Part II – Metrics Suggesting Centrality or Complexity of Themes

Rees W. Morrison

2022-02-04

Part I of this series of five posts (working notes for myself in fact) provides an overview of the Themes analyzed in this introductory investigation of centrality and complexity. This second post lays out the metrics that underlie the analysis. I present all 20 of them, with the brief introduction of each metric covering similar topics and order: first, why I gathered the metric, next the source of the metric, and then any calculations that were carried out. Finally, a two-by-two layout of plots shows the results for the four preceding metrics, each of whom loosely shares a common attribute.

The table below lists each metric, its source and its (preliminary) classification regarding whether it suggests centrality, complexity or a mixture between the two. The right-most column shows the attribute by which I sorted the 20 metrics into five groups. The following pages present the metrics in attribute groups of four.

Table 1: 20 Metrics

Metric	Source(s)	Class	Attribute
Synonyms	https://thesaurus.com	Mixture	frequency
OED Frequency Bands	https://www.oed.com/	Mixture	frequency
1000 Most Common Words	three combined	Centrality	frequency
Google Book Ngrams	https://books.google.com/ngrams	Centrality	frequency
Google Trillion Words	https://norvig.com/ngrams/count_1w.txt	Mixture	internet
Google Search Results	browser search with Google	Centrality	internet
Google Trend Percents	https://support.google.com/trends/	Centrality	internet
Hashtags	https://www.wordtracker.com/	Centrality	internet
Quotations	http://www.quotationspage.com/ + kaggle data	Complexity	language
Poem Titles	three combined	Complexity	language
Expressions	multiple online sources	Centrality	language
Book Titles	https://www.worldcat.org/advancedsearch	Mixture	commercial art
Readability	R program functions	Complexity	language
Song Titles	http://www.mldb.org/	Centrality	commercial art
Movie Titles	https://www.imdb.com/search/title/	Centrality	commercial art
Play Synopses	http://www.playdatabase.com/	Complexity	commercial art
PhD Dissertations	https://biblioboard.com/opendissertations/	Complexity	complexity
Code of Fed. Regs.	https://www.ecfr.gov/search	Complexity	complexity
UN Documents	https://digitallibrary.un.org/search?	Complexity	complexity
College Majors	author classification	Complexity	complexity

Synonyms

One attribute of a Theme is the number of words in English that are deemed to have a similar meaning. The more synonyms that exist, therefore, the more people have used the idea and developed nuances; thus

both centrality and complexity could correlate to the number of synonyms available.

For each Theme, we found the number of synonyms that a leading dictionary gives for the concept. The source, Thesaurus.com, identifies "most relevant" synonyms (colored red on the website) and less-close synonyms (colored orange). The metric for this analysis consists of the sum of both kinds of synonyms for each Theme.

The plot in the quartet plot below depicts for each Theme the aggregated number of its synonyms.

Oxford English Dictionary (OED), Frequency Bands

As with Google Trillion words, frequency of use could correspond to centrality as well as complexity. Centrality because it is much discussed and written about, and complexity because relatively many ideas, implications, extensions, comparisons, and applications come to mind about the concept.

The Oxford English Dictionary (OED) assigns each of its words a band from 1 to 8 according to how frequently the OED has determined that the term appears per million words. Specifically, Frequency Band 6 contains words that occur between 10 and 100 times per million words in typical modern English usage. Frequency Band 7 contains words that occur between 100 and 1,000 times per million words (i.e., at least ten times more frequently than Band 6 words appear), while Frequency Band 8 covers words that occur more than 1,000 times per million words. This highest band includes the most common English words, such as determiners (e.g., the, a, an, this, that), pronouns, principal prepositions, and conjunctions (e.g., and, but, that, if).

For each Theme, the metric is its OED Frequency Band. We also compiled the earliest year the OED has found of written use of the Theme term, and plotted both year and band.

1,000 Most Common English Words, Rank

The same reasoning that led to metrics from Google's Trillion Word data base and the OED frequency bands supports the inclusion of rankings of Themes among the thousand most common words in English. I combined three sources to create the list used here. One is 1000 Most Common Words, a second was Wikipedia, and the third came from Smart Words.

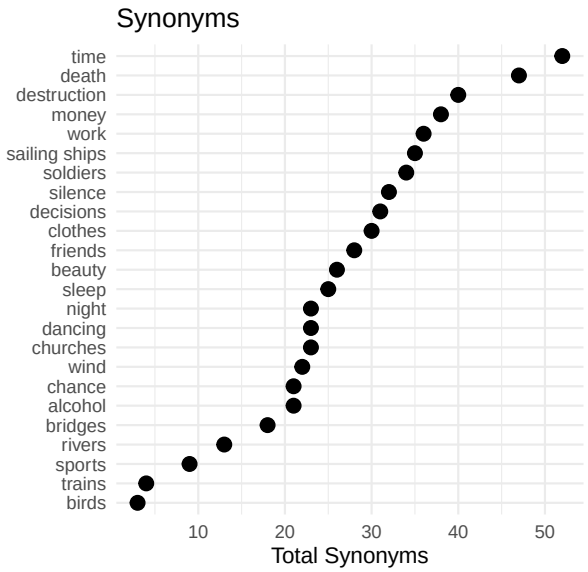
If the Theme was expressed as a plural, such as "Soldiers", I determined the ranking of the singular form. Any Theme that was not included in the composite list received a ranking of 1001. For each Theme, the plot below shows its rank on the compiled list.

Google NGram Viewer, Percentage of Book References

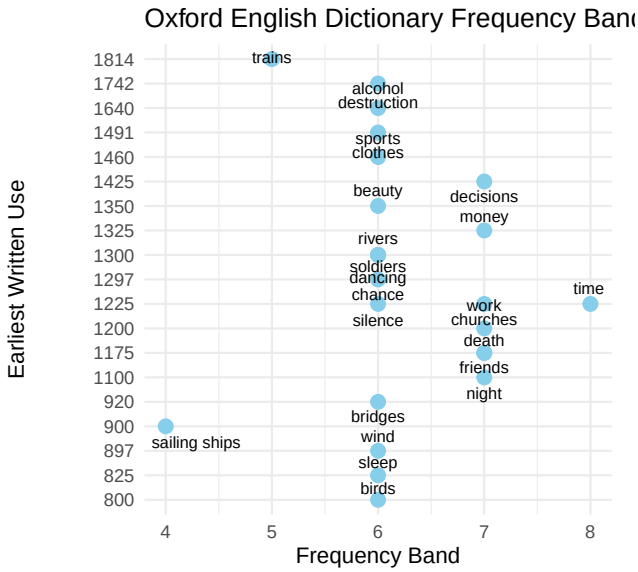
What people write about in books gives insight into what is important in their thinking. For Themes, therefore, the metric suggests complexity more than centrality.

The Google Ngram Viewer charts the frequencies of search strings using a yearly count of n-grams found in sources printed between 1500 and 2019 in Google's text corpora. I searched for each Theme and added together the percentages given for lower case, initial caps, and all caps of the word for the latest year, 2000.

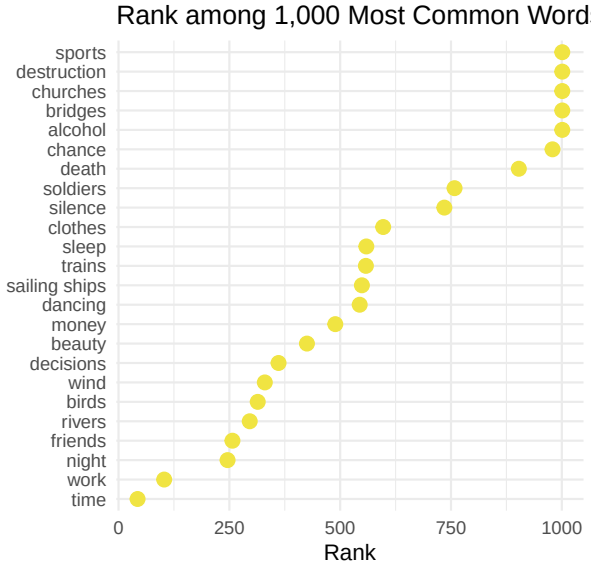
Here are the four preceding plots, whose metrics share an attribute of Theme frequency.



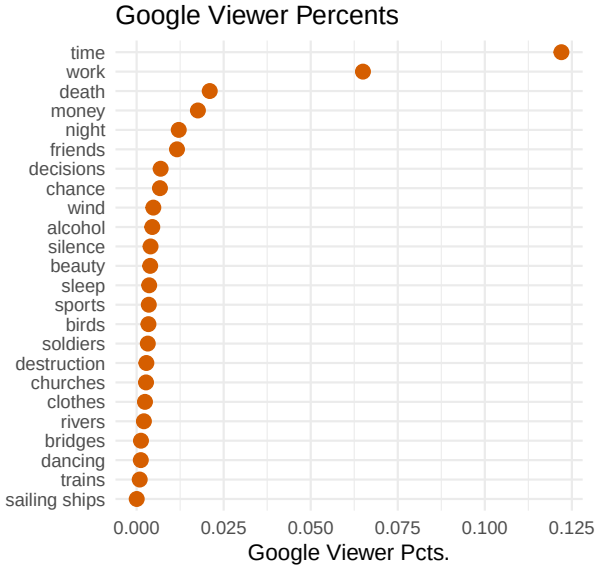
Source: Thesaurus.com



Source: Princeton Pub. Lib. access to OED Online
https://www.oed.com/



Sources: https://1000mostcommonwords.com/
https://en.wikipedia.org/wiki/Most_common_words_in_English
s://www.smart−words.org/500−most−commonly−used−english−words.html



Source: https://books.google.com/ngrams/

Google's Trillion-Word Database, Frequency

A metric of how common a Theme is on the internet could point us toward its centrality, because it is foremost in the minds of many writers, or perhaps to its complexity, because more has been written about the concept than about a less-frequently used concept.

One source of data about the frequency with which words have been used comes from the database created by Google's Peter Norvig. In 2006, Google made available 1,024,908,267,229 words (one trillion, 24 billion words) and counts for words that appeared at least 40 times. The metric used here simply searched for each Theme word and included the count.

The quartet plot after the next three metrics displays the count for each Theme word.

Google Search Results, Number Returned

Another hypothesis for the relative centrality of a Theme holds that the more people have published a term on the internet, the more relevant it is to their lives: the more central. Thus, the number of hits for a Theme found by a powerful search engine crudely reflects its relative currency and salience.

For this metric, I entered a search into Google for each Theme. When the results page returned, I recorded the number of results from the top left of my screen (e.g., for Thinking it said "About 992,000,000 results". A later post will incorporate comparable search data results from Bing.

The data collected appear in the plot below, where each Theme sits to the left of the point that shows the Google results hits.

Google Trends, Internet Search Requests

Google Trends provides access to a largely unfiltered sample of search requests made using Google. To make comparisons between terms easier Google Trends normalizes the search results to the time and location of a query by the following process:

Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. The resulting numbers are then scaled from 0 to 100 based on a topic's proportion to all searches on all topics.

Numbers represent search interest relative to the highest point on the chart for the United States and the year. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for the term.

The metrics are plotted in the corresponding chart below.

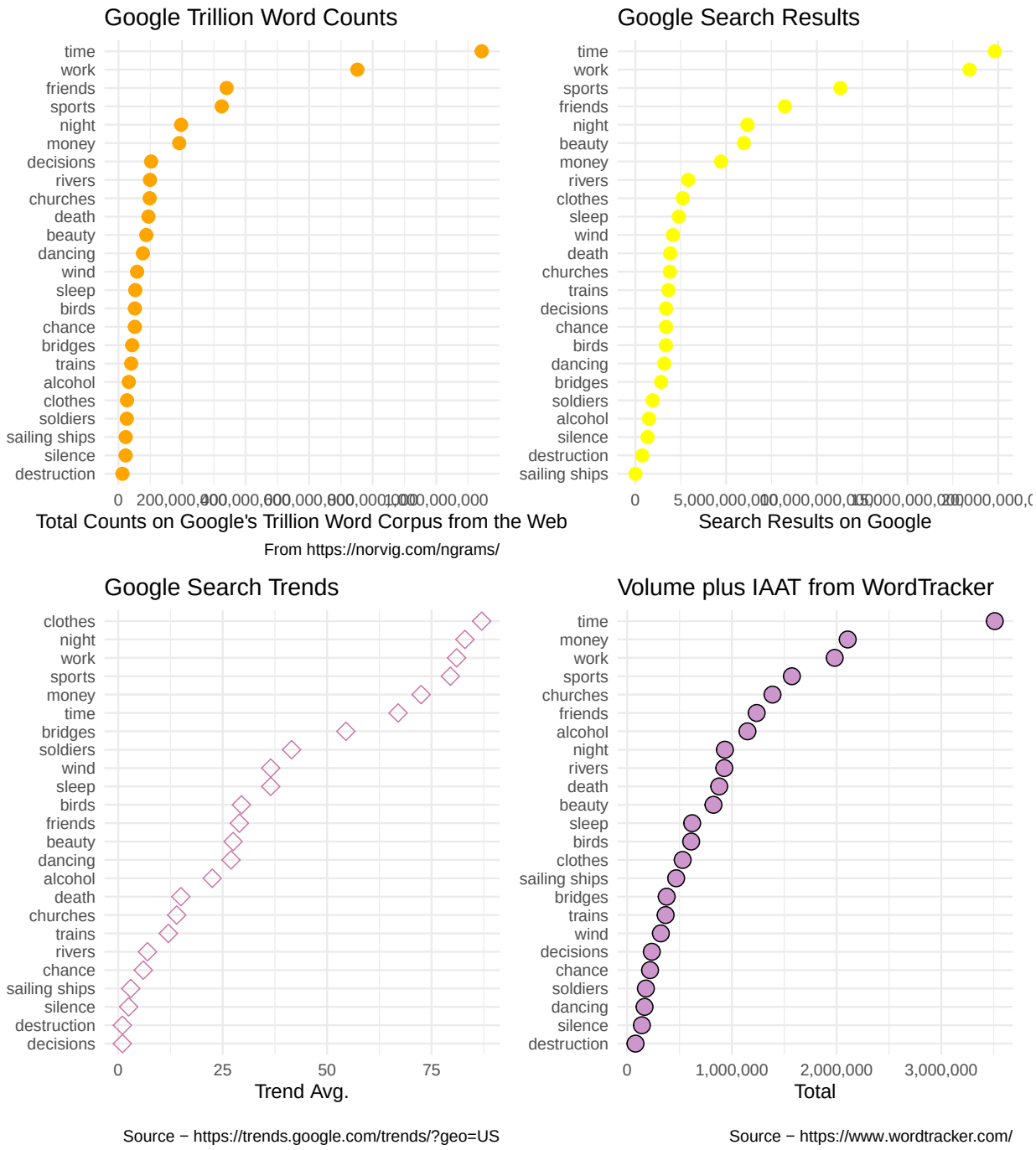
SEO Hashtags

Those who maintain an online presence want their site or advertisement to attract viewers and persuade them to read or to click and buy. Thus, among many Search Engine Optimization (SEO) techniques, online publishers choose hashtags and keywords to capture terms that will lure searches by search engines to display their company high in the results. SEO is the effort to game this interchange. Thus, a Theme term in prominent use online in the form of hashtags rests on centrality: what people want to buy, not anything heady like complexity.

One of the online services that advises marketers regarding search terms, WordTracker, provides two sets of data for each Theme word search: *Volume* ("the number of times a search for each keyword appears in our database") and *IAAT* ("The raw number of pages that have the keyword both in the title tag and also in anchor text in a backlink from an external domain."). Because the correlation between these to numbers is only 0.500, I added them together.

The lower right plot in the quartet presents the data.

Here are the four preceding plots, whose metrics share an attribute of a Theme's presence on the internet.



Quotations, Number of Appearances

People say something quotable (or are credited with the bon mot) because they are witty and pithy, or because they deliberately craft something memorable in a speech or writing. Quotations may derive from deep insights (complexity) or from a humorous twist on an everyday topic (centrality), but I lean toward the number of quotations using a Theme as an indicator more of complexity.

The host of The Quotations Page has compiled seven sources of quotations. In the Fall of 2021, the home page claimed 28,000 quotations in English from more than 3,400 authors. Using the site's search engine, I

recorded how many quotes have each Theme term.

Additionally, a Kaggle dataset includes 22 MB of JSON formatted quotations. From its 36,937 unique quotes, I recorded how many have each Theme term. Adding the two results together for each Theme generated this metric.

The plot in the combination graphic hereafter displays the total of the counts from the two sources.

Poem Titles, Count of Instances

By their nature of compression, allusion, meter and rhyme, poems suggest that if a Theme appears in the title, it more likely than not pertains to a deliberate, thoughtful aspect of the poem: ergo, an indicator of the Theme's complexity.

Three online databases of poems provided the data for this metric: Poetry.com, Poetry Foundation, and Poetry.org. I added together the results returned by searches on them for each Theme.

The plot that follows shows a point for each Theme corresponding to the number of combined results.

Figurative Expressions, Count

Theme terms appear in figurative, idiomatic expressions. For an example with the Theme of Clothes, "hung out to dry" refers literally to drying clothes on a line, but the expression conveys metaphorically abandoning someone in a fraught situation. For each Theme I compiled idiomatic expressions that use the term or an associated term (such as "shirt" for Clothes, or "barn burner" for Fire). The non-literal phrases came from more than 38 online sources. Expressions become commonplace because they capture an idea that many people adopted as a way of speaking indirectly – they point toward centrality.

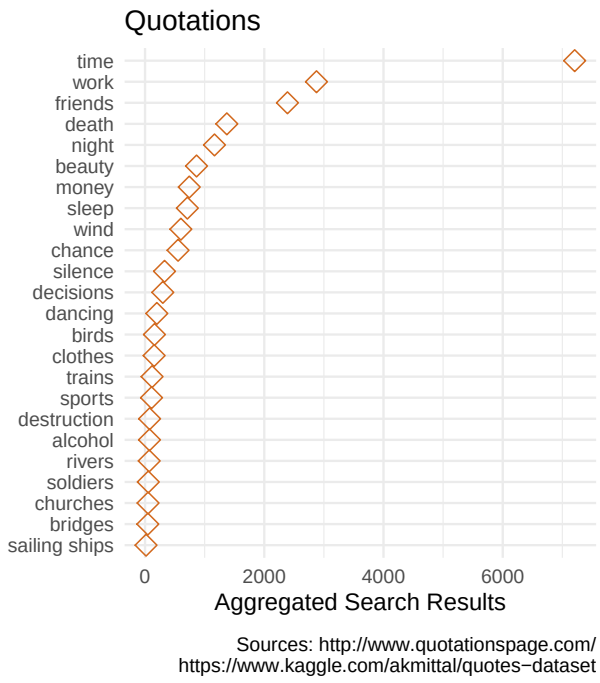
Counting the number of unique expressions (combining slight variations on them) resulted in the 1661shown in the graphic below as the total per Theme.

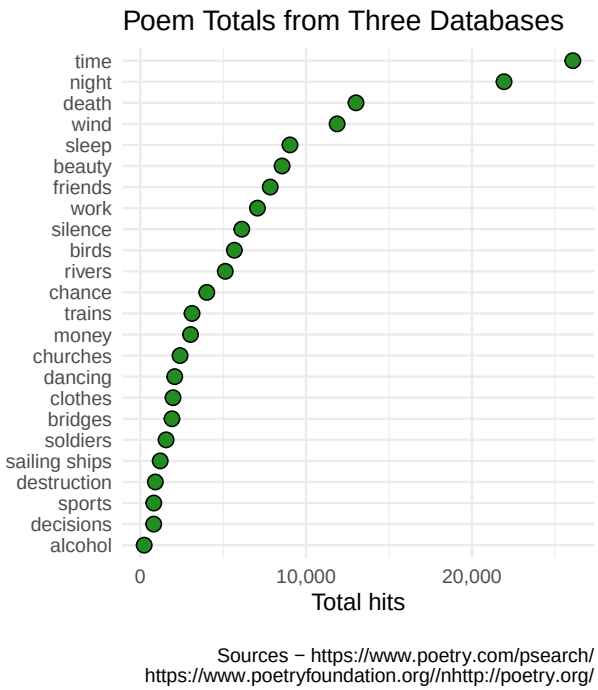
Readability of Posts, School Grade Level

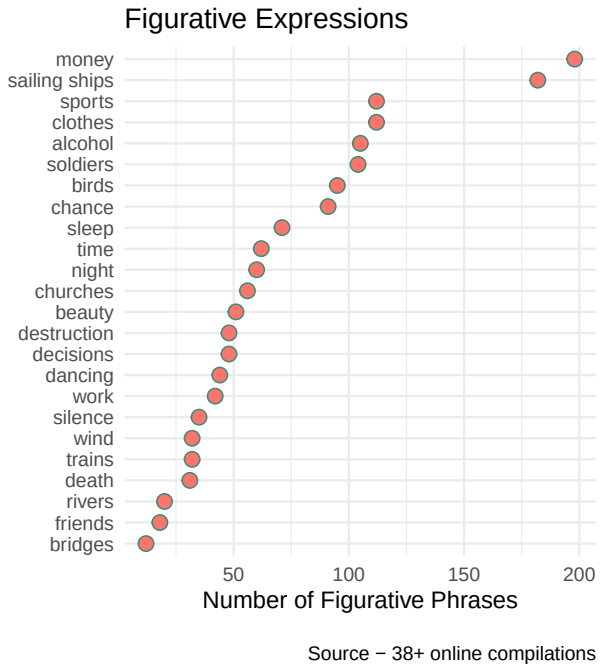
It seemed plausible to me that how people write about Themes might match to some degree the centrality or the complexity of the Theme. Thus, I searched for algorithms that assess the level of a piece of writing. Of the multiplicity of readability measures that have been developed, I chose three that make sense for adult non-fiction writing: Flesch-Kincaid Grade Level, SMOG, and the Coleman-Liau Index. I calculated the scores of those three measures for each Theme's blog posts (with some adjustments to the content, such as to remove lines of poems and lyrics of songs), converted the three measures all to a level of schooling (such as 10.2 for into 10th grade), and averaged the levels. The higher the grade level, the more complex the treatment. This assumes that I write consistently, with a similar vocabulary breadth and style, so it is very subjective and unlikely to be included in the next round of analysis.

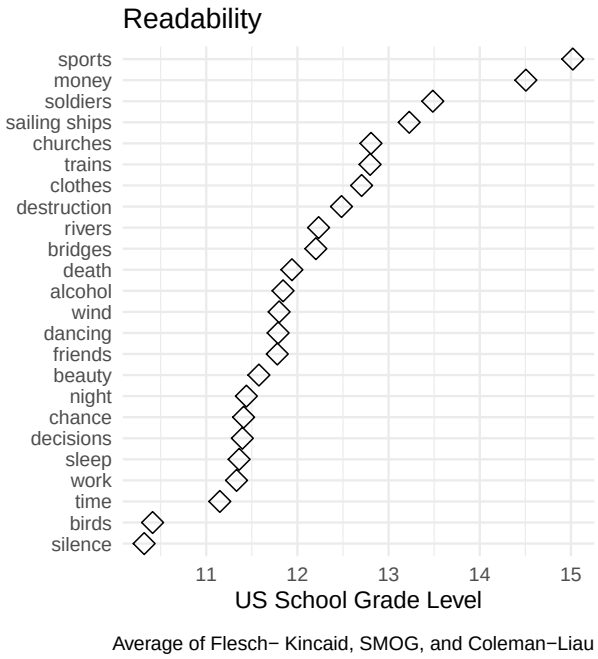
This metric appears in the lower right plot.

Here are the four preceding plots, whose metrics share an emphasis on Themes used in literary language.









Book Titles, Count of Instances

When an author and publisher come up with a book's title, it may be an effort primarily to appeal to the market segment that is likely to buy the book (tends toward centrality) or it may be a title chosen by the author (might be complexity, e.g., "The Color Purple" or "Catcher in the Rye").

The online site WorldCat.org lets users search Format "Book", Language "English" and specify Audience "Any Audience", as well as Content "Any Content." Before running searches on Themes, I removed Dissertations (because I have a separate metric for them) and used singular versions of Theme terms, except

Trains. The results on the website appear in this format: "Results 1-10 of about 95,810".

The metrics collected are the number of titles found on WorldCat.org and the plot in the group of four below presents them.

Song Titles, Count of Instances

As with the titles of books, the choice of a title for a song may be artistic or it may be commercial: what will attract a prospective buyer of the record. Obviously, too, those who pick song titles do not want to duplicate an earlier title. If the title caters to the pragmatic, money-making style, centrality reins; if the title derives from a less facile, more complicated source, it suggests more the complexity of the Theme. On the whole, given the competitive pressures surrounding the efforts to produce hit songs, I favor centrality.

To obtain a metric for how often Theme terms have appeared in the titles of songs, I searched The Music Lyrics Database. On the header you click "Search" and then pick "title." On July 26, 2021, MLDB.org claimed "Lyrics: 236,453, Albums: 23,932, Artists: 11,115." After submitting a search, multiple pages of results show at the bottom of the page, but the highest number on the bottom right provides the data (I multiplied the number of total pages of results [less 1] by 30, and added the number of results on the final page).

Plotted below are the results.

Movie Titles, Count of Instances

In the same vein as song titles, the decision to include a Theme term in a movie title might have several motivations. But metrics based on a huge number of movie titles suggest that centrality drives more of the decisions – the title should resonate immediately with many prospective movie-goers. To the degree that titles of movies derive from the title of a book (or play, musical, or short story) the movie adapts, we will see overlap with book titles (or play synopses, as per below) and the choice of title is more pre-determined.

The source of these metrics is the database of more than 80,000 movies maintained by IMDb.com. The search looked only in Titles, Feature Films (not TV, series, etc) and excluded Adult Films.

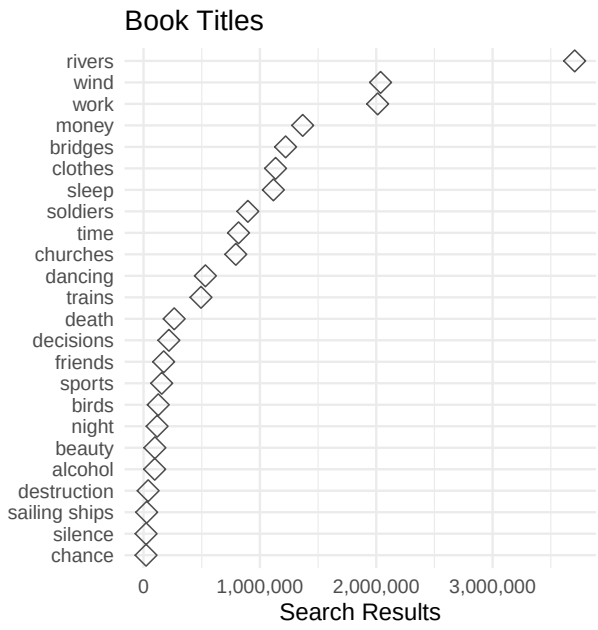
Below, in the next mosaic of four plots, is a plot of the resulting data.

Play Synopses, Count of Instances

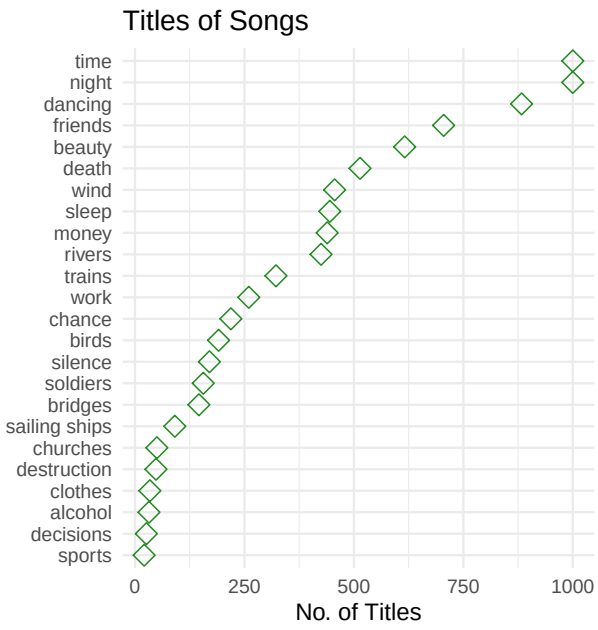
To explore our hypotheses of centrality, complexity, or both, we can add synopses of plays. Plays suggest more commonly an indicator of complexity as they deal with human psychology and emotions. Here we deal not with titles, as in books, songs and poems, but with brief overviews of the plays.

The source for summaries of plays in English is Play Database which stated in early Winter of 2021 that its "Current Totals" included 12,498 plays by 5,653 playwrites. The Navigation Bar offers a choice for "Search." I entered each of the Theme terms into that search function and accepted all defaults. The maximum results returned is 200, so the data is truncated for the six Themes in the upper right of the graphic.

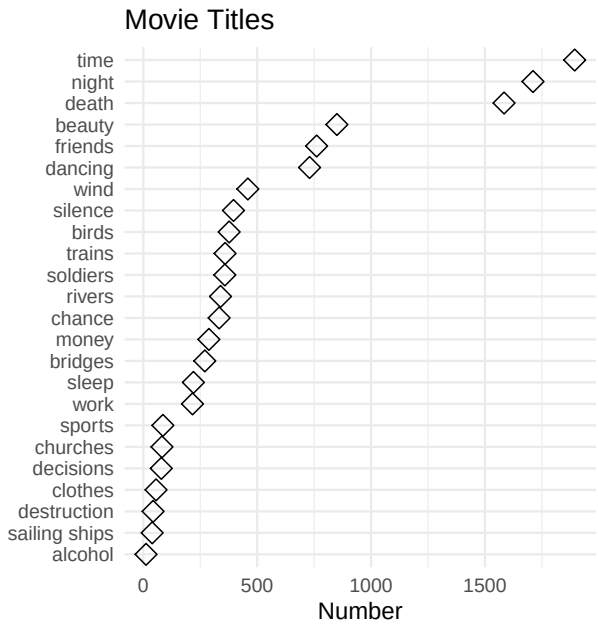
Here are the four preceding plots, whose metrics apply to Themes that share a source in commercial art.



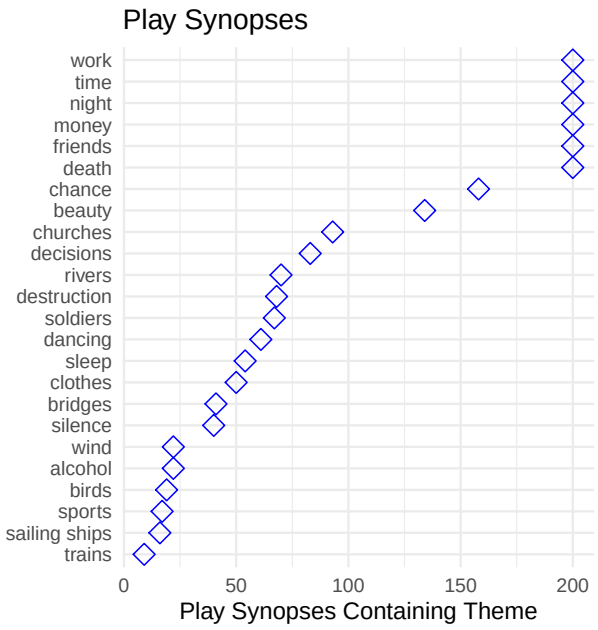
Source −− https://www.worldcat.org/advancedsearch



Source − Music Lyrics Database, http://www.mldb.org/



Source − https://www.imdb.com/search/title/



Source −− http://www.playdatabase.com/login.asp

PhD Dissertations, Count of Instances

The exhaustive, original and long research that becomes a dissertation, evincing the depth, coverage, and quality that merits the award of a doctorate decree, surely suggests complexity.

For this metric, EBSCO Open Dissertations has obtained from 320+ universities around the world more than 1.4 million theses and dissertations (I don't know the breakdown). For each Theme term, the EBSCO site returned results in the form of "Search Results: 1 - 10 of 3,527". I used the final figure.

The quartet below displays the findings in the top left of the quadrant.

Code of Federal Regulations (CFR), Count of Instances

Government regulations and laws focus mostly on serious topics (by which I mean in comparison to the typically more emotional or psychological topics of songs and poems for example). Thus, in classifying this metric, I lean toward complexity.

On the United States government's website for the Code of Federal Regulations, you can query the contents all 70,000+ pages of the CFR. I searched with Find, Title "All titles," and Date "current." The search engine returns, for example, "Showing the most relevant 50 of 454 matching results." The final figure became the metric for the Theme term. The maximum values returned are 10,000 matching results (four Themes hit that ceiling). A future extension of this project might gather comparable metrics from Eurolex, an online compendium of European Union documents.

The graph shows the results for all Themes.

United Nations Documents, Count of Instances

My reasoning for how to interpret documents published in the Code of Federal Regulations applies similarly to documents published by the United Nations. That is, complex topics dominate quotidian topics.

For this metric, I searched the UN's Digital Library. I chose all document types and added all the results that displayed on the left pane (it breaks the results down by several kinds of documents).

The metric lends itself to a scatter-plot, as included below, sorted by increasing numbers of documents.

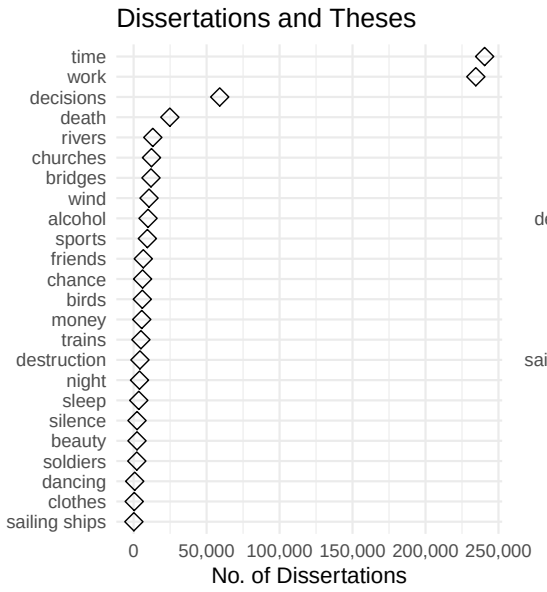
College Majors, Distribution

The 24 Themes generated 412 Subthemes. I categorized each Subtheme according to which of 13 college major I thought would be most likely to address it in courses. The table below lists the Themes, subthemes, major, and the blog post where they are discussed, all sorted alphabetically by major. I regard this metric as weakly complex.

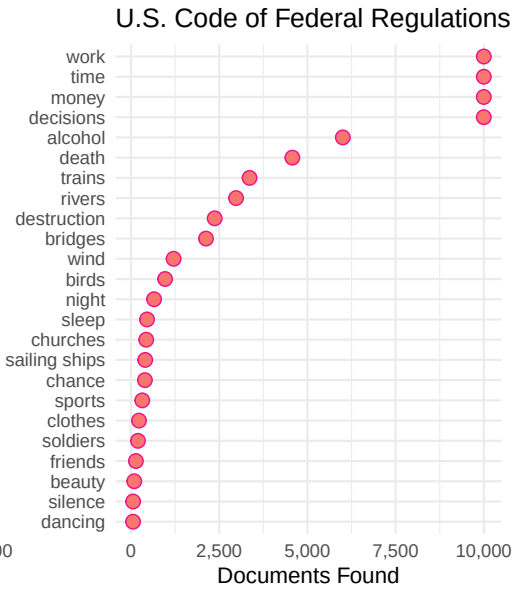
The analysis calculated for each Theme the result of dividing its number of Subthemes by its number of majors. A Theme with seven majors and 14 Subthemes has a score of 2 (14 divided 7); a theme with six majors and 14 Subthemes has a score of 2.33. My hypothesis is that the higher the score, the more complex the Theme, since on this normalized basis, a higher score means more diverse academic areas would deem it of interest.

After all my work to figure out the most common majors, standardize their names, classify Subthemes by those majors – a process I did again after waiting a month, and then picking the consensus major – I tested my process with two friends. So many disputes arose, and so many different interpretations of the task, that I have concluded that this metric lacks credibility. It is not sufficiently empirical. To compound the shakiness of the foundational data by calculating a complexity score makes even less sense. Nevertheless, I have left it in for this series as a record and because with a refined methodology it might deserve inclusion as a metric.

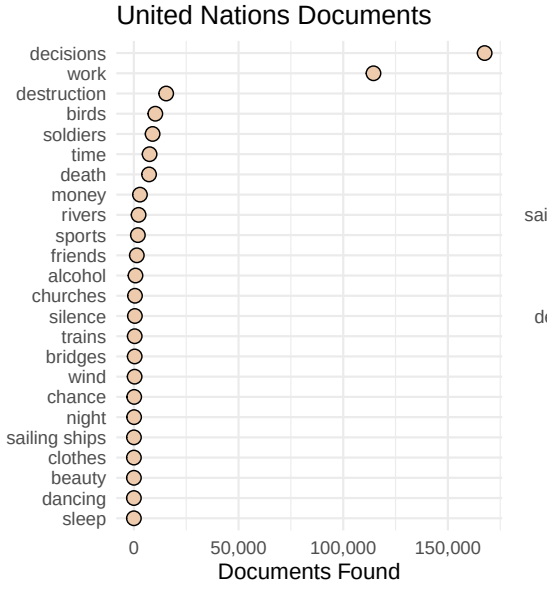
Here are the four preceding plots, whose metrics apply to Themes that lean toward complexity.



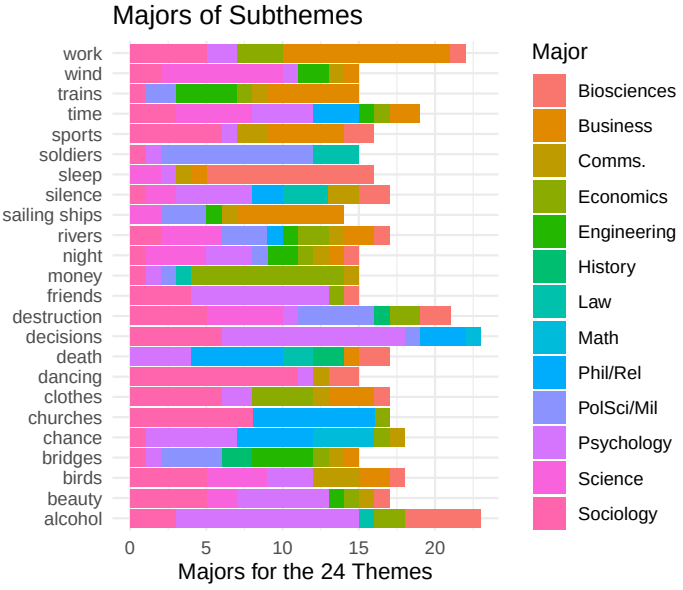
Source: https://biblioboardcom/opendissertations/



Source: https://www.ecfr.gov/search?



Source: https://digitallibrary.un.org/search?



Source: Author's classification

Normalize the Metrics

Once all of the data described above had been collected for the 24 Themes, I created a single dataframe and converted the raw numbers of each metric into a normalized score. This is an important step so that numbers which vary significantly in size, such as the frequency band digits of the Oxford English dictionary, which range from 0 to 8, can be compared to the number of Google search hits, which soar into the billions. The R function scale() subtracts the average of the metric (the mean) from each value and divides that result by the *standard deviation* of the metric. By that calculation, the metrics (a.k.a. variables or features) are centered around zero and have roughly a variance of one (unit variance).

Another method of normalization converts each value into a score from 0 to 1, where the highest value

would have the highest normalized score and the lowest would have the lowest normalized score – but all of them would fall into the zero-to-one range. Another series may test this and other methods of standardizing the metrics.

To give a sense of the normalization, the small table below shows five Themes and three metrics, the latter in their absolute values and to the right their scaled values.

Theme	Book Titles	CFR	Google Hits	Book Titles Scaled	CFR Scaled	Google Scaled
alcohol	95810	6004	105	-0.7163616	0.8835747	0.7439213
chance	21439	394	91	-0.8005721	-0.6596951	0.4529346
beauty	97502	92	51	-0.7144457	-0.7427731	-0.3784559
silence	21851	59	35	-0.8001056	-0.7518511	-0.7110121
work	2012242	10000	42	1.4536223	1.9828449	-0.5655187