Statistical Literacy: Ratio-Based Grammar Research

Milo Schield
New College of Florida

Abstract:
Confounding is a major problem in dealing with statistics obtained from observational studies. The influence of confounders can be blocked in two ways: physical (effect size and study design), and mental (selection and ratios). Ordinary English is typically used to describe and compare ratios. Today's students need guidance on the proper use of English to form descriptions and comparison that are accurate. They have no idea of whether "the percentage of men who smoke" is the same as "the percentage of men among smokers". They have no idea of whether "men are more likely to smoke than [are] women" is the same as "smokers are more likely [to be found] among men than [among] women." or "men are more prevalent among smokers than women."

In 2023, Kendall-Hunt published a statistical literacy textbook that provides provisional classifications and grammatical structures for describing and comparing ratios. In order to provide guidance that is generally accepted, it must be empirically based on common usage. This requires research using a large corpus of text based on written and spoken English.

This paper identifies the data sources used in generating these classifications and structures. And it summarizes the provisional classifications and grammatical rules obtained from this investigation. While this paper indicates the connection between these data sources and the results, it does not – and cannot – detail the thinking and choices involved in obtaining the results from these sources.

Keywords: quantitative rhetoric, quantitative literacy, statistical illiteracy

1. Statistical Literacy Textbook and Ratios

In 2023, Kendall-Hunt published *Statistical Literacy: Critical Thinking about Everyday Statistics*. This textbook is used in teaching Statistical Literacy (Math 1300) at the University of New Mexico where it is now being required for all those majoring in statistics – along with the traditional population-inference course (Math 1350). This textbook is also used in teaching Statistical Literacy (STAN 2720) at New College in Florida.

This statistics textbook is different: less than a 30% overlap with traditional introductory statistics textbooks where statistics are defined as the properties of samples and where the focus is on the use of randomness in population inference: the derivation (binomial distribution, normal distribution, central limit theorem) and the results (confidence intervals and hypothesis tests).

This statistical literacy textbook views statistics broadly: statistics are numbers in context – where the context matters. Since everyday statistics are socially constructed, they can be influenced. In this book, the myriad of influences are classified into four groups: confounding, assembly, randomness and error/bias. The first letter of each of the groups form CARE which is part of good advice in dealing with statistics: "take CARE".

Confounding is arguably a major problem in dealing with statistics obtained from large observational data. This book argues that there are two ways to block the influence of confounders: physically (effect size and study design) and mentally (selection, ratios and standardization).

Ratios are a simple way to mentally take into account the influence of a related factor. The different ways describing ratios can be classified into two groups; those using common preposition (the ratio of men to women was two to one) and those using the preposition 'per' (the number of COVID deaths per thousand population was bigger in Czechia than in South Africa).

Comparing ratios adds a new level of complexity. Named ratios (e.g., chance, rate, percentage, etc.) are useful in making comparison of ratios. These named ratio families are classified into two groups based on whether they used ordinary prepositions or the preposition 'per'. Those using the 'per' preposition were classified into three grammars: percent, percentage and rate grammars. Those using ordinary prepositions were classified into two grammars: ratio and chance. This paper presents the research underlying these choices.

The results of organizing statements describing ratios into to grammatical families are shown in Figure 1. The two families are those on the left using the common prepositions ('of', 'in', 'for', and 'to') and those on the right (the 'per' family) using the less-common preposition: 'per'.

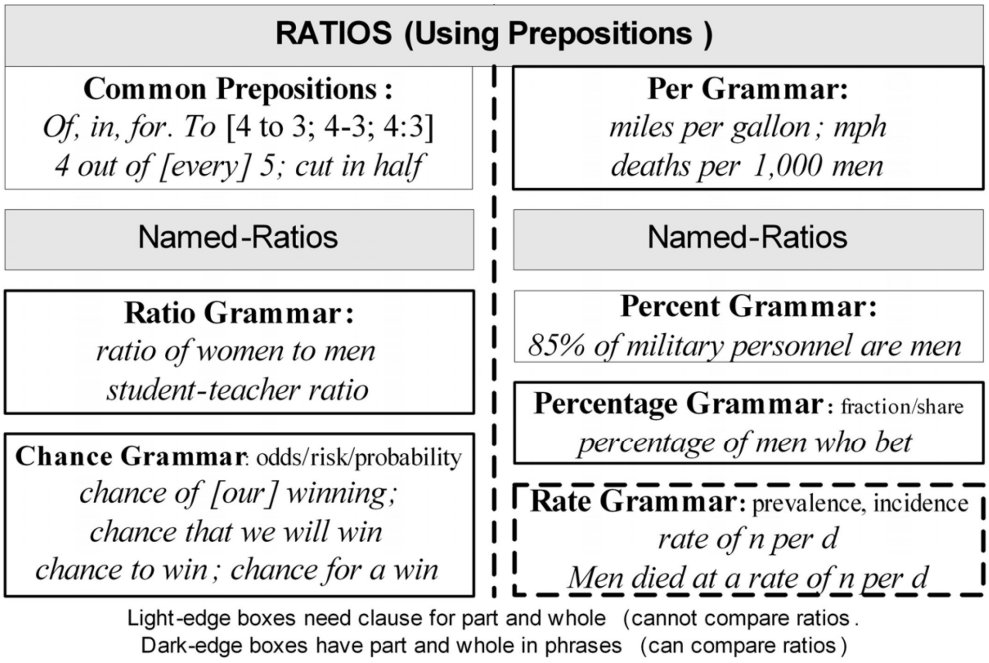


Figure 1: Two Kinds of Ratio-Based Grammar

Notice that a named-ratio family can have multiple members. The 'chance' family includes odds, risk, likelihood and probability. The 'rate' family includes prevalence and incidence. The 'percentage' family includes fraction and share. These families are constructed based on their grammatical similarity. The family name is chosen based on which of the members is the most common in everyday usage.

This classifications are based on research presented in these papers: Schield (2001a, 2001b, 2005a, 2005b, 2011a, 2011b, 2015, and 2024), and Schield and Burnham (2007). These papers include a call for further research by trained grammarians (Schield 2020) and recommendations on teaching grammar in schools (Schield 2024).

Students need help in dealing with the small grammatical details involved in distinguishing part and whole. They are not used to close reading. They soon realize that small differences in wording (syntax) can create big differences in meaning (semantics). Learning this when dealing with quantities may help prepare them for dealing with subtle distinctions involving ordinary words, phrases and statements.

Appendix A gives more detail on the grammar for each of these families.

2. Text-Based Data Sources

The empirical research used to generate the aforementioned results involves three different text-based data sources:

- The Collins-COBUILD corpus
- The English-Corpora (previously the BYU Corpora)
- Google Ngrams accessing the Google Books corpus

Appendix B introduces the Collins-COBUILD corpora. This was arguably the first massive corpus that was machine accessible, that classified every word by part of speech, and that had powerful search capabilities. This was the corpus used to develop everything. Access was expensive. The cheapest was a one-year subscription. This corpus was accessed twice: 1999 and 2010.

Appendix C introduces the English-Corpera. This corpus doesn't have the depth of analysis as did the COBUILD corpera. But the English-corpera was much cheaper, it had multiple sources, and they included more data that did the COBUILD copera. This was used to handle special studies or anything after 2011 when our second one-year subscription to the COBUILD corpus expired.

Appendix D introduces Google Ngrams. Rather that present an in-depth analysis of how to use this powerful tool, this section presents some of the results that were used to identify what form or which one of a group was the most common in everyday usage.

Appendix E shows the results of searching for various words and phrases in the Harvard Business Review. This corpus was not used to find text. It was used to see if the prevalence in the HBR was at all similar to that in COBUILD, in the English-Corpora, or in Google Ngrams.

3. Student Success

While algebraic notation focuses attention on the mathematical relationships, we don't speak algebraically in every day usage. An increasing fraction of our students speak English as their second language. They need special help in dealing with the peculiarities of English.

Some students – especially those in non-quantitative majors – have difficulty with algebraic math. This difficulty handicaps them in dealing with quantitative aspects of an increasing fraction of the college courses that use statistics as evidence.

By giving those students a course that is accessible, topical and one that they see as being valuable, students can find that they are good in dealing with math. Here's some quotes from students taken from Schield (2022):

Here are some of the brief comments: "I like the critical thinking aspect of the class." "I liked that the things we learned were applicable to life - I caught myself several times using information I learned from this class in daily conversation." "I like the critical thinking. I feel like I haven't been able to have some critical thinking in my other classes." "I like reading tables and graphs and I like questions that require critical thinking because I feel like these things are beneficial to know in the real world outside of school." "This is the first time I feel like I'd actually use a math class outside of the classroom regularly." "I enjoyed critical thinking and the news stories. Both provide beneficial knowledge I will take with me into my everyday life." "The critical thinking. I don't have many chances in other classes to be a critical thinker."

Here are some of their more in-depth comments:

> "I like the content and critical thinking aspect of the class. As someone who had to drop the regular stats class I was very happy to have this class as an option. I feel like the content has the overall goal of helping me with critical thinking. There are many times now where I am looking out for parts and whole in people's statements and critically thinking about what people are saying more often. So I would say that the content is also more applicable for life goals. I would have to say that the regular stats class is not very useful to me because the

content will not matter later in my life. Taking that class back in the spring I just was so confounded on why I was needing that class for general ed. But like I said I really appreciate having this class as an option."

"The set up and idea behind the class as a whole are my favorite part, tackling statistics from a different angle that is much more engaging for those who find math subjects to be typically challenging is a brand new approach and one that I think would be beneficial for a broader group of students. Attempting to explain the workings behind statistics has personally allowed me to understand the material much better than I had previously."

"This course is an answer to my prayers, I am a music major and horrible at math so fulfilling my math requirement has been hard. This is the first math class I actually liked. I loved the format and the material is about things I can apply to everyday life. The textbook is fantastic and helped me a lot accompanied by the instruction. I would recommend this class for anyone."

"Definitely not a repeat of AP statistics. The math problems make way more sense than a statistics class. It really helped me begin to think critically about ALL of the statistics regular I hear on the news."

4. Conclusion

Statistical literacy is arguably quantitative rhetoric. (Schmit, 2010) In today's world where statistics are everywhere, learning how to read, interpret and communicate quantitative ideas is increasingly important. In describing and comparing ratios using ordinary English, small differences in syntax can make big differences in semantics. Students – especially those who are not native English speakers – have difficulty communicating quantitatively. Providing empirically-based provisional rules and guidelines is necessary for them to function effectively in today's world of big data and quantitatively-based arguments. Hopefully this first step in text-based research involving quantitative descriptions and comparison will encourage others to verify these results or provide a better set of recommendations and guidelines.

Acknowledgements

Thanks to the W. M. Keck Foundation for providing the financing needed to access the COBUILD corpus, to Tom Burnham (Burnham 2016, 2018) for his years of effort in analyzing how ordinary English is used to describe and compare ratios, and to the many students and teachers who have dealt with this material and provided helpful feedback and suggestions.

References

Burnham, Tom (2018). Remembrance. http://www.statlit.org/pdf/2018-Tom-Burnham-RIP.pdf

Burnham, Tom (2016) Accomplishments www.statlit.org/pdf/2016-Burnham-Accomplishments.pdf

COBUILD Corpus. Overview https://en.wikipedia.org/wiki/COBUILD

COBUILD Corpus. Details. The Bank of English (COBUILD Corpus). https://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5678579368

English-Corpora. Access at https://www.english-corpora.org. Website history at https://www.english-corpora.org/byu-corpora.asp

Google Ngrams: https://books.google.com/ngrams/

Schield, Milo (2000). Statistical Literacy: Difficulties in Describing and Comparing Rates and Percentages. *2000 American Statistical Association Proceedings of Section on Statistical Education*, pp. 76-81. See www.StatLit.org/pdf/2000SchieldASA.pdf.

Schield, Milo (2001a). Describing Rates and Percentages in Tables. Handout, 2001 Business of Communications conference. See www.StatLit.org/pdf/2001SchieldBusOfComm.pdf.

Schield, Milo (2001b). Statistical Literacy: Reading Tables of Rates and Percentages, *2001 American Statistical Association Proceedings of Section on Statistical Education*, [No pages numbers given]. See www.StatLit.org/pdf/2001SchieldASA.pdf.

Schield, Milo (2005a). Statistical Literacy and Chance. *2005 American Statistical Association Proceedings of the Section on Statistical Education.* [CD-ROM] 2302-2310. See www.StatLit.org/pdf/2005SchieldASA.pdf

Schield, Milo (2005b). Quantity Words Without Numbers: Why Students use "Many", 2005 QR Conference> www.StatLit.org/pdf/2005SchieldCarleton.pdf

Schield, Milo and Thomas Burnham (2007). Grammar of Statements Involving "Chance", *2007 American Statistical Association Proceedings of the Section on Statistical Education.* [CD-ROM] pp. 2235-2242. See www.StatLit.org/pdf/2007SchieldBurnhamASA.pdf

Schield, Milo (2011a) Describing Arithmetic Relations Using Informal Grammar. *2011 American Statistical Association Proceedings of the Section on Statistical Education.* [CD-ROM] P. 911-925. See www.StatLit.org/pdf/2011SchieldASA.pdf

Schield, Milo (2011b). Describing Arithmetic Relations Using Informal Grammar: Appendices. www.StatLit.org/pdf/2011SchieldASA-WB.pdf

Schield, M. (2015). Statistical Phrases in the Harvard Business Review Abstracts. Technical paper. www.statlit.org/pdf/2015-Schield-HBR-Statistical-Words.pdf

Schield, Milo (2020) Call for Research on Using Ordinary English to Describe and Compare Ratios. www.statlit.org/pdf/2020-Schield-Call-Research-English-Grammar-Ratios.pdf

Schield, M. (2022). Statistical Literacy Math1300 Year 1 at UNM. *Proceedings of the Section on Statistics and Data Education*. P. 2035-2065. www.statlit.org/pdf/2022-Schield-ASA.pdf

Schield, M. (2024). Using English to help students understand quantitative ideas. *The ATEG (Assembly for the Teaching of English Grammar) Journal*. Vol 32. P 21-30. Copy at www.statlit.org/pdf/2024-Schield-ATEG.pdf

Schmit, J. (2010). Teaching Statistical Literacy as a Quantitative Rhetoric Course. *American Statistical Association* 2010 *Proceedings of the Section on Statistical Education*. p. 2372-2386. www.statlit.org/pdf/2010SchmitASA.pdf

Appendix A: Describing Rates, Percentages and Ratios

The following are excerpts presenting the templates and rules from the Schield (2023) Statistical Literacy textbook (with minor changes). The page numbers are those from that textbook.

PERCENT GRAMMAR:
p. 182: Percent grammar describes a part-whole ratio when there is no other named ratio keyword and the % symbol (or "percent") is followed by "of" or a verb (are). Part, in part-whole ratios, designates the group (men) which if applied to the whole (soldiers) gives the part within that whole (male soldiers).

What can we learn from these questions and their answers?
#1: In every fraction question, there is a whole and a part. The key is to figure out what is part and what is whole. Note: Part may involve something physical, but in ratios or fractions, it always refers to the size or quantity involved.
#2: The questions, "What fraction", "What proportion" and "What percentage" are the same except for units. "What fraction?" gives a decimal fraction. "What percentage?" gives a percent. "What proportion?" can be answered either way.

P. 183: A part-whole percentage is the size of the part as a percentage of the size of the whole.

Grammar of Percent Statements

Of	_____% of _____ are _____ # whole part	20% of students are juniors
Among	Among _____, ___% are _____ whole # part	Among students, 20% are juniors
Among and of	Among _____, ___ % of _____ are _____ whole # whole part	Among students, 20% of men are juniors

Figure 2: Template for Percent Grammar Statements

General Rules involving Ratio Grammar Statements	
1	A part-whole ratio must always have a top (part) and a bottom (whole)
2	*Among* always introduces a whole. [Among W, X% of W are P]
3	Leading prepositional phrases introduce a whole or whole delimiter
4	Leading modifiers take on the part-whole status of what they modify.
Specific Rules for Percent Grammar Statements	
1	"X% of" or "X% are" indicates a part-whole ratio.
2	'X% of' always introduces a whole. ; 'X% are' always introduces a part
3	The part is the subject or predicate opposite the "X%"
4	Relative clauses (trail) take on part-whole status of what they modify

Figure 3: Rules for Percent Grammar Statements

P. 192: MOST GRAMMAR: When 'most' is followed by 'of' or a verb and follows the rules for percent grammar, it indicates a majority. A majority is more than half (50%).

- Most [of the] military personnel are men.
- Among military personnel, most are men.

When 'most' is an adjective preceded by a definite article (such as 'the' or 'a'), it indicate a plurality—not a majority. A plurality is the largest or most numerous.

- In 2019, the most common US baby boy name was Liam.

PERCENTAGE GRAMMAR:
P 193: **Percentage grammar** describes part-whole ratios using the keywords *percentage*, *proportion* or *fraction*. The rules are the same for each.

Defining percentage, proportion and fraction gets technical. For now, focus on their similarities: the keyword is typically preceded by "the." The percentage of men who smoke is 30%. The proportion of the population who are elderly is increasing. The fraction of adults who are bilingual is small.

Percentage grammar is commonly used in the titles for tables and graphs, and in comparisons of percentages.[217] Comparing two percentages using percent grammar creates a very long statement. To read most tables and graphs of percentages, you must understand percentage grammar.

Percentage grammar is different from percent grammar. First, the word "percentage" never follows a number. Percentage typically follows an appositive (The percentage). It may follow an adjective (The highest percentage; an immoral percentage). Second, decoding part-whole in percentage grammar is more difficult.

Percent should only be used when it follows a number, as in 38 percent (38%). Otherwise, the correct word is percentage. Percent is a unit of measure (20%), while percentage is the ratio being measured. Percents are units of measure like inches, seconds or volts, while percentages are attributes or characteristics that are being measured like height, time or voltage. This may seem like a picky distinction, but it's important in understanding statistical evidence. To indicate the part and whole, percent grammar requires a clause while percentage grammar typically needs just a few phrases.

	"Of" is	Leading Preposition	Subject: Whole and Part	Predicate
1	whole		The percentage of __ who* are __ {whole} {part}	is ___% ##
2	absent	Among _____, {whole}	the percentage who* are ____ {part}	is ___ % ##
3	whole2	Among _____, {whole1}	the percentage of __ who* are __ {whole2} {part}	is ___ % ##
4	part	Among _____, {whole}	the percentage of _____ {part}	is ___ % ##

* Other relative pronouns beside *who* include *that, which, what, when* and *where*.

Percentage grammar without "percentage": The <##>% of <whole> who are <part> is <predicate>.

Figure 4: Template for Percentage Grammar Statements

Relative pronouns (who, that and which as well as what, where or when) refer to prior nouns or noun phrases and introduce relative clauses. Who refers to people; that refers to people, animals and things; which refers just to animals and things.

P. 195: What percentage of men are MIS majors? This question can be asked in three ways:

1. What is the percentage of male students who are MIS majors?
2. What percentage of male students are MIS majors?
3. What percentage of students who are males are MIS majors?

The first question ("what is the percentage") involves percentage phrase grammar. The second and third questions ("what percentage") involve percent grammar even though they use the keyword "percentage." But they don't use "the percentage."

P. 198 Sports Grammar: Sports reports often use sports grammar as a short form of percentage grammar. A good example is "the percentage of passes completed" or "the percentage of completed passes." Both phrases mean "the percentage of passes that were completed." A "pass completion" cannot occur without a "pass", so "pass" is the natural whole and "completion" is the natural part.

When there is no natural whole, this abbreviated form is ambiguous.

- percentage of female smokers; percentage of working males
- percentage of infant deaths; percentage of single women

Portion or Ingredient grammar: <Whole> is X% <part>. (The human body is 65% water). This grammar applies to portions that are measurable (Butter is 80% fat). To avoid ambiguity, do not use ingredient grammar with attributes of individual subjects (Texans are 35% Mexican). Use attribute grammar (35% of Texans are Mexican). Exceptions (He got 85% right; 85% of his answers were right).

P. 216: There are other forms using percentage grammar without the keyword percentage (the 55% of college students who are female is increasing). Other forms of percentage grammar use "to" indicating the part: "Percent of births to unmarried women" instead of the "percentage of births that involve unmarried women".

RATE GRAMMAR:
P. 219: Rates come in four kinds: frequency, prevalence, incidence and growth. The growth rate is common in business and economics.

1. Frequency or velocity (rate): events per unit time (heart rate or speed.228
2. Prevalence (rate): the ratio of two counts: group count divided by population count at the same time (unemployment rate).
3. Incidence (rate): a relative frequency such as number of events in a time interval per group size (2020 birth rate per 1,000 population).
4. Growth rate: percentage or count change in amounts per unit time.

Rate grammar can get complex. Here are three things to know.

#1: The most obvious thing to note is time. Prevalence is the only rate that is cross-sectional: at a moment in time. The other three rates—frequency, incidence and growth—all involve a time period or interval. To say "the immigration rate doubled" may be incomprehensible. With four rates, we have no idea of which one is being used without more context.

#2: Second, some rates can go negative or exceed 100%; others cannot. Growth rates (interest rates) can exceed 100% and can go negative. Prevalence rates (obesity and unemployment rates) cannot go negative or exceed 100%. Incidence rates cannot go negative, but as ratios, they can exceed unity.

#3: Third, rates and ratios are similar mathematically, but they are quite different grammatically. We say the female-to-male ratio is 4 to 3 (but not the female-to-male rate). We say the unemployment rate, the homicide rate or the mileage rate (but not the unemployment ratio, the homicide ratio or the mileage ratio). We say the ratio of doctors per 10,000 population in England (not the rate of doctors). We can get combinations. "The male-female ratio of suicide rates is increasing."

#4: Fourth, some of the adjectives modifying the part adjective preceding the keyword rate should be expressed in the possessive voice. The male death rate probably means "the male' death rate" or "the death rate of males."

P. 222: Rates can be described using two different grammatical forms. **Phrase rate grammar** describes a rate using just phrases. **Clause rate grammar** describes a rate using an entire clause: a verb separates the part and whole.

Here is a rate using phrase grammar: "In 1995, the accidental death rate for males was 91 per 100,000", the main verb is 'was'. All that follows was is the numerical value, so both the part and the whole must be before was.

Here is a rate using clause grammar: Male students are majoring in Arts or Science at a rate of 48 out of 200. The main verb separates the whole (male students) from the part (majoring in Arts or Science) so the entire clause is needed.

P. 223 Phrase-Based Rate Grammar

General Rules for Per and Rate Grammar

1. Any countable following the keyword *per* is always a whole.
2. Named rate grammar include *rate*, *incidence*, and *prevalence*.
   Incidence is a ratio involving a time interval (death rate, birth rate).
   Prevalence is a ratio taken at a moment in time (unemployment rate).
3. In phrase rates, *rate of* introduces either a part or whole.
   The word preceding *rate* is a part if it is countable*.
   Rate of introduces a whole if the word before a rate is countable*.
   Rate of introduces the part if the word before rate is *not* countable*.

*Uncountable: appositive (the), quantity (8%), evaluation (low) or possessive (men's).

Figure 5: Rules for Per and Rate Grammar Statements

Grammar template for phrase-based rate descriptions

R1: The _____rate of _____ is____ per _____.
part whole # #

R2: Among _____, the rate of _____ is ___ per ___.
whole part # #

Figure 6: Template for Phrase-Based Rate Grammar

P 226: Clause-based rate grammar:

Grammar template for clause-based rate descriptions

Clause rate grammar can have either the whole or part as the subject. [1]

Main clause rate grammar: "at a rate"

R3: Subject (whole) verb (part) at a rate of N per M.
Men die accidentally at a rate of 91 per 100,000 per year.

R4: Subject (part) verb (intransitive)
Accidental death occurs [among/to/for men] at a rate of N per M [whole]

Subordinate-clause rate grammar: "rate at which". [2]

R5 The rate at which subject (whole) verb (part) is N per M.
The rate at which males died accidentally is 91 per 100,000 per yr.

R6 The rate at which subject (part) verb (intransitive) is N per M whole.
The rate at which accidental deaths occur among/to/for males is …

[1] Other forms are possible. E.g., "The U.S. rate is 61 deaths per 100,000 males/year."
[2] Subordinate clause forms follow grammatically from the main clause descriptions.

Figure 7: Template for Clause-Based Rate Grammar Statements

P. 127: CHANCE GRAMMAR. P. 127: Chance grammar describes part-whole ratios using the keywords chance, risk, odds, probability and likelihood. Since all five keywords use a common syntax, they form a family: a family this textbook calls the chance grammar family.

According to Google nGrams, risk is more common than chance since 1980 while chance is more common than the other three. Since risk is often associated with just negative or bad outcomes, chance was chosen to name this family. Chance grammar is arguably the most common of all the named-ratio grammars. However, it typically implies a future event whereas rate and percentage focus on past or present events. This may be why chance grammar is not used in the titles of graphs or tables.

P 228: Chance grammar has two forms: clause and phrase. Let's start with clause. It is easier to decode; the subject is a concrete (students), and it is more common.

C1. _____ have a __% chance to/of _____
{whole} # {part}
Startups have a 10% chance to succeed (of succeeding)
Startups have a 10% chance of [being a] success.

Figure 8: Template for Clause-Based Chance Grammar Statements

In the phrase form, the chance phrase and the numerical percentage are always separated by the main verb. One is the subject; the other is the predicate. Chance phrase grammar is similar to percentage grammar in form. In both cases, the named ratio is followed by a preposition (of) or a relative pronoun (that, who).

C2: The chance to/of/at _____ in/among/for _____ is ____%.
{part} {whole} ##
The chance of success (succeeding) among startups is less than 30%.
The chance of a head when randomly flipping a fair coin is 50%

C3: _____ chance of _____ is ____ %.
{Whole possessive} {part} ##.
Our team's chance of winning is 70%.

C4. The chance that/of _____ is|verb _____ is ____%.
{whole} {part} ##
The chance that a randomly selected male is a senior is about 20%.
The chance of a randomly selected male being a senior is about 20%.

Figure 9: Template for Phrase-Based Chance Grammar Statements

P 228: There are other forms. "Our team had a 70% chance of winning."

In the phrase forms, the keyword chance is followed by an infinitive (to), a preposition (of) or a relative pronoun (that). The relative pronoun always introduces a clause; the others can introduce either a phrase or a clause.

When the infinitive or preposition introduces a phrase, the phrase is the uncertain outcome (part). When the chance grammar introduces a clause, the result is determined by the meaning of the words or is simply ambiguous.

P 229: Exceptions and ambiguities make phrase grammar problematic. Exception: "The chance that success will be ours." *Success* is the subject and the uncertain outcome. Ambiguity: "What is the chance that a head on the first flip is followed by a head on the second?" When statistical educators were asked this question, half chose 50%, a fourth chose 25% and rest did not respond.[233] The question was ambiguous.

Chance is either the premise or the conclusion, or the statement is ambiguous.

- Premise: An outcome is unlikely if due to chance
- Speculative conclusion: An outcome is unlikely to be due to chance.
- Ambiguous: This outcome is unlikely due to (by) chance

Note the ambiguity in this statement: “First convince us that a finding is not due to chance, and only then, assess how impressive it is.”

Chance grammar is ambiguous in a second way. It may simply describe an existing or historic rate or percentage in a group. Or it may entail an inference: a generalization, a prediction, a specification or a causal claim.

RATIO GRAMMAR

P 230: A ratio is a quantitative relationship between two things involving or implying multiplication or division. Ratio grammar describes a ratio using prepositions (nine men for every woman), per (miles per hour, 5 per cent), or the keyword ratio (the male-female ratio was 9:1).

Preposition-based ratios can be classified by whether the prepositions are common (of, in, for, over, by, to) or uncommon (per).
Common prepositions: Proper (Part/whole):

- Of: Literal (one of 9 kids; page 1 of 3); Figurative (one of two sales)
- In: Literal (lost one game in six); Figurative (one in five crimes)
- Out of: Literal (oldest out of six). Figurative (four out of five crimes).

Common prepositions: Improper (Value/Value):

- In: Six hits in five innings. For: Split the shares three for two.
- For every: For every women there are three men.
- Over: Blood pressure is 200 over 150 (200/150).
- By: Animals entered the ark two-by-two.
- To240 or "-to-": "two to one" or "two-to-one." "To" introduces the base.
- Symbols for "to": the dash (We won 3-2) and the colon (score of 3:2).

Fractions can also be classified by whether they are literal or figurative. Literal: 40 of the 50 doctors sampled recommend Crest toothpaste. Figurative: "four out of five doctors recommend Crest toothpaste" where doctors is a larger figurative group than the five that were mentioned. In proper fractions, the literal and figurative use the same grammar. The "part out of whole" is a common way to teach children fractions. It allows children to move from a literal use to a figurative use using familiar words. The following ratios use the uncommon preposition "per:"

Per preposition: Proper (Part/whole):

- Prevalence: Six unemployed per 100 civilian workers.
- Incidence (combo): one escape per year per 10 million prisoners

P 231: Per preposition: Improper (Value/Value):

- Per cent: Sale: 50% off. Interest rate of 6 percent per year.
- Average: Average income per family; Federal debt per capita
- Exchange: $ per ounce of gold; 82 ounces of silver per ounce of gold
- Incidence (velocity): deaths per year, miles per hour
- Per by symbols: the slash (2 / hundred). Typically for small ratios.
  Other: accidents per million flights, miles per gallon; gallons per mile

Ratio Grammar Using the Keyword Ratio

The use of the keyword ratio clearly indicates the presence of a mathematical ratio: a numerator (N or top) divided by a denominator (D or bottom): where the denominator is the unit of measure—the denomination (Euro, Dollar, etc.)

There are two ways to present a general ratio when using the keyword ratio:

- Trailing modifier: the ratio of Blacks to Asians (the ratio of N to D).
- leading modifier: the male to female ratio (the N to D ratio)

Note that "of" introduces the numerator while "to" introduces the denominator.

The numerator and denominator of a ratio can be indicated as follows:

1 Of N following *ratio* (ratio of N) identifies N as the numerator.
  E.g., The ratio of N to D, the ratio of men to women.
2 *To, for or per* D following *ratio of N* identify the denominator
  E.g., ratio of Dollars to Euros, ratio of defaults per 100 loans.
3 A hyphen (-), slash (/) or colon (:) in a phrase modifying *ratio*
  indicates the presence of *to* (e.g., price-earnings ratio, Euro/Dollar ratio).
4 Substantive adjectives that modify *ratio* and precede *to* identify the numerator
  (e.g., price-to-earnings ratio)
5 If an adjective phrase modifying *ratio* is a single word that is countable,
  then it typically indicates the numerator; the denominator must be
  inferred from the context. (e.g., the debt ratio)

Figure 10: Rules for Ratio Grammar

Appendix B: COBUILD Word Banks Corpus

Grammatical rules should be based – whenever possible – on contemporary use. In 2004, the best machine-readable source of a written and spoken English where words were identified by their part of speech was arguably the Harper –Collins' Word Banks corpus.

Harper Collins is arguably a world leader in publishing books (reference and textbooks) involving English grammar. COBUILD is an acronym for Collins Birmingham University International Language Database. This a British research facility set up at the University of Birmingham in 1980 and funded by Collins publishers. See Cobuild (corpus).

This COBUILD corpus is large (200 million words in 2010). More importantly it classified every word into the various parts of speech. Most importantly, it allows online access with very powerful search options allowing users to focus in on selected grammatical patterns.

That corpus was first accessed in 2004 and provisional rules were identified. Those rules were taught and tested. In 2009-10, this corpus was accessed more systematically. This time 205 words or phrases were searched with 175 having their content downloaded.

File Naming:
Establishing a simple but accurate file naming system was critical when dealing with over a thousand downloads. The reports were divided into three categories:

- F1: These involve part-of-speech tags and the search word (the null word).
- F2: These involve part-of-speech tags, the null word and other words.
- F3: These had no POS tags but involved the null word and other words.

Establishing the location of words relative to the null used left (before) or right (after)
So, L1 was left one word from the null. R2 was the second word right of the null.

To understand the entire process, first consider the types of reports (and their names) generated for 'rate' and then consider the list of the 205 search phrases.

RATE	ID	Content	---------- file name ----------	Col 1	Col 2	Col 3	Col 4
F1 (3)	2040	word (1), tag(1)	2040Rate-Freq1tL1wN0.txt	tag L1	**word N0**	Freq	
		word (1), tag(1)	2040Rate-Freq1wN0tR1.txt	**word N0**	tag R1	Freq	
F2 (2)	2040	word (2), tag(1)	2040Rate-Freq2wL1tL1wN0.txt	**word L1**	tag L1	**word N0**	Freq
2 words+tag		word (2), tag(1)	2040Rate-Freq2wN0wR1tR1.txt	**word N0**	**word R1**	tag R1	Freq
F3 (4)	2040	word (2)	2040Rate-Freq3wL1wN0.txt	**word L1**	**word N0**	Freq	
No tag		word (3)	2040Rate-Freq3wL1wN0wR1.txt	**word L1**	**word N0**	**word R1**	Freq
		word (2)	2040Rate-Freq3wN0wR1.txt	**word N0**	**word R1**	Freq	
		word (3)	2040Rate-Freq3wN0wR1wR2.txt	**word N0**	**word R1**	**Ford R2**	Freq
List	2040	Rate	2040Rate-List.txt	Lines of actual text. Read in MS Word Pad			

L1: word first left of node *N0 The node (key word)* *R1: Word first right of node*
t symbol for tag *w symbol for word*

Figure 11: Names and Types of Content in Summary Reports for Rate

To better understand these reports, here are these same summary reports with the 10 most common entries for each one.

The following F1 reports showed the most common parts of speech for nearby positions.

Rate-Freq1-tL1-wN0

*2040Rate-Freq1tL1wN0.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

tag     word    Freq
------------------------------------------
NN      rate    28083
DT      rate    11587
JJ      rate    11346
JJR     rate    1586
VVN     rate    1199
NP      rate    866
VVG     rate    765
JJS     rate    733
NNS     rate    721
POS     rate    427
```

Rate-Freq1-wN0-tR1

*2040Rate-Freq1wN0tR1.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

word    tag     Freq
------------------------------------------
rate    IN      24243
rate    SENT    5665
rate    ,       4946
rate    NN      3994
rate    NNS     2652
rate    VBZ     2598
rate    CC      1600
rate    VVD     1321
rate    MD      1248
rate    VVZ     1218
```

Rate-Freq1-tL1-wN0-tR1

*2040Rate-Freq1tL1wN0tR1.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

tag     word    tag     Freq
------------------------------------------
NN      rate    IN      8557
DT      rate    IN      6616
JJ      rate    IN      5378
NN      rate    SENT    3267
NN      rate    ,       2293
NN      rate    VBZ     1802
NN      rate    NN      1764
JJ      rate    SENT    1413
NN      rate    NNS     1240
DT      rate    ,       1169
```

Figure 12: F1 Summary Reports with Top 10 Content for Rate

The F2 reports showed the most common word-POS combinations for nearby positions.

Rate-Freq2-wL1-tL1-wN0

*2040Rate-Freq2wL1tL1wN0.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

word    tag     word    Freq
------------------------------------------
the     DT      rate    6011
interest        NN      rate    4470
a       DT      rate    2162
exchange        NN      rate    1993
growth  NN      rate    1822
any     DT      rate    1486
heart   NN      rate    1381
unemployment    NN      rate    1181
annual  JJ      rate    1086
tax     NN      rate    998
```

Rate-Freq2-wN0-wR1-tR1

*2040Rate-Freq2wN0wR1tR1.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

word    word    tag     Freq
------------------------------------------
rate    of      IN      14618
rate    .       SENT    5461
rate    ,       ,       4946
rate    is      VBZ     2587
rate    for     IN      1988
rate    in      IN      1984
rate    and     CC      1289
rate    at      IN      1147
rate    was     VBD     1111
rate    to      TO      976
```

Rate-Freq3-wL1-wN0

*2040Rate-Freq3wL1wN0.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

word    word    Freq
------------------------------------------
the     rate    6011
interest        rate    4471
a       rate    2162
exchange        rate    2006
growth  rate    1822
any     rate    1486
heart   rate    1381
unemployment    rate    1181
annual  rate    1086
tax     rate    998
```

Figure 13: F2 Summary Reports with Top 10 Content for Rate

The F3 reports showed what words were most common for nearby word positions.

Rate-Freq3-wL1-wN0-wR1

*2040Rate-Freq3wL1wN0wR1.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Frequency limit: 0

word    word    word    Freq
------------------------------------------
the     rate    of      3720
a       rate    of      1210
any     rate    ,       777
annual  rate    of      701
growth  rate    of      596
the     rate    at      458
high    rate    of      432
The     rate    of      411
subscription    rate    .       346
interest        rate    on      330
```

Rate-Freq3-wN0-wR1-wR2

*2040Rate-Freq3wN0wR1wR2.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Sort:  word in right context
# Frequency limit: 0

word    word    word    Freq
------------------------------------------
rate    of      return  772
rate    in      the     698
rate    of      inflation       645
rate    of      growth  614
rate    at      which   609
rate    of      interest        596
rate    .       The     592
rate    ,       the     443
rate    for     the     359
rate    of      the     343
```

Rate-Freq3-wN0-wR1

*2040Rate-Freq3wN0wR1.txt - Notepad

File Edit Format View Help

```
# Frequency list
# Corpus: preloaded/wbo-english.conf
# Query:  word,[word="(?i)rate" & tag="NN"]
# Sort:  word in right context
# Frequency limit: 0

word    word    Freq
------------------------------------------
rate    of      14618
rate    .       5461
rate    ,       4946
rate    is      2587
rate    for     1988
rate    in      1984
rate    and     1289
rate    at      1147
rate    was     1111
rate    to      976
```

Figure 14: F3 Summary Reports with Top 10 Content for Rate

In each case, the focus was on what was most common. We used the most common words as our examples for a given key word.

Even if our rules weren't all-encompassing, they should address the most common uses.

Finally, we went to the actual text in which these words or phrases appeared. Normally, we downloaded the node in the middle of a 100 word character string.

Seeing 100 word text as single lines involved using MS Wordpad in landscape mode with an 8 point Courier (fixed width) font is still daunting.

```
usbooks                                                judge ruled that these
individuals could be admitted . Selected International Comparisons Indicators
Canada France Germany* Japan United Kingdom United States Percentage of low-birth-
weight babies 1990 6 5 6 6 7 7 Infant mortality rate ( per 1,000 live births )
1990 6.8 7.4 7 4.6 7.9 9.1 Teen birth < rate /NN > ( per 1,000 teens ) Selected
years 23.1 ( 1988 ) 9.5 ( 1988 ) 10.3 ( 1988 ) 3.5 ( 1989 ) 31.8 ( 1989 ) 54.8 (
1988 ) Percentage of appropriate age group enrolled in secondary education 1988 --
-1989 93 83 85 96 79 88 Percentage
```

WordBanks did have an option to select only the sentence containing the search phrase. However, this sometimes left out relevant context from an adjacent sentence. So, we normally selected the context manually in short single lines as follows:

```
Brbooks  We 'll have to give you a season ticket at this < rate /NN > !
brbooks    base rates were linked directly to ` bank < rate /NN > " , but
brbooks   the position is a net liability at the year-end < rate /NN > #3,670.
Sunnow  PLAYER THIERRY HENRY 374 MINUTES ON PITCH, CONVERSION < RATE /NN > % 21 %
indnews    The revised rates of interest are: ( < Rate /NN > % per annum )
brmags  the falling fertility < rate /NN > 'problem '
usbooks   they say the < rate /NN > 's 500 bucks a key .
times  infant mortality has fallen to a < rate /NN > of 6.9 deaths per 1,000.
brbooks   The < rate /NN > at which a business sells its products
```

We then grouped these lines based on their similarities having the same structure. This requires judgement. Tom and I worked independently. When we were done, we compared our results. I tended to focus on what was more visually distinctive in syntax: clause vs. phrase. Tom tended to focus on what was most mentally similar semantically. I agreed that semantics is more fundamental than syntax. But I argued that syntax was generally more readily accessible than semantics – to which Tom agreed. We jointly agreed in saying "Small differences in syntax can create big differences in semantics."

Having shown the process for a single search word (rate), now consider the scope of the process for this research. The goal was to study all words used to describe and compare quantities: whether counts or measurements individually or as ratios.

In our second access to Cobuild's corpus, we searched on 205 phrases. We selected 182 phrases for downloading and created 852 summary reports for a total of 1,034 downloads.

Downloading all this data was an intensive manual effort. But this was just the first step in combining our analysis for various structures involving the same keyword into two or three sub-groups.

The next step was to group keywords having similar structures into families. All of this requires time (lots of time) and judgement (lots of judgement)!

Scope of Downloads

The complete list of all words and phrases searched is presented in the following five pages. Along with the search word or phrase (SEARCH), this report shows the number of lines of text (LINES), the identification number of that search (ID), the type of data involved (noun, preposition, CQL, etc), the size of the listing downloaded (List), the number of reports by type (F1, F2, and F3), the total number of summary reports (ALL), the ID # for this search in the first use of WordBanks and the "search criteria".'

12/2009 - 9/2010 — **Schield-Burnham Grammar Research** — Cobuild Corpus
182 Names with Lists — **1,034 Downloads** — 852 summaries

LINES	ID	SEARCH	TYPE	List	F1	F2	F3	ALL	Old ID	Search Criteria	1,034 =852+182
5,657,565		SUMMARY		Width	176	149	435	852			219 Search Names
											182 Search w download
202,197	2000	# tagIN 3#	CQL	**S100**			1	1		[tag="CD"] [tag="IN"] []{0,3} [tag="CD"] within <s>	
6,400,000	2001	XslashX	CQL	**S100**			2	2		".*/.*"	
22,040	2001	x# slash x#	CQL	1				0		[tag!="CD"] "/" [tag!="CD"]	
309,183	2002	XdashX	CQL	**S100**			1	1		".*-.*"	
780,976	2002	x# dash x#	CQL	1				0		[tag!="CD"] "-" [tag!="CD"]	
1,800,000	2003	x# colon x#	CQL	***1***				0		[tag!="CD"] ":" [tag!="CD"]	
1,452,000	2003	CD colon CD	CQL	***1***				0		[tag!="CD"] ":" [tag!="CD"]	
100,093	2004	CD3 to CD	CQL	1				0		[tag="CD"] []{0,3} "to" [tag="CD"] within <s>	
11,568	2005	CD3 by CD	CQL	1			1	1	2010	[tag="CD"] []{0,3} "by" [tag="CD"] within <s>	
81,641	2006	CD3 in CD	CQL	***1***			1	1	2010	[tag="CD"] []{0,3} "in" [tag="CD"] within <s>	
7,554	2007	CD3 out of CD	CQL	1			2	2	2010	[tag="CD"] []{0,3} [word="out"] [word="of"] [tag="CD"] within <s>	
49,957	2008	CD3 for CD	CQL	***1***			1	1	2010	[tag="CD"] []{0,3} "for" []{0,3} [tag="CD"] within <s>	
232,397	2010	**per**	Preposition	**S10**	**3**	2	4	9			
57,564	2011	per (~cent)	CQL	1	3	2	4	9		per [word!="cent"]	
5,185	2015	times a		1			2	2	1004		
236	2016	times each		1			2	2	1004		
124	2017	times every		1			2	2	1004		
6,494	2020	**ratio**	Noun Sing.	1	3	2	4	9			
1,997	2021	ratio of	Phrase	1	3	2	5	10		[word="(?i)ratio" & tag="NN"]	
946	2022	AdjRatio	CQL	1		2	2	4		[tag="JJ"] "ratio" within <s>	
2,479	2023	NounRatio	CQL	1		2	2	4		[tag="N*"] "ratio"	
612	2024	To_Ratio	CQL	1				0		to []{0,4} "ratio" within <s>	
2,808	2028	Ratio_PRE	CQL	1			1	1		ratio [tag="IN"] within <s>	
1,684	2030	ratios	Noun plura	1	3	2	4	9			
59,991	2040	**rate**	Noun Sing.	1	3	2	4	9			
14,618	2041	rate of	Phrase	1	3	2	4	9			
1,087	2042	at a rate	Phrase	1	3	2	4	9			
609	2043	rate at which	Phrase	1	3	2	5	10			
1,817	2044	rate_perXcent	CQL	1			2	2		rate []{0,9} "per" [word!="cent"]	
11,345	2045	AdjRate	CQL	1	3	2	2	7		[tag="JJ"] "rate" within <s>	
28,086	2046	NounRate	CQL	1	1			1		[tag="N*"] "rate"	
24,312	2048	rate_PRE	CQL	1	3	2	2	7		rate [tag="IN"]	
36,627	2049	rate_!PRE	CQL	1	3	2	2	7		rate [tag!="IN"] within <s>	
45,549	2050	rates	Noun plura	1	3	2	4	9			
1,655	2060	prevalence	Noun Sing.	1	3	2	4	9			
3,202	2070	incidence	Noun Sing.	1	3	2	4	9			
75,810	2080	**percent**	Noun Sing.	1	3	2	4	9			
23,653	2081	percent of	Phrase	1	3	2	5	10			
	2082	percentXof	CQL							percent [word!="of"]	
	2083	% attributed|able to	CQL							percent [word!="attributed"|attributable"] "to"	
	2084	accounted for x% of	CQL							"accounted for" [Word="CD"] "percent of"	
130,347	2090	**%**	noun	1	3	2	4	9			
23,788	2091	% of	Phrase	1	3	2	5	10			
	2092		CQL							"%" [word!="of"]	
	2093	% attributed to	CQL							"%" [word!="attributed"|attributable"] "to"	
	2094	accounted for X% of	CQL							"accounted for" [Word="CD"] "% of"	
174,670	2100	**per cent**	Phrase	**S10**	**3**	2	4	9			
43,558	2101	per cent of	Phrase	1	3	2	4	9			
131,111	2102	per CentXof	CQL	1	3	**2**	4	9		per "cent" [word!="of"]	
	2103	% attributed to	CQL							per cent [word!="attributed"|attributable"] "to"	
	2104	accounted for X% of	CQL							"accounted for" [Word="CD"] "per cent of"	

Figure 15: Downloads and Summaries of Statistical Phrases: Page 1

LINES	ID	SEARCH	TYPE	List	F1	F2	F3	ALL	Old ID	Search Criteria	1,034
159	2110	percentile	Noun Sing.	1	3	2	4	9			
14	2120	percentiles	Noun plura	1	3	2	4	9			
13,608	2130	**percentage**	Noun Sing.	1	3	2	4	9			
5,396	2131	Percentage of	Phrase	1	3	2	4	9			
537	2132	percentage_who	**CQL**	1			3	3		percentage []{0,10} "who" within <s>	
718	2133	percentage_that	**CQL**	1			3	3		percentage [] {0,10} "that" within <s>	
1,040	2140	percentages	Noun plura	1	3	2	4	9			
10,601	2150	**percentageXpts**	**CQL**	1	1	2	2	5		percentage [word!="point" & word!="points"]	
	2151	**percentagexXpts**	**CQL**							percentages [word!="points"]	
7,120	2170	portion	WordUNS	1				0			
2,393	2180	portions	WordUNS								
4,035	2190	**fraction**	Noun Sing.	1	3	2	4	9			
401	2200	fractions	Noun plura	1	3	2	4	9			
52,776	2210	**share**	Noun Sing.	1	3	2	4	9			
47,297	2220	shares	Noun plura	1	3	2	4	9			
9,095	2230	proportion	Noun Sing.	1	3	2	4	9			
2,779	2240	proportions	Noun plura	1	3	2	4	9			
87,635	2250	**chance**	Noun Sing.	1	**3**	2	4	9			
33,161	2251	Chance to	Phrase	1	3	2	4	9			
54,404	2252	chance ~to	**CQL**	1				0		chance [word!="to"]	
17,041	2253	chance of	Phrase	1	3	2	4	9			
2,874	2254	chance for	Phrase	1	3	2	4	9			
2,211	2255	chance that	Phrase	1	3	2	4	9			
1,561	2256	chance at	Phrase	1		1	3	4			
1,541	2257	chance in	Phrase	1			4	4			
25,598	2260	chances	Noun plura	1	3	2	4	9			
1,848	2261	chances to	Phrase	1	2	2	4	8			
7,085	2263	chances of	Phrase	1	2	2	4	8			
824	2264	chances for	Phrase	1	3	2	3	8			
386	2265	chances that	Phrase	1	3	2	1	6			
54,307	2270	risk	**noun**	1	1	2	3	6			
17,112	2271	**risk of**	phrase	1		2	3	5			
1,294	2272	risk that	phrase	1				0			
13,709	2280	risks	Noun plura	1	2	2	4	8			
11,374	2290	odds	Noun plura	1	2	2	3	7			
4,198	2300	**likelihood**	**noun**	1	2	2	3	7			
2,361	2301	likelihood of	phrase	1							
783	2302	likelihood that	phrase	1							
2,682	2310	probability	**noun**	1	2	2	4	9			
910	2311	probability of	phrase	1		1	3	5			
340	2313	probability that	phrase	1	2	2	2	7	2010		
380	2320	Probabilities	**noun**	1	2	2	4	9			
74,813	2330	likelyADJ	adjective	1	2	2	3	8			
53,513	2331	likelyJJ-to	CQ								
3,238	2332	likelyJJ-that	CQ								
15,668	2340	likelyADV	adverb	1	2	2	3	8			
1,759	2350	prevalent	adjective	1	2	2	3	8			
323	2351	more prevalent	phrase	1				1	2010		

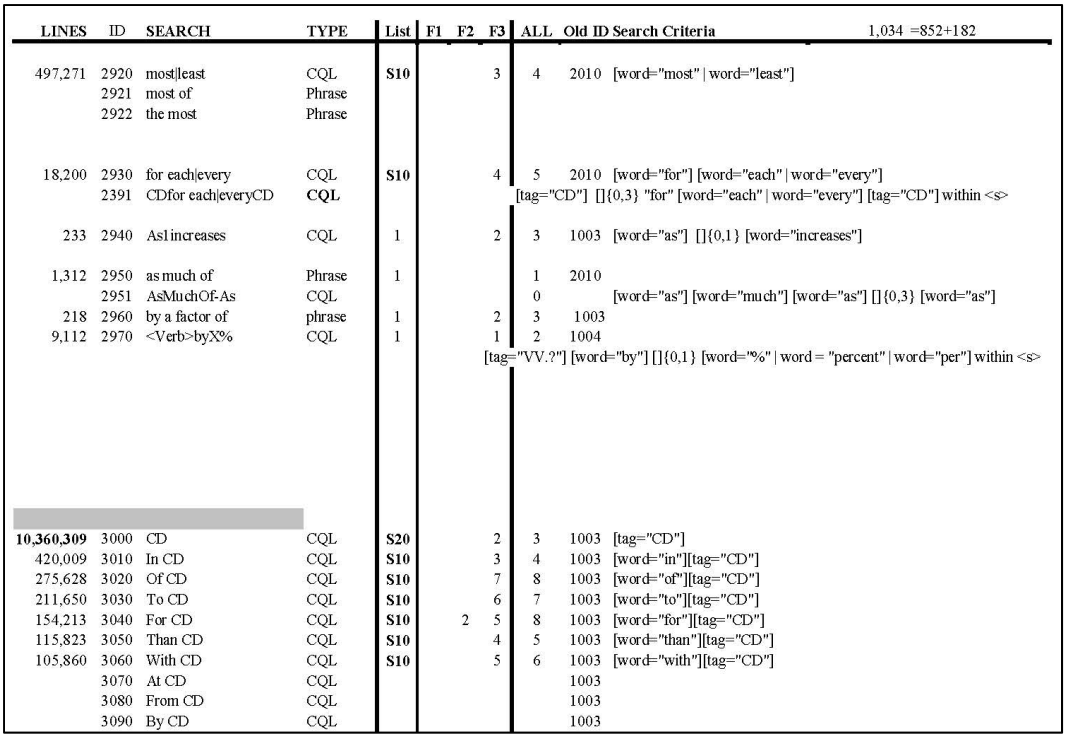
Figure 16: Downloads and Summaries of Statistical Phrases: Page 2

LINES	ID	SEARCH	TYPE	List	F1	F2	F3	ALL	Old ID	Search Criteria 1,034 =852+182
30	2352	less prevalent	phrase	1				1	2010	
166	2353	most prevalent	phrase	1				1	2010	
4,190	2360	risky	adjective	1	2	2	3	8		
83	2361	AsRisky	phrase	1			2	3		
183	2362	more risky	phrase							
198	2363	less risky	phrase							
28	2364	most risky	phrase	1				1	2010	
412	2370	riskier	WordUNS	1	2	2	3	8		
90	2371	riskier than	phrase	1	2	2	2	7		
3,935	2380	probable	WordUNS	1			2	3		
14,867	2390	frequently	WordUNS	1			3	4		
915	2391	more frequently	Phrase							
618	2392	most frequently	Phrase	1			2	3	2010	
90,604	2400	likely	WordUNS	**S10**	2	2	4	9	2010	
437	2401	likely as	phrase	1				1	2010	
1,190	2402	likely than	phrase	1				1	2010	
14,242	2403	more likely	phrase	**S10**				1	2010	
3,638	2404	less likely	phrase	1				1	2010	
1,821	2405	as likely	phrase	1				1	2010	
6,580	2406	most likely	phrase	1	2	2	4	9	2010	
489	2407	least likely	phrase	1			4	5	2010	
7,069	2408	mostLeast-likely	CQL							[word="most" | word="least"] [word = "likely"]
54,509	2410	likely to	phrase	**S10**	2	2	3	8	2010	
9,875	2411	more likely to	phrase	1	2	2	2	7	2010	
2,653	2412	less likely to	phrase	1				1	2010	
2,517	2413	most likely to	phrase	1	1	1	4	7	2010	
335	2414	least likely to	phrase							
12,528	2415	moreL likely to	CQL	1	2	2	3	8	2010	[word="more" | word="less"] [word = "likely"] [word= "to"]
2,852	2416	mostL likely to	CQL	1	2	2	3	8	2010	[word="most" | word="least"] [word = "likely"] [word= "to"]
3,833	2420	likely that	phrase	1			3	4	2010	
	2421	more likely that								
	2422	less likely that								
95	2423	most likely that	phrase	1			2	3	2010	
	2424	least likely that								
331	2430	proportionately	word	1		1	4	6	1003	
7,163	2440	at risk		1	1	1	3	6	1004	
	2441	more|less at risk								[word="more" | word="less"] at risk
	2442	at risk of								
	2443	CD - at risk								
	2444	CD - more at risk								
	2445	prone								
	2446	more|less prone								
	2447	CD - more prone								
	2448	apt|prone|inclined								
	2449	more|less|apt|prone|inclined								,[tag="CD"] []1,3 [word="more" | word="less"] [word="prone" | word="apt" | word="inclined"] [word="to"] within <s>
51,802	2500	common	WordUNS	**S10**			2	3	2010	
310	2501	common as	Phrase	1			3	4	2010	
397	2502	common than	Phrase	1			2	3	2010	
2,046	2503	more common	Phrase	1			2	3	2010	
350	2504	less common	Phrase	1			2	3	2010	

Figure 17: Searches, Downloads and Summaries of Statistical Phrases: Page 3

LINES	ID	SEARCH	TYPE	List	F1	F2	F3	ALL	Old ID	Search Criteria 1,034 =852+182
459	2505	as common	Phrase	1			2	3	2010	
3,973	2506	most common	Phrase	1			2	3	2010	
29	2507	least common	Phrase	1			2	3	2010	
7,590	2510	**frequent**	phrase	1			2	3	2010	
47	2511	frequent as	phrase	1			2	3	2010	
43	2512	frequent than	phrase	1			2	3	2010	
683	2513	more frequent	phrase	1			2	3	2010	
	2514	less frequent	phrase							
	2515	as frequent	phrase							
338	2516	most frequent	phrase							
	2517	least frequent	phrase							
118,728	2520	**often**	WordUNS	**S10**			3	4	2010	
18,170	2521	often as	phrase	1			2	3	2010	
1,889	2522	often than	phrase	1			2	3	2010	
4,266	2523	more often	phrase	1			2	3	2010	
366	2524	less often	phrase	1			2	3	2010	
1,763	2525	as often	phrase	1			2	3	2010	
1,447	2526	most often	phrase	1			2	3	2010	
7	2527	least often	phrase	1			2	3	2010	
15,894	2530	top CD	CQL	1			3	4	1003	
739	2540	bottom CD	CQL	1			3	4	1003	
7,743	2550	rank	word	1			2	3	1003	
9,752	2560	ranks	word	1			2	3	1003	
	2561	ranks	verb							
6,474	2570	ranked	word	1		2	2	5	1003	
4,893	2580	ranking	word	1			2	3	1003	
4,016	2590	rankings	word	1	2	2	2	7	1003	
15,691	2600	First2ndPlace	CQL	1	2	2	3	8		[word="first"|word="second"|word="third"|word="fourth"|word="last"] "place"
258	2610	First2ndRate	CQL	1			3	4	1003	
160,712	2620	top	word	**S10**			2	3	1003	
595,842	2630	first	word	**S10**			2	3	1003	
	2690	distribution	word	1	2	2	2	7		
	2691	distribution of	phrase	1	2	2	2	7		
2,766	2800	IncreasedBy5CD		1			2	3	1003	[word="increased"] [word="by"] []{0,5} [tag="CD"] within <s>
209	2810	DecreasedBy5CD		1			2	3	1003	[word="decreased"] [word="by"] []{0,5} [tag="CD"] within <s>
9,448	2820	CD5Increase		1			2	3	1003	[tag="CD"] []{0,5} [word="increase"] within <s>
515	2830	CD5Decrease		1			2	3	1003	[tag="CD"] []{0,5} [word="decrease"] within <s>
61	2840	CD-Fold		1			2	3	1003	[tag="CD"] [word="fold"] within <s>
904,905	2850	CD-Of	**CQL**	**S10**			3	4	2010	[tag="CD"] []{0,3} "of" within <s>
117,642	2900	times	WordUNS							
3,280	2901	times as	Phrase	1			2	3	2010	
	2902	times as1 as	CQL							
2,464	2903	times more	Phrase	1			2	3	2010	
403	2904	times more than	Phrase							
135	2905	times less	Phrase	1			2	3	2010	
50	2906	times less than	Phrase							
	2907	more times	Phrase							
	2908	less times	Phrase							
1,303	2909	more than CD times	CQL	1			3	4	2010	[word="more"][word="than"][tag="CD"][word="times"]
1,005,582	2910	more|less	CQL							[word="more" | word="less"]

Figure 18: Searches, Downloads and Summaries of Statistical Phrases: Page 4



LINES	ID	SEARCH	TYPE	List	F1	F2	F3	ALL	Old ID	Search Criteria 1,034 =852+182
497,271	2920	most|least	CQL	S10			3	4	2010	[word="most" | word="least"]
	2921	most of	Phrase							
	2922	the most	Phrase							
18,200	2930	for each|every	CQL	S10			4	5	2010	[word="for"] [word="each" | word="every"]
	2391	CDfor each|everyCD	CQL							[tag="CD"] []{0,3} "for" [word="each" | word="every"] [tag="CD"] within <s>
233	2940	As|increases	CQL	1			2	3	1003	[word="as"] []{0,1} [word="increases"]
1,312	2950	as much of	Phrase	1				1	2010	
	2951	AsMuchOf-As	CQL					0		[word="as"] [word="much"] [word="as"] []{0,3} [word="as"]
218	2960	by a factor of	phrase	1			2	3	1003	
9,112	2970	<Verb>byX%	CQL	1			1	2	1004	
										[tag="VV.?"] [word="by"] []{0,1} [word="%" | word = "percent" | word="per"] within <s>
10,360,309	3000	CD	CQL	S20			2	3	1003	[tag="CD"]
420,009	3010	In CD	CQL	S10			3	4	1003	[word="in"][tag="CD"]
275,628	3020	Of CD	CQL	S10			7	8	1003	[word="of"][tag="CD"]
211,650	3030	To CD	CQL	S10			6	7	1003	[word="to"][tag="CD"]
154,213	3040	For CD	CQL	S10		2	5	8	1003	[word="for"][tag="CD"]
115,823	3050	Than CD	CQL	S10			4	5	1003	[word="than"][tag="CD"]
105,860	3060	With CD	CQL	S10			5	6	1003	[word="with"][tag="CD"]
	3070	At CD	CQL						1003	
	3080	From CD	CQL						1003	
	3090	By CD	CQL						1003	

Figure 19: Searches, Downloads and Summaries of Statistical Phrases: Page 5

This data was used to generate these three articles:

- Describing and comparing rates and percentages. (Schield 2000b)
- Describing percentages presented in tables and graphs (Schield, 2001b)
- Describing percentages using Chance grammar (Schield and Burnham, 2007)

These articles were the basis of the recommended templates for describing and comparing rates and percentages in chapters 4, 5 and 6 in the textbook.

Since this process is based on judgment and since we are not trained grammarians, we have generated a call for those trained in such matters to use our data, to conduct their own analysis and see what they find. (Schield, 2020)

Unfortunately my colleague, Thomas Burnham, has passed away. (Burnham, 2018)

For years I wanted to connect with grammar experts on our process and our provisional rules with no success. I proposed a math-grammar paper for the 2024 NCTM-NCTE conference, but it was not accepted. In looking for other venues I heard about the Assembly for the Teaching of English Grammar (ATEG). I proposed a paper for their 2024 conference but was rejected. However, the editor of their journal asked me to write an article. I agreed. With the help of the editor, the final paper was much more accessible to school teachers than my initial draft. (Schield, 2024)

Appendix C: English-Corpora

Accessing this corpus is easy. And it is free for limited usage. See English Corpora.

The first step is select a corpus. Figure 20 shows some of the choices.

Corpus	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)		20.0 **billion+**	20 countries	2010-**yesterday**	Web: News
iWeb: The Intelligent Web-based Corpus		14 **billion**	6 countries	2017	Web
Global Web-Based English (GloWbE)		1.9 **billion**	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 **billion**	(Various)	2014	Wikipedia
Coronavirus Corpus		1.5 **billion**	20 countries	2020-2023	Web: News
Corpus of Contemporary American English (COCA)		1.0 **billion**	American	1990-2019	Balanced
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows

Figure 20: English-Corpora: Choice of Corpus

While the News on the Web (NOW) corpus is the largest, the Corpus of Contemporary American English (COCA) was most commonly used in this analysis since it spanned a wider time interval.

The next step is to access the search. To select on Collocates, select the plus (+) sign.

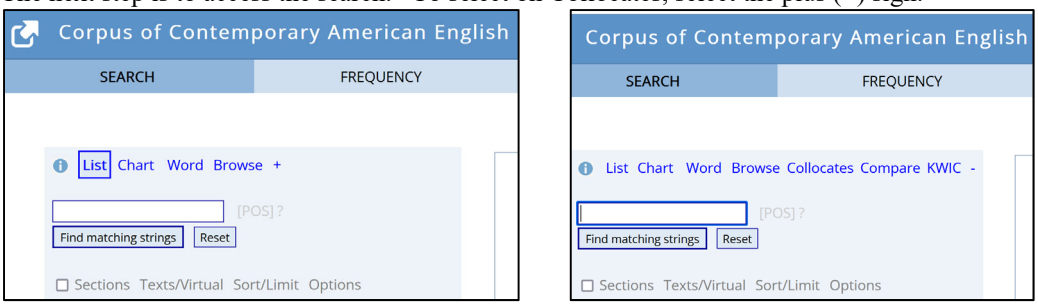


Figure 21: English-Corpora: Search Box and Collocates

Select the Collocates hyperlink to get these options. The numbers indicate the position of words with respect to the node. Enter the search word or phrase in the word/phrase text box. To get the first collocate to the left of the node, select the "1" left of the "0".

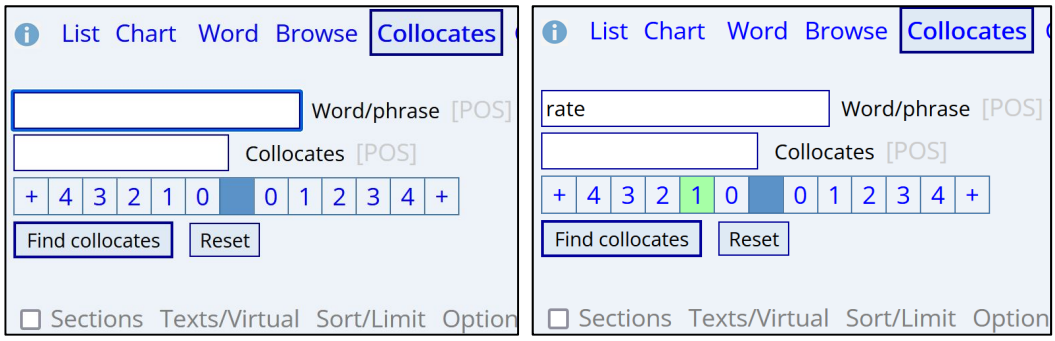


Figure 22: English-Corpora: Search Box, Collocates and Choices

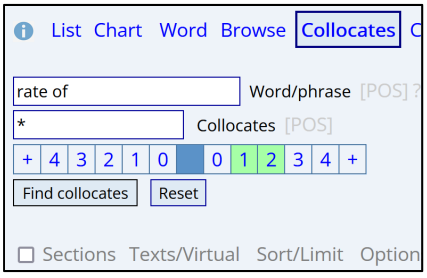
To search, press the "Find Collocates" command button.
Figure 23 shows the search results for L1 (on the left) and R1 (on the right):

RE-USE WORDS	FREQ	ALL
UNEMPLOYMENT	4909	26258
HEART	4256	193188
INTEREST	4181	147609
TAX	4168	153160
GROWTH	3250	109173

RE-USE WORDS	FREQ	ALL
OF	26631	23605687
THAN	1529	1476138
AMONG	804	267841
INCREASES	671	35717
INCREASE	454	114929

Figure 23: Collocate Output for Rate: L1 on the Left; R1 on the Right

Notice that the preposition 'of' in R1 is much more common than the subsequent R1 words. Now enter 'rate of' as the node to see what is most common in R2.

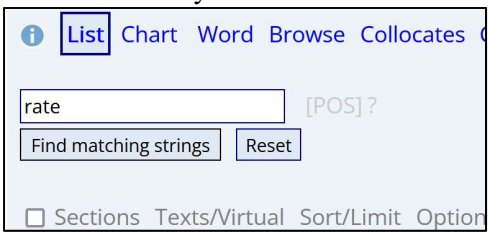


	FREQ	ALL
RETURN	860	132702
GROWTH	773	109173
INFLATION	549	22219
CHANGE	493	327055
INCREASE	452	114929
INTEREST	278	147609
SPEED	184	65302
DECLINE	172	32033
UNEMPLOYMENT	126	26258
APPROXIMATELY	99	32199
EXPANSION	86	25245
PROGRESS	80	59701
INFECTION	74	22084

Figure 24: Collocate Output for Rate of: Search on the Left; R2 on the Right

There are a lot more choices, but this is a start.
FREQ: This is the number of times this word or phrase appeared within this search.
ALL: This is the number of times this word or phrase appeared in the Corpus.
Sorting is done by FREQ – not by ALL.

To see how many matches there are for a given node or search, select LIST.



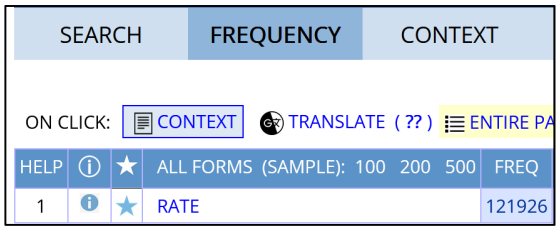


Figure 25: Collocate Output for Rate: List command on the Left; List Match Count on the Right

Of the 121,926 matches for 'rate', the 4,909 matches for 'unemployment rate' are a small fraction (4%), but the 26,631 matches for 'rate of' are a larger fraction (22%).

Appendix D: Google Ngrams

Google Ngrams is a free service based on Google's corpus of written English. Texts are dated so that time-based comparisons are available. This easily accessed corpus allows one to quickly identify which phrase of several similar phrases was the most common. It seemed advisable to use the most common phrase whenever possible.

Here are some examples that were relevant to this grammatical analysis:

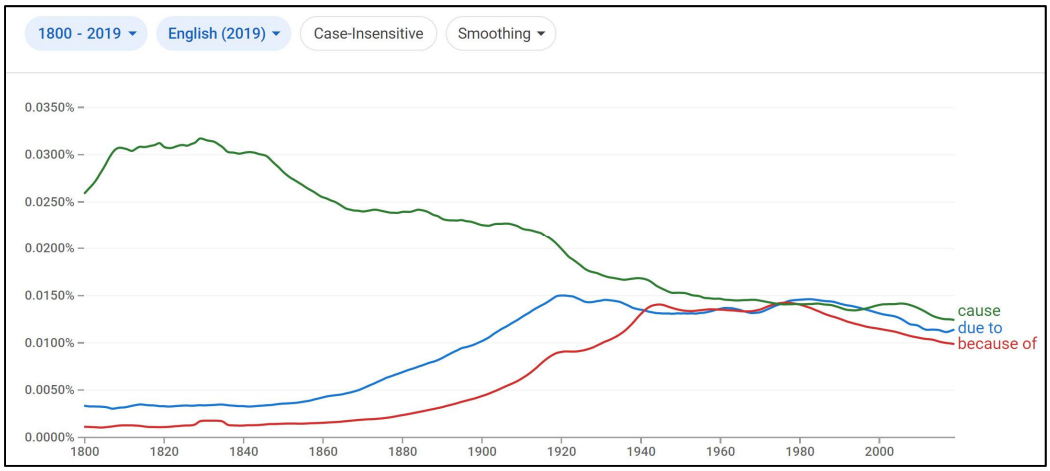


Figure 26: Prevalence of Cause, Due To and Because

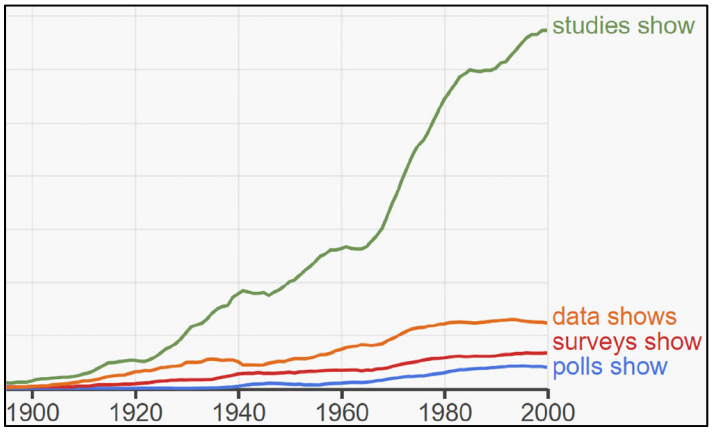


Figure 27: Prevalence of "Studies Show" vs "Data Shows"

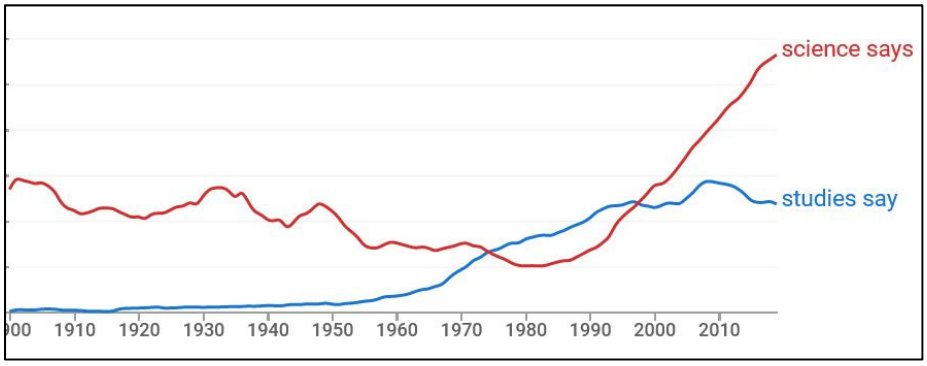


Figure 28: Prevalence of "Science Says" vs. "Studies Say"

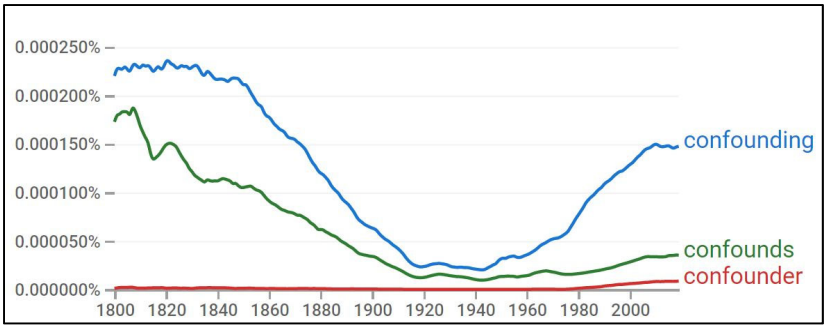


Figure 29: Prevalence of Confounding, Confounds and Confounder

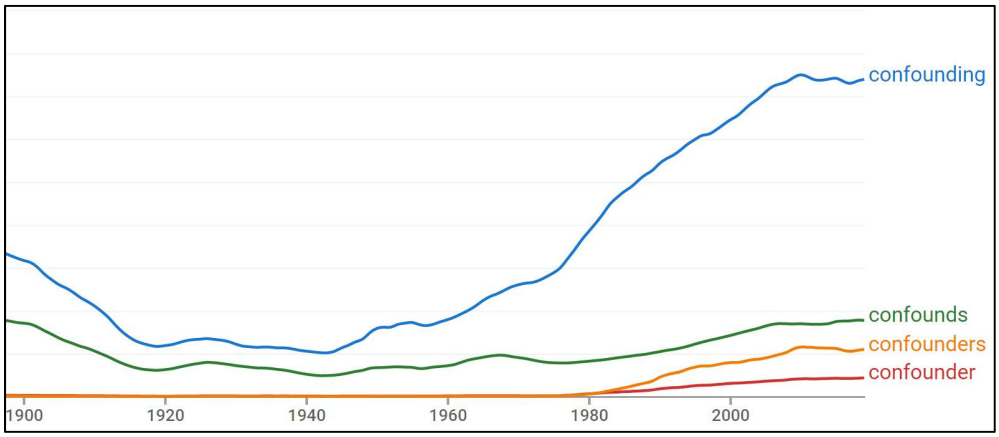


Figure 30: Prevalence of Confounding, Confounds, Confounders and Confounder

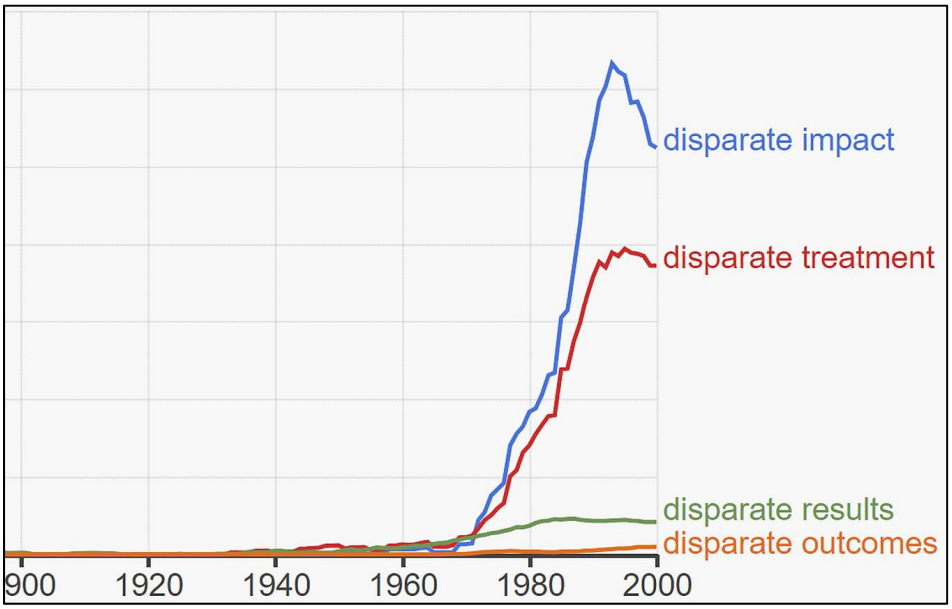


Figure 31: Prevalence of "Disparate Impact" vs "Disparate Treatment"

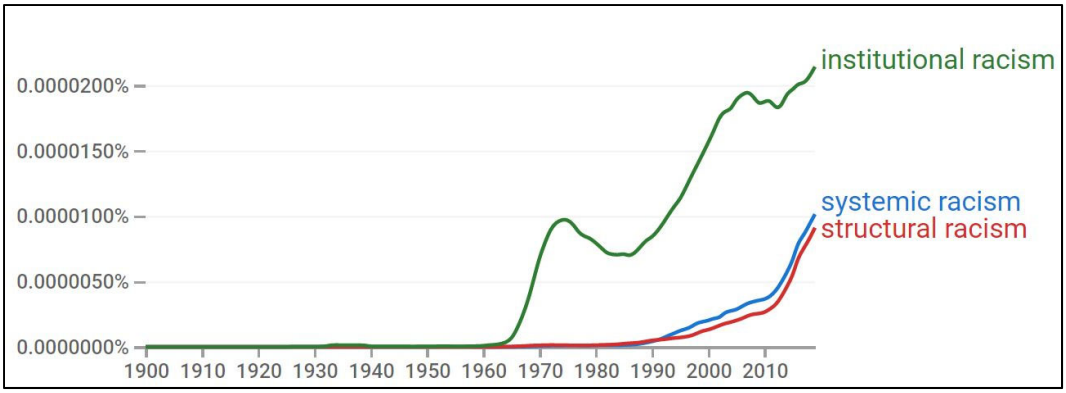


Figure 32: Prevalence of Institutional, Systemic and Structural Racism

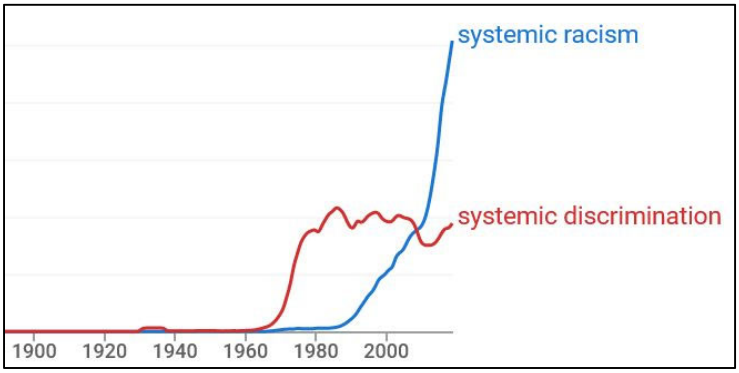


Figure 33: Prevalence of "Systemic Racism" vs "Systemic Discrimination

Appendix E: Harvard Business Review

The titles and abstracts of the 42,000+ articles in the Harvard Business Review <www.HRBR.org> were searched for various phrases between 12/2014 and 9/2015.

16,081* COMPARE & CHANGE

COMPARE	10,943*
8,864	more
1,940	"more than"
1,479	less
354	"less than"
36	"times as"
23	"times more"
0	"times less"
485	"at least"
18	"at most"
18	unequal
17	"no more than"
3	"no less than"

CHANGE	5,138*
3,956	increase (es, ed)
29	increase (es, ed) by
13	increase (es, ed) as
279	decrease (es, ed)
4	decrease (es, ed) by (as)
569	cut
334	"the more"

3,267* ----------NAMED RATIO GRAMMARS----------

PER/PERCENT	989*
714	per
275	percent
173	"per cent"
152	"percent of"
75	"per cent of"

PERCENTAGE	189*
189	percentage
119	"percentage of"
34	"the percentage"
8	"a percentage"
11	"percentage points"
8	"percentage point"
1	"percentage that"
0	"percentage who"

RATES	1,578*
1,521	rate
253	"rate of"
29	prevalence
28	incidence

CHANCE	511*
511	chance
150	"chance to"
64	"chance of"
5	"chance that"
2	"chance in"
2	"by chance"
1	"if by chance"
0	"due to chance"

LIKELY GRAMMAR	938*
938	likely
187	"more likely"
39	"less likely"
8	"more prevalent"
2	"less prevalent"
16	"as likely"
545	"likely to"
3	"likely as"
19	"likely than"

CONTROL/CONFOUND	580
234	"control of"
137	standardize (ed)
64	"take into account"
49	"taking into account"
22	"taken into account"
14	"control for"
17	"adjust(ing, ed) for"
4	"controlling for"
1	"controlled for"
1	"took into account"
8	"common cause"
3	"sampling bias"
15	confound (ed)
9	confounding
1	confounder
1	"another explanation"
0	"alternate explanation"
0	"lurking variable"
0	"effect size:
0	z-score

STUDY DESIGN	435
400	experiment
22	"clinical trial"
7	"longitudinal study"
3	"randomly assigned"
0	"random assignment"
0	"observational study"
1	"controlled study"
0	"clinical study"
2	"clinically proven"

# SAMPLE	240*
145	sample
88	random (ly)
7	randomness
2	"random sample"
0	"stratified sample"
0	"cluster sample"
0	"systematic sample"
0	"convenience sample"
0	"sample statistic"

INFERENTIAL	33
7	ANOVA
7	"statistical significance"
4	"statistically significant"
4	t-test
3	"standard error"
2	chi-squared
2	"hypothesis test"
1	"sampling error"
1	"margin of error"
1	"statistical power"
1	p-value
0	"not statistically significant"
0	"statistically insignificant"
0	"sampling distribution"
0	"null hypothesis"
0	"alternate hypothesis"
0	"research hypothesis"
0	"reject the null"
0	"rejection region"
0	"confidence level (interval)"

Figure 34: Frequency of Phrases in Harvard Business Review Abstracts

There were just 22 instances of clinical trial, 11 instances of statistical significance or statistically significant, and one instance of p-value. Whereas there were 135 instances of "[take|taking|taken] into account" and 25 instances of confound [ed|ing]. However this disproportionate result may suffer from selection bias. Statistical inference may be more common in the body of the articles than in the title or abstract: the basis for all of this data.