness, disgust, and negative polarity. These findings suggest that certain clinical concepts occupy affectively dense regions of the corpus, functioning as points of emotional concentration within the Freudian vocabulary. Enrichment analysis complements the correlation results by indicating not only longitudinal associations between keyword frequency and emotion categories, but also the contextual concentration of affective categories around specific psychoanalytic keywords.

Principal component analysis of keywords in the affective space revealed a clear multivariate organization (Fig. 3C). The first two principal components explained 84.2% of the total variance, with PC1 accounting for 64.9% and PC2 for 19.3%. PC1 primarily separated clinical-symptomatic terms, such as "symptom", "neurosis", "hysteria", and "anxiety", from terms more closely associated with dream-interpretive, elaborative, and metapsychological fields, such as "dream", "desire", "censorship", "pleasure", and "unconscious". PC2 contributed to the differentiation of specific affective components, including disgust, anger, fear, anticipation, and positive polarity.

Taken together, Fig. 3 indicates that Freudian concepts are not randomly distributed in the affective space. Instead, they form clusters compatible with clinical-symptomatic, dream-interpretive, metapsychological, and affective fields. This organization suggests that Freud's work contains a

measurable affective-conceptual structure that can be investigated through text mining and lexicon-based emotion analysis.

(A) Correlation map between conceptual keywords and NRC emotion categories across the volumes. Correlations were estimated using Spearman's rho, based on the relative frequency of keywords and the proportional representation of NRC emotion categories by volume. Red tones indicate positive correlations, blue tones indicate negative correlations, and values close to white indicate weak or absent monotonic association. (B) Keyword-by-emotion NRC enrichment analysis. Point size represents the magnitude of emotional hits associated with the volumes in which the keyword is present, whereas color intensity represents $-\log_{10}(\text{FDR})$ after correction for multiple testing. Only enriched and statistically significant associations are shown. (C) PCA of keywords in the NRC affective space. Black points represent conceptual keywords, and red vectors represent the NRC categories that most contribute to the structure of the reduced space. The first principal component mainly separates clinical-symptomatic terms, such as symptom, neurosis, hysteria, and anxiety, from terms more closely associated with dream-related and elaborative fields, such as dream, desire, censorship, pleasure, and unconscious. The second component contributes to distinguishing specific affective axes, including disgust, anger, fear, anticipation, and positive polarity. Taken together, the panels indicate that psychoanalytic concepts are not randomly distributed in the affective space of the corpus, but form clusters compatible with distinct clinical, affective, and metapsychological fields.

Volumes are organized into a longitudinal affective cartography

Principal component analysis was performed on the editorial volumes using only the NRC affective profile as the input matrix. The first two principal components explained 75.1% of the emotional variation across volumes, with 53.9% attributed to PC1 and 21.2% to PC2. This high proportion of explained variance indicates that much of the lexicon-based affective variation in the collection can be summarized by a limited number of principal axes.

PC1 separated volumes more strongly associated with fear, sadness, and negative polarity from those closer to surprise, joy, positive polarity, and trust. This axis can be interpreted as a broad affective gradient, distinguishing volumes with greater negative lexical load from those with stronger contributions from positive or elaborative categories. PC2 mainly distinguished volumes oriented toward anger and disgust from those more influenced by anticipation and positive polarity.

Thus, the distribution of volumes in the PCA space suggests that the affective organization of the collection is not homogeneous. Instead, the volumes are structured by lexicon-based affective gradients that distinguish different thematic and editorial moments in Freud's work.

Principal component analysis of the volumes of Freud's work using only the normalized proportions of NRC Emotion Lexicon categories as input variables. Each point represents one volume, and the red vectors indicate the direction and contribution of the affective categories to the ordering of volumes in the reduced space. The first principal component explained 53.9% of the variance, whereas the second

explained 21.2%, totaling 75.1% of the affective variation represented by the first two axes. PC1 separates volumes with greater contributions from fear, sadness, negative polarity, and lower trust from those more closely associated with surprise, joy, positive polarity, and trust. PC2 mainly distinguishes volumes oriented toward anger and disgust from those influenced by anticipation and positive polarity. The distribution of volumes suggests that the lexicon-based affective organization of the collection is not homogeneous, but structured by gradients that distinguish different thematic and editorial moments in Freud's work. Because the unit of analysis corresponds to editorial volumes rather than individual texts, essays, or chapters within each volume, this PCA should be interpreted as an exploratory cartography of the affective organization of the collection.

Discussion

The results of this study show that text mining can function as a complementary strategy for investigating the internal organization of large psychoanalytic corpora. By analyzing the 23 volumes of Freud's Complete Works in Portuguese, it was possible to identify lexical, conceptual, and affective patterns that do not replace hermeneutic or clinical reading, but expand its scale of observation. This perspective is aligned with distant reading, in which large textual collections are analyzed to reveal structural patterns that exceed the close reading of isolated works, and with literary macroanalysis, which proposes the use of computational methods to investigate regularities distributed across extensive corpora (Moretti, 2013; Jockers, 2013). At the same time, the study is situated within digital humanities by treating computational tools not as substitutes for interpretation, but as instruments capable of reconfiguring the relationship between reading, scale, and textual evidence (Hayles, 2012).

In this sense, the main contribution of the present study is not merely the application of quantitative techniques to a classical corpus, but the demonstration that Freud's work, as represented in the Portuguese edition analyzed here, exhibits a longitudinal lexical-affective organization. Its central concepts vary across volumes, are associated with specific affective fields, and contribute to the formation of distinct textual affective signatures within the collection. These findings position text mining as a methodological tool capable of engaging with the interpretive tradition of psychoanalysis by offering an exploratory, reproducible, and comparative layer for the study of the textual development of Freudian concepts.

The first axis of results indicates that the computational approach was able to recover a lexical architecture compatible with the classical conceptual organization of Freud's work. The emergence of terms and expressions related to dream, sexuality, neurosis, symptom, childhood, unconscious, and mental life suggests that the corpus was not structured merely around generic lexical frequency, but around recognizable theoretical nuclei of psychoanalysis. This point is methodologically important because it functions as a form of internal validation of the analysis. Before supporting more complex affective interpretations, the textual procedure recovered semantic fields expected from the Freudian tradition, particularly those related to dream theory, infantile sexuality, the clinic of neuroses, and

unconscious processes (Freud, 1900/1953; Freud, 1905/1953; Freud, 1915/1957; Laplanche & Pontalis, 1973).

At the same time, the longitudinal distribution of keywords shows that these concepts do not appear uniformly across the collection, but intensify at different editorial moments, suggesting shifts in lexical emphasis throughout the work. This reading is consistent with the understanding of Freud's writings as a conceptual system in transformation, in which notions such as dream, desire, repression, unconscious, symptom, and sexuality are continually reworked in relation to distinct clinical and metapsychological problems (Ricoeur, 1970; Lear, 2005). Thus, lexical analysis not only confirms the thematic coherence of the corpus, but also provides quantitative evidence that Freud's production can be read as a conceptual trajectory in motion, in which particular vocabularies become more or less salient according to the clinical, metapsychological, or interpretive focus of the volumes.

The second axis of results expands this lexical reading by showing that Freud's work is organized not only by conceptual recurrence, but also by systematic variation in lexical affective tonality. The combination of lexical polarity and NRC Emotion Lexicon categories makes it possible to move beyond a simple opposition between positive and negative terms toward a more complex representation of textual affectivity, involving trust, anticipation, fear, sadness, anger, disgust, joy, and surprise. This distinction is relevant because sentiment and emotion lexicons do not measure the psychological states of the author, but rather semantic and affective associations encoded in the vocabulary used in the text (Pang & Lee, 2008; Taboada et al., 2011; Mohammad & Turney, 2013). Accordingly, the oscillations observed across the volumes should be interpreted as changes in the affective profile of the discourse, not as direct inferences about Freud's emotional life.

The alternating presence of negative, positive, and anticipatory categories is particularly coherent with a corpus devoted to the elaboration of psychic conflict, suffering, desire, symptom formation, and interpretation. Thus, lexicon-based affective analysis does not add only a descriptive layer to the study. It also provides evidence that conceptual displacements in the work are accompanied by changes in the affective texture of the vocabulary, reinforcing the hypothesis of a longitudinal lexical-affective trajectory.

The third axis of results deepens this cartography by showing that psychoanalytic concepts not only vary in frequency, but also differ in the type of affective vocabulary they mobilize. The association between keywords and NRC categories suggests that clinical terms such as symptom, neurosis, hysteria, and anxiety are closer to an affective field marked by fear, sadness, negativity, and disgust, whereas terms such as dream, desire, pleasure, censorship, unconscious, and reality occupy regions closer to interpretive, elaborative, and metapsychological dimensions. This separation is theoretically relevant because Freudian psychoanalysis does not treat affect as an accessory element of psychic life, but as a central component of symptom formation, conflict, repression, and unconscious elaboration (Freud, 1915/1957; Freud, 1926/1959; Green, 1999).

At the same time, the proximity of dream, desire, and unconscious to less directly clinical-negative affective fields is consistent with the role of these concepts in the construction of a theory of

interpretation, in which dreams and related formations function as privileged pathways to desire and unconscious processes (Freud, 1900/1953; Laplanche & Pontalis, 1973). The central contribution of this result is therefore not to assign fixed emotions to isolated concepts, but to show that, within the corpus analyzed, different Freudian notions cluster according to specific affective-semantic profiles. This organization reinforces the hypothesis that Freud's work contains an internal lexical-affective structure, in which clinical-symptomatic, dream-interpretive, and metapsychological concepts are distinguished not only by their theoretical function, but also by the affective textual fields in which they appear.

The fourth axis of results shifts the analysis from the level of concepts to the macrostructural level of volumes, showing that the collection presents a measurable affective organization at the editorial scale. The PCA based exclusively on NRC affective categories showed that the volumes are not randomly distributed in the multivariate space, but are ordered along affective gradients that distinguish sets of texts closer to fear, sadness, and negative polarity from those more associated with trust, positive polarity, joy, surprise, and anticipation. This interpretation is consistent with the use of principal component analysis as an exploratory technique for reducing the dimensionality of complex data, preserving as much variation as possible, and revealing latent structures among observations described by multiple correlated variables (Abdi & Williams, 2010; Jolliffe & Cadima, 2016).

In the context of this study, the volumes function as comparable documentary units, and their position in the affective space indicates similarities and differences in the lexical-affective composition of the collection. Therefore, PCA should not be interpreted as definitive proof of rigid emotional periods in Freud's work, but as evidence that different editorial moments share textual affective signatures. This finding reinforces the central argument of the article: Freud's work can be described not only as a chronological sequence of texts, but also as a lexical-affective cartography in transformation, in which concepts, affective tonalities, and volumes are reorganized in a structurally detectable way.

Several methodological limitations must be considered when interpreting these findings. First, lexicon-based analyses depend on the coverage and predefined categories of the dictionaries used. For this reason, they do not necessarily capture irony, ambivalence, polysemy, negation, technical shifts in meaning, or the specific conceptual value that certain words assume within a theoretical tradition (Pang & Lee, 2008; Taboada et al., 2011). This point is particularly important in a psychoanalytic corpus, where terms such as desire, repression, censorship, unconscious, pleasure, or anxiety have conceptual density that exceeds their ordinary lexical use. The affective results of this study should therefore be understood as affective-semantic profiles of the vocabulary, not as direct measures of the author's affects, clinical intentions, or final psychoanalytic meanings.

Second, the study analyzed the Portuguese edition of Freud's Complete Works published by Editora Imago, rather than the German originals. This means that the lexical and affective patterns identified here are mediated by translation, editorial decisions, terminology choices, and the semantic structure of Portuguese. The results should therefore be interpreted as a computational cartography of the Portuguese textual corpus analyzed, not as a direct reconstruction of Freud's original German wording.

This limitation does not invalidate the analysis, since translated corpora are themselves historically influential objects of reception and interpretation. However, it restricts the degree to which the findings can be generalized to Freud's original language. Future studies should compare the Portuguese corpus with German originals and other translations in order to assess which lexical-affective patterns are stable across languages and which are translation-dependent.

Third, the use of editorial volumes as documentary units allows a broad longitudinal view, but may smooth internal variation among chapters, essays, and texts written at different moments. Some volumes contain texts from multiple years or heterogeneous theoretical contexts, and their treatment as a single unit may obscure finer conceptual shifts. To reduce these risks, the study adopted conservative textual curation, z-score normalization within each affective category, FDR correction in enrichment analyses, and a triangulation strategy combining independent methods, including lexical frequency, n-grams, directed networks, polarity analysis, NRC categories, correlations, contextual enrichment, and PCA. The convergence among these analytical layers does not eliminate the limitations, but reduces the likelihood that the observed patterns result only from lexical noise, scale bias, or the isolated effect of a single method.

For these reasons, the strength of this study does not lie in offering a definitive interpretation of the theoretical evolution of Freud's work. Rather, it demonstrates that there is a recurrent, measurable, and methodologically traceable lexical-affective structure that can guide future analyses at more granular levels, including individual texts, chapters, chronological periods, original-language corpora, translations, or comparisons with other psychoanalytic authors.

Despite its exploratory character, this study offers a methodological and interpretive contribution to digital humanities approaches to psychoanalysis. By combining distant reading, text mining, and affective-semantic analysis, the proposed approach does not reduce Freud's work to frequencies or automatic classifications. Instead, it creates an additional layer of evidence capable of orienting new interpretive questions. The results can be understood as a starting point for an augmented reading of Freud's work: computational analysis identifies patterns, proximities, and displacements at scale, while psychoanalytic interpretation remains necessary to examine the conceptual density of each term, text, or period (Moretti, 2013; Jockers, 2013; Hayles, 2012).

The main implication of the study is that Freudian concepts do not merely circulate across the volumes, but are distributed within organized lexical and affective fields. This makes it possible to formulate more refined hypotheses about the relationship between language, affect, and theoretical construction in psychoanalysis. Future studies may test this cartography using more granular textual units, such as chapters, essays, or chronological periods, and may compare it across authors, languages, translations, or psychoanalytic traditions. Thus, the contribution of this article is not to close the interpretation of Freud's work, but to show that large psychoanalytic corpora can be investigated in a reproducible, comparative, and theoretically oriented way, opening a promising agenda for dialogue between psychoanalysis, computational textual analysis, data mining, and digital humanities.

Conclusion

The findings suggest that Freud's work, as represented in the Portuguese edition analyzed in this study, is characterized not only by the recurrence of central psychoanalytic concepts, but also by a longitudinal reorganization of lexical, conceptual, and affective fields. Psychoanalytic keywords varied across the volumes, were differentially associated with NRC Emotion Lexicon categories, and clustered in affective spaces compatible with clinical-symptomatic, dream-interpretive, and metapsychological fields. In addition, the PCA of editorial volumes indicated that different parts of the collection share comparable lexical-affective signatures, suggesting that the textual trajectory of Freud's work can be described as a computational cartography in transformation. Computational analysis therefore does not replace theoretical or clinical interpretation in psychoanalysis. Rather, it provides an additional layer of quantitative evidence that helps map how concepts and affective vocabularies are articulated throughout Freud's production. By doing so, this study opens a methodological path for the large-scale, reproducible, and comparative analysis of psychoanalytic corpora within digital humanities and computational textual studies.

Declarations

Funding.

No funding was received for conducting this study.

Author Contribution

K.F. and A.F. wrote the main manuscript text. All authors reviewed the manuscript.

Acknowledgement

This work is dedicated to the memory of Kazuo Ferreira Fukutani.

References

1. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, 2(4), 433–459.
2. Aydin, C., & Possati, L. M. (2025). Less and more than data: a Lacanian inquiry into self-formation in the age of data mining. *AI & SOCIETY*, 40(8), 6123–6134
3. Assoun, P.-L. (1983). *Introdução à epistemologia freudiana*. Rio de Janeiro: Imago.
4. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.

5. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
6. Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder: Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18), 3710–3715.
7. Carvalho, P., & Silva, M. J. (2015). SentiLex-PT: Principais características e potencialidades. *Oslo Studies in Language*, 7(1), 425–438.
8. Freud, S. (1953). The interpretation of dreams. In J. Strachey (Ed. & Trans.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vols. 4–5). London: The Hogarth Press. Original work published 1900.
9. Freud, S. (1953). Three essays on the theory of sexuality. In J. Strachey (Ed. & Trans.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 7, pp. 123–246). London: The Hogarth Press. Original work published 1905.
10. Freud, S. (1957a). Repression. In J. Strachey (Ed. & Trans.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 14, pp. 141–158). London: The Hogarth Press. Original work published 1915.
11. Freud, S. (1957b). The unconscious. In J. Strachey (Ed. & Trans.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 14, pp. 159–215). London: The Hogarth Press. Original work published 1915.
12. Freud, S. (1959). Inhibitions, symptoms and anxiety. In J. Strachey (Ed. & Trans.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 20, pp. 75–176). London: The Hogarth Press. Original work published 1926.
13. Garcia-Roza, L. A. (2008). *Freud e o inconsciente* (24. ed.). Rio de Janeiro: Jorge Zahar.
14. Green, A. (1999). *The fabric of affect in the psychoanalytic discourse*. London: Routledge.
15. Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. Chicago: University of Chicago Press.
16. Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), 1–13.
17. Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.
18. Jockers, M. L. (2015). Syuzhet: Extracts sentiment and sentiment-derived plot arcs from text. R package.
19. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.

20. Laplanche, J., & Pontalis, J.-B. (1973). *The language of psycho-analysis*. London: The Hogarth Press.
21. Lear, J. (2005). *Freud*. New York: Routledge.
22. Meunier, J.-G. (2015). *L'analyse textuelle assistée par ordinateur (L'ATAO)*. Montréal: Presses de l'Université du Québec.
23. Mezan, R. (2014). *Freud: pensador da cultura* (8. ed.). São Paulo: Companhia das Letras.
24. Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, motivation, and emotion. In *Oxford Handbook of Affective Computing* (pp. 1–15). Oxford: Oxford University Press.
25. Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26–34).
26. Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.
27. Moretti, F. (2013). *Distant reading*. London: Verso.
28. Ooms, J. (2026). *pdftools: Text extraction, rendering and converting of PDF documents*. R package.
29. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1-135.
30. Ricoeur, P. (1970). *Freud and philosophy: An essay on interpretation*. New Haven: Yale University Press.
31. Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3), 37.
32. Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. Sebastopol: O'Reilly Media.
33. Silva, M. J., Carvalho, P., Costa, C., & Sarmiento, L. (2010). Automatic expansion of a social judgment lexicon for sentiment analysis. Technical Report TR 10 – 08. University of Lisbon, Faculty of Sciences, LASIGE.
34. Silva, M. J., Carvalho, P., & Sarmiento, L. (2012). Building a sentiment lexicon for social judgement mining. In *Computational Processing of the Portuguese Language, PROPOR 2012. Lecture Notes in Computer Science*. Berlin: Springer.
35. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307.

Figures

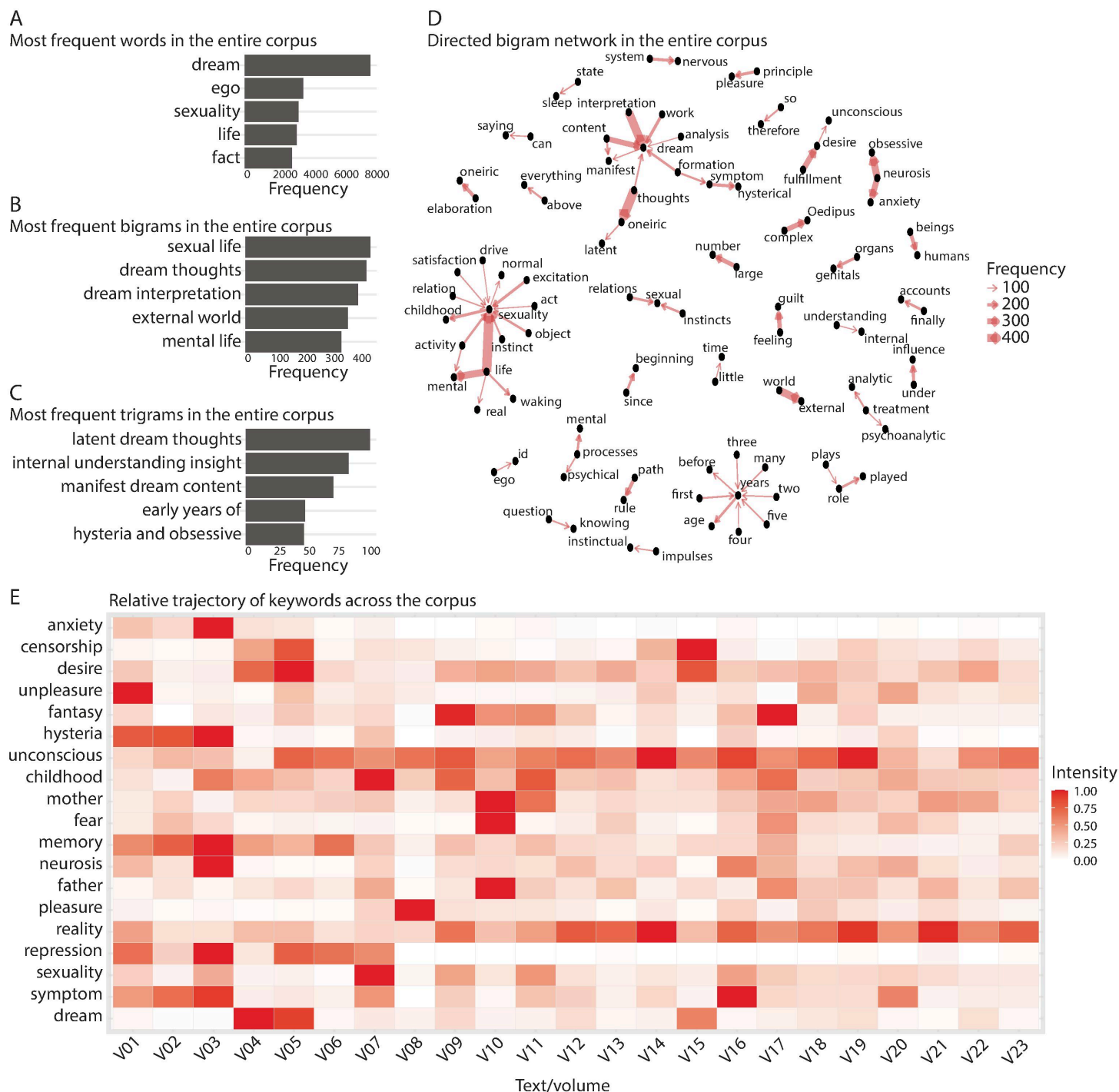


Figure 1

Global lexical profile and longitudinal distribution of keywords in Freud's work.

(A) Most frequent words in the complete corpus after textual curation, removal of editorial paratexts, and lexical filtering. (B) Most frequent bigrams identified across the corpus, representing recurrent lexical associations between pairs of consecutive terms. (C) Most frequent trigrams, highlighting compound expressions and recurrent semantic units in the corpus. (D) Directed network of the most frequent bigrams, in which nodes represent words and edges indicate the order of occurrence between consecutive terms; edge thickness is proportional to bigram frequency. (E) Relative trajectory of

conceptual keywords across the volumes after grouping lexical variants by concept. Each cell represents the relative intensity of a keyword in a given volume, scaled from 0 to 1 within each concept, allowing the identification of thematic concentration peaks throughout the work. The analysis highlights the global predominance of terms related to dreams, sexuality, psychic life, symptoms, neurosis, and interpretive processes, while also revealing longitudinal oscillations in the distribution of the main psychoanalytic concepts.

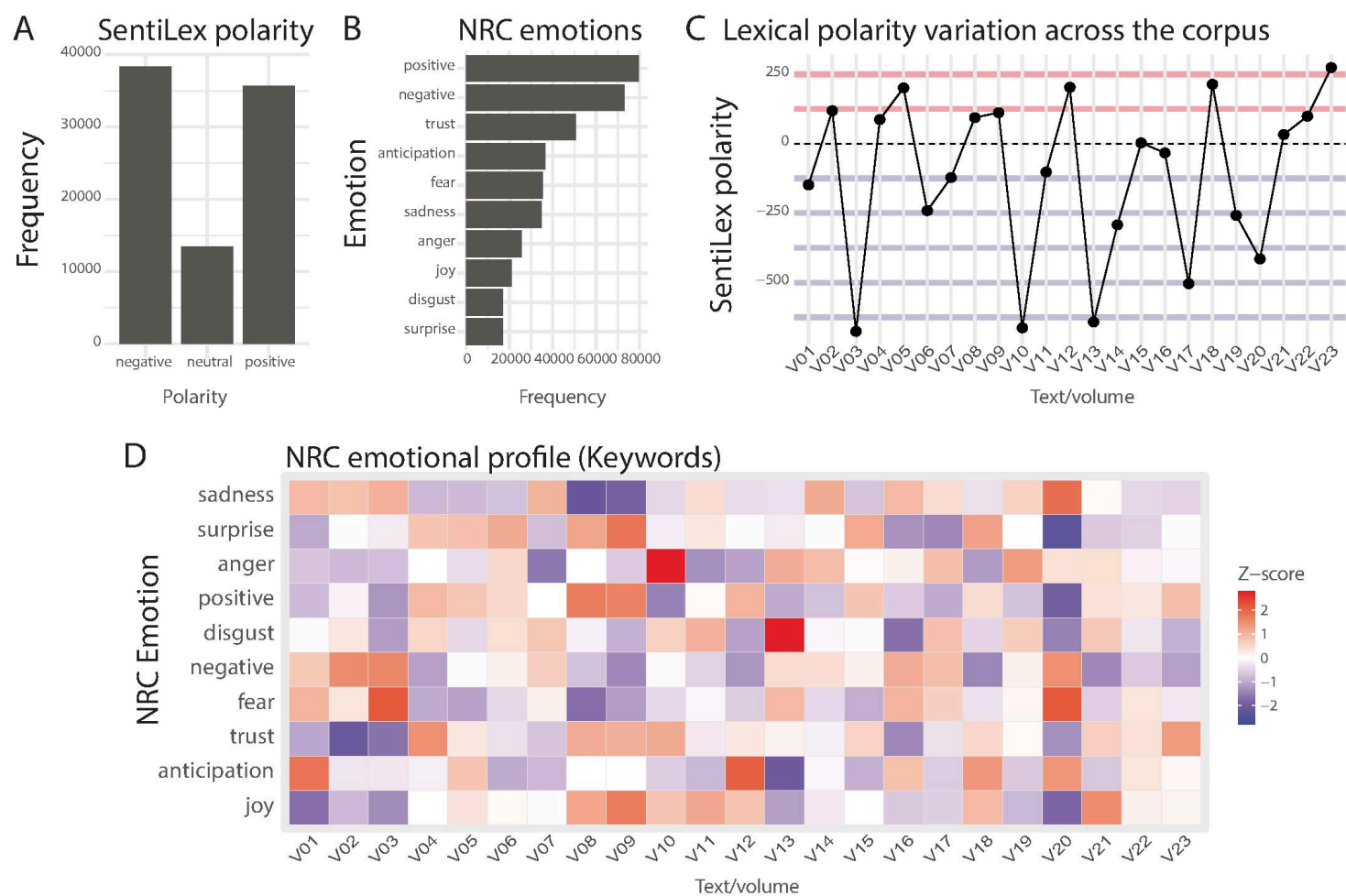


Figure 2

Lexical polarity and NRC emotional profile across Freud's work.

(A) Global distribution of lexical polarity according to SentiLex, classified as positive, negative, and neutral terms. (B) Global frequency of NRC Emotion Lexicon categories in the complete corpus. (C) Longitudinal variation in lexical polarity across the volumes, calculated as the difference between the number of positive and negative terms in each volume. The dashed horizontal line indicates balance between positive and negative polarity; values above zero indicate positive predominance, whereas values below zero indicate negative predominance. (D) NRC emotional profile across the volumes, normalized by z-score within each emotion category. Positive values indicate volumes with a higher relative proportion of a given emotion compared with the mean of that same emotion across the corpus, whereas negative values indicate lower relative representation. This normalization reduces the effect of

globally abundant categories and highlights volume-specific variation in each affective category across the editorial sequence. The results indicate that Freud’s work does not show a uniform affective predominance, but rather alternating patterns of positivity, trust, anticipation, fear, sadness, anger, negativity, and other NRC categories across the corpus.

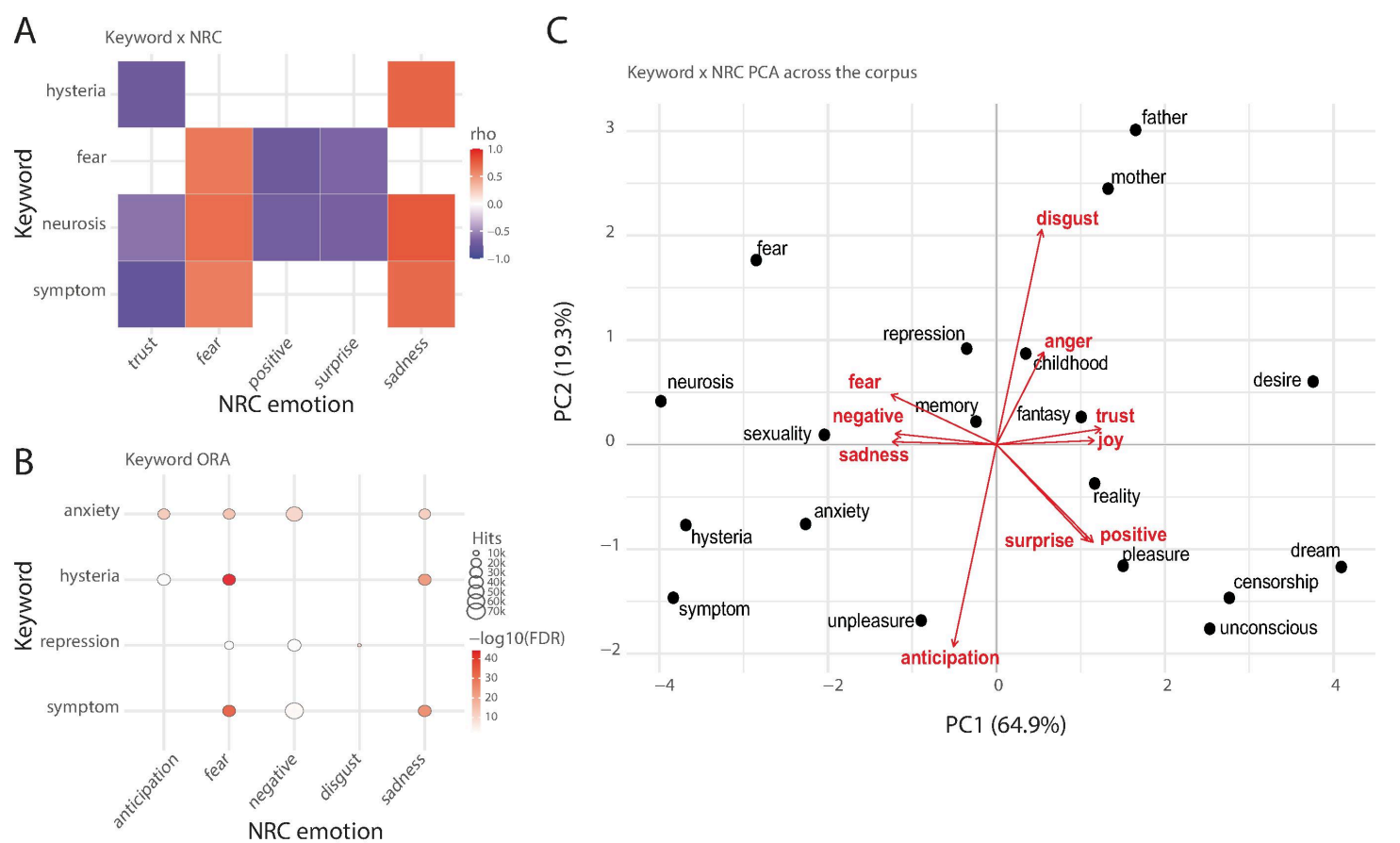


Figure 3

Association between psychoanalytic keywords and NRC emotion categories.

(A) Correlation map between conceptual keywords and NRC emotion categories across the volumes. Correlations were estimated using Spearman’s rho, based on the relative frequency of keywords and the proportional representation of NRC emotion categories by volume. Red tones indicate positive correlations, blue tones indicate negative correlations, and values close to white indicate weak or absent monotonic association. (B) Keyword-by-emotion NRC enrichment analysis. Point size represents the magnitude of emotional hits associated with the volumes in which the keyword is present, whereas color intensity represents $-\log_{10}(\text{FDR})$ after correction for multiple testing. Only enriched and statistically significant associations are shown. (C) PCA of keywords in the NRC affective space. Black points represent conceptual keywords, and red vectors represent the NRC categories that most contribute to the structure of the reduced space. The first principal component mainly separates clinical-symptomatic terms, such as symptom, neurosis, hysteria, and anxiety, from terms more closely associated with dream-related and elaborative fields, such as dream, desire, censorship, pleasure, and unconscious. The second component contributes to distinguishing specific affective axes, including disgust, anger, fear,

anticipation, and positive polarity. Taken together, the panels indicate that psychoanalytic concepts are not randomly distributed in the affective space of the corpus, but form clusters compatible with distinct clinical, affective, and metapsychological fields.

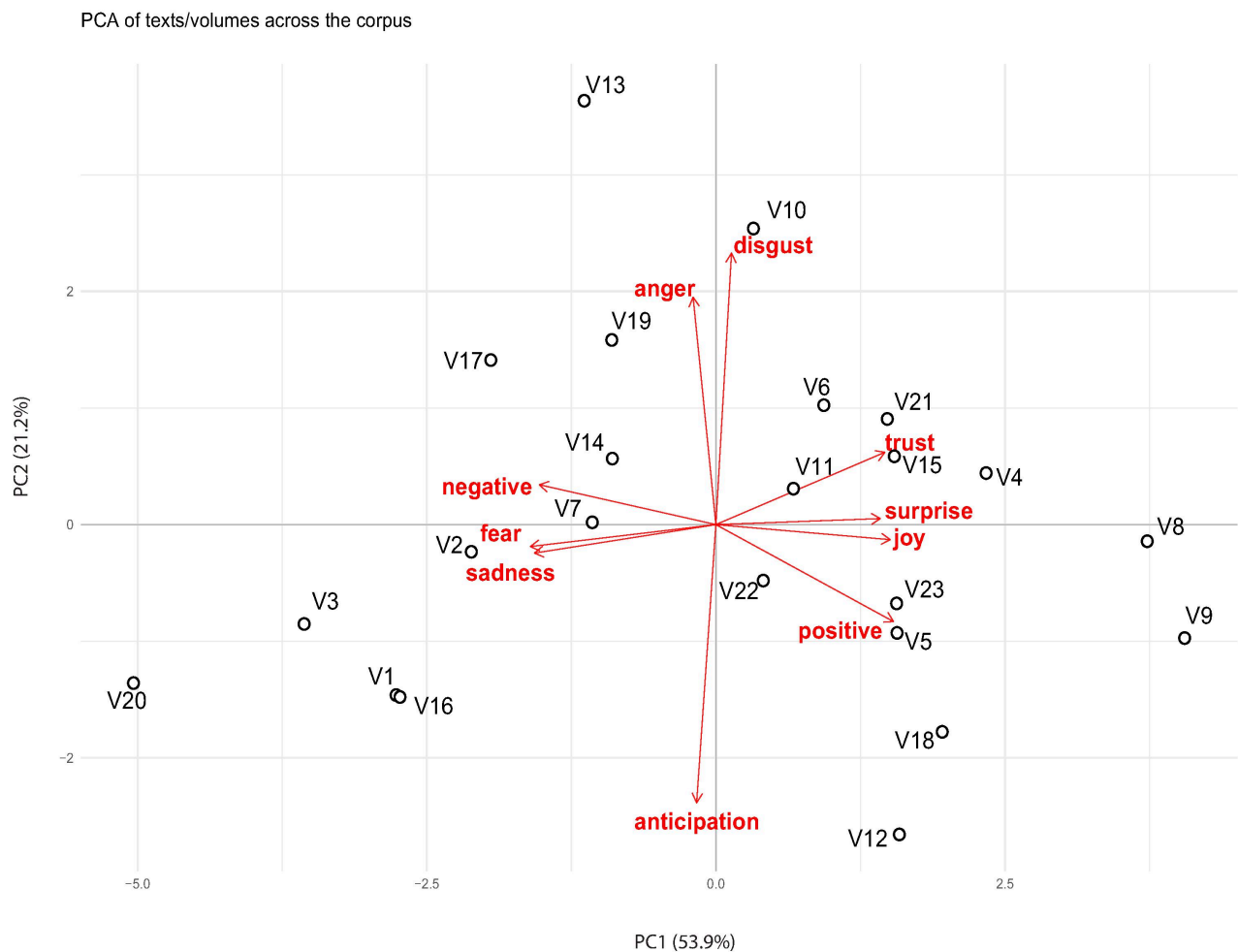


Figure 4

PCA of volumes based exclusively on the NRC affective profile.

Principal component analysis of the volumes of Freud’s work using only the normalized proportions of NRC Emotion Lexicon categories as input variables. Each point represents one volume, and the red vectors indicate the direction and contribution of the affective categories to the ordering of volumes in the reduced space. The first principal component explained 53.9% of the variance, whereas the second explained 21.2%, totaling 75.1% of the affective variation represented by the first two axes. PC1 separates volumes with greater contributions from fear, sadness, negative polarity, and lower trust from those more closely associated with surprise, joy, positive polarity, and trust. PC2 mainly distinguishes volumes oriented toward anger and disgust from those influenced by anticipation and positive polarity. The distribution of volumes suggests that the lexicon-based affective organization of the collection is not homogeneous, but structured by gradients that distinguish different thematic and editorial moments in Freud’s work. Because the unit of analysis corresponds to editorial volumes rather than individual

texts, essays, or chapters within each volume, this PCA should be interpreted as an exploratory cartography of the affective organization of the collection.