

# Extraction of Key Concept Relevance Graphs From Fourteen Decades of Psychoanalytic Journal Publications

Sheryl Brahnham\*  
Information Technology  
& Cybersecurity  
Missouri State University  
Springfield MO, USA  
sbrahnham@missouristate.edu

Rick Brattin  
Information Technology  
& Cybersecurity  
Missouri State University  
Springfield MO, USA  
RickBrattin@MissouriState.edu

Andrew Crofford  
Information Technology  
& Cybersecurity  
Missouri State University  
Springfield MO, USA  
crofford000@live.missouristate.edu

Justin Freres  
Information Technology  
& Cybersecurity  
Missouri State University  
Springfield MO, USA  
Freres75@live.missouristate.edu

**Abstract**— Automatically tracing the history of concepts and terms by text mining diachronic corpora is a new multidisciplinary area of research. An essential step in the pipeline of such research is calculating the relevance of terms, with most work relying on simple frequency measures. In this study, we extract relevance graphs of key psychoanalytic concepts from a corpus built for this task, one that represents the entire history of psychoanalysis. Different measures of relevance are examined and shown to tell a different story; when combined, however, they provide a clearer picture of a term's significance over time.

**Keywords**—term relevance, text mining, frequency, TF-IDF, psychoanalytic concepts, PEP-Web, glossary building

"Full/Regular Research Paper" CSCI-ISBD

## I. INTRODUCTION

Recently, studies across several fields (computer science, engineering, science, and the digital humanities) have demonstrated the value of text mining literary and scientific corpora for the purposes of tracing the history of important concepts and of constructing historical narratives of different disciplines. One of the challenges posed by assembling such histories is figuring out how concepts emerge and change over time [1].

To identify topic changes in artificial intelligence, Zhang et al. [2] conducted a topic-based bibliometric study in the journal *Knowledge-Based Systems* (KnoSys) from 1991 to 2016. The authors produced a Latent Dirichlet Allocation model to profile areas of activity in KnoSys and to predict possible future trends. Similarly, Faust [3] quantified changes and diversification in topics covered in the journal *Computers in Biology and Medicine* (CBM) by extracting all author keywords listed in 1990 and mapping their changes by normalizing the number of appearances for a specific keyword in yearly intervals to 2017. Results were visualized with graphs containing trend lines. Trevisani and Tuzzi [4] constructed a diachronic corpus to track the history of concepts in several scientific fields. Their methods involved matching n-grams with glossary entries and applying normalization of temporal trajectories to raw frequencies. Of

interest here is the differences in history readings that were made possible depending on the type of data normalization methods used to account for the fluctuating size of texts across time. Finally, in the digital humanities, Gavin [1] analyzed changes in concepts on the EEBOText Creation Partnership corpus dated from 1640-1699 using semantic analysis.

In each of the studies mentioned above, calculating the relevance of terms was an essential step in the pipeline. In general, relevance was measured with term frequency (often normalized) or by counting the number of documents containing a term. In this study, we track changes in concepts using both these methods, but we also explore and compare several measures of relevance.

The primary goal of this study, however, is to map the relevance of several key psychoanalytic concepts across the entire history of psychoanalysis, which began in the 1880s with Sigmund Freud's early publications. This discipline was chosen for several reasons. First, psychoanalytic concepts and terms have permeated modern culture and have had a major impact on the humanities, the arts, and psychology. Most people living in the West today are familiar with such concepts as the Oedipus complex, castration, dream analysis, the ego, and the oral and anal developmental stages in early childhood. Second, psychoanalysis, contrary to popular belief, is alive and well as a therapeutic modality, with many journals currently dedicated to psychoanalytic practice and research. Finally, there exists a huge digital archive of psychoanalytic writings on the Psychoanalytic Electronic Publishing Website (PEP-Web).

PEP-Web is a goldmine of research opportunity, valuable not only to clinicians and researchers in the field of psychoanalysis but also potentially to many others in such diverse fields as sociology, history, linguistics, and the digital humanities. The value of PEP-Web is that it is an online archive that intends to hold the entire psychoanalytic literature (which now spans 149 years), according to the articles "About the PEP Archive" and "History of PEP" on <https://www.pep-web.org/>. The idea for a searchable archive of the psychoanalytic literature was first proposed by Paul Mosher in 1991. Interest snowballed,

and a searchable CD archive was actualized in 1998, funded with startup loans from the American Psychoanalytic Association (APsaA) and the Institute of Psychoanalysis (London). The CD archive was transferred to PEP-Web in 2002 and is now funded, on a subscription basis, by over 100 universities and by many individuals and public agencies worldwide. As of 2020, the PEP archive is said to contain the entire contents of seventy-eight psychoanalytic journals, totaling 122,000 articles, as well as one hundred book classics in psychoanalysis. There are few archives, if any, in other disciplines that cover as much of the literature as PEP-Web does across the entire history of psychoanalysis.

Carrying out the research goals of this project required that a diachronic corpus be built, as described in Section II, based on the English documents contained in the PEP-Web archive. Three measures of relevance, detailed in Section III, were then applied to sixteen psychoanalytic concepts selected by comparing their rankings in terms of the number of documents containing the terms on PEP-Web (a rough measure of relevance) and the number of times seven psychoanalysts included the terms in their lists of the top ten all time psychoanalytic concepts. In Section IV, relevance graphs for four concepts (*ego*, *repression*, *transference*, and *unconscious*) were extracted that show the vicissitudes of these concepts across ten decades, and differences in readings are provided for two measures. In addition, sparklines for both measures are displayed for all sixteen terms from 1880-2020. In Section V, we suggest a couple of ways these visualizations can be applied.

## II. CORPUS CONSTRUCTION AND DESCRIPTION

Given that this study's goal is to map out changes in the relevance of some important psychoanalytic concepts by data mining English texts produced in this field across most of its history, the first step was to build a corpus suitable to the task. An often-quoted truism by Biber et al. [5] is that a corpus is "not simply a collection of texts" (246) but instead requires a carefully planned process for obtaining texts appropriate to research goals and for arranging, storing, and cataloging texts once so gathered [6].

As noted in the Introduction, the corpus built for this study is based on the PEP-Web archive. Because the size of this archive is over 122,000 documents, web crawling robots were programmed for automatically analyzing all the available data on PEP-Web. A valid subscription was obtained that provided full access to the website. According to PEP-Web's robot.txt file, robots are only prohibited from accessing the root directory, so systematically analyzing documents via a robot is allowed. All data and raw text extracted from PEP-Web and processed were solely used for this research project.

PEP-Web assigns each of its documents a unique identifier, or ID (e.g., *ijp.027.0099a*). The first group of letters composing an ID (located before the first period) provides information regarding the journal/book/video source of the document, which proved useful in determining the document's language. Thus, the problem of systematically accessing all English texts on PEP-Web boiled down to collecting all the relevant document IDs. Searching for common English words, such as *the*, using the PEP-Web search engine and extracting document IDs from the

results proved futile because all searches were limited to a listing of 5,000 sources.

Two possible methods were eventually discovered for extracting document IDs from the archive: 1) by going through each journal, book, and video collection and 2) by clicking on each author on PEP-Web's author list to access their publications. The simplest method was to use the author list, which provided the added advantage of listing all authors in a standard format so that author and co-author first and last names could be extracted and stored in a database for future analysis along with links to their writings. It was discovered by searching for some concepts via the PEP-Web search engine that not all documents in the archive were associated with an author. Inspection of these orphan articles revealed that most of these works were reports from various psychoanalytic organizations, task forces, and committees. As these documents offered little value for psychoanalytic concept analysis, the loss of these documents was considered acceptable.

A database was created to store the names of all authors and the document IDs of their works, along with other information, such as document URLs, the full reference of each document, date of publication, the journal's name and unique identifier, as well as the language of the journal (the language of the seventy-eight journals was inspected manually). A total of 30,302 authors were collected from the author list, along with 81,044 document IDs.

The next step in the process of creating the corpus required that all document URLs collected from the author list and located in English sources be visited and scrapped. Each HTML webpage was processed by removing all HTML and java code. The text was extracted in UTF-8 format and cleaned by removing document headings, table of contents, all references and bibliographies, footnotes, notes, text markups (indicating bold, italics, etc.), and all new paragraph and new-line markers. Short text files less than 500 bytes (~75 words or less) were ignored as they mostly contained one or two sentences expressing the general theme of a recently published book or an obituary announcement. Also culled were books of author correspondences and all articles still in embargo (a waiting period of three to five years after publication before becoming fully accessible on PEP-Web—until then only the abstract is provided).

In this way, a final corpus containing 71,609 texts in English was produced. With numbers and punctuation removed and contractions expanded, the resulting PEP-Web corpus (referred to henceforth as PEP-Web-20, so named because documents were accessed on PEP-Web in February of 2020) contains a total of 285,865,003 running words, with 454,759 types (unique words).

PEP-Web-20 was further organized like the Brown corpus, which contains a file (*cat.txt*) listing each document's file name and category membership (adventure, belles lettres, etc.). Because the goal of this project is to explore changes in relevance over time, the PEP-Web-20 *cat.txt* file lists the decade associated with each document, starting in the 1880s and extending to the end of the 2010s. Thus, each document falls into one of fourteen categories.

TABLE I. PEP-WEB-20 CATEGORIES: NUMBER OF DOCUMENTS PER DECADE AND OTHER STATISTICS

Decade	Document Count	Word Count	Average Tokens Per Document	Word Count (Analysis)*
1880	7	30034	4290.57	29873
1890	69	260331	3772.91	259028
1900	53	731457	13801.08	726439
1910	492	1934352	3931.61	1924570
1920	1889	4419562	2339.63	4394523
1930	2091	6096685	2915.68	6060015
1940	3185	6949485	2181.94	6906526
1950	4097	13017829	3177.41	12927688
1960	4579	17584210	3840.19	17451254
1970	5780	25343092	4384.62	25110210
1980	9158	38575708	4212.24	38170328
1990	14575	55508884	3808.50	54858041
2000	14615	66265868	4534.10	65497156
2010	11019	49147506	4460.25	48555933

\* See IV.B-C for details on token count per decade in text used in the analysis.

As evident in Table 1, categories are not balanced. Changing the timespan of each category from ten to five years was considered but rejected because statistical measures, as noted in section III, can be applied that offset imbalances in the number of documents contained within a sub-corpora or category. It will also be observed that the number of publications in psychoanalysis has risen significantly, at least until the 2000s, when activity plateaus. The drop that follows in the teens is most likely due to the embargo period of three to five years imposed on articles included in the PEP-Web archive.

### III. MEASURES OF TERM RELEVANCE

Determining the relevance of a term across a corpus subdivided into unbalanced categories requires that statistical measures be applied that remove bias based on variations in the number of documents within each category/sub-corpora. An obvious candidate for expressing a term's relevance is to look at the frequency of a term in a corpus. Frequency is the number of times a word type occurs within a document or group of documents and is often denoted as the Absolute (or raw) Frequency (AF). When corpora (or parts of a corpus) are compared, AF must be normalized so that a word's frequency has the same significance across corpora of different sizes. Normalized frequency, more commonly referred to as Relative Frequency (RF), is expressed as a ratio of AF to the whole, i.e., to the total number of words in a corpus. As noted by Vaclav Brezina [7], in corpus linguistics, RF is multiplied by a basis for normalization (a unit of  $x$  number of words). Thus, RF is expressed as

$$RF = \frac{AF}{\text{number of tokens in corpus}} \times \text{basis for normalization.} \quad (1)$$

Brezina [7] also stresses the importance of combining RF with a measure of dispersion, which provides some indication of the distribution of the word in a corpus. A measure of dispersion is not usually necessary when looking at function words as they are fairly evenly distributed throughout texts, but when examining specialized terms, such as psychoanalytic terms, a measure of dispersion is necessary. The standard measure of

dispersion is the Standard Deviation (SD), which expresses the amount a given value varies around the mean frequency of the corpus:

$$SD_p = \frac{\sqrt{\text{sum of squared distances from the mean}}}{\text{total number of corpus parts}}. \quad (2)$$

The formula for the standard deviation of a population is presented here because of the size of the PEP-Web-20 corpus and because it represents all articles published outside the embargo period in all the English psychoanalytic journals available on PEP-Web.

When comparing frequencies from different corpora with documents of various sizes, the Coefficient of Variation (CV) is a better measure because it measures the amount of variation relative to the mean RF of a word. In other words, it allows for the comparison of different SD values. CV can be represented as

$$CV = \frac{\text{standard deviation}}{\text{mean}}. \quad (3)$$

Finally, a measure of relevance that is commonly used in information science is Term Frequency-Inverse Document Frequency (TF-IDF). This measure increases proportionally to the number of times a word appears in a document. TF-IDF is also offset by the number of documents in the corpus that contain the word. In other words, this measure compensates for documents with many occurrences of a term and makes adjustments for function words and other words that appear more frequently in general.

The first part of TF-IDF, Term Frequency (TF), is a local weighting of a term that is proportional to the frequency of the term. If this weighting is zero, it is equivalent to AF. The second part of TF-IDF, the Inverse Document Frequency (IDF), is a global measure that modulates TF, i.e., it diminishes the weight of common or more frequent words and increases the weight of terms that occur more rarely but which are, nonetheless, prominent within a given corpus or sub-corpus. IDF approaches zero for terms occurring in most documents and is larger for those appearing in fewer documents. TF-IDF is computed by multiplying TF by IDF and possibly normalizing the resulting documents to unit length. The formula for the unnormalized weight  $w$  of a term  $i$  in document  $j$  in a corpus of documents is

$$w_{i,j} = \text{frequency}_{i,j} \times \log_{10} \frac{\text{total documents in corpus}}{\text{document\_frequency}_{i,j}} \quad (4)$$

TF-IDF can be calculated in many ways. Reported in this paper are values produced using Quanteda [8], a framework for quantitative text analysis in R.

### IV. TERM RELEVANCE ANALYSIS

In addition to the R package Quanteda for TF-IDF computation, corpus analysis was performed in Python using some functions available in the Natural Language ToolKit (NLTK) [9]. NLTK is well-recognized in computational linguistics and research in natural language processing and is valued for its consistency, extensibility, and modularity [10].

#### A. Term Selection

Among the many invaluable resources provided by PEP-Web is the "PEP Consolidated Psychoanalytic Glossary" [11]

(hereafter, referred to as the "consolidated glossary") edited by David Tuckett and Nadine A. Levinson. This glossary merges glosses taken from five published psychoanalytic dictionaries: i) The Language of Psychoanalysis [12], ii) The Edinburgh International Encyclopaedia of Psychoanalysis [13], iii) A Glossary of Psychoanalytic Terms and Concepts [14], iv) A Dictionary of Kleinian Thought [15], and v) Psychoanalytic Terms and Concepts [16]. Corresponding glosses in several languages are also supplied in the consolidated glossary from the EPF Glossary of Psychoanalysis in Europe [17]. The consolidated glossary contains approximately 1,550 glosses (this calculation involved grouping variations in spelling and expanding some compound entries). Each gloss is hyperlinked throughout the 122,000 articles and books on PEP-Web.

The method for selecting a small but representative sample of important psychoanalytic terms for investigation considered the following: an ordered list of the number of documents referencing each gloss in the consolidated glossary (this number was obtained by searching each term with a robot on the full PEP-Web archive using their search engine) and a list of concepts deemed most important by eight psychoanalysts (an email was sent to analysts and candidates at two psychoanalytic institutes located in the state of Missouri asking the analysts to list the top ten most important psychoanalytic terms and five concepts having the most contemporary relevance). Expert opinion was consulted because many of the most frequent glosses in the consolidated glossary were generic terms, such as *nature*, *being*, *position*, *mother*, and *thinking*, words not generally identified as psychoanalytic terms.

The psychoanalysts listed forty-five concepts they thought were most important in psychoanalysis. Some concepts were phrased in sentences so were matched as closely as possible to the glosses by an advanced psychoanalytic candidate. A total of seventeen concepts were mentioned by more than one analyst, with all but one (*free association*) ranking in the top 10% of the ordered list of glosses (i.e., as having more than 13,464 document references). Here are the seventeen concepts, with the number of document references in parentheses followed by the number of analysts listing the concept/gloss in the top ten: *aggression* (21092, 2), *defense/defence*\* (31503, 4), *dream*\* (25589, 3), *conflict* (33631, 3), *drive* (20282, 2), *ego* (39555, 4), *super ego/superego/super-ego* (16614, 3), *id* (14604, 3), *fantasy/phantasy* (25751/6815, 3), *free association* (6976, 2), *oedipus complex* (8704, 3), *projection*\* (16736, 2), *repetition* (16677, 2), *repression* (16548, 2), *symptom*\* (16596, 2), *transference/countertransference/counter-transference* (59322, 6), and *unconscious*\* (48235, 7). Of the top ten concepts listed above, those deemed most relevant in the contemporary scene are starred, with *conflict*, *defense*, *dream*, *projection*, *symptom*, and *unconscious* considered relevant today by more than one analyst. All but one of the above glosses, *free association*, were selected for relevance analysis.

### B. Text Preparation Pipeline

Before term relevance could be evaluated, the text in the PEP-Web-20 corpus had to undergo further cleaning and preparation. All text was made lowercase, British and English differences in spelling were normalized, and the text was lemmatized and stemmed using the Porter Stemmer. However,

because none of the NLTK stemming packages captured all the relevant inflectional forms of the sixteen analytic terms under examination, code was written to reduce them separately. Next, the text was tokenized by splitting words based on the space character. Finally, the following stop words were removed 'pp', 'p', 's', and 't'.

### C. Results

The pipeline resulted in new token counts for each category (see last column Table I). Table II presents the analysis of term relevance for the sixteen psychoanalytic terms presented in section IV.A. Each term receives three rows in the table, one for each of the following relevance measures described in section III: AF, RF, and TF-IDF. RF is calculated with a basis of 1000 words. This value was selected to be proportional to the average number of words per document post pipeline, which is 3783. Relevance measures for each decade are presented in columns, in addition to mean, SD, and CV, as described in section III. A dash value for AF or RF indicates that the term did not appear within a given category. Accordingly, these values are not included in the mean or standard deviation calculations so as not to skew the results. The TF-IDF values for these decades are set to zero to indicate a total lack of relevance during that decade.

Figure 1 compares RF and TF-IDF measures for four psychoanalytic terms from 1920 to 2020. The 1920s is a significant decade for several reasons. During this time, interest in psychoanalysis rose in Britain and America, as evident in Table I, where psychoanalytic texts in English nearly quadrupled in the 1920s compared to the decade before. This spike in scholarly activity in the 1920s was most likely stimulated by the establishment of The British Society for Psychoanalysis in 1919, the American Psychoanalytic Association in 1911, and the Institute for Psychoanalytic Education and Training in Berlin in 1923. Finally, of interest to this study, all sixteen psychoanalytic terms under consideration here were present in the literature by 1920.





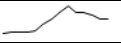
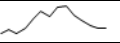


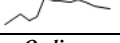
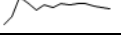

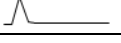
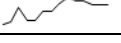

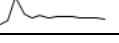

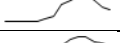
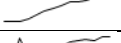
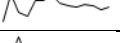
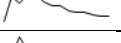
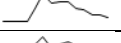
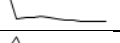
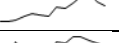
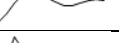
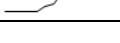
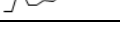
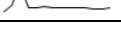

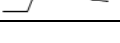

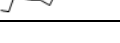

The four graphs in Figure 1 demonstrate how RF and TF-IDF capture different aspects of term relevance. As already noted, RF describes the extent to which a given term appears in a corpus with respect to the total number of tokens in the corpus. Another meaningful aspect of term relevance is the extent to which multiple documents reference the term. As a term saturates the corpus, it becomes less relevant with respect to the amount of novel contribution it brings. The TF-IDF statistic modulates RF by adjusting the value placed on document-level term frequencies according to the number of other documents that also reference the term. Overly saturated terms are of lower relevance. This may be shown at the corpus-level by averaging document-level TF-IDF values as we did for each of the ten decades represented in Figure 1.

The RF line in Figure 1.A demonstrates a rapid decline of the term *repression* from 1920 to 2020, while the TF-IDF line shows a somewhat slower rate of decline in term relevance. Post-analysis shows that an average of 32% of documents includes the term *repression* across these ten decades with little variation. In a sense, the TF-IDF metric rewards the term for lower saturation levels, which is visible in its slower rate of decline across time.

TABLE II. AF, RF, AND TF-IDF PER TERM PER DECADE AND OTHER STATISTICS

	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	SD	Mean	CV
Aggression AF	-	14	28	76	561	4084	5619	8847	10189	15208	19133	23136	22686	14686	8347	9559	0.87
RF	-	0.05	0.04	0.04	0.13	0.67	0.81	0.68	0.58	0.61	0.50	0.42	0.35	0.30	0.26	0.40	0.65
TF-IDF	0.00	0.22	0.43	0.17	0.31	1.12	0.96	0.95	0.91	1.02	0.87	0.74	0.70	0.64	0.35	0.65	0.54
Conflict AF	-	31	122	866	2116	4133	4637	10023	13876	20814	34014	41747	39358	24031	14802	15059	0.98
RF	-	0.12	0.17	0.45	0.48	0.68	0.67	0.78	0.80	0.83	0.89	0.76	0.60	0.49	0.23	0.59	0.39
TF-IDF	0.00	0.36	1.14	0.86	0.58	0.82	0.68	0.84	0.82	0.87	0.85	0.73	0.69	0.62	0.26	0.70	0.37
Defense AF	1	208	100	342	1043	3347	4039	11425	15567	20346	33104	39057	39081	26851	14596	13894	1.05
RF	0.03	0.80	0.14	0.18	0.24	0.55	0.58	0.88	0.89	0.81	0.87	0.71	0.60	0.55	0.29	0.56	0.51
TF-IDF	0.12	1.01	1.09	0.46	0.42	0.94	0.70	0.99	0.95	0.87	0.88	0.70	0.68	0.69	0.26	0.75	0.35
Dream AF	1	119	6866	5378	8903	9385	12110	22289	25530	35214	53831	56497	66552	45171	21869	24846	0.88
RF	0.03	0.46	9.45	2.79	2.03	1.55	1.75	1.72	1.46	1.40	1.41	1.03	1.02	0.93	2.18	1.93	1.13
TF-IDF	0.12	0.60	28.39	2.99	2.03	2.07	2.09	2.21	2.04	2.03	2.02	1.51	1.65	1.57	6.89	3.67	1.88
Drive AF	-	9	35	72	251	1217	2012	5671	10014	10889	16072	20080	20105	15170	7638	7815	0.98
RF	-	0.03	0.05	0.04	0.06	0.20	0.29	0.44	0.57	0.43	0.42	0.37	0.31	0.31	0.17	0.27	0.64
TF-IDF	0.00	0.13	0.58	0.15	0.15	0.44	0.46	0.74	0.96	0.79	0.80	0.67	0.68	0.68	0.29	0.52	0.56
Ego AF	3	120	70	938	5685	11397	10229	28750	40982	38223	45539	41907	38504	24025	17458	20455	0.85
RF	0.10	0.46	0.10	0.49	1.29	1.88	1.48	2.22	2.35	1.52	1.19	0.76	0.59	0.49	0.72	1.07	0.68
TF-IDF	0.36	0.90	0.96	1.33	1.82	2.55	1.67	2.27	2.08	1.59	1.37	0.97	0.98	0.90	0.60	1.41	0.42
Fantasy AF	-	79	457	1548	3546	5944	6321	11907	16146	23611	37250	52572	49906	30523	17975	18447	0.97
RF	-	0.30	0.63	0.80	0.81	0.98	0.92	0.92	0.93	0.94	0.98	0.96	0.76	0.63	0.19	0.81	0.23
TF-IDF	0.00	0.70	4.04	1.52	1.05	1.35	1.08	1.18	1.19	1.16	1.10	1.01	0.95	0.88	0.85	1.23	0.69
Id AF	-	-	-	24	956	1770	1597	3513	4989	3822	5937	4322	4025	2673	1730	3057	0.57
RF	-	-	-	0.01	0.22	0.29	0.23	0.27	0.29	0.15	0.16	0.08	0.06	0.06	0.10	0.16	0.59
TF-IDF	0.00	0.00	0.00	0.08	0.62	0.67	0.48	0.61	0.71	0.47	0.51	0.28	0.27	0.25	0.25	0.35	0.71
Oedipus Complex AF	-	-	-	1	8	248	623	3331	5484	9091	15578	17177	14200	8389	6284	6739	0.93
RF	-	-	-	0.00	0.00	0.04	0.09	0.26	0.31	0.36	0.41	0.31	0.22	0.17	0.14	0.20	0.71
TF-IDF	0.00	0.00	0.00	0.01	0.01	0.17	0.24	0.57	0.72	0.80	0.83	0.65	0.62	0.53	0.32	0.37	0.88
Projection AF	1	8	23	100	331	591	754	1685	2597	3746	6033	9421	13232	9840	4265	3454	1.23
RF	0.03	0.03	0.03	0.05	0.08	0.10	0.11	0.13	0.15	0.15	0.16	0.17	0.20	0.20	0.06	0.11	0.52
TF-IDF	0.12	0.12	0.38	0.18	0.19	0.25	0.22	0.29	0.35	0.37	0.37	0.36	0.40	0.40	0.10	0.29	0.35
Repetition AF	6	110	200	473	1611	2259	2803	4504	5770	7870	12860	18066	19424	15588	6786	6539	1.04
RF	0.20	0.42	0.28	0.25	0.37	0.37	0.41	0.35	0.33	0.31	0.34	0.33	0.30	0.32	0.06	0.33	0.18
TF-IDF	0.32	0.65	1.77	0.50	0.54	0.58	0.54	0.51	0.52	0.51	0.52	0.48	0.48	0.52	0.33	0.60	0.55
Repression AF	2	215	481	1796	4179	4226	3844	7055	7280	8594	13117	15245	14595	10757	5119	6528	0.78
RF	0.07	0.83	0.66	0.93	0.95	0.70	0.56	0.55	0.42	0.34	0.34	0.28	0.22	0.22	0.27	0.50	0.53
TF-IDF	0.16	1.13	3.29	1.52	1.01	0.95	0.69	0.82	0.71	0.66	0.69	0.55	0.55	0.56	0.72	0.95	0.76
Superego AF	-	-	-	8	1653	4621	3406	6825	9483	9465	12496	10997	11780	6506	4018	7022	0.57
RF	-	-	-	0.00	0.38	0.76	0.49	0.53	0.54	0.38	0.33	0.20	0.18	0.13	0.21	0.36	0.58
TF-IDF	0.00	0.00	0.00	0.03	0.94	1.32	0.74	0.96	1.04	0.86	0.80	0.55	0.60	0.49	0.42	0.60	0.71
Symptom AF	61	640	355	1108	2933	4262	4229	6369	7974	9288	12714	16514	18967	15704	6252	7223	0.87
RF	2.04	2.47	0.49	0.58	0.67	0.70	0.61	0.49	0.46	0.37	0.33	0.30	0.29	0.32	0.64	0.72	0.89
TF-IDF	1.27	1.55	2.30	1.03	0.88	0.98	0.72	0.72	0.73	0.69	0.65	0.57	0.64	0.71	0.46	0.96	0.48
Transference AF	2	5	92	631	2077	2411	2172	10677	14174	26510	55581	89161	75364	46418	29699	23234	1.28
RF	0.07	0.02	0.13	0.33	0.47	0.40	0.31	0.83	0.81	1.06	1.46	1.63	1.15	0.96	0.50	0.69	0.73
TF-IDF	0.24	0.11	1.43	0.91	0.79	0.78	0.53	1.44	1.37	1.69	1.75	1.52	1.32	1.21	0.50	1.08	0.47
Unconscious AF	6	134	687	2547	5710	6694	7815	12853	14833	21032	36286	57590	71646	51040	22966	20634	1.11
RF	0.20	0.52	0.95	1.32	1.30	1.10	1.13	0.99	0.85	0.84	0.95	1.05	1.09	1.05	0.28	0.95	0.30
TF-IDF	0.32	0.89	3.20	1.62	1.03	1.05	0.92	0.88	0.82	0.88	0.89	0.83	0.90	0.91	0.64	1.08	0.59

TABLE III. SPARKLINES FOR RF AND TF-IDF FOR ALL SIXTEEN TERMS FROM 1880 THROUGH 2010

	<i>Aggression</i>	<i>Conflict</i>	<i>Defense</i>	<i>Dream</i>	<i>Drive</i>	<i>Ego</i>	<i>Fantasy</i>	<i>Id</i>
RF								
TF-IDF								
	<i>Oedipus Complex</i>	<i>Projection</i>	<i>Repetition</i>	<i>Repression</i>	<i>Superego</i>	<i>Symptom</i>	<i>Transference</i>	<i>Unconscious</i>
RF								
TF-IDF								

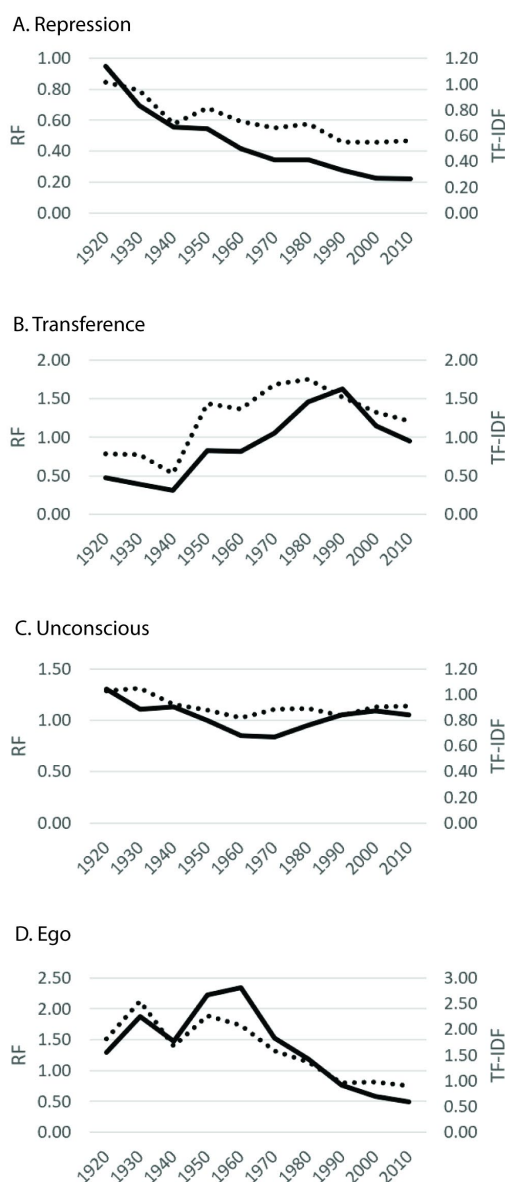


Fig. 1. Comparison of RF and TF-IDF (1920-2020) for *Repression*, *Transference*, *Unconscious*, and *Ego* (solid lines represent RF; dotted lines TF-IDF; dates on the x-axis represent decades; right and left labels on the y-axis provide RF and TF-IDF value ranges, with exact values found in Table II)

Figure 1.B presents the relevance measures for the term *transference*. Both lines show an increase in term relevance over the period of 1940 to around 1980/1990. From 1940 to a peak in the 1990s, AF rose from 2,172 occurrences to close to 90,000. During the same time, however, the percentage of documents that included the term *transference* grew from 17% to 51% by the end of the 1980s and reached a high of 57% in the 1990s. The TF-IDF line accordingly turns downward after 1980 and drops off considerably after that because of the high saturation

of the term *transference* in the literature as a whole. The decline of RF compounded by the relatively higher saturation of the term is making it less relevant.

Figure 1.C indicates that the RF of the term *unconscious* drops by about 33% from its high point in the 1920s to its lowest level in 1970s, and then recovers much of what it lost. The percentage of documents referencing *unconscious* steadily increases from 46% in the 1920s to 64% in 2010s. The moderation of RF is visible in this figure. The TF-IDF failed to follow the u-shape form of the RF due to the higher corpus saturation in the later decades.

Finally, Figure 1.D provides an example of the detrimental influence of low RF on the TF-IDF measure. The term *ego* quickly dropped in RF in the 1970s after a peak in the 1960s with the ascendancy of ego psychology in America. From 1970 on, the percentage of documents referencing *ego* dropped quickly from 59% in the 1960s to 38% in the most recent decade. In this case, the lower saturation of documents was not enough to overcome the influence of low relevance, and the TF-IDF is drawn down as well.

Table III presents RF and TF-IDF sparklines for each of the sixteen terms for readers interested in examining differences in these measures for other key psychoanalytic concepts.

## V. APPLICATIONS

In section IV, the more detailed analysis of four key psychoanalytic terms showed that graphs of relevance do more than provide an overview of a term's significance over time; such graphs also pose questions in need of answers regarding the reasons for the ebbs and flows of a term's popularity and pinpoint historical periods for further investigation by scholars in the field.

In addition, we believe that relevance graphs, along with a discussion of them, would prove useful in glossaries and provide valuable historical context for glosses. We also suggest that online glossaries produce relevance sparklines that scholars (especially authors of textbooks) could include when first presenting a key term. Relevance sparklines could be followed in parentheses with information regarding the time span of the graph and breaks in usage as follows: *repetition* (RF: [1880-2010]).

## VI. CONCLUSION

In this study, relevance graphs of sixteen key psychoanalytic concepts were extracted with changes mapped across the entire history of psychoanalysis. A corpus, called PEP-Web-20, containing over 286 million words taken from journal articles, books, and video transcripts, was built and organized into fourteen classes representing the decade in which the document had been published. Concept relevance was analyzed using three statistical measures: absolute frequency (AF), relative frequency (RF), and Term Frequency-Inverse Document Frequency (TF-IDF). Results demonstrated that AF and TF-IDF provide different insights into the histories of terms, as illustrated by our readings of the AF and TF-IDF relevance graphs for four terms *ego*, *repression*, *transference*, and *unconscious*.

Future work will include investigating other measures of relevance, building relevance graphs for several hundred glosses

in the "PEP Consolidated Psychoanalytic Glossary," suggesting new glosses based on relevance analysis, performing sentiment analysis of psychoanalytic terms over time, and examining changes in the meaning of some concepts by analyzing collocations.

Although the PEP-Web-20 corpus cannot be shared due to copyright law, all generated statistical data are available upon request from the first author.

#### REFERENCES

- [1] M. Gavin, C. Jennings, L. Kersey, and B. Pasanek, "Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading," *Debates in the digital humanities 2019*, M. K. Gold and L. F. Klein, Eds., Minneapolis, MN: University of Minnesota Press, 2019, pp. 333-366.
- [2] Y. Zhang, H. Chen, J. Lu, and G. Zhang, "Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016," *Knowledge-Based Systems*, vol. 133, pp. 255-268, 2017/10/01/ 2017.
- [3] O. Faust, "Documenting and predicting topic changes in Computers in Biology and Medicine: A bibliometric keyword analysis from 1990 to 2017," *Informatics in Medicine Unlocked*, vol. 11, pp. 15-27, 2018/01/01/ 2018, doi: <https://doi.org/10.1016/j.imu.2018.03.002>.
- [4] M. Trevisani and A. Tuzzi, "Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories," *Knowl. Based Syst.*, vol. 146, pp. 129-141, 2018.
- [5] D. Biber, S. Conrad, and R. Reppen, *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press, 1998.
- [6] M. Nelson, "Building a written corpus: What are the basics," in *Routledge handbooks in applied linguistics*, A. O'Keeffe and M. McCarthy Eds. New York: Routledge, 2010, pp. 53-65.
- [7] V. Brezina, *Statistical corpus linguistics: A practical guide*. Cambridge: Cambridge University Press, 2018.
- [8] C. D. Manning and P. Raghavan, *Introduction to information retrieval*. New York: Cambridge University Press, 2008.
- [9] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc, 2009.
- [10] M. Lobur, A. Romanyuk, and M. Romanyshyn, "Using NLTK for educational and scientific purposes," in *2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, 23-25 Feb. 2011 2011, pp. 426-428.
- [11] D. Tuckett and N. A. Levinson, *PEP consolidated psychoanalytic glossary*. London: Psychoanalytic Electronic Publishing, 2016.
- [12] J. Laplanche and J. B. Pontalis, *The language of psychoanalysis*. New York/London: W. W. Norton, 1973.
- [13] R. Skelton, Ed. *The edinburgh international encyclopaedia of psychoanalysis*. Edinburgh: Edinburgh University Press, 2006.
- [14] B. Moore and B. D. Fine, *A glossary of psychoanalytic terms and concepts*. New York: American Psychoanalytic Association, 1968.
- [15] R. D. Hinshelwood, *A dictionary of Kleinian thought*. London: Free Association Books, 1989.
- [16] E. L. Auchincloss and E. Samberg, *Psychoanalytic terms and concepts*. New Haven: Yale University Press, 2012.
- [17] G. Junkers, *EPF glossary of psychoanalysis in europe*. Karnac Books, 1997.