TOWARD A MODEL OF LANGUAGE ACQUISITION THRESHOLD

ABSTRACT
We demonstrate how the paradigm of complex networks can be used to model some aspects of the process of second language acquisition. In learning a new language, knowledge of 3000-4000 most frequent words appears to be a significant threshold, necessary to transfer reading skills from L1 to L2[1]. We show that this threshold corresponds to the transition from the Zipf's law to non-Zipfian regime in the rank-frequency plot of words of the English language. Using a large dictionary, we then construct a graph representing the dictionary, and study topological properties of subgraphs generated by k most frequent words of the language. Clustering coefficient of these subgraphs reaches a minimum in the same place as the crossover point in the rank-frequency plot. We conjecture that the coincidence of all these threshold may indicate a change in the language structure which occurs when the vocabulary size reaches about 3000-4000 words.

KEY WORDS
Modelling and simulation methodologies, language acquisition, random graphs, complex networks

1 Introduction

In the past decade, complex networks composed of a large number of interacting components became an important paradigm in modelling of natural, social, and technological phenomena. Many hundreds of publications appeared, reporting various properties of large-scale complex networks and attempting to describe their topology and dynamics using a variety of tools drawn from diverse disciplines including graph theory, probability and statistics, as well as statistical physics. Examples of successful applications of this paradigm include models of collaboration networks, food webs, traffic networks, complex networks in genomics and proteomics, power grids, and many others.[2]

Given the complex nature of human languages, it is not surprising that the network paradigm has been utilized to study linguistic phenomena. For example, it has been demonstrated that co-occurrence of words in sentences can be described in terms of a scale-free graph exhibiting the so-called small-world effect [8]. Similarly, terms of a thesaurus can be viewed as nodes of a large graph, with graph edges representing relationships between terms. It has been found that the degree distribution of this graph also exhibits many features typical to scale-free networks [5].

[1]L1 refers to the first language, and L2 the second language. Those are common abbreviations used in lingustic literature.

In this paper, we will investigate another linguistic phenomenon, namely the process of second language acquisition. We will show that the paradigm of complex networks can be applied to model some aspects of this highly complicated process.

In the past, learning a foreign language was viewed mainly as a matter of mastering the language's grammar, with a relatively minor importance attached to the vocabulary development. Contemporary language acquisition specialists, however, recognize the central importance of the vocabulary, and in the last two decades a lot of research effort went into the study of vocabulary learning strategies, determining what it means to "know a word", and methods of testing vocabulary knowledge and use [12].

One of the first questions which one encounters while learning a new language is "how much vocabulary do I need to know?". Of course, the most ambitious goal would be to know all words of the language. Such goal, however, is usually impossible to attain, as even native speakers do not know all of the language. While comprehensive dictionaries of English can easily contain over 10^5 word families, it has been demonstrated that educated native speakers of English know only a fraction of this lexicon – about 20000 word families [6].

Many language scholars agree that the significant threshold in the language learning occurs perhaps around 3000-4000 word families. It turns out that once this threshold is reached, learners can understand over 90% of the running words in a typical text [3]. Such high coverage of the text, in turn, appears to be a necessary condition for transferring reading skills from the first to the second language [9]. In what follows, we will call this threshold a *linguistic threshold*, to be referred to as T_l.

The goal of this work is to shed some light on the aforementioned threshold. We will show that some aspects of the language structure also exhibit thresholds located very close to T_l.

2 Zipf's law

In 1932, George Zipf [15] found that in a large text corpus there exists a striking approximate relation between the frequency of the occurrence of the word and its rank in the list of all words. By rank r we mean the position of the word in the list of all words arranged by decreasing frequency. If $f(r)$ is the frequency of occurrence of the word with rank r, then the Zipf's law states that

$$f(r) = \frac{A}{r^z}, \tag{1}$$

where A is the normalization constant and z is the exponent which usually takes a value slightly larger than 1. While the Zipf's law is only approximately true, and better phenomenological models for rank-frequency statistics of words have been proposed, we will use Zipf's law as a starting point for subsequent considerations. For simplicity, let as assume that the value of the exponent z is exactly 1, and let the total number of words in the language be N. The normalization constant A will then be given by

$$A = \left(\sum_{i=1}^{N} \frac{1}{i} \right)^{-1} = \frac{1}{\Psi(N+1) + \gamma}, \tag{2}$$

where Ψ is the digamma function, defined as the logarithmic derivative of the gamma function $\Psi(x) = \frac{d}{dx} \ln \Gamma(x)$, and $\gamma = 0.57721566\ldots$ is the Euler-Mascheroni constant.

Let us first assume that the learner of a foreign language learns new words following the frequency list, starting from the most frequent words and moving down the list. If the learner knows k top-ranking words, then the text coverage, or the fraction of known words is

$$C(k) = \sum_{r=1}^{k} f(r) = A(\Psi(k+1) + \gamma). \tag{3}$$

The asymptotic expansion of the digamma function is given by

$$\Psi(k+1) \sim \ln k + \frac{1}{2k} - \sum_{n=1}^{\infty} \frac{B_{2n}}{2nk^{2n}}, \tag{4}$$

where B_{2n} are Bernoulli numbers. One expects, therefore, that for large k the leading term of $C(k)$ should be

$$C(k) \sim A \ln k, \tag{5}$$

meaning that the text coverage should roughly be a linear function of the logarithm of the vocabulary size. This already is a bad news, as it requires exponentially growing effort to keep the coverage increasing at a constant rate. Yet in reality the situation is even worse. Figure 1 show the percentage text coverage as a function of vocabulary size (based on data from [3]), plotted in semi-logarithmic coordinates. Up to about 4000 words, the plot follows eq. (5) rather well, but for larger vocabulary sizes, the actual coverage is *smaller* than what Zipf's law (eq. 1) with $z = 1$ would predict. The reason for this behavior of $C(k)$ is the deviation from Zipf's law which can be observed for low-frequency words. In [11], M. Montemurro studied word-frequency distribution of English words using a large corpus consisting of 2606 books in English. He found that words for which the rank is below 3000-4000 obey Zipf's law regardless of the text length. Above this limit, there seems to be another power law analogous to eq. (1), although with a much larger exponent z, close to $z = 2.3$ and possibly even larger. This is illustrated in Figure 2. The point above which the Zipf's law is no longer valid will be called *Zipfian threshold*, to be referred to as T_z.

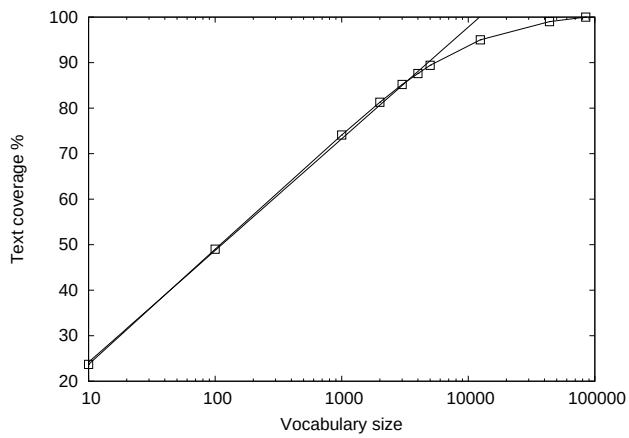

Figure 1. The percentage text coverage as a function of vocabulary size (based on data from [3]). Straight line represents the least square fit to the first seven data points.

The origin of the crossover from Zipf's law to non Zipfian behavior remains unknown. It is worth mentioning, however, that it is possible to encompass both these regimes within a single a framework of a semi-empirical model based on a single differential equation, originally used to describe re-association in folded proteins [13]. The starting point is the observation that that the Zipf's distribution (1) satisfies

$$\frac{df}{dr} = -\lambda f^q, \tag{6}$$

where $\lambda = zA^{-1/z}$ and $q = 1 + 1/z$. M. Montemurro [11] suggested the following generalization of the above equation

$$\frac{df}{dr} = -\mu f^p - (\lambda - \mu) f^q, \tag{7}$$

with p, q, μ and λ being positive real parameters. By fitting solution curves to the data, one can find values of these parameters, and the resulting $f(s)$ line exhibits behavior similar to the rank-frequency distribution shown in Figure 2. This suggest a possible connection between the equation (7) and the mechanism leading to the formation of the language, although details of such a connection remain unknown.

3 Self-hosting

The evidence provided so far seems to support the idea that some sort of structural change takes place when the vocabulary size reaches 3000-5000. But why would this threshold be significant in the process of language learning?

A possible explanation may be related to the concept of "self-hosting" known in the theory of computer languages. A computer language compiler is self-hosting if it is natively implemented in its own language. This is also known as bootstrapping [10]. It has been demonstrated that

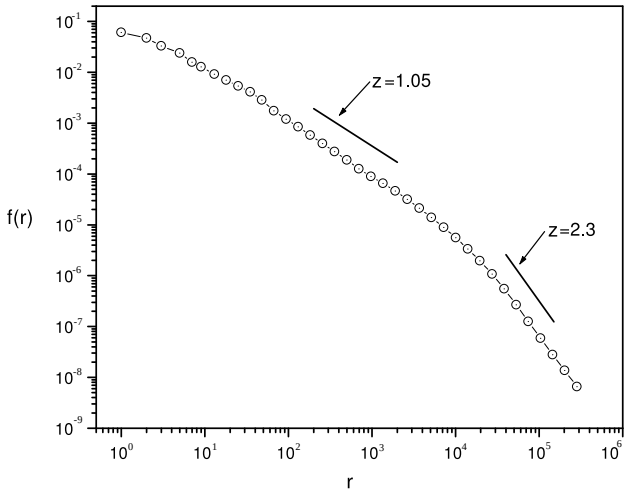

Figure 2. Frequency-rank plot for a large corpus comprising 2606 books in English [11]. Figure adopted from http://arxiv.org/abs/cond-mat/0104066.

the size of the source code for a self-hosting compiler can be surprisingly small. For example, the "Obfuscated Tiny C Compiler (OTCC)" is only 446 lines of code with one statement per line, yet it is able to compile itself [1].

Everyone who learns a new language knows that single-language dictionaries such as, for example, "Oxford English Dictionary", are not useful at the beginning, since one does not know enough vocabulary to understand word definitions. A bilingual dictionary must be used instead. At some point however, the single-language dictionary becomes more useful than the bilingual one - a clear sign that the knowledge of the language reached a level capable of "bootstrapping" or "self-hosting", that is, defining unknown words in terms of already known words of the new language.

In order to model this phenomenon we decided to use a large dictionary of English language available from Project Gutenberg web site, also known as *The Gutenberg Webster's Unabridged Dictionary* [7]. The dictionary has then been converted into a large graph with vertices representing individual words. If one word occurs in the definition of another word, then these two words (vertices) are linked with an edge.

For the purpose of this project, the dictionary has been altered in the following ways:

- Entries consisting of more than one word were omitted.
- All senses of a word were considered to be a single vertex in the resulting graph.
- All definitions for abbreviations, prefixes and suffixes were omitted. Abbreviations within definitions were deleted.

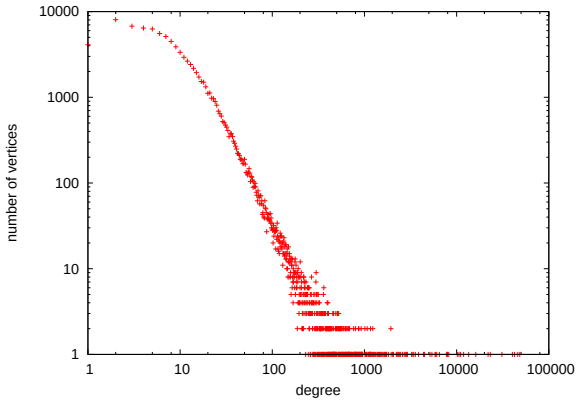

Figure 3. Degree distribution of $\mathcal{G}$.

- Entries that were simply alternate spellings of another entry or pointers to other entries were deleted.
- All pronunciations, references to illustrations, and other miscellaneous items were deleted.

The resulting graph, to be referred to as $\mathcal{G}$, has about 10^5 vertices and 10^6 edges (exactly 93 062 vertices and 1 124 654 edges). Its degree distribution, illustrated in Figure 3, appears to obey a power law for all except very small and very large degree values. It is interesting to note that similar distribution have been observed in a number of recently investigate complex networks, including, among others, collaboration graphs, such as collaboration of scientists or film actors [2]. Even more importantly, a very similar distribution has been reported in the graph representing English thesaurus [5].

4 The model

The person learning English knows at a given moment only a subset of all words represented by vertices of $\mathcal{G}$. Let W denote the set of known words, and let $\mathcal{G}_W$ be the subgraph of G generated by W. We will assume that the learner learns new words in the order dictated by the frequency list, starting with most frequent words and progressing toward less frequent words. The set of k top-ranking words will be denoted as $W(k)$. We can now consider a family of graphs $\mathcal{G}_{W(k)}$ with $k \in \{1, 2, \ldots, N\}$.

In order to study properties of $\mathcal{G}_{W(k)}$, one clearly needs a frequency list for words of the English language. We used the frequency list obtained from the American National Corpus, a large electronic collection of American English texts consisting of 22 million words [4]. We generated graphs $\mathcal{G}_{W(k)}$ using that list and investigated how the topological structure of these graphs changes with the number of vertices k.

While thinking of subgraphs $\mathcal{G}_{W(k)}$ in the context of the aforementioned self-hosting or bootsraping, one would expect that small subgraphs should be disconnected, and at

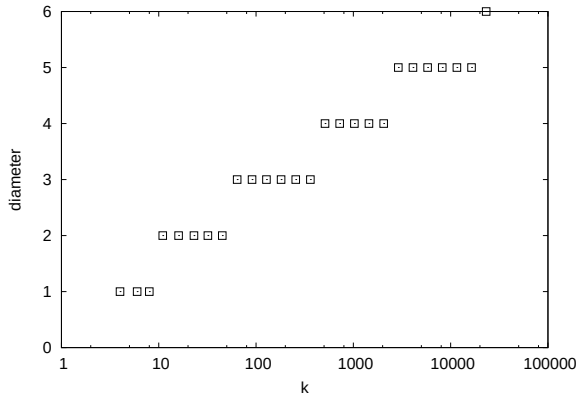

Figure 4. Diameter of $\mathcal{G}_{W(k)}$ as a function of k.

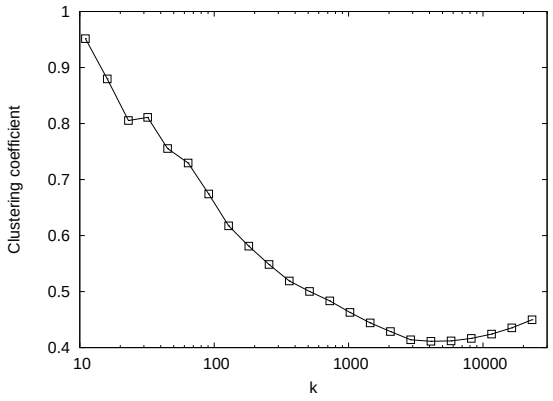

Figure 5. Clustering coefficient of $\mathcal{G}_{W(k)}$ as a function of k.

some point, when the vocabulary is large enough for bootstrapping, they should become connected. Yet this simple expectation turned out to be completely wrong, as the first feature that became immediately apparent was that $\mathcal{G}_{W(k)}$ is connected even for very small values of k. This is a consequence of the fact that functional words such as *in, for, the, of,* etc. occupy the top of the frequency list, and they appear in essentially all definitions. So one has to turn to other characterizations of the graph topology.

Since $\mathcal{G}_{W(k)}$ is connected, its diameter is finite, and we can investigate how it varies as a function of k. Recall that the diameter of a connected graph G is the maximum distance between two vertices, where by the distance $d(u, v)$ between two vertices u and v we mean the number of edges in the shortest path linking u and v. The diameter is thus defined as

$$\mathrm{diam}(G) = \max_{u,v \in V(G)} d(u, v), \tag{8}$$

where $V(G)$ is the set of vertices of G. As shown in Figure 4, the diameter grows approximately linearly with the logarithm of the number of vertices – strikingly similar to what one observes in classical random graphs [2]. The diameter, therefore, does not reveal any structural change in the topology of $\mathcal{G}_{W(k)}$ as k increases. We will have to turn to yet another quantity characterizing complex networks, namely the clustering coefficient.

5 Clustering coefficient

Originally introduced in [14], the clustering coefficient represents the average probability that two neighbours of a given vertex are also a neighbour of one another. More formally, given a vertex v of a graph G, the local clustering coefficient is defined as

$$c_v(G) = \frac{\text{number of edges between neighbours of } v}{\binom{\deg(v)}{2}},$$

where $\deg(v)$ is the degree of v, that is, the number of edges connected to to v. Clustering coefficient can thus be understood as the ratio of the number of edges that exist in the neighbourhood of v to the maximum number of edges that could exist in that neighbourhood of v, which happen to be $\binom{\deg(v)}{2}$. The clustering coefficient $c(G)$ of the whole graph G is then defined as the average of $c_v(G)$ over all vertices v belonging to G.

We computed the clustering coefficient for $\mathcal{G}_{W(k)}$ and plotted the result as a function of k, as shown in Figure 5. The interesting feature of this graph is the fact that the clustering coefficient initially decreases with the growing vocabulary. When the vocabulary reaches about 4000 words, the trend reverses, and the coefficient starts increasing again. The point at which the clustering coefficient reaches its minimum will be called a *clustering threshold*, defined as $T_c \approx 4000$.

This behavior of $c(\mathcal{G}_{W(k)})$ can be explained as follows. Initially, new words which are being added to the vocabulary "probe" new (i.e., previously not covered) regions of $\mathcal{G}$, and they remain relatively far from each other, since the graph $\mathcal{G}$ is rather large. Since each consecutive word has a smaller frequency than the previous one, it contributes smaller number of links to the clustering coefficient than the previous word. Therefore, $c(\mathcal{G}_{W(k)})$ initially decreases. When T_c is reached, the number of words is large enough that all vertices of $\mathcal{G}$ are in a close proximity of a known word. New words no longer "discover" new areas of $\mathcal{G}$, but rather end up in the proximity of previously known words, thus increasing the clustering coefficient. Reaching T_c, therefore, is equivalent to covering all important areas of the "semantic space", and further increase of T_c corresponds to obtaining finer and finer coverage of that space.

Of course, it is quite remarkable that T_c coincides with the linguistic threshold T_l and the Zipfian threshold T_z - all of them appear to be around 4000 words.

6 Conclusions and future directions

We presented a compelling evidence that the process of language acquisition is strongly nonlinear, exhibiting a thresh-

old which can be observed in several aspects of the process, including text coverage, rank-frequency distribution of the vocabulary, as well as the topological structure of the dictionary. Clearly, the presented evidence is not conclusive, but the authors hope that it may stimulate further research in this field. One possible approach, which is currently under investigation, is to use some other types of graphs representing vocabulary, especially graphs where semantic relationship between words are more precisely defined than those in the dictionary-based graph $\mathcal{G}$. Such semantic networks have been constructed for several languages, and they could easily be used to generate graphs in which edges representing relationships such as synonyms, hypernyms, hyponyms, etc.

We should also point out that although the change in the topology of subgraphs of $\mathcal{G}$ has been observed at T_c, representing minimum of the clustering coefficient, it is not entirely clear how it is related to the concept of "self-hosting" or bootstrapping. We plan, therefore, to construct a different graph, with directed edges from the word being defined to the word occurring in the definition. If one starts removing bottom-ranking words from the graph, not much should change at first - most entries will be defined in terms of entries still remaining. But at some point, one expects that too many entries will contain unknown words - corresponding to the loss of "bootstrapping" ability. This work is currently in progress and will be reported elsewhere.

7 Acknowledgements

The authors acknowledge financial support form the Natural Sciences and Engineering Research Council of Canada in the form of the Discovery Grant (H.F.) and the USRA award (C.P.).

References

[1] F. Bellard. Obfuscated tiny C compiler. http://fabrice.bellard.free.fr/otcc/.

[2] S. Bornholdt and H. G. Schuster, editors. *Handbook of Graphs and Networks*. Wiley-VCH, Weinheim, 2003.

[3] J. B. Carrol, P. Davies, and B. Richman. *The American Heritage Word Frequency Book*. Houghton Mifflin, New York, 1971.

[4] American National Corpus. http://americannationalcorpus.org.

[5] A. de Jesus Holanda, I. Torres Pisa, A. Souto Martinez O. Kinouchi, and E. E. Seron Ruiz. Thesaurus as a complex network. *Physica A*, 344:530–536, 2004.

[6] R. Goulden, P. Nation, and J. Read. How large can a receptive vocabulary be? *Applied Linguistics*, 11:341–363, 1990.

[7] Project Gutenberg. The gutenberg webster's unabridged dictionary. http://www.gutenberg.org/etext/673.

[8] R. Ferrer i Cancho1 and R. V. Solś. The small world of human language. *Proc. Roy. Soc. Lond. B*, 268:2261–2265, 2001.

[9] B. Laufer. How much lexis is necessary for reading comprehension? In P. J. L. Arnaud and H. Béjoint, editors, *Vocabulary and Applied Linguistics*, pages 126–132. Macmillian, London, 1992.

[10] O. Lecarme, M. Pellissier Gart, and M. Gart. *Software Portability*. McGraw-Hill, 1986.

[11] M. A. Montemurro. Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*, 300:567–578, 2001.

[12] I. S. P. Nation. *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge, 2001.

[13] C. Tsallis, G. Bemski, and R. S. Mendes. Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Phys. Lett. A*, 257:93–98, 1999.

[14] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[15] G. K. Zipf. *Human Behavior and the Principle of least Effort*. Addison-Wesley, 1949.