Type-Token Theory and Bibliometrics

(DRAFT—please do not cite or distribute)

Jonathan Furner
Professor in the Department of Information Studies
University of California, Los Angeles
furner@gseis.ucla.edu

Introduction

The terms "type" and "token" were introduced by the American pragmatist philosopher Charles Sanders Peirce (1839–1914) in 1906 (Peirce, 1906, pp. 505–506). Peirce's distinction has proven useful in various fields as a model of the supposedly pervasive relationship between repeatable, instantiable, abstract objects (such as the single word "the") and their concrete instances (such as the numerous individual occurrences of that word). While the importance of probability theory for quantitative analyses of people's document-handling activities—analyses, for example, of the productivity of authors, or the citedness of publications—has long been recognized, the common understanding that the probability distributions of values of bibliometric variables may be treated as distributions of sets of tokens over sets of types (i.e., publications over authors, or citations over publications) is a more recent phenomenon, dating back only to the 1980s. The goal of this paper is to examine critically the assumption that the application of type–token theory to bibliometrics is warranted.

In the second section, the metaphysical foundations of type–token theory are reviewed, and a distinction is made between two different, though possibly complementary understandings of the type–token relationship: one in which this relationship is conceived as roughly equivalent to that between kinds and individuals, and another in which occurrences are identified as forming a third category that consists of neither types nor tokens.

In the third section, the history is traced of attempts to apply type–token theory in empirical studies of language use (in the field of quantitative linguistics) and document use (in the field of quantitative bibliography, i.e., bibliometrics). This section begins with an overview of some of the assumptions made and notation used in the description of probability distributions in general, and power-law distributions in particular. The discovery (and regular rediscovery) of a power-law regularity in the distribution of word-tokens over word-types—usually known as Zipf's law of word frequency—is highlighted as one of the most important catalysts for the development of bibliometrics as a scientific endeavor.

Lastly, in the final section, the utility and impact of the application of type–token theory to bibliometrics is assessed, and the prospects for future developments evaluated. The conclusion is reached that, while the importance of the type–token distinction for bibliometrics has at times been overplayed, opportunities for broadening the scope of type–token bibliometrics remain under-explored.

Types and Tokens in Metaphysics

"The world is everything that is the case." How many words? Eight, if we're counting word-*tokens*; six, if we're counting word-*types*, since two of those word-types—"the" and "is"—occur twice. Each word-token stands for, signifies, represents, denotes a particular word-type—viz., the type whose essential formal features are shared by the token. Tokens are said to instantiate types; they exemplify, embody, manifest, fall under, belong to types; they're occurrences, instances, members of types. Tokens are treated as individuals, singles, particulars, substances, objects; they're concrete, real, material. Types, on the other hand, are like sorts, kinds, forms, properties, classes, sets, universals; they're said to be abstract, ideal, immaterial.

The relationship between types and tokens is sometimes characterized as ontologically fundamental, in that the two categories are among those that comprise the basic elements of reality.[1] How is the type–token relationship precisely to be distinguished from other dichotomies said to be ontologically fundamental, such as the kind–individual relationship? The goal of this first section is to suggest one way in which types and tokens may be distinguished from properties and substances, kinds and individuals, abstracta and concreta, and universals and particulars. We begin by describing each of these dichotomies in turn.

Properties and substances

Some metaphysicians describe a world comprised of properties and substances. Typically, a thing X is said to be a *property* iff[2] there is something Y such that X is predicable of (i.e., is attributable to, is characteristic of) Y; X is a *substance* iff there is something Y such that Y is predicable of X. A few examples of properties are redness, wisdom, and meaningfulness; those things

[1] See Wetzel (2006, 2009) for comprehensive overviews of philosophical approaches to the study of concepts of type and token.
[2] i.e., if and only if.

that are red, wise, meaningful, etc. (i.e., that "have the property of" being red, wise, meaningful, etc.) are substances.

Kinds and individuals

Some metaphysicians describe a world comprised of kinds (a.k.a. categories, classes, sorts) and individuals. A thing X is a *kind* "iff there is something Y such that Y is an instance of X and Y is distinct from X"; while X is an *individual* "iff X is an instance of something Y (other than itself) and X itself has no instances (other than itself)" (Lowe, 1983, pp. 50–51). For example, the mountain kind is instantiated by individual mountains, the artifact kind by individual artifacts, the kind kind by individual kinds, and so on.

For any kind X—the mountain kind, the artifact kind, the kind kind, or any other kind—we may ask: What are the individually necessary and jointly sufficient *identity conditions* for instances of that kind? To put it this way is actually to conflate two separate questions:[3]

1.) What properties individuate (i.e., serve to distinguish) all instances of that kind from all instances *of a different kind*? For example: On what criteria are mountains to be distinguished from non-mountains? Among the properties that have been suggested as such criteria are high elevation, high relative relief, steep slope gradient, large land volume, small summit area, and short inter-valley distance.[4]

2.) What properties individuate any instance of that kind from any other instance *of the same kind*? For example: On what criteria is any one mountain to be distinguished from any other mountain (assuming we have already identified both as instantiations of the mountain kind)? The single property that is most commonly suggested as such a criterion is that of each instance's precise spatio-temporal coordinates (i.e., being located in a specific spatio-temporal position).

Abstracta and concreta

Some metaphysicians describe a world comprised of *abstracta* (a.k.a. abstract objects) and *concreta* (a.k.a. concrete objects). Treating concreta as a kind—i.e., the concretum kind—we may ask: What are the individually necessary and jointly sufficient identity conditions for instances of

---

[3] Some authors make a distinction between *individuation conditions* (addressed by the first question) and identity conditions (addressed by the second question).

[4] See, e.g., Gerrard (1990, pp. 3–5); for the limitations of this approach, however, see Smith and Mark (2003).

that kind? Among the properties that are commonly said to distinguish instances of the concretum kind *from instances of other kinds* are the following: (a) *materiality*: i.e., being constituted by matter; (b) *spatio-temporality*: i.e., occupying space and persisting through time; (c) *causal efficacy*: i.e., having the capacity to enter into causal relationships; (d) *endurability*: i.e., having the capacity to undergo and survive change; and (e) *physical form*: i.e., having size, shape, and color. The single property that is most commonly said to distinguish individual concreta *from one another* is—as in the case of individuals—the precise *spatio-temporal coordinates* of each concretum.

Universals and particulars

It is no trivial matter to determine the precise nature of the relationships between the property–substance distinction, the kind–individual distinction, and the abstractum–concretum distinction. Are any two of these six purportedly fundamental categories identical? For example, are the categories of kinds and properties equivalent, such that X is a kind iff X is a property? At first, it might seem as if this situation would only be complicated further if we were to allow an additional distinction to be made between *universals* and *particulars*. It is rare, however, for the universal–particular distinction to be characterized in a uniquely different way from all others. Some metaphysicians define universals in the same way as they do properties, and particulars in the same way as substances; others define universals in the same way as they do kinds, and particulars in the same way as individuals. An anonymous contributor to a standard dictionary of philosophy takes the former approach, for example;[5] while Jonathan Lowe (2006) is one who takes the latter route, constructing a "four-category ontology" in which substantial universals (e.g., the mountain kind) are instantiated by substantial particulars (e.g., individual mountains), and non-substantial universals (e.g., redness) are instantiated by non-substantial particulars (e.g., the redness of my shirt).

Types and tokens

Again, adding the *type–token* distinction to the mix might appear to complicate the situation even further. Is the type kind identical to the property kind, the kind kind, the abstractum kind, and/or the universal kind? And is the token kind identical to the substance kind, the indvidual kind, the

[5] "Things are particulars and their qualities are universals. So a universal is the property predicated of all the individuals of a certain sort or class. Redness is a universal, predicated of all red objects." (Flew 1979, p. 334).

concretum kind, and/or the particular kind? To address these questions, it is instructive to turn to the originator of the type–token distinction in the form in which it has been understood since the early twentieth century.

Writing in 1906, C. S. Peirce introduced the terms *type*, *token*, *tone*, and *instance*, defining them in the following way:

> A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words.[7] There will ordinarily be about twenty *the*s on a page, and of course they count as twenty words. In another sense of the word 'word,' however, there is but one word 'the' in the English language; and it is impossible that this word should lie visibly on a page or be heard in any voice, for the reason that it is not a Single thing or Single event. It does not exist; it only determines things that do exist. Such a definitely significant Form, I propose to term a *Type*. A Single event which happens once and whose identity is limited to that one happening or a Single object or thing which is in some single place at any one instant of time, such event or thing being significant only as occurring just when and where it does, such as this or that word on a single line of a single page of a single copy of a book, I will venture to call a *Token*. An indefinite significant character such as a tone of voice can neither be called a Type nor a Token. I propose to call such a Sign a *Tone*. In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type, and thereby of the object the Type signifies. I propose to call such a Token of a Type an *Instance* of the Type. Thus, there may be twenty Instances of the Type 'the' on a page. (Peirce, 1906, pp. 505–506; emphases in original).

As might be expected, Peirce draws his examples of types and tokens from the domain of *semeiotic*, his theory of signs. Presentations of the concept of *sign*, frequently varying in some large or small respect, abound in his papers; but one recurrent idea is a model relating entities of three kinds—*objects*, *representamina*, and *interpretants*. A representamen is a sign standing for some object; and an interpretant is a separate sign, for the same object, that is created "in the mind of a person" by a representamen (Peirce, 1897/1932, p. 228). "Representamen," "interpre-

[7] Peirce's footnote in the original: "Dr. Edward Eggleston originated the method."

tant," and "object" may be understood as corresponding loosely to "symbol" (or "term," "signal"), "thought" (or "concept," "sense"), and "referent," respectively, in later formulations of semiotic triangles by others.

For Peirce, each word "on a single line of a single page of a single copy of a book" is a "Single" object. All "Single" (i.e., individual) objects or events are to be known as *tokens*; *types* are "definitely significant Form[s]" that "determine" (or are "embodied" by) tokens; and the type that a token embodies is the type of which that token is said to be an *instance*.[9] Both token and type are said to "signify": A token is a sign both of the type of which it is an instance, and of the "object" signified by the type. It may be tempting to infer from this that "token" and "type" should be understood merely as synonyms for "representamen" and "interpretant," respectively. Such a reading is undermined by at least two factors, however.

One relatively insignificant objection is that "object" seems to be used in at least two different ways in the quoted passage—to refer both to the kind of thing that a token is, and to the kind of thing that is signified by both token and type—whereas it is not the primary function of representamina to stand for themselves.

The second difficulty is more important to address. Peirce talks of "twenty Instances of the Type 'the' on a page," and of the multiple occurrences of words "on … a single page of a single copy of a book." But he does not clarify how we should count the words on the pages of multiple copies of the same book. For example: Suppose we have two copies of the same page from the same book, each copy showing twenty instances of "the." Do we have forty instances of "the" in total, or still only twenty?[10]

The source of this difficulty is that there is a difference between instantiation-by-tokenization and instantiation-by-occurrence. In the case of the two copies of the same page from the same book, for example, we may count twenty tokens of the type "the," while simultaneously counting forty occurrences. The type–occurrence relationship would appear to correspond to the kind–individual relationship discussed earlier; the type–token relationship, on the other hand, is something new. To reduce ambiguity, then, "token" should be used as a name for the products of events of only one of these two kinds of instantiation, not both.

---

[9] On other occasions, Peirce used "sinsign" instead of "token," and "legisign" instead of "type."
[10] Williams (1936) was one of the first to stress the significance of this ambiguity, but his resolution is different from the one presented here.

It would appear that Peirce's type–token distinction is orthogonal, rather than equivalent, to his representamen–interpretant distinction. In the quoted passage, the focus is on representamina, and on simple linguistic symbols in particular: strictly speaking, the definitions given are of "*word*-type" and "*word*-token." We should be alert to the possibility of the type–token distinction's applying not only to words, but also to (a) more-complex linguistic symbols such as sentences; (b) aggregates of linguistic symbols such as the full texts of books and other textual documents; (c) non-symbolic signs such as icons and indexes;[11] (d) interpretants—concepts, propositions, beliefs, and other mental states; and (e) objects or referents (including events, properties, relationships, and states of affairs)—both natural and artifactual.

In this light, the type–token relationship begins to look a little more like the kind–individual relationship. That there is a difference, however, is demonstrable if we return to the token–occurrence contrast noted above. The latter distinction makes sense only when applied to signs. We can distinguish sensibly among word-types, word-tokens, and word-occurrences, but not among bird-types, bird-tokens, and bird-occurrences. To extend the type–token distinction to referents in general would, it seems, be one step too far.

We are left, then, with one view of the world in which kinds (e.g., the bird kind) are instantiated by individuals (e.g., Alex the parrot, 1976–2007[12]); and another in which types (e.g., the word "bird," and the book *Bird by Bird* by Anne Lamott) are instantiated by tokens (e.g., the seventeenth word of this paragraph, and the 1994 edition of Lamott's work), which in turn are instantiated by occurrences (e.g., the set of ink marks on my print-out of this paper, and my copy of the book). These two views may easily be reconciled if we equate kinds and types, equate individuals and occurrences, and allow for intermediate tokenization of signs only.

We shall return to this interpretation after considering, in the next section, the role of the type–token distinction in statistical linguistics and statistical bibliography. As a preliminary to that discussion, it may be helpful first to review some theoretical, conceptual, and terminological aspects of the statistical approach.

---

[11] Peirce (1911/1998, pp. 460–461) defined three main classes of sign: *icons*, "which serve to represent their objects only in so far as they resemble them in themselves"; *indices*, "which represent their objects independently of any resemblance to them, only by virtue of real connections with them"; and *symbols*, "which represent their objects, independently alike of any resemblance or any real connection, because dispositions or factitious habits of their interpreters insure their being so understood."

[12] See http://en.wikipedia.org/wiki/Alex_(parrot).

Types and Tokens in Linguistics and Bibliometrics

Power-Law Distributions

The field of statistics is concerned with *random variables*, i.e., observable properties (of events, cases, etc.) whose values are not predictable. Random variables whose possible values may be specified in a list of finite length are known as *discrete*; those that can take any numerical value are known as *continuous*.

In statistics, a *probability distribution* is "a description of the possible values of a random variable, and of the probabilities of occurrence of these values" (Upton & Cook, 2008). Any probability distribution is specifiable by a function $p_X(x)$ that relates each possible value x to the probability of occurrence $P(X=x)$ of that value—a.k.a. a *probability mass function* (pmf) for discrete variables, or a *probability density function* (pdf) for continuous variables. For the discrete random variable X whose possible values are $x_1, x_2, x_3, \ldots, x_M$, where M_X is the total number of possible values, the pmf $p_X(x)$ may be given by $f_x = n_x / N_X$, where n_x is the *absolute frequency* of occurrences of the value x, N_X is the total number of events, and f_x is thus the *relative frequency* of occurrences of the value x. To visualize in graphical form the probability distribution specified by such a pmf, one might simply plot values of the variable X on the abscissa (x-axis) of a histogram, against the absolute frequencies of occurrence n_x of each value on the ordinate (y-axis).[14] This way of characterizing a probability distribution, however, says nothing about the properties of the relation between values of X and their expected frequencies of occurrence; as a result, a probability distribution function typically specifies such a relation explicitly. Some commonly instantiated types of probability distribution include the discrete *uniform* distribution (which describes, for example, the rolls of a fair die; pmf $p_X(x) = 1 / M_X$), the *normal* or Gaussian distribution[15] (which describes, for example, people's heights), and the Pareto distribution[16] (which describes, for example, people's incomes).

A number of different methods of classifying general families of distributions have been defined by statisticians. Some distributions (e.g., the uniform and normal distributions) are sym-

[14] This presentation assumes the "frequency" interpretation of probability due to Venn (1876; see also Hájek 2011), which defines a value's probability as the limit of its relative frequency in a large number of trials.

[15] Named for the German mathematician Carl Friedrich Gauss (1777–1855).

[16] Named for the Italian economist Vilfredo Pareto (1848–1923; see Pareto, 1895, 1896/1965).

metric; others (e.g., the Pareto distribution) are asymmetric, a.k.a. *skew*. Among the skew distributions, some (e.g., the Pareto distribution) are *heavy-tailed* (i.e., they have tails that are longer and/or fatter than the tail of an exponential distribution); while others are light-tailed (i.e., they have tails that are shorter and/or thinner). The *Zipf* (a.k.a. zeta) distribution (pmf $p_X(x) = c \cdot x^{-a}$, where a and c are constants whose values depend on context)[17]—like the skew, heavy-tailed Pareto distribution of which it is the discrete version—is an example of a *power-law* distribution. In general, power-law distributions describe variables where events characterized by a large x are so rare, and events characterized by a small x are so common, that the probability of occurrence of a given value x is inversely proportional to a power (i.e., a in the pmf given above) of that value.

Power-law distribution functions can been fitted to empirical datasets on many different kinds of phenomena, both natural and social.[18] Power-law relationships have been observed not only in distributions of incomes of people, but also in distributions of magnitudes of earthquakes, populations of settlements, frequencies of occurrence of words, productivities of authors, and frequencies of occurrence of journal titles in bibliographic references or citations, among many others; see Table 1 for a summary.[19]

The last three in this list (again among others) have long been studied by *bibliometricians* interested in applying statistical techniques as a means of understanding people's document-related activities. Which words are used the most in German-language publications? Who in the field of biochemistry has been cited most often by philosophers? In which journals have papers about nanotechnology most frequently appeared? These are a small sample of the kinds of questions that may be answered simply by counting the number of times each value of a defined variable occurs in a given bibliographic dataset, and then comparing those counts to find the most frequently occurring values. Various bibliometric "laws," implying the existence of some sort of causal relationship between the values of a variable and their probabilities of occurrence, have been proposed as determinants of the distributions of probabilities—Zipf's law of word frequency (Zipf, 1929, 1932, 1935, 1949), Lotka's law of scientific productivity (Lotka, 1926), and Bradford's law of scattering (Bradford, 1934) are traditionally the "big three"—but it should al-

[17] Named for the American linguist George Kingsley Zipf (1902–1950; see Zipf, 1929, 1932, 1935, 1949).
[18] The degree of "goodness of fit" may be calculated by comparing the observed data with the data that would be expected if the function were accurate.
[19] See Newman (2005) and Clauset, Shalizi, and Newman (2009) for comprehensive reviews of the properties of power-law distributions and their occurrence in the natural and social worlds.

ways be borne in mind that here we are observing mere statistical regularity, or conformance to patterns, not the operation of laws in any way analogous to the laws of physics. In any case, it is even debatable which (if any) of these empirical datasets really are best-fitted by a power-law distribution, regardless of the values that are computed for its parameters. In some cases, the regularities observed are characteristic only of the middle range of the values of the defined variable, while some other distribution (e.g., the lognormal distribution) is a better fit for values in the upper or lower range.

(Table 1 here)

Three Different Terminological Approaches

The terminology used to discuss power-law distributions in general, and the bibliometric laws in particular, varies in accordance with the writer's interpretation of the nature of these distributions' contexts.

One approach, as taken above, is to talk of sets of *events* (a.k.a. individuals, cases, or objects), each characterized by a particular categorical variable (a.k.a attribute, or property), which takes *classes* (a.k.a. kinds, or categories) as values. We might say, "A set of events is distributed over a set of classes," and tally the events that *constitute* (belong to, are members of) each class, in order to produce a set of class-specific event-counts that take numerical values n_x representing the *size* of each class.

An alternative is to speak of sets of *items*, each characterized by a particular categorical variable that takes *sources* as values. We might say, "A set of items is distributed over a set of sources," and tally the items *produced* (generated) by each source, in order to produce a set of source-specific item-counts that take numerical values n_x representing the *productivity* of each source.

Thirdly, the terminology of classes and events (or sources and items) can be mapped to types and tokens, so that we consider sets of *tokens*, each characterized by a particular categorical variable that takes *types* as values. We might say, "A set of tokens is distributed over a set of types," and tally the tokens that *signify* (stand for) each type, in order to produce a set of type-specific token-counts that take numerical values x representing the *incidence* (a.k.a. prevalence) of each type.

Two Different Conceptual Approaches

The possibility that any presentation of a given distribution may involve any or any combination of these terminological approaches is not the only potential source of confusion for students of bibliometrics. The class/event relationship manifested in any sample dataset can be represented by either or both of two plots: a (class-)*rank*–(class-)*size* plot,[20] in which classes of events are listed on the *x*-axis in rank order (from largest to smallest), and the frequency of events in each class plotted on the *y*-axis; and a (class-)*size*–(class-)*frequency* plot, in which the various sizes of classes are listed on the *x*-axis (from smallest to largest), and the frequency of classes of each size plotted on the *y*-axis.[21] The plots in Figures 1 and 2 are derived from the data presented in Table 2. It is important to recognize that the two plots "are not contradictory or competing descriptions; rather they are complementary ways of summarizing the same data" (Herdan, 1960, p. 87). In Figures 3 and 4, the same data is plotted on a double-log scale, producing the straight line that is typical of power-law distributions.

(Figures 1, 2, 3, and 4, and Table 2 here)

To take the example of a random variable *X*, each value *x* of which is a different word-form: in a rank–size plot (e.g., Figure 1), the word-forms are listed on the *x*-axis in descending order of frequency of occurrence (a.k.a. "size"), and the frequency of occurrence of each word-form plotted on the *y*-axis; whereas in a size–frequency plot, the various sizes of word-forms (i.e., the various frequencies of occurrence) are listed on the *x*-axis, and the frequency of word-forms of each size plotted on the *y*-axis. In the case of the size–frequency plot (e.g., Figure 2), it is useful to think of class-sizes (e.g., the various possible frequencies of word-occurrence) as classes in their own right, and word-forms as the individual events in each class. In this way, we can conceive of the random variable *X* a little differently, such that each of its values *x* is a different class-size (e.g., a different frequency of word-occurrence).

Suppose, then, we are dealing with a population of *N* sources (e.g., word-forms), for each of which we can observe a value *x* of the random variable *X*, which is equal to the number of items (e.g., word-occurrences) produced by that source, i.e., the source's productivity. In this context, we can make the following observations, using notation similar to that adopted by Burrell (1991), among others.

---

[20] A.k.a. a (class-)*rank*–(event-)*frequency* plot.

[21] Rank–frequency and size–frequency plots are sometimes known as Zipfian and Lotkaian plots, respectively, after the authors with whom they were originally associated.

The number of sources that each have a productivity of *exactly x* is given by n_x; the combined productivity of those sources that each have a productivity of exactly x is given by $x \cdot n_x$; and the total productivity of all sources is given by $M = \sum x \cdot n_x$. The mean productivity (i.e., the average number of items per source) is given by $\mu = M / N$. The probability that a randomly selected source has a productivity of exactly x is given by $P(X = x) = f_x = n_x / N$; and the probability that a randomly selected item is the product of a source that has a productivity of exactly x is given by $g_x = x \cdot n_x / M$.

The rank of a source with a productivity of exactly x is given by r_x, and is equal to the number of sources that each have a productivity of *at least x*. The combined productivity of those sources that each have a productivity of at least x is given by R_x. The probability that a randomly selected source has a productivity of at least x is given by $P(X \geq x) = \Phi_x = r_x / N$, which is known as the *tail distribution function* (tdf) of X. The probability that a randomly selected item is the product of a source that has a productivity of at least x is given by $\Psi_x = R_x / M$, which is known as the *tail moment function* (tmf) of X. Plotting Φ_x against Ψ_x for all values of x produces a Leimkuhler curve[22] (see Figure 5).

The probability that a randomly selected source has a productivity of *at most x* is given by $P(X \leq x) = 1 - \Phi_x$, which is known as the *cumulative distribution function* (cdf) of X. The probability that a randomly selected item is the product of a source that has a productivity of at most x is given by $1 - \Psi_x$, which is known as the *cumulative moment function* (cmf) of X. Plotting $1 - \Phi_x$ against $1 - \Psi_x$ for all values of x produces a Lorenz curve[23] (see Figure 6).

The Leimkuhler and Lorenz curves are graphical representations of inequality (a.k.a., concentration, diversity, dispersion, richness). They allow us to find, for any given fraction of the total number of sources, what fraction of the total number of items are accounted for—i.e., to make statements like "the least-frequently occurring 50% of word-forms account for only 20% of word-occurrences," or "the most-frequently occurring 10% of word-forms account for 70% of word-occurrences." When $\Phi_x = \Psi_x$ (and $1 - \Phi_x = 1 - \Psi_x$) for all values of x , the amount of inequality is zero, and the curve is a straight line drawn from (0, 0) to (1, 1). The Gini index[24] G is a single-valued measure of the inequality of a probability distribution, given by the ratio of A (the

[22] Named for the American engineer Ferdinand F. Leimkuhler (b. 1928; see Leimkuhler, 1967).
[23] Named for the American economist Max Otto Lorenz (1876–1959; see Lorenz, 1905).
[24] Named for the Italian statistician Corrado Gini (1884–1965; see Gini, 1914).

area between the Leimkuhler [or Lorenz] curve and the 45° line of equality) to $A + B$ (the total area above [or below] that line).[25]

Having reviewed this statistical material, we are now ready to focus directly on the role played by type–token theory in the development of statistical approaches to linguistics and bibliography. We shall see that it is Zipf's work, not Peirce's, that has proved the more influential in both domains.

Zipf, Peirce, and Type–Token Theory: A Historical View

Zipf's law appears to have been first stated by the French stenographer Jean-Baptiste Estoup (1868–1950), in French, in 1916 (Estoup, 1916; see also Lelu, 2014), and first stated in English by E. U. Condon of Bell Telephone Labs in 1928. "While studying some data on the relative frequency of use of different words in the English language," writes Condon (1928, p. 300), "I noticed a rather interesting functional relationship …"

The Harvard linguist George Kingsley Zipf (1902–1950) developed the idea in a series of publications, beginning in 1929 with his doctoral dissertation, "Relative frequency as a determinant of phonetic change" (published as Zipf, 1929), in which, acknowledging the help of Estoup, he proposes (p. 4) a phonological "Principle of Frequency": the ease with which a word may be pronounced is "inversely proportionate to the relative frequency of that word … among its fellow words … in the stream of spoken language." In other words, "as usage becomes more frequent, form becomes … more easily pronounceable." Zipf uses statistical data on the frequency of occurrence of words supplied by Godfrey Dewey's *Relativ* [sic] *Frequency of English Speech Sounds*,[27] in which Dewey analyzes 100,000 word-occurrences in English text (instantiating just over 10,000 different words), and presents further statistical data, including some on Chinese, purportedly in support of his phonological thesis, in *Selected Studies of the Principle of Relative Frequency in Language* (Zipf, 1932).

---

[25] The Gini index G (a.k.a. Gini coefficient) is equivalent to Herdan's "Lorenz factor" L (Herdan, 1960, pp. 48–50). Herdan points out (p. 50, emphasis in original) that "for the lognormal distribution the *Lorenz factor depends only upon the value of the logarithmic standard deviation,* □ and can be read off immediately from a numerical table giving values of L for specified values of □," characterizing this result as one "of great importance" for quantitative linguistics.

[27] Godfrey Dewey's father was Melvil Dewey, the creator of the Dewey Decimal Classification. Another English "frequency dictionary" that came to be widely used was *The Teacher's Word Book of 30,000 Words* (Thorndike & Lorge, 1944).

Zipf's next major work, *The Psycho-Biology of Language* (Zipf, 1935), presents "in full" the results of his decade-long study of "speech as a natural phenomenon … investigated, in the manner of the exact sciences, by the direct application of statistical principles" (p. v). Here he argues not only that "the more complex any speech-element phonetically, the less frequently it occurs" (p. v), but also that "the length of a word … is closely related to the frequency of its usage—the greater the frequency, the shorter the word" (p. v), and that "if the number of different words occurring once in a given sample is taken as x, the number of different words occurring twice, three times, four times, n times, in the same sample, is respectively $1/2^2$, $1/3^2$, $1/4^2$, … $1/n^2$ of x, up to, though not including, the few most frequently used words; that is, we find an unmistakable progression according to the inverse square, valid for well over 95% of all the different words used in the sample" (p. vi). This evidence, Zipf says, "points quite conclusively to the existence of a fundamental condition of equilibrium between the form and function of speech-habits, or speech-patterns, in any language" (p. vi). By the time he came to write *Human Behavior and the Principle of Least Effort* (Zipf, 1949)—in which he again acknowledges the pioneering work of Estoup[28]—Zipf had generalized from this idea to a general theory of all kinds of human behavior, not just linguistic behavior, purporting to explain such behavior by reference to a fundamental principle that people tend, when required to carry out a task, to expend the least possible effort that is consistent with an adequately effective performance.

Whatever has been made of the explanation that Zipf infers from the evidence (and contemporary reviews were not wholly kind[29]), only a few have denied that the empirical relationship that he establishes between word frequency and rank is something to be explained. Over the years, however, the reliability of the data used, and the validity of conclusions drawn, have been called into question. Gustav Herdan (1960), for example, mounts a sustained attack, arguing that not only is Zipf's "law" not a law in the theoretical sense,[30] but that it is not even empirically

---

[28] "The first person (to my knowledge) to note the hyperbolic nature of the frequency of word usage was French stenographer J.-B. Estoup who made statistical studies of French …" (Zipf, 1949, p. 546).

[29] See, for example, E. Prokosch's coruscating review of *Selected Studies* … in *Language*: "An adequate review would consist in the two words 'utterly worthless,' and to say more seems waste of space. But … [t]he censure should be directed not so much against him as against those … who should have performed the duty of advising the Harvard University Press against accepting this book for publication. Zipf's book constitutes a disgrace to American scholarship …" (Prokosch, 1933, p. 92).

[30] "That the decrease of frequency [of word-occurrences] should be related to an increase in rank [of word-forms] follows not from any natural property of language structure, but merely from the fact that

true.[31] Herdan asserts (pp. 33, 35) that "[i]t is difficult to understand why the Zipf law should have attained such notoriety, … since it is not … of much practical use to the linguist, and mathematically a triviality. … [It] is the product of a period when quantitative methods were a novelty in linguistics. What was an achievement then is quite obsolete now." Herdan does allow (p. 38) that "the Zipf Law, although unsuitable for the scientific description of linguistic distributions, has its uses when it comes to the mechanical handling of word masses. … [I]t is often sufficiently close to the actual distribution to be of service in the technology of language, and we may regard it as a useful technological device." But he then goes on to argue that, in any case, the lognormal distribution is a much closer fit than the Zipf distribution is to word-count data.

Zipf did not use the terminology of "type" and "token" in his work, preferring instead simply to talk of the number of times words occur (or are used). The late 1930s and early 1940s saw the emergence of a research program in language behavior regarded as scientific by its proponents,[32] and the opportunity to relate Zipf's work to Peirce's gradually became apparent. One of the first to note the applicability of Peirce's terminology to discussions of Zipf's rank–frequency relationship was Wendell Johnson (Johnson, 1939), who discusses the *type–token ratio* (TTR)[33] and mentions that Zipf refrains from using the term—but Johnson does not cite

---

the word with the highest frequency is given the lowest rank, and as the frequency decreases the words are given correspondingly higher ranks. Thus the inverse relation between frequency and rank which is at the basis of the so-called Zipf law is one of our own making." (Herdan, 1960, p. 35).

[31] "… [A]ll kinds of exceptions have had to be suggested to make the 'law' fit the actual observations. According to some investigators, it does not hold for high-frequency words, nor does it hold for the low-frequency words, but seems to fit only the distribution of words of intermediate frequency. Considering that no definition is given … for high- and low-frequency …, it is evident that we cannot speak here of a law. … [T]he simple and straightforward relation between vocabulary and occurrence which it suggests [is] just not … true." (Herdan, 1960, pp. 35-37).

[32] See Sanford (1942) for an early review of research on "the existence, consistency, and significance of individual differences in the mode of verbal expression" (p. 811). Sanford draws attention to a development towards "a quantitative analysis and description of linguistic events … a quantitative science of language" (p. 813).

[33] "This is a measure of vocabulary 'flexibility' or variability, designed to indicate certain aspects of language adequacy. It expresses the ratio of different words (types) to total words (tokens) in a given language sample. If in speaking 100 words (tokens) an individual uses 64 different words (types), his TTR would be .64." (Johnson, 1944, p. 1). The value of the TTR tends to decrease as the sample size increases. Johnson explains how a *cumulative TTR curve*—possibly helpful in predicting TTRs for larger samples (cf. Chotlos, 1944)—can be plotted "by computing successive TTRs as increments are added to the sample" (Johnson, 1944, p. 2). Chotlos (1944) finds that the *bilogarithmic TTR*—i.e., the ratio of the logarithm of the number of types to the logarithm of the number of tokens—is constant for samples of different sizes from the same

Peirce. In a 1944 paper, Johnson notes that the effectiveness of the science-of-language program "depends upon the development of highly reliable and differentiating measures, by means of which specified aspects of language behavior might be systematically observed in relation to one another and to other variables" (Johnson, 1944, p. 1), and identifies the TTR as just such a measure. Even simpler, Johnson says, is the notion of *type frequency*, i.e., "the frequency of occurrence of each different word, or type" (p. 3)—but instead of compiling mere lists of the most-frequently occurring types in sample texts, à la Godfrey Dewey, the aim of the language behaviorists of the 1940s was to compare sets of type-frequency data for multiple individual language-users or group representatives, with a view to identifying characteristic patterns, group differences, changes over time, correlations with other variables, etc., while also distinguishing among types of different grammatical or semantic kinds.

In his overview of "highly reliable and differentiating measures," Johnson also discusses the concept of *proportionate vocabulary*: "How many different words or types make up 25, or 50, or 75 per cent of a given language sample?" (p. 4). He explains how to plot a curve representing the observed percentages of types (*x*-axis) that account for certain percentages of tokens (*y*-axis), and notes (citing Zipf, 1935) that this curve can be expressed (a) mathematically, and (b) in terms of rank as well as in terms of frequency.[34]

John B. Carroll appears to have been one of the first to mention both Peirce and Zipf in the same work. In his study of psychological aspects of linguistic behavior,[35] Carroll (1944) draws on the work of the semiotician Charles Morris to define and focus on a category of *linguistic response* that is broader than that implied by "word" or "phoneme," encompassing "communicative habits which do not specifically involve the speech mechanism; namely, non-vocal gestures, expressive movements, and other conventionalized responses" (p. 104).[36] Carroll points out (p. 107) that "it is necessary to introduce a distinction between the terms *response* and

---

text, and hence can be used as a single-valued characteristic of the style of a text. "This fact [is] one of the most remarkable in the field of quantitative linguistics …" (Herdan, 1960, p. 26).

[34] "[A] curve that is fitted to word-frequencies as a function of rank, the most frequent word having the lowest rank number, 1, represents in an alternative way the same phenomenon that is discussed here in terms of proportionate vocabulary." (Johnson, 1944, p. 5)

[35] "Our study is concerned, in the first instance, with the characteristics of verbal responses, the frequency with which these responses are emitted, the sequences in which they are patterned, and the general conditions of their occurrence." (Carroll, 1944, p. 102)

[36] For Morris, semiosis is a process that involves three entities: the sign-vehicle, the designatum, and the interpretant (see Carroll, 1944, p. 106). Cf. Peirce's representamen, object, interpretant.

response-type" that mirrors Peirce's type–token distinction.[37] However, Carroll cites Ogden and Richards (1936, Appendix D) as his source for Peirce's distinction.[38]

Meanwhile, in the course of his analysis of kinds of linguistic resource-types, Carroll (p. 113) describes his Phrase Completion Test, "in which the subject must give his first response to incomplete phrases like 'Hounds and _____'; 'And as for _____.'" He reports (p. 113) that "when a distribution is made of the responses to these items, it is found that two or three different responses constitute the majority of all the running responses, while a relatively large number of infrequent responses constitute the remainder of the responses," then notes (citing Zipf, 1935) that "in general these distributions, when frequency is plotted against descending rank order of frequency, follow roughly a Zipf-type curve."

The Moravian statistician and linguist Gustav Herdan (1897–1968) made a series of major contributions to the emerging field of quantitative (a.k.a. statistical) linguistics in the 1950s and 1960s, including three pioneering textbooks (Herdan, 1956, 1960, 1966), one of which (1960) was called *Type–Token Mathematics*. The concept of type–token duality, mined later by Egghe (see, e.g., Egghe, 2003), was central to Herdan's view of the field; yet he preferred to cite the distinction made by the Swiss linguist Ferdinand de Saussure (1857–1913) between *langue* and *parole* (roughly, abstract linguistic rules and concrete speech acts) as historical precursor, rather than Peirce (see, e.g., Saussure, 1916/1983).

Charls Pearson and Vladimir Slamecka's *Semiotic Foundations of Information Science: Final Project Report* (1977), drawing on Pearson's research from 1974 onwards, appears to be the earliest work in library and information studies (LIS) to cite both Zipf and Peirce on types and tokens, and is followed by further elaborations by Pearson and by his erstwhile colleague Pranas Zunde (see, e.g., Zunde, 1984). LIS writers began to cite Herdan around the same time (see, e.g., Pratt, 1975), but did not straightaway pick up on the applicability of type–token theory

---

[37] "The response-type is conceived here as an abstraction, a learned uniformity in linguistic behavior which has certain dynamic properties and which hence functions as a unit in behavior. In speaking of a linguistic response, on the other hand, we refer to a specific behavioral occurrence of a linguistic response-type. For example, the lexical form *dog* may be taken as a response-type, while a particular utterance of the sounds [d□g] would constitute a linguistic response. This distinction is quite similar to C. S. Peirce's distinction between *token* and *type* …, and is made in order to avoid the confusion between the specific and the generic usages of the term response often encountered in psychological writings." (Carroll 1944, p. 107).

[38] A few years later, Osgood (1952) discusses the TTR, cites Zipf and Morris, and mentions (but does not cite) Peirce.

to bibliometrics. Herdan's work was sufficiently well-known in bibliometric circles to be listed in J. Vlachý's bibliography of works relating to Lotka's law in volume 1, issue 1 of *Scientometrics* in 1978, and cited in J. J. Hubert's monumental review of "linguistic indicators" that appeared in 1980 (Hubert, 1980; see also Hubert, 1981).

By the late 1980s, Tague and Nicholls (1987, p. 155) were characterizing Zipf's law explicitly as "the distribution of a set of tokens over a set of types" (p. 155). Tague and Nicholls give several examples of other kinds of type–token pairs: author–publication, author–citation, publication–citation, and key–access (the last apparently indicating the distribution of search-term occurrences over search-term forms). From this time onwards, the terminology of types and tokens has become standard in bibliometrics. However, citations to Peirce's original work are still relatively rare.

Recent Developments

In a 1990 article summarizing the contributions made in his Ph.D. dissertation, Leo Egghe refers to the means by which sources such as authors, journals, etc., produce bibliographic items as "information production processes" (IPPs; Egghe, 1990, p. 17), and distinguishes one-dimensional bibliometrics—which "deal[s] with the sources or items separately (i.e., when they are not linked with each other)" (p. 18)—from two-dimensional or *dual* studies that examine the quantitative relationships between sources and items. Egghe asserts that every bibliometric problem can be addressed using either of two complementary approaches—"one looking at (sources, items), in that order, and the other looking at (items, sources), in the reverse order" (p. 19). Following Herdan (1960, pp. 14–15),[41] Egghe calls this "the *duality principle*," and compares it with the duality procedure in geometry, where "every time one obtains a theorem proving a relation between points and lines (in that order), one can formulate the dual theorem by interchanging the words lines and points" (p. 19). Egghe goes on to advocate for three- and even four-dimensional studies that involve more than one set of sources and/or more than one set of items (e.g., journals as well as authors and papers), and for examinations of the temporal aspects of IPPs.

By 2003, Egghe could write that "the dual approach" to bibliometrics—i.e., type–token (T/T), source–item, or Lotkaian bibliometrics—"is very well known" (Egghe, 2003, p. 603; see

[41] "This principle asserts for language that if in any valid proposition of language the words *type* (linguistic form, e.g., phoneme, morpheme) and *token* (frequency of occurrence, probability) are interchanged, the resulting proposition is also valid." (Herdan, 1960, p. 15).

also Egghe, 2005). In the same paper, Egghe introduces, as a "more important" part of informetrics (p. 604), what he calls type/token–taken (T/T–T) informetrics, which "studies the *use* of items rather than the items [themselves]" (p. 603; emphasis added) by describing the source–item relationship "as it is experienced by users (information professionals as well as information seekers)" (p. 606). Egghe proposes that, rather than focus only on distributions of sets of items over sets of sources, and on (e.g.) finding the probability that a randomly-selected word-form occurs j times, we also consider distributions of sets of sources over sets of items, and (e.g.) finding the probability that a randomly-selected word-occurrence is the product of a word-form that occurs j times. His rationale is that, in doing so, we will be better able to understand the ubiquitous scenario in which, for every value of j, the probability that a given item is the product of a source with a productivity of at least j is greater than the probability that a given source has a productivity of at least j.

For Egghe, the "taken" (i.e., use) component of his T/T–T formulation is a "third level" (p. 605) that "has never been studied" (p. 604). Quentin Burrell, however, argues that Egghe's proposal "adds little new to the theoretical framework of informetrics" (Burrell, 2003, p. 1263). Burrell identifies two random variables whose distributions form the core of Egghe's proposal: the variable X, each value of which denotes the productivity of a randomly chosen source, and whose distribution is defined by $f(j)$; and the variable Y, each value of which denotes the productivity of the source from which a randomly selected item comes, and whose distribution is defined by $g(j)$. Burrell shows that, in fact, the distribution of variable Y is "nothing more than the proportional tail-moment distribution of X" (p. 1261), while the relation between the distributions of X and Y "is illustrated by the familiar Leimkuhler curve of concentration" (p. 1261)—as we may confirm by comparing Egghe's definitions of $f(j)$ and $g(j)$ with the definitions of f_x and g_x given in the section on Two Different Conceptual Approaches, above.

Discussion and Conclusions

Not All Sources are Types

Bearing in mind the terminological distinctions noted above, we have now seen that to characterize as types and tokens the classes and events of interest to bibliometricians, and to other seekers of statistical regularities in human behavior, is a relatively recent phenomenon. It is also, we might conclude, a tactic that confuses rather than clarifies—for the simple reason that the type–

token distinction is quite different from both the source–item distinction and the class–event (a.k.a. kind–individual) distinction. To take the example of authors and publications: it is no stretch to see how each author may be conceived as the source of each of the items they produce, nor to understand their publications as events that belong to the class of those that share the property of being authored by the same person. It is more difficult, however, to grasp the rationale for treating each author as a type that is tokenized by publications, in the same way in which word-forms are tokenized by word-occurrences. The kinds of things that we typically consider to be tokenizable are representamina (words, sentences, texts, etc.) and interpretants (concepts, propositions, works, etc.). The terms we use to talk about these kinds of things are essentially ambiguous: context may make our meaning clear, but if it does not, then we can clarify only by specifying whether our subject is type or token. No such issue arises with authors and publications: we seldom mistake the class for the event. So, for a bibliometrician to invoke, sweepingly, the type–token distinction that works for words, but not for birds, is misleading at worst, and simply unnecessary at best.

Not All Type–Token Relations are Power Laws

Mitzenmacher (2004) provides a comprehensive review of the various explanations that have been given over the years for the apparent prevalence of power-law (and lognormal) distributions in empirical data. He identifies three families of generative models for power-law distributions, each of which received particular attention in the 1950s before their later rediscovery: *preferential attachment* models (see, e.g., Simon, 1955), *optimization* models (see, e.g., Mandelbrot, 1953), and *multiplicative process* models (see, e.g., Champernowne, 1953). Almost half a century before Mitzenmacher's review, Herdan disputes the assumption that large numbers of heavy-tailed distributions can be explained by the same model: "Simon's claim [in Simon, 1955] to have provided a uniform mathematical explanation of these distributions rests upon an insufficient realization of the differences in form between the distributions, and suffers from a neglect of considering the relations between some of them which makes it highly unlikely, if not mathematically impossible, that one mathematical model should fit them all" (Herdan, 1960, p. 207). Herdan's view is not only that the contextual differences between, for example, the distribution of word-occurrences and the distribution of personal wealth are sufficiently significant to warrant a search for explanations of different kinds, but also that closer inspection of individual datasets reveals patterning that fits just as closely with a distribution of some other (non-power-law) kind

entirely. Difficulties in distinguishing between instances of power-law and instances of lognormal distributions, especially, persist.

Not All Type–Token Relations Have Been Studied by Bibliometricians

The article in which Peirce originally presented his ideas on types and tokens was published to little notice from the wider philosophical community. It was the British philosopher Frank Ramsey (1903–30) who set the ball rolling in 1923 with his influential review of Ludwig Wittgenstein's *Tractatus Logico-Philosophicus*, in the course of which Ramsey uses the type–token distinction to explain aspects of Wittgenstein's picture theory of language (Ramsey, 1923; see also Nubiola, 1996, for a detailed account of Ramsey's role in the dissemination of Peirce's thought). Since then, the metaphysical status of types and tokens has been the subject of much philosophical work (see, e.g., Wetzel, 2009; Hilpinen, 2012), very little of which has been recognized as having implications for LIS in general or bibliometrics in particular. One of the directions taken in philosophy of language, philosophy of literature, and philosophy of art has been to explore the ramifications of sentences, propositions, pictures, etc.—as well as aggregations of such phenomena at various levels—having type–token ambiguity (see, e.g., Stevenson, 1957; Jacquette, 1994; Howell, 2002). In LIS, meanwhile, a homegrown variation on type–token theory has emerged in the modeling of resource description data, where an analogous distinction between works and items is drawn in standards such as the *Functional Requirements for Bibliographic Records* (FRBR; IFLA, 1998). It is clear that the work carried out in philosophy is relevant to the ontology of bibliographic phenomena that forms the core of contemporary library cataloging and classification theory, and vice versa, but the connections have received little attention from either side. Even more conspicuous by its absence is a bibliometric perspective on FRBR and related models. What probability distribution functions best describe empirical data on numbers of works, expressions, manifestations, and items, and what explanations can be given for the processes producing such distributions?

Not All Bibliometricians Are Data Scientists (Yet)

It is already somewhat of a cliché that we live in an age of "big data." One upshot of the increasing scholarly interest in practical questions to do with the most effective means of managing large datasets has been a corresponding surge in the level of attention given to philosophical questions about the nature of data, data models, database records, etc. There is a long story that remains to be told about the development of standard database structures based on the modeling

of entities and relationships, attributes and values, etc., against the backdrop of philosophical ideas about the ontological status of substances and properties, kinds and individuals, and—yes—types and tokens. The field of bibliometrics is both a participant in, and a contributor to the telling of this story. And there is much more to do than fitting power-law functions to distributions of links among websites. Bibliometricians are the natural pioneers of a science of data use that makes use of type–token theory in as judicious a manner as did the language behaviorists three-quarters of a century previously.

Cited References

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermann's Geographische Mitteilungen*, 59(1), 74–77.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *137*(3550), 85–86.

Bradford, S. C. (1948). *Documentation*. London, England: Crosby Lockwood.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics*, *21*(2), 181–194.

Burrell, Q. L. (2003). Type/token–taken informetrics: Some comments and further examples. *Journal of the American Society for Information Science and Technology*, *54*(13), 1260–1263.

Carroll, J. B. (1944). The analysis of verbal behavior. *Psychological Review*, *51*(2), 102–119.

Champernowne, D. G. (1953). A model of income distribution. *Economic Journal*, *63*(250), 318–351.

Chotlos, J. W. (1944). Studies in Language Behavior, IV: A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, *56*(2), 75–111.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

Condon, E. U. (1928). Statistics of vocabulary. *Science*, *68*(1733), 300.

Corbet, A.S. (1941). The distribution of butterflies in the Malay Peninsula (Lepid.). *Proceedings of the Royal Entomological Society of London, Series A: General Entomology*, *16*(10–12), 101–116.

Dewey, G. (1923). *Relativ* [sic] *frequency of English speech sounds*. Cambridge, MA: Harvard

University Press.

Egghe, L. (1990). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science*, *16*(1), 17–27.

Egghe, L. (2003). Type–token/taken informetrics. *Journal of the American Society for Information Science and Technology*, *54*(7), 604–610.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Bingley, England: Emerald.

Estoup, J.-B. (1916). *Gammes sténographiques: Méthode et exercices pour l'acquisition de la vitesse* (4th ed.). Paris, France: Institut Sténographique de France.

Fleming, T. P., & Kilgour, F. G. (1964). Moderately and heavily used biomedical journals. *Bulletin of the Medical Library Association*, *52*(1), 234–241.

Flew, A. (Ed.). (1979). *A dictionary of philosophy*. London, England: Pan.

Gerrard, A. J. (1990). *Mountain environments: An examination of the physical geography of mountains*. London, England: Belhaven Press.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del R. Istituto Veneto di Science, Letter ed Arti*, *73*(2), 1203–1248.

Gutenberg, B., & Richter, C. F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, *34*(4), 185–188.

Hájek, A. (2011). Interpretations of probability. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Stanford, CA: Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. http://plato.stanford.edu/entries/probability-interpret/

Herdan, G. (1956). *Language as choice and chance*. Groningen, The Netherlands: Noordhoff.

Herdan, G. (1960). *Type–token mathematics: A textbook of mathematical linguistics*. The Hague, The Netherlands: Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Berlin, West Germany: Springer.

Hilpinen, R. (2012). Types and tokens: On the identity and meaning of names and other words. *Transactions of the Charles S. Peirce Society*, *48*(3), 259–284.

Howell, R. (2002). Ontology and the nature of the literary work. *Journal of Aesthetics and Art Criticism*, *60*(1), 67–79.

Hubert, J. J. (1980). Linguistic indicators. *Social Indicators Research*, *8*(2), 223–255.

Hubert, J. J. (1981). General bibliometric models. *Library Trends*, *30*(1), 65–81.

International Federation of Library Associations and Institutions. Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records*. Munich, Germany: K. G. Saur.

Ishimoto, M., and Iida, K. (1938). Observations sur les séismes enregistrés par le microsismographe construit dernièrement (I) [In Japanese]. *Bulletin of the Earthquake Research Unit, University of Tokyo*, *17*(2), 443–478.

Jacquette, D. (1994). The type–token distinction in Margolis's aesthetics. *Journal of Aesthetics and Art Criticism*, *52*(3), 299–307.

Johnson, W. (1939). *Language and speech hygiene: An application of general semantics; Outline of a course*. Chicago, IL: Institute of General Semantics.

Johnson, W. (1944). Studies in Language Behavior, I: A program of research. *Psychological Monographs*, *56*(2), 1–15.

Lamott, A. (1994). *Bird by bird: Some instructions on writing and life*. New York, NY: Pantheon Books.

Leimkuhler, F. F. (1967). The Bradford distribution. *Journal of Documentation*, *23*(3), 197–207.

Lelu, A. (2014). Jean-Baptiste Estoup and the origins of Zipf's law: A stenographer with a scientific mind (1868–1950) [In Spanish]. *Boletín de Estadística e Investigación Operativa*, *30*(1), 66–77. http://www.seio.es/BEIO/files/BEIOVol30Num1Feb2014-HyE.pdf

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–325.

Lowe, E. J. (1983). Instantiation, identity, and constitution. *Philosophical Studies*, *44*(1), 45–59.

Lowe, E. J. (2006). *The four-category ontology: A metaphysical foundation for natural science*. Oxford, England: Oxford University Press.

Mandelbrot, B. (1953). An informational theory of the statistical structure of languages. In W. Jackson (Ed.), *Communication theory* (pp. 486–502). Woburn, MA: Butterworth.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*(2), 226–251.

Motomura, I. (1932). Statistical method in animal association [In Japanese]. *Dōbutsugaku Zasshi* [*Zoological Magazine*], *44*(2), 379–383.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Nubiola, J. (1996). Scholarship on the relations between Ludwig Wittgenstein and Charles S. Peirce. In I. Angelelli and M. Cerezo (Eds.), *Studies on the history of logic: Proceedings of the III. Symposium on the History of Logic* (pp. 281–294). Berlin, Germany: Walter de Gruyter.

Ogden, C. K., and Richards, I. A. (1936). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism* (4th ed.). London, England: Kegan Paul, Trench, Trübner.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197–237.

Pareto, V. (1895). La legge della domanda. *Giornale degli Economisti*, 2nd series, *10* (January), 59–68.

Pareto, V. (1896/1965). La courbe de la répartition de la richesse. In G. Busino (Ed.), *Écrits sur la courbe de la répartition de la richesse* (pp. 1–15). Genève, Switzerland: Librarie Droz.

Pearson, C. & Slamecka, V. (1977). *Semiotic foundations of information science: Final project report*. Atlanta, GA: School of Information and Computer Science, Georgia Institute of Technology.

Peirce, C. S. (1897/1932). On signs: Ground, object, and interpretant. In C. Hartshorne & P. Weiss (Eds.), *The collected papers of Charles Sanders Peirce, Volume 2: Elements of logic* (pp. 227–229). Cambridge, MA: Harvard University Press.

Peirce, C. S. (1906). Prolegomena to an apology for pragmaticism. *The Monist*, *16*(4), 492–546.

Peirce, C. S. (1911/1998). A sketch of logical critics. In The Peirce Edition Project (Ed.), *The essential Peirce: Selected philosophical writings, Volume 2 (1893–1913)* (pp. 451–462). Bloomington, IN: Indiana University Press.

Pratt, A. (1975). The analysis of library statistics. *Library Quarterly*, *45*(3), 275–286.

Prokosch, E. (1933). [Review of the book *Selected studies of the principle of relative frequency in language,* by G. K. Zipf]. *Language*, *9*(1), 89–92.

Ramsey, F. P. (1923). [Review of the book *Tractatus logico-philosophicus*, by L. Wittgenstein].

Mind, *32*(128), 465–478.

Sanford, F. H. (1942). Speech and personality. *Psychological Bulletin*, *39*(10), 811–845.

Saussure, F. de (1916/1983). *Course in general linguistics* (C. Bally & A. Sechehaye, Eds.; R. Harris, Trans.). London, England: Duckworth.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42(3–4), 425–440.

Smith, B., & Mark, D. M. (2003). Do mountains exist? Towards an ontology of landforms. *Environment and Planning B: Planning and Design*, *30*(3), 411–427.

Stevenson, C. L. (1957). On "What is a poem?" *Philosophical Review*, *66*(3), 329–362.

Tague, J., & Nicholls, P. (1987). The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing & Management*, *23*(3), 155–170.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Teachers College, Columbia University.

Upton, G., & Cook, I. (Eds.). (2008). *A dictionary of statistics* (2nd ed.). Oxford, England: Oxford

University Press. http://www.oxfordreference.com/view/10.1093/acref/9780199541454.001.0001/acref-9780199541454

Venn, J. (1876). *The logic of chance* (2nd ed.). London, England: Macmillan.

Vickery, B. C. (1948). Bradford's law of scattering. *Journal of Documentation*, 4(3), 198–203.

Vlachý, J. (1978). Frequency distributions of scientific performance: A bibliography of Lotka's law and related phenomena. *Scientometrics*, *1*(1), 109–130.

Wetzel, L. (2006). Types and tokens. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*, Stanford, CA: Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. http://plato.stanford.edu/entries/types-tokens/

Wetzel, L. (2009). *Types and tokens: On abstract objects*. Cambridge, MA: MIT Press.

Williams, D. C. (1936). Tokens, types, words, and terms. *Journal of Philosophy*, *33*(26), 701–707.

Willis, J. C., & Yule, G. U. (1922). Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, *109*(2728), 177–179.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in*

Classical Philology, *40*, 1–95.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston, MA: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Zunde, P. (1984). Empirical laws and theories of information and software sciences. *Information Processing & Management*, *20*(1–2), 5–18.

Table 1. Some empirical phenomena that purportedly follow a power-law distribution.

Common name (if applicable)	Early sources	Classes	Events	Event-count
The Pareto law	Pareto (1895, 1896/1965)	Persons	Dollars	Wealth
The rank–size rule	Auerbach (1916)	Settlements	People	Population
Zipf's law	Estoup (1916); Condon (1928); Zipf (1929, 1932, 1935, 1949)	Words	Occurrences	Occurrence-count
The Willis–Yule distribution	Willis & Yule (1922)	Taxa	Subtaxa	Subtaxon-count
Lotka's law	Lotka (1926)	Authors	Publications	Productivity
Bradford's law	Bradford (1934, 1948); Vickery (1948)	Journals	Citations	Citedness
The Gutenberg–Richter law	Ishimoto & Iida (1938); Gutenberg & Richter (1944)	Earthquakes	Joules	Magnitude
The species abundance distribution (SAD)[47]	Corbet (1941)	Species	Individual organisms	Abundance
—	Fleming & Kilgour (1964)	Journals	Uses	Use-count

[47] The distribution of individual organisms over species was originally modeled as a geometric distribution (Motomura, 1932), and has since been modeled most frequently as either a logarithmic or a lognormal distribution.

Table 2. Sample data consistent with a power-law distribution. Each value in the column headed x represents a different class-size, and each value in the column headed n_x is the number of classes that have the corresponding size x. We might imagine a text comprising 1374 word-occurrences, distributed over 404 word-forms, so that 271 of those word-forms occur once, 53 occur twice, and so on.

x	n_x	$x \cdot n_x$
1	271	271
2	53	106
3	23	69
4	13	52
5	8	40
6	6	36
7	4	28
8	3	24
9	2	18
10	2	20
11	2	22
12	1	12
13	2	26
14	1	14
15	1	15
17	1	17
18	1	18
20	1	20
21	1	21
22	1	22
25	1	25
29	1	29
33	1	33
40	1	40
50	1	50
67	1	67
100	1	100
200	1	200
Sum	**404**	**1374**

Figure 1. A partial rank–size plot derived from the data in Table 2.

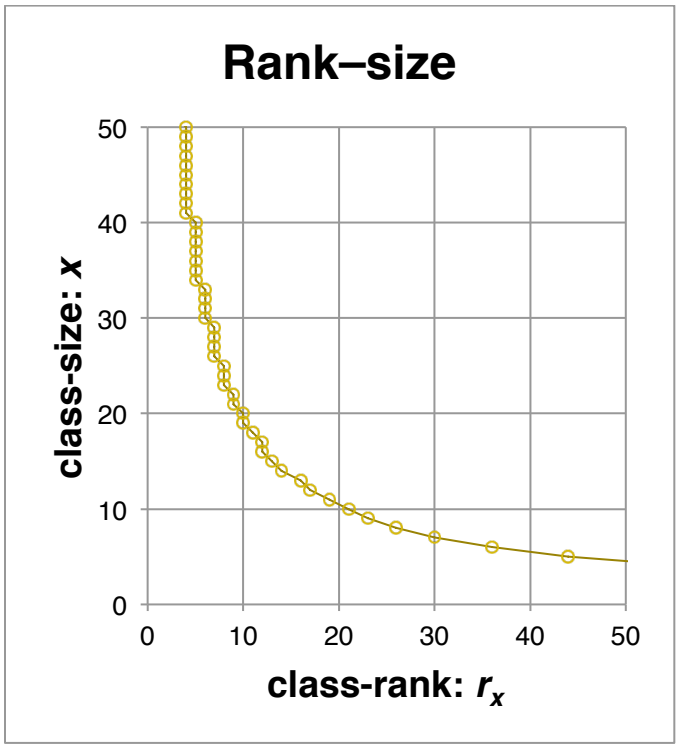

Figure 2. A partial size–frequency plot derived from the data in Table 2.

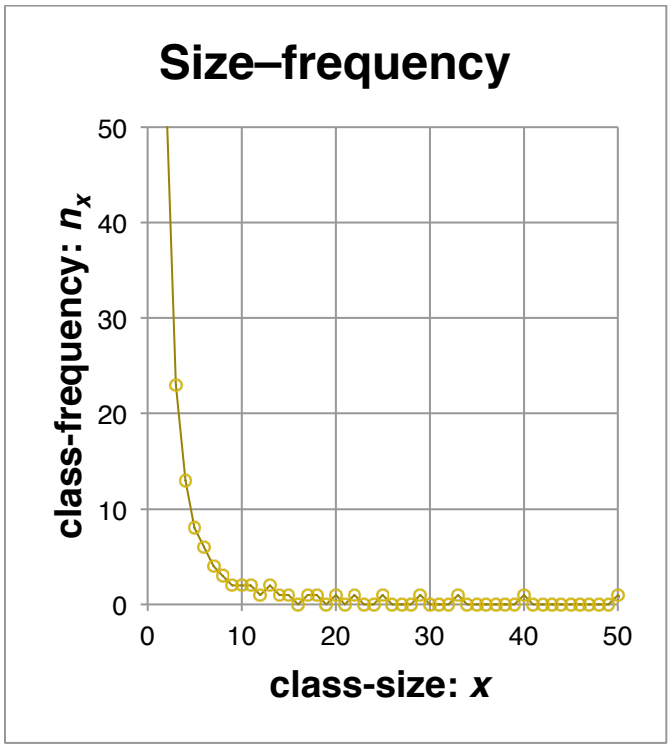

Figure 3. The data from Figure 1 plotted on a double-log scale.

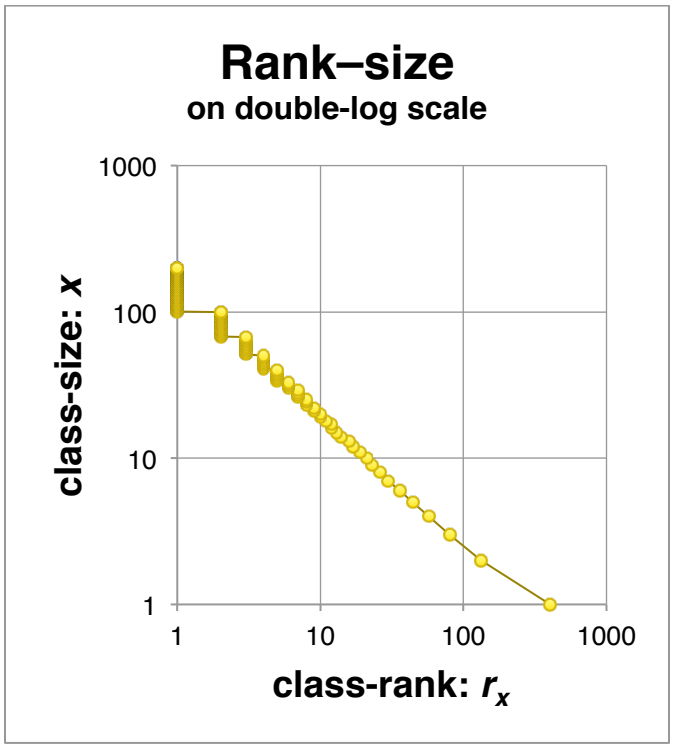

Figure 4. The data from Figure 2 plotted on a double-log scale.

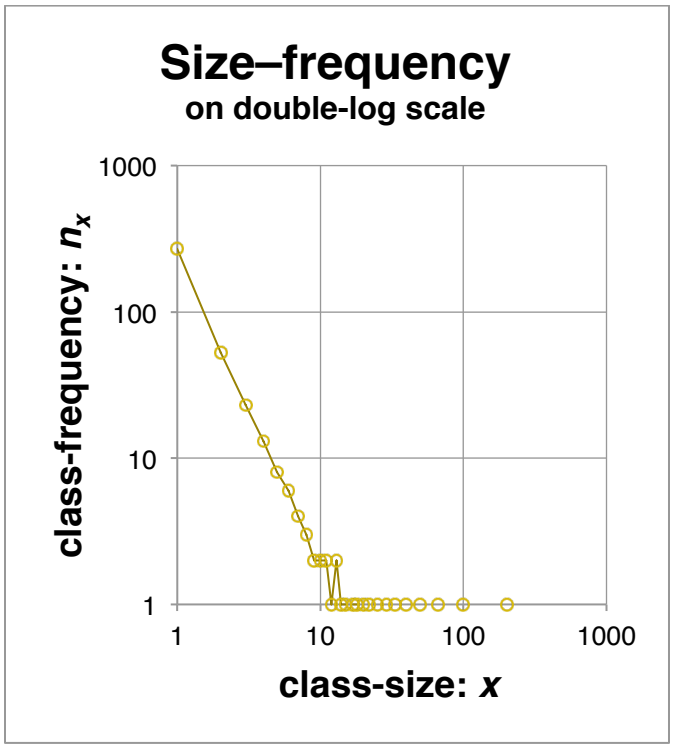

Figure 5. The Leimkuhler curve derived from the data in Table 2.

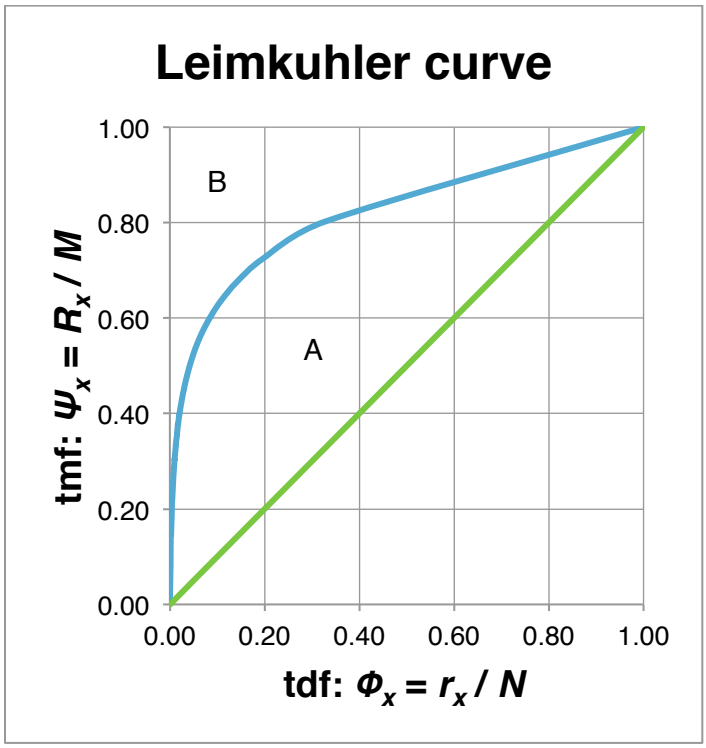

Figure 6. The Lorenz curve derived from the data in Table 2.