LANGUAGE LEARNING, POWER LAWS, AND SEXUAL SELECTION

TED BRISCOE

Computer Laboratory
University of Cambridge
JJ Thomson Ave
Cambridge CB3 OFD, UK
ejb@cl.cam.ac.uk

I discuss the ubiquity of power law distributions in language organisation (and elsewhere), and argue against Miller's (2000) argument that large vocabulary size is a consequence of sexual selection. Instead I argue that power law distributions are evidence that languages are best modelled as dynamical systems but raise some issues for models of iterated language learning.

Keywords: Zipf Curve, Iterated Learning Model, Small World Distribution, Evolutionary Linguistics, Diathesis Alternation

1. Introduction

A diagnostic of a power law distribution is that a log-log plot of frequency against rank yields a (nearly) straight line. For instance, Zipf (1935) plotted word token counts in a variety of texts against the inverse rank of each distinct word type and showed that typically such plots approximate a straight line. The characteristic 'Zipf curve' of word frequency against rank deviates from this line because the relative frequency of very common word types, such as the English determiners *the* and *a*, tend to be more similar than the power law predicts, as also does the relative frequency of very rare words in the tail of the distribution. Zipf's 'law' is often expressed as:

$$c(w) \propto \frac{1}{r(w)^B} \tag{1}$$

where $B > 1$, the exponent, defines the slope of the plot, frequency $c(w)$ is the token count of word type w in text, and rank $r(w)$ is the position of word type w in the list of word types sorted in descending order of frequency, $c(w)$. Guiraud's (1954) related law states that the number of word types V in a text is proportional to the length of that text N:

$$V \propto N^A \tag{2}$$

Although the models of power law distributions of which I am aware have a dynamical component, they have received little attention from evolutionary linguists. I know of only one argument, due to an evolutionary psychologist (Miller, 2000), which utilises Zipf's observation about word frequencies in attempting to explain large, redundant vocabularies in terms of sexual selection. I argue against this explanation in §4, but, before doing so, I discuss the ubiquity of power law distributions in §2, some relevant models of them in §3, return to Miller's argument in §4, and then discuss some issues power law distributions raise for evolutionary models of iterated language learning in §5.

2. Manifestations of Power Law Distributions

Power law distributions are very different from normally distributed phenomena, such as height, which yield the characteristic 'Bell Curve'. The factors that influence a person's height, such as nutrition and genetic inheritance, combine in a more linear manner so that (relatively minor) variation in height is normally distributed around a mean that can be accurately estimated from a representative sample of the population. Zipf (1949) noted that Pareto's observations about the distribution of wealth in the population could also be modelled using a version of his 'law'. What makes wealth different from height intuitively is that the factors that influence the amount of money we have combine non-linearly and there are strong (positive) feedback effects (i.e. 'the rich get richer'). We now know that power law distributions are good approximations of many other non-linguistic phenomena, such as the distribution of people within cities, citations amongst scientists, accesses of web pages, species within habitats, authors amongst scientific articles, actors within films, links between web pages, activation of genes, size of earthquakes, number of sexual partners, and many more (e.g. Albert & Barabasi,2002).

There are similar results regarding extrinsic properties of languages: for instance, the distribution of languages within language families approximates a power law (Wichmann, 2005). In terms of inherent properties of language, Zipf also showed that plotting the length or the number of meanings of word types against their frequency also yields similar distributions. With the increasing availability of annotated electronic corpora, these observations have been extended to many other areas of language organisation, such as the frequency of contiguous sequences of words (bigrams, trigrams, and more generally ngrams), of grammatical rules, of construction types, of lexical relations between word types, as well as the length of constituents, and the association of verbs with constructions (e.g. Sharman, 1989; Manning & Schutze, 1999; Korhonen 2002; Yook *et al.*, 2001).

Given these results, it is tempting to speculate that all distributions relating to language will approximate power laws. However, this is not always the case; or, at least, deviations from the approximation can have significant practical import: for instance, Preiss *et al.* (2002) show that the average frequency mass assigned to the first WordNet sense of high frequency verbs is about 45%, the second around 25% and the third around 5% (ignoring the tail of infrequent senses which account for the last 25% of the mass). This is still a skewed distribution, but a prediction based on the power law assumption would underestimate the frequency of the second sense in order to better approximate the tail. The empirical data though tells us that we ignore the second sense at our peril if we wish to capture the bulk of the behaviour of tokens of such verbs in texts. Arguably, this specific observation reduces to the more general one that we tend to see 'Zipf curves' rather than straight lines. However, here the lack of fit between curve and line is made more severe by the lower number of types – even polysemous verbs rarely have more than 15 senses, while there will be in excess of 50K word types in a typical text corpus of around 1M word tokens.[a] Similarly for external features of languages, Schulze and Stauffer (2006) point out that the distribution of speakers amongst languages is better approximated by a lognormal distribution than a power law.

As well as being careful about the goodness of fit between a power law and the more frequent types of any linguistic distribution, we also need to consider the nature of the tail. All such distributions are characterised by a long tail consisting of a high proportion of very infrequent types. However, these also often bunch together curving downwards graphically in a manner closer to an exponential distribution than a power law. Ferrer i Cancho & Sol'e (2001) demonstrated using the 100M word British National Corpus that the tail of frequency-rank word plots of singly-authored subcorpora are more bunched than plots of multiply-authored text which, therefore, more closely approximate a power law. Zipf's original examples are mostly singly-authored and relatively small text samples, so this result suggests that vocabulary distributions for idiolects or I-language may differ from those of the general (E-)language.

3. Models of Power Law Distributions

So far I have used the term 'distribution' ambiguously between the linguistic and probabilistic sense. The most important insight about such distributions with large numbers of rare events (e.g. Baayen, 2001) is that it is unwise to convert a frequency-rank plot into a probability-rank plot via maximum likelihood (i.e. relative frequency) estimation, and treat the result as a probability distribution.

[a]There have been many attempts to model 'Zipf Curves' more accurately beginning with Mandelbrot (1953) and Simon (1955) and continuing more recently with Church & Gale (1995) who use mixtures of Poisson distributions to model word and ngram distributions for applications such as information retrieval and speech recognition. I ignore these here as they aren't relevant to the specific goals of this paper.

Since the counts of the tail are very low, statistical estimation theory tells us that they will be unreliable. A rare word, for instance, may suddenly become fashionable (e.g. the frequency of *egregious* and *serendipity* has increased markedly in the last five years) and thus increase in relative frequency over a given time period.[b] Since, we always see a long tail of rare events no matter how much (more) text we sample, and the number of types grows in proportion to the size of this sample (Guiraud's law), power law distributions are often described as 'scale-free'. In statistical terms, power law distributions which remain invariant over different sample sizes are a strong indication that we may be sampling from a statistically unrepresentive non-stationary (i.e. dynamical) system.

One of the most intuitive models of such a system, due to Bak (1996), is that of a sand pile built up on a flat surface of finite size by addition of sand grains. When the pile reaches 'self-organized criticality' (i.e. the slopes of the sides are steep), a new grain will trigger a landslide. Frequency-rank plotting of the size of the landslides produces a power law distribution. What is apparent about this rather simple, though mathematically complex, model is that it is a dynamical system in which the addition of each individual grain of sand represents a discrete time step. Most landslides are small, but occasionally individually-unpredictable larger ones occur.

Baayen (1991), following in the tradition of Mandelbrot (1953) and Simon (1955), develops a stochastic Markovian model of phonotactically legal Dutch word strings and relates it to empirical data on similarities between words by phonological form and by relative frequency. He finds that to model these effects accurately, it is necessary to add a second 'dynamical' stochastic model which introduces or removes word types with probability proportional to their token frequency. This has the effect of increasing overall frequency-based and decreasing form-based similarity. For present purposes, it is indicative that the second dynamical word 'birth-death' processs is required even though it says nothing directly about the relationships between word types.

Albert & Barabasi (2002) provide a recent survey of work on 'small world' networks in which most nodes of a network can be reached by any other in a small number of (node) steps, though the overall number of nodes can be arbitrarily high. They define a dynamical algorithm for generating such networks, by continuously adding new nodes and attaching them to old nodes with probability proportional to their number of existing links. They prove that such networks evolve to a scale-free organisation obeying a power law distribution in which there is a long tail of nodes with low numbers of links and a small number of 'popular' nodes with many links. They also prove that both 'growth', the dynamical component, and 'preferential attachment' are necessary for this pattern to emerge. Such

[b] Such effects can be monitored, for example, using the 'top 20' on-line dictionary queries published by Cambridge University Press, http://dictionary.cambridge.org/top20/top20_0205.asp

networks have been applied to models like that of Baayen (1991), described above (e.g. Bornholdt & Ebel, 2001), and to lexical semantic organisation (e.g. Yook *et al.*, 2001).

4. Power Laws and Sexual Selection

Miller (2000:369f), in the context of a more general argument that human language evolved by sexual selection, argues that large vocabulary size, in comparison with those of other (artificial and natural) animal communication systems, evolved through sexual selection. Women preferred men with large active vocabularies but needed to acquire large passive vocabularies themselves to assess the trait. Miller offers, as evidence for the non-functional nature of much of this vocabulary, Zipf's observation that vocabulary distributes like a power law and contains many near synonyms:

> ...any of the words we know is likely to be used on average about once in every million words we speak... Why do we bother to learn so many rare words that have practically the same meanings as common words, if language evolved to be practical? (Miller, 2000:370)

He argues that human variation in vocabulary acquisition correlates with intelligence and has a heritable component, and thus is an (indirect) fitness indicator, triggering an 'arms race' in which advertising excessive vocabulary size is a 'display' of fitness akin to the peacock's tail, precisely because it does not contribute usefully to communication.

It is plausible that human language evolved under selection for 'communicative success'. Otherwise, it is hard to understand how our cognitive abilities adapted to support acquisition of large vocabularies and complex grammatical systems (e.g. Briscoe, 2000). 'Communicative success' is defined in terms of parity of form-meaning mappings between agents supporting accurate sharing of meaning. It is essentially neutral about the extent to which this ability is exploited for specific social acts. Language is certainly useful for courtship and seduction, as Miller argues at length. But it is also useful for trading, teaching, bonding, lying, and much more. To demonstrate that vocabulary size and distribution, or any other linguistic trait, is under sexual rather than natural selection, it is necessary to show that it wouldn't evolve in any other way and doesn't contribute to communicative success. Miller fails on both counts.

In §2 we saw that power law distributions manifest themselves in many areas of linguistic organisation. For instance, there is a tail of rare long constituents in text samples (Sharman, 1989). However, there is no evidence that 'display' of such forms is a particular feature of courtship, nor that such forms are non-functional. As we saw in §3 models predicting such distributions need only a dynamical component and no element of natural or sexual selection whatever. Evidence of power law distributions in both idiolects and language forces us to conclude that both are best modelled as dynamical systems – rather than well-formed sets, as in generative linguistics (e.g. Sampson, 2001:165f) – but nothing more.

If vocabulary size were non-functional, we might expect there to be many truly synonymous words. What we find in the organisation of vocabulary is that partially synonymous words have different distributions in terms of specificity of reference, syntactic potential, or genre and register. There is, in fact, considerable evidence that children avoid hypothesising synonyms in language acquisition (e.g. Clark, 2003) and that language users adhere to the convention of preemption by synonymy, except where discourse or syntactic context triggers a non-synonymous reading (e.g. Briscoe *et al.*, 1995; Copestake & Briscoe, 1995). For instance, *cow*, unlike *chicken*, is not generally used to refer to the meat because of the existence of *beef*. However, in an appropriate context *cow* can be used this way and triggers an implicature of 'disgust':

> There were five thousand extremely loud people on the floor eager to tear into roast cow with both hands and wash it down with bourbon whiskey. (Tom Wolfe, 1979. *The Right Stuff*, Farrar, Straus and Giroux, New York (p. 298, Picador edition, 1991))

Similarly, the word *stealer*, formed by the fairly productive derivational rule of agentive *+er* nominalisation, is blocked by *thief*, except in syntactic contexts where the specificity of reference is narrowed:

> He is an inveterate *stealer / thief / stealer of Porsche 911s

These and many similar observations suggest that partial synonymy is communicatively useful and actively exploited to convey meaning.

To understand why we have so many words and how the cognitive ability to cope with them (co-)evolved, consider the likely environment of adaptation for language. In a foraging, scavenging or hunter society, the ability to discriminate – and thus name more and more species, according to nutritional value, location, method of capture or harvesting, and so forth – would be of value for survival because it would allow efficent transmission of these skills to kin as well as survival over larger and more varied habitats. Modern hunter-gatherers are known to have large vocabularies specialised in this way (Diamond, 1997). This may not have been the sole driver for increasing vocabulary size, but it has the advantage that it predicts that vocabulary will be to a large extent organised by specificity of reference. It is useful not only to be able to talk about plants in general but also species and subgroups (e.g. by location or edible part) in order to discriminate the edible, find the source, and harvest effectively. Once we accept such a pressure to name in an increasingly complex and multifaceted environment, then the tendency for there to be smaller numbers of high frequency words of generic reference and a larger number of rarer words with highly specific denotations is just a case of the structure of vocabulary mirroring (our perception of) this environment.

5. The Real Challenge – Iterated Learning

One achievement of recent evolutionary models of language is the demonstration that treating languages as complex adaptive systems responding to conflicting selection pressures (e.g. Briscoe, 2000) leads to insightful acccounts of typological and other linguistic universals without the need to invoke innateness. These accounts rely heavily on the iterated learning model (ILM, e.g. Kirby, 2001) in which linguistic traits must undergo repeated relearning by successive generations of language learners acquiring their language from that of the previous generation. For instance, Kirby (2001) demonstrates that languages in the ILM evolve to have compositional structure in which only high frequency irregular form-meaning mappings are stable, given the following assumptions:

1. an *invention strategy* for form-meaning pairs,
2. a *production bias* to express meanings using short forms,
3. an *inductive bias* to learn small grammars and lexicons,
4. a *learning period* in which not all form-meaning pairs appear
5. and *environmental structure* which favours some meanings

In the simulation, initial (proto)languages are holistic and non-compositional but chance regularities which emerge in form-meaning mappings are acquired by learners, who then reliably exemplify them for the next generation of learners, because regularities are, by definition, more frequent in data. Thus, over time the language evolves to be mostly compositional and regular. However, (short) irregular mappings can survive provided they are associated with meanings which are expressed frequently and, therefore, also occur reliably during the learning period.

This instantiation of the ILM neatly explains the observation that irregularity correlates with high frequency in attested languages: children would continue to say *goed* into adulthood if *went* were not a high frequency form. The corollary, however, is that rare unpredictable properties of language which do not follow from some regularity manifest during the learning period should be unstable and, therefore, rarely observed.

Rare word-meaning associations are unpredictable and may also influence lexico-grammatical behaviour. For example, the verb *obsess* is a stable lexeme of English, but does not appear in any of the 40 or so case studies of child-directed speech in CHILDES[c]. It is transitive but usually appears in the passive in adult speech accompanied by a PP headed by *by*, *with* or *over*. However, vocabulary acquisition continues through adulthood, so the ILM (and other models) simply predict that such vocabulary will be acquired later (and less universally).

---

[c]http://childes.psy.cmu.edu/data/

Marked but predictable constructions, such as multiple centre-embeddings, which Sampson (2001:21) estimates occur once in every 250K words on average, are also not counter-examples if one believes that they are a consequence of learners acquiring, on the basis of more frequent constructions, grammatical rules which correctly predict the appropriate form-meaning mapping for center-embedded constructions.

A more challenging case for the ILM is diathesis alternation, in which verbs of certain semantic classes semi-predictably occur in alternant constructions often with predictable meaning changes. For instance, *eat* can appear in intransitive and transitive constructions but when it occurs intransitively the theme of the action is 'understood'. However, verbs with similar senses, such as *devour* or *consume* do not undergo this alternation. There is evidence that children learn at least some of these alternation rules by around three years old because they produce errors, such as *Don't fall my dolly down*, an apparent overapplication of the causative-inchoative alternation, and because novel alternant constructions with pseudo-verbs can be elicited from children of this age under experimental conditions in which they have only heard the pseudo-verb in one construction (Conwell and Demuth, 2007). Nevertheless, the rate at which errors of this type naturally occur in children's spontaneous speech also suggests that alternation rules are learnt conservatively and only rarely overapplied. There are on the order of 100 such alternation rule types in English, when productive meaning change is taken into account.

Korhonen *et al.* (2006) describe Valex, a lexicon of over six thousand English verbs constructed from nearly one billion words of automatically-parsed English text containing occurrences of these verbs. In Valex, verbs are associated with one or more of over 150 verb-headed constructions based on the frequency with which they have been observed with them in the automatically-parsed text. Because the automatic parses are somewhat noisy, the final lexicon is constructed using statistical filtering and smoothing techniques which make use of extant manually-created dictionaries and of information about the prototypical constructions associated with verbs in specific semantic classes, such as verbs of motion, transfer of possession, and so forth. Our most accurate lexicon created this way has a F-measure of 87.3%, which equals the accuracy of manually-created dictionaries (see Korhonen *et al.* for details of the evaluation, filtering and smoothing), but includes information about the relative frequency of any given construction occurring with any of these verbs.

Figure 1 shows a plot of the inverse rank of these verb-headed constructions against their log frequency for all the verbs in Valex before and after filtering. The unfiltered line shows the output of the automatic parser and the filtered line shows the predicted distribution of the most accurate lexicon. Both exhibit highly-skewed distributions, especially for the higher frequency constructions; for example, the most common transitive construction has over twice as many counts as

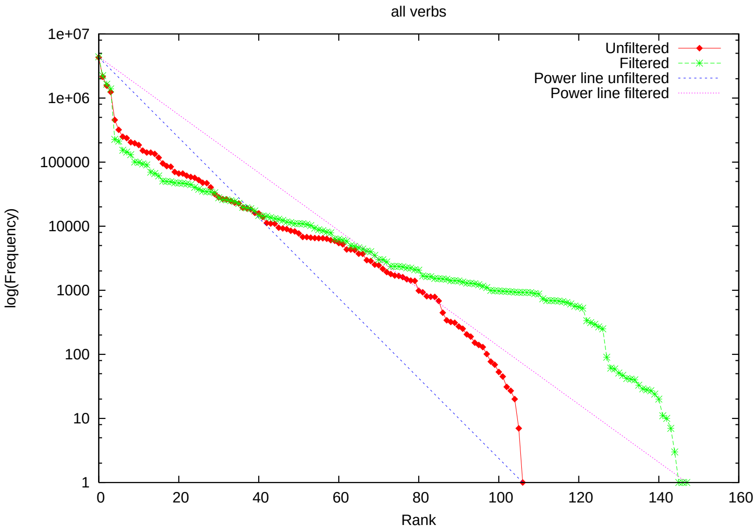

Figure 1. Overall Distribution of Verb-headed Constructions

the next most common intransitive construction, while the sentential complement construction (with or without a complementiser *that*, as in *I believe (that) the world is round*) is ranked fifth but has more than one order of magnitude less counts than the transitive construction.

Figure 2 shows a similar plot of the inverse rank of verb-headed constructions against their log frequency for the verb *believe* before and after filtering. The plot for *believe* is prototypical of that for other verbs of varying frequencies which are associated with multiple constructions. Both the overall distribution for all verbs and that for *believe* loosely approximate to power laws, shown as straight lines on the graphs, though we make no stronger claim than that these distributions are highly-skewed with a few very frequent constructions and a tail of far less frequent ones. Crucially, however, the correlation between the overall distribution of verb-headed constructions and the specific distribution for any given verb is low. For instance, although the sentential complement construction is ranked fifth in the overall distribution and predicted to occur about one tenth as often as the transitive construction, *believe* occurs most frequently in this construction with the transitive in second place. There are various ways to measure the correlation between the overall and verb-specific distributions, all of which lead to similar conclusions (see

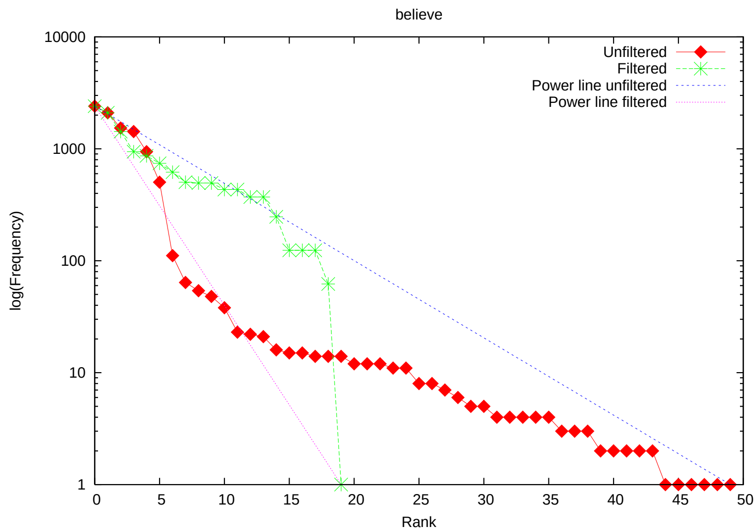

Figure 2. Distribution of Verb-headed Constructions for *believe*

e.g. Korhonen, 2002). One which is straightforward to interpret is the Spearman Rank Coefficient, which measures the similarity between the frequency ranked order of constructions between the distributions and yields a value between 0 (no correlation) and 1 (perfect correlation). According to this measure, the similarity of the filtered overall distribution and that for *believe* is 0.73, while the average similarity of all the verb specific distributions to the overall distribution is only 0.48. The correlation improves when we compare verb-specific distributions to those prototypical for the same semantic class of verbs (e.g. manner of motion or propositional attitude), but typically only to around values of 0.7 on the same measure. Therefore, even if the learner has access to such classes in some way, predicting verb-construction associations from such imperfect correlations would be highly errorful.

This comparative lack of correlation, taken together with the fact that analysis of the CHILDES database shows that child-directed speech only reliably exemplifies the common verb-construction associations and not the longer tail of rarer associations (e.g. Buttery & Korhonen, 2005), suggests that children may not have reliable evidence for the existence of most alternation *rules* – assuming that evidence would be several exemplars of the same alternation involving several different verbs. For example, returning to the causative-inchoative alternation, a child would need to be exposed to at least the following four sentences (or analogues) to have the minimum evidence from which to infer the existence of a rule which relates the a) and b) cases.

a) The Duke of York marched his men up the hill.
b) The men marched up the hill.
a) The cowboy galloped his horse across the prairie.
b) The horse galloped across the prairie.

From such evidence, a child might induce, for example, that manner of motion verbs *may* participate in the causative-inchoative alternation and thus make appropriate inferences about external/internal causation of motion in the face of novel verbs in these constructions, and may even occasionally and sometimes inappropriately use such a rule to generate novel utterances, as with *Don't fall my dolly down*.

The ILM in its current form predicts that over successive generations such rules should tend to become more productive in the evolving language if there is enough data to support their induction in the input to any child. However, this doesn't seem to be the case with such semi-productive alternation rules, they appear to remain stably semi-productive. For example, for most British English speakers it is inappropriate to extend dative movement from *give* to *donate* despite the existence of the alternation rule, and of both verbs within the 'transfer of possession' semantic class, for about the last thousand years.

It may be that such semi-productive rules are acquired later in life, despite the occasional apparent overapplication errors in children's speech and despite some experimental evidence supporting their early acquisition. This is a general strategy that proponents of ILM-style explanations can take. But on the other hand, there must also be some learning 'bottleneck', caused by limited exposure to data during the learning period, for ILM accounts of linguistic evolution to work. Cases like this pose interesting challenges for the approach because they suggest that linguistic data is distributed in such a fashion that there may still be a 'poverty of stimulus' issue during the sensitive period for acquisition, though this is not to suggest, of course, that the correct solution is necesarily a more nativist account of acquisition.

More empirical work on language acquisition is needed to determine whether the ILM's predictions hold up for such specific cases. For example, in the case of diathesis alternations it would be very useful to know whether children's input provides the necessary data for acquisition of the rules and whether their output provides evidence of productive use of such rules, yielding verb-construction combinations that they haven't been exposed to in caretaker input.

Acknowledgements

I would like to thank the anonymous reviewers for their helpful comments, and Paula Buttery and Anna Korhonen for analysis and plots of data from the Valex lexicon.

References

Albert, R. & A. Barabasi (2002) 'Statistical machanics of complex networks', *Reviews of Modern Physics, vol.74,* 47–97.

Baayen, H. (1991) 'A stochastic process for word frequency distributions', *Proceedings of the Assoc. for Computational Linguistics,* Morgan Kaufmann, Menlo Park, CA, pp. 271–278.

Baayen, H. (2001) *Word Frequency Distributions,* Kluwer, Dordrecht.

Bak, P. (1996) *How nature works: the sicence of self-organized criticality,* Copernicus Press, New York.

Bornholdt, S. & Ebel,H. (2001) 'World Wide Web scaling exponent from Simon's 1955 model', *Physical Review, vol.64,* 035104.

Briscoe, E.J. (2000) 'Evolutionary perspectives on diachronic syntax' in (eds) Pintzuk, S., Tsoulas, G. and Warner, A. (eds.), *Diachronic Syntax: Models and Mechanisms,* Oxford University Press, Oxford, pp. 75–108.

Briscoe, E.J., A.A Copestake & A. Lascarides (1995) 'Blocking' in St. Dizier, P. and Viegas, E. (eds.), *Computational Lexical Semantics,* Cambridge University Press, Cambridge, pp. 273–302.

Buttery, P. & A. Korhonen (2005) 'Large-scale analysis of verb subcategorization differences between child directed speech and adult speech', *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes,* Saarland University.

Clark, E. (2003) *First Language Acquisition,* Cambridge University Press, Cambridge.

Church, K. & W. Gale (1995) 'Poisson Mixtures', *Natural Language Engineering, vol.1,* 1–36.

Conwell, E. & K. Demuth (2007) 'Early syntactic productivity: evidence from dative shift', *Cognition, vol.103,* 163–179.

Copestake, A.A. & E.J. Briscoe (1995) 'Regular polysemy and semi-productive sense extension', *Journal of Semantics, vol.12,* 15–67.

Diamond, J. (1997) *Guns, Germs and Steel: The Fate of Human Societies,* Random House, New York.

Ferrer i Cancho, R. & R. Sol'e (2001) 'The small world of human language', *Proceedings of the Royal Society B: Biological Sciences, vol.268(1482),* 2261–2265.

Guiraud, H. (1954) *Les Charact'eres Statistiques du Vocabulaire,* Press Universitaires de France, Paris.

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

Korhonen, A. (2002) *Subcategorization Acquisition,* Computer Laboratory, University of Cambridge, Techical Report UCAM-CL-TR-530.

Korhonen, A., Y. Krymolowski & E.J. Briscoe (2006) ‘A large subcategorization lexicon for natural language processing applications’, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC06),* Genova, Italy.

Mandelbrot, B. (1953) ‘An informational theory of the statistical structure of language’ in W. Jackson (eds.), *Communication Theory,* Butterworths, London.

Manning, C. & H. Schutze (1999) *Foundations of Statistical Natural Language Processing,* MIT Press, Cambridge MA.

Miller, G. (2000) *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature,* William Heinemann, London.

Preiss, J., A. Korhonen & E.J. Briscoe (2002) ‘Subcategorization Acquisition as an Evaluation Method for WSD’, *Proceedings of the Language Resources and Evaluation Conference (LREC02),* Morgan Kaufmann, menlo Park, Ca., pp. 1551–1556.

Sampson, G. (2001) *Empirical Linguistics,* Continuum, London.

Sharman, R. (1989) *Observational Evidence for a Statistical Model of Language,* IBM UKSC Report 205.

Simon, H. (1955) ‘On a class of skew distribution functions’, *Biometrika, vol.42,* 435–440.

Schulze, C. & D. Stauffer (2006) ‘Recent developments in computer simulations of language competition’, *Computing in Science and Engineering, vol.8,* 86–93.

Wichmann, S. (2005) ‘On the power law distribution of language family sizes’, *Journal of Linguistics, vol.41,* 117–131.

Yook S., H. Jeong, A-L Barabasi & Y. Tu (2001) ‘Weighted evolving networks’, *Physical Review Letters, vol.86,* 5835–5838.

Zipf, G. (1935) *The Psycho-Biology of Language: An Introduction to Dynamic Philology,* Houghton-Miflin, New York.

Zipf, G. (1949) *Human Behavior and the Principle of Least Effort,* Addison-Wesley, Cambridge, MA.