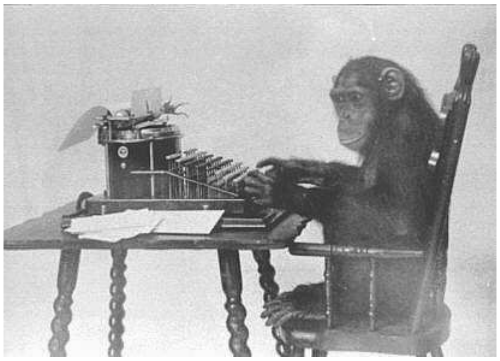
Not So Randomly Typing Monkeys –
Rank-frequency Behavior of Natural and Artificial Languages

Algorithms for Information Networks – Project report

Andreas Krause and **Andreas Zollmann**



Abstract

Power laws arise through many natural processes. Zipf showed that the frequencies of words, as they appear in Shakespeare's *Hamlet*, follow a power law distribution. Mandelbrot explained this effect as a result of an underlying information-theoretic optimization problem. Miller invoked doubt by showing that a very simple mechanism could also explain the presence of power laws: A monkey typing words with uniformly and independently selected letters would also produce word frequencies following a power law. In consequence, several other researchers proposed and investigated rank-frequency distributions of randomly generated text.

In this paper, we first present a literature overview over this exciting topic. We then propose a class of Hidden Markov Models (HMMs) which generalizes the models previously investigated, generating power law, log-normal and other behavior. We extend a result of Conrad and Mitzenmacher for computing the power law exponent of zero order Markov processes to a setting which captures random walks in d-regular graphs. In an extensive empirical evaluation, we investigate convergence of rank frequency distributions for randomly generated text to those of natural language corpora, for increasing orders of Markov Processes and HMMs with an increasing number of hidden states. Our analysis uses four real-world corpora: the Reuters corpus, Shakespeare's Hamlet, Goethe's Faust and source code of the Linux kernel.

1 Introduction

Zipf's law is a mathematical model of the relationship between the frequency and frequency-rank of words in natural language text. Estoup [4] and Zipf [15] independently observed this relationship after manually counting the frequency of each word in a large amount of text. They discovered that the frequency f_i of the ith ranked word (in terms of frequency) is roughly given by the formula $f_i = ci^{-\alpha}$, for some parameter $\alpha > 0$ and normalizing constant c, which is now know as Zipf's law, or power law with exponent α.

Apart from word frequencies, power laws have been found to accurately model a wide range of real-world phenomena, including the distribution of incomes [1] [12], city sizes [16] and the structure of the internet [5]. There is no a priori reason why these phenomena should exhibit power law behavior. Hence, in an effort to obtain a deeper understanding of these phenomena, there is a significant amount of literature devoted to finding an explanation of Zipf's law along the lines of the Central Limit Theorem for the normal distribution. In this paper, we review and generalize some of this literature, concentrating on the aforementioned rank-frequency distribution of words.

1.1 Literature overview

The first explanation of Zipf's law was offered by Zipf himself [16]. He argued that authors want to minimize the length of text required to communicate an idea, even if this introduces ambiguities. On the other hand, readers want to minimize the effort required to understand (disambiguate) the text. Zipf claimed that this trade-off, called the *principle of least effort*, leads to Zipf's law. Although his arguments were not mathematically rigorous, recent empirical evidence [6] supports Zipf's hypothesis.

Zipf's explanation was superseded by Mandelbrot [8], who used information theory to formally derive a process generating Zipf's law. Mandelbrot regarded language design as an optimization problem, in which the aim is to maximize the average amount of information communicated per average word cost. Suppose the ith most probable word w_i occurs with probability p_i. Then the information which words

[1] Actually, Pareto studied the distribution of incomes before Zipf's law was discovered. See [1] for details.

communicate is given by the entropy of the generating process, namely $-\sum_i p_i \lg p_i$. Generally, we want more frequent words to have a smaller cost (length), since they are used more often. Mandelbrot solved the optimization problem by encoding each word w_i by the integer i, which has cost $\log_n i$, where n is the size of the alphabet.

Mandelbrot's result suggests that the evolution of the English language is influenced by an underlying optimization process. Although this appears to be a compelling explanation, a subsequent paper by Mandelbrot [9] hinted that his own optimization process had a much simpler explanation. In this paper, Mandelbrot describes a Markov model for generating random text by way of a thought experiment, in which a monkey randomly types on a keyboard with n letters (the alphabet) and one space bar, where each key is pressed independently with probability inferred by the above optimal encoding. Mandelbrot proved that the random text generated by the monkey is asymptotically distributed according to Zipf's law. Miller [10] considered a special case of Mandelbrot's experiment in which the monkey hits the keys *uniformly at random*. Although there is no underlying optimization in this experiment, the limiting behavior is still Zipf's law. The reason for this is purely statistical - the number of words of length l, which is n^l, increases exponentially in l, while the probability that such a word occurs, which is $\left(1 - \frac{1}{n+1}\right)^l \frac{1}{n+1}$, decreases exponentially in l. A simple calculation shows that it is the trade-off between these two properties that leads to Zipf's law. Miller concluded that Zipf's law can be derived "*without appeal to least effort, least cost, maximal information, or any branch of the calculus of variations*".

Perline [13] considered the monkey-at-a-typewriter experiment with fixed word size l and an arbitrary probability distribution over the letters. This model leads to a log-normal distribution in the following way. Let p_i be the probability that the monkey types the ith letter of the alphabet, and let $X_j = p_i$ if this is the jth letter of the current word. Then $Y = X_1 X_2 \ldots X_l$ gives the probability that the monkey types a randomly chosen l-letter word. Since the X_i's are i.i.d and $\ln Y = \sum \ln X_i$, $\ln Y$ converges to the normal distribution by the Central Limit Theorem, and so Y converges to the log-normal distribution. Perline also extended this argument to hold for words up to some fixed size l.

Although this result holds for any fixed l, Mitzenmacher [11] notes that this sequence of log-normal distributions need not converge to a log-normal distribution. In fact, under weak assumptions on the probability distribution of letters, this sequence converges to the Zipf's law [3].

Markov processes for generating random text have been considered before, e.g. by Mandelbrot [9] as mentioned above, and by [7] who considered the case of a particular two-parameter family of Markov processes. In [14], Shannon hypothesized that text generated by Markov processes of increasing order estimated from natural language would increasingly resemble natural language text. To the best of our knowledge, nobody derived results about rank-frequency distributions for random text generated by arbitrary Markov models. Furthermore, no class of models able to describe the process of generating random text both with bounded and unbounded word lengths has been proposed so far.

1.2 Organization and Contributions

In the remainder of this paper, we present a particular class of Hidden Markov Models as a generalization of the models described above. We show that these models can generate power laws, log-normal distributions and other behavior. Furthermore, we show how to explicitly compute the power law exponent for a specific class of Markov processes which captures the case of random walks in d-regular graphs.

In an extensive empirical evaluation, we investigate convergence of rank frequency distributions for randomly generated text by Markov processes of increasing order to those of natural language corpora, aiming at verifying Shannon's theory. We also investigate similar convergence effects for HMMs estimated from natural language text, where the number of hidden states is gradually increased. Intuitively, with HMMs we have much finer control over the number of parameters involved in the estimation, and hence we would expect a slower and more detailed picture about the convergence of the rank-frequency distributions. The reason for this is that the number of parameters grows exponential in the order of the Markov process, whereas they only grow polynomially in the number of hidden states of the HMM.

Motivated by the hypothesis that languages consist of a repertoire of structural words as well as an ever growing collection of content words, along with the assumption that structural words have short length, we analyze mixtures of processes. In these mixtures, one component generates text of bounded length, and the other component generates text of unbounded length. We show that the rank frequency distributions of these mixtures correspond more closely to the observed natural rank-frequency distributions. Our analysis uses four real-world corpora: the Reuters corpus of news articles, Shakespeare's *Hamlet*, Goethe's *Faust* and C source code of the Linux kernel.

2 A more general model

In this section we propose a more general framework for Miller's [10], Perline's [13] and Mitzenmacher's [11; 3] thought experiments. We will consider what happens if the key strokes are generated by a Hidden Markov Model. A pair of sequences of random variables $(X_t)_{t\in\mathbb{N}}$ over a finite set of states S and $(O_t)_{t\in\mathbb{N}}$ over a finite alphabet A and joint probability $\Pr$ is called a Hidden Markov Model (HMM) if

$$\begin{aligned}\Pr(X_{t+1} = x_{t+1} | X_t = x_t) &= \\ \Pr(X_{t+1} &= x_{t+1} | X_1 = x_1, \ldots, X_t = x_t)\end{aligned}$$

and the O_t are conditionally independent of all other variables given X_t for all $t \geq 1$. We will assume stationarity, i.e. $\Pr(X_{t+1} = y | X_t = x) = P(x, y)$ for all $x, y \in S$ and $\Pr(O_t = o | X_t = x) = Q(x, o)$ for all $x \in S$, $o \in A$ and $t > 0$. The HMM is fully specified by the transition matrix P, the emission matrix Q and an initial distribution ν over S. A sample from $\Pr$ is called a trajectory of the HMM. Let us associate a special element $\perp$ of both S and A with the blank character separating words, and define this state to be absorbing, i.e. $P(\perp, \perp) = 1$. We furthermore require that $\Pr(\inf_t X_t = \perp < \infty) = 1$, $Q(s, \perp) = 0$ for $s \neq \perp$ and $Q(\perp, \perp) = 1$, i.e. trajectories will almost surely reach the absorbing state, and from there only emit the character $\perp$. For any trajectory, we call the sequence of emissions $o_1, \ldots, o_t$ up to and not including the first occurrence of $\perp$ a word (note that this might include the empty word). This definition ensures that there is a one-to-one correspondence between trajectories of the HMM and words.

Consider how the examples in the literature fall into this framework. Let μ be a distribution on $A \setminus \{\perp\}$, and $p > 0$. The corresponding HMM will have two states, $S = \{1, \perp\}$ and we can associate with it the transition matrix

$$P = \begin{pmatrix} 1-p & p \\ 0 & 1 \end{pmatrix}$$

In the state 1, the emission probabilities will be $Q(1, a) = \mu(a)$ for $a \in A \setminus \{\perp\}$, and $Q(\perp, \perp) = 1$. If we let the initial distribution be $\nu_\mu = ((1-p), p)$, then the distribution over the words generated by this HMM is exactly the distribution of words generated by Miller's experiment (if μ is uniform) or Mitzenmacher's generalization.

Of course this model captures the case where the underlying process is specified by a Markov chain as considered in [9; 7]. In this case we set $S = A$ and each state s emits symbol s with probability 1. Such a chain can easily be transformed such that it has exactly one absorbing $\perp$ state.

More interestingly, the concept of HMMs allows us to formalize Perline's model [13], where letters are chosen i.i.d. at random, but only words of a fixed length l or of length up to l are considered. We can formulate this process as an HMM very similarly to the i.i.d. case, but instead of a single 1 state, we will have states $1, \ldots, l$. For the case of words with length exactly l, state $t+1$ follows state t with probability 1 for $1 \leq t \leq l-1$, and state l is deterministically followed by $\perp$. We furthermore require that no state $s \neq \perp$ may emit the $\perp$ symbol. For the case of words up to length l, we simply add an edge from states $1, \ldots, l-1$ to $\perp$ with probability $p > 0$.

2.1 When do we get power laws?

We want to identify properties under which the word frequency distribution of a given HMM follows a power law. In [3], Conrad and Mitzenmacher prove that in the i.i.d. case, power laws arise. In [7], Kanter and Kessler show that this is also the case for a very specific Markov chain controlled by two parameters. For the case of fixed word lengths, lognormal behavior arises [13].

It is tempting to try to generalize Conrad's and Mitzenmacher's proof to the case of Markov chains. The following observation indicates that this might be possible. In [11], Mitzenmacher presents a simple proof that for a two letter alphabet and unequal probabilities q and q^2, the word frequency distributions follow a power law. Here we present a similar proof, but for the case of a more complex Markov chain.

Let $A = S = \{1, 2, \perp\}$,

$$P = \begin{pmatrix} q & q^2 & 1-q-q^2 \\ q^2 & q & 1-q-q^2 \\ 0 & 0 & 1 \end{pmatrix},$$

where $q > 0$, $1 - q - q^2 > 0$ and let Q deterministically emit the current state. Furthermore choose $\nu = (1, 0, 0)$. Then every word w generated by the process has probability $q^k(1 - q - q^2)$ for some $k \geq 0$. Using the terminology from Mitzenmacher's proof, we call k the pseudo-rank of w. There is one word of pseudo-rank 0 ("a"), one of pseudo-rank 1 ("aa"), two words of pseudo-rank 2 ("ab" and "aaa") and a simple induction yields that there are F_{k+1} words of pseudo-rank k where F_k is the k-th Fibonacci number (where $F_0 = 0$ and $F_1 = 1$). This can be seen from the fact that the words with pseudo-rank k are generated from the words of pseudo-rank $k-1$ by appending the last character twice and from the words of pseudo-rank $k-2$ by appending the last character flipped. Using the identity $\sum_{i=1}^{k} F_i = F_{k+2} - 1$ it follows that when $F_{k+2} - 1 < j \leq F_{k+3} - 1$, the j-th frequent word has pseudo-rank k. Since $F_k = \Phi^k/\sqrt{5} + o(1)$ where $\Phi = (1 + \sqrt{5})/2$ is the golden ratio, for the frequency f_j it follows that

$$q^{\log_\Phi(\sqrt{5}(j+1))-2}(1-q-q^2) < f_j \leq q^{\log_\Phi(\sqrt{5}(j+1))-3}(1-q-q^2)$$

where we can directly read off bounds for the power law exponent.

The above observation leads to the following generalization of Conrad and Mitzenmacher's Theorem:

Proposition 1 (based on Theorem III-A from [3]). *Let $\mu_1 = (p_1, \ldots, p_{n-1}, p_\perp)$ be a distribution over A, where the n-th element is associated with the blank $\perp$. Furthermore, for $1 < j < n$ let $\mu_j = (p_{\pi_j(1)}, \ldots, p_{\pi_j(n-1)}, p_\perp)$ where π_j is a permutation of the integers $1, \ldots, n-1$, and let $\mu_n = (0, \ldots, 0, 1)$. Then the rank frequencies of the trajectories of the Markov process defined by*

$$P = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix},$$

with start distribution $\mu_0 = e_j$ (the j-th unit vector) for any $1 \leq j < n$ follow a power law distribution with explicitly computable power law exponent.

Proof. In [3], Conrad and Mitzenmacher prove the following statement. Let A be a multiset of positive integers $a_1 > \cdots > a_n$, each with multiplicity $w_1, \ldots, w_n$, and enumerate the elements as $\sigma_1, \ldots, \sigma_N$ where $N = \sum_i w_i$. For each real $\nu \geq 0$, let c_ν be the number of distinct N-tuples $(m_1, \ldots, m_N)$ of nonnegative integers, such that $\sum_i m_i \sigma_i \leq \nu$. Let D be the least common multiple of the denominators of the ratios a_i/a_1 which are assumed to be rational, and let $r = D/a_1$. Furthermore let $f(x) = \sum_i w_i x^{a_i}$ and x_0 be the unique positive solution of $f(x) = 1$. Define

$$A' = \frac{1}{r(1 - x_0^{1/r}) x_0 f'(x_0)}.$$

Then it holds that

$$\sum_{\nu \leq t} c_\nu \sim A' \cdot (1/x_0)^{\lfloor rt \rfloor / r}.$$

The proof of this Theorem by Conrad and Mitzenmacher is very involved, and involves techniques from complex analysis. They also extend this result to the case of irrational ratios, by passing to appropriate limits. After stating this general theorem, they specialize it to the case where c_ν counts the number of words with probability $p_1^\nu(1 - \sum_i w_i p_i)$, where each letter is struck independently of the previous letters. In this case, the a_i are simply defined by $1 = a_1 > \cdots > a_n$ where $a_j = \frac{\log p_j}{\log p_1}$.

To prove the generalization for permutated Markov processes, we only have to observe, that c_ν counts also the number of words with probability $p_1^\nu(1 - \sum_i w_i p_i)$, where each letter is generated according to this more general process. This can be seen by considering that word probabilities of words of length $n + 1$ are constructed by multiplying word probabilities of words of length n by some entry of the j-th row of the transition matrix P, where j is the respective n-th letter. As the base case, the unique word with length 0 (the empty word) always has probability $1 - \sum_i w_i p_i$. Hence, the assumption of independent key presses is not necessary; the fact that word probabilities are constructed by the same factors as in the independent case is sufficient. □

Note that this extension applies to the interesting special case of random walks in d-regular graphs, where the random walk ends with a probability p which is the same for every node in the graph. It remains true for the directed case, where each node has exactly d outgoing edges, and each node distributes probabilities $p_1, \ldots, p_d$ over all outgoing edges.

One might wonder whether all Markov models lead to power law behavior. Consider the following counterexample of a Markov Chain for which the distribution neither follows a power law nor a log-normal distribution. Let A, S, Q, ν be defined as above and

$$P = \begin{pmatrix} q & q^2 & 0 \\ 0 & q & q^2 \\ 0 & 0 & 1 \end{pmatrix},$$

where $q + q^2 = 1$. All words w generated by this process have probabilities $q^k \cdot q^3$ for $k \geq 1$. We again call k the pseudo-rank of w. In this case, a simple induction yields that there are exactly k words of pseudo-rank k. Hence when $k(k-1)/2 < j \leq k(k+1)/2$, the j-th frequent word has pseudo-rank k. We find

$$q^{(1+\sqrt{4j-1})/2+3} < f_j \leq q^{(1+\sqrt{4j+1})/2+3}$$

and hence we do not get power law behavior. In this case, the number of words with pseudo-rank k grows only polynomially in k.

3 Experiments

In this section we provide extensive evidence that rank frequency distributions from a variety of natural and artificial text corpora follow power law behavior.

Initially, in Section 3.1, we introduce the four natural text corpora used in our analysis. In Section 3.2 we reproduce the asymptotic results about randomly typing monkeys discussed in [10] empirically on finite data. We will then, in Section 3.3, investigate rank frequency behavior for several first order Markov models. In Section 3.4, we will discuss the convergence of the rank-frequency distributions for Markov models of increasing order. These investigations will be paralleled in Section 3.5 for the case of Hidden Markov Models with an increasing number of hidden states. Section 3.6 proposes a mixture model of two first order Markov processes, one generating words of bounded length, and one generating words of arbitrary length, and shows that this mixture model achieves rank frequency behavior very close to the true natural distribution. Finally, in Section 3.7, we compare the different proposed models.

The overall goal of our analysis was to verify the asymptotic results discussed in the literature empirically on finite data and to investigate rank-frequency distributions for scenarios that are difficult to analyze theoretically. These scenarios are meant to fill a gap between the crudest model of natural language—the memoryless "monkey" case—and the behavior of actual natural language in order to investigate the validity of Zipf's law for higher-order but yet artificial models of natural language.

3.1 Zipf's law across different corpora

We performed experiments with Zipf's original corpus—Shakespeare's *Hamlet*—, Goethe's *Faust*, a corpus of Reuters news articles from August 20-22, 1996, and part of the Linux kernel (version 2.6.11.7, all C files in `kernel`, `kernel/irq`, `kernel/power`, `fs/ext3`, `fs/afs`, `init` and `lib` directories). For all our analyses, we converted the text to lower case and removed all characters apart from the letters A through Z and the space character. Figure 3.1 gives the rank-frequency distributions of the corpora. Barring the ten most frequent words, all three purely natural-language corpora follow a nice Zipf distribution, Reuters yielding the straightest curve. The curve of the Linux kernel, on the other hand, seems to follow a log-normal distribution. Consisting mainly of C-language and partly of natural-language comments, the vocabulary size (and thus maximum word-length) is severely restricted. Thus a more log-normal behavior of the rank-frequency distribution is expected.

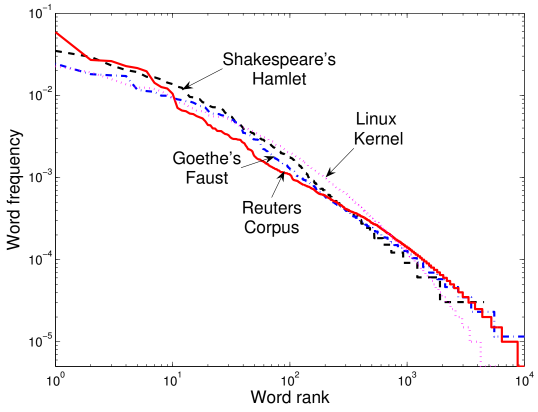
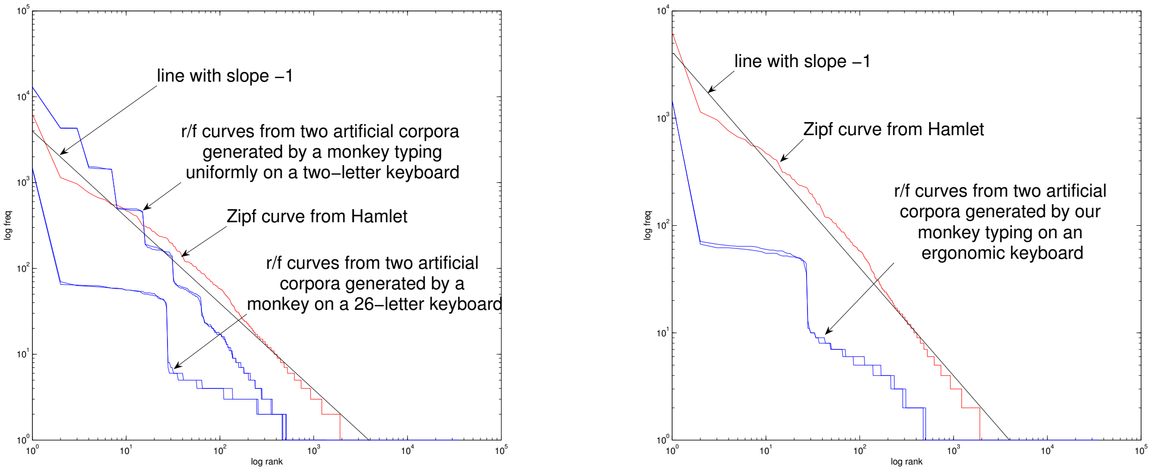


Figure 1: Comparison of corpora.



(a) *Miller's monkey for two and 26 letters.*

(b) *Monkey on an ergonomic keyboard.*

Figure 2: Rank-frequency curves of several zero- and first-order Markov models.

3.2 Uniform letter probabilities

The first experiment concerns Miller's scenario of a monkey typing letters with uniform probabilities. Figure 2(a) shows the rank-frequency distribution resulting from generated data for the original two-letter case as well as for all 26 letters of the alphabet.[2] For comparison, the graph also contains the rank-frequency distribution of *Hamlet*—and a line with slope -1. One can see that the graph shows a step-behavior as opposed to the much smoother Zipf curve. This is due to the fact the number of words of a given length—which, in the uniform case, directly determines their expected frequencies—increases exponentially in the word length, as indicated in Section 1. When the number of words in the corpus grows large, words of greater length are expected to occur in the corpus, and thus the number of steps in the graph increases, so that on a bigger scale the graph looks more and more linear. In the two-letter case words of greater length are produced earlier; thus, more steps can be seen in the corresponding graph than in the 26-letter case, whose only distinguishable step is the one from rank 1 to rank 27 representing the 26 words of length one. Also, the slope in the two-letter case is steeper than the slope in the 26-letter case, as predicted by the theory.

3.3 First order Markov processes

One might wonder whether the power-law behavior of the rank-frequency curve is valid for more sophisticated but yet artificial language models. We studied the case of a first-order Markov process underlying the generation of letters. The model has 27 states for the letters of the alphabet and the space key. As previously, generation of the space key is interpreted as the end of a word.

The first model explored is only a slight generalization: a monkey typing on an *ergonomic* keyboard, separated into two halves. We assume the monkey uses its left hand to type on the left keyboard half, and its right hand for the other half, thereby alternating letter-by-letter between its hands. The space key is assumed to be reachable by both hands and each letter is again equally likely. Note that this model is an HMM in the spirit of Section 2, having two states for each keyboard half with emission probabilities for the respective symbols and one absorbing space key state. It also can be viewed as a first-order Markov model with transition probability $1/14$ of a letter in one half of the keyboard transitioning to a letter in the other half (or space), and probability zero for all other transitions. Figure 2(b) shows the resulting rank-frequency distribution, strikingly similar to the above 26-letter case of Miller's monkey.

In the next experiments, the transition matrix governing the transition probabilities from a given state to the next one was assigned (i) randomly from an elementwise uniform distribution with subsequent normalization and (ii) as the maximum-likelihood estimate from the *Hamlet* corpus. Figure 3 shows the rank-frequency curves from the generated data and from the original data. Unsurprisingly, the data generated according to the maximum likelihood estimate from *Hamlet* yields a curve much closer to the original rank-frequency curve as compared to data from random transition matrices. Striking, however, is the similarity of the rank-frequency curves from the different random transition matrices. The irregular transition matrix estimated from actual natural language leads to a good approximation of a line with slope -1, whereas the truly random transition matrices yield a curve that is more similar to the zero-order uniform case of the memoryless Miller monkey, with the difference that the large step from rank 1 to rank 27 is much smoother in the random first-order Markov model case.

The question arises whether different initial-state distribu-

[2] Here and in Figure 2(b), two (nearly identical) graphs are shown for each distribution from generated data, reflecting the small variance in results from different runs.

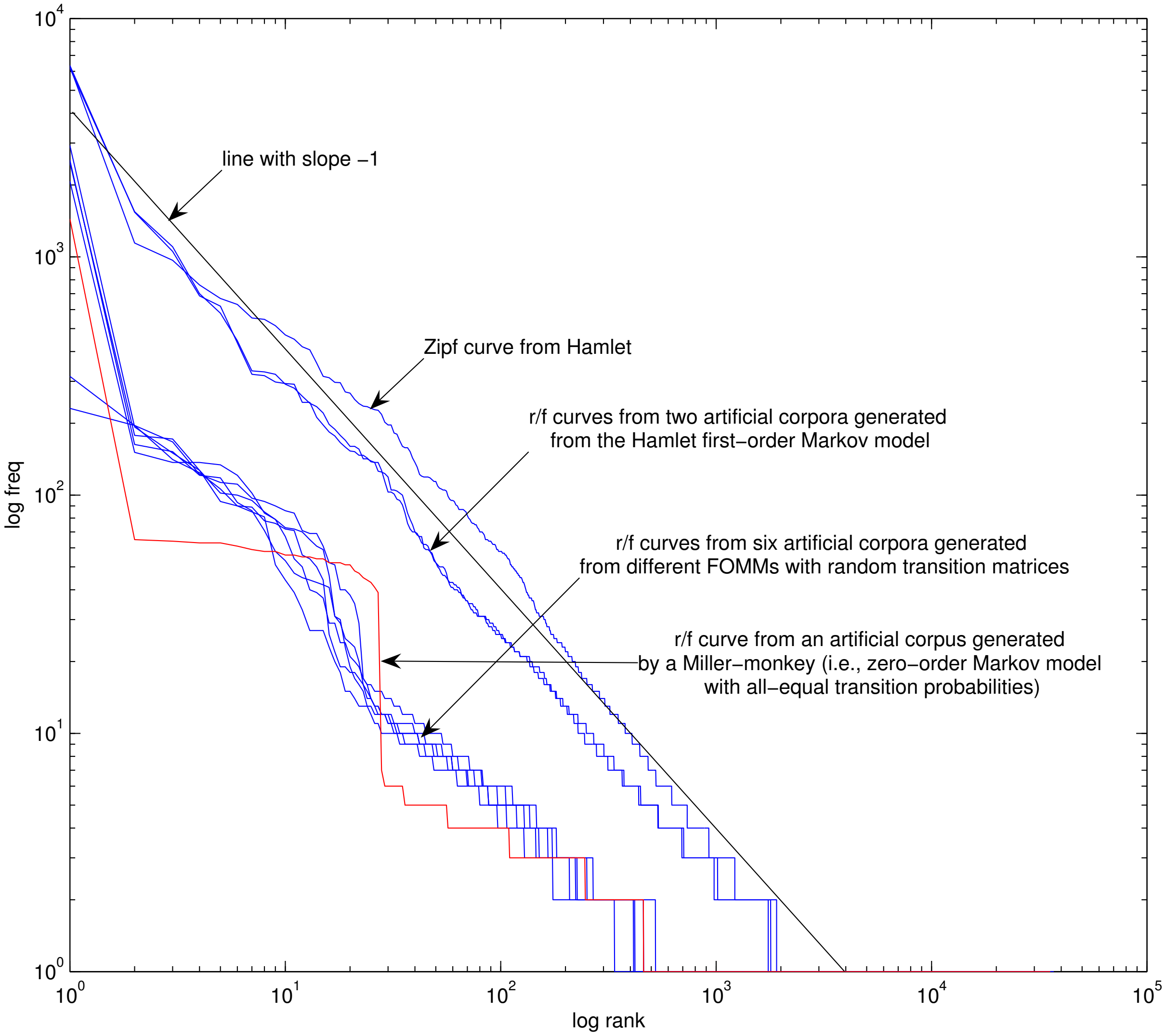


Figure 3: First-order Markov models.

tions can lead to fundamentally different types of generated corpora for a given transition matrix. The answer is no: As long as there is a non-zero probability path from each state to the space state, the initial distribution will only have influence on the first word in the corpus. After that word has been generated, we are always in the space state, which caused the end of the first word, and thus from now on the initial distribution is irrelevant. In our experiments, the initial state for the first-order Markov model was chosen to be the space state.

3.4 Convergence of Markov processes with increasing order

In his groundbreaking paper [14] on information theory, Shannon hypothesized that text generated by Markov models of increasing order estimated from natural language text should increasingly resemble the underlying natural language. For reasons of data-sparseness, it has been impossible to estimate a sufficiently close Markov model to verify this claim. A weaker claim is that there is a sequence of Markov models of increasing order whose rank-frequency distribution converges to the rank-frequency distribution of natural language. We verified this claim empirically on the Reuters corpus by estimating higher-order Markov models (HOMMs) from the data. Table 1 lists the most frequent words in the original data set and in artificial corpora generated by first-, third-, and sixth-order Markov models. While the first-order model still gives clear precedence to one- and two-letter words as in the case of Miller's uniform zero-order model, the top-50 words generated by the third-order model are already a nearly-perfect permutation of the original top-50 words. The sixth-order model yields words at ranks 1-50 with even nearly identical positions as in the Reuters data.

Figure 4(a) displays the rank-frequency distributions of the Reuters corpus and corpora generated from Markov Models of orders 1 to 6. Clearly, the graphs become increasingly close to the true distribution and are nearly indistinguishable from it for orders 5 and 6. In order to quantify the convergence behavior, we computed the L_2 distance between the true and the artificial curve for each Markov model (Figure 5(a)). The distance is monotonically decreasing with increasing Markov order and appears to converge to zero asymptotically.

Overfitting issues and corpus homogeneity In the above experiments, we tested models estimated from a corpus against the very same corpus. This is objectionable since the convergence results might be due to overfitting. Figure 3.4 shows the rank-frequency graphs of two different sections of the Reuters corpus (the section used in the experiment above and a held-out section) together with the rank-frequencies of data generated from two first-order Markov models trained from these sections. The Reuters corpus is strikingly homogeneous in terms of rank-frequencies. The graphs for the training and the held-out section are nearly indistinguishable. Therefore, testing against the held-out section in the above experiments would have resulted in approximately the same results. As one would expect, the corpus homogeneity also leads to strongly similar rank-frequency behavior of the corresponding trained first-order Markov models.

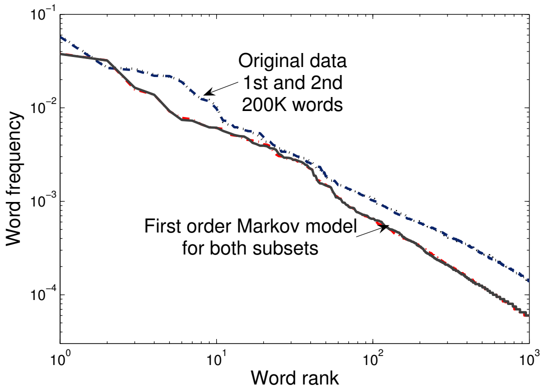


Figure 6: Corpus homogeneity.

3.5 Convergence of Hidden Markov Models with increasing number of hidden states

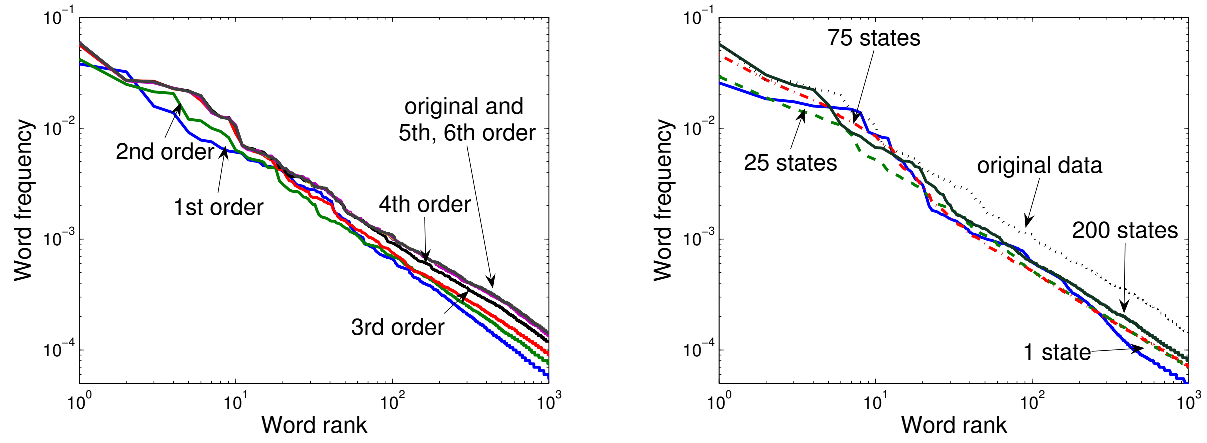
Analogously to our analysis in Section 3.4, we analyzed how the rank frequency distributions changes if increasingly complex Hidden Markov Models are estimated from the same corpus of natural language text. For this experiment, we considered the first 10000 words of the Reuters corpus. For each number of hidden states k, we generated random initial values for the transition matrix $P \in \mathbb{R}^{k\times k}$, the observation matrix $Q \in \mathbb{R}^{k\times 27}$ and the initial distribution $\nu \in \mathbb{R}^{27}$ as specified in Section 2. The parameters were chosen uniformly at random from the interval $(0, 1)$ and then normalized to probability distributions. These matrices were then used as initial values for the Baum-Welch parameter learning algorithm for Hidden Markov Models [2]. Since this expectation maximization method is only guaranteed to find local maxima, we used several random restarts. For each experiment, we performed 100 iterations, which provided a tolerance of approximately 0.1% of the absolute log-likelihood. We then used the learned models to generate artificial corpora of 1.5 million words each from which we estimated the rank-frequency distributions. Figure 7(a) shows the log-likelihoods with standard error bars for an increasing number of hidden states, and Figure 7(b) shows the corresponding Akaike Information Criterion (AIC) score. This common model selection criterion, along with computational restrictions, led us to only consider HMMs up to size of 200 hidden states.

Figure 8 shows an example of a learned Hidden Markov Model with ten hidden states. Only edges and emissions with probabilities greater than 10 percent are drawn. It is interesting to see that there are two states which clearly account for most of the vowels. Furthermore, the transitions are illuminating: There is one state emitting 'ct' which leads with high probability to a state emitting 'hr', apparently accounting for the common consonant combinations of 'ch' and 'tr'. There is a state which likely emits spaces and the letter 's', with a self-loop. This intuitively makes sense, since many words start and end with the letter 's'.

Figure 4(b) presents the rank-frequency curves for 1, 25, 75 and 200 states, along with the original data. Although the convergence is not as drastic as for the higher order Markov

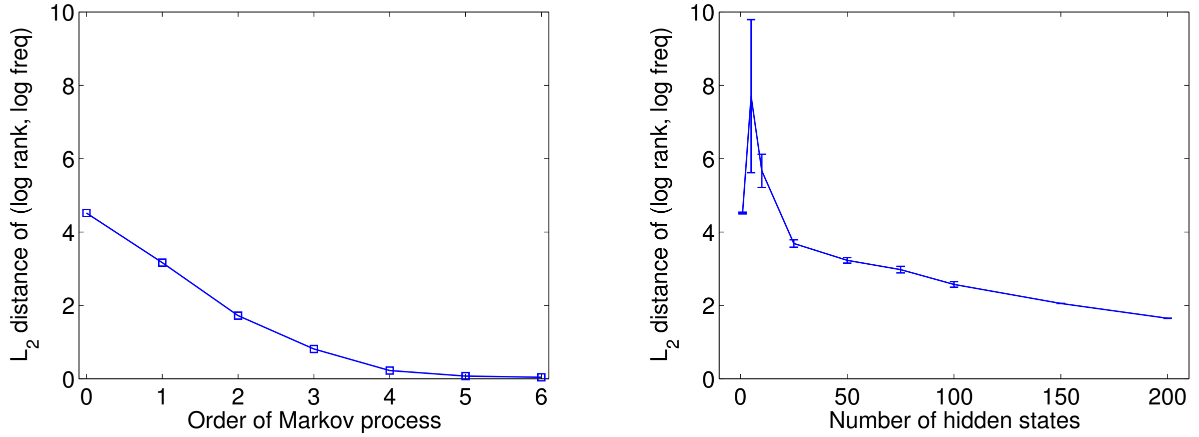
Rank	IND	HMM10	HMM50	HMM100	HMM200	1OMM	3OMM	6OMM	Orig
1	e	s	the	the	the	s	the	the	the
2	t	a	in	to	to	t	of	of	of
3	a	o	on	a	of	d	to	to	to
4	o	he	a	on	a	f	in	in	in
5	s	on	te	in	in	a	a	a	a
6	i	an	tho	s	s	e	and	and	and
7	n	the	s	of	it	o	said	said	said
8	r	or	and	an	on	n	on	on	on
9	l	re	th	it	ang	r	s	s	s
10	d	ar	it	at	and	y	for	for	for
11	c	e	ang	ind	said	g	at	at	at
12	h	i	to	is	as	w	it	that	that
13	u	in	at	ing	by	an	that	it	it
14	m	en	if	and	at	in	is	was	was
15	p	che	of	as	is	m	was	is	is
16	f	tre	is	ang	thas	l	by	with	with
17	g	to	an	che	for	te	with	from	by
18	w	le	as	he	he	p	from	by	from
19	y	os	ed	ow	its	th	its	percent	percent
20	b	ton	ce	whe	ant	he	as	as	as
21	ee	as	sa	cor	an	the	be	its	its
22	k	ta	whe	sas	that	h	he	tuesday	tuesday
23	et	tan	aid	yhe	be	on	an	million	million
24	te	me	he	was	all	be	but	be	year
25	ea	ir	che	said	or	as	new	year	be
26	ie	ol	aig	by	was	at	had	will	will
27	v	hy	we	qhe	we	pe	has	but	but
28	ae	u	sand	waid	thay	st	tuesday	he	has
29	se	er	er	t	tue	by	would	which	he
30	oe	hs	thed	thin	will	me	year	an	an
31	ei	tor	e	sto	ants	ce	not	has	which
32	es	ss	u	tho	mon	c	up	have	have
33	eo	al	ped	ats	t	we	have	would	market
34	en	co	sed	or	they	ay	they	are	had
35	ne	ont	sang	its	u	ts	ther	market	are
36	re	at	t	mar	wat	se	market	were	new
37	er	thy	ho	no	inter	is	we	not	were
38	aa	ot	ins	py	con	id	which	new	would
39	tt	es	p	ther	sail	ad	this	had	not
40	at	cre	ons	they	ther	ar	are	company	company
41	st	she	ther	than	fors	re	u	after	this
42	ot	con	who	thit	dor	al	or	this	after
43	to	pe	sat	has	co	ss	first	up	up
44	ta	te	its	be	thesdas	pr	his	u	we
45	oo	tar	ser	stry	if	of	no	we	they
46	tn	tin	ip	af	are	to	mill	first	first
47	an	is	af	that	fow	it	their	billion	billion
48	ar	whe	con	all	mare	de	per	they	u
49	ti	ten	se	may	cost	ve	been	one	one
50	ss	ti	wo	haid	inted	nd	who	their	or

Table 1: Most frequent words for different models, based on initial 10K words of Reuters data set.



(a) *Convergence of HOMMs.* (b) *Convergence of HMMs.*

Figure 4: Convergence of rank frequency distributions.



(a) *Convergence of HOMMs.* (b) *Convergence of HMMs.*

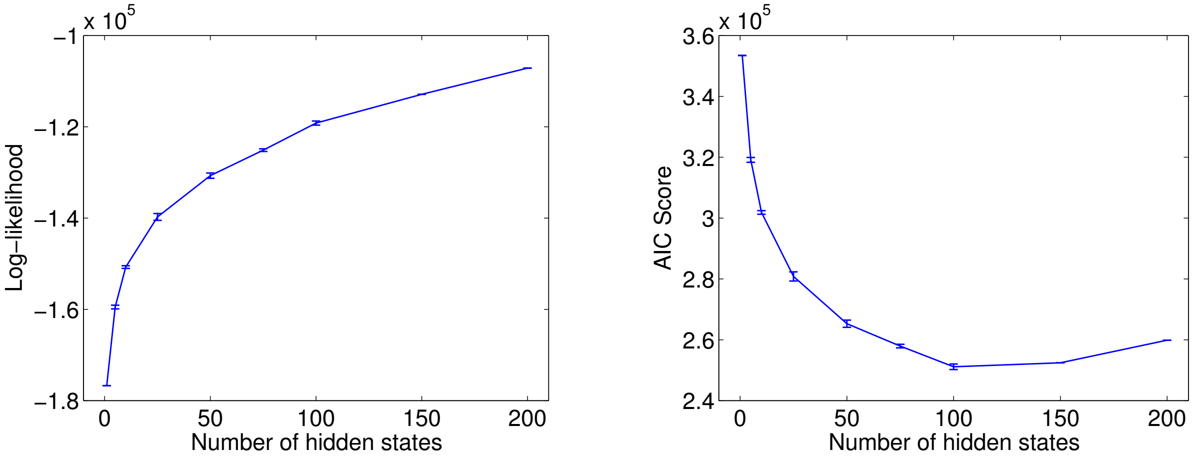
Figure 5: Convergence in L_2 distance.

processes, it can be seen that the curve for 200 hidden states much more closely approximates the true distribution than the curves for fewer states. The curve for one state (the i.i.d. case) favors the very short words (i.e. the one-letter words). The curve for 75 states approximates the true distribution better than the curve for 25 states for the high-frequency words (rank 1-20) and than provides comparable performance for the lesser ranked words.

We again computed the L_2 distance between the true rank-frequencies and the rank-frequencies for artificial text for different numbers of states. The results are presented in Figure 5(b). This figure also provides the standard error bars for up to 100 states – the last two experiments could not be repeated due to high computation time. The graph shows a monotonically decreasing L_2 distance, with the only exception that the i.i.d. case matches the true distribution more closely than even the 10 state Hidden Markov Model. We believe that this issue – and also the high variability for these exceptional cases of 5 and 10 states – is due to problems with the EM learning procedure.

3.6 Structural words

Comparing the original Zipf curve for the Hamlet corpus with the line of slope -1 as presented in Figure 2(a), we identify a slight “bump” caused by the most frequent words (approximately up to rank 100). One possible hypothesis for the reason of its occurrence is the presence of structural words, or “stop words”. As opposed to content words, these structural words have only grammatical purpose, and their repertoire is limited, but they occur very often. Guided by the assumption that structural words are very short, we decided to model the generation of the text as a two component mixture of processes: One component generates, with probability p, words up to a fixed length k, and the other component, with probability $1-p$, words of arbitrary length. Each component uses transition probabilities P formalized as a first order Markov



(a) *Log-likelihoods for HMMs.*

(b) *AIC Score for HMMs.*

Figure 7: Model selection criteria.

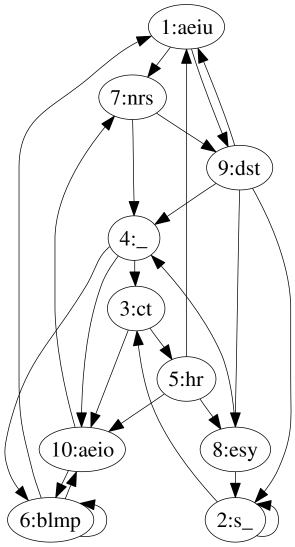


Figure 8: Example HMM with 10 states. Shown are only edges and emissions with probability greater than $.1$.

model, which was estimated from the Hamlet corpus. The mixture can be modeled as a Hidden Markov Model: The hidden state is a product state $s = (c, j, a)$, consisting of an indicator $c \in \{1, 2\}$ of the mixture component, the counter $1 \leq j \leq k$ of the number of emitted letters (for the bounded length case) and the actual current emitted letter a, hence the emission probabilities are deterministic. This mixture model is parameterized by p and k. We experimented with varying values of these parameters, and Figure 3.6 presents results for $p = 0.9$ and $k \in \{2, 3\}$. Whereas for $k = 2$ the log-normal bounded word behavior dominates, for $k = 3$ the curve very closely approximates the true rank-frequency distributions. This result shows the power of Hidden Markov Models to capture more complex stochastic behavior. It also demonstrates that a simple modified first order model can approximate the true rank-frequency distribution very well.

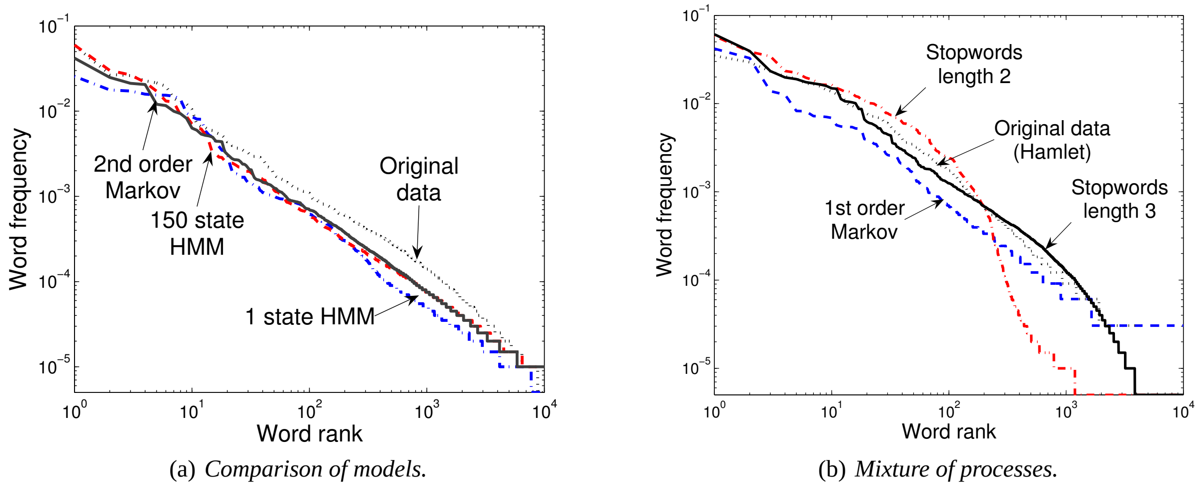
3.7 Comparison of Models

Looking back at Figure 5, the convergence results are clearest for the case of Markov models of increasing order. Table 1 supports this impression: While the top-50 words of the sixth-order Markov model are nearly identical to the top-50 words in the original corpus, many unnatural words still occur in the data generated by the 200-state HMM. However, this comparison is not quite fair: A 200-state HMM has only approximately $200^2 + 200 * 27 + 200 = 45600$ parameters, while a sixth-order Markov model has $27^7 = 10.5$ billion! An appropriate comparison is the 150-state HMM, which has approximately 26000 parameters, to the second order Markov process with approximately 20000 parameters. Figure 9(a) shows the rank-frequency distributions of these models and the original (Reuters) data. The curves of the two models are conspicuously similar.

More rigorous is again the comparison of the L_2 distances in rank-frequencies between models and true data (Figure 5). Although the complex 150-state HMM has more parameters than the simple second-order Markov model, its L_2 distance is slightly higher. The same happens for the case of the the HMM with 25 states, which has almost as many parameters as the first order Markov model but a higher L_2 distance. We believe that this is due to the fact that in the Markov model case it is possible to explicitly compute the maximum likelihood estimates of the parameters, whereas for the HMM, the EM algorithm can get stuck in local minima. Independently of this observation, we can conclude that one can clearly see a convergence behavior of the rank-frequencies also in the case of Hidden Markov Models, and the picture is more detailed than that from the higher order Markov chains, since one has a more fine-grained control over the number of parameters.

4 Conclusions and Future Work

We presented a literature overview discussing power-law and log-normal distributions arising from human language



(a) *Comparison of models.*

(b) *Mixture of processes.*

Figure 9: Model comparison and structural / content words.

and randomly generated text. We then showed how Hidden Markov Models can capture the behavior of the previous models for randomly generating text and indicated how they can lead to various rank-frequency distributions, such as power laws and log-normal distributions. We proved an extension to Conrad's and Mitzenmacher's theorem [3] which allows to explicitly compute power law exponents for Markov chains for which the rows of the transition matrix are permutations of each other. This captures the important special case of random walks in d-regular graphs. A promising perspective for future work would be to extend this result for general higher order Markov processes or even Hidden Markov models.

The focus of this paper was on empirical analysis of rank-frequency behavior for randomly generated text. We showed how the artificial rank-frequency distributions converge to the true distribution for increasing order of the Markov process, as well as for an increasing number of hidden states in the HMM setting. We also proposed a mixture of processes generating both bounded and unbounded length words, which even in the first order case closely approximates the true rank-frequency distribution for the Reuters corpus. This mixture can be formulated in our HMM framework. We believe that our results shed more light on the rank-frequency behavior of text with a varying amount of randomness.

5 Acknowledgements

We would like to thank David Abraham for his collaboration in the initial part of the project, and Reuters for making their news archive available.

References

[1] L. A. Adamic. Zipf, power-laws, and pareto - a ranking tutorial. www.hpl.hp.com/research/idl/papers/ranking/ranking.html.

[2] L. E. Baum. A maximization technique occuring in the statistical analysis of probabilisitic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[3] B. Conrad and M. Mitzenmacher. Power laws for monkeys typing randomly: The case of unequal probabilities. *IEEE Transactions on Information Theory*, 50(7):1403–1414, 2004.

[4] J. Estoup. Gammes sténographiques. Technical report, Institut Stenographique de France, 1916.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, 1999.

[6] R. F. i Cancho and R. Sole. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

[7] I. Kanter and D. A. Kessler. Markov processes: Linguistics and Zipf's law. *Phys. Rev. Lett.*, 74(22):45594562, 1995.

[8] B. B. Mandelbrot. *Communication Theory*. New York Academic Press, 1953.

[9] B. B. Mandelbrot. On recurrent noise limited coding. In E. Weber, editor, *Information Networks, the Brooklyn Polytechnic Institute Symposium*, pages 205–221, 1955.

[10] G. A. Miller. Some effects of intermittent silence. *Amer. J. Psychology*, 70:311–314, 1957.

[11] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[12] V. Pareto. Cours d'economic politique. Technical report, Dronz, Geneva, Switzerland, 1896.

[13] R. Perline. Zipf's law, the central limit theorem, and the random division of the unit interval. *Phys. Rev. E*, 54(1):220223, 1996.

[14] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

[15] G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press, 1932.

[16] G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley, 1949.