POLYTECHNIC UNIVERSITY OF THE PHILIPPINES
COLLEGE OF COMPUTER AND INFORMATION SCIENCES
STA. MESA, MANILA

Filipino Native Language Identification Using Markov Chain Model and Maximum Likelihood Decision Rule

Submitted By

Garcia, Jedidiah P.

Gecarane, Jay Leonarth F.

Periabras, Redentor A.

Sablan, Paul Vincent D.L.

Prof. Ria Sagum

Subject Professor

Table of Contents

Keywords: language identification, maximum likelihood, native language, language modeling

I. Introduction and its Background

The Philippines is an archipelago that consists of more than 7,000 islands [1]. It is the reason why the Philippines is considered as multilingual country or country with different languages. It is also the reason of having a difficult correlation of each other and makes the country as one.

Filipino language is characterized by variation within every province. Varieties which share similar features diverge from one another to different degrees. Divergent varieties are often referred to as dialects. In some cases, the varieties may be distinct enough that some would consider them to be separate languages. In other cases, the varieties may be sufficiently similar to be considered merely characteristic of a particular geographic region, social grouping, or historical era. Sometimes speakers may be aware of dialect variation and be able to label a particular dialect with a name. The variation may go largely unnoticed or overlooked.

Language identification is one of the pre-processing unit in natural language processing. It is the task of determining an author's language through statistical computing and works by identifying patterns [2]. It became increasingly important, as more and more textual data is making its way all. Language identification for most of the national language of every country already exists but as said, the Filipino language for example is characterized by variations. These variations may affect and cause failure with succeeding pre-processing units of NLP.

Using a language model is one of the popular approach to identify a language. Some of the known modeling techniques are through character n – gram, markov models, naïve baiyes classifiers, support vector machines, and neural networks.

N-gram is the base of language modeling. It is where words or letters are grouped into *n* from the start up to a document or word's termination while moving one step at a time. With a lot of training data, n-gram is able to identify unique characteristics of a language by identifying probability of a word or letter given *n-1* words or letters. Knowing the dialects of the Filipino language, this may work but the fact that these dialects are somehow based from Tagalog, the structure of words may still show potential similarities.

The project aims to create a native language identification tool that recognizes 3 of the 8 major dialects/languages in the Philippines. These languages are Cebuano, Kapampangan and Pangasinense. The Filipino Native Language Identification will use *markov chain* for language modeling and *maximum likelihood decision rule* as method for identifying the native language.

II. Review of Related Works

A. Native Languages

Native language which was also known as *first language* or *mother tongue* is a language that a person has been exposed to from birth [3]. The first language of a child is a part of the personal, social and cultural identity [4]. It brings the reflection and learning of successful social patterns of acting and speaking. One can have two or more native languages, thus being a native bilingual or indeed multilingual. These are usually people from India, Philippines, Malaysia, Singapore and South Africa, where most of them speak more than one language.

A native language is said to be:

1. Based on origin, the languages learned by an individual first;
2. Based on internal identification, the languages an individual is identified as speaker of;
3. Based on external identification, the languages an individual is identified as speaker of, by others;
4. Based on competence, the languages one knows best; and
5. Based on function, the languages one uses most [5].

In Philippines there are more than a hundred languages separately spoken all over the different regions [6] but there are 8 major dialects: Bikol, Cebuano, Hiligaynon or Ilonggo, Ilocano, Kapampangan, Pangasinan, Tagalog, and Waray [7]. These languages are used in different regions of the Philippines as shown in Figure 1.

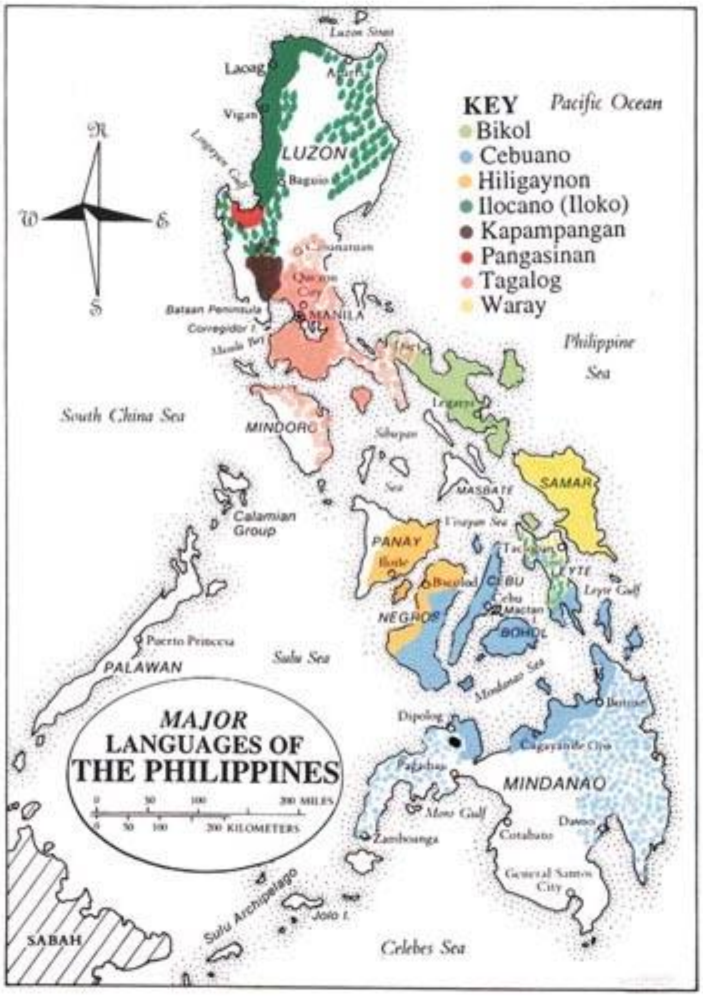


Figure 1. The Major Languages of the Philippines

Although there are 8 varieties, we picked following languages to be identified in our language identification tool assuming that these languages are more distinct from each other.

1. **Cebuano**

   The Cebuano language, often colloquially referred to by most of its speakers simply as Bisaya ("Visayan"; not to be confused with other Visayan languages), is an Austronesian regional language spoken in the Philippines by about 20 million people, mostly in Central Visayas, eastern Negros Island Region, western parts of Eastern Visayas and most parts of Mindanao, most of whom belong to the Visayan ethnic group [8].

Cebuano Language uses “ABAKADA” as their alphabet consisting 15 consonants and 3 vowels. Vowels i/e and o/u are interchangeable meaning that there is no difference in pronunciation between the two of them. The structure of words are likely the same with Tagalog words excluding the use of foreign letters like c (hard c is replaced by k then soft c is replaced by s), f(replaced by p), j(replaced by dy), q (replaced by kw) ,v(replaced by b) ,x (replaced by ks) and z except with the use of foreign words [9].

Cebuano is not a truly-written language, which means that generally, a Cebuano word is spelled just as it is pronounced. Each syllable and vowel is pronounced separately and distinctly. One Cebuano syllable will consist of either a vowel (V), a vowel with a consonant (CV) / (VC), or a vowel between two consonants (CVC). This makes reading in Cebuano much easier than English because the words can be easily broken into syllables [10].

2. Kapampangan

Kapampangan is an Austronesian language spoken in Indung Kapampangan by the ethnic group known as Bangsang Kapampangan or Kapampangan people. It is located in the northern island of Luzon in the Philippines. The language Kapampangan is also called as Amanung Sisuan which means “breastfed” or “nurtured language” [11].

Standard Kapampangan has 21 phonemes, 15 consonants and five vowels. Like other Philippine languages, Kapampangan is a predicate-initial language. The predicate is followed by pronouns and/or adverbs and optionally followed by one or more phrases [12]. Kapampangan seem to be more liberal in converting nouns to verbs as compared to Tagalog and other Philippine Languages. Kapampangan also seem to have no problem in turning adjectives into verbs. Kapampangan speaker can be spotted even if he is speaking Tagalog the moment he verbalizes and adjective [11].

3. Pangasinan

Pangasinan language or Pangasinense is one of the major languages of the Philippines. It is commonly used in the province of Pangasinan and also understood in some places in Benguet and Nueva Ecija and Aeta of Zambales [13]. The Pangasinense belongs to the Malayo – Polynesian languages. It is similar to other languages such as Indonesian in Indonesia, Malaysian in Malaysia, Hawaiian in Hawaii and Malagasy in Madagascar. The Pangasinense is closely related to four small Southern Cordillera languages including Ibaloy, Karaw, Kalanguya and Ilongot [14].

Pangasinense is an agglutinating language [15]. It is capable of forming morphologically complex words by stringing together sequences of affixes before and/or after a root morpheme [16]. Hence, it is also capable of undergoing morphophonemic processes and manifests evidence of morphophonemic changes by sequencing of morphemes and phonemes.

The Pangasinense is rich in adjectives which are capable of making various morphophonemic changes including assimilation and reduplication. The morphophonologically processed adjectives are being recognized by the native speakers which helps in the registry of Pangasinense lexicon [17].

B. Statistical Language Model

It is a model that specifies a priori probability of a particular word sequence in the language of interest. Given an alphabet or inventory of units **Σ** and a sequence of $\boldsymbol{W = w_1,w_2,\ldots,w_n \in \Sigma}$, a language model can be used to compute the probability of *W* based on parameters previously estimated from a training set. Most commonly the inventory **Σ** (also called **vocabulary**) is the list of unique words encountered in the training data [18]. Usually used as training data is corpora – a collection of writings, conversations, speeches, etc. that is used to study and describe language.

C. Language Identification

This is the task of identifying the language used by an author in his or her contexts. The first known approach in language identification is *text categorization* [19] but the best-known model is *per-language character frequency* [20] also known as *n-gram* approach. Variants on this basic method include *Bayesian Models for character sequence prediction* [21], *dot products of word frequency vectors* [22] and *information theoretic measure of document similarity* [23] [24]. *Support vector machines(SVMs)* and *kernel methods* we're also applied to the task of language identification [25] [26]. Notice that these approaches make use of notions of such as "**word**", typically based on the naïve assumption that the language uses white space to delimit words.

D. Native Language Identification

NLI is a fairly recent, but rapidly growing area of research. While some early research was conducted in the early 2000s, most work has only appeared in the last few years. This surge of interest, coupled with the inaugural shared task in 2013 have resulted in NLI becoming a well-established NLP task. The NLI Shared Task in 2013 was attended by 29 teams from the NLP and SLA areas. While there exists a large body of literature produced in the last decade, almost all of this work has focused exclusively on English.

NLI are already implemented in other countries such as Kingdom of Saudi Arabia and Finland. Just like the national languages, native ones are also needed to be identified because for some ways or another although they are derived from their national language, they will always differ in their context or even in their spellings.

In Arabic native languages, Sadat presented a comparative study on dialect identification of Arabic language using social media texts; which is considered as a very hard and challenging task. They studied the impact of the character n-gram Markov models and the Naive Bayes classifiers using three n-gram models, unigram, bi-gram and tri-gram. Their results showed that the Naive Bayes classifier performs better than the character n-gram Markov model for most Arabic dialects. Furthermore, the Naive Bayes classifier based on character bi-gram model was more accurate than other classifiers that are based on character uni-gram and tri-gram. Last, their study showed that the six Arabic dialect groups could be distinguished using the Naive Bayes classifier based on character n-gram model with a very good performance [27].

In ASEAN region, one popular research is the ASEAN MT, a practical network-based service on ASEAN languages text translation. They believe that the significance of communication has increased gradually and will become extreme especially after 2015 when the ASEAN Community begins [28]. Native language identification would be a great help to this kind of project.

III. Aims and Objectives

A. Aims

The varieties of Filipino language have effects and relation to the estate of life, gender, labor, level of education and beliefs. Also the territory, location or space has a part of varieties of Filipino language. The Philippines that consist of different islands are composed of its own native languages.

The Filipino Native Language Identification Tool aims to build a tool that can identify at 3 of the 8 major dialects namely Cebuano, Kapampangan and Pangasinense by using *markov chain* for language modeling and *maximum likelihood decision rule* as method for identifying the native language under the computational linguistics particularly the natural language processing. It can be used in automatic segmentation of input text into blocks of individual languages in case of multiple dialect documents.

B. Objectives

GENERAL

Create a tool that would identify the native language used in Filipino texts either as Cebuano, Kapampangan, or Pangasinense.

SPECIFIC

1. Identify and build a model to each language.
2. Match the Filipino text to the language model through with Markov chain.
3. Select the best matching profile as evaluation function.
4. Assign the matching or winning language label to the document.
5. Evaluate the accuracy of the tool in identifying the native language.

IV. Scope and Limitations

1. The tool will only accept inputs of text such as alphabets, punctuations and numbers that are used in a certain language.
2. It can identify only 3 Filipino major languages such as Cebuano, Kapampangan and Pangasinense.
3. It will only be able to identify monolingual texts.

V. Terminology

- **Austronesian Languages** – are language family that is widely dispersed throughout Maritime Southeast Asia, Madagascar and the islands of the Pacific Ocean, with a few members in continental Asia.

- **N-gram** - a contiguous sequence of *n* items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles**.**

- **Markov Model** - is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the current state not on the events that occurred before it (that is, it assumes the Markov property). Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable.

- **Maximum Likelihood –** also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum [29].

VI. Methods

A. System Architecture

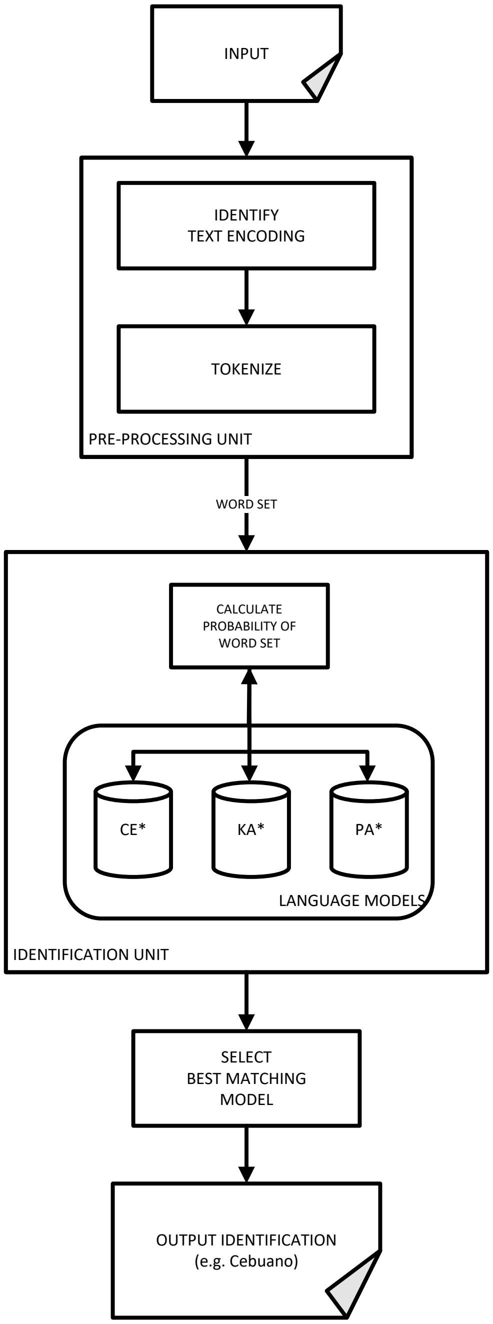


Figure 2. System architecture which shows the process of identifying a document in Filipino Native Language

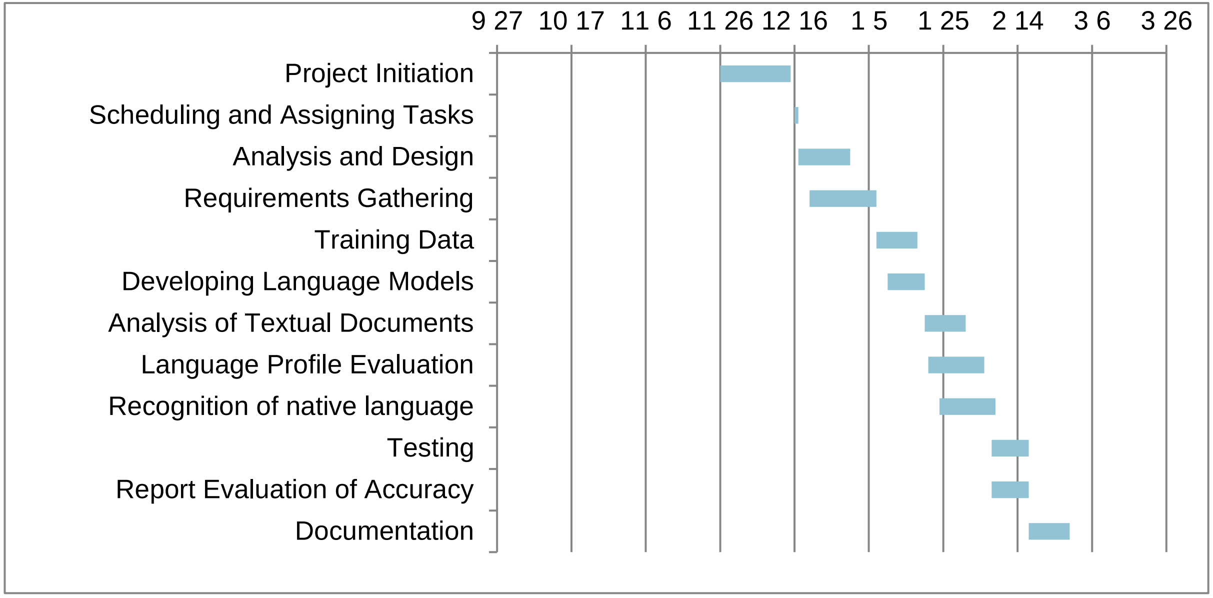
B. Experimental Methodology

In this project, the markov chain language model and maximum likelihood decision rule will be used as an approach for identifying native languages. The training data from the corpus will be organized and classified through markov process. These training data will be used for training. While there will be another collection of text for testing purposes. The markov chain process, is a stochastic process that undergoes transitions from one state to another state space while maximum likelihood decision rule consists of parametric model for all probabilities and denotation of vector parameters which consists of class – conditional distributions. These approaches will help the project tool to test and to train data to achieve higher accuracy.

VII. Project Schedule and Gantt Chart

Task Name	Start	End	Days	Status
Project Initiation	11 26	12 15	19	Complete
Scheduling and Assigning Tasks	12 16	12 17	1	Complete
Analysis and Design	12 17	12 31	14	Complete
Requirements Gathering	12 20	1 7	18	Complete
Training Data	1 7	1 18	11	Complete
Developing Language Models	1 10	1 20	10	Complete
Analysis of Textual Documents	1 20	1 31	11	Complete
Language Profile Evaluation	1 21	2 5	15	Complete
Recognition of native language	1 24	2 8	15	Complete
Testing	2 7	2 17	10	Complete

Report Evaluation of Accuracy	2 7	2 17	10	Complete
Documentation	2 17	2 28	11	Not started



VIII. Assignments

Function	Name	Role(s)
Data Gathering	Garcia, Jedidiah P.	Gather, define, and analyze requirements.
Documentation Control	Gecarane, Jay Leonarth F.	Prepare the necessary documents
Lead Programmer	Periabras, Redentor A.	Responsible for developing and maintaining the program
Data Gathering	Sablan, Paul Vincent D. L.	Gather, define, and analyze requirements.

IX. References

[1] "Geography of the Philippines," Wikipedia, 21 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/Geography_of_the_Philippines. [Accessed March 2017].

[2] "Native Language Identification," Wikipedia, 26 April 2016. [Online]. Available: https://en.wikipedia.org/wiki/Native-language_identification. [Accessed February 21 2017].

[3] L. Bloomfield, Language, Motilal Banarsidass Publ, 1935.

[4] T. Hirst, "The Importance of Maintaining a Childs First Language".

[5] "First Language," Wikipedia, 1 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/First_language. [Accessed 8 February 2017].

[6] P. M. Belvez, "Varieties of Filipino," National Commision for Culture and the Arts, 30 April 2015. [Online]. Available: http://ncca.gov.ph/subcommissions/subcommission-on-cultural-disseminationscd/language-and-translation/varieties-of-filipino/. [Accessed 8 February 2017].

[7] "Major Languages of the Philippines," CSUN.edu, [Online]. Available: http://www.csun.edu/~lan56728/majorlanguages.htm. [Accessed 8 February 2017].

[8] "Cebuano Language," Wikipedia, 2011. [Online]. Available: http://en.wikipedia.org/wiki/Cebuano_language. [Accessed 30 January 2017].

[9] T. Mariking, "Learning Cebuano," 7 November 2005. [Online]. Available: http://www.leaningcebuano.com/files/CebuanoStudyNotes.pdf. [Accessed 30 January 2017].

[10] "Living Cebu," 30 May 2002. [Online]. Available: http://www.livingincebu.com/pdf/cebuano/cebuano_language_objectives.pdf. [Accessed 30 January 2017].

[11] M. R. M. Pangilinan, "An Introduction to Kapampangan Language," Tokyo, 2014.

[12] M. R. Pangilinan and K. Hiroaki, "Motivations for Pamamakmul Amanu "Word Swallowing in Kapampangan"," in *Cross-Linguistic Perspective on the Information Structure of the Austronesian Languages*, Tokyo, Japan, 2013.

[13] "Pangasinan Language," Wikipedia, 2011. [Online]. Available:

http://en.wikipedia.org/wiki/Pangasinan_language. [Accessed 30 January 2017].

[14] R. S. Himes, "The souther Cordilleran group of Philippine languages," *Oceanic Linguistics,* vol. 37, no. 1, pp. 120-77, 1998.

[15] R. G. Gordon, "Ethnologue: Languages of the World, Fifteenth edition," 29 August 2009. [Online]. Available: http://www.ethnologue.com/web.asp. [Accessed 30 January 2017].

[16] C. J. Hall, "An Introduction to Language and Linguistics: Breaking the Language Spell," New York, 2005.

[17] M. A. B. Austria, "Assimilation and Reduplication in Pangasinan Adjectives: A Morphophonemic Analysis," SlideShare, 2012. [Online]. Available: http://www.slideshare.net/shinathrun/assimilation-and-reduplication-in-pangasinan-adjectives. [Accessed 31 January 2017].

[18] D. M. Bikel and I. Zitouni, Multilingual Natural Language Processing Applications, From Theory to Practice, IBM Press.

[19] E. M. Gold, "Language Identification in the limit," *Information and Control,* pp. 447-474, 1967.

[20] W. B. Cavnar and J. M. Trenkle, "N-gram based categorization," in *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*, Las Vegas,, 1994.

[21] T. Dunning, "Statistical Identification of the Language," Computing Research Laboratory, 1994.

[22] M. Darnashek, "Gauging Similarity with N-grams: Language-independent categorization of text," *Science,* pp. 267:843-848, 1995.

[23] J. A. Aslam and F. Meredith, "An information-threoretic measure for document similarity," in *Proceedings of 26th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, Canada, 2003.

[24] B. Martins and M. J. Silva, "Language Identification in web pages," in *Proceedings of the 2005 ACM symposium on Applied Computing*, Santa Fe, USA.

[25] O. Teytaud and R. Jalam, "Kernel-based text categorization," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2001)*, Washington DC, USA, 2001.

[26] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins, "Text Classification

using string kernels," *Journal of Machine Learning Research,* no. 2, pp. 419-444, 2002.

[27] Sadat, "Automatic Identification of Arabic Language Varieties and Dialects in Social Media," 2004.

[28] "Network-based ASEAN Languages Translation Public Service," ASEAN MT, [Online]. Available: http://aseanmt.org. [Accessed 2 February 2017].

[29] E. W. Weisstein, "Maximum Likelihood," MathWorld - A Wolfram Web Resource, [Online]. Available: http://mathworld.wolfram.com/MaximumLikelihood.html. [Accessed 8 February 2017].

[30] A. K. Singh and G. Jagadeesh, "Identification of Languages and Encodings in Multilingual Document".

[31] D. Tran and D. Sharma, "Markov Models for Written Language Identification".