Describing Quantitative Relationships Using Informal Grammar

Milo Schield, W. M. Keck Statistical Literacy Project. Minneapolis, MN.

Abstract: This paper investigates the informal grammar used to describe quantitative relationships. It excludes the use of formal-grammar: ratio-nouns such as *percent, percentage, rate, chance, odds, risk, likelihood* and *probability*. The approach is empirical – examining the words and phrases found in ordinary usage. This paper investigates implicit relationships involving the quantitative functions of words (e.g., *of, to, per, out of, of every, of each* and *out of every*) when placed between two numbers. It investigates explicit relationships involving (1) the quantitative functions of tendency adjectives (*likely, apt, liable* and *prone*), verbs (*tend* and *inclined*) and nouns (*inclination, tendency* and *propensity*); (2) the difference between non-temporal and temporal comparisons and (3) the properties of coordinated comparisons (*the more x, the more y*).
Keywords Statistical literacy, syntax, semantics.

INTRODUCTION AND APPROACH

If statistical literacy is to involve literacy, it must address the use of ordinary English to express quantitative relationships – both implicit and explicit. Any relationship between two numbers is implicitly the basis for a relationship, whereas an explicit relationship needs only one number.

The data is obtained from the Harper-Collins' WordBanks corpus. The Appendices have the details. The data is based on 3,213 downloaded files based on 505 unique selections. The numbers in braces [1234] are the ID numbers of specific downloaded files.

The associated statistics reflect the underlying data (which may not be representative of anything) and the computer-generated tags (which to our knowledge have not been evaluated for accuracy). In the worst case, these statistics may be so inaccurate that they actually obscure the syntactic relationships. Since we have no better source and we have no evidence that these statistics are inaccurate, they are presented when they helped us identify the underlying semantic relationships – and in case they may help others identify relationships that we have missed.

One approach is to take ratio-keywords involving quantities (formal grammar) and see what patterns emerge in the associated sentences. See Schield (1999) and Schield and Burnham (2000).

This paper take a more empirical or investigative approach. The first task is to investigate the words involved in an implicit quantitative relationship. We narrow our focus to short phrases involving two numbers separated by a single word (Section 1), by two words (Section 2) and by three words (Section 3). The goal is to identify any quantitative functions of these words.

1. ONE WORD BETWEEN TWO NUMBERS

WordBanks has 1.2 million phrases [2620] containing one non-numeric word with two adjacent numbers. This generally requires that both numbers have the same units (e.g., page 1 of 9), but not always (12 June 1940). In this three-word node, here is the distribution of tags for the center word: NP 31%, : 19%, , 15%, CC 8%, JJ 7%, IN 6%, TO 5%, NN 3%, (2%,) 2%, SYM 0.7%, NNS 0.5% and other tags (0.7%). Table 2 has a list of the WordBanks part-of-speech tags. After omitting coordinating conjunctions and selected non-words such as commas, parentheses and symbols (3 # 4 oz, 3 $ 20), this is the distribution of center tags: NP 43%, : 27%, JJ 9%, IN 9%, TO 7%, NN 4%, and other tags (2%). The following analyzes the matching phrases by tag.

1.1 # NP

In the 291,043 matching phrases [2621] with a proper noun singular (tag=NP) between two numbers, 99% of the center words are months: 12 June 1940. Of the rest, time of day (am, pm) is most common. It seems that none of these cases involve an arithmetic relationship.

1.2 # IN

There are 84,451 sentences with two numbers separated by words tagged as IN. [2625] The most common of these tagged words are *of* (18.328), *in* (16,660), *for* (5,249), *by* (3,700), *with* (1,705),

over (1,659), *from* (1,311), *at* (897), *through* (785), *after* (543), *on* (496) and *per* (368). In the following paragraphs, the search was on words, not on tags, so the counts may differ from these.

1.2A # of #

There are 19,682 phrases with two numbers separated by *of* [2631]. **(1) Literal part-whole ratio (*of* introduces the whole)**: "in one of two ways", "page 1 of 3", "was one of four children," "lost two of three games" and "massacre of all but 15,000 of 80,000 Herero people". (2) **Figurative part-whole ratio (*of* introduces the whole)**: "Almost one of four car sales in the USA" "nine of 10 white mothers living in cities obtained prenatal care in the first trimester," "Nearly nine of 10 graduating seniors rate their education," "Islam, a religion whose adherents account for one of five human beings on the planet," and "About one of four loans now are zero-percent deals" and "Panoz is one of [the] 2% of smokers who are allergic." (3) **Figurative part-whole ratio (*of* introduces the part)**: "average death rate per 1,000 of 32.43," (4) **Other arithmetic relations**: "we have about a .25 [quarter] of 1% rate of positives," "It was ten of [before] seven," (5) **Non arithmetic relations involving a date**: "the sale in 1995 of 4.75 acres," "convicted May 26 of 161 counts of first-degree murder" and "the Corsa 3 [car] of 1995," "don't expect a duplication in 2003 of 2002" and "his Top Ten [songs] of 2004" (6) **Other non-arithmetic relationships**: "[board] lengths, two of thirteen feet," "one [Megalodon tooth] of 5 inches will cost" and "Twelve of Six Flags' 13 parks". Analysis: Signs of a figurative relationship are L1 "approximate" words such as *about* and *nearly*, and phrases such as *more than* or *less than.*]

1.2B # in #

There are 21,118 phrases with two numbers separated by *in* [2630]. In 47% of the phrases, the second number is between 1901 and 2011 (a date); in 21% the first number is one; in 26% the first number is "million" or "billion". In 17% the first number is "million or "billion" and the second number is between 1901 and 2011. The R1 words include the comma (18%), *to* (10%), *and* (6.3%) and the period (4.2%). Of the 34% of R1 tags that are nouns (N*), 78% are common plural (NNS). (1) **Literal part-whole ratio (*in* introduces the whole)**: "St. Louis two [goals] in 19 [games]," "bagging five [goals] in five [tries]" "We've lost one [game] in 12," (2) **Figurative part-whole ratio (*in* introduces the whole)**: "about one in three people," "nearly one in four crimes" "only one in 10 (12 %) [of those Britons surveyed]" "Fewer than two in five (37 per cent) [Britons]," "Almost three in four (70%) [people surveyed]," "Nearly three in four (71 %) employers" "risks of failure (roughly 3 in 10)" and "one in seven Americans". (3) **Non-arithmetic relationships involving a date**: "only 25 in 1997" or "flight 103 in 1988." (4) **Other non-arithmetic relationship**: "was 69-50 in four seasons" and "record for strikeouts, 263 in 178 innings".

1.2C # for #

There are 7,561 phrases with two numbers separated by *for* [2632]. (1) **Out of**: Common phrases include "finished 2 for 3", "went 2 for 3 yesterday", "shot 6 for 25" A whole sentence: "Burrell is 12 for 40 (.300) with five homers…" In these cases the number preceding *for* is always smaller than the number following and *for* indicates *out of*. (2) **Exchange**: A different use: "split the shares two for one" or "a 3 for 2 offer." In these cases, the second number can be smaller than the first and *for* indicates an exchange. (3) **Result**: Example: "hitting 35 of 54 for 64.8%." In this case, *for* indicates a result. (4) **Related duration**: "no. 1 for two weeks", "October 7 for six performances" or "$1 million for one year".

1.2D # by #

There are 4,149 phrases with two numbers separated using *by*. [2633] (1) **Arrangement**: Common examples: "eliminated one by one" meaning "one after another" in sequence. *By* can indicate together or side-by-side: "The animals went in two by two". (2) **Rectangle**: *By* can also be used to distinguish height from width: "each 800 by 1024 pixels," "The board was a 2 by 4 (2" by 4")" It may also distinguish width (distance parallel to the street) from depth (distance perpendicular to the street): "a standard city lot is 50' by 90' or 75' by 150'. (3) **Multiplied by**: Note that "x" is often used as a symbol for *by*. E.g., "Multiply three by 2" = 3 x 2.

1.2E # with #

There are 4,797 phrases with two numbers separated by *with*. [2634] (1) **Along with**: Sports examples are very common: "hit .321 with 10 homers" or "led 61-56 with 25 seconds". Non-sports

examples include "voted 139-1 with one abstention" and "ending 2002 with two million". Conclusion: In each of these cases, *with* introduces something that happened, existed or was true at the same moment in time. *With* doesn't seem to indicate a quantitative relationship in these cases.

1.2F # over #

There are 1,865 phrases with two numbers separated by *over*. [2635] (1) **Duration (during)**. In about 80% of the matches, the word following the number after *over* involved a unit of time such as *years*, *months*, *days*, *weeks* or *decades*. Examples: "spend $100 million over three years". In such cases, *over* is short for *spread over (during) the following time interval*. In another 5% of the matches, the word following the number after *over* is time-related such as *seasons*, *innings*, *games* or *quarters*. For example, "WorldCom had overstated a key measure of earnings by more than $3.8 billion over five quarters" or "Molitor batted .306 over 21 seasons." (2) **More than**: "At one minute past eight over 300,000 visitors hit the website," "In 1998 over 12,000 people were exposed" "Six people: three [with ages] over seventy" or "Woosie shot a three over 74." Here *over* means *more than.* (3) **Compared to**: "revenues climbed by $700 million in 2002 over 2001." (4) **Ratio**: "Your blood pressure is 195 over 100" where the first number is the diastolic (the maximum) and the second number is the systolic (the minimum) for the same person at the same time. This can be written using a slash: "195/100."

1.2G # from #

WordBanks has 3,069 phrases with two numbers separated just by *from*. [2636] (1) **Reference value in a change: To-from**: "the number declined to 81 from 85," "raise the legal dropout age to 18 from 16," "The Argos will trim their roster today to 40 from 50," "September index fell to 48.1 from 54.9 in August," "percentage of women holding so-called clout titles from executive vice president up to CEO increased to 7.9% in 2002 from 1.9% in 1995" and "expand capacity to 14,000 from 12,300." **By-From-To**: "Deaths from other causes decreased [by 22] from 39 to 17" (2) **Range of times: From start to finish** "The Vietnam War cost roughly $500 billion from 1964 to 1972" "was president of Local 234 from 1989 to 1996" and "Babe Ruth's nine-year stretch of 421 [home runs] from 1920-28" (3) **Out of**: "Jockey Ben Russell won four times yesterday, giving him 50 wins from 252 starts," (4) **Originating in or taken from**: "Penfold Grange vintage [wine] offerings include six from 1971," "I've cooked up four meals for two [people] from one medium-sized chicken," "[the lawyer] stole $2.76 million from 39 clients" and "Subtract [take] 5 from 23." (5) **Distance**: "made a bogey five from [a distance of] 20 feet." (6) **Mixture of the preceding**: "leap in homicides to 41 in 2003 from 27 in 2002."

1.2H # at #

WordBanks has 2,192 phrases with two numbers separated just by *at*. [2637] (1) **Related time**. Of the words preceding the leftmost number, 27% are the names of months such as "Posted online: Monday, November 01, 2004 at 0248 hours" where the number preceding at is the *year* and the number following is the time. 14% (2) **Related amount**: "The S&P 500 was up 5.82 at [the final value of] 864.37" or "borrow $100 million at [an interest rate of] 10%." Conclusion: In none of these cases does *at* indicate a quantitative relationship between the two numbers.

1.2I # per #

WordBanks has 875 phrases with two numbers separated just by *per*. [2638] Only 368 of these are tagged as IN. (1) **Part-whole ratio** (must have same units with the first number smaller than the second): "7 per 1,000 live births," "6 per 100,000 population," "CCA said its escape rate was less than one per 10,000 inmates" (2) **Incidence rates** (different units): "six [pilot accidents] per 10 million flights." (3) **Exchange ratios** (typically different units): "Gold prices hit a five-year low of Rs [Indian rupees] 4,335 per 10 gm."

1.3 # TO #:

Among the 61,793 matching phrases with two numbers separated by the word *to* [2626], the 17 most common L1 words are *from* (16283), *of* (3246), *in* (3112), *for* (1913), "comma" (1909), *aged* (1724), *about* (1579), "left parenthesis" (1417), *by* (1266), *next* (983), *ages* (768), *the* (750), a (727), "colon" (676), *at* (594), *and* (553), *within* (504). Three patterns are observed. (1) **Range**: Common examples: "ranging from one to five," "scale of 1 to 10," "in three to five years," "for six to eight weeks," "By three to four weeks after vaccination," "next [last, past] two to three

years", and "every three to five years." These phrases indicate a range of values (from minimum to maximum). (2) **Change/Difference.** Examples: "reduced from three to two". This phrase indicates a change in values (from old to new). Now consider this phrase: "increased from three to four". This phrase may indicate a change in values (from old to new) or it may indicate the resulting range of values (from minimum to maximum). In such cases, the change-value form may be confused with the range-value form unless additional context is provided. E.g., "rates increased from three to four percent in six months." (3) **Ratio comparison**: The number preceding *to* is generally bigger than the number following. "odds of 10 to 1 [against winning]" "His army outnumbered Charles's by two to one." "The majority of voters – by two to one – wanted …" "The score was 3 to 2," "savers outnumber borrowers by seven to one." The score proceeding *to* is typically that associated with the subject or speaker of the statement.

Note that *from* (either explicit or implicit) is necessary for a range or change/difference but is inappropriate for ratio comparisons so cases preceded by *from* should be analyzed separately.

1.3B # - TO - #:

WordBanks has 36,103 phrases where two numbers are separated by "- to -". [2007] (1) **Ratio** (certainly if the first number is at least as big as the second). Examples: "the gender ratio of people who read print newspapers is about 1-to-1," "risers outnumbering decliners by 10-to-seven," "the 10-to-four majority," "drug DUI arrests outnumber those involving alcohol more than 2-to-1," "at odds of 21-to-1" and "a 3-to-1 edge in delegates". (2) **Range** (possible if the second number is larger than the first). Examples: "18-to-34 age group," "16-to-24-year-olds" and "Spanish flu killed an estimated 20-to-50 million people."

1.4 # NN #

Few of the 37,649 matching phrases having two numbers separated by a word tagged as a common noun (tag-NN) make any sense. [2628] Of these nouns, 16% involve month abbreviations without a period, 14% involve the letter *r* [Rupee], and 10% involve the percent sign (%).

1.5 # JJ #

Of the 25,411 matching phrases [2624] with a word tagged as an adjective between two numbers, 97% of these "words" are numbers. No arithmetic relationships. This tagging of numbers as adjectives often occurs when the number is followed by a quantitative noun such as *million.*

1.6 # : #

Of the 11,192 matching phrases [2622] with a colon between two numbers, 97% are followed by an indicator of time: am or pm. This indicates these colons are commonly used to distinguish hours from minutes with hours before the colon and minutes after. While there is a 60 to 1 relationship between hours and minutes, this is fixed by the units. Of those that don't distinguish hours and minutes, 84 are preceded by "ratio of."

1.7 # - #

WordBanks has 6.5 million matches involving a minus sign [2002]. Of these, 24% involve just a single minus sign, 9% involve two minus signs and 1.7% involve three minus signs. The rest of the matching node "words" involve a minus sign with additional characters. The dash and hyphen are both represented in 7 bit code as minus signs. Generally, if the minus sign separates two numbers (e.g, 0-2 or 2004-05) or two parts of a date (e.g., 1-Jan), it is considered a dash and may be tagged as a "number" (CD). But if it separates two words that don't involve dates, it is considered a hyphen and the entire unit may tagged as a proper noun (NP), common noun (NN) or an adjective (JJ).

Examples of common proper nouns: e-mail, al-Qaeda, Wal-mart, Coca-cola, sun-times and african-american. Examples of common nouns: half-time, semi-final. line-up, vice-president, free-kick, team-mate and make-up. Examples of common adjectives include long-term, so-called, full-time, all-round, two-year, well-known and high-profile.

Trying to identify when the minus sign or dash indicates a ratio is best done by looking at phrases involving the keyword *ratio.* That is outside the scope of this paper. In looking at the words in the L1 position, these include the numbers 2 (10,788), one (7,907), 2002 (4,006), 6 (2,854), 10 (2,361) and 3 (2,336). In looking at the words in the R1 position, these include the

numbers 1 (17,454), 11 (17,336), 12 (14,692) and 10 (14,055). (1) **Ranges.** Examples: "the Bronte sisters Charlotte (1816 - 55)," and "bake for 15 - 20 minutes."

1.8 # SYM

Of the 8,348 phrases with two numbers separated by a symbol [2629], the most common symbols are the equal sign (56%), right bracket (17%), times "x" (9%), copyright c (5%), "r" [Rupees] (4%), slash (3.3%), plus sign (2.1%) and vertical bar (1.8%). The slash between two numbers [2637] is sometimes considered a part of the first number as in "1/ 4 cup". Conclusion: Of all these symbols, only the slash and plus indicate an arithmetic relationship between two numbers.

1.8a # / # <word>

There are 9,411 matching phrases [2627] with a slash between two numbers followed by a noun or preposition (tag= .N). The most common trailing words in the node are *cup, in, innings, teaspoon, tsp, out, commission* and *of*. (1) **Ratio or fraction**: "bittersweet chocolate 1-1/2 tablespoon," and "pitched 1-2/3 innings." (2) **Figurative ratio** where the slash stands for *per*: "Rate: 1.7 / 1,000 for serious injuries." The word *rate* or *ratio* could distinguish these from the fractions (1/4 oz).

1.9 Summary

In phrases with one word between two numbers, certain prepositions (*to, of, in, for, by, over, from* and *per*) often indicate quantitative relations of more than one type so that additional syntax or a knowledge of the semantics is required to distinguish the meanings. It seems that *in* and *of* have similar quantitative functions or roles.

2. TWO WORDS BETWEEN TWO NUMBERS

We did not obtain the number of phrases containing two non-numeric words between two numbers. Of such phrases, 79,000 involve a preposition as the second word [2650, 2683], 43,000 have a preposition as the first word [2670], and 16,000 have *to* as the second word [2680].

2.1 SECOND WORD IS A PREPOSITION

WordBanks has 79,168 two-word phrases that end with a preposition and are between two numbers [2650 and 2683]. In the 27,584 such phrases analyzed, the most common center words are *out of* (4231), *goals in* (1476), *months of* (954), *hits in* (780), *times in* (521), *% in* (502), *boats with* (461), *runs on* (424), *games in* (349), *runs in* (342), *crore in* (335), *points from* (272), and *points in* (266). Note that all these examples involve a noun-preposition except for *out of* where *out* is tagged as an adverb (RB). In the 27,584 phrases analyzed, the top 12 second-word prepositions in descending order are *in* (38%), *of* (23%), *with* (6.5%), *from* (5.3%), *on* (5.2%), *for* (4.4%), *at* (2.4%), *over* (2.4%), *by* (2.1%), *after* (1.9%), *per* (1.3%) and *under* (1.0%). Of these, just the first two (*in* and *of*) are investigated along with *per* based on the results of our preceding analysis. Our goal is to see if any new forms or functions emerge that indicate quantitative relationships.

2.1A # <word> in

The 20 most common words [2684] following the first number – and preceding *in* –are *goals* (9%), *hits* (6%), *times* (5%), *runs* (3%), *games, crore, points* (2%), *%, people, wins, years, months, percent, assists, win* (1%), *day, wickets, million, run,* and *tries* (0.8%). (1) **Non part-whole ratio - geographically**. "two apartments in one building", "52 locations in 30 states," "326 adults in 12 states". (2) **Non part-whole ratio - temporal**: "two goals in three," "six hits in seven innings", "two goals in four minutes", "15 goals in 55 starts", "three birdies in four holes", "three times in 30 days", "zero exports in 1994-95," "6.8 percent in 1998"or "three murders in 1993". (3) **Part-whole ratio** (where *in* indicates the whole): "one person in three", "three times in six," "one patient in 13", "one night in three", or "four women in ten." (4) **Temporal comparisons**: "[in] 1996 than in 1995" "two more in 2006". Analysis: To indicate a part-whole relation like *of*, the two numbers must have the same units and the same subject.

2.1B # <word> of

The eight most common words [2685] following the first number – and preceding *of* – are *out* (52%), *months* (11%), *%* (2%), *percent, groups, sets, instead* and *quarters* (1%). Examples using *out of* include "four out of five people", "eight out of nine games" and "three out of four women". Examples using other first words include"53% of 2004 models", "10 percent of 1990 levels", "12 groups of [size] four [each]", "three sets of 15 reps [each]", "four instead of three" and "first two

quarters of 2004". **Two distinct uses appear:** (1) where the last number distinguishes the group but does not indicate the size of the group (as in years), and (2) where *of* is preceded by *sets* or *groups* so that *of* indicates their constituent parts.

2.1C # <word> per #

The 20 most common words [2689] following the first number – and preceding *per* – are *parts* (28%), *cases* (7%), *deaths, births* (4%), *part, problems, people, runs* (2%), *litres, crimes, women, defects, cars* (1%), *cents, claims, km, abortions, doctors, divorces,* and *milligrams* (0.6%). Examples include "379 parts per million", "55 cases per 1,000 households", "92 deaths per 100,000 workers", "107 births per 1000 women", "charged 25 cents per 100 pounds". In some cases the units are quite different: "76 complaints per 100 cars", "one adult per two children", "50 milligrams per 100 millilitres", and "0.9 walks per nine innings". **Two distinct uses appear**: (1) where *per* indicates a part-whole relationship between two closely related terms so the part can never exceed the whole (50 centenarians per 100,000 people) and (2) where *per* indicates a ratio of exchange (price per pound).

2.2 FIRST WORD IS A PREPOSITION

WordBanks has more than 43,000 two-word phrases beginning with a preposition that appear between two numbers [2670]. These prepositions include *of* 68%, *in* 7%, *from* 6%, *out* 5%, *for* 3%, *on* 3% and *with* 3%. The tags for the second word include appositive (DT 47%), proper noun (NP 15%), possessive pronoun (PP$ 11%), adverb (RB 6%), adjective (JJ 6%) or preposition (6%). Of these, just the first four (*of, in, from* and *out*) are investigated along with *per* based on the previous results. Our goal is to see if any new forms emerge that indicate quantitative relationships.

2.2A # of <word> #

WordBanks has 21,000 two-word phrases located between two numbers and beginning with *of.* [2671] The most common tags of the second word are the appositive (DT 68%), possessive pronoun (PP$ 21%), adverb (RB 7%) and adjective (JJ 1%). The most common second words are *the* 65%, *his* 9%, *only* 5%, its 4%, their 4%, her 2%, *those* 1%, *about* 1%, *every* 1% and *just* 1%. Common appositives are *the*, *only*, and *every*; common adverbs are *only* and *just*; and common adjective are *fiscal*, *several* and *early*. *Of only* is 3.8 times as common as *of every* while *of each* and *of any* are rare. *One* precedes almost all (99%) of the *of only* phrases but only half (51%) of the *of every* phrases. It appears that there is a major difference between these two phrases. Examples: "Augusta is one of only 10 metro areas to …", "have to win six of every eight undecided voters", "he will hire one of any ten first-class … lawyers."

2.2B # out <word> #

WordBanks has 4,996 two-word phrases located between two numbers and beginning with *out.* [2674] Of these *out* is tagged as a preposition (IN 37%), an adverb (RB 34%), a particle (RP 16%) and a common noun (NN 12%). *Out* is almost always (99.3%) followed by *of.* E.g., "almost three out of four (71.2%) are in work." Adding *out* to *of* emphasizes the part-whole relationship.

2.2C # in <word> #

WordBanks has 4,079 two-word phrases located between two numbers and beginning with *in.* [2672] The most common tags for the second word are DT (1,697), NP (1,066), CD (390) and RB (274). The most common words tagged DT are *every* (393), *the* (348) and *a* (311). Example: "1 in every 100 infants dies before his or her first birthday." The most common words tagged NP are the months (64%) and *game* (19%). E.g., "# in July 2001" Sometimes the word is a number: "fatal errors occur in 1 in 600,000 transfusions" with "000" is treated as a separate number.

2.2D # from <word> #

WordBanks has 1,941 two-word phrases located between two numbers and beginning with *from.* [2673] The most common tags for the second word are JJ (1,234: all numbers), NP (240: names of currencies), RB (115: *just* 42, *about* 35 and *only* 20), CD (106) and DT (94). E.g., "dwindled to fewer than 100 from about 1,000" and "down 1,000 from the 2001 total."

2.2E # per <word> #

WordBanks has 697 two-word phrases located between two numbers and beginning with *per.* **[2678]** In all these *per* is tagged as a preposition (IN). The word following *per* is a noun (NN,

89%), a number (CD 4%), a common noun (NP, 3%) an adjective (JJ 1.7%) or other. The most common words following *per* are *cent* (87%), cent-related (cent/ 4%), numbers (*1, 100*), time units (*month, annum, day, year*), other units (*cylinder, share, child* or *pound*) or *per every* (0.3%). Examples: "to buy Treasury 10 *per cent* 2003", "compared with 16 *per cent* three months ago", "2.72 [deaths] *per 100* million miles driven", and "1 [miscarriage] *per every* 10 live births".

2.3 SECOND WORD IS *EVERY* or *EACH*

Notice in the preceding sections, the word *every* has appeared as the second word in these phrases: *of every* (1%), *in every* (14%), and *for every* (8%) where the percentages are the fraction of phrases with two numbers separated by two words where the first word is the preposition indicated that have *every* as the second word. There are 847 instances of two-word phrases located between two numbers that involve a preposition followed by *every* or *each* [2640]. The most common prepositions are *in* 51%, *of* 31%, and *for* 16%. Of these two word phrases, 98% (835) involve *every*; only 2% (12) involve *each.* Note: second words *any* or *only* were not analyzed.

2.3A # in every

WordBanks has 429 instances of *in every* between two numbers. [2641] Examples: "Down's Syndrome affects one in very 1000 births", "[twins] about 1 in every 89 births", "[identical twins] 1 in every 60,000 cases", and "Infertility affects one in every four couples to some degree in the UK".

2.3B # of every

WordBanks has 266 instances of *of every* between two numbers. [2642] Examples: "four of every five Americans approved", "one of every three babies born has an unwed mother" and "Losing weight could prevent one of every six cancer deaths in the US".

2.3C # for every

WordBanks has 133 instances of *of every* between two numbers. [2643]. Examples: "[US] 104.8 boys born in 2000 for every 100 girls", "Number of mothers dying of pregnancy-related complications 1915: 9 for every 1,000 live births"

Analysis: The differences between *in every, of very* and *for every* seem subtle. When the units of the two numbers differ, *for every* is the more common choice.

2.3D # <word> each

WordBanks has only 12 instances of involving *each* as the second word between two numbers. [2640]. The first words are *for* (6), *on* (2) and one each of *from, in , of* and *with*. Five of the examples involving *for each one* are literal: "the inscription '1972-1973' on each one," "Tickets … with $2 from each one purchased donated," "Russian matrioshka dolls: there's five or seven in each one," "Kelly gives marks out of ten for each one [person]," "it's going to cost between 15 and 20 for each one [refrigerator disposal]," Three of the examples involving *for each* with other value are literally true: "A trillion dollars is $1000 billion with each billion $1000 million," "She had a 32 [golf score] on each nine [holes]" and "one[home allowed] for each five acres." Four of the examples involving *each* are to be taken figuratively or on-average: "Australia also won more medals per capita – one for each 408,000 people," "taxpayers coughed up Pounds 2,138 [on average] for each one [library loan]," "Thus, 75 of each 100 persons," and "death rate … is 1.73 for each 1000 lives." Most of these figurative or on-average usages involving *each* sound a bit odd.

Analysis: When *each* is the second word between two numbers, it is generally meant to be taken literally. When *every* is the second word, it is generally meant to be taken figuratively or to indicate what is expected on average.

Note that the units of both numbers are often the same when using *every* while they are generally different when using *each.* It seems that *of every* is more likely to indicate a part-whole ratio that cannot exceed unity while *each* can indicate a rate or ratio.

2.4 # out of

WordBanks has 4,959 phrases with two numbers separated by *out of*. [2651] Note that the *out* in *out of* is sometimes tagged as a preposition (IN) and sometimes as an adverb (RP). The tags of the first word following such phrases are NNS (1,942), IN (432), JJ (389), NN (370), NP (330), DT (214), SENT (144), VVD (106) and VVP (95). The most common nouns (N..) are *people*, *women*,

games, cases, patients, times, Americans, children, men, seats, voters and *families*. The most common prepositions (IN) are *of, for* and *in.* The most common verbs are *said, is, had* and *believe*. In at least 95% of the phrases the first number was less than or equal to than the second. E.g., "two out of three", "four out of five" and "three out of four." In those phrases where the first number was bigger than the second, there were two distinct causes: (1) the use is facetious: "I don't care if these refs get nought out of 10 or 11 out of 10. They must be allowed to think for themselves" or "Bo Derek, who became an international star as the 11 out of 10 woman in the film 10." (2) Word Banks tagged the thousands as a separate number: "Israel's promise to free 900 out of 8 000 Palestinian prisoners" or "South Africans comprised 5 849 out of 111 590 settler arrivals"

It seems that adding *out* to *of* is used to emphasize that the first quantity is *part of* the second. A third-grade teacher in Fairfax Station Virginia said this usage (*m out of n*) was the simplest and easiest usage for her students to understand part-whole fractions.

2.5 SECOND WORD IS *TO*

WordBanks has 15,938 two-word phrases ending with *to* that are between two numbers [2680]. In each case, the word *to* is properly tagged as *TO*. As noted previously in section 1.3, this phrasing can indicate a range (from min to max), a change or difference (from start to finish) or a ratio.

2.5A WordBanks has 3,277 two-word phrases that end with *to*, and are between two numbers where the first number is preceded by *from* [2681]. Of these, 58% have a second number that is larger than the first.

2.5B WordBanks has 11,396 two-word phrases that end with *to* and are between two numbers where the first number is NOT preceded by *from* [2682].

- Of these, 62% have a second number that is larger than the first. In many of these there is an implicit *from.* Examples: "save 15 % to 50 %", "increases of 7 % to 8 %", "rate of 15 percent to 20 percent", "children ages 6 months to 23 months", or "an estimated 20 % to 30 %". Another use is when the first number and second have different units: "in the 15 years to 2003", "increased by four million barrels to 283.4 million".

- Of these 38% have a second number that is smaller than the first. Some of these numbers involve a difference in the type of units: "within 18 months to two years", "at ten minutes to four", "fell by 9 percent to 3.7 million" or "punishable by 30 days to six months". Some numbers involve the same type of units, but have a difference in scale: "rose by 800,000 barrels to 123.8 million". Of those numbers involving the same units and same scale, some involve a range, difference, comparison or change: "turnout of 117.5 million to 121 million", "terms of five years to 10 years", "fiscal 2003 compared to 2002", "were outgained 369 yards to 331", "plunged 90 percent to 6". Finally there are some numbers involving the same units and scale that may involve a ratio: "[win] by two sets to one", "scoring three tries to one" and "a three games to one edge".

  In an analysis of 300 such matches, about half (154) had explicitly identical units while a sixth (54) had similar units based on the context.

2.6 SUMMARY

In phrases with two words between two numbers, certain two-word phrases (*out of* and *in|of|for every|each*) clearly indicate quantitative relationships – generally part-whole fractions. A third-grade teacher in Fairfax Station Virginia said the phrase "*m out of n"* was the simplest and easiest way for her students to understand part-whole fractions.

3. THREE-WORDS BETWEEN TWO NUMBERS

Consider three-word phrases located between two numbers. Based on previous findings, the focus is on *to* and on new forms involving previously selected prepositions.

3.1 THIRD WORD IS *TO*

WordBanks has 100,093 three word phrases that are between two numbers and that end with *to* [2691]. In scanning the node, no new quantitative relationships were noted. Examples: "1. Preheat the oven to 450", "one and a half to two", and "six children ages 3 to 11".

3.2 THIRD WORD IS *EVERY* OR *EACH*

Now consider **three-word phrases** located between two numbers that involve a word and a preposition preceding *every* or *each* [2645]. Of these, 95% involved *every* while 5% involved *each.* The most common prepositions are *of* (52%), *for* (34%) and *in* (9%).

Of those phrases involving *every*, the most common preceding prepositions are *of* (54%), *for* (33%) and *in* (10%). Examples: "one out of every four people," "one computer for every five students" and "one death in every three". Of the phrases with *of every*, 95% involved *out of every*.

Of those phrases involving *each*, the most common preceding prepositions are *for* (50%), *on* (13%) and *after* (7%). Examples involving *each*: "two weeks after each one", "four birdies on each nine", "whizzing for 10 seconds after each one is added" and "[water] pressure increases 1 atmosphere with each 10 m (33 ft.) of depth".

Analysis: As mentioned earlier in section 2.3D, this use of *every* is generally meant figuratively – a rhetorical emphasis for any subgroup while this use of *each* is generally meant literally. . But some examples involving *each* are probably meant figuratively: "one adult for each four 2-year-olds" and "rose a further 1.30 percent for each 1 percent change in the market", "15 injuries for each 1000 workers" and "20 cats for each 100 people in Tasmania."

4. OTHER SYNTAX INDICATING TENDENCY

While nouns like *percentage, rate* and *chance* generally indicate part-whole ratios or tendencies, there are several other syntactical means of indicating a general tendency. These involve adjectives (*likely, apt, liable* and *prone*), verbs (*tend* and *inclined*) and nouns (*inclination, tendency* and *propensity*). The two sets of nouns differ in their breadth of applicability. According to web Thesaurus, the latter (*inclination*, *tendency* and *propensity*) require a physical being while the former (*percentage*, *rate* and *chance*) can be applied to abstract mathematical relationships.

4.1 Adjective: *Likely*

Likely is the most common of the adjectives. *Likely* has two main forms: *likely to* [54K, 2410] and *likely that* [3.8K, 2420]. See also *unlikely to* [9.7K, 2431] and *unlikely that* [2.6K, 2432].

The 25 most common words following *likely to* are *be* (27%), *have* (5%), *get* (2%), *take* (1%), *become, make, go, see, come, find, remain, lead, develop, die, suffer, do, continue, cause, happen, occur, give, face, use, increase, fall,* and *vote.* The six most common tags for the words following *likely to* are VV (65%: *get, take, make, become, go, see),* VB (27%: *be*) and VH (5%: *have*). These verb tags contain 97% of the *likely to* phrases. Most words following *likely to* involve something optional (*get, become, die, vote*) or that could be optional (*to be*). Being optional – either in reality or in our awareness –creates the uncertainty that underpins the use of *likely*. Thus, *likely to* often introduces the part: the source of uncertainty. E.g., "Women are more likely to smoke than men." But some of the words following *likely to* – such as *happen, be* and *occur* – necessarily refer back to something prior – usually the subject of the clause. That makes the existence or state of that referent the part. E.g., "Smoking is more likely [to be found] among women than among men."

Analyzing the phrases following *likely that* is much more complex. Generally the predicate contains the uncertain or variable element while the subject sets the context.

4.2 Other Adjectives: *Apt, Liable* and *Prone*

These adjective phrases – *apt to* (2.7K), *liable to*, (1.3K) and *prone to* (870) – are much less common [see 2443] than *likely to.*

The 20 most common words following these phrases are *be, the, pay, a, have, do, get, make, become, flooding, take, depression, violence, prosecution, this, go, tax, such, injury,* and *being.* The 8 most common tags for the words following these adjective phrases are VV (*pay, do, get, make*), NN (*depression, flooding, violence, prosecution*), JJ (*such, violent, sudden, severe*), VVG (*making, getting, falling*), VB (*be*), NNS (*errors, earthquakes*), DT (*the* and *a*) and VH (*have*).

Examples of words following *apt|liable|prone to*: *be hit, the disease, pay capital gains, a fine, have the weapons, do so, get into trouble* or *make mistakes.*

4.3 Verbs: *Tend* and *Inclined*

These verb phrases *tend to* [present tense 15K VVP] and *inclined to* [past participle 1.6K VVN] are more common [see 2444] than the adjective phrases but much less common than *likely to.*

Tend to (85%) appears more than five times as frequently as *inclined to* (15%). In the phrase *tend to, tend* is tagged as a participle (VVP) 90% of the time and as a base verb (VV) the remaining 10%. In the phrase *inclined to, inclined* is tagged as a verb present tense (VVN) 92% of the time and as an adjective (JJ) the remaining 8%.

The words following these phrases (*tend to* and *inclined to*) are tagged mainly as verbs: base verbs (VV 70%), *to be* (VB 20%) or *to have* (VH 4%). The 10 most common words following these phrases are all verbs: *be, have, think, get, do, go, take, believe, make and see*. Examples: *teens tend to be mall shoppers*, *girls tend to outperform boys* and *I'm inclined to agree.*

4.4 Nouns: *Inclination, Tendency* and *Propensity*

Of these noun phrases [2445], *tendency to* (78%) is more common than *inclination to* (14%) and *propensity to* (9%). All three nouns are tagged as common noun singular (NN). The most common tags on the first word following these phrases are base verbs (VV 80%), common nouns (NN 9%), to be verbs (VB 5%) and adjectives (JJ 2%). The 10 most common words following these phrases are *be* (5%), *make* (1.6%), *go, take, get, think, become, do, see* and *use* (0.8%). Common nouns following these phrases include *violence, bully, insulin* and *depression.* Common adjectives following these phrases include *high, low, emotional* and *wild.* Examples: "their tendency to romanticize," "little inclination to change" and "a propensity to psychosomatic symptoms"

4.5 Analysis:

Although *apt*, *liable* and *prone* are all adjectives they can be used as either predicate adjectives (She was apt to win) or noun adjectives (the prone position). *Inclined* can also be used as a noun adjective (an inclined plane). Requiring *to* after each of these eliminates them as noun adjectives. Examples: "make computers less prone to breakdown", "older riders are more apt to wear helmets" and "men are more prone to stray than women".

According to web Thesaurus, "*Apt* is for general probabilities, *likely* is for specific probabilities; *liable* and *prone* indicate a probability arising as a regrettable consequence. Use *likely* if you mean 'probable, expected'; use *liable* if you mean 'bound by law or obligation'."

5. IMPLICIT AND EXPLICIT (SYNTATCTIC) COMPARATIVES

Comparing is a basic mental operation. As mentioned previously, a comparison can be implicit in the two numbers being related. These implicit comparisons may involve score comparisons or division comparisons. The explicit comparisons involve unique verbal syntax.

- Score comparisons require two numbers representing actual quantities being compared (grew from 2 to 3, won by a score of 4 to 3). The relation is implicit in the underlying numbers.
- Division comparatives require two numbers expressing ratios using prepositions (one in four, 12 per 1,000) or prepositional phrases (one [out] of every two). These need not be literal scores, but may be small integers chosen to readily indicate the ratio.
- Syntactic comparisons do not require two numbers (they may involve one or none); they use grammatical comparatives such as "more/less than", "bigger/smaller than", "times as much as", "% more/less than" and "times more/less than" to express an arithmetic relationship between two numbers. Syntactic comparisons focus on the relationship – qualitatively or quantitatively. E.g., three is 50% (one) more than two; 3% is one percentage point more than 2%.

5.1 Score Comparisons

The most common method uses *to* as in "we won 3 to 2" where the preposition *to* follows the score associated with the subject and introduces the score of the opponent unless otherwise indicated. *See* prior discussion of *to* in sections 1.3, 2.5 and 3.1. There are various short-hand techniques for presenting a comparison of two scores. These include the dash (we won: 4-3) and the colon (We won by a score of 4:3). In both cases, the dash and colon signify *to.*

Dash as ratio indicator: WordBanks has 275,628 matches with *of* followed by a number [2602]. Of these, 12% involve a number that is probably a year (between 1900 and 2050), 1% involve a range of years (e.g., 1940-2011), a few involve a date-month (e.g., 12-June), and 87% involve a non-date number. Of the 275,628 matches, 1,019 (<1%) were followed by a dash. Of these, none of the downloaded phrases had the form: of # - #. But several of the numbers had the form "20-1" treated as a single number. WordBanks has 1,248,000 matches with two numbers separated by a non-number [2620]. Of these 4% of the numbers were separated by a minus sign (dash). More work will be required to understand the role of the dash as a ratio indicator.

Colon as ratio indicator: Unfortunately looking for *of* followed by a number including a colon [2602] yields two distinct cases: (1) **time of day** where hours precedes the colon and minutes follow, and (2) **ratio of similar quantities**. The first can be eliminated when the first number exceeds 24 or the second exceeds 60. In the 22 cases involving a colon as part of a number, all could indicate a time of day based on their numeric values.

5.2 Division Comparisons

The most obvious form of a division comparison involve *each* or *every* preceded by a preposition such as *in*, *of* or *for*. [2640] While this form always involves a division comparison of two numbers, these numbers are seldom actual scores. Typically the denominator is the smallest integer such that the numerator is an integer. Thus a ratio of 18 out of 56 would be expressed as three out of seven. Comparing ratios with different denominators is not easy unless they have a common numerator such as *one*. This ratio (3/7) might be expressed as "more than one out of three."

Use of the slash. In X% of the lines involving a slash, the slash is preceded and followed by a number. In the lines where the slash is not preceded – and not followed – by a number, the most common use is to function as parenthesis or commas.

5.3 Syntactic Comparatives

Comparing is a basic mental operation. Comparisons can be ordinal (*12 is more than 10*) or quantitative (*12 is two more than 10*). There are many ways to express such comparisons.

Consider a simple comparison of heights*: Jim is taller than most men; Jim is a tall man; Jim is tall*. Now consider a comparison involving a group: *Doctors like Crest the most; More doctors like Crest; Doctors like Crest more*. In these comparisons, the basis of the comparison is implicit. (*More doctors like Crest than nurses like Crest).* In the last comparison, there is an additional ambiguity about whether the comparison involves frequency (*More doctors*) or severity (*Doctors like Crest more strongly than they like any other toothpaste*).

In English, there is a separate syntax for one class of comparatives that typically end with *er*. These comparatives can be either adjectives (*a faster car*) or adverbs (*This car goes faster*). The comparative adjectives [2801] are often followed by nouns (43%). The comparative adverbs [2802] are often followed by adjectives (40%), adverbs (30%) or verbs (20%).

In the WordBanks corpus, there are 1,042 comparative adjectives [2801] tagged as JJR. These involve 1,307,216 lines or uses. Here are the most prevalent with a cumulative total of uses: more (40%), better (48%), less (53%), higher (57%), further (61%), lower (64%), greater (66%), older (69%), worse (71%), younger (73%), smaller (75%), larger (76%), bigger (78%), easier (80%), earlier (81%), fewer (82%), stronger (83%), longer (84%), later (85%), closer (86%), wider (86%), cheaper (87%), harder (88%), deeper (88%), safer (89%), broader (89%) and tougher (90%). Of the words in the R1 position, 43% are tagged as nouns (e.g., *more time*, *more people*) while 31% are tagged as prepositions and subordinating conjunctions (e.g., *more of, more than*). Of comparative adjectives tagged as a preposition or subordinating conjunction, *than* (73%) is most common.

In the WordBanks corpus, there are 32 comparative adverbs [2802] tagged as RBR. They have 903,042 uses. Here is the list ordered by prevalence along with their cumulative total of uses: more (46%), later (62%), better (70%), earlier (77%), less (83%), longer (89%), further (93%), closer (94%), faster (95%), sooner (96%), worse (97%), harder (98%), farther (98%), higher (99%), lower (99%), louder (99%), smarter (99%), quicker (99%), easier (100%), deeper, healthier, stronger, tougher, heavier, tighter, wider, poorer, slower, leaner, rosier, gloomier and firmer.

Of these 32 comparative adverbs, only one (*sooner*) does not appear in the 1,042 comparative adjectives. The five most common pairs of syntactic comparatives for both adjective and adverbs are: more/less, better/worse, higher/lower, greater/fewer and bigger/smaller. Of the prepositions and subordinating conjunction words in the R1 position, *than* is most common (54%).

Even though *than* is the most common subordinating conjunction immediately following a syntactical comparative, there is no requirement that *than* follow immediately. A separate analysis [2803, 2804, 2805 and 2806] found that 28% of the lines containing a syntactical comparative are followed by *than* in the same sentence: 16% with no gap, 4% with a one-word gap, 6% with a gap of two to nine words and 2% with a gap of 10 to 99 words within the same sentence.

The presence of a comparative in a sentence does not require the explicit use of *than.* E.g., He wanted more money and a higher title. Consumer debt got bigger while savings got smaller. Since sentences involving comparatives need not involve *than*, this 28% prevalence of a trailing *than* is empirical evidence that *than* has a special significance in forming certain kinds of comparatives.

6. TIME-CHANGE COMPARISONS

Comparisons may involve changes over time. There is no syntactical marker for these words so this group is more difficult to identify. One way is to find situations where they are commonly used. Situations involving numbers seem promising but excluding dates and times is difficult. One way to accomplish both is to focus on percent changes. Three situations with percent changes seem promising: (1) *A #%* [*increase*], (2) *An* [*increase*] *of #%, (3)* [*increased*] *by #%.* Each situation is analyzed separately. The results are then analyzed together. Situations that were not analyzed include "at a # %", "to # %" and "the # %".

6.1 Time Comparisons in R1 Using "a #%" ___

WordBanks has 23,472 matching lines [2650]. The R1 words either *do* – or *do not* – require a time-change. The most common R1 words that *do not* indicate a time-change are *stake, share, chance, discount, pay, premium, annual, interest, rate, return, tax, loss* and *profit*.

The 25 most common R1 words that may indicate a time change are *increase, rise, drop, fall, jump, reduction, decline, gain, growth, cut, improvement, decrease, surge, lift, raise, boost, slump, slide, plunge, leap, advance, dip, increased, rollback* and *contraction*. Note that a time change can be either a change in focus from one subject or a temporal change within a given subject.

6.2 Time Comparisons in N0 Using "An __ of X%"

WordBanks has 4,768 matching lines [2670]. The words located in the node before *of* were classified into three groups:

- Comparatives that may indicate a time-change: *increase, rise, drop, fall, gain, decline, growth, loss, jump, reduction, decrease, cut, improvement, hike, expansion, leap, dip, etc.*
- Superlatives that may indicate a time-change: *high, minimum, maximum, low, peak, etc.*
- Others: *average, rate, return, yield, premium, margin, market-share, total, share, target, etc.*

6.3 Time Comparisons in L1 Using ___ "By X%"

WordBanks has 16,661 matching lines [2811]. Of the words in the L1 position, 54% are tagged as verbs (e.g., *inflation* [*increased*] *by X%*), 33% are nouns (e.g., *inflation increased* [*prices*] *by X%*), 9% are participles or adverbs (e.g., *inflation is* [*up*] *by X%*) and 4% are other.

In this situation, it appears that all the time-change words in L1 are verbs, participles or adverbs. The verb tenses include past (*increased*), present (*increase*), past-perfect (*had increased*), present-perfect (*increases*), and gerund (*increasing*). The only participles or adverbs are *up* and *down*.

A more focused situation is where "*by # PCT*" must be preceded by a verb, participle or adverb. WordBanks has 9,436 lines [2812]. This yields 417 unique words with 306 unique lemmas.

Based on this approach, here are the most common time-change lemmas followed by their cumulative percentage of uses: rise (20%), increase (35%), fall (49%), grow (62%), drop (66%), up (70%), decline (73%), reduce (75%), jump (77%), soar (79%), cut (80%), shrink (81%), expand (82%), decrease (83%), climb (84%), surge (85%), dip (85%), down (86%), slump (87%), im-

prove (87%), plunge (88%), back (88%), plummet (89%), support (89%), follow (89%), slip (90%). Note that these 26 lemmas are less than 9% of the 306 lemmas, but they involve 90% of the uses in the WordBanks corpus.

In this approach, the most common form of the lemma appears to be the past tense: 64% of the uses among the top four lemmas. Note the top four words in the list of comparative adjectives: increased, rose, fell and grew. These happen to be the past-tense forms of the four most common lemmas and they provide 4,542 lines for analysis when followed by "by #". [2813]

6.4 Time-Comparisons Conclusions

Table 1 summarizes the most common words generally used to indicate a change in time for these three situations. Notice the high degree of overlap and of rank order between the three cases.

Table 1: Time-Comparatives

Rank	an X% __	an _ of X%	_ by X%
1	Increase	Increase	Rise
2	Rise	Rise	Increase
3	Drop	Drop	Fall
4	Fall	Fall	Grow
5	Jump	Gain	**Up**
6	Reduction	Decline	Drop
7	Decline	Growth	**Down**
8	Gain	**Loss**	Decline
9	Growth	Jump	Reduce
10	Cut	Reduction	Jump
11	Improvement	Decrease	**Soar**
12	Decrease	Cut	Cut

Rank	an X% __	an _ of X%	_ by X%
13	Surge	Improvement	Shrink
14	Lift	Hike	Expand
15	Raise	Expansion	Decrease
16	Boost	Leap	Climb
17	Slump	Dip	Surge
18	Slide		Dip
19	Plunge		Slump
20	Leap		
21	Advance		
22	Dip		
23	Increased		
24	Rollback		
25	Contraction		

One difference is that the *up-do*wn words found in the 3rd situation are not appropriate in the first two situations. A second difference is that the words in the first two situations are generally nouns, whereas those in the third situation are generally verbs. In many cases the noun and verb are identical (e.g., *increase*) or there are equivalents (e.g., *growth* vs. *grow*), but in others there is no equivalent (e.g., *soar*) or the equivalent has a different meaning (e.g., *lost*).

One way to investigate whether the particular approach biased the outcome is to take the top 10 time-change comparatives, apply them to a new situation and see what order emerges.

Consider selecting on 10 top time-change comparatives that are followed immediately by a number. WordBanks has 57,674 matching lines [2820]. Here are the lemmas and their prevalence: up (40%), down (17%), rise (14%), fall (13%), drop (5.5%), gain (4.8%), increase (2.6%), grow (2.4%), decline (1.2%) and improve (0.4%). *Up* and *down* rank much higher here than in 2812.

6.5 Comparisons Conclusions

Some comparatives are generally used to indicate temporal comparisons (*increase, rise, drop* and *fall* along with other active verbs such as *cut*). Other comparatives are generally used to make non-temporal comparison (*more, better, less,* etc.): comparisons at the same moment in time.

- Non-temporal comparatives can be used to make temporal comparison: "*Jill weighs 2# more now than a year ago*" instead of "*Jill's weight increased by 2# during the past year*."
- Temporal comparatives can be used in making non-temporal comparisons: "*Eating nuts cuts risk of diabetes"* or "*playing video games increases risk of school dropout*" instead of "*People who eat nuts have a lower risk of diabetes [than those who don't]*" or "*People who play video games have a higher risk of being school dropouts [than those who don't]."*

The use of non-temporal comparatives to make temporal comparisons doesn't create any problem if the time referents are explicit. But the use of temporal comparatives to make non-temporal comparisons can be extremely problematic. In some cases, the context tells the average reader temporal comparatives are being used for non-temporal comparisons. E.g., "as weight increases among adults, height increases." In other cases the item being compared is a one-time event so a

temporal comparison is impossible. E.g., "Getting married decreases the risk of premature death." This still leaves cases where the role of time is ambiguous. E.g., "Drinking red wine decreases the frequency of migraines." Knowing whether the data source is a longitudinal (before-after) or cross-sectional study may eliminate some ambiguity. Because of these cross-over usages, it is difficult to compare the prevalence of time-ordered comparisons with the prevalence of time-independent comparisons.

7. COORDINATED COMPARATIVES

Comparisons may involve simultaneous changes. These can involve syntactical comparatives: "*The more X, the more Y*." These can involve the time-change comparatives: "*As X increases, Y increases*" or "*Y increases as X increases*". A necessary condition for all three is the presence of two comparatives within the same sentence. A necessary condition for the last two is the presence of *as* either before the first comparative or between the two comparatives.

Lines involving two syntactic comparatives (*more* or *less*) each preceded by *the* are easy to identify. [2946: 5,347 lines] The most common form is "*The more you get involved, the more likely*…" or "*The more things change, the more they stay the same.*" One problem is false hits: lines where the comparatives are unrelated: "*The richer man married the smarter woman*."

Lines involving the two comparatives, *more* or *less*, where the first is preceded by *as* are less common [2943, 48 lines]. Lines involving the two comparatives, *more* or *less*, where *as* is found between the two comparatives has not been analyzed. Lines involving two time-comparatives are harder to identify since time-comparatives are not tagged separately. Another variation involves only a single comparative and the phrase *for each* or *for every*. E.g., *Y increases by y for each additional unit in X*. This was not analyzed [2930].

CONCLUSION

English contains many grammatical devices for indicating the relationship between two quantities. In phrases between two numbers, certain prepositions (*to, of, in, for, by, over, from* and *per*) are often used to indicate quantitative relations. Additional syntax and/or semantics may be required to distinguish these. The tendency adjectives (*likely, apt, liable* and *prone*), tendency verbs (*tend* and *inclined*) and tendency nouns (*inclination, tendency* and *propensity*) can all indicate quantitative relationships. As for comparisons, there are some where one cannot distinguish a temporal (longitudinal) from a non-temporal (cross-sectional) comparison. (E.g., eating nuts cuts cancer risk). One cannot tell whether the comparison results from a shift-in-focus change between different subjects or from a temporal change within a single subject. This confusion applies to the coordinated comparatives as well.

Statistical educators should urge students to avoid ambiguity in forming quantitative comparisons. Words that normally indicate a change in time should not be used to describe non-temporal comparisons. Thus, action verbs should not be used to describe differences between groups. Instead of "Eating nuts cuts cancer risk" say "Cancer risk - lower in nut-eaters" or "Nut-eaters have lower cancer risk.

REFERENCES

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

Schield, Milo and Thomas Burnham (2001). Statistical Literacy: Reading Tables of Rates and Percentages, *2001 ASA Proceedings of the Section on Statistical Education.* See www.StatLit.org/pdf/2001SchieldASA.pdf.

ACKNOWLEGEMENTS

To Thomas V. V Burnham for assistance and comments.

APPENDIX: ACCESSING WORDBANKS

For more details, see www.StatLit.org/pdf/2011SchieldASA-WB.pdf.

Table 2: Tag Set Summary

Tag Description Examples

$ dollar Seen only: $ -$ $A Z$
`` opening quote mark ` ``
" closing quote mark ' "
(opening parenthesis ({
) closing parenthesis) }
, comma ,
: colon or ellipsis : ; ... - --
/ Slash /
SENT sentence terminator . ! ?

CC conjunction, coord and & but or both nor either plus neither yet versus vs. v. et minus less times whether so 'n
CD numeral, cardinal
DT determiner the a an this that some no all any those these another each every both either neither half many
EX existential there there
FW foreign word

IN preposition/sub-conj aboard about above across afore after against albeit along[side] although amid[st] around as at atop because before behind below beneath beside[s] between betwixt beyond by 'cause circa 'cos cum despite down during en except for from if in inside into lest like minus near[est] next notwithstanding of on off once onto opposite out[side] over past per plus re respecting sans since so **than** though thro through[out] thru 'til till toward[s] under[neath] unless unlike until unto up[on] versus via vis-à-vis vs. whereas whereupon whether while whither with[in] without

IN/that that -- most or all are nominalizer

JJ adjective or ordinal #
JJR adjective, comparative
JJS adjective, superlative

LS list item marker

Note: this tag summary is not an official document. It is based on our experience with the documentation and the data.

MD modal auxiliary ca[n't] can cannot could 'd 'll may might must need ought shall should will wo[n't] would NOTE: the [n't] are separate words!!!

NN noun, common, singular or mass
NNS noun, common, plural
NP noun, proper, singular
NPS noun, proper, plural

PDT pre-determiner: all such half quite nary
POS genitive marker ' 's
PP pronoun, personal: 'em he her[s[elf]] him[self] I it[self] me mine myself one[self] our[selves] she theirs them[selves] they us we you[rself]
PP$ pronoun, possessive: his their her its my our your[s]

RB adverb
RBR adverb, comparative
RBS adverb, superlative

RP particle up out off down over around on away through along in aside upon open apart unto whole

SYM symbol *] [/ + = @ _ > | ~ \ < and some single letters
TO "to" as preposition or infinitive marker
UH interjection

In verbs, the x stands for B (to be), H (to have) or V (all others).

Vx verb, base form
VxD verb, past tense
VxG verb, present participle or gerund
VxN verb, past participle
VxP verb, present, not 3rd person single
VxZ verb, present, 3rd person singular

WDT WH-determiner: that what whatever which whichever
WP WH-pronoun what who[m][ever]
WP$ WH-pronoun, possessive whose
WRB Wh-adverb: how[ever] when[ever] where[by] wherever why