A CRYPTOGRAPHIC APPROACH TO LANGUAGE IDENTIFICATION: PPM

Ebru Celikel
Ege University International Computer 35100 Bornova/Izmir, Turkey

Keywords: Language identification, Statistical modelling, PPM

Abstract: The problem of language discrimination may arise in situations when many texts belonging to different source languages are at hand but we are not sure to which language each belongs to. This might usually be the case during information retrieval via Internet. We propose a cryptographic solution to the language identification problem: Employing the Prediction by Partial Matching (PPM) model, we generate a language model and then use this model to discriminate languages. PPM is a cryptographic tool based on an adaptive statistical model. It yields compression rates (measured in bits per character –bpc) to far better levels than that of many other conventional lossless compression tools. Language identification experiment results obtained on sample texts from five different languages as English, French, Turkish, German and Spanish Corpora are given. The rate of success yielded that the performance of the system is highly dependent on the diversity, as well as the target text and training text file sizes. The results also indicate that the PPM model is highly sensitive to input language. In cryptographic aspect, if the training text itself is kept secret, our language identification system would provide security to promising degrees.

1 INTRODUCTION

With ongoing improvements in network facilities, the number of computers getting connected to Internet is increasing everyday. This widespread use of Internet makes diversity of languages available online. Hence, there should be a sound way of determining the correct language of a retrieved text. This would not only help discriminate texts belonging to different source languages, but also facilitate text categorization, natural language processing and information retrieval.

Of many techniques used to determine the correct language, exploiting the statistical characteristics of disputed languages is quite common. Frequency values for the alphabet size, average word length, statistics of vowel/consonants throughout the text might all be used to generate the characteristics of different languages (Ganesan and Sherman, 1998). Using the frequency of short word occurrences (Klukowski, 1991 and Ingle, 1991), the independent as well as joint probability of symbols (Rau, 1974), n-grams (Beesley 1988; Cavner and Trenkle, 1994), n-graphs, i.e. group of n-words (Batchelder, 1992), diacritics and special characters (Newman, 1987) may all help extract the linguistic characteristics. Exploiting statistics might be extended to incorporate syllable characteristics (Mustonen, 1965) and morphology and syntax (Ziegler, 1991) of texts under suspect. Being a statistical approach, Hidden Markov Model (HMM) is employed in language discrimination studies, too (Preez and Weber, 1996).

Another approach in language discrimination is to use artificial intelligence system with neural networks. In this way, the system can learn characteristics of a language and the language of unknown texts might be determined (Braum and Levkowitz, 1998).

Our approach towards language identification incorporates the Prediction by Partial Matching (PPM) algorithm. The premises of our study are cryptographic, as well as statistical: It is cryptographic in that, PPM is basically a lossless compression tool which is used to encode texts into formats that are hard to recover, as long as the algorithm itself is hidden. It is also statistical because, PPM extracts a statistical model out of a given text (called training text) and uses this model to decide whether the text in question belongs to the same language as that of the text from which the model was extracted. Teahan used PPM technique to discriminate the language of Bibles written in six

Celikel E. (2005).
A CRYPTOGRAPHIC APPROACH TO LANGUAGE IDENTIFICATION: PPM.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 213-219
DOI: 10.5220/0002556102130219

languages as English, French, German, Italian, Latin and Spanish (Teahan, 2000).

2 COMPONENTS

The scheme we designed and developed for language identification is made up of an adaptive statistical model PPM, followed by an encoder, Arithmetic Coding. In the following subsection, these components are introduced. We then give the details of Corpora we used during our implementation.

2.1 PPM Model

PPM is a compression technique, making use of a statistical model of the input text. It was first introduced by Cleary and Witten (Cleary and Witten, 1984), and was improved by Moffat (Moffat, 1990). The algorithm reads the upcoming symbol from the input text, assigns a probability to it and sends it to an adaptive encoder, i.e. the Arithmetic Coder. The probabilities assigned to each symbol are determined by blending together and adaptively updating the several fixed order context models (Nelson, 1991).

PPM is a bijective algorithm. By employing fixed input symbol sizes, it is capable of interpreting any collection of bytes as valid compressed input. This property has positive implications for compression efficiency and security: there is no way to distinguish random data from valid output (Nodeworks Encyclopedia URL).

PPM employs a context of length k to predict and assign a probability to the up-coming symbol. If the upcoming symbol is a new one that has never occurred so far, then the length-k context cannot predict it. In this case, an escape symbol is issued and the model is transferred to a lower level, i.e. length k-1 context. This reduction may continue until a non-first-time symbol is reached. When the upcoming symbol is determined successfully, Arithmetic Coder is then used to encode it (Teahan, 1998).

Arithmetic Coding is an encoding technique providing compression rates close to the language entropy. Arithmetic Coding is considered to be optimal because, on the average, it is not possible to encode better than the entropy rate (Witten, Moffat and Bell, 1999). The algorithm for Arithmetic Coding is as follows:

```
procedure ArithmeticCoding()
set low to 0.0 and high to 1.0
while there is more input symbol do
    get an input symbol
    code_range = high - low
    high=low + range * high_range(symbol)
    low = low + range * low_range(symbol)
end of while
output low
end procedure
```

Arithmetic Coding is achieved by representing a stream of input symbols as a floating point number between 0 and 1, instead of replacing an input symbol with a new symbol. As the text is encoded, this interval narrows as much as the frequency of the symbols in the source text. Source message symbols with high frequency narrows the interval less as compared to the low frequency source message symbols. In this manner, high frequency source message symbols require less number of bits in the output sequence.

2.2 Language Identification System

The language identification system we designed and developed is made up of two components as a statistical model generator and an encoder (Fig. 1). There are two inputs to the system: One being the text whose language is already known and the other is the text in question (the disputed text), i.e. whose language is to be determined. Within the system, the former is called the training text and is fed into the PPM model, and the latter is called the plaintext and is fed into the encoder itself.

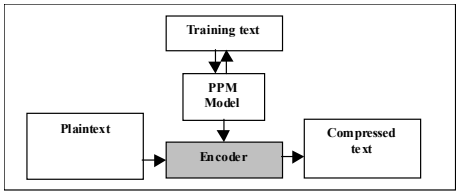

Figure 1: The Language identification system

During implementation, we incorporated the PPM algorithm as follows: We used *training text* to gather symbol frequencies and out of these frequencies, we com-pressed different texts whose languages are of question. When the languages of the *training text* and disputed text are the same, the compression performance of PPM is expected to be better. But there are some effects as linguistic characteristics, type and length of texts, etc. that might affect the performance of our system.

In order to evaluate our scheme's rate of success in determining the right language, we employed the cross-entropy. The cross-entropy H(L,M) is defined as an upper bound to the actual entropy H(L) of a

language by using a probability model M. We need the cross-entropy because, the exact probability distribution of a language is never certain. Hence, H(L) is just a theoretical value. The cross-entropy of a language L calculated using a model M is formulated in Eq. 1. The entropy and cross-entropy are both measured in bits per character – bpc.

$$H(L,M) = -\sum p_M(x_1, x_2, ..., x_m) \log p_M(x_1, x_2, ..., x_m) \quad (1)$$

In Eq. 1, the entropy of a language L, i.e. H(L) is calculated as follows (Eq. 2):

$$H(L) = \lim_{m\to\infty} -\frac{1}{m} \sum p(x_1, x_2, ..., x_m) \log p(x_1, x_2, ..., x_m) \quad (2)$$

Entropy H in Eq. 1 is the average number of bits per symbol needed to encode a message (Shannon, 1948).

$$H(P) = -\sum_{i=1}^{k} p(x_i) \log p(x_i) \quad (3)$$

Assuming that the probabilities summing to 1 are independent and the k possible symbols have a probability distribution P=p(x_1), p(x_2), …, p(x_k), the entropy can be calculated as below (Eq. 3):

For this study, we incorporated Corpora from five European languages as English, French, Turkish, German and Spanish. These languages differ from each other to varying degrees. For example, while English, French and Spanish are more alike, Turkish is totally different from these languages and German has some common linguistic characteristics with the above mentioned group of three. The alphabet size for these languages are 26 letters for English and French, 29 letters for Turkish and 30 letters for both German and Spanish.

All languages of concern in this study are analytical, agglutinative and fusional (having affixes). Analytical languages either does not combine concepts into single words at all (like in Chinese) or does so economically as is the case in English and French. The sentence itself is of primary concern in analytical languages, while the word is of minor interest. Turkish is synthetic and a free constituent order language, morphologically extendible with the its rich set of derivational and inflectional suffixes. In a synthetic language, the concepts cluster more thickly, the words are more richly chambered, but there is a tendency to keep the range of concrete significance in the single word down to a moderate compass (Sapir, 1921). German has extensive use of inflectional endings and compound words are quite common (German Linguistic URL). Spanish has quite the same linguistic characteristics as French, with higher average word length.

Our Corpora consists of these five different languages and each language has a group of seven texts. The texts within each language have been deliberately selected from different essay categories to reflect the changes of the style into our language discrimination implementation. These categories are novel, technical document, poetry, manual, theatre text, Holy book (Bible or Qoran) and a dictionary/encyclopedia. We first based our Corpora construction on text files from the standard English Cor-pus Canterbury (Canterbury Corpus). Modeling after the Canterbury Corpus, we then compiled Corpora from the other four source languages as French, Turkish, German and Spanish. The texts in Turkish Corpus are from Celikel (Celikel, 2004), and the rest of the texts in other languages are all from Internet. The sizes and contents of each Corpus are listed in Tables 1 through Table 5:

Table 1: English Corpus

ENGLISH		
File	File size (bytes)	Explanation
E1	152,089	Novel: Lewis Carroll's "Alice in Wonderland"
E2	426,754	Technical document
E3	481,861	English poetry
E4	4,227	GNU manual
E5	125,179	Theatre text of the play "As You Like it"
E6	4,047,392	Bible in English
E7	2,473,401	World Fact Book of CIA

Table 2: French Corpus.

FRENCH		
File	File size (bytes)	Explanation
F1	871,286	Novel: Jules Verne's "20000 Leagues under the Sea"
F2	66,049	Technical document
F3	185,205	French poetry
F4	32,428	GNU manual
F5	135,477	Theatre text of the play "Tartuffe" by Molière
F6	4,669,107	Bible in French
F7	51,521	French dictionary of computers

Table 3: Turkish Corpus

TURKISH		
File	File size (bytes)	Explanation
T1	167,799	Novel: Ataturk's Discourse
T2	9,664	Technical document
T3	59,386	Turkish poetry
T4	18,526	GNU manual
T5	113,545	Theatre text of the play "Galilei Galileo"
T6	937,532	Quran in Turkish
T7	765,624	Online Philosophy terms disctionary

Table 4: German Corpus

GERMAN		
File	File size (bytes)	Explanation
G1	716,800	Novel
G2	25,872	Technical document
G3	105,868	German poetry
G4	4,622	GNU manual
G5	100,712	Theatre text of the play "Faust"
G6	4,359,878	Bible in German
G7	157,289	Online dictionary of medical terms

Table 5: Spanish Corpus

SPANISH		
File	File size (bytes)	Explanation
S1	63,170	Novel "Oración Cívica" by Gabino Barreda
S2	25,295	Technical document
S3	129,005	Spanish poetry
S4	5,865	GNU manual
S5	158,991	Theatre text of the play "Las Mocedades Del Cid"
S6	4,126,848	Bible in Spanish
S7	216,267	Online ophtalmology dictionary

3 RESULTS

To discriminate among languages, we applied the PPM model on texts from each Corpus. During implementation, we repeated the language identification experiments on each text file. Within each language set, we employed each of seven text files as the training text to PPM to compress the texts of the whole Corpora. Since there are five different languages, it makes 7x5=35 runs for each text; since there are seven texts within each language set, it makes 7x5x7=245 runs for each language; and since there are five different languages, it makes 245x5=1,225 runs in total.

In order to evaluate the performance of our language discriminator, we used the accuracy rate measure (Eq. 4). In this formula, the successes are the cases when both the training text and the

disputed text belong to the same language. All other cases are considered to be as failure.

$$\text{Accuracy rate} = \frac{\text{\# of successes}}{\text{total \# of experiments}} \quad (4)$$

In Table 6 below, language determination experiment results with training texts as-signed from the English Corpus are given. The rows having the lowest bpc rates for English at each row are depicted as success cases. If so, we conclude that the text in question is in English, as well. Unfortunately, this is not the case all the time: There are some rows in Table 6, where the lowest compression rates with PPM are obtained with training texts in other languages than English. These cases are called as fails and are depicted as bold cells. The reason why the language identifier fails might be caused by the insufficient size of the training text or the compressed text (Table 6). According to Table 6, there are 18 failures and 31 successes, yielding the accuracy rate of 63.27%. This is the accuracy rate performed by the PPM model on English.

Repeating the language identification experiments on French texts, we obtained the results in Table 7. Here, the effect of training text size on the performance of the iden-tification scheme is obvious: In cases when the training text size is small (files 2, 4, 5 and 7) the number of failures are greater, while with larger training text (file 6), the system has only two failures. Moreover, when the compressed text is not in French but its size is relatively large, the language identifier system incorrectly decides that it was written in French. This is the effect of compressed text size on the overall system performance. The accuracy rate is 46.94% for French with 23 successes and 26 failures.

Applying our scheme with training texts in Turkish this time, we obtained the bpc values in Table 8. The number of failures for Turkish is 39 while the number of suc-cesses is only 10 out of 49 experiments. Hence, the accuracy rate of our model on Turkish is only 20.41%. This performance degradation in the identification scheme is probably due to the different linguistic characteristics of Turkish.

Table 6: Language identification experiments with training text from English Corpus

File	Language				
	English	French	Turkish	German	Spanish
1	1.36	2.04	2.12	1.78	2.74
2	1.95	2.16	3.78	2.60	2.48
3	2.35	2.64	2.86	2.94	2.69
4	2.86	**2.71**	2.75	3.88	3.72
5	2.45	2.02	2.53	2.63	**1.79**
6	1.61	1.78	1.95	1.76	1.80
7	1.51	2.43	1.91	1.80	2.01
TT	E1	152,089 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.22	2.08	2.17	**1.83**	2.84
2	1.34	2.24	3.81	2.65	2.54
3	2.38	2.73	2.93	3.04	2.78
4	2.69	**2.77**	2.62	3.73	3.61
5	2.54	2.09	2.6	2.71	**1.85**
6	1.63	1.8	1.97	1.77	1.81
7	1.52	2.44	1.93	1.86	2.08
TT	E2	426,754 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.23	2.09	2.18	**1.84**	2.84
2	1.99	2.25	3.88	2.70	2.59
3	1.70	2.72	2.93	3.08	2.75
4	2.91	2.83	**2.78**	3.88	3.71
5	2.39	2.07	2.61	2.73	**1.85**
6	1.61	1.80	1.98	1.78	1.82
7	1.53	2.52	1.94	1.87	2.07
TT	E3	481,861 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.19	1.99	2.01	**1.73**	2.55
2	1.93	2.00	3.52	2.39	2.27
3	2.33	2.51	2.67	2.76	2.53
4	0.88	2.48	**2.62**	3.58	3.44
5	2.47	1.88	2.40	2.45	**1.68**
6	1.60	1.76	1.91	1.74	1.78
7	1.49	2.26	1.87	1.70	1.92
TT	E4	4,227 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.17	2.03	2.11	**1.78**	2.73
2	1.95	2.14	3.77	2.58	2.47
3	2.31	2.62	2.83	2.93	2.66
4	2.85	**2.69**	2.73	3.83	3.70
5	1.47	1.99	2.52	2.60	1.77
6	1.60	1.78	1.94	1.76	1.80
7	1.50	2.41	1.91	1.79	2.00
TT	E5	125,179 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.35	2.25	2.27	**1.97**	3.11
2	2.17	2.45	3.98	2.87	2.78
3	2.52	2.71	3.03	3.28	2.94
4	3.06	**3.13**	2.92	4.10	4.00
5	2.43	2.25	2.71	2.89	1.96
6	1.43	1.87	2.04	1.86	1.88
7	1.63	2.76	2.01	1.98	2.23
TT	E6	4,047,392 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.49	2.20	2.32	**1.92**	3.09
2	2.11	2.41	3.98	2.86	2.80
3	2.58	2.92	3.14	3.25	2.98
4	3.02	3.05	**2.89**	3.96	3.80
5	2.77	**2.28**	2.76	2.91	2.01
6	1.70	1.85	2.03	1.81	1.86
7	1.07	2.69	2.02	2.03	2.26
TT	E7	2,473,401 bytes			

Table 7: Language identification experiments with training text from French Corpus

File	Language				
	English	French	Turkish	German	Spanish
1	2.45	1.54	2.19	1.86	2.92
2	2.13	2.03	3.85	2.72	2.67
3	2.41	2.39	2.95	3.09	2.83
4	3.41	2.10	2.98	3.92	3.72
5	2.76	1.79	2.63	2.74	1.89
6	**1.66**	1.78	1.99	1.79	1.85
7	**1.56**	2.23	1.95	1.89	2.15
TT	F1	871,286 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.27	1.99	2.07	1.76	2.65
2	1.98	1.18	3.69	2.50	2.39
3	**2.37**	2.51	2.77	2.85	2.61
4	3.21	2.39	2.78	3.74	3.57
5	2.57	1.88	2.47	2.54	**1.74**
6	**1.61**	1.77	1.93	1.75	1.79
7	**1.50**	2.27	1.89	1.75	1.97
TT	F2	66,049 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	3.40	1.97	2.13	**1.79**	2.75
2	2.04	2.01	3.80	2.61	2.53
3	2.42	1.60	2.87	2.97	2.70
4	3.34	2.34	2.90	3.88	3.71
5	2.65	1.81	2.55	2.65	**1.80**
6	**1.65**	1.76	1.96	1.76	1.80
7	**1.52**	2.30	1.91	1.80	2.04
TT	F3	185,205 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.25	1.97	2.06	**1.75**	2.60
2	1.96	1.94	3.64	2.46	2.34
3	**2.36**	2.48	2.74	2.83	2.58
4	3.08	1.22	2.69	3.56	3.32
5	2.54	1.83	2.46	2.51	**1.72**
6	**1.60**	1.76	1.92	1.75	1.79
7	**1.50**	2.13	1.88	1.74	1.95
TT	F4	32,428 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.31	1.97	2.10	**1.77**	2.69
2	2.00	1.98	3.72	2.55	2.44
3	**2.39**	2.44	2.85	2.90	2.65
4	3.25	2.30	2.86	3.83	3.63
5	2.59	1.23	2.51	2.59	1.77
6	**1.61**	1.76	1.94	1.75	1.80
7	**1.51**	2.28	1.90	1.78	2.00
TT	F5	135,477 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.61	2.08	2.28	**1.95**	3.18
2	2.28	2.16	4.10	2.90	2.89
3	2.66	2.51	3.08	3.25	3.02
4	3.69	2.29	3.17	4.19	4.01
5	2.93	1.84	2.74	2.89	2.02
6	1.72	1.60	2.06	1.87	1.96
7	**1.64**	2.48	2.02	1.99	2.33
TT	F6	4,669,107 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.27	1.99	2.07	**1.76**	2.64
2	**1.96**	2.01	3.65	2.47	2.36
3	**2.37**	2.53	2.78	2.86	2.61
4	3.10	2.30	2.74	3.64	3.47
5	2.56	1.90	2.48	2.55	**1.74**
6	**1.61**	1.77	1.93	1.75	1.79
7	1.50	1.08	1.89	1.75	1.96
TT	F7	51,521 bytes			

For German, we repeated the experiments and gathered the figures in Table 9. Whenever the training texts are in German, the performance of the PPM model in discriminating the languages of texts are not quite good. It has 36 failures and only 13 successes, yielding an accuracy rate of 26.53%.

Lastly, we run our language discriminator with training texts assigned from the Spanish language set and recorded the results on Table 10. According to the values in Table 10, our scheme yielded 34 failures versus 15 successes. So, the accuracy rate indicated by our system for Spanish is 30.61%.

To observe the overall success rate of our scheme on the five languages of concern, we have plotted the accuracy rate vs. language chart in Fig. 2.

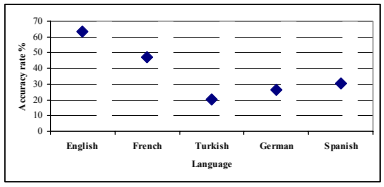

Figure 2: Accuracy rate of the language identification system for different languages.

According to Fig. 2, the language identifier employing the PPM model achieves the best accuracy rate with training texts in English (with 63.27%). The system indicates accuracy rates of 46.94% with French, 30.61% with Spanish, 26.53% with German and the worst performance as 20.41% with Turkish. The difference in performance levels might probably occur due to the characteristics of source languages, as well as the category, i.e. the type, together with the different time span each text belongs to.

Table 8: Language identification experiments with training text from Turkish Corpus

File	Language				
	English	French	Turkish	German	Spanish
1	2.33	2.03	1.20	1.78	2.74
2	**2.00**	2.15	3.23	2.60	2.49
3	**2.40**	2.62	2.64	2.93	2.68
4	3.42	2.73	2.73	3.93	3.84
5	2.63	2.00	2.33	2.61	**1.78**
6	**1.61**	1.78	1.92	1.75	1.80
7	**1.52**	2.44	1.86	1.80	2.01
TT	T1	167,799 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.22	1.99	1.99	**1.74**	2.58
2	1.94	2.03	0.94	2.42	2.31
3	**2.34**	2.53	2.63	2.79	2.55
4	3.18	**2.54**	2.60	3.67	3.54
5	2.51	1.90	2.36	2.48	**1.69**
6	**1.60**	1.76	1.90	1.74	1.78
7	**1.49**	2.30	1.86	1.72	1.93
TT	T2	9,664 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.29	2.02	2.01	**1.76**	2.68
2	**1.98**	2.11	3.33	2.55	2.44
3	2.38	2.59	1.25	2.88	2.63
4	3.34	**2.66**	2.74	3.86	3.74
5	2.59	1.98	2.33	2.56	**1.75**
6	**1.61**	1.77	1.91	1.75	1.79
7	**1.51**	2.40	1.87	1.77	1.98
TT	T3	59,386 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.22	2.00	2.02	**1.74**	2.59
2	**1.94**	2.04	3.36	2.42	2.31
3	**2.34**	2.54	2.68	2.81	2.57
4	2.91	2.52	0.89	3.25	3.15
5	2.51	1.92	2.41	2.49	**1.71**
6	**1.60**	1.77	1.91	1.74	1.78
7	**1.49**	2.29	1.87	1.73	1.94
TT	T4	18,526 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.31	2.03	1.97	**1.77**	2.72
2	**2.00**	2.14	3.15	2.59	2.48
3	**2.39**	2.61	2.52	2.92	2.66
4	3.41	**2.71**	2.75	3.93	3.82
5	2.61	1.99	1.26	2.59	1.77
6	**1.61**	1.77	1.90	1.75	1.79
7	**1.51**	2.44	1.85	1.79	2.00
TT	T5	113,545 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.44	2.08	2.07	**1.83**	2.90
2	**2.08**	2.28	3.40	2.76	2.66
3	**2.47**	2.75	2.62	3.08	2.81
4	3.58	**2.88**	2.90	4.11	4.04
5	2.75	2.11	2.33	2.75	**1.86**
6	1.63	1.80	1.44	1.78	1.82
7	**1.54**	2.59	1.93	1.86	2.12
TT	T6	937,532 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.42	2.08	1.98	**1.82**	2.87
2	**2.06**	2.22	3.07	2.69	2.59
3	**2.47**	2.71	2.65	3.06	2.79
4	3.48	2.81	2.75	3.88	3.87
5	2.73	2.09	2.29	2.72	**1.86**
6	**1.63**	1.79	1.96	1.77	1.82
7	1.54	2.52	1.33	1.89	2.08
TT	T7	765,624 bytes			

Table 9: Language identification experiments with training text from German Corpus

File	Language				
	English	French	Turkish	German	Spanish
1	2.42	2.09	2.18	1.34	2.84
2	**2.08**	2.25	3.83	2.34	2.58
3	**2.50**	2.73	2.93	2.84	2.75
4	3.41	**2.86**	2.95	2.89	3.80
5	2.74	2.08	2.61	2.49	**1.83**
6	**1.64**	1.80	1.99	1.78	1.81
7	**1.54**	2.54	1.94	1.83	2.07
TT	G1	716,800 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.25	2.00	2.05	1.73	2.61
2	1.96	2.06	3.62	0.94	2.09
3	**2.35**	2.56	2.74	2.77	2.58
4	3.16	**2.57**	2.72	3.11	3.53
5	2.54	1.93	2.45	2.45	**1.71**
6	**1.60**	1.77	1.92	1.74	1.79
7	**1.50**	2.31	1.88	1.71	1.95
TT	G2	25,872 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.34	2.04	2.12	1.75	2.73
2	**2.01**	2.16	3.81	2.42	2.49
3	2.41	2.64	2.85	1.53	2.67
4	3.34	**2.73**	2.89	3.28	3.79
5	2.64	2.01	2.54	2.31	**1.78**
6	**1.62**	1.78	1.95	1.74	1.79
7	**1.51**	2.45	1.91	1.77	2.01
TT	G3	105,868 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.21	1.99	2.02	1.72	2.55
2	**1.94**	2.01	3.51	2.27	2.27
3	**2.33**	2.52	2.68	2.74	2.53
4	3.04	2.48	2.57	1.00	2.95
5	2.50	1.89	2.41	2.42	**1.68**
6	**1.60**	1.76	1.91	1.74	1.78
7	**1.49**	2.26	1.87	1.69	1.92
TT	G4	4,622 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.33	2.03	2.11	1.74	2.71
2	**2.00**	2.15	3.79	2.39	2.47
3	**2.40**	2.63	2.85	2.64	2.65
4	3.34	**2.71**	2.78	3.31	3.78
5	2.62	2.00	2.53	1.39	1.77
6	**1.61**	1.78	1.94	1.74	1.79
7	**1.51**	2.44	1.90	1.76	2.00
TT	G5	100,712 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.62	2.21	2.30	1.97	3.05
2	**2.21**	2.41	4.05	2.67	2.79
3	**2.66**	2.91	3.09	2.76	2.91
4	3.65	**3.09**	3.17	3.36	4.16
5	2.95	2.23	2.75	2.39	**1.93**
6	1.72	1.88	2.07	1.55	1.87
7	**1.60**	2.74	2.03	2.02	2.20
TT	G6	4,359,878 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.32	2.03	2.12	1.77	2.71
2	**2.00**	2.13	3.76	2.45	2.46
3	**2.40**	2.62	2.85	2.88	2.65
4	3.33	**2.70**	2.89	3.44	3.75
5	2.62	1.99	2.53	2.55	**1.78**
6	**1.61**	1.78	1.94	1.75	1.79
7	1.51	2.42	1.91	0.91	1.99
TT	G7	157,289 bytes			

Table 10: Language identification experiments with training text from Spanish Corpus

File	Language				
	English	French	Turkish	German	Spanish
1	2.28	2.02	2.08	1.76	1.4
2	1.98	2.1	3.67	2.51	2.23
3	2.37	2.58	2.78	2.87	2.44
4	3.22	**2.61**	2.79	3.74	2.91
5	2.57	1.96	2.49	2.54	**1.64**
6	1.61	1.77	1.93	1.75	1.78
7	1.51	2.37	1.89	1.76	1.91
TT	S1	63,170 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.23	2	2.04	1.74	2.52
2	1.95	2.04	3.6	2.20	0.92
3	2.35	2.55	2.73	2.81	2.53
4	3.1	**2.54**	2.7	3.66	3.08
5	2.53	1.92	2.44	2.5	**1.69**
6	1.6	1.77	1.92	1.74	1.78
7	1.5	2.3	1.88	1.73	1.91
TT	S2	25,295 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.33	2.04	2.13	**1.78**	2.40
2	2.01	2.15	3.77	2.58	2.28
3	2.40	2.63	2.86	2.92	1.49
4	3.33	2.69	**2.89**	3.86	3.09
5	2.62	2.01	2.54	2.60	**1.62**
6	1.61	1.78	1.95	1.75	1.78
7	1.52	2.44	1.91	1.79	1.96
TT	S3	129,005 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.21	1.99	2.02	**1.73**	2.48
2	1.94	2.00	3.51	2.38	2.17
3	2.33	2.52	2.68	2.77	2.50
4	3.01	2.45	**2.55**	2.94	1.02
5	2.49	1.89	2.41	2.46	**1.66**
6	1.60	1.76	1.91	1.74	1.78
7	1.49	2.25	1.87	1.71	1.90
TT	S4	5,865 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.31	2.03	2.11	**1.77**	2.47
2	2.00	2.14	3.74	2.56	2.34
3	2.39	2.62	2.83	2.90	2.43
4	3.31	**2.68**	2.87	3.87	3.22
5	2.60	1.99	2.52	2.58	1.08
6	1.61	1.78	1.94	1.75	1.78
7	1.52	2.42	1.91	1.79	1.97
TT	S5	158,991 bytes			

File	Language				
	English	French	Turkish	German	Spanish
1	2.56	2.29	2.29	**1.91**	2.38
2	2.23	2.46	4.02	2.86	2.56
3	2.62	2.96	3.08	3.19	2.49
4	3.62	**3.11**	3.15	4.19	3.17
5	2.91	2.27	2.75	2.84	1.68
6	1.70	1.92	2.06	1.82	1.56
7	1.61	2.80	2.02	1.96	2.23
TT	S6	4,126,848 bytes			

File	English				
	English	French	Turkish	German	Spanish
1	2.34	2.05	2.13	**1.79**	2.54
2	2.03	2.19	3.80	**2.60**	2.26
3	2.41	2.67	2.87	2.95	2.60
4	3.26	2.74	2.86	3.81	3.06
5	2.63	2.03	**2.55**	2.62	1.76
6	1.62	**1.79**	1.96	1.76	1.80
7	1.52	2.45	1.91	1.79	1.16
TT	S7	216,267 bytes			

4 CONCLUSION AND FUTURE WORK

We have designed and implemented a scheme incorporating the Prediction by Partial Matching (PPM) model. We applied this scheme which is both a cryptographic and statistical tool, in language identification problem. We have implemented our system on texts from five different languages as English, French, Turkish, German and Span-ish. Results revelaed that, the cryptographic tool we introduced achieves accuracy rates of 63.27% , 46.94%, 20.41%, 26.53% and 30.61% for English, French, Turkish, German and Spanish, respectively. Experiments with the scheme can be extended with equally sized training as well as disputed text sizes to eliminate the size effect. Implementations further be carried on with more languages belonging to various language families. The performance of our system can be compared with that of the existing automatic language identification methods by running the algorithms on the same data set.

REFERENCES

Batchelder, E.O., 1992. A learning experience: Training an artificial neural network to discriminate languages. Technical report.

Beesley, K.R., 1988. Language identifier: A computer program for automatic natural-language identification on on-line Text. In proceedings of the 29th annual conference of the American translators association, 47-54.

Braum, J., Levkowitz, H., 1998. Automatic language identification with perceptually guided training and recurrent neural networks. In Int'l conf. on spoken language processing (ICSLP 98), Sydney, Australia.

Canterbury Corpus http://www.Corpus.canterbury.ac.nz

Cavner, W.B., Trenkle J. M., 1994. N-gram based text categorization. In proceedings of the 3rd annual symposium on document analysis and information retrieval, 261-269.

Celikel, E., 2004. *The compression and modelling of turkish texts*, PhD Thesis, Ege University, International Computer Institute, Izmir/Turkey.

Cleary, J.G., Witten, I.H., 1984. Data compression using adaptive coding and partial string matching. In IEEE transactions on communications. 32(4), 396-402.

Ganesan R., Sherman A. T., 1988. Statistical techniques for language recognition: An introduction and guide for cryptanalyst. Cryptologia XVII:4, 321-366

German Linguistic URL: http://www.infoplease.com/ce6/society/A0858390.html.

Ingle, N. C., 1991. A language Identification Table. In The Incorporated Linguist, Vol. 15(4), 98-101.

Kulikowski, S., 1991: Using short words: A language identification algorithm. Unpublished Technical Report.

Moffat, A., 1990. Implementing the PPM data compression scheme. In IEEE transactions on communications, 38(11), 1917-1921.

Mustonen, S., 1965. Multiple discriminant analysis in linguistic problems. In statistical methods in linguistics, no: 4, Skriptor Fack, Stockholm.

Newman, P., 1987. Foreign language identification: First step in the translation process. In Proceedings of the 28th annual conference of the American translators association, 509-516.

Nelson, M., 1991. Arithmetic coding + statistical modeling = data compression part 1 – arithmetic coding. Dr.Dobb's Journal.

Nodeworks Encyclopedia URL: http://pedia.nodeworks.com/P/PP/PPM/PPM_compression_algorithm/

Preez, J., Weber, D., 1996. Automatic language recognition using high-order HMMs, In Inter-national

conference on spoken language processing (ICSLP), Sydney, Australia.

Rau, M.D., 1974. Language identification by statistical analysis. Master's thesis, Naval post-graduate school.

Sapir, E., 1921. *Language: An introduction to the study of speech*.

Shannon, C.E., 1948. A mathematical theory of communication. In Bell system technical journal. vol. 27, 623-656.

Teahan, W.J., 1998. Modeling English text. PhD Thesis, Univ. of Waikato, NZ.

Teahan, W.J., 2000. Text classification and segmentation using minimum cross-entropy. In proceedings of RIAO'2000. Vol. 2, Paris, France, 943-961.

Witten, I., Moffat, A. and Bell, T.C., 1999. Managing Gigabytes Compressing & Indexing Documents and Images, 2nd ed., Morgan Kauffman Pub., CA, USA.

Ziegler, D.V., 1991 The Identification of Languages Using Linguistic Recognition Signals. PhD thesis, SUNY Buffalo.