
Hamptonese and Hidden Markov Models

Mark Stamp and Ethan Le

Department of Computer Science, San Jose State University, San Jose, CA 95192
`stamp@cs.sjsu.edu`

Summary. James Hampton was a quiet and secretive man who left behind a monumental work of visionary art, along with a strange handwritten script. Hampton’s script, or Hamptonese, bears no significant resemblance to any known written language. In this paper we analyze Hamptonese using hidden Markov models. This analysis shows that Hamptonese is not a simple substitution for English and provides some evidence that Hamptonese may be the written equivalent of “speaking in tongues”.

1 James Hampton and Hamptonese

James Hampton was born in South Carolina in the year 1909. Hampton’s father left his family to pursue his calling as an itinerant gospel singer and self-appointed Baptist minister. At the age of 19, Hampton moved to Washington, DC, where he struggled to find gainful employment during the Great Depression. Hampton was drafted into the Army in 1942 and served in Guam and Saipan. After his discharge from the Army in 1945, Hampton was employed as a janitor for the General Services Administration in Washington, DC, where he worked until his death in 1964.

Beyond the thumbnail sketch in the previous paragraph, very little is known about Hampton’s life. In his post-war years, he lived in a small apartment in DC and rented a nearby garage. During this time period, he apparently had no family or close friends. Shortly after Hampton’s death (due to stomach cancer), the owner of his rented garage discovered that it housed a collection of artwork that Hampton had dubbed “The Throne of the Third Heaven of the Nations’ Millennium General Assembly”.

Art critic Robert Hughes has written that Hampton’s Throne “. . . may well be the finest work of visionary religious art produced by an American” [14]. The Throne is part of the permanent collection of the Smithsonian Museum

of American Art in Washington, DC. Hampton is pictured with his Throne in Figure 1.



Fig. 1. Hampton and his Throne

Hampton's Throne is a remarkable collection of some 180 individual pieces. Each piece is constructed of discarded items—broken furniture, burnt out light bulbs, jars—carefully wrapped in gold and silver foil, with the foil itself taken from discarded items, such as cigarette packages. The artwork is extremely fragile, being held together with tacks, pins and tape.

The Throne is clearly a religious work as it includes several plaques with Biblical references. Hampton was a believer in the doctrine of “dispensationism” [5], and the “7 dispensations” figure prominently in his work. For more information and speculation on Hampton's artwork and his life see [14, 15].

Along with the Throne, several notebooks were discovered in Hampton's garage. These notebooks contain some scattered English text, but consist primarily of a script of unknown origin that we refer to as Hamptonese. In total, there are 164 pages of Hamptonese. These 164 pages can be separated into two sets, one of which includes tombstone-shaped drawings, Roman numerals, Hamptonese characters and significant amounts of English text, including phrases such as “THE OLD AND NEW COVENANT RECORDED BY ST. JAMES” and “fear not”. Many of these pages deal with the “7 dispensations” and some are similar to writing that appears on plaques included with the Throne.

The second set of Hamptonese text consists of 100 pages of essentially “pure” Hamptonese. This set has no drawings and only a handful of interspersed English words or letters, such as “Jesus”, “Virgin Mary”, “NRNR”

and “Revelation”. Each of these pages contains 25 or 26 lines of Hamptonese, with a highly variable number of symbols per line. There is no punctuation or paragraphs and it is not clear that the pages should be read top-to-bottom and left-to-right (although the ragged right margin might indicate that left-to-right is appropriate).

Each page in the pure Hamptonese set is headed by “ST James” with “REVELATiON” appearing at the bottom. Hampton also numbered the pages, but his numbering is inconsistent with duplicates, gaps and some pages having two number. Curiously, a few pages appear to have had certain lines extended after the page was initially written, and in a few places it appears that at least some lines may have been written from bottom to top, due to the location of strategic gaps within the text. In addition, water damage significantly affects a few pages and some pages are marred by ink blotches.

It is not clear whether Hampton wrote his Hamptonese over a relatively short period of time, or over a more extended timeframe. There is a reference to Truman’s inauguration on one page, which implies that Hampton could have spent well over a decade writing his Hamptonese text. The writing style does change substantially over the 100 pages of pure Hamptonese, becoming far clearer—and easier to transcribe—as it progresses.

The 100 pages of “pure” Hamptonese is the basis for the analysis presented in this paper. A partial page of Hamptonese appears in Figure 2. Scanned images of all 100 pages are available online in JPEG and TIFF format at [12].

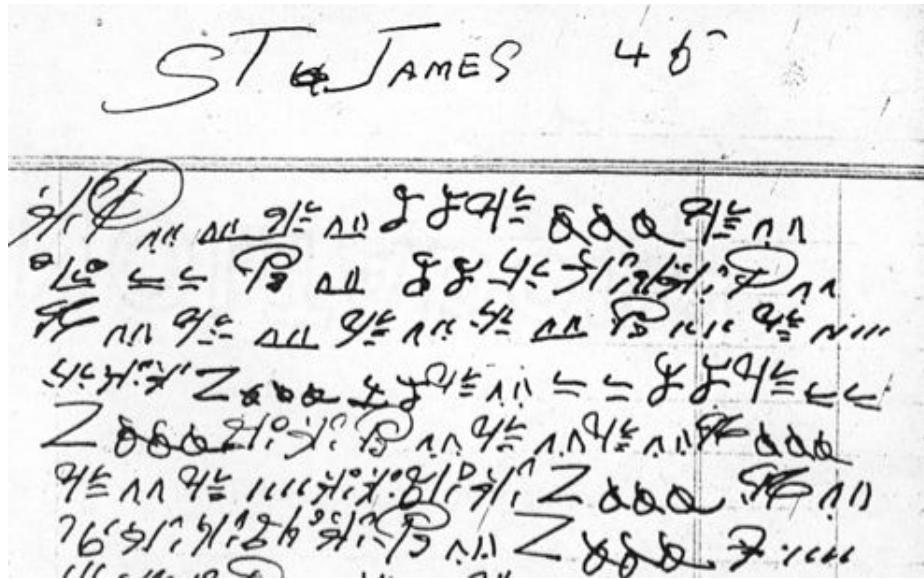


Fig. 2. Example of Hamptonese

A comparison of Hamptonese with the encyclopedic collection of written languages at Ominglot [1] reveals no similar script. The only published analysis of Hamptonese that we are aware of is [9], which makes strong claims regarding the number of “consonants” and “vowels” in Hamptonese based on a small sample of the text. We believe that these claims are unsupported by the limited data and analysis provided.

2 Transcription

Before analyzing Hamptonese, we first transcribed it into a computer-friendly form. We believe that our transcription is reasonably accurate, but given the nature of Hamptonese, there are certain to be some errors.

We took a conservative approach, choosing to distinguishing symbols that might simply be variants or “typos”. When doing analysis, it is always possible to group selected symbols together.

The Hamptonese symbols, the corresponding transcription keys and frequency counts appear in Table 1. A complete transcription of the 100 pages of Hamptonese is available at [12].

In the next section we present elementary entropy calculations which indicate that in some sense the information in Hamptonese is comparable to that in English. We then give a brief introduction to hidden Markov models followed by a discussion of the utility of this technique for analyzing English text. This is followed by a hidden Markov model analysis of Hamptonese and a discussion of our findings.

3 Entropy of Hamptonese

Shannon’s entropy is the classic measure of information or uncertainty [8]. In terms of bits, entropy is computed as

$$H(X) = - \sum_{x \in X} P(x) \log(P(x)),$$

where the logarithm is to the base 2, and $0 \log(0)$ is taken to be 0. Another useful formula is that for conditional entropy [13],

$$H(X | Y) = - \sum_x \sum_y P_{X,Y}(x, y) \log(P_Y(y | x)). \quad (1)$$

It is well-known—and intuitively clear—that conditioning reduces entropy, i.e., $H(Y | X) \leq H(Y)$, with equality holding if and only if X and Y are independent.

If English consisted of random selections from the 27 symbols (letters and space) then the uncertainty of English (at the level of individual symbols) would be

key	Hamptonese	count	relative frequency	key	Hamptonese	count	relative frequency
2	2	529	0.0180565	14	1111	1313	0.0448169
ee	88	227	0.0077482	76	76	871	0.0297300
95	95	195	0.0066560	96	96	453	0.0154623
dc	dc	709	0.0242004	EE	33	415	0.0141653
vv	111	6318	0.2156535	M	11	581	0.0198314
F	7	493	0.0168277	N	96	541	0.0184660
g	8	360	0.0122879	g7	82	46	0.0015701
GG	18 18	1365	0.0465918	Gi	61	889	0.0303444
Ki	31	2138	0.0729768	d3	61	510	0.0174079
d4	61	51	0.0017408	Y3	91	3578	0.1221285
Y4	91	644	0.0219818	qL3	81	1113	0.0379902
qL4	91	23	0.0007851	4L	44	272	0.0092842
uL	44	45	0.0015360	J1	1	186	0.0063487
JJ	87	587	0.0200362	LL	5 5	1014	0.0346111
nn	77	337	0.0115029	P	7	754	0.0257364
PL	1	113	0.0038571	P1	7	302	0.0103082
P2	18	665	0.0226986	o3	0 0 0	299	0.0102058
q3	88 88	558	0.0190463	S	3	329	0.0112298
3	8	89	0.0030379	T	7	67	0.0022869
A	96	21	0.0007168	A-	96	182	0.0062122
44	44	105	0.0035840	I	1	10	0.0003413
total						29297	1.0000000

Table 1. Hamptonese frequency counts

$$H(X) = - \sum 1/27 \log(1/27) = \log(27) \approx 4.75,$$

i.e., each letter would contain about 4.75 bits of information.

Of course, the distribution of English letters is not uniform, as is also true of the phonemes in phonetic English. Similarly, Hamptonese characters are not uniformly distributed. We have computed the empirical entropy for individual symbols for English letters and phonemes, as well as for Hamptonese characters. We have also computed the entropy for two and three consecutive symbols using the conditional entropy formula (1). For example, the term associated with the string abc in (1) is given by

$$P(abc) \log(P(c|ab)).$$

Our entropy results appear in Table 2.

	Entropy		
	1 symbol	2 symbols	3 symbols
English letters	4.11	3.54	2.68
English phonemes	4.76	3.87	3.00
Hamptonese	4.41	3.63	3.08

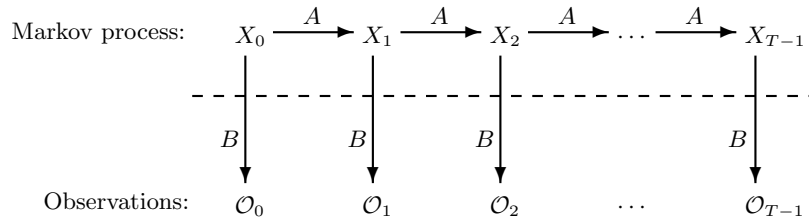
Table 2. Entropy comparison

The results in Table 2 indicate that for strings of three symbols or less, the information contained in Hamptonese is comparable to that of English. Of course, this does not prove that Hamptonese is a language, but it does suggest that it is reasonable to pursue the analysis further.

4 Hidden Markov Models

Hidden Markov models (HMMs) provide an efficient algorithmic solution to certain problems involving Markov processes. In a Markov process (of order one) the next state depends only on the current state and a fixed set of transition probabilities. HMMs generalize a Markov process to the case where the states cannot be directly observed. In an HMM, we observe some properties of the system that are related to the underlying states by fixed probability distributions.

A schematic view of an HMM is given in Figure 3, where $A = \{a_{ij}\}$ is the (row stochastic) matrix driving the underlying Markov process, the X_i are the states of the Markov process, the $\mathcal{O}_i$ are the observations, T is the length of the observed sequence, $B = \{b_{ij}\}$ is a (row stochastic) matrix that relates the states to the observations, and the dashed line is a “curtain” between the observer and the underlying Markov process. More precisely, a_{ij} is the probability of a transition to state j , given that the system is in state i , and b_{ij} is the conditional probability of observing “ j ” given that the Markov process is in state i . The matrices A and B , together with an initial state distribution π define the model, which is denoted $\lambda = (A, B, \pi)$.

**Fig. 3.** Hidden Markov Model

Efficient algorithms exist for solving the following three HMM problems.

Problem 1: Given the model $\lambda = (A, B, \pi)$ and a series of observations $\mathcal{O}$, find $P(\mathcal{O} | \lambda)$, that is, find the probability of the observed sequence given the (putative) model.

Problem 2: Given the model $\lambda = (A, B, \pi)$ and the observations $\mathcal{O}$, determine the most likely state sequence. In other words, we want to uncover the hidden part of the HMM.

Problem 3: Given the observations $\mathcal{O}$, “train” the model to best fit the observations. Note that the dimensions of the matrices are fixed, but the elements of A , B and π can vary, subject only to the row stochastic condition.

For example, consider speech recognition—an application where HMMs have been applied with great success. We could use the solution to Problem 3 to train a model to best match the word “no” and train another model to best match the word “yes”. Then given a spoken word, we could use the solution to Problem 1 to determine whether it was more likely “yes”, “no” or neither. In this scenario we do not need to solve Problem 2. However, a solution to Problem 2 (i.e., uncovering the hidden states) might provide additional insight into the underlying speech model.

In this paper we are interested in solving Problem 3. But first we must determine the sense in which the solution will “best fit” the observations. Perhaps the most intuitively appealing approach is to find the state sequence which yields the largest overall probability. In fact, this is precisely the solution found by dynamic programming. However, for HMMs, we instead maximize the expected number of correct states. For this reason, the HMM algorithm is sometimes referred to as the expectation maximization (EM) algorithm. This definition of “best” has a curious side-effect, namely, that invalid state transitions can occur in the optimal solution.

The HMM solution method (for Problem 3) is iterative, and can be viewed as a discrete hill climb on the parameter space defined by $\lambda = (A, B, \pi)$. As is typical of hill climbing techniques—discrete or not—we are only assured of finding a local maximum.

Further details on the HMM algorithms are beyond the scope of this paper. For more information, the standard introduction is [7], which also discusses the speech recognition application in some detail. The paper [11] gives a reasonably thorough introduction to HMMs with the emphasis on the algorithms and implementation.

5 HMMs and English Text

In this section we present two HMM experiments, one of which applies to letters in English text, and the other to phonetic English. In the following section we present analogous results for Hamptonese.

Cave and Neuwirth [3] apparently were the first to apply HMMs to English text. They selected the Brown Corpus [2] as a representative sample of English. This corpus of more than 1,000,000 words was carefully compiled (in the early 1960's) so as to contain a diverse selection of written English. Cave and Neuwirth eliminated all numbers, punctuation and special characters, and converted all letters to lower-case, leaving 27 distinct symbols—the letters plus inter-word space. They then assumed that there exists a Markov process with two hidden states, with the observations given by the symbols (i.e., letters) that appear in the Brown Corpus. This results in an A matrix that is 2×2 and a B matrix that is 2×27 . They then solved HMM Problem 3 for the optimal matrices.

In repeating the Cave and Neuwirth experiment, using 10,000 observation and about 200 iterations, we obtain the results in Table 3, where the leftmost two columns give the initial B matrix, and the rightmost two columns give the final B . Note that the B matrix was initialized with approximately uniform random values. It is crucial that the matrix not be initialized to precisely uniform values, since that is a stationary point from which the algorithm cannot climb.

The results in Table 3 clearly show that hidden state 1 represents the “vowel state”, while hidden state 0 is the “consonant state”. Although it may not be surprising that consonants and vowels are distinguished, it is worth emphasizing that we made no *a priori* assumptions on the nature of the hidden states. This nicely illustrates the ability of HMMs to distill the most statistically significant information from the data.

Cave and Neuwirth [3] obtain and interpret results for English text HMMs having up to 12 hidden states. For example, with three hidden states, the consonant state splits into a “pre-vowel” consonant state (i.e., consonants that tend to occur before vowels) and a “post-vowel” consonant state.

We have also computed HMM results for English phonemes instead of letters. For this phoneme HMM we first made a phonetic transcription of the Brown Corpus using the CMU pronouncing dictionary [4]. Neglecting accent marks, the pronouncing dictionary has 39 symbols. With two hidden states, using about 50,000 phonemes and running the HMM for about 200 iterations, we obtained the results in Table 4. Additional results are available at [12].

In the English phoneme HMM we again see a clear separation of the observation symbols into the two hidden states. Interestingly, in this phoneme HMM, one hidden state evidently represents consonant sounds and the other represents vowel sounds, analogous to the English text HMM.

6 HMMs for Hamptonese

We have computed HMM experiments on Hamptonese similar to those presented in the previous section for English. In our experiments, the number of distinct Hamptonese symbols was 42 and we computed results for 2, 3, 4 and

letter	Initial B		Final B	
	state 0	state 1	state 0	state 1
a	0.0372642	0.0366080	0.0044447	0.1306242
b	0.0386792	0.0389249	0.0241154	0.0000000
c	0.0358491	0.0338276	0.0522168	0.0000000
d	0.0353774	0.0370714	0.0714247	0.0003260
e	0.0349057	0.0352178	0.0000000	0.2105809
f	0.0344340	0.0370714	0.0374685	0.0000000
g	0.0400943	0.0370714	0.0296958	0.0000000
h	0.0344340	0.0347544	0.0670510	0.0085455
i	0.0349057	0.0370714	0.0000000	0.1216511
j	0.0391509	0.0366080	0.0065769	0.0000000
k	0.0363208	0.0356812	0.0067762	0.0000000
l	0.0353774	0.0403151	0.0717349	0.0000135
m	0.0344340	0.0366080	0.0382657	0.0000000
n	0.0410377	0.0370714	0.1088182	0.0000000
o	0.0396226	0.0398517	0.0000000	0.1282757
p	0.0377358	0.0338276	0.0388589	0.0000047
q	0.0377358	0.0398517	0.0011958	0.0000000
r	0.0344340	0.0403151	0.1084196	0.0000000
s	0.0358491	0.0366080	0.1034371	0.0000000
t	0.0377358	0.0352178	0.1492508	0.0134756
u	0.0349057	0.0361446	0.0000000	0.0489816
v	0.0405660	0.0370714	0.0169406	0.0000000
w	0.0377358	0.0384615	0.0286993	0.0000000
x	0.0382075	0.0370714	0.0035874	0.0000000
y	0.0382075	0.0389249	0.0269053	0.0000003
z	0.0382075	0.0338276	0.0005979	0.0000000
space	0.0367925	0.0389249	0.0035184	0.3375209

Table 3. English text HMM results

5 hidden states. The final B matrix for a typical case with two hidden states appears in Table 5.

Suppose that we require a threshold of a factor of 10 in order to assign a particular symbol to either state 0 or 1. Then the results in Table 5 show that the five symbols 14, M, q3, vv and o3 belong to state 0, while the 15 symbols 2, 96, F, g7, d3, d4, Y3, Y4, qL3, qL4, uL, P, P1, P2, and S, are in state 1. However, the remaining 22 symbols cannot be assigned to either state. This is in sharp contrast to the HMM results for English letters and phonemes, where—using the same criteria—virtually all symbols are assignable to one state or the other.

We can reasonably interpret an HMM solution as providing a “fingerprint” of the underlying data. Of course, the rows of the solution matrices can appear in any order and a relabeling of the observations will reorder the columns. But qualitative differences in the results—such as those mentioned in the previous

phoneme	Final B		phoneme	Final B	
	state 0	state 1		state 0	state 1
AH	0.0000000	0.2969759	EY	0.0000957	0.0460653
Z	0.0468621	0.0013497	F	0.0272705	0.0000000
AO	0.0000000	0.0353722	R	0.0772887	0.0000000
T	0.1115366	0.0164165	UW	0.0024273	0.0420961
W	0.0264391	0.0000000	N	0.1239146	0.0000000
AA	0.0000000	0.0443532	B	0.0270710	0.0000000
ER	0.0104758	0.0458586	G	0.0110797	0.0005440
K	0.0533438	0.0000000	M	0.0451959	0.0000000
D	0.0662517	0.0139417	EH	0.0000000	0.0763638
V	0.0367820	0.0000000	AE	0.0000000	0.0781199
IH	0.0000000	0.1464058	S	0.0804303	0.0101116
IY	0.0079505	0.0596027	L	0.0653828	0.0000000
OW	0.0000000	0.0273946	NG	0.0132029	0.0000000
SH	0.0161295	0.0000000	HH	0.0211180	0.0000000
AW	0.0001282	0.0131526	TH	0.0039602	0.0006984
AY	0.0001135	0.0282771	JH	0.0123050	0.0000000
P	0.0381176	0.0030524	CH	0.0094782	0.0000000
ZH	0.0007316	0.0000000	Y	0.0104094	0.0000000
DH	0.0545078	0.0000000	UH	0.0000000	0.0118911
OY	0.0000000	0.0019568			

Table 4. English phoneme HMM results

paragraph—indicate fundamentally different data. By comparing the results of the English HMMs with the Hamptonese HMM, we can conclude that Hamptonese characters do not represent English letters or phonemes.

7 Conclusions

There are several possible explanations for our Hamptonese HMM results. For example, it might be the case that we have insufficient Hamptonese data available. However, we can obtain HMM results on English with as few as 10,000 symbols and we have more than 29,000 Hamptonese characters available.

If the quantity of data is sufficient, then perhaps its quality is insufficient, i.e., the Hamptonese data is excessively noisy. This could be due to either poor transcription or errors inherent in Hampton’s writings.

Another possible problem is incorrect interpretation of the Hamptonese data. It is conceivable that combinations of characters must be interpreted as the basic symbols. In this case the individual Hamptonese characters would have little or no semantic meaning and hence the HMM could not classify them. A potential example of this is given by the character “Ki”, which occurs both individually and as “Ki Ki”. The latter could certainly be considered as a distinct character.

symbol	Final B		symbol	Final B	
	state 0	state 1		state 0	state 1
2	0.0000000	0.0323915	14	0.0947011	0.0052195
ee	0.0059456	0.0091797	76	0.0504303	0.0132991
95	0.0013590	0.0108613	96	0.0071626	0.0220520
dc	0.0104561	0.0351127	EE	0.0141455	0.0141818
vv	0.4873287	0.0000000	M	0.0430382	0.0014101
F	0.0003230	0.0299307	N	0.0148755	0.0213175
g	0.0095264	0.0144809	g7	0.0005227	0.0024017
GG	0.0205943	0.0672325	Gi	0.0214469	0.0374094
Ki	0.0716056	0.0740697	d3	0.0000000	0.0312281
d4	0.0000000	0.0031228	Y3	0.0000000	0.2190866
Y4	0.0000000	0.0394331	qL3	0.0000000	0.0681507
qL4	0.0000241	0.0013892	4L	0.0068542	0.0112139
uL	0.0010389	0.0019307	J1	0.0055702	0.0069672
JJ	0.0089622	0.0288284	LL	0.0395888	0.0306004
nn	0.0108177	0.0120475	P	0.0000000	0.0461686
PL	0.0007135	0.0063528	P1	0.0000000	0.0184919
P2	0.0000000	0.0407190	o3	0.0230629	0.0000000
q3	0.0390770	0.0031463	S	0.0000000	0.0201452
3	0.0013213	0.0044007	T	0.0012893	0.0030790
A	0.0005299	0.0008652	A-	0.0044791	0.0075884
44	0.0027660	0.0042336	I	0.0004434	0.0002603

Table 5. Hamptonese HMM results

Perhaps Hamptonese is a cipher. At the extreme, Hamptonese could be the result of one-time pad encryption, in which case we have little hope of ever deciphering it. But given Hampton’s background we might assume that he could only have developed a fairly weak encryption system—the non-randomness of Hamptonese could be taken as some limited evidence of this¹. Our HMM experiment has ruled out that Hamptonese is a simple substitution cipher for English. After more cryptanalysis of Hamptonese, we might be able to make an argument to the effect that Hamptonese is probably not a good cipher, and we could probably break a bad cipher. Then if we cannot “break” Hamptonese, it is probably not a cipher.

Of course, there is a very real possibility that in spite of its language-like appearance, Hamptonese is simply the written equivalent of “speaking in tongues”. But even if this is the case, it is not necessarily the end of the story. In the late 19th century, a French mystic, Hélène Smith, claimed that when in a trancelike state, spirits transported her to Mars, where she was able to communicate with Martians. To substantiate her claims, she produced written “Martian” messages in an unknown script. Her Martian writing appar-

¹ There is an obvious risk of underestimating James Hampton. His Throne clearly shows that he was capable of achieving the unexpected.

ently was consistent and had the appearance of a genuine language. Reportedly, Smith's Martian messages could be interpreted as a simple variation on French [6, 10].

References

1. S. Ager, Omniglot: A guide to writing systems,
<http://www.omniglot.com/>
2. The Brown Corpus of Standard American English,
<http://www.cs.sjsu.edu/faculty/stamp/Hampton/BrownCorpus.gz>
3. R. L. Cave and L. P. Neuwirth, Hidden Markov models for English, in *Hidden Markov Models for Speech*, J. D. Ferguson, editor, IDA-CRD, Princeton, NJ, October 1980.
4. The CMU Pronouncing Dictionary,
<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
5. M. Esposito, L. Fleetwood and J. Brooks, Dispensationalism,
<http://www.endtimes.org/dispens.html>
6. A. G. Hefner, Smith, Helene,
http://www.themystica.com/mystica/articles/s/smith_helene.html
7. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, February 1989,
<http://www.cs.ucsb.edu/~cs281b/papers/HMMs%20-%20Rabiner.pdf>
8. C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October 1948,
<http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>
9. D. J. Stallings, Hamptonese statistics,
<http://www.geocities.com/ctesibos/hampton/hamptonese.html>
10. D. J. Stallings, Idiolect,
<http://www.geocities.com/ctesibos/hampton/idiolect.html>
11. M. Stamp, A revealing introduction to hidden Markov models,
<http://www.cs.sjsu.edu/faculty/stamp/Hampton/HMM.pdf>
12. M. Stamp and E. Le, Hamptonese,
<http://www.cs.sjsu.edu/faculty/stamp/Hampton/hampton.html>
13. W. Trappe and L. C. Washington, *Introduction to Cryptography with Coding Theory*, Prentice Hall, 2002.
14. M. Walsh, The miracle of St. James Hampton,
<http://www.missioncreep.com/tilt/hampton.html>
15. F. Weaver, James Hampton's Throne Of The Third Heaven Of The Nations Millenium General Assembly,
<http://www.fredweaver.com/throne/thronebody.html>