Statistics of Morphological Finite-State Transition Networks Obey the Power Law

Alexander Troussov, Brian O'Donovan

IBM Dublin Software Lab, Airways Ind. Est., Cloghran, Dublin 17, Ireland
{atrousso, Brian_ODonovan}@ie.ibm.com

Abstract

Finite-state devices are widely used in natural language processing, yet little if anything is known about metrics and topology of finite-state transition graphs. Here we study numerically the structure of directed state transition graphs for several types of finite-state devices representing morphology of 16 languages. In all experiments we have found that distribution of incoming and outcoming links is highly skewed and is modeled well by the power law, not by Poisson distribution typical of classical random graphs. The power-law form of degree distribution is regarded as a signature of self-organizing systems, and it has been previously found for numerous real world networks in communication, biology, social sciences and economics.

1 Introduction

Finite-state devices, including finite-state automata and transducers, are widely used in natural language processing to produce morphological information. Constructed as applications of formal finite-state techniques, they can be considered as networks where nodes represent states and arcs (labeled by characters) represent the transitions. Examination of their graph-metrics and topology is essential for efficient computer implementation of finite-state processing, including per-node optimization. It might also lead to new quantitative methods in language typology as we argue below.

In computational linguistics semantic and co-occurrence networks were already studied. In these networks nodes correspond to words. In semantic networks the links show semantic relations between words. In co-occurrence networks links represent the fact that words occur beside each other in a corpus. We are not aware of similar investigations applied to finite-state transition networks, representing language morphology. In [Leslie 1995] the average out-degree of (random, non-deterministic) automata is shown to be a good predictor for the expected number of states in the determinized automaton, the same technique is used in [van Noord 2000].

In the Introduction we remind the basics of finite-state processing in morphological applications and provide the rationale – why applying of modern random networks theory might be of interest for applications in finite-state processing.

In the second section – Random Networks and Related Work – we briefly outline methods and results of this relatively new theory to identify which of them are related to the study of finite-state devices. We argue that one particular metric studied for random networks - degree distribution – is of special interest for the initial investigation.

In the third section we describe the morphological data used in our experiments, and in the fourth - our cross-linguistic experimental study of the degree distribution, which we have found well approximated by the power-law.

In Discussion we put forward additional considerations about consequences of power-law behavior in view of our experiments.

1.1 Finite-state devices used in morphology

In our experiments we analyzed two major types of finite-state devices, used in natural language processing for word verification and producing morphological information. In both devices word verification is regarded as a process of moving from an initial input state to an acceptance state in a space of character transitions.

Finite-state automata. The input list of words (surface forms), is compiled into a letter tree, which is then minimized to reuse common postfixes. Each word can be loaded with additional information (its part-of-speech categories, etc.), which can be attached to the leaves (the terminals) of the letter tree. In this case two postfixes can be merged only if they lead to exactly the same information. Finite-state automata (FSAs) constructed this way, are acyclic and deterministic (for each state and each character there can be only one or zero output links labeled by this character).

Lexical transducers. In our experiments we also analyzed IBM lexical transducers that implement two-level morphology rules. Some of them have cycles and are non-deterministic.

1.2 Transition networks of finite-state devices

All finite-state devices considered here are finite-state transition networks and can be viewed as the compact description of morphology in the form of algorithm. If information about conditions and/or intentional descriptions of the transitions is ignored, we are left with a graph, which in case of deterministic devices exactly represents control flow of the algorithm.

These transition networks are easily large enough to be used for the investigation of power-law effects. They range in size from 31,219 to 429,577 nodes. This is similar in size to networks described in [Albert and Barabasi 2001].

1.3 Possible outcomes from the study of structure of finite-state transition networks

Applications to computer implementation. Programming structures, used to navigate through a finite-state net, have different performance and different memory requirements. For example, deterministic finite-state devices are most efficiently implemented with transition tables that enable rapid selection of links between states, where these links are stored in an array indexed by characters from the input language. However, efficiency comes at the expense of considerable memory overheads. Thus, the efficiency of the usage of transition tables depends primarily on the distribution of nodes according to the traffic they experience during corpora processing. Transition tables can be efficiently used if this distribution is a fat-tailed distribution like power-law distribution, so that a small number of nodes are responsible for a significant portion of traffic.

The paper Troussov et al. (2003) provided classification of nodes according to their traffic-related role in an FSA and a formal procedure for assigning a format to each node based on this role. However, this optimization model was based on the heuristic assumptions about the structure of dictionaries. Here we validate most of these assumptions.

Linguistics application. Comparison of the full-form lexicon given as an annotated list of word forms with its FSA representation in case of Indo-European languages, suggests that the latter form is more suitable for extracting implicit morphological information. For example, [Daciuk 1998] provides algorithms for morphological FSAs with the aim to "*generalize the knowledge contained in the lexicon so that accurate prediction of morphological information for unknown words be possible", "little or no linguistic knowledge is required for that process".*

Graphmetrics and topology of transition networks might bear the following features, needed for developing new quantitative approaches in computational linguistics:

- generic language feature, which might be used to characterize and detect intelligent language like features in an input signal (see for example [Israeloff et al. 1996], [Elliott et al. 2000]);
- language specific features, which might give way to new quantitative approaches for morphological phenomena and for language typology.

Zipf's law is considered as an indication of language-like behavior. Indicators, measured on finite-state transition networks – (which result from compilation of the lexis of a language into a minimized FSA) – probably, provide a more powerful formalism, because they involve not only frequencies of words, but also wordformation processes.

2 Random Networks and Related Work

2.1 Graph theory and random graphs

In the theory of random graphs, the simplest and the most intensively studied one-vertex characteristic is degree. The in-degree, k_i, is the number of incoming arcs of a vertex in an directed graph. The out-degree, k_o, is the number of its outgoing arcs. The degree, k, is the number of nearest neighbors of a vertex. For directed graphs the vertex degree is the total number of its connections $k = k_i + k_o$ (this holds true under additional condition that there are no arcs with both start and end point in this node).

Total distributions of vertex degrees of an entire network, $P(k_i;k_o)$ - the joint in- and out-degree distribution, $P(k)$ - the degree distribution, $P_i(k_i)$ – the in-degree distribution, and $P_o(k_o)$ - the out-degree distribution - are basic statistical characteristics of random networks.

Random graphs were first studied in the late 1950s by Erdös and Rényi. In general terms, a random graph is a graph in which properties such as the number of nodes, edges, and connections between them are determined in a random way. In Erdös and Rényi simplest classical model the graph has a fixed number of vertices, which are connected, at random, by edges. The degree in classical random graphs follows binomial distribution which for large N can be replaced by a Poisson distribution $P(k) = e^{-\lambda} \lambda^k / k!$.

2.2 Random massive networks

During the last decade random networks became an interdisciplinary area of research with a strong influence from statistical physics. Empirical and theoretical studies were applied to numerous real world (both natural and man made) networks in communications, biology, social sciences and economics. Examples in-

clude telephone networks, WWW, Internet, ecological networks (food webs), citation networks, co authorship networks.

In [Dorogovtsev et al. 2002] networks as objects of statistical physics are classified into:

1. equilibrium vs. nonequilibrium (for example, classical random graphs of Erdös and Rényi are equilibrium nets, WWW or citation index – nonequilibrium).
2. according to the form of degree distribution (distribution function $P(k)$, which gives the probability that randomly selected node has exactly k edges): rapidly decreasing degree distributions vs. fat-tailed degree distributions.
3. uncorrelated or correlated networks.

Standard indicators or measurements that characterize the structure of a graph are:

- The statistical distribution of links (characterizing homogeneity and scaling properties of the graph);
- The mean or maximum intervertex distance (i.e. the length of the shortest path between a pair of vertices), giving an idea of its size, or diameter;
- The clustering index (a measure of independence of neighboring links).

The following less frequently used characteristic is important for finite-state transition networks (as we consider them as control flow networks):

- the traffic (the number of trajectories passing through each vertex or arc, and so identifying the most active hubs).

The following phenomena were found in many real networks (see [Albert and Barabasi 2001]):

- Small path length (known as the small-word concept due to Stanley Milgram's famous conclusion that, there is a path of acquaintances between most pairs of people in USA with typical length of about six);
- Large degree of clustering;
- Power-law tail degree distribution (scale-free concept).

2.3 Applications of random networks to computational linguistics

Methods of random networks theory were already successfully applied to the study of lexical-semantic resources like WordNet - a database of word meanings with basic semantic relations between them, such as synonymy, hyponymy etc. See, for example, ([Motter et al. 2002], [Sigman and Cecchi 2002] and [Ferrer and Solé 2001]). The major focus was on the small-world concept – despite the large network size, the distance among any pair of nodes is relatively small. Degree distribution was found to follow the power-law, and its positive correlation with word frequencies was demonstrated. Studies of co-occurrence networks [Dorogovtsev and Mendes 2001a] also showed highly skewed link distributions.

Finite-state devices are efficient computational tools, and the theory of finite-state processing is algorithmically oriented. We are not aware of papers where finite-state transitions graphs were investigated in the framework of modern random network theory. One of the reasons behind this, is that it is probably difficult to consider such networks as "real" ones, which might have crosslinguistic underlying structure common for different types of finite-state devices and governed by relatively simple laws.

In this paper we consider morphological finite-state transition networks as equilibrium nets, because their formal construction method doesn't provide us with the idea how they grow up. The initial investigation is focused at the degree distribution for the following reasons:

1. Although the degree of a vertex is a local quantity, degree distribution often determines some important global characteristics of random networks (see [Dorogovtsev and Mendes 2001]).
2. Standard definitions of intervertex distance and of clustering coefficient are more suitable for undirected nets; while it seems that directedness is important characteristic of the computational process provided by finite-state transitions networks.
3. Metric structure, determined by intervertex distances, highly relevant to the study of semantic relations and polysemy, seems a priori less relevant to morphology compared to network dynamics (described by degree distribution) and topology.

The paper by Albert and Barabasi (2001) states, that most network models studied before ignore the network's directedness and generic features of complex directed networks are not fully investigated.

3 Description of Experimental Data

3.1 Morphological data

Finite-state transition networks of IBM morphological dictionaries were used for experiments. All of the dictionaries provide lexical coverage sufficient for spell-checking. They also provide fine-grained morphosyntactic information. For the purpose of this paper, it is necessarily to describe the types of glosses in the dictionaries, because this directly affects topology of the network through minimization stage which eliminates some nodes.

All dictionaries provide inflectional morphology. Derivational morphology is provided only for Finnish.

3.2 Description of the dictionaries

In full-form lexicons, used as the source word lists for FSAs construction, all case variations are explicitly presented.

Germanic languages: English, German, Dutch, Swedish, Norwegian, Danish. Dictionaries contain word formation elements used for compounding (save for English), e.g. in German dictionary we have *Schul-* as an allomorph of the noun *die Schule* when used in initial or middle position of a compound.

Romance languages: French, Italian, Spanish, Portuguese. Clitics are present in the dictionaries.

Ideographic languages: Chinese traditional and simplified. Chinese FSAs are compiled from the lists of words provided with glosses: part-of-speech and "frequencies". The frequencies are used for statistical word segmentation because Chinese language has no orthographic word boundaries. Implementation is complicated by the fact that Chinese is an ideographic language with a repertoire of thousands of characters. In [Goetz and Wunsch 2001] binary search was suggested for implementation of finite-state devices for ideographic languages. We use the polymorphic node structure suggested in [Troussov et al. 2003] .

Lexical transducers. Languages: Finnish, Turkish, Czech, Polish, Thai. Thai dictionary contains words and collocations and is used for word segmentation. Other languages provide inflectional and derivational morphology based on two-level morphology rules.

4 Experimental work

Degree distribution

We studied separately distribution of in-, out- and full-degree. In all our experiments, performed for two types of deterministic finite-state devices representing morphology of 16 languages, we have found that degree distribution in all state transition graphs that were considered is well approximated by the power-law. The random variable x is said to be distributed according to the power-law with the exponent γ if its probability density function satisfies $P(x) \sim x^{-\gamma}$.

A power-law implies that small events are common, whereas large events are rare. In linguistics, such distribution appears in Zipf's law, other instances of power-law are Gutenberg-Richter distribution of earthquake magnitudes, Kolmogorov's law in turbulence, Pareto's law in economics. [Adamic] exposes relations between different forms of power-law.

The usual way to fit power-law distribution is to perform a linear regression on a log-log plot of the cumulative distribution function. Fig. 1 and Appendix 1 provides some samples of log-log plots for degree distribution, exponents are given in the Table 1.

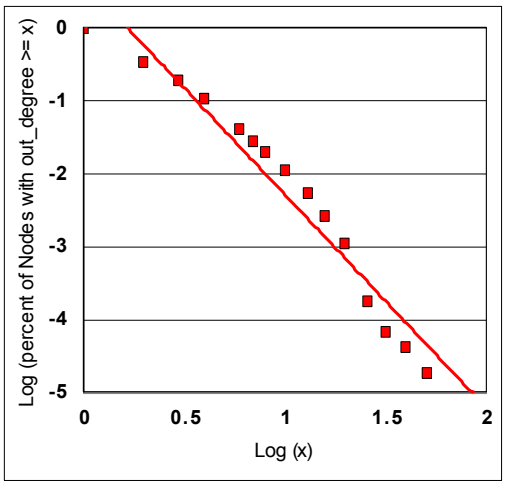


Fig. 1. The log-log (base 10) plot of power-law approximation of the out-degree distribution for English with data binned into exponentially expanding bins so that they will appear evenly spaced on a log scale (the method is discussed in [Adamic]).

Table 1. The exponents of power-law approximation of the distribution of in- and out-degree found for our experimental data. Quantification was done with the data binned into exponentially expanding bins (the method is discussed in [Adamic]).

Language	Finite-state device	In-Degree exponent	Out-Degree exponent
Chinese simplified	FSA	2.47	2.53
Chinese traditional	FSA	2.55	2.56
Czech	FST	2.44	2.69
Danish	FSA	2.62	3.19
Dutch	FSA	2.71	3.35
English	FSA	2.52	3.18
Finnish	FST	2.37	2.35
French	FSA	2.50	3.38
German	FSA	2.67	3.32
Italian	FSA	2.43	3.14
Norwegian	FSA	2.56	3.17
Polish	FST	2.56	2.96
Portuguese	FSA	2.42	3.20
Spanish	FSA	2.45	3.20
Swedish	FSA	2.58	3.09
Thai	FST	2.82	2.95

Traffic

Nodes of a finite-state device experience traffic when the device is used to process corpora. The traffic is defined as a the number of visits to a node while processing a corpus. Fig.2 shows that the distribution

of traffic for English dictionary is Zipf-like distribution.

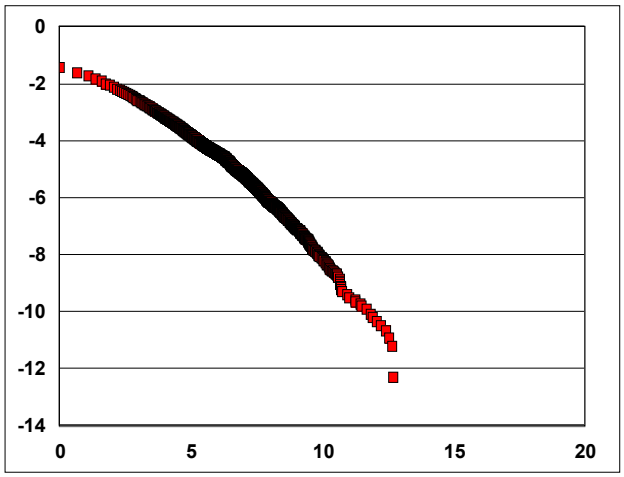


Fig. 2. The log-log (natural log) plot of cumulative distribution function for nodes traffic in English dictionary.

5 Discussion

Fitness of the degree-distribution to power-law and other approximations

Our experiments demonstrate, that the in-degree and full-degree distribution of finite-state transition networks are well approximated by the power-law. In-degree in all our experiments cleanly fit a power-law over about three decades. Empirical distribution is considered to exhibit a regime of power-law decay if log-log plot fits visually to a strait line at least for two decades preferably on both axes (see e.g. [Dorogovtsev and Mendes 2003]).

Deterministic finite-state transition nets, considered in this paper, have finite node's capacity of the out-flowing links, because these links are uniquely labeled by characters from the dictionary alphabet. For example, English dictionary contain 96 unique characters (including case variations). However, out-degree degree distribution also demonstrate limited power-law regime, which is quite different from the Poisson distribution typical for classical random graphs.

Many of real massive networks studied recently show power-tail degree distribution. Some other networks have exponential or a coherent mixture of the power law and exponential degree distribution. In our plots there are deviations from the power-law, i.e. curvature in the log-log plots, which is typical for many other distributions regarded as well approximated by the power law. However, search for other models is necessary for linguistic interpretation of parameters of the distribution (e.g. to say that there is the difference in degree distribution for Germanic and Romance languages).

Consequences of power-law behavior

Peaked distributions characterizing random networks have a strong central tendency, that is a tendency to cluster around some particular value. In such cases it may be useful to characterize the distribution by a few numbers that are related to its moments, including the mean (or by alternative estimators – median and mode). For example, in the height of human individuals the ratio between the tallest human and the average is less than 2. In systems exhibiting power law distribution, the mean and median are typically not very useful.

Physicists attribute the property of self-organization (self-optimization) to networks, whose links follow the power-law. "*Power laws are considered as one of the most striking signatures of complex self-organizing systems*" [Laherrere and Sornette 1998]. Preferential linking (vs. random linking) is one of the mechanisms that explains such self-organization [Albert and Barabasi 2001]; in this model the likelihood of receiving new edges increases with the node's degree.

We can therefore expect persistence of a self-organized structure in the morphological finite-state transition networks. The structure may emerge due to:

a) generic algorithmic nature of finite-state processing (caused by determinization and minimization procedures);

b) generic similarity in the source data used for construction of finite-state transition graphs – i.e. properties of writing systems, morphological properties of languages, distribution of word length and the distribution of letter frequencies.

From our results it is difficult to correlate structural characteristics with language typology because the content of the source is not homogeneous (e.g. word-formation elements are present in the dictionaries of Germanic languages, but there are no wordformation elements for French or Spanish). Word lists compiled from corpora (with or without part-of-speech annotation) might be more suitable for finding such correlations.

6 Conclusions and Future Work

Experiments, made on the morphology of 16 languages, represented by several types of deterministic finite-state devices, show that an important structural property of finite-state transition graphs – degree dis-

tribution – is well approximated by the power law. Such behavior is considered in statistical physics as a typical signature of self-organizing systems.

Our experiments are not sufficient to correlate structural properties of state transition graphs with language typology due to non-homogeneity of the source data. Finding such correlations might eventually give way to new quantitative approaches in computational linguistics. Consideration of a single structural property – the degree distribution – is probably insufficient.

Future works:

- use homogeneous source data like word lists compiled from corpora (with or without part-of-speech annotation) and represented by deterministic minimized FSAs;
- to correlate graph theoretical metrics with traffic;
- to correlate structural characteristics and traffic with typological language characteristics (Does the topological complexity of the finite-state transition graph correlate with the morphological complexity of the represented language?)
- describe the formation of state transition graphs in terms of their determinization/minimization and in terms of lexica growth (e.g. during first or second language acquisition) to find similarity with other non-equilibrium networks for possible hypothesis about how state transition graph becomes specifically structured.

7 References

[Adamic] Adamic, L. Zipf, power-laws, and Pareto–a ranking tutorial. http://www.parc.xerox.com/istl/groups/iea/papers/ranking/ranking.html

[Albert and Barabasi 2001] Albert, R., and Barabasi, A.-L. Statistical Mechanics of Complex Networks. *cond-mat/0106096* (June 2001). http://arxiv.org/abs/cond-mat/0106096

[Daciuk 1998] Daciuk, J. Incremental construction of finite-state automata and transducers, and their use in the natural language processing. Rozprawa doktorska ETI, (1998)

[Dorogovtsev and Mendes 2001a] Dorogovtsev, S.N., and Mendes, J.F.F. (2001). Language as an evolving word web. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1485):2603–2606. http://arxiv.org/abs/cond-mat/0105093

[Dorogovtsev and Mendes 2003] Dorogovtsev, S.N., and Mendes, J.F.F. Evolution of networks. Oxford U. Press, New York, 2003.

[Dorogovtsev et al. 2002] Dorogovtsev, S.N., Mendes, J.F.F., and Samukhin, A.N. Modern architecture of random graphs: constructions and correlations. arXiv:cond-mat/0206467 v1 24 Jun 2002. http://arxiv.org/abs/cond-mat/0206467

[Elliott et al. 2000] Elliott, J., Atwell, E., and Whyte, B. Increasing our Ignorance of Language: Identifying Language Structure in an Unknown 'Signal'. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, (2000) pp. 25–30. acl.ldc.upenn.edu/W/W00/W00-0705.pdf

[Ferre and Solé 2001] Ferrer I Cancho, R. and Solé R.V., The small world of human language, Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1485):2661–2665. www.santafe.edu/sfi/publications/Working-Papers/01-03-016.pdf

[Goetz and Wunsch 2001] Goetz, T., Wunsch, H. An Abstract Machine Approach to Finite State Transduction over Large Character Sets. Finite State Methods in Natural Language Processing 2001. ESSLLI Workshop, August 20-24, Helsinki. http://www.research.ibm.com/IE/publications/goetz_wunsch.pdf

[Israeloff et al. 1996] Israeloff, N. E., Kagalenko, M., and Chan, K. Can Zipf Distinguish Language From Noise in Noncoding DNA? Volume 76, Number 11, Physical Review Letters, 11 March 1996.

[Laherrere and Sornette 1998] Laherrere, J., and Sornette, D. Stretched Exponential Distributions in Nature and Economy: Fat tails with characteristic scales. European Physics Journal B, 2 (1998), pp.525–539. http://arxiv.org/PS_cache/cond-mat/pdf/9801/9801293.pdf

[Leslie 1995] Leslie, Ted. 1995. Efficient approaches to subset construction. Master's thesis, Computer Science, University of Waterloo.

[Motter et al. 2002] Motter, A. E., de Moura, A. P. S., Lai, Y.-C., and Dasgupta, P. Topology of the conceptual network of language. Physical Review E, 65, 065102, (2002) cactus.eas.asu.edu/partha/Papers-PDF/ 2002/PhyReview.pdf

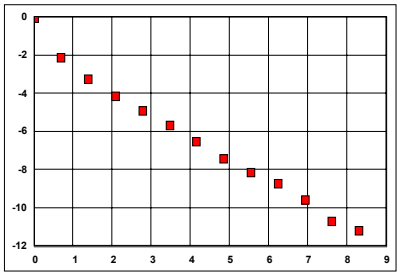
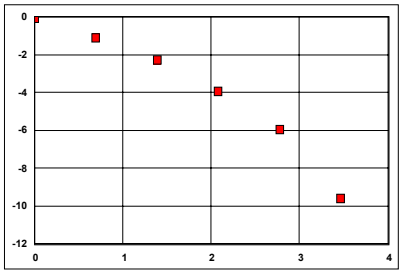
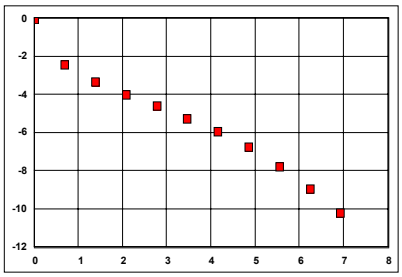
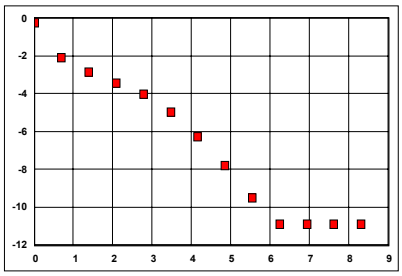
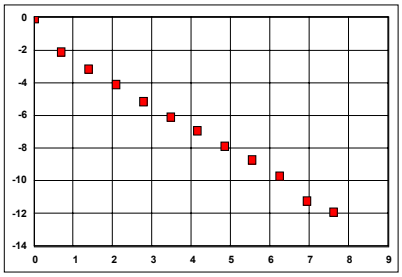
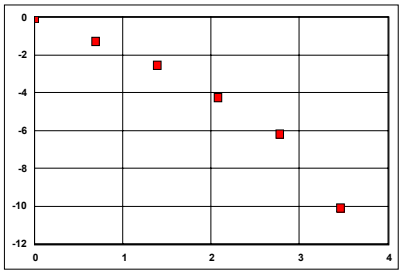
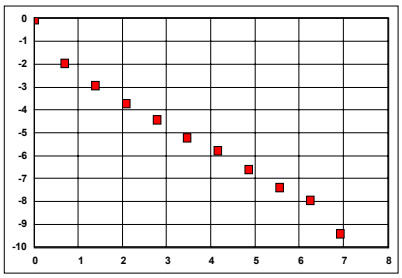
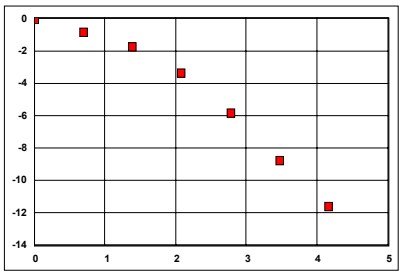
[Sigman and Cecchi 2002] Sigman, M., and Cecchi, G.A. Global organization of the lexicon. cond-mat/0106509. asterion.rockefeller.edu/guille/Papers/lexicon.pdf

[Troussov et al. 2003] Troussov, A., O'Donovan, B., Koskenniemi, S., and Glushnev, N. Per-Node Optimization of Finite-State Mechanisms for Natural Language Processing. Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science N 2588, Springer-Verlag, (2003), pp.223–227. http://brianodonovan.tripod.com/fsopt.pdf

[van Noord 2000] van Noord, Gertjan. Treatment of Epsilon Moves in Subset Construction. Computational Linguistics 26(1): 61-76 (2000)

Appendix 1. Samples of In- and Out-Degree Distribution

Table 2. In- and out-degree distributions for 6 sample languages (out of 16 examined ones): 5 languages, whose morphology is represented by FSAs (English, Chinese as an ideographic language, Germanic language and Romance language) and Finnish – an agglutinative language, whose morphology is represented by a lexical transducer implementing two-level morphology. Shown are log-log (natural log) plots of the cumulative distribution function for in- and out-degree distributions, with data binned into exponentially expanding bins (the method is discussed in [Adamic]).

	In-degree	*Out-degree*
English. FSA.		
Chinese traditional. Ideographic language, FSA.		
German. FSA.		
Spanish. FSA.		
Finnish. Agglutinative language, finite-state transducer.