Scale-Free Networks in Phonological and Orthographic Wordform Lexicons

Christopher T. Kello and Brandon C. Beltz
George Mason University

To appear in I. Chitoran, C. Coupé, E. Marsico, & F. Pellegrino (Eds.), *Approaches to Phonological Complexity* (Mouton de Gruyter)

1. Competing Constraints on Language Use

Languages are constrained by the physical, perceptual, and cognitive properties of human communication systems. For instance, there are upper bounds on the amount of time available for communication. These bounds constrain the lengths of phonological and orthographic codes so that communication can proceed apace. There are also constraints on the amount of linguistic information that can be condensed into a given span of perception or production (Liberman, 1967). These constraints place lower bounds on the amounts of speech activity needed for phonological and orthographic codes.

Constraints on languages often work in opposition to one another, perhaps the most famously proposed example being Zipf's *principle of least effort* (Zipf, 1949). On the one hand, memory constraints produce a tendency towards using fewer numbers of words to reduce memory effort needed to store and access them. A vocabulary that requires minimal memory effort on the part of the speaker is one that uses a single word for all purposes. On the other hand, ambiguity constraints produce a tendency towards using larger numbers of words to reduce the number of meanings per word, and thereby reduce effort needed to disambiguate word meanings. A vocabulary that requires minimal disambiguation effort on the part of the listener is one that uses a different word for every distinct concept. The principle of least effort states that natural languages are constrained to minimize both speakers' and listeners' efforts, and only by balancing them can effective communication be achieved.

It is generally accepted that language usage must strike a balance between these two kinds of effort. However, Zipf controversially claimed that the principle of least effort is responsible for a particular kind of *scaling law* (also known as a power law) that appears to be true of word usage throughout the world. The scaling law states that the probability of using a given word W in language L is approximately inversely proportional to its frequency rank,

$$P(W_L) \approx r^{\alpha},$$

where $\alpha \approx -1$. For instance the highest ranked word in English (THE) is about twice as likely to occur as the second highest ranked word (OF), which is about twice as likely as the fourth highest, and so on.

This scaling law in the distribution of word frequencies means that a few words are used very often and most words are used rarely. This dichotomy creates a combination (balance) of high frequency words requiring little memory effort (because they are general-purpose words used often in many different contexts), and low frequency words requiring little disambiguation effort (because they are specialized words with particular meanings and contexts). The connection between word frequency and word meaning is evident, for instance, in the fact that closed-class words tend to be the most frequent of their language, and also appear in the most general contexts (e.g., the English word THE may be followed by virtually any noun, adjective, or adverb, albeit some words follow more frequently than others). Rare words are often from highly specialized domains and therefore appear in very particular contexts (e.g., terms specific to a given profession).

Zipf's law transparently corresponds to a continuous balance across the frequency range, from minimizing memory effort in the few frequent, context-general words, to minimizing disambiguation effort in the many rare, context-specific words (see Morton, 1969). This balance is present at all measureable scales because the function between word frequency and frequency rank is the same regardless of the scale at which these variables are measured (i.e., the relation is invariant over multiplication by a common factor).

The idea that Zipf's principle of least effort leads to this scaling law makes some intuitive sense, but Zipf never gave a rigorous proof of it. More problematically, other candidate hypotheses came to light that appeared to provide simpler explanations. Mandelbrot (1953), Miller (1957), and Li (1992) each showed that scaling law frequency distributions could be obtained from texts composed of random letter strings. Their proofs have led many researchers to discount such distributions as inevitable and therefore trivial facts of language use.

However, others have pointed out that corpora composed of random strings have important differences with natural language corpora (Tsonis, Schultz, & Tsonis, 1997). For instance, the most frequent random strings are necessarily those of middling length, whereas in natural languages these tend to be the shorter words. Random strings also cannot speak to the relationship between word frequency and word meaning. More generally, random strings do not have the capacity for structure that is requisite of real wordforms. Thus it appears that random strings exhibit scaling laws because string frequency has a particular relationship with string length, but this relationship is not what creates scaling laws in real word frequencies.

1.1. Criticality in Language Use

Spurred by the inadequacies of random string accounts, Ferrer i Cancho and Solé (2003) conducted an information theoretic analysis to investigate Zipf's hypothesized connection between the principle of least effort and scaling law frequency distributions. The authors showed that, under fairly general assumptions, the balance of memory effort and disambiguation effort can be shown to produce a scaling law in the frequency distribution of word usage. Their analysis was motivated by theories of *critical phenomena* that were developed in the area of physics known as *statistical mechanics* (Huang, 1963; Ma, 1976).

The aim of statistical mechanics is to describe the probabilistic, ensemble (global) states of systems with many interacting components. Ferrer i Cancho and Solé (2003) modeled communication systems by treating language users as system components and word usage as the result of component interactions. From this perspective, ensemble states correspond to distributions of word usage, and the authors focused on two kinds of distributions that often constitute opposing *phases* of a system's behavior. One phase is characterized by *high entropy* in that systems may exhibit different behaviors with roughly equal probability (i.e., a flat probability distribution). The other is characterized by *low entropy* in that some behaviors may occur more often than others (i.e., a peaked probability distribution).

In this framework, the high entropy phase corresponds to minimizing disambiguation effort in that many different words are used in order to distinguish among many different meanings (i.e., a relatively flat probability distribution of word usage). The low entropy phase corresponds to minimizing memory effort in that only one or a few words are used for most meanings (i.e., a relatively peaked probability distribution of word usage). As explained earlier, an effective communication system is one that strikes a balance between these two opposing phases.

Theory from statistical mechanics is useful here because it has been shown that, when complex systems transition between phases of low and high entropy, the transition often occurs abruptly rather than gradually (Ma, 1976). In thermodynamic terms, low memory effort and low disambiguation effort may be two opposing phases of the communication system that have a sharp *phase transition* between them. Systems poised near phase transitions are said to be in *critical states*, and critical states are known to universally exhibit scaling laws in their behaviors, including scaling law distributions like Zipf's law (Bak & Paczuski, 1995).

Thus evidence of Zipf's law suggests that communication systems tend to be poised near critical states between phases of low memory effort and low disambiguation effort. To investigate this hypothesis, Ferrer i Cancho and Solé (2003) built a very simple, information theoretic model of a communication system, and they optimized the model according to two opposing objectives: To minimize the entropy of word usage on the one hand (minimize memory effort), while also minimizing the entropy of meanings per word on the other hand (minimize disambiguation effort). These entropies are opposed to one another and the model contained a parameter that governed their proportional influence on communication.

Model results revealed a sharp transition between the phases of low memory effort and low disambiguation effort. Moreover, Zipf's law was obtained when communication was poised near this phase transition. These simulation results provide a theoretically grounded explanation of Zipf's law, but one might question whether the authors have built a bridge too far: why would theories of critical phenomena developed for physical systems apply to systems of human communication? The answer is that systems in critical states exhibit general principles of behavior that hold true regardless of the particular kinds of components that comprise the system, a phenomenon known as *universality* in theoretical physics (Sornette, 2004). Thus interacting atoms or interacting words or interacting people may all share certain principles of emergent behavior in common.

2. Competing Constraints on Wordform Lexicons

If principles of criticality are general to language systems, then scaling laws analogous to Zipf's law should be found in language systems wherever there is a phase transition between low and high entropy. In the present study, we adopt and adapt Ferrer i Cancho and Solé's (2003) information theoretic analysis to investigate an analogously hypothesized phase transition in language systems.

The language domain that we focus on is wordform lexicons. For the sake of simplicity let us represent wordforms as linear strings of phonemes or letters. The appearances of words in speech or text can be coarsely represented as such strings, in which case wordform lexicons consist of all strings that appear as wordforms in a given language (token information about individual appearances is discarded). Language users must know their wordform lexicons to communicate, and thus communication constraints should apply to lexicon structure, just as they apply to word usage (the latter being defined in terms of token information instead of lexicon structure). We investigated two competing constraints on lexicon structure that are analogous to the ambiguity and memorability constraints hypothesized for Zipf's law, namely, *distinctiveness and efficiency constraints*.

On the one hand, the mutual distinctiveness of wordforms in a lexicon should be maximal in order to minimize the chance of confusing them with each other during communication. We consider wordforms as distinctive to the extent that they consist of substrings unique to them. For instance, the English orthographic wordform YACHT is distinctive because substrings like YACH, ACHT, YAC, and CHT are not themselves English wordforms (note that substrings are position-independent). By contrast, the wordform FAIRED is less distinctive because FAIR, AIR, AIRED, IRE, and RED are all wordforms themselves. A maximally distinctive lexicon is one that uses the most unique substrings possible, which minimizes the amount of substring overlap among wordforms.

On the other hand, the efficiency of lexicon structure should be maximal in order to minimize the resources needed to represent them. In terms of substrings, a maximally efficient lexicon is one that uses the fewest substrings necessary to distinguish among all wordforms. This means that substrings are reused across wordforms as much as possible. If one allows homophones or homographs to occur without limit (i.e., using the same wordforms to represent multiple word meanings, as in /mit/→MEAT or MEET for homophones, and WIND→/wɪnd/ or /waɪnd/ for homographs), then a maximally, overly efficient lexicon would use only one string to code all words.

We define these competing constraints in terms of all substrings (i.e., wordforms of all sizes) because there does not appear to be any privileged scale of substring analysis. One can see this in the fact that, collectively speaking, languages of the world use all scales of substrings to express their phonological, orthographic, and morphological structures. In English, for instance, some inflectional

morphemes are expressed as single letters (e.g., -s for pluralization), whereas others conveying whole word meanings are expressed by strings as large as the wordforms themselves. Between these extremes one can find morphological structures expressed as substrings at any given scale, in any given position.

Because distinctiveness and efficiency constraints are defined over all substrings, an analysis of any given language will include substrings that are not linguistically relevant to the wordforms containing them. For instance, the wordform RED does not correspond to a linguistic unit in the wordform FAIRED, yet it is included below in our analysis of an English orthographic wordform lexicon. Conversely, substrings will not capture all possible morphological structures (e.g., infixes in languages like Hebrew). One-to-one correspondence between substrings and linguistic structures is not necessary for our analysis because substrings are not meant to capture *all* the factors that might help to shape a wordform lexicon; this would not be feasible. Substrings are only meant to capture one facet of the hypothesized balancing act between distinctiveness and efficiency, albeit a salient one.

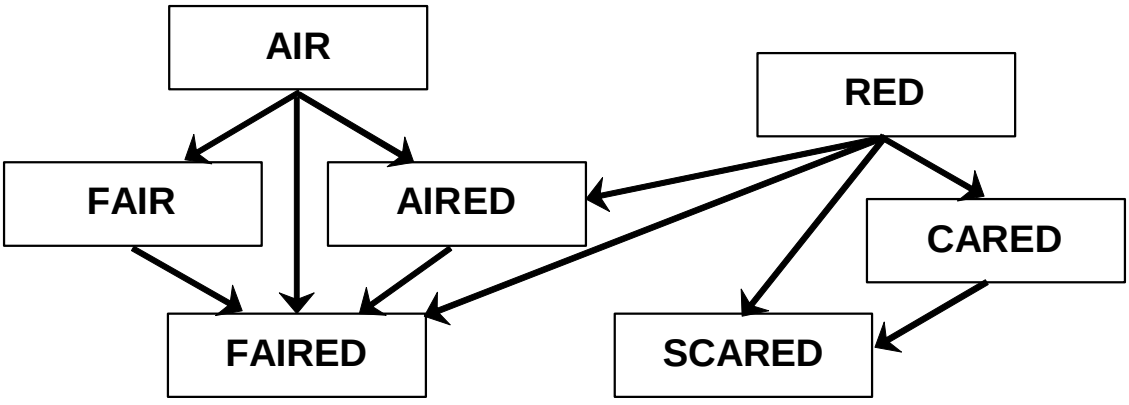
The face validity of our analysis can be seen in the functional importance of balancing distinctiveness and efficiency: If distinctiveness is over-emphasized, then structure will not be sufficiently shared across wordforms. If efficiency is over-emphasized, then structure is not sufficiently heterogeneous across wordforms. Our research question is whether the need to balance these competing constraints poises wordform lexicons near a phase transition between states of low and high entropy. If so, then a scaling law is predicted to occur in the distributions of substrings that comprise wordform lexicons.

2.1. Scale-Free Wordform Networks

We explain how a scaling law is predicted in the next section, but it is helpful to first point out that our prediction corresponds to what is commonly referred to as a *scale-free network*. To illustrate by contrast, note how the word frequency distributions following Zipf's law do not have any explicit connections among the words. This is because only frequency counts are relevant to Zipf's law. Substring frequency distributions are different because substring counts are related to the substring structure of wordform lexicons.

For instance, each substring count for the English wordform RED corresponds to its connection with another English wordform (RED is a substring of FAIRED, REDUCE, PREDICT, and so on). These connections form a structure that can be formalized as network (i.e., directed graph) for which each node is a different wordform, and one node is linked to another whenever the former is a substring of the latter. A small piece of the network created from an English wordform lexicon is diagrammed in Fig 1.

Fig 1. Piece of English Orthographic Wordform Network



The inclusion of all substring relations among wordforms creates a densely interconnected network with a tree-like branching structure from shortest to longest wordforms. The shortest wordforms

serve as the tree trunks; they have no incoming links because no wordforms are contained within them. The longest and most unique wordforms are at the branch tips; they have no outgoing links because they are not substrings of other wordforms. The progression from trunks to tips is highly correlated with wordform length, but not strictly tied to it: Some longer but common substrings are more trunk-like than shorter but unusual substrings (e.g., SING is more root-like than YO).

This wordform network is relevant to our research question because the links are directly related to the distinctiveness and efficiency of the wordform lexicon. In particular, distinctiveness increases as the number of incoming links decreases, and efficiency increases as the number of outgoing links increases. Thus wordform networks serve as tools for conceptualizing and analyzing the hypothesized distinctiveness and efficiency constraints on lexicon structure.

In terms of the network formalism, our predicted scaling law can be found in the counts (i.e., degrees) of outgoing links per node (i.e., the number of times that a given wordform appears as a substring of another wordform in the lexicon). Rather than use the frequency rank distribution as for Zipf's law, network link distributions are often expressed in terms of the *cumulative probability distribution*: The probability of choosing a wordform node at random whose number of outgoing links is $\geq k$ is predicted to be

$$P(\geq k) \approx k^{\gamma},$$

where $\gamma \approx -1$ is typically referred to as a scale-free network. The cumulative probability distribution is a popular means of expressing scale-free networks, in part because exponents can be more directly and reliably estimated from it (see Kirby, 2001).

Casting our predicted scaling law as a scale-free network is also potentially useful because scale-free networks have attracted a great deal of attention in recent years throughout the sciences. Many systems in nature and society can be represented as networks, and it turns out that such networks are often scale-free. For instance, scale-free network structures have been found in computer networks (Barabasi, Albert, & Jeong, 2000; Albert, Hawoong, & Barabasi, 1999), business networks (Wasserman & Faust, 1994), social networks (Barabasi, Jeong, Neda, Ravasz, Schubert, & Vicsek, 2002), and biological networks of various kinds (Jeong, Tombor, Albert, Oltvai, & Barabasi, 2000; Sole, 2001).

In the context of language, Steyvers and Tenenbaum (2005) found that semantic networks of words have scale-free structures when constructed using either behavioral or encyclopedic methods. They built one semantic network from word association data by linking any two word nodes for which one was given as an associate of the other (e.g., a participant might associate the word NURSE with DOCTOR). Two other networks were similarly built using encyclopedic methods, one based on a thesaurus and the other on an on-line encyclopedia. All three methods yielded semantic networks whose link distributions obeyed a scaling law.

Semantic networks have the connotation of spreading activation across the nodes via their links, and many other networks also entail transmission of information or materials among the nodes. However, it is important to clarify that our wordform networks do *not* come with an assumption of spreading activation or information transmission among wordforms. We employ the network formalism only for its structural properties.

2.2. Information Theoretic Analysis

To show how a scale-free wordform network is predicted in the balance of distinctiveness and efficiency constraints, we parallel Ferrer i Cancho and Solé's (2003) information theoretic analysis that showed how Zipf's law can be predicted from Zipf's principle of least effort.

We represent a wordform network as a binary matrix $\mathbf{A} = \{a_{ij}\}$. Each row i represents a wordform w_i, where $1 \leq i \leq n$ and n is the number of words in the lexicon. Each column j also represents a wordform numbered from 1 to n. Each $a_{ij} = 1$ if w_i is a substring of w_j (wordforms are treated as substrings of themselves, i.e., $a_{ij} = 1$ for all $i = j$), and $a_{ij} = 0$ otherwise. The probability that wordform w_i appears as a substring, relative to all other wordforms, is given by (all sums are from 1 to n),

$$P(w_i) = \sum_j a_{ij} \Big/ \sum_k \sum_l a_{kl}.$$

The efficiency of a wordform lexicon is defined in terms of the entropy of the substring probability distribution,

$$H_n(w) = -\sum_i P(w_i) \log_n P(w_i).$$

$H_n(w) = 0$ when a single wordform is used for all words, and $H_n(w) = 1$ when all wordforms appear as substrings of other wordforms equally often (the upper boundary is 1 because the log is base n).

The distinctiveness of a wordform w_i is defined in terms of its diagnosticity, that is, the amount that uncertainty is reduced about the identity of a word W given that it contains w_i. The negative of this amount can be quantified by the entropy over the probability distribution of wordforms conditioned by the presence of w_i,

$$H_n(W|w_i) = -\sum_j P(w_j|w_i) \log_n P(w_j|w_i).$$

$H_n(W|w_i) = 1$ when the presence of w_i provides no information about the identity of W, and $H_n(W|w_i) = 0$ when the presence of w_i assures the identity of W. Each conditional probability is given by

$$P(w_j|w_i) = a_{ij} \Big/ \sum_k a_{ik}.$$

Finally, the overall distinctiveness of a wordform lexicon is defined as the average distinctiveness over wordforms (the average is used to normalize both $H_n(w)$ and $H_n(W|w)$ between 0 and 1),

$$H_n(W|w) = \sum_i H_n(W|w_i) \Big/ n.$$

The balancing of distinctiveness and efficiency now translates into the simultaneous minimization of $H_n(w)$ and $H_n(W|w)$. These constraints are in opposition to each other because $H_n(W|w) = 1$ when $H_n(w) = 0$, i.e., when a single wordform is used for all words. However, when wordforms appear as substrings equally often, $H_n(w) = 1$, there is no guarantee that substrings will be as diagnostic as possible, $H_n(W|w) = 0$. This is true because wordforms may be equally "overused" as substrings. Thus these constraints are not isomorphs of each other. The balance of minimizing $H_n(w)$ versus $H_n(W|w)$ is parameterized by $0 \leq \lambda \leq 1$ in

$$\Omega(\lambda) = \lambda H_n(w) + (1-\lambda) H_n(W|w).$$

In their parallel analysis, Ferrer i Cancho and Solé (2003) created matrices $\mathbf{A}_\lambda$ that minimized $\Omega(\lambda)$ at numerous sampled values of $0 \leq \lambda \leq 1$ (see also Ferrer i Cancho, 2006). They showed that at $\lambda \approx 0.4$, a sharp transition existed in the values of their entropic measures that were analogous to $H_n(w)$ and $H_n(W|w)$. Moreover, they found that the frequency of word usage was distributed according to Zipf's law at the transition point. Thus it appears that this point is a phase transition exhibiting a scaling law.

Our analysis parallels Ferrer i Cancho and Solé's (2003) in order to make the same kind of scaling law prediction, but in terms of substring structure in a wordform lexicon, rather than word usage in communication. Thus our analysis predicts a scaling law in the distribution of outgoing links across wordform nodes, that is, it predicts a scale-free network. This scale-free network is predicted at the transition point between phases of lexicon distinctiveness versus lexicon efficiency.

3. Empirical Evidence for Scale-Free Wordform Networks

Our predicted scaling law is relatively straightforward to test. It simply requires the creation of wordform networks from real languages, and the examination of their structure for a scaling law in their link distributions. We begin with networks created from phonological and orthographic wordforms in English, and we then report the same analyses for four other languages.

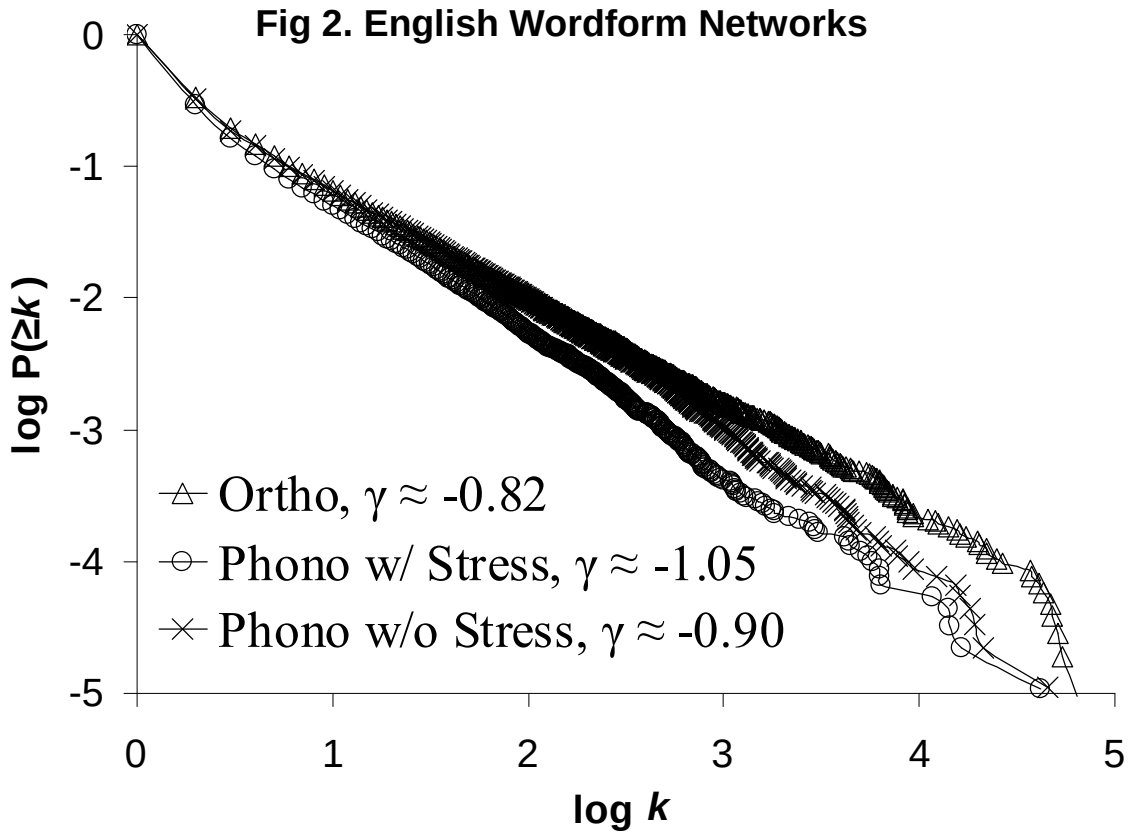
3.1. English Wordform Networks

A total of 104,347 printed words and 91,606 phonetically transcribed words were drawn from the intersection of the Carnegie Mellon University pronunciation dictionary and the Wall Street Journal corpus. The letter strings comprised an orthographic wordform lexicon, and the phoneme strings were used to create two different phonological wordform lexicons, one with lexical stress markings on the vowels (primary, secondary, and tertiary) and one without stress markings. The frequency of wordform usage was not part of the wordform lexicons.

A wordform network was created for each of the three lexicons. Each node in each network corresponded to an individual wordform, and within each network one node was linked to another if the former wordform was a substring of the latter. For the stress-marked lexicon, one wordform was a substring of another only if both the phonemes and stress markings of the former were contained in the latter. Each node i of a network had k_i outgoing links, where $1 \leq k_i \leq n$ and n is the total number of wordforms in the corresponding lexicon. As mentioned earlier, the predicted scaling law is usefully expressed in terms of the cumulative probability distribution, which is linear under a logarithmic transform (the intercept is zero),

$$\log P(\geq k) \approx \gamma \log k.$$

This expression facilitates visualization and analysis of the data.



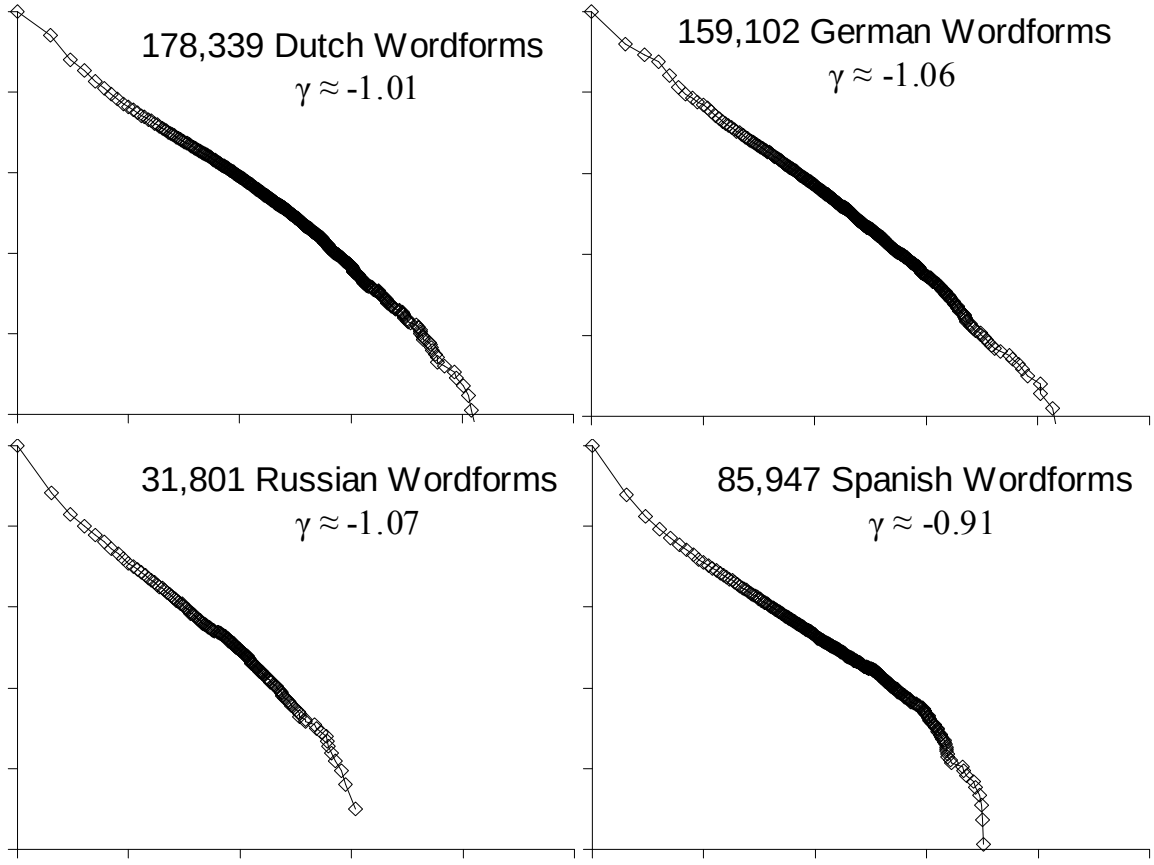
Cumulative probability distributions for the three wordform networks are plotted on a log-log scale in Fig 2. Clear evidence for a scaling law can be seen in the negative linear relation between log $P(\geq k)$ and log k. The exponent of the scaling relation for each distribution was estimated by the slope of a linear regression line fit to the data between $1 \leq \log k \leq 3$. In theory, scaling laws range over all scales (i.e., the entire distribution), but empirical observations rarely if ever achieve this ideal because of limited amounts of data and other practical limitations. These limitations typically show up in cumulative probability distributions as deviations in the tails from the scaling relation. These deviations are slight for the wordform networks plotted in Fig 1, but to avoid them exponents were estimated from the middle of the distribution.

The estimated exponents are close to the canonical value of –1 for scale-free networks. The exponent estimate for the orthographic wordform network is slightly more negative than the others, indicating that it is slightly more densely interconnected (and likewise for the phonological network without stress versus with stress). These differences in density are under investigation but they may be partly due to the differences in morphological transparency among the lexicons: In English, orthographic wordforms represent morphological structure more directly, e.g., SIGN is a substring of SIGNATURE in the orthographic network but not the phonological networks. Differences aside, the results generally confirm the predicted scaling law.

3.2. Wordform Networks in Other Languages

The same wordform network analyses were also conducted on orthographic wordform lexicons for Dutch, German, Russian, and Spanish. These particular lexicons were chosen only because they were readily analyzable and downloadable at ftp://ftp.ox.ac.uk/pub/wordlists. These languages represent a sample of the Indo-European language family. In terms of their morphological structure, they are mostly characterized as synthetic languages (i.e., high morpheme-to-word ratios). Comparing these

Fig 3. Other Wordform Networks



languages to English, which is more of an isolating language (i.e., low morpheme-to-word ratio) provides an initial gauge of the degree to which language type influences the results of our analyses.

The cumulative probability distribution for each wordform network is plotted in Fig 3 with the lexicon size and the estimated scaling exponent (the axes are the same as in Fig 2). All four languages show evidence of a scaling relation in the center of their link distributions with estimated exponents near –1. Estimates varied slightly across languages, as did the amount of deviation in the tails of the distributions.

The wordform statistics of the orthographic lexicons are reported in Table 1. *N* is the number of wordforms analyzed, *M* and *SD* are the mean and standard deviation of wordform lengths respectively, and γ is the estimated scaling exponent of the link distributions. Evidence for the isolating quality of English morphology is reflected in its shorter mean wordform length compared with the other languages (fewer and smaller morpheme combinations), which are more synthetic by comparison. The slightly less negative scaling exponent for English may be due to its isolating quality, but this possibility requires further investigation. For our current purposes, it is sufficient that all the languages exhibit a scaling law as predicted.

Table 1. Summary Statistics for Orthographic Lexicons

	N	*M*	*SD*	γ
English	104,347	7.3	2.3	-0.82
Dutch	178,339	10.2	3.0	-1.01
German	159,102	11.9	3.5	-1.06
Russian	31,801	8.1	2.4	-1.07
Spanish	85,947	8.9	2.5	-0.91

3.3. Ruling Out an Artifactual Explanation

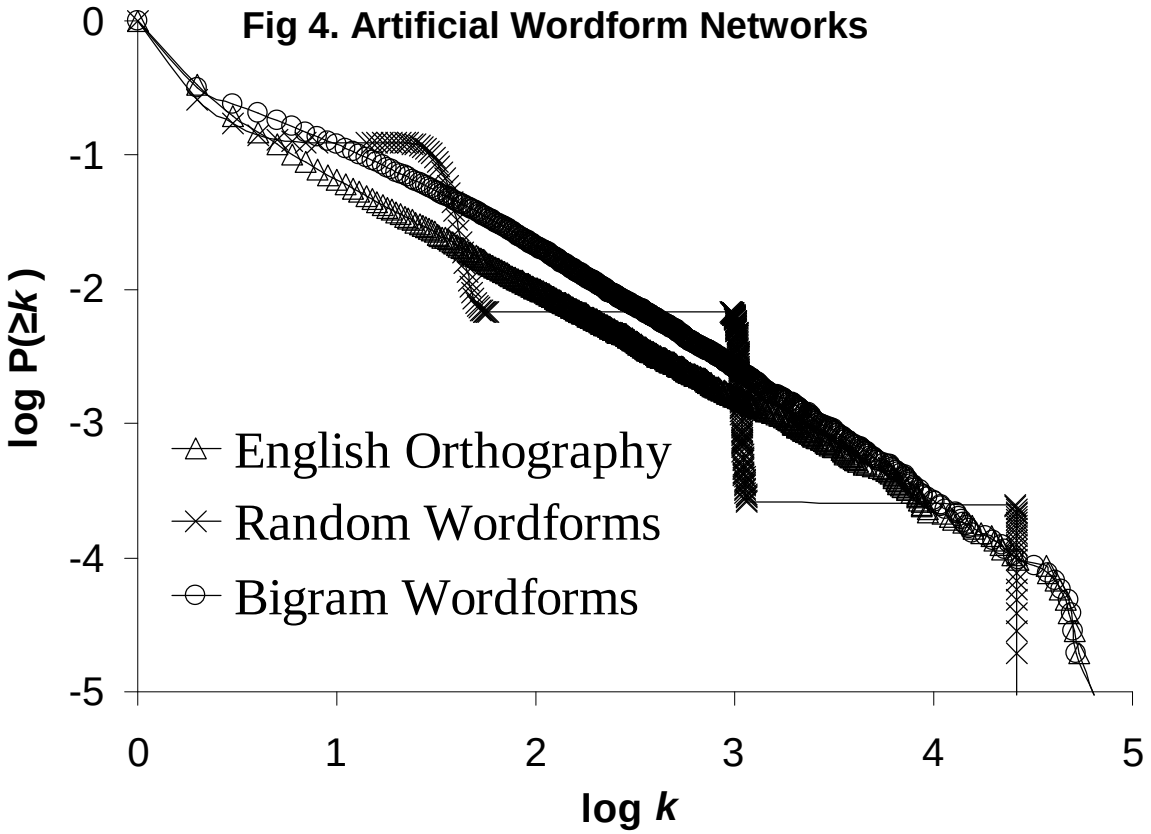
All together, our network analyses appear to provide considerable evidence for the scaling law predicted to occur in the balance of distinctiveness and efficiency constraints on the structure of wordform lexicons. But before coming to this conclusion, we must first determine whether these results may be an inevitable and therefore trivial property of wordform networks created from substring relations. In particular, it may be that lexicons composed of variable-length random letter strings also produce the predicted scaling law. This may seem possible because, even for random letter strings, shorter wordforms will tend to have more outgoing links compared with longer wordforms, and the longest wordforms will have no outgoing links. Thus variations in wordform length alone may be sufficient to create the predicted scaling law.

We tested this artifactual explanation by creating a wordform lexicon comprised of random letter strings using essentially the same method as used by Mandelbrot (1953), Miller (1957), and Li (1992). Each wordform was incrementally built up by repeatedly adding a letter with probability $p = 0.82$, or completing the wordform with probability $1–p = 0.18$. Each letter was chosen at random with equal probability, and the completion probability was chosen so that the average wordform length would be the same as that for our corpus of English orthographic wordforms. A total of 104,347 random wordforms were created, which is the size of our English orthographic wordform lexicon.

The cumulative probability distribution for the random wordform network is plotted in Fig 4. The graph shows that the distribution does not at all resemble the scaling relation observed for the English orthographic wordform network, whose distribution is also plotted for purposes of comparison. Instead of a scaling relation, the random wordforms yielded a tiered distribution that is indicative of characteristic numbers of outgoing links per node. For instance, the majority of nodes had only one or a few outgoing links, but a second large group of nodes had 30-35 links. Hardly any nodes had between 6 and 18 links. Five other random lexicons were generated and each one resulted in a similarly tiered distribution.

The failure of random wordform lexicons to yield a scaling relation shows that our results with real lexicons were not an artifact of length variability in wordform lexicons. It therefore appears that

the observed scaling relations reflect a property of the structural relations among wordforms in natural languages. To provide further support for this conclusion, we tried to recreate the scaling relation by creating an artificial wordform lexicon using the bigram frequencies of English orthography. Wordforms were again built up incrementally, except that the probability of each letter being chosen was conditioned on the previous letter, and the conditional probabilities were estimated from the Wall Street Journal corpus. So for instance, if the letter Q happened to be chosen as the first letter of a given wordform, there was a 97% chance that the second letter would be U. This method created a wordform lexicon that mimicked the statistical properties of English wordforms.



The cumulative probability distribution for the bigram wordform network is also plotted in Fig 4. This distribution is much closer to the predicted scaling law in that the tiers are gone and the slope of the overall descent is near -1. However, there is a "bump" over most of the center of the distribution that deviates from the nearly perfect linear relation of the English wordform network. This result indicates that the statistical structure of English wordforms did, in fact, play a role in the observed scaling relation. However it also suggests that not all relevant aspects of wordform structure are captured by bigram frequencies because the scaling relation was not entirely recovered. Work is underway to determine whether more of the scaling relation can be recovered with artificial lexicons that more closely mimic the statistical structure of English.

4. Conclusions

In this chapter, theories of criticality were used to predict a heretofore unexamined scaling law in the structure of phonological and orthographic wordform lexicons. Evidence for the predicted scaling law was found in the wordforms of five different languages, and analyses of artificial lexicons showed that the scaling law is not artifactual. The law is hypothesized to emerge from the balance of two competing constraints on the evolution of wordform lexicons: Lexicons must be as distinctive as possible by minimizing substring overlap among wordforms, while also being as efficient as possible by reusing substrings as much as possible. A phase transition is hypothesized at the balance of these high

and low entropy phases, respectively. Empirical and theoretical work on critical phenomena predicts a scaling law distribution near the hypothesized phase transition.

The predicted scaling law distribution was expressed in terms of scale-free networks in which wordforms were connected whenever one was a substring of another. In general, some of these substring links reflect the linguistic structure that underlies wordforms. For instance, root morphemes like FORM will often be substrings of their inflected and derived forms like FORMED and FORMATION, respectively. Also, monosyllabic wordforms like /fit/ are substrings of multisyllabic wordforms like /dɪˋfit/. However, substring relations do not always respect linguistic structure, and not all linguistic structure is reflected in substring relations. For instance, LAND is a substring of BLAND even though there is no morphological relation between them, and /ɹid/ is not a substring of /ɹɛd/ even though the latter verb is the past tense of the former.

This partial correspondence between our wordform networks and linguistic structure makes their relationship unclear. Substring relations among wordforms fall within in the purview of linguistics, but they do not appear to have a place in current linguistic theories. Nonetheless, the observed scaling relations are lawful and non-trivial, as we have argued, and may be universal as well. If so, then it may prove informative to investigate whether and how scale-free wordform networks may be accommodated by linguistic theory.

For instance, there are some phonological processes that may fit with our explanation of scale-free wordform networks. Processes like assimilation, elision, syncope, and apocope may generally help to make wordform lexicons more efficient by creating more overlap among wordforms, whereas processes like dissimilation, epenthesis, and prothesis may help to make wordform lexicons more distinctive by creating less overlap among wordforms.

Finally, similar ideas have been explored in Lindblom's Theory of Adaptive Dispersion (Lindblom, 1986; Lindblom, 1990) and in Ohala's Maximum Use of Available Features (Ohala, 1980). In Lindblom's theory, for instance, the phonological contrasts of a language are chosen to simultaneously 1) maximize the number of contrasts, 2) maximize the distinctiveness of contrasts, and 3) minimize articulatory effort. Constraint 2 is analogous to distinctiveness as we have defined it, except that phonological contrasts are more fine-grained than substrings. Constraints 1 and 3 stand in opposition to Constraint 2, and phonological systems must strike a balance between these opposing constraints, analogous to how lexicons must strike a balance between distinctiveness and efficiency. The similarities between Lindblom's theory and ours suggest possible avenues of fruitful exchange. In one direction, Lindblom's theory may benefit from principles of critical phenomena. In the other direction, our analysis may benefit from the inclusion of articulatory effort, which clearly has an important influence on the structure of wordforms. Such theoretical exchanges exemplify the kind of transdisciplinary work that is currently going on throughout the complexity sciences.

References

Albert, R., Hawoong, J., Barabasi, A.L.
1999 Diameter of the World Wide Web. *Nature,* 401, 130.

Bak, P. & Paczuski, M.
1995 Complexity, contingency, and criticality. *Proceedings of the National Academy of Sciences,* 92, 6689–6696.

Barabási, A., R. Albert, & Jeong, H.
1999 Mean-field theory for scale-free random networks. *Physica A,* 272(1), 173-187.
2000 Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A,* 281(1-4), 69-77.

Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T.
2002 Evolution of the social network of scientific collaborations. *Physica A,* 311(3-4), 590-614.

Fellbaum, C.
1998. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferrer i Cancho, R.

2006 When language breaks into pieces: A conflict between communication through isolated signals and language. *BioSystems,* 84, 242-253.

Ferrer i Cancho, R., & Solé, R.V.
2003 Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Science*, 100(3), 788-791.

Huang, K.
1963 *Statistical Mechanics.* New York: Wiley.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z., & Barabási, A.
2000 The large-scale organization of metabolic networks. *Nature*, 407(6804), 651-654.

Kirby, S.
2001 Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2), 102-110.

Li, W.
1992 Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory,* 38(6), 1842-1845.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M.
1967 Perception of the Speech Code. *Psychological Review,* 74(6), 431-461.

Lindblom, B.
1986 Phonetic universals in vowel systems. In J.J. Ohala and J.J. Jaeger (Eds), *Experimental Phonology*. Orlando, Florida: Academic Press.
1990 Phonetic content in phonology. *PERILUS*, 11.

Ma, S.
1976 *Modern theory of critical phenomena.* Reading: Benjamin/Cummings.

Mandelbrot, B.
1953 An Informational Theory of the Statistical Structure of Language. In W. Jackson (Ed), *Communication Theory*. London: Bettersworths.

Miller, G.
1957 Some effects of intermittent silence. *American Journal of Psychology,* 52, 311-314.

Morton, J.
1969 Interaction of information in word recognition. *Psychological Review,* 76(2), 165–178.

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

Ohala, J.J.
1980. Chairman's introduction to symposium on phonetic universals in phonological systems and their explanation. In *Proceedings of the Ninth International Congress of Phonetic Sciences*, 1979, 184-185. Copenhagen: Institute of Phonetics, University of Copenhagen.

Oudeyer, P.
2005 The self-organization of speech sounds. *Journal of Theoretical Biology,* 233(3), 435-439.

Solé, R. V.
2001 Complexity and fragility in ecological networks. *Proceedings of the Royal Society B: Biological Sciences,* 268(1480), 2039-2045.

Sornette, D.
2004 *Critical phenomena in natural sciences: chaos, fractals, selforganization, and disorder: concepts and tools* (2nd ed.). Berlin; New York: Springer.

Steyvers, M. & J. B. Tenenbaum
2005 The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science,* 29(1), 41-78.

Tsonis, A., Schultz, C., & Tsonis, P.
1997 Zipf's law and the structure and evolution of languages. *Complexity*, 2(5), 12-13.

Wasserman, S. & Faust, K.
1994 *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Zipf, G. K.
1949 *Human Behaviour and the Principle of Least Effort*. New York: Hafner.