Genetics and Power Laws

Lev Vygotsky was a pioneer in the genetic approach to psychology [5, 6]. The genetic approach prioritizes the history of human beings for understanding their present behavior. For a humanities scholar as Vygotsky originally was - and especially a Marxist scholar, a historical approach could be taken for granted. But Vygotsky did much more than apply historical analysis in a humanities-influenced way. His genetic approach was a new kind of historical analysis spanning an enormous range of scales:

1. Phylogenesis: biological evolution of the human species
2. Social history: the transferred knowledge of humanity and particular cultures
3. Ontogenesis: the psychological development of an individual person
4. Micro-genesis: creation of short-term behaviors

Vygotsky emphasized the *explanatory* value of a genetic perspective, that it elucidates how and why an organism behaves the way it does. He contrasts the genetic approach with classical psychology which emphasizes characteristics of the organism that can be observed now. In Vygotsky's words, the genetic approach studies the *process* by which the organism is formed, rather than the *product* of that process.

By way of example, Vygotsky mentioned biological genetics (phylogenetics) – still relatively young at that time – and contrasted it with descriptive biology – description and classification of organisms from their observable features. Both biological methods were popular at that time, but there was no parallel "genetic method" in psychology. Vygotsky was arguing that the genetic approach should be just as valuable in psychology. Since Vygotsky's time there has been a revolution in biology, with genetics assuming a dominant position in essentially every corner of the field. The value of genetics is now understood for both phenotypic analysis – understanding characteristics of the human species – and for the individual organism. By contrast, the genetic approach has not gained anywhere near as much ground in psychology. Developmental psychology and human learning science are (necessarily) genetic forms of inquiry, but Vygotsky's point was that the genetic perspective should be foundational to *any* branch of psychology. Given how much people develop cognitively and socially from birth, it's rather remarkable that this perspective isn't followed universally. Even more so when one recognizes that similar ideas have recurred in the work of others. In particular, Piaget's book "Genetic Epistomology" discusses the social history of scientific knowledge, and complements his major life's work on ontogenesis. But for the most part, cognitive and social psychology today consider the human-social complex as it exists now ignoring its development.

But suppose we want to pursue a genetic analysis of a person or group, how do we start? A genome is an unambiguous blueprint for a particular organism. By contrasting with other organisms, we can learn much about the history of the phenotype as well – from degree of gene mutation, we can recreate an approximate evolutionary tree. We can even go back beyond the birth of our species to our common ancestors with primates, mammals and simpler life forms.

We have no such genomic blueprint for human behavior. But we have better tools than ever to gather and record human experience. We can observe the events which are critical in shaping "psychological evolution" and also their effects on later behavior. Projects such as "MyLifeBits" have been proposed to record a substantial part of an individual's experiences over their lifetime. Furthermore, there is surprising evidence for the genetic character of behavior in “small” observations, such as a single work by a particular author.

Power Laws

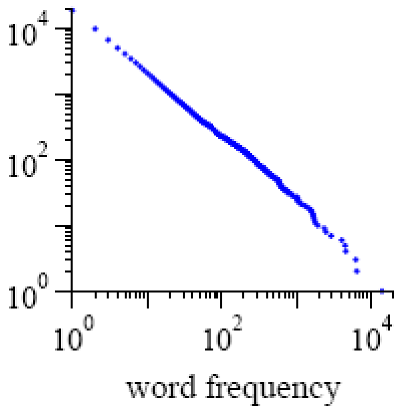
Take a reasonably large corpus of texts in English, such as the works of William Shakespeare or all the Associated Press news stories in a year, or even a single work (James Joyce’ Ulysses for instance). Count the number (frequency) of occurrences of each word in the corpus, and then sort in decreasing order of that number. The position in this order is called the rank of the item. The most common (rank 1) word will almost surely be “the,” the second most common (rank 2) will be “of,” followed by “to”, “and” etc. A remarkable relationship holds between the rank of each item and its frequency in the corpus, which is

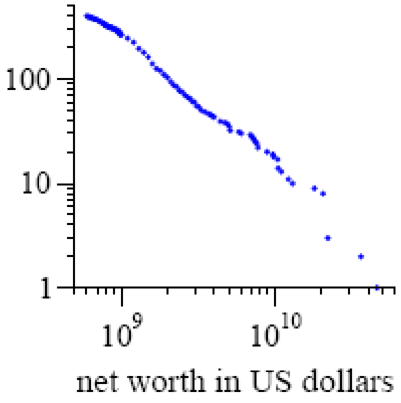
$$f(r) \approx {}^{c}/_{r^p}$$

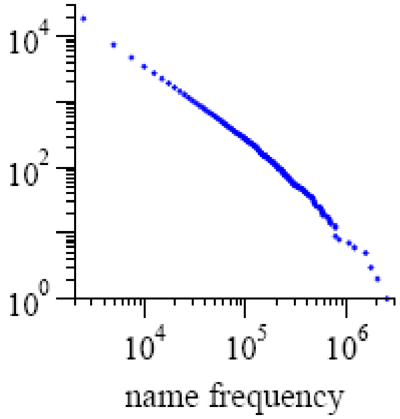
Where r is the rank, $f(r)$ is the frequency of the item of rank r, and c is a constant. The exponent p depends weakly on the corpus, but is usually very close to 1. This is called a *Power Law* because the frequency is proportional to a (negative) power of the rank. This particular power law is called Zipf’s law and was discovered by George Zipf in 1935 [7, 8]. Zipf’s law for texts is already surprising. First that such a simple law exists for the English language, and even more so that it applies to corpora generated in so many different ways, including those by a single author. But that is just the tip of the iceberg. The first Power Law published seems to be Pareto’s law for personal incomes, published in 1896 [4]. Similar power laws apply to corpora of texts in other languages. Power Laws are also ubiquitous in the social sciences, and include [3]:

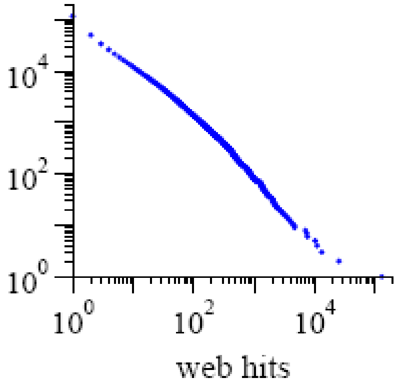
1. The frequency of personal names
2. The populations of cities
3. Number of citations to a scientific paper
4. Number of papers written by an author
5. Sizes of wars
6. Sales of books, music albums, most other items

Some samples of power law plots are given in the figures below.









The label on each plot defines the quantity that has been measured from a large number of examples. The quantity is sorted in descending order on the y-axis, and the x-axis value is the rank of order of the item.

In the last couple of decades, many more relationships have been found by studying the web [1]:

1. The number of links into a web page.
2. The number of pages in a web site.
3. The number of visitors to a web site.
4. Email address book sizes
5. Number of user's Facebook friends
6. The popularity of Facebook apps.

There are many other examples of Power Laws. They occur widely in nature as well (size of craters on the moon, the size of earthquakes,…).

Explanations

Several models have been developed to explain Power Laws, and there are a variety of models although "preferential attachment" models are now the most common explanation. In preferential attachment, a collection of objects is built incrementally, with the objects grouped together somehow (such as people in cities). Each new object is randomly "attached" to a group in proportion to its size. Preferential attachment is a natural model for genetic processes with a small "mutation rate". For instance, consider a new generation being born to residents of a set of cities. Most will grow up in the city they were born in, but a few will leave. Those that leave are more likely to go to a large city than a small one. This process leads to the observed power law for populations of cities. A similar model was used to explain the distribution of words in a corpus (Zipf's law), but very different models have been proposed as well. Zipf's problem was in fact the source of fierce debate over the "right" model to explain it (). We will take a stand on this question: if we accept Vygotsky's "genetic" framework, we are led naturally to look to it for a genetic explanation. Vygotsky was very cautious himself about claiming that similar laws govern his four genetic domains, in fact he thought it to be unlikely. But even if the mechanisms are very different, there are very general notions of "genetic" transmission which are enough to produce Power Law behavior. First, let's look at classical genetic transmission.

Phylogenesis

One of the earliest observations of a Power Law in nature is Yule's law. For this law, $f(r)$ is a count of the number of species in a genus r, where all the genera lie in some higher biological class. The frequency $f(r)$ satisfies a power law with exponent ≥ 1. Yule's law was published in 1925 based on experimental work by the biologist J.C.R. Willis. Yule developed a simple model to explain the development of new species, and showed that it agreed with the observed distributions in nature. Yule's model was based on *a pure birth process*, where new species appear in proportion to the size of the genera they are contained in, and never die. This is a fair approximation to reality, and later research has "filled the gaps" in the model, showing that power laws also occur with weaker assumptions. This helps explain the high incidence of power laws in phylogenetic systems, even when the dynamics of those systems are different.

Yule's system is a preferential attachment system. If a genus contains a large number of species, it is proportionately more likely to generate a new species by mutation (most of the time that mutated species will lie in the same genus, and so the size of that genus grows by one).

Literary Theory

Before we get to scientific models of texts, we will make a quick detour through literary theory. Why do this? Well we are looking to better understand human behavior, and that is a subject considered all across the sciences and humanities. Even if we are interested mostly in scientific explanations at the end, it's good to be pragmatic (and humble) about the limitations of the scientific approach to human behavior, that it its claims must be "repeatable" and statistically significant, and that confounding factors must be eliminated somehow. This makes is very hard to study a lot of interesting phenomena – like everyday human communication or the history of writing – without filtering out the most interesting effects. Realistically, scientists spend a great deal of their careers developing intuitions and hypotheses about the world, and these intuitions may be far from scientific. These are rarely if ever discussed in scientific texts. This doesn't matter too much, because scientific problems by definition (or rather by the nature of the scientific method) are repeatable, context-independent, and usually have a short statement. But if we believe human behavior is by its nature complex and difficult to decompose (and we will give some arguments later why that should be so), then a scientific approach is not ideal.

In the humanities however, authors can write about impressions and their own personal view of reality. Not all the ideas generated are going to be good ones, but instead of shooting them down for lack of laboratory studies, we can look at which ideas have had a significant impact. From a pragmatist perspective, these ideas are "good" by definition. William James would even define them as "true," although that is too strong even for most other pragmatists. We don't have enough space to do more than scratch the surface on literary theory, but that is enough for now. For much of the 20th century, two of the major movements in literary theory were **structuralism**, and **post-structuralism**.

Structuralism emphasized the importance of "structures" in texts. Structuralism dissects texts in a way closely related to **semiotics**, an earlier theoretical movement. Structuralism holds that fundamental structures exist in language that are independent of particular texts, but on the other hand it is necessary to study those texts to understand such structures. It breaks with classical linguistics based on formal grammar and semantics, and the idea that meaning is entirely "in" a text. An example of a structuralist principle "outside" of texts was Ferdinand de Saussure's 1916 notion of *paradigm* – that words or phrases have similar meanings when they are used in similar contexts (the context being the surrounding words in other texts). Furthermore, the meanings of words are influenced by these alternatives. The structuralist perspective is remarkably evocative of "corpus linguistics" in computational text analysis that evolved from the 1980s. Today, many practical problems in text processing are based on corpus methods, and in fact paradigmatic meaning is the dominant notion, although it is not named as such.

Post-structuralism is a closely-related movement to structuralism which is probably easiest to define in terms of its time period and key authors (who began life as structruralists) rather than key ideas. It is often demarcated by Julia Kristeva's concept of **inter-textuality**, that entreats an analyst to find meaning "between" texts. The simplest form of inter-textuality is **allusion** to earlier texts. One can allude to fictional or non-fictional works with a word or phrase: *Romeo*, *Prodigal son, Waterloo, the lady doth protest too much, the play's the thing,…* But Kristeva's point was that other forms of inter-textuality were far more common, in fact omnipresent in the creation of new texts.

To describe inter-textuality, we have to jump back in time to Vygotsky's Russia. One of the leading literary theorists of the 20th century was a(nother) Russian named Mihkail Bakhtin. Bakhtin was a contemporary of Vygotsky, although he apparently had little or no contact with him. Bakhtin developed the idea of polyphony (also translated as multi-vocality) in texts. For Bakhtin, a text expresses many "voices" – those of the author, implicitly that of the reader, those of the characters in the story etc. But aside from these obvious voices, there are many others that follow from the author's background – nationality, culture, education. These voices for Bakhtin do more than shape the reader's experience of the text – they carry its meaning. That is, to fully understand a voice, one has to know something about the speaker's experience – and for collective voices something about the experience of that entire group. Bakhtinian polyphony is therefore a form of historical analysis – but rather than a social history and a history of texts, the social and literary history is distilled into a collection of voices. Namely the voices that animate a particular text.

A "voice" can be recognized by a listener who knows the speaker. If the document is a text (which has no literal voice), the voice can only be a property of the words chosen by the author. Voices can be given a concrete (scientific) interpretation. They can and have been explored using statistical analysis. In the simplest case, a voice can be recognized by the frequency of words in the text. Similar analysis can be used to characterize the other types of voice above: nationality and region, culture, educational background. This doesn't yet give us any reason to expect Power Law behavior, but it weakly suggests that we might see similar statistical behavior in a corpus of texts written by an individual as we would for a collection by a larger social group of authors. The individual's "voice" is a mixture of the voices of larger social groups.

Moving forward from Bakhtin, inter-textuality was a concept introduced by Julia Kristeva in 1966. Kristeva built on Bakhtin's idea of polyphony, and argued that authors do not just assume the voices of other authors – they actually "borrow and transform" from other texts. This is close to the now-familiar idea of authoring by "remixing" earlier works. Authors before the 20th century could safely use excerpts from classical Latin and Greek. In the 20th century, the classics were replaced by mass media and movies "I'll be back," "here's looking at you, kid," and "ET phone home" have displaced Plato and Cicero. And yet Shakespeare has not left the stage, and remains one of the strongest voices in English. Or to be more precise Hamlet, Beatrice, Cordelia and MacBeth are such voices. Politicians make frequent references to earlier “great” speeches like Lincoln, Churchill, Kennedy, and orators like Martin Luther King. The imitation is complex, involving not just choice of words and phrases, but intonation, pauses, gesture, imagery etc.

Intertextuality was further developed by Roland Barthes, who in 1973 (S/Z) claimed there were no original works, and that

> *“A text is... a multidimensional space in which a variety of writings, none of them original, blend and clash. The text is a tissue of quotations... The writer can only imitate a gesture that is always anterior, never original. His only power is to mix writings, to counter the ones with the others, in such a way as never to rest on any one of them”*

Barthes’ was the strongest vision of text as “remix” with the author borrowing from a vast landscape of earlier texts. Barthes also argued for the superiority of academic or “writerly” texts whose interpretation required considerable effort by the reader – in effect the reader becomes a new author of the work, hence its “writely” character. Because of these works, Barthes name has been appropriated by Landow to argue for hypertext as a natural evolution of linear texts. Landow argues that many texts have non-linear narratives, shifting between scenes and back and forth in time, and that hypertext allows a more consistent representation of this non-linearity. Indeed, he suggests that existing texts are really hypertexts in compromise linear form. With fairly natural assumptions, the distribution of public hypertext (i.e. the web) follows a power law. If we accept that texts really are linearized hypertexts, and assuming the “units” (which Barthes called lexia) are not too large, then we would expect to see power laws for word distributions in texts as well. Whether or not we follow this strong form of intertextuality, all of the discussion to this point has argued that authors largely reproduce the texts they have encountered themselves in their own writing. If they did so randomly, we can easily derive power law statistics. Writing is evidently not random, but as long as the process that actually leads to written texts is an unbiased re-use of earlier texts, it will produce power laws. On the other hand, if authors somehow produced new texts “independently” of earlier texts, it’s very difficult to see how Zipf’s law could arise.

Among the derivations of Zipf’s law, Herbert Simon’s account is one of the best known. It owes no debt to post-structuralism (it was written in 1955, and preceded that movement by decades), but the parallels are interesting. Here is a short exposition:

Simon’s Model

Simon’s derivation was based on the analysis of the function $g(i)$ which is the number of words occurring exactly i times in the corpus. This is an alternative to the rank-vs-frequency function $f(r)$ we saw earlier. To relate the two, we introduce a function $h(i)$ which is the number of words occurring i *times or more* in the corpus. Then $h(i) = r$ where r is the rank of the last item whose frequency is at least i. The functions $f(r)$ and $h(i)$ are approximate inverses of each other, and so if

$$f(r) \approx {}^{c}\!/_{r^{\alpha}} \quad \text{then} \quad h(i) \approx {}^{c'}\!/_{i^{1/\alpha}}$$

And to relate back to $g(i)$, we can notice that $g(i) = h(i) - h(i+1)$. This is well-approximated by the negative derivative $-dh/di$, i.e.

$$g(i) \approx {c''}\Big/{i^{\frac{1}{\propto}+1}}$$

And so a power-law form for $f(r)$ implies a power law form for $g(i)$ and vice-versa. The only difference is that the exponent $\propto$ in the rank-frequency form corresponds to an exponent of $\beta = \frac{1}{\propto} + 1$ in the alternate form. So while most texts exhibit a rank-frequency exponent very close to 1, in Simon's form the exponents will be close to 2.

To explain Simon's model, we consider building the corpus up one word and a time. Let k be the number of words that have been added so far. We generalize the function $g(i)$ from before to $g(i,k)$ which is the number of words occurring exactly i times when the corpus has size exactly k.

Simon's model makes two assumptions:

Assumption 1
The probability that the $(k+1)^{st}$ word is a word that has already occurred exactly i times is proportional to

$$ig(i,k)$$

That is, to the total number of occurrences of all words that have appeared exactly i times.

Assumption 2
There is a constant probability that the $(k+1)^{st}$ word will be a new word – a word that has not appeared in the first k words.

The assumptions are simple enough to state, and assumption 2 is quite intuitive. The reader may be concerned that words are always added and never taken away – this is OK for small corpora, but if we are talking about the evolution of the English language, we need to both make room for new words ("Watergate") and the removal of old ones ("forsooth"). This has been done in extensions to Simon's model with little effect on the functional form of the power law.

Assumption 1 needs some justification, and Simon devotes a section of his paper to it. This is the piece of most interest to us.

Justification for Simon's Model

In section IV of his paper, Simon considers two mechanisms for generation of a new text or set of texts:

- Association: sampling from earlier segments of the corpus
- Imitation: "sampling segments of word sequences from other works he has written, from works of other authors, and, of course, from sequences he has heard."

In both cases, Simon is considering idealized random sampling since his model is a probabilistic one. This is clearly not realistic, but on the other hand we are only considering first-order statistical properties. More complicated selection methods will "look" random as long as the sampling they do is unbiased – that is, words appear in the writer's output as frequently as they do in her perception. Under such assumptions, the association mechanism by itself should produce clean power-law statistics.

But association by itself cannot explain the strong statistical similarities between texts by different authors, or indeed between any two texts in English. Simon argues that texts in practice make use of both mechanisms, and that imitation is responsible for the lion's share of the distribution of words in most texts. Simon's imitation mechanism is clearly a form of intertextuality. Simon was arguing that authors "borrow" from other works in a strong enough sense that the word counts in the author's new work are derivable from them.

Furthermore, he argues for "stratified sampling" in the imitative process. Simon did not clarify what he meant by "strata" but from the context it appears to be close to "subject" or "topic." But the selection process is fairly clear, and involves selection first of "related" texts or sections of texts in some (not necessarily random) fashion, and then sampling of words within those texts according to their frequency in those sections of texts.

The apparent vagueness in Simon's definition of strata is probably deliberate. He was interested in the statistical properties of texts, and for his analysis to be plausible, one doesn't need to know what the strata comprise. His archetype seems to be the scientific text, where it is fairly obvious that authors borrow theorems, equations, terminology and ideas from other authors. A "strata" therefore is a well-defined scientific topic. But any other reasonable definition of strata should work, and Simon's use of James Joyce's challenging novel Ulysses must have given him pause to take the idea of "topic" too far.

Borrowing from Simon's Model

Simon's model seems to leave the door open to a wide class of "models" of writing, and to most versions of intertextuality. For instance, we should be free to equate strata with Barthes' lexia, and the authoring process with anything from literal quotation to radical surgery – just so long as the word frequencies are in an aggregate sense preserved. And we can discuss both "readerly" texts whose authors strive for clear communication – such as scientific works – and works of literature which demand much more of the reader to tie the text being read to those that it draws from. In fact Simon's paper uses James Joyce's Ulysses as a test case. Joyce's Ulysses was for its time one of the most challenging works in literature. It was a difficult text for its use of stream-of-consciousness, parody and allusion. James' work is often analyzed by post-structuralists and is sometimes credited as an inspiration for that movement.

Nevertheless, Ulysses demonstrates the same robust power-law statistics as other works, or other corpora by single authors, or the English language as a whole. This doesn't "prove" Simon's explanation, but it is encouraging to see this kind of agreement between a highly-evolved but non-"scientific" literary theory and an empirical model.

The Genetic Perspective

We have found an interesting parallel between the post-stuctural literary perspective on texts and a scientific/empirical one. A version of preferential attachment seems to be at work whereby authors borrow "strata" or "lexia" from earlier texts in proportion to the popularity of those texts. Preferential attachment is often apparent in classical genetics, but is that the only sense in which our model of text generation is "genetic"?

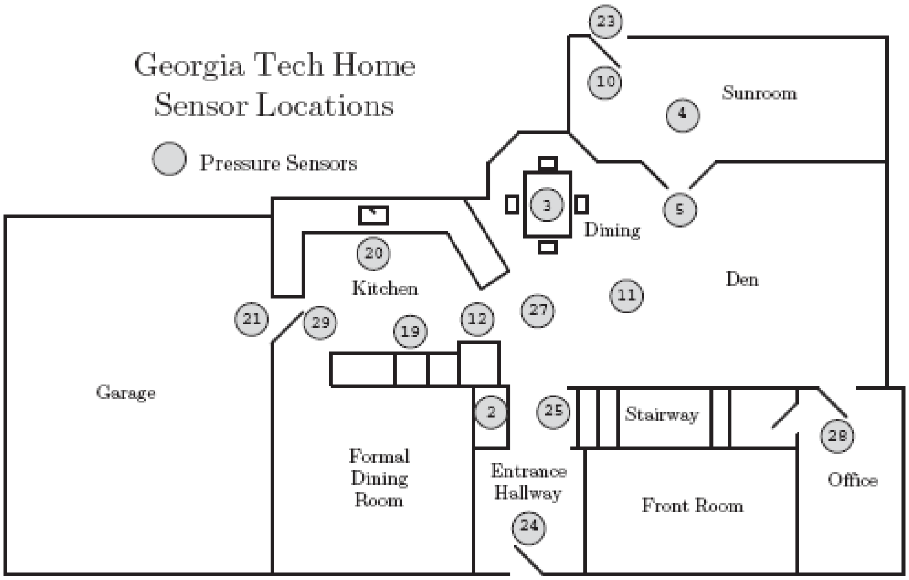
Writing and literature lie on Vygotsky's "social-historical" genetic plane. And for technological cultures, they are the dominant form of knowledge archival and transfer. In onto-genesis, an individual human being develops in response to the world, and in literate cultures human beings develop in response to many texts that they read. In other words the human being develops, and texts help mediate that development.

On the social-historical plane, these roles are reversed. Texts, corpora and lexia are the units being reproduced on the social-historical plane. They are in fact the most naturally "genetic" objects discussed in this paper, since they can be coded as bits like a genome, or could be coded *as* a genome if one really wished to. We have argued in this paper that texts are "reproduced" through writing, that authors borrow liberally from other authors, so that texts acquire their own "lineage" over time. This reproduction mechanism can be complex and we have not explored it in depth. But as long as it is unbiased, i.e. as long as words in written texts occur with similar frequency to those that the author has read, the system of readers/writers will preserve the power law statistics of words.

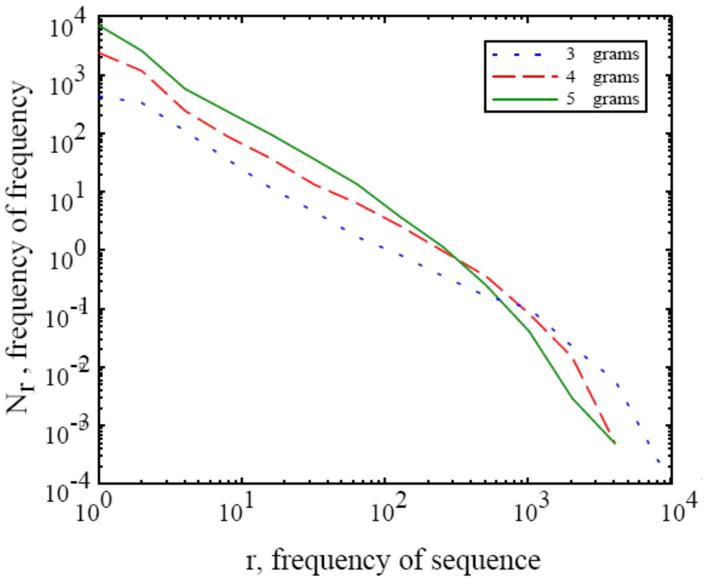
Language as Action

Vygotsky viewed language as a tool, a means of "mediating action." Indeed for him, it was the most important means of mediation for human beings, and the key to higher-level human reasoning. This principle will be very important later in the course. If we take this principle literally, we would expect to see many examples of "language-like" action. Some are obvious, such as musical performance, dance, theater or the skillful execution of a craft or sports play. These performances exhibit varying amounts of spontaneity and innovation by the performer – that depends on the form – but all show the depth of the performer's skill and training. That is, they use highly-evolved "units" of performance that the performer has acquired through practice, performance and perhaps mental rehearsal. The units may be isolated in time (e.g. an arpeggio) or may shape the entire performance (the dramatic arc of a play).

For a less-obvious example, we can look at the movements of residents in a house. A typical house has a relatively small number of "landmarks," e.g. by the stove, the refrigerator, near closets etc. Several "smart home" projects explored the use of proximity sensors for modeling user movement. Such models are useful for automatic light control among other things. The sequence of landmarks is a string very much like the characters in a word. When short sequences of landmarks are analyzed with a rank-frequency plot, a Zipf-like straight line emerges. Figure shows the home and sensor layout we studied in [2].



Each of the 26 sensor locations represents a “symbol”. As users walk about the house, they create a sequence composed of these sensor locations. Subsequences of a discrete sequence of length are called n-grams. So 2-grams are consecutive pairs of symbols, 3-grams are consecutive triples etc. For movement about the house, n-grams represent patterns or “habits” of movement. A rank-frequency plot for 3-, 4- and 5-grams is given below.



From the plot, there is a good degree of linearity (power law behavior) over at least 3 orders of magnitude of rank, for all grams. This particular plot is in Zipf’s original form, with the number of occurrences of the string on the x-axis, and the number of strings with that number of occurrences on the y-axis. The exponent is evidently quite close to 2, which corresponds to an exponent of 1 in rank-frequency form.

This result is quite suggestive. It is consistent with an (individual) onto-genetic development of habits related to movement about the house. This is an extreme generalization of the development of a personal statistical "voice" for a writer.

References

1. Adamic, L. and B. Huberman, *Zipf's law and the internet.* Glottometrics, 2002. **3**: p. 143-150.
2. Aipperspach, R., E. Cohen, and J. Canny. *Modeling Human Behavior from Simple Sensors in the Home*. in *IEEE Conf. on Pervasive Computing*. 2006. Dublin Ireland.
3. Newman, M.E.J., *Power Laws, Pareto distributions and Zipf's law.* Contemporary Physics, 2005. **46**(5): p. 323-351.
4. Pareto, V., *Cours d'Economie Politique*. 1896, Geneva: Droz.
5. Vygotsky, L.S., *Mind in Society*. 1978, Cambridge, MA: Harvard University Press.
6. Wertsch, J., *Vygotsky and the Social Formation of Mind*. 1985, Cambridge, MA: Harvard U. Press.
7. Zipf, G.K., *Human behavior and the principle of least effort*. 1949, Cambridge, MA: Addison-Wesley.
8. Zipf, G.K., *The Psychobiology of Language*. 1935, Boston: Houghton Mifflin.