UNPUBLISHED MANUSCRIPT
Do not *cite* or *alter* this document in any way

Written	By
June 3, 2011	Erin E. Dahlgren

The Statistical Turn in Linguistics

Erin E. Dahlgren

Our story begins on August 18th, 1851, in little nook of London called Camden Street, where a forty-five year old Cambridge man is sitting down at his writing desk to pen a letter. He writes of his wife and seven children who are presently residing in the little town of Broadstairs in Kent. He writes of an old friend in Alvescot, of another Cambridge man who calls himself a Puseyite, and finally comes to write the following:

> I wish you would do this: run your eye over any part of those of St. Paul's Epistles which begin with $\Pi\alpha\nu\lambda o\varsigma$—the Greek, I mean—and without paying any attention to the meaning. Then do the same with the Epistle to the Hebrews, and try to balance in your own mind the question whether the latter does not deal in longer words than the former.

This Cambridge man closes his letter with the following:

> If scholars knew the law of averages as well as mathematicians, it would be easy to raise a few hundred pounds to try this experiment on a grand scale. I would have Greek, Latin, and English tried, and I should expect to find that one man writing on two different subjects agrees more nearly with himself than two different men writing on the same subject. Some of these days spurious writings will be detected by this test. Mind, I told you so.
>
> With kind regard to all your family, I remain, dear Heald,
>
> Yours sincerely,
> A. De Morgan

This is the very Augustus De Morgan, mathematician, whose logical laws are now standard in many modern philosophic and semantic proofs[1]. Here, his suggestion for a statistical-comparative treatment of language marks the first that this author has found in the whole history of linguistics. Two things in particular stand out: this suggestion is made by a mathematician, not by a linguist nor by a philologist. Two, this Cambridge man is not interested in language on its own, but in the detection of authentic authorship—arguably for the sake of a *history.* This is the only record we have of De Morgan's interest in textual-linguistic statistics; this theme does not appear in any other letters, in any of his formal works, and there is no mention as to whether the study proposed here was ever actually carried out. There are also no records showing that other British academics at the time where entertaining similar ideas, and we may be convinced by his final line: *Mind, I told you so*, that De Morgan himself knew of no others either.

[1] $\neg(A \wedge B) \rightarrow \neg A \vee \neg B, \neg(A \vee B) \rightarrow \neg A \wedge \neg B$

A year has passed, it is now 1852, and we find ourselves about 475 miles from Camden Street in the Prussian Province of Saxony, in a town called Wernigerode. There is employed a librarian of peculiar interests. Fascinated by language and numbers since his youth, and having been trained in comparative philology, he now finds himself comfortably settled outside of the Berlin academic scene and asking several unusual questions:

> What can we use to calculate a *distance* between two languages?
> And for this, what aspects of languages should we use?

This librarian hastens to begin his experiments with three languages he likely encountered while at University: Gothic, Latin, and Greek. He gathers their sound inventories and applies the following original method:

> Take the relative frequency of each phoneme (= each minimally contrastive sound), for each language, and express these frequencies in *percents*.
>
> Take the relative frequencies for two languages at a time and find the *percent difference* for each phoneme, between the languages.
>
> Take the differences corresponding to consonant phonemes, sum (average) over all individual differences; take the differences corresponding to vowel phonemes, sum (average) over all individual differences.
>
> The lowest possible resulting number is zero and will arise when the two languages are identical (the relative frequencies are exactly the same, for every phoneme). The highest possible resulting number is 200 and will result when no one sound from one language occurs in another.

The following emerged:

	Greek-Latin	*Greek-Gothic*	*Latin-Gothic*
Consonants	46	80	78
Vowels	64	102	96

Easily seen here is that the differences in the vowel systems of any two languages are always greater than the differences in the consonant systems. A second observation: Of the three comparisons, Greek and Latin are the most similar in their sound systems and Greek and Gothic are the farthest; we can see this more easily by summing the vowel-consonant differences for each column.

What this librarian has done, which may not be so obvious, is to give a concrete measure of variation or deviation, between languages. If enough language pairs were tested and enough linguistic aspects involved (not only phoneme frequency, but morpheme, word, phrase-structure frequencies), upper and lower bounds for a standard deviation within language could be estimated.[2]

Our librarian, though, makes a point to tell us that his "numerical method" suffers from many weaknesses. One such weakness is the measure's blindness to sample size—there is no way to compare the relative distances found from mere samples to what the distances may actually look like from an ideal, infinitely large sample. This weakness, though, is far lesser than what the method accomplishes: Ernst Förstemann, comparative linguist and librarian, has just announced to the world that just as structured forms and sounds are legitimate aspects of a language, *so too* are relative frequencies.

Our story now jumps forward roughly 30 years, to a point in history that does not, in fact, concern language whatsoever. It is 1881 and Simon Newcomb, director of the United States Naval Observatory's Nautical Almanac Office, is flipping through the pages of his logarithm books (at the time used to carry out basic mathematical calculations). Three years previous he had begun at this post with a particular project in mind: to recalculate all of the major astronomical constants, albeit with the help of his friend and assistant George William Hill. While engaged in this pursuit, he read widely in economic theory and would later be accredited for the first formulation of the equation of exchange between money and goods[3].

On this occasion, something strikes him about his books: the earlier pages were far more worn than later pages. He considered, might numbers from an arbitrary set of data, like a list of page numbers needed for a set of logarithmic calculations, tend to begin with *1*, more than any other digit? At the time, and even at present, such a hypthothesis sounds farfetched. Yet Newcomb wrote up a mathematical formulation and quickly moved to other work.

In a sense, it is somewhat ironic that Newcomb's digit-frequency hypothesis arose from physically handling logarithm books. Frank Benford, apparently unaware of Newcomb's work, came to formulate and convincingly defend (with much more data than Newcomb) the very same digit-frequency hypothesis, but not until 1938. The hypothesis was such that the digit-frequencies followed a power-law distribution, phrased by Newcomb in terms of logarithms:

> The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable. (Newcomb, p. 40)

[2] By this particularly insightful librarian, the only other language that was added into these comparisons was German.

[3] $M \cdot V = P \cdot Q$,
M = average amount of money in circulation, V = velocity of money (spending), P = price level, Q = an index on expenditure.

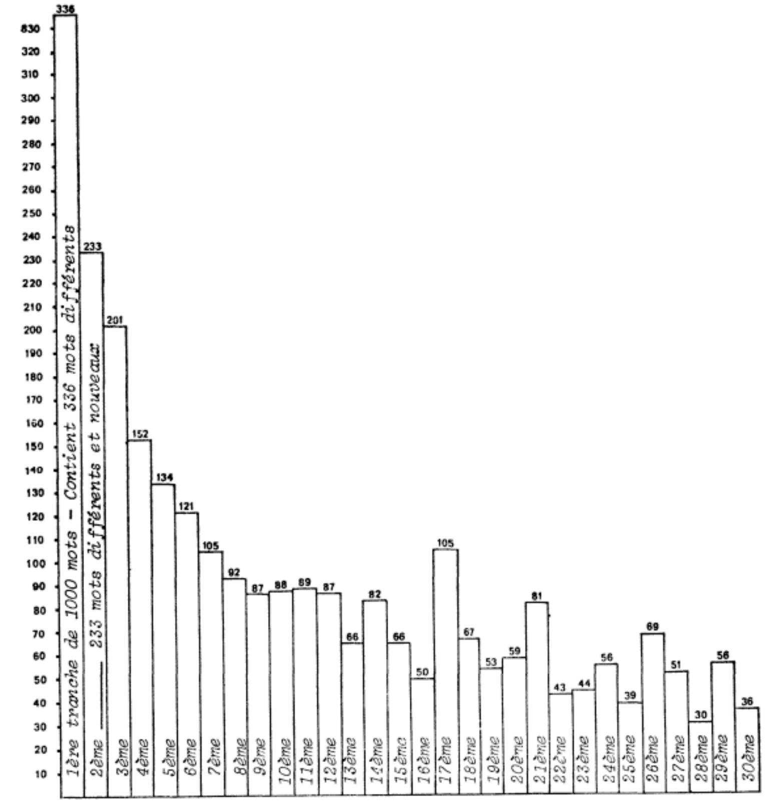
As to why, no one was quite sure. What was evident and subsequently intriguing was that a distributional law held consistent for token items where there was no clear necessity for them to behave in such a way. The birth of a short but powerful wave of discovery, concerned with these mysterious token frequencies, had just begun.

The year 1896 invites two key players into the scene of our history: a Scotsman and an Italian. The first is a man very much interested in applying statistical techniques to social problems. He was a student (and for a short time an employee) of Karl Pearson, and his closest friends were interested in statistical problems as well. It is very likely that this Scotsman knew of English social and genetic statistician Sir Francis Galton, who for some time had been mathematically justifying his cousin Charles Darwin's ideas. This Scotsman is George Udny Yule, who in 1896 chooses to make his first entrance into the statistical arena with an analysis of pauperism and social-economic relief.

The second man, though, is not at this time entering into the world of intellects. He is shaking it. Vilfredo Pareto's inverse power law of economic wealth, showing how 20% of the population earned 80% of a nation's worth, was shocking sociologists, provoking criticism from statisticians, and fueling a new field called *microeconomics*. At the heart of this work is the determination of the conditions for economic *equilibrium*, as one American economist just one year after its appearence aptly noted. Pareto's law carries with it a conceptual richness that Newcomb's law could not provide: namely that power laws can and perhaps should be associated with competing forces and their eventual balance.

Evidence that this work quickly reached the Continent comes from the excitement of a doctoral student at the University of Wisconsin-Madison, who while there would publish a paper that would introduce a curve, his own name attached, for describing income inequalities. From 1905 onward, Max Otto Lorenz can be seen advancing the economic power law that rocked Europe.

We will leave Mr. Yule for a moment, though he does, quite prominently, have a place in this tale. Let us take ourselves now to Paris, France, specifically to the *Institut Stenographique*. It is 1916 and there a gentleman named Jean-Baptiste Estoup is employed as its General Secretary. He writes a paper that year titled *Les Gammes Stenographiques*, and another one year later (*Les mots usuels. Leur nombre et leur fréquence*), both of which gain some if little traction. Both carefully treat the problem of word frequency in a realm that intensely requires this knowledge: for the compression of words into a short-hand. The practice of stenography dates back to the 4th century BC in Ancient Greece, and had been a necessary aspect of Roman, imperial Chinese, and Medieval English culture, and even for Sir Isaac Newton himself. It was not until Mr. Estoup that we find this visual approach to the problem:



Estoup himself describes the phenomenon statistically but not mathematically:

> *Ce graphique montre clairement que le pourcentage des mots différents et nouveaux trovés dand chaque tranche décroit très rapidement au fur et à mesure que l'on avance dans le décompte.*
>
> This visual clearly shows the percentage of different words found in each very rapidly shrinking count as one advances along the slices.
>
> *Dans le* 2^e *mille la proportion des mots nouveaux est de 23%, dans le* 10^e *mille elle est d'environ 9%, dans le* 20^e *mille de 6%, dans le* 30^e *mille de 4%.*
>
> In the second thousand (second slice) the proportion of different words is 23%, in the tenth thousand it is around 9%, in the twentieth thousand 6%, in the thirtieth thousand 4%.

Here are the preconditions, the data-outline, for a power-law describing word frequency in language. The data is present, the description clear, but importantly the latter lacks the mathematical rigour of Newcomb's or Pareto's. To most of the academic world, except for mention in the annual bibliographic record of the Royal Statistical Society and in the footnotes of several statistical papers of the decade, this work has been largely forgotten.

In the same year that Estoup sits down to write *Les Gammes Stenographiques*, an American boy of age 14 is preparing to enter the local High School of a remote town in northern Illinois. After four years, by an excellent record and most likely a good deal of luck, he finds himself traveling to Harvard University in Cambridge, Massachusetts.

In 1924, this young man makes his first travels abroad—crucially to the University of Berlin, Germany, the very school where Ernst Förstemann was trained more than half a century before. Back on the Continent, in 1929, the following dissertation appears out of Harvard University:

Relative Frequency as a Determinant of Phonetic Change

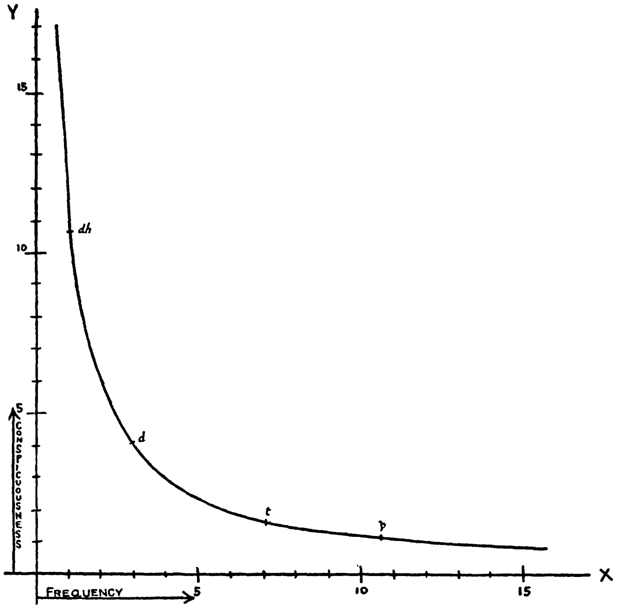
This title, in the context of this history, does not seem so new, and thus why it merits special mention here could be questionable. We have encountered *Relative* notions from De Morgan, we have encountered *Frequency* from Newcomb, Benford, Pareto, and Estoup, and the statistical treatment of *Phonetic*s from Förstemann. The notion we have not encountered, which perhaps is the most crucial notion in our history, is *Determinism.* One of the author's first sentences is such:

> As time passed, however, and collateral facts began to accumulate rapidly, the firm outlines of a powerful law of language became ever more manifest, until, after having presented my data to others better qualified in physics, psychology, and biology, I was encouraged to take up the whole question of mutation in language itself.

Two points are particularly interesting here: first, that in the conception of a law for language, other disciplines of thought were consulted, mainly the most law-driven disciplines of the day. Second, this bold author tells us that he won't stop at his "powerful law of language" but will take up broader, causal processes (*mutation*, evolutionary, or historical change). A thesis appears on the fourth page:

> The foregoing theory of the inter-dependence of form and frequency of usage I shall now try to demonstrate a posteriori with the thesis:
>
> *Principle of Relative Frequency.* The accent, or degree of conspicuousness, of any word, syllable, or sound, is inversely proportionate to the relative frequency of that word, syllable, sound, among its fellow words, syllables, sounds, in the stream of spoken language.
> As usage becomes more frequent, form becomes more accented, or more easily pronounceable, and vice versa.

After tabular data presented from Latin, Vedic Sanskrit, and counts of other languages the philologist of the day would recognize, the following abstract geometric representation appears. There is no statistical approximation from the data involved in this curve, in contrast to Estoup's completely naturalistic representation of the trend. This here is a *theoretical ideal*, described by the author not as a logarithm (as Newcomb would have it), not as a state of equilibrium (as Pareto would have it), nor as a trend (as Estoup might have it), but here as a completely geometric entity: a hyperbola. (p.89) The author deems no mathematical derivation necessary.



The dissertation we have before us gives a radical proposition, linguistically and even philosophically: frequency, just as abstract to us now as it must have been then, is a *deterministic* cause for certain processes, here evolutionary processes of language. In suggesting this alone, with wide-spread language data (much more than Estoup could offer), this marks a definitive change in conceptualizing language: statistics could be an object for asking philosophical questions, not just descriptive questions, about language. This, at the age of 27, is Mr. George Kingsley Zipf.

There is one very small detail of Zipf's 1929 thesis that I would like to mention because I fear it will go unnoticed. On the second page, there exists a long footnote where a list of individuals, 11 to be exact, are thanked for their help in acquiring the cross-linguistic data that the thesis requires. The very last is a name we should recognize: *Mr. J.B. Estoup, Paris.* Nowhere else is Estoup mentioned. The irony lies in that Estoup's recognition of a power relation was essentially equal to that of Zipf's, and for some years after the dissertation appeared, the law was titled *Estoup-Zipf.* In every modern reference, Estoup's name has vanished and only Zipf's remains.

After the release of his dissertation, Zipf, now assistant professor of German linguistics at Harvard, published papers growing more and more in creativity until his premature death in 1950.[4] Of note is his *Psycho-Biology of Language*, which he wrote in 1935 after a period of years verifying his *Principle of Relative Frequency* with more varied language data. Here he declares, perhaps overly confident:

> Much as a physicist might investigate the forces of the mind by viewing linguistic phenomena in the stream of speech as manifestations of the forces of the mind in the process of functioning.

The attempt here is to collect frequency distributions describing not only words but phonemes, word accent, sentence structure, and what is called 'the stream of speech'. A hopeful Zipf predicts to access *meaning* and *emotional intensity* via these results.

In 1941 comes *National Unity and Disunity: the Nation as a Bio-Social Organism.* Of interest to an older Zipf is the individual's social freedom within a community. The "saturation-point" of a community is defined as the asymptote to a series distribution of the population. More controversial is the mathematization of personal liberty:

> A function of one's security which in turn is equivalent to the degree of one's regimentation in "organization":
>
> $$L \times S^q = Constant,$$
>
> where L = personal liberty, S = strength-security

Then, a year before his death, he publishes a sort of magnum opus, a huge synthesis of the themes he has treated over the years—language, behavior, and economy: *Human Behavior and the Principle of Least Effort.* This work became immensely influential not for the content contained in any one section (for data could be found in many of his earlier works), but for its ability to convince its readers of the deep-philosophic nature of the inverse power-law distribution by showing its appearance in so many areas of human behavior: for various

[4] He was only 48 years old.

subsections of language, for economies, and for human populations. The conclusion: There is nothing inherently special about the power-laws that result from these human domains. The reason they appear in so many humanly driven phenomena is because of something *humans* are doing. According to Zipf, this underlying cause of the power law distribution—which he had so many times used as a cause in and of itself for different processes—is a human *conservation of effort.*

Zipf at the end of his career is telling us that his law is more about humans than language itself.

Scotsman George Udny Yule, who we had left in the background, now re-enters our story. From his 1896 entrance into the world of statistics up until 1944, he has made himself a major player: following Sir Francis Galton and Karl Pearson, he contributes to a mathematics of the Theory of Evolution (1925)[5], he philosophizes on the problems in correlation analysis, he produces work on periodicity in 'disturbed series'—the same year that Zipf's dissertation emerges, and he writes a textbook of modern statistics that is lucid and still useful today. By 1944, though, he has caught wave of a topic that likely would have ignited his younger interests in social statistics. By George Udny Yule, the following appears: *Statistical Study of Literary Vocabulary.* This is an attempt to give statistics of author style, and it involves a power-law description of language.[6]

Only one year later, a review of this book appears from Zipf himself:

> ... it suffers to no small extent from his admitted ignorance (p.32) of the investigations of others whose statistical studies and theoretical analyses of the same or related problems in many different languages had already taken on sizable proportions during the 1930s.

Fortunately, the review is not completely caustic:

> ... in this book ones sees just how the problems appear to one of the great present-day masters of statistical method.

After such a critique, one immediately wonders whether George Udny Yule had the courage to contribute anything else to linguistic science. This author notes

[5] Notably, this work involved a power-law distribution.

[6] That same year, J. Richard Reid, an American, publishes *A French Word-Frequency Distribution Curve*, in the purely linguistic journal *Language*. This marks the point where statistical treatments of language find a place in general linguistics. It also should be noted that Mathematician B. B. Mandelbrot, the father of fractal geometry and a key player in economic theory, began contributing to the statistical linguistics literature in 1951, a year after Zipf's death. His work was placed under the umbrella of the new field,"Information Theory", though it crossed many boundaries. The celebrated father of information theory, Claude Shannon, was centered at the American Bell Systems Laboratories.

that one statistical linguistic work by Yule can be found to have been published after this episode, in 1946, on the statistical prediction of errors for the copying of manuscripts. Notably this appeared in the *Journal of the Royal Statistical Society*, safely within the confines of the statistics community.

Twenty years later, another statistician, similar to Yule in his gravitation towards social problems, his primary training in statistical mathematics, and parallel interests in genetic biology, steps back from the progression we find ourselves in. He admits that statistics have been *useful* to language study. But he continues with the more important question:

How are statistics *meaningful* to language study?

Gustav Herdan's answer brings statistical linguistics back to a deeply philosophic root: to Ferdinand Saussure's 1922 distinction—for reasons this author knows not, untouched by Zipf and his predecessors—between the individual and the speaking collective. For Herdan, not only must there be made a distinction between the individual language and Language, capital *L*, but also a distinction between the individual speech utterance the language sample. This conceptual tie to purely linguistic-philosophic conceptions of language marks the definitive point of entry for statistical linguistics into the broader milieu of language study.

In the present day, we see that statistical, mathematical perspectives on language are making places for themselves at the institutional level. The journals following give evidence as to the historical progression of this phenomenon:

The historical emergence of statistical-mathematical language studies:

Title	*Founding*
Mathématiques et sciences humaines	1962
American Journal of Computational Linguistics	1974
Forum der Gesellschaft für Linguistische Datenverarbeitung	1983
AJCL → renamed, Computational Linguistics	1984
Machine Translation	1989
Journal of Quantitative Linguistics	1994
International Journal of Computational Linguistics and Chinese Language Processing	1996
Journal of Logic, Language, and Information	1992
LDV → renamed, Gesellschaft für Sprachtechnologie und Computerlinguistik	2008
International Journal of Computational Linguistics and Applications	2010

It has taken us a long time since August 18th, 1851 on Camden Street. Based on the progression we can witness above and niches that have yet to be filled, this historical tale will likely project into the future for quite some time.

Bibliography

Augustus De Morgan

1851	*To Rev. W. Heald.* Memoir of Augustus De Morgan, Correspondence, Longmans, Green, and Co.: London	London

Ernst Förstemann

1852	*Numerische lautverhaältnisse im, Griechischen, Lateinischen un Deutschen* Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen. 1. 163-179	Berlin
1853	*Numerische lautbeziehungen des griech., latein. und deutschen zum sanskrit* Zt. f. Vergl. Sprachforschung. 2. 35-44	
1853	*Numerische lautverhältnisse in griechischen dialecten* Zt. f. Vergl. Sprachforschung. 2. 401-414	

Simon Newcomb

1881	*Note on the frequency of use of different digits in natural numbers* American Journal of Mathematics. 4. 39-40	Baltimore

Frank Benford

1938	*The law of anomalous numbers* Proceedings of the American Philosophical Society. 78.4 551-572	Philadelphia

George Udny Yule

1896	*On the Correlation of Total Pauperism with Proportion of Out-Relief* The Economic Journal 6.24 613-623	St. Andrews

Vilfredo Federico Damaso Pareto

1896	*Cours d'economie politique*	Lausanne

Jean-Baptiste Estoup

1916	*Les Gammes Stenographiques* Gauthier-Villars, Institut Stenographique de France	Paris
1917	*Les mots usuels. Leur nombre et leur fréquence* Journal de la Société de Statistique de Paris	Paris

Godfrey Dewey

1923	*Relative frequencies of English speech sounds* Harvard University Press	Boston

George Kingsley Zipf

1929	*Relative Frequency as a Determinant of Phonetic Change* Harvard Studies in Classical in Philology	Cambridge
1932	*Selected Studies and the Principle of Relative Frequency of Language* Harvard University Press	Cambridge
1935	*Psycho-Biology of Language* The Riverside Press: Houghton Mifflin Company	Cambridge
1941	*National Unity and Disunity: the Nation as a Bio-Social Organism* The Principia Press Inc.	Bloomington, Indiana
1945	*Review of: The Statistical Study of Literary Vocabulary, by G. Udny Yule* American Literature 17.3 286-287	??
1949	*Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* Addison-Wesley Press Inc.	Cambridge

George Udny Yule

1944	*The Statistical Study of Literary Vocabulary* Cambridge University Press	Cambridge
1946	*Cumulative sampling: a speculation as to what happens in copying manuscripts* Journal of the Royal Statistical Society 109 44-50	St. Luke's

Benoit B. Mandelbrot

1951	*Adaptation d'un Message a la ligne de transmission. I & II* Comptes Rendus 232	Paris
1953	*Contribution a la Theorie Mathematique des Jeux de communication* Institut de Statistiques de l'Universit de Paris	Paris
1953	*An informational theory of the statistical structure of languages* in Communication Theory, PUBLISHER, p.486	Illinois
1953/1954	*Simple games of strategy occurring in communication through natural language* Transactions of IRE 3 124-137	
1965	*Information Theory and psycholinguistics,* Scientific Psychology: Principles and Approaches p.550 Basic Books Inc.	New York

Claude Shannon

1948	*A Mathematical Theory of Communication* Bell System Technical Journal 27 379-423	Massachusetts

J. Richards Reid.

1944	*A French word-Frequency Distribution Curve* Language 20.4 231-237	Washington

Gustav Herdan.

1966	*Language as Choice and Chance* P. Noordhoff Inc.	Groningen

Ferdinand de Saussure

1922	*Cours de linguistique generale* Paris 2nd edition	Paris