AUTOMATIC LANGUAGE IDENTIFICATION FOR METADATA RECORDS: MEASURING THE EFFECTIVENESS OF VARIOUS APPROACHES

Ryan Charles Knudson BA, MA, MS

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2015

APPROVED:

Jiangping Chen, Major Professor
Rada Mihalcea, Committee Member
Brian O'Connor, Committee Member
Haj Ross, Committee Member
Suliman Hawamdeh Chair of the Department of Library and Information Sciences
Costas Tsatsoulis, Interim Dean of the Tolouse Graduate School

Knudson, Ryan Charles. Automatic Language Identification for Metadata Records: Measuring the Effectiveness of Various Approaches. Doctor of Philosophy (Information Science), May 2015, 92 pp., 11 tables, 9 illustrations, references, 56 titles.

Automatic language identification has been applied to short texts such as queries in information retrieval, but it has not yet been applied to metadata records. Applying this technology to metadata records, particularly their title elements, would enable creators of metadata records to obtain a value for the language element, which is often left blank due to a lack of linguistic expertise. It would also enable the addition of the language value to existing metadata records that currently lack a language value. Titles lend themselves to the problem of language identification mainly due to their shortness, a factor which increases the difficulty of accurately identifying a language. This study implemented four proven approaches to language identification as well as one open-source approach on a collection of multilingual titles of books and movies. Of the five approaches considered, a reduced *N*-gram frequency profile and distance measure approach outperformed all others, accurately identifying over 83% of all titles in the collection. Future plans are to offer this technology to curators of digital collections for use.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Imagine an era when language developed in early humans and branched into multiple mutually unintelligible languages. In such an era, the likelihood of encountering a person who did not communicate with your language would have been low. Today, however, influenced by factors such as the printing press, mass communication, and digital technologies, the likelihood of encountering a person who does not understand your language is much higher.

Knowing the language of a message is of highest import in communication. Whether the message is spoken or written, without knowledge of which language is being used, little can be done with said message. Knowing the language of a digital message is similar, although semantic features may not play as big a role if they are not required for action to be taken on the message. For example, in information retrieval, or IR, correctly identifying a query's language is necessary for retrieving relevant documents, or documents in the query's language. Knowing the language of a message is requisite to responding to the message or taking appropriate action with said message.

Of the 6-7 billion inhabitants of the earth, there are an estimated 6,909 languages used among them (http://www.nationsonline.org/oneworld/languages.htm (One World Nations Online). English has been and continues to be the dominant language online. However, this is changing. The emergence of online documents, including digital libraries, web pages and sites, blogs, news, etc., in languages other than English has grown exponentially in recent years. According to Internet World Stats, the rate of growth for the top ten most used languages on the Internet demonstrates a trend that, if it continues, will soon result in English losing its place as the

dominant language of the Internet. For example, the Chinese language has experienced an Internet growth of 1478.7% from 2000 to 2011, while English has experienced a significantly smaller Internet growth of only 301.4% (http://www.internetworldstats.com/stats7.htm). This proliferation of various languages in the digital world led to the development of research areas such as multi-lingual information retrieval (MLIR) as well as, in part, machine translation (MT). It also served as an impetus to further research in language identification. Concerning MT, language identification is requisite to translation. See, for example, https://translate.google.com, in which Google lists an option to "Detect language" for source language input before translation. This implies the use of a language identification program to identify the language before meaningful translation can be reached. In cases in which a user did not know the source language, yet, nevertheless, desired a translation in a language that he or she could understand, language identification is essential. Of course, this excludes cases in which, as is another option, the user knows the language and can manually supply it to the translation engine.

Problem

Language plays a key role in many, perhaps even most, endeavors involving information. This is particularly true in cataloging and metadata generation. Many metadata schemas include a "language" element of the document or object that describes the metadata record. For example, in the MARC (MAchine-Readable Cataloging) standards, in field 008 for books, which has 40 character positions, positions 35-37 are specifically used for language codes—a three letter code assigned to every known language. Consequently, English is represented by "eng," French by "fre," and German by "ger." Professional catalogers for the Library of Congress, who

utilize MARC standards in their cataloging, are often highly specialized polyglots who collaborate among themselves if necessary for identifying a language of an object or dealing with any other difficulty that arises. Although cases may be rare in which a cataloger encounters problems that cannot be remedied easily through collaboration, these types of problems do occur (S. Miksa, personal communication, September 29, 2014).

What tools could a cataloger in a small, public library who used even the Dublin Core metadata schema utilize to identify the language of an object, digital or otherwise, if such a need arose? How would one who lacked linguistic specialization or a network of collaborative, polyglot peers go about identifying the particular language of an object for which he or she had no familiarity? This problem is clearly stated by Bade (2002) when he said that there are "too few catalogers in the country to do a greatly increasing load of publications in an increasing number of languages" (p. 10). It is a reasonable assumption that the title of an object is reflective of the contents of the same object, and since metadata is most often created in the language of the collection users, e.g., English, the elements other than the title, which may not be in English, is not considered for supplying an accurate value to the language element of a metadata record.

In 1994, Ted Dunning declared: "Given the following 20 character strings,

e pruebas bioquimica

man immunodeficiency

faits se sont produi

it is hardly surprising that a person can identify the languages as Spanish, English, and French, respectively" (p. 1). He further claims, based on this example, that language identification does not necessarily require language understanding, in that even a person who speaks no French or

Spanish can still determine these to be two of the languages in question of the third and first items. While this may very well be the case with a select few languages such as Dunning's example above, consider the following example.

Given the following 20 character strings,

ca urmare publicarea

ezzel a magyar nyelv

sig inte som en egen

would it be ”hardly” surprising for a person to recognize these strings as Romanian, Hungarian, and Swedish, respectively? I believe the answer is that it would be just the contrary—quite surprising for a person to accurately identify the three languages above. Just as in Dunning (1994), the above texts are selected to purposefully exclude any diacritical marks or special characters peculiar to the languages. Even so, consider how many persons with little or no knowledge of Romanian, Hungarian, and Swedish could accurately identify the above languages. The answer is very few, if any. In three cases, I present the three strings above to three separate acquaintances, and in every case, no single string is accurately identified. The fact is, the few persons who could accurately identify the above strings would likely be either polyglots or trained linguists.

With the aforementioned increase in languages present on the World Wide Web and in digital media, it is necessary, more than ever, to be able to accurately identify the languages of these texts in order to process them for any purpose, including obtaining accurate language identification for metadata records creation.

A recent project funded by the Institute of Museum and Library Services (IMLS) (www.imls.gov/), required a group of researchers, including me, to ascertain the language of two million metadata records as part of a plan to translate only English records into both Spanish and Chinese. For this purpose, I developed a language identification program to separate records determined to be English from records determined to be non-English. Experiments with open-source language identification programs proved to result in less accurate language identification than reached by the in-house program.

Identifying the language of metadata records may be for one of two reasons. Firstly, supplying a value for metadata schema that contain a "language" element; and secondly, translation efforts, whether manual or by machine.

Now take a moment to consider the former reason above. Few multilingual digital collections have more than approximately five languages, and even fewer have polyglot staff members to identify languages or translate materials. The World Digital library, e.g., has a collection including contents in over 100 languages and a fully functional interface in seven languages ("About the World", n.d.). Another multilingual digital collection is the International Children's Digital Library, which includes 4619 books in 59 languages and a user interface available in five languages ("Library fast facts", n.d.). However, most libraries and digital collections, as noted above, do not have a multilingual staff necessary for these types of collections.

Consider a public library in suburban Texas where a recent immigrant from Bulgaria has just moved to this suburban area and wishes to locate some readings in his or her mother tongue—Bulgarian. When the recent immigrant searches the collection for Bulgarian titles, with the help of a library staff member, only the metadata records which have an accurate value in the

language element will be returned by such a search. As mentioned above, the language element is not always present in metadata records. Let us say an estimated ten documents in the collection are in fact in Bulgarian, but only three of the ten have a Bulgarian value in the language element. Recall is a commonly used measure in information retrieval that refers to the number of relevant documents retrieved by a search divided by the number of relevant documents in a collection, or: $R = rdr/rdc$ where *R* is recall, *rdr* is relevant documents returned by a search, and *rdc* is relevant documents in collection. The recall value of the hypothetical search above for Bulgarian documents would then be .30, or 30%, which is quite low in terms of state-of-the-art information retrieval. The hypothetical Bulgarian user would then be incapable of using more than three of the ten Bulgarian documents present in the hypothetical collection, short of manually inspecting each title in the entire collection, or performing an exhaustive search.

This problem is indubitably occurring with a frequency correlated to the burgeoning of documents and digital materials in various languages from around the globe. This demonstrates a gap in effective search and retrieval in any digital collection, a gap in which users are incapable of retrieving relevant documents to a query consisting of the language of the content of desired documents.

Research question

In light of the problem above, I answer the following question: Of the various approaches to automatic language identification, which one is most effective for the accurate language identification of metadata records, specifically, the title elements?

To answer this question, five approaches that are suited for language identification are implemented on a test collection for "effectiveness." Effectiveness is measured by accuracy—the number of correctly identified samples/the number of total cases. Answering this question is requisite to determining and supplying an effective and automated method of language identification of metadata records for digital libraries wishing to improve their users' language-related searches.

Significance of the study

Language identification is used on a number of domains, but as of yet, has not been applied to title elements of metadata records. Accurate language identification becomes more difficult to achieve as the number of languages to be identified increase and as the length of the text segments to be identified shortens. For example, if the task is to determine whether a text belongs to one of two languages, e.g., Spanish or English, it is relatively easy, assuming one has sufficient data for training. The task increases in difficulty, however, if the texts to be identified belong to one of three, four, or even 50 or more languages (King, Radev, & Abney, 2014). Also, the length of the text affects the difficulty of language identification (Dunning, 1994). It is much easier to identify the language of a novel-length text than it is to identify the language of a single word or short phrase.

Metadata records embody both of these difficulties associated with language identification, i.e., they are often short segments of text and their language is limited only by the number of languages in the world today. Metadata records, especially the title element of such records, may be in any language of the world, whether the language is still alive and in use today or dead and no longer in use, e.g., Sanskrit. The number of possible languages that could appear in a

library collection is limited only by the number of languages extant on the planet today, whether the languages are thriving and in use or antiquated. The length of the title elements, of special import because the reasonable assumption can be made that the title of an object is most certainly reflective of the language of said object's contents, is quite small. Knudson (2014), considering 800 metadata records from two different digital collections, found the title elements of the records to be an average length of 8.815 words. This illustrates the problem of accurately identifying the language of such elements, in that the shorter the text to be identified, the more difficult it is to accurately identify it. Also of note is the nature of titles, i.e., they are often incomplete, grammatical sentences, but are often comprised of phrases or even single words.

The growing prevalence of library, whether digital or print, objects in multiple languages result in what Bade (2002) calls linguistic errors in cataloging. These errors are due to the lack of linguistic training that modern catalogers receive. Most modern libraries invariably contain some objects that are in a language other than the language of the library. If and when metadata is generated for such objects, it would benefit users at all levels if the metadata were more exhaustive and thorough.

Language identification has been researched with short texts such as queries (Ceylan & Kim, 2009), news text (Cavnar & Trenkle, 1994), etc. However, language identification has not been applied to metadata records or their title elements, a fact that clearly indicates the significance of this study.

Consequently, I have devised a tool—an ope*N*-source language to identify the language of metadata records for digital libraries. I am making this tool available to digital libraries to

provide a more robust retrieval for language-based searches in their collections. The initial identifier uses 21 European languages, with future plans of extending the models to include more languages. This study provides the means for digital libraries to supply an accurate language element to the records in their collections. This, in turn, allows for better information retrieval by users wishing to find documents in a particular language. It also allows a more robust description of the collections' objects. Users of said collections can then achieve more exhaustive results to queries based on the language of the content of desired objects.

Limitations of the study

The limitations of this study are:

1) The languages considered include only 21 European languages;

2) The data used for training the language identifiers are mainly political in nature.

The first of these limitations could easily be overcome by training the language identifiers to identify languages other than European languages. Expanding the dataset to include no*N*-European languages may greatly diminish this limitation.

The specific domain of training data often restricts the test data to be of the same domain. For example, if trying to determine the language of online chatroom logs, the best training data for this task would be taken from chatrooms. The Europarl corpus, used for training in these experiments, is of a political domain. This study reduces this would-be domain limitation by implementing character-level *N*-gram language models rather than word-level. The approaches implemented have been proven and tried in the literature.

Definition of terminology

Throughout the remainder of this thesis, certain concepts will be used which will be based on the following definitions:

Effectiveness—the score of a language identification approach. Effectiveness is the total number of correct classifications / the total number of considered items. It is measured in "accuracy."

Language Identification—a form of text classification designed to determine to which language a particular text belongs. In this study, language identification strictly refers to textual language identification. Various approaches to language identification are commonly used and are discussed in a later section.

N-Gram—a sequence of 1-*N* tokens of text. These tokens may be characters, words, phrases, etc. Most uses of this term throughout refer to character *N*-grams, i.e., 1-*N* characters of text. Cases where tokens other than characters constitute the *N*-grams under discussion are explicitly stated.

Language Model (LM)—a statistical model comprised of sequences of tokens of text, often providing probabilities for the occurrence of the last token of an *N*-gram in light of previously occurring tokens.

N-Gram Frequency Profile—a profile built using a language sample. The profile consists of all character *N*-grams with *N*=1-5 sorted in descending order of occurrences within the language sample. This results in a list of highest to lowest occurring *N*-grams in the language sample, or training data (Cavnar and Trenkle, 1994).

Out-of-Place Measure—an ad hoc statistic computed by taking the difference between ranks of highest-occurring *N*-grams from a training *N*-gram frequency profile and a test *N*-gram frequency profile (Cavnar and Trenkle, 1994).

Distance Measure—a measure for determining how close a test text is to a particular language from training data. This should be done by comparing the *N*-gram frequency profile of the test text to the *N*-gram frequency profiles of the known languages developed in training. In this study, distance measure is calculated by summing the Out-of-Place Measures between the test *N*-gram frequency profile, and a training *N*-gram frequency profile. A distance measure is calculated for each language during training (Cavnar and Trenkle, 1994). For example, in the training data for English and Spanish, the *N*-gram—"es" is the 16th and 22nd highest occurring *N*-gram, respectively. In a test string, the *N*-gram—"es" is the 4th highest occurring *N*-gram. This means that, for this particular *N*-gram—'es', the distance measure for English would be 16-4, or 2, and the distance measure for Spanish would be 22-4, or 8. Of course, as more *N*-grams were considered, the distance measures would be incremented by the new values, and the distance measure for each profile would be the sum of all out-of-place measures for each profile.

Research plan

The purpose of this study is to experiment with different methods of language identification and determine the most suitable for digital libraries to use for identifying the language of metadata records, specifically, the title element. Five approaches, four of which have been tested and proven to be effective in the literature, are implemented, along with various sub-approaches using different-sized segments of language for modeling. These approaches are: An *N*-gram approach, a modified *N*-gram approach, a vector-space model approach, a naïve Bayes

approach, and an open-source approach. In the near future, the best of these approaches will be supplied to, and suggested for, use by digital libraries for identifying the language of their digital objects.

Summary

This chapter presents a problem that exists in digital libraries and their collections, a problem related to poorer retrieval than is necessary if a language element easily obtained for objects within their collections. A solution to this problem is to provide an effective automatic language identifier to obtain an accurate value for supplying the language element of metadata records. Implementation by digital libraries of such an identifier would improve retrieval in cases where a user desires to retrieve objects whose contents are in a particular language.

The following pages contain a review of literature related to language identification and metadata records; the methods used for developing three language identification programs, each with varying numbers of approaches suitable for identifying the languages of metadata records; a pilot study which prompted this investigation; an analysis of the results of the approaches implemented; a comparison of these programs' accuracies with the accuracy of an open-source language identification program, and a discussion.

CHAPTER 2

LITERATURE REVIEW

Introduction

This chapter contains a review of literature about two things: language identification and the use of language in cataloging metadata records. First, I present a brief history of language identification, comprised of commonly-used and effective approaches to language identification, along with recurring problems encountered in the field of language identification. Then I present two popular metadata schemas, Dublin Core and Metadata Object Description Schema (MODS) with a focus on the schemas' "language" element.

Language identification

I consider textual language identification in this study. Textual language identification is different from spoken-language identification, a popular research area in recent years. I include any and all forms of text in any language for which a language model (LM) may have been built or readily available.

"Textual language identification," hereafter "language identification," has been used in many Web and digital applications, ranging from the language identification of the contents of Web pages (Choong, Mikami, & Nagno, 2011; Martins & Silva, 2005) to determining which language a query in an online search belongs (Gottron & Lipka, 2010).

Language identification has been called "a solved problem," but only if dealing with few and clearly distinct languages with clean datasets (McNamee, 2005). However, as the number of languages increase and the length of texts to be identified becomes shorter, i.e., the

identification is at a finer level than the document level, language identification is still an unsolved problem (Baldwin & Lui, 2010; Hughes, Baldwin, Bird, Nicholson, & MacKinlay, 2006). Language identification may be a special form of text classification. Text classification is the task of assigning a text to one or more categories based on a set of predefined features or characteristics. The recent increase in digital text and information causes a higher demand for classifying texts. Text classification is prevalent in recent research, whether it be to classify documents in electronic format (Kim, Han, Rim, & Myaeng, 2006; Liu, Chen, Zhang, Ma., & Wu 2004); to improve relevance of results in a search engine; to filter spam in email servers; to detect opinions (Missen & Boughanem, 2009); to mine reviews of movies (Zhao & Li, 2009), or products (Balahur & Montoya, 2008), or political sentiments (Durant & Smith, 2006); or to analyze sentiment in texts (Polpinij & Ghose, 2008). Each of these types of text classification requires knowledge of the language of the text. Commonly-used approaches include naïve-Bayes and Rocchio classifiers. It is evident, due to the burgeoning of information in multiple languages in all digital arenas that further research into text classification, specifically, language identification, is essential.

Language identification has a long history. Early approaches to language identification were manual, i.e., humans, perhaps librarians, were required to compare uncatalogued texts to guides containing characteristics of a list of languages. Ingle (1976) developed a table comprised of common words in many languages. The table was consulted by an evaluator who looked for the listed common words in the text to be identified. Newman (1987) also implemented a manual language-identification scheme to differentiate between 16 European languages by considering characters and diacritics frequently used on particular letters in the languages. She

also claimed that one did not need be familiar with a language to accurately identify it by following this procedure. Although this method using commonly occurring characters or words for language identification may be intuitive, the cost of employing humans to manually evaluate texts for language identification is proportionate to the amount of unclassified text, which, as noted above, is increasing daily through the Web and all digital media. Another problem with this method is that the accuracy of classification drops inversely in proportion to the size of the text to be identified. In other words, smaller test texts are less accurately identified than larger texts, which is to be expected (Baldwin & Lui, 2010). Grefenstette (1995) compared two language identification schemes, one of which utilized a list of short words, similar to what Ingle (1976) proposed, and the other utilized character trigrams, a method which I explain further below. Grefenstette used these two schemes to determine which of 10 European languages a text belonged. He built his dataset using the European Corpus Initiative (ECI) CD-ROM, made available by elsnet (http://www.elsnet.org/eci.html). He extracted the second-million characters from ten European languages available on this CD-ROM. He then split the data into sentences by tokenizing at sentence-final punctuation ('.', '!', '?'). See Table 1

Accuracy using trigram-character and short-word approaches for English text (Grefenstette, 1995) for a sample of Grefenstette's results for the English language results of this study.

Table 1

Accuracy using trigram-character and short-word approaches for English text (Grefenstette, 1995)

Number of Words	Trigram Approach Accuracy	Short-Word Approach Accuracy
1 or 2	78.9%	52.6%
3-5	97.2%	87.7%
6-10	99.5%	97.3%
11-15	99.9%	99.8%
More than 50	100.0%	100.0%

Notice in Table 1

Accuracy using trigram-character and short-word approaches for English text (Grefenstette, 1995) the sharp drop in accuracy when considering short texts, i.e., those of "1 or 2" words. A language identification method cannot be restricted to use with only large amounts of text. If this were the case, the above "short-words" approach would be suitable. Consequently, other methods must be used in cases where there is a very short text. Information retrieval requires the use of language identification for queries. Spink, Wolfram, Jansen, and Saracevic (2001) found the average query length for online searches to be 2.4 words in length. Language identification is also necessary for determining a value for a language element in a metadata scheme, where the "title" element is the main consideration for such a value. Knudson (2014), who performed language identification on 800 metadata records, found that the average length of the "title" element of the records was 8.8 words. This paucity of words, coupled with a higher number of languages included in the language-identification models, may severely diminish the classification accuracy of a program on such texts.

Not only are the size of texts to be identified and the number of languages considered problematic for language identification; but also important are the fine-grain distinctions between similar or related languages, i.e., languages belonging to the same linguistic family. King, Radev, and Abney (2014) investigated this problem using six sets of similar languages, ranging in number from two to three languages, in an attempt to boost classification performance. They found that a simple naïve Bayes classifier using character and word *N*-gram features yielded the best performance, with an average accuracy of 87.5% over the six tasks. Automatic language identification is a supervised machine learning algorithm that assigns a text to a class, or a language, based on predefined sets of features belonging to this class or language. While language identification of Latin-based script languages is a well-researched field, the spawning of Asian, African, and other languages on the World Wide Web requires, and even demands that research into language identification continue (Choong, Mikami, & Nagno, 2011).

Critical decisions in language identification

When performing language identification, the two most important decisions to be made relate to: 1) the unit of linguistic measurement to be considered, e.g., “characters,” “words,” “pairs of words,” etc., and 2) the algorithm to be implemented for the process, e.g., “*N*-gram frequencies,” “decision trees,” naïve Bayes, etc. These decisions are both logically connected to what type of text is to be identified. In the following text, I discuss the unit of linguistic measurement, as well as different algorithms used for language identification.

Unit of linguistic measurement in language identification

When performing language identification, a crucial decision is determining the unit of linguistic measurement. This decision is influenced by the type and domain of text to be identified, along with the expected length of the texts to be identified. While the choices for unit are many, characters or words are often the choice. Various features have been used in language identification in the past. Table 2 lists studies conducted in language identification using these various features of language.

Table 2

Studies of language identification using varying features of language

Characters of a Language	Words of a Language	*N*-gram Frequencies of a Language
Mustonen (1965)	Ingle (1976)	Beesley (1988)
Newman (1987)	Henrich (1989)	Henrich (1989)
	Kulikowski (1991)	Cavnar and Trenkle (1994)
	Batchelder (1992)	Dunning (1994)
	Souter, Churcher, Hayes, Hughes, and Johnson (1994)	Souter *et al.* (1994)

Dunning (1994), in his study on language identification, reviewed commonly-used approaches of that day. One such approach, the unique strings approach, proposed that *N*-length strings of characters that are peculiar to a language, be used to identify a language. Dunning pointed out the weakness of this approach is the reliance on a very limited number of strings, i.e., the *N* number of strings determined to belong exclusively to a given language, strings which may not show up in a text.

Another approach is to use common words of a language to identify the language. These words are the *N* most frequent words in a language, which are also known as “stop words.” In English, these words would include a, and, the, of, and so on. Dunning (1994) stated that the problem of

this approach, though it may perform well in identifying the language of sufficiently large texts, is again that smaller texts are more likely to be misclassified. Another approach is to use a dictionary, therefore including the majority of words in a given language in the identifier's language model. This approach would be less effective with highly inflectional languages unless there were access to a stemmer. A stemmer is an algorithm that reduces the inflected forms of words to base forms, or stems, in order to facilitate less computationally intensive linguistic processing of such words. For example, the stem of the English words "house," "housed," "houses," and "housing" might be "hous." This approach makes it difficult to classify slang, abbreviations, etc. and likely suffers lower accuracy with shorter texts.

Cavnar and Trenkle (1994) and Dunning (1994) experimented with what would become a more effective method of language identification: an *N*-gram approach. An *N*-gram is a sequence of *N* consecutive tokens of a text, whether these tokens be characters, words, or even sentences. In the case of language identification, these tokens are usually characters or words. *N*-gram models are often used to predict the last gram of a string from the *N* prior one(s) using empirical probabilities calculated from training data. Jurafsky and Martin (2009) refer to such a statistical mechanism as a language model (LM).

N-gram language modeling is based on Zipf's Law, which states that the probability of words or other items starts very high and then tapers off, creating what is known as a Zipfian distribution (Black, 2009). For language, this means that the probability of items in a language, ranging from characters to clauses or even sentences, is high for many reoccurring items, and will taper off to a long tail in the distribution of rarer items. Take, for example, the English alphabet. Analysis of English text shows that "e" is the most commonly occurring letter, while "z" is far less common.

So a sample of English text will have large numbers of the “e” character and far smaller numbers of the “z” character. Zipf’s law also applies to English words, where words such as “the,” “and,” and “to” occur far more often than words such as “transubstantiate,” “migratory”, and “unsustainable.”

To illustrate *N*-grams, consider the string “WORD.” The string is four characters long. Generally, when padded with spaces, a text of length *k* will have *k + 1 N*-grams, whether they are bigrams (comprised of pairs of grams), trigrams (comprised of triplets of grams), and so on (Cavnar & Trenkle, 1994). Below are the bigram, trigram, and quadgram profiles for this example string of text:

Bigrams: _W, WO, OR, RD, D_;

Trigrams: _WO, WOR, ORD, RD_, D_ _;

Quadgrams: _WOR, WORD, ORD_, RD_ _, D_ _ _.

N-gram language-modeling operates under the Markov assumption, which states that a future event in a system (in the case of language, the next occurring character or word) can be predicted by *N*-most recent past events, or previous characters or words in language modeling. So, to estimate the likelihood of the sentence: “Sarah left in my car,” belonging to English, one could combine the probabilities of each *N*-gram’s occurrence within the English training text. If word bigrams were chosen, then the probability of the above sentence’s belonging to English could be computed as follows:

P(<begin>S< end>) = p(w1 | <sentence start>) p(w1 | w2)…p(<end> | wN),

*wh*ere “*S* “= sentence, “*w*” = word, and each probability corresponds to the probability assigned to identical bigrams within the English language model.

N-gram frequency algorithm for language identification

Cavnar and Trenkle (1994) performed language classification by constructing *N*-gram profiles for a number of languages and using an ad hoc statistic they called the "out-of-place measure" to determine which category should be used to classify a test text. In essence, they used a ranked list of highest occurring *N*-grams from training samples, with *N* = 1-5, to build profiles for each category. These profiles were padded to avoid overlapping of words with other words. They then similarly constructed a profile for the test data, and compared the profile with those of the category profiles, classifying the test profile to the category profile having the smallest distance measure. With even the most frequently occurring 400 *N*-grams, several features particular to a language arise. For example, many stop words (the *N* most frequently occurring words of a given language), common prefixes and suffixes, etc., are present in a profile of the top 400 occurring *N*-grams . *Figure 1* illustrates Cavnar and Trenkle's (1994) method of language identification and calculation of the ad-hoc statistic: the out-of-place measure.

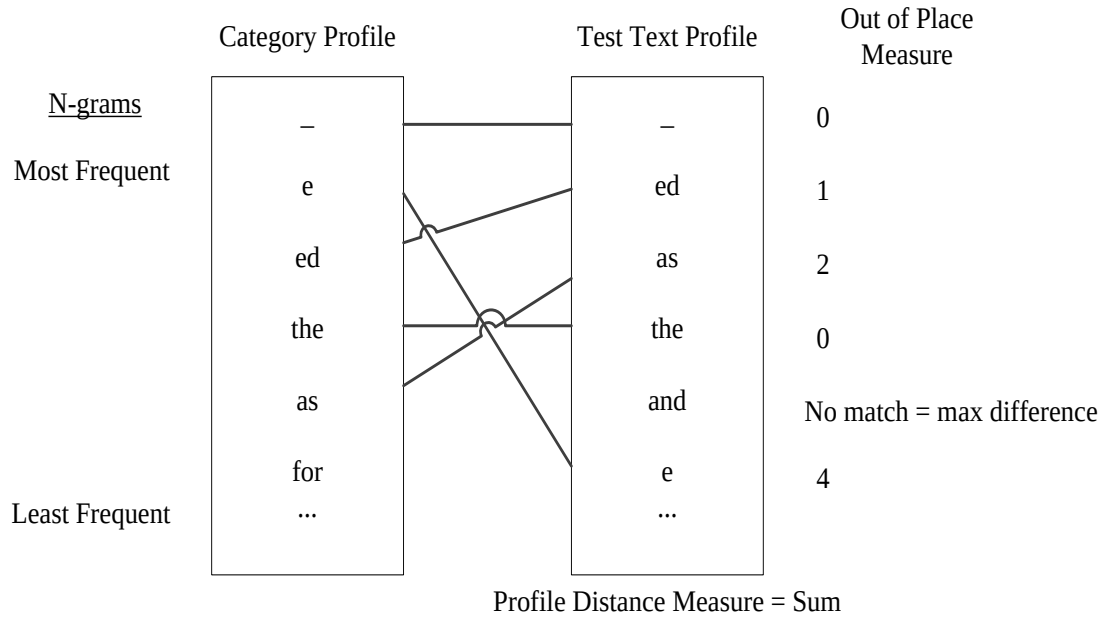

Figure 1. Cavnar and Trenkle's (1994) language identification procedure. These N-grams are for illustrative purposes only and not part of any N-gram profile of this study.

As *Figure 1* demonstrates, the "out-of-place measure" for each of the *N*-grams in the test data is simply the difference in the rank of the frequency of the *N*-grams. For example, the "4" shown for "e" simply means that the category profile's rank of "e" was four places in difference from the test data's rank of the same letter, or *N*-gram. Adding these "out-of-place measures" together, the sum totals to a "profile distance measure," and the lowest "profile distance measure" is judged to be the correct category for classification.

To illustrate this method more clearly, let us say the *N*-grams in the *Figure 1* category profile belong to the English profile constructed with English language training data. This would mean that in the English *N*-gram frequency profile, the highest occurring six *N*-grams are "_", "e," "ed," "the," "as," and "for," in that order. The highest occurring six *N*-grams in the test text would then be "_," "ed," "as," "the," "and," and "e." The out-of-place measure for each class is simply the difference between ranks of all *N*-grams occurring in the test data and all *N*-gram frequency

profiles for all predetermined classes (in this case, including English). If the *N*-grams in *Figure 1* were the only six *N*-grams in the test string, then the distance measure between the English *N*-gram frequency profile and the test string's *N*-gram frequency profile would be the sum of every test text *N*-gram's out-of-place measure with the English *N*-gram frequency profiles. This means that the profile distance measure between *Figure 1*'s test text and the hypothetical English *N*-gram frequency profile would total 13—each of the out-of-place measures added together. Keep in mind that since the *N*-gram "and" occurs in one profile but not the other, the maximum distance is selected, which in this hypothetical example, is six, since there are six *N*-grams total. This distance measure would then be compared to the distance measures between all other *N*-gram frequency profiles and the smallest number would be deemed the "winner," or the correct class for the test text.

Cavnar and Trenkle (1994) experimented with *N*-gram profiles ranging from the top 100 occurring *N*-grams to the top 400 occurring ones, and they found that using *N*-gram profiles of the top occurring 400 *N*-grams resulted in 99.8% accuracy for language identification, misclassifying only 7 of 3478 total documents. Note that these documents were entire newswire articles, which suggests that they were lengthier than queries or titles. They even used the same *N*-gram profiles to classify the subject of test texts, although a bit less accurately, at up to 80%. Since 1994, researchers have recognized the effectiveness of *N*-gram modeling and have utilized, with success, this approach to language identification. The effectiveness of an *N*-gram approach is attested to by the frequency with which it is seen in language identification problems in the literature, for example, cf.: Brown, 2012; Choong, Mikami, & Nagno, 2011;

Deepamala & Ramakanth Kumar, 2012; Hornik, Mair, Rauch, Geiger, Buchta, & Feinerer, 2013; etc.

Dunning (1994) developed and successfully implemented an *N*-gram LM for classifying Spanish and English texts. His model was also applied to genetic sequences, modelling them after languages specific to the animal or organism to which they belonged, with some success. Unlike Cavnar and Trenkle's (1994) implementation of the ad hoc "out-of-place measure," Dunning implemented the Bayes theorem, an algorithm often used in classification, and Laplace, or 'add-one' smoothing. Dunning achieved an accuracy of 92% when using only 20 bytes of test text and 50 bytes of training text and up to an accuracy of 99.9% when the test text is increased to 500 bytes.

Naïve bayes algorithm for language identification.

Another commonly-used algorithm for language identification is the Naïve Bayes approach. This seems suitable since language identification is, after all, a classification task. Naïve Bayes uses features of language, e.g., single words, to estimate the probability of a test text belonging to a preprocessed training text by summing the probabilities of features in the test text seen in the training text. This can be shown as:

$$p(L_e \mid T) = \frac{p(L_i)}{p(T)} \prod_i p(w_i \in T \mid L_i),$$

where *T* is test text, *L* is one language, and *w* is word. This algorithm will be further explained in Methodology. This method has been successfully implemented in language identification. Gottron and Lipka (2010) used a Naïve Bayesian approach and reached accuracy levels of 87.90% for identification of character *N*-grams with *N=1* and as high as 99.44% accuracy for

identification of character *N*-grams with *N=5*. Their data consisted of query-type strings with an average length of 7.2 words of on average 6-7 characters long taken from newswire texts.

Vector-space model algorithm for language identification

Another approach to language identification seen in the literature is a vector-space model approach, which assigns a value to a language's vector-space as well as the test text's vector space and compares the values using the cosine similarity measure. This approach has also been successful in accurately identifying the language of a text. Gottron and Lipka (2010) used a vector space model to identify the language of news headlines at an accuracy level of 54.68% for character *N*-grams with *N=1* and at an accuracy level of 75.37% for character *N*-grams with *N=5.*

Dunning (1994) attempted to distinguish Spanish and English texts ranging in length from 10 bytes to 500 bytes. While Dunning's 99.9% accuracy using Bayesian method with Markov modeling may seem impressive, just as many of the achieved accuracies using various algorithms would, such accuracies would be difficult to achieve were the number of languages considered to increase even slightly (Baldwin & Lui, 2010). For example, if one were to run the same controlled tests on a group of documents comprised of five or six different languages, rather than just two, this level of accuracy would likely not be reached. This accuracy would also be less likely with shorter texts to be identified. The problem of identifying the language of relatively short texts is discussed in Dunning (1994), Baldwin and Lui (2010), and Ceylan and Kim (2009). Ceylan and Kim (2009) worked on identifying the language of search engine queries ranging in length from two to three words.

As noted above, there are an estimated 6,909 languages being used in the world today. And while many of them do not have sufficient textual data to build robust language models, cases may arise in which a digital or print object belongs to a language unfamiliar to a cataloger. These cases are less likely to occur in large libraries, e.g., the Library of Congress, than they are in small-scale libraries (S. Miksa, personal communication, September 29, 2014).

Method of evaluation of language identification

To evaluate how effective language identification is, the measure used is often accuracy. This accuracy is reflective of how many texts out of the total number a program accurately identifies. So, for example, if there are 50 texts to be identified, and a program accurately identifies the language of 30 of these texts, the accuracy is 30/50, or 60%. This is the most common, but not the only measure for evaluation of language identification. For example, Trieschnigg, Hiemstra, Theune, Jong, and Meder (2012) used measures of precision and recall to evaluate language identification of minority and very similar languages in the Dutch Folktale Database. They defined precision as the proportion of predictions for a particular language that were correctly identified. They defined recall as the proportion of documents in a particular language that were correctly identified.

Metadata schema and the language element

According to Reitz, metadata is:

> Literally, "data about data." Structured information describing information resources/objects for a variety of purposes. Although AACR2/MARC cataloging is formally metadata, the term is generally used in the library community for nontraditional schemes such as the Dublin Core Metadata Element Set, the VRA Core Categories, and the Encoded Archival Description (EAD). Metadata has been categorized as descriptive, structural, and administrative. 'Descriptive metadata' facilitates indexing, discovery, identification, and selection. 'Structural metadata' describes the internal structure of

> complex information resources. 'Administrative metadata' aids in the management of resources and may include rights management metadata, preservation metadata, and technical metadata describing the physical characteristics of a resource. For an introduction to metadata, please see Priscilla Caplan's *Metadata Fundamentals for All Librarians* (American Library Association, 2003). Also spelled "meta-data." (2004)
>
> In this study, the metadata considered is "descriptive metadata," which is defined as:
>
> Data about an information resource that is intended to facilitate its discovery, identification, and selection. Descriptive metadata is also used to bring together all the versions of a work in process called collocation, and for acquisition purposes. When viewed as metadata, traditional library cataloging is descriptive, as are such schemes as the Dublin Core Metadata Element Set and the VRA (Visual Resources Association Core. Descriptive metadata is also used for evaluation, both narrative (reviews, etc.) and formal (content ratings; for linkage (relationships between a resource and other things); and for usability (Reitz, 2004).

Many metadata schemes are utilized by libraries for describing the documents of their collections, e.g., Dublin Core is a popular choice for this purpose among libraries. Hillman (2005) stated that often there is a clear choice for determining the value of a metadata element, but that there is occasionally some judgment required from the cataloguer creating the metadata. In the case of a language element value, unless the cataloguer is certain of the language, often the language element, which is not required, is left out, leaving the metadata record lacking this particular element. The following will demonstrate a few metadata formats that make use, or in some cases do not make use, of the language element.

Dublin Core

Dublin Core was the product of collaboration between Yuri Rubinsky, Stuart Weibel, and Eric Miller (all presenting papers) and Terry Noreault and Joseph Hardin at the 2nd International World Wide Web Conference held in Chicago in October of 1994. This collaboration led to an event called OCLC/NCSA Metadata Workshop held in Dublin, Ohio in March of 1995. The aim of

this event was to create a schema that would facilitate easier search and retrieval on the Web ("Dublin Core Metadata", 2014).

Dublin Core has been used since, although it has its opponents, e.g., Beall (2004) claimed that Dublin Core had outlived its usefulness and proposed that MODS was more appropriate for metadata. However, Dublin Core is still utilized by a number of libraries around the world. For example, the University of North Texas' Portal to Texas History (http://texashistory.unt.edu/) uses Dublin Core for its metadata.

Simple Dublin Core has 15 elements, including a language element, but Qualified Dublin Core has expanded upon these. A list of the 15 elements of Simple Dublin Core taken from Zeng and Qin (2008) is shown in Table 3. The language element is not required in Simple Dublin Core, yet can be repeated in cases where an object is in more than one language. One problem with not requiring this element lies in the possibility of a user searching for objects in a particular language. The only results that would show up for this sort of search would be those that had a valid entry for the language element; all others would be lost. Consider a user of a collection who only understands Bulgarian. When a search is conducted, perhaps through an advanced search option for only the language, how many objects will go unnoticed by the user if there are numerous cases of metadata records lacking a correct language element?

Table 3

Simple Dublin core elements in three categories: taken from Zeng and Qin (2008)

Category:	Content	Intellectual Property	Instantiation
Elements:	Title	Creator	Date
	Description	Publisher	Format
	Type	Rights	Identifier
	Subject	Contributor	Language
	Source		
	Relation		
	Coverage		

Metadata Object Description Schema (MODS).

Another recently and popular metadata schema is the Metdata Object Description Schema, or MODS. One goal in the creation of this schema is to allow for a rich description compatible with existing libraries' large numbers of records in MARC (Guenther, 2003). One motivation for creating MODS was to find a compromise between the complexity of MARC and the simplicity of Dublin Core. Another motivation was to accommodate the inclusion of electronic resources (McCallum, 2004). Included in the twenty elements of MODS, just as in Dublin Core, is a language element. This element reflects the language of the content of the object being described by the metadata ("Metadata Object Description Schema", 2013). Just as in Dublin Core, the language element is not required but is optional. This means that it may or may not be available in the metadata record of an object. The case noted above in which a user searches for all objects within a collection in one particular language is applicable here as well. If the collection a user were searching were described using MODS, the chances of retrieving all relevant documents in a search for a specific language may be lessened contingent upon whether the language element were available for the relevant objects. The twenty MODS elements, taken from Zeng and Qin (2008), are listed below.

1. Titleinfo
2. Name
3. TopeOfResource
4. Genre
5. OriginInfo
6. Language
7. PhysicalDescription
8. Abstract
9. TableOfContents
10. TargetAudience
11. Note
12. Subject
13. Classification
14. RelatedItem
15. Identifier
16. Location
17. AccessCondition
18. Part
19. Extension
20. RecordInfo

In many instances, the language element is left blank, or it is incorrectly identified (Bade, 2002). There is also often an option that can be inserted similar to “not available” in instances where the language of a title is unknown to the cataloguer. It is for this reason that it is important to automate this process, not only to lessen the workload of cataloguers, but also to ensure that correct values are supplied and input into the language element’s place. Ensuring the inclusion of the language element facilitates more fruitful and relevant results in cases where a user searches a collection based on the language of the documents or digital objects.

Summary

This chapter shows various approaches to language identification and identifies problems encountered in language identification, including large numbers of languages to be identified

and short texts to be identified. It also illustrates the use of language in descriptive metadata, i.e., Dublin Core and MODS.

Metadata title elements are often short, averaging 63.81125 characters, or 8.815 words per element (Knudson, 2014), and usually consist of fragmented language, e.g., phrases or single words. This makes the title elements of metadata records an ideal text to be researched regarding language identification. This shortness of text, accompanied by the vast number of languages which could comprise a metadata title element, present a problem which has not yet been addressed in the literature, a problem that is often ignored, according to Bade (2002). The following chapter outlines a methodology for reaching a viable solution to this problem. It also illustrates a pilot study to this research in which language identification is used for metadata records in an effort to facilitate machine translation of the records.

CHAPTER 3

METHODOLOGY

Introduction

The purpose of this research is to determine a suitable method for identifying the language of title elements in metadata records. The result of this identification can be used to provide a value for the "language" element for metadata records applying different metadata schemas, or to help machine translation systems to process the title element differently from other elements in cases where the title element is in another language. One issue of determining the language of metadata records is the vast number of languages in which they may occur. Languages that have relatively small amounts of textual data available for training are often more difficult to identify than languages with larger amounts of textual data available (Hughes, Baldwin, Bird, Nicholson, & MacKinlay, 2006; King & Abney, 2013). The nature of metadata is such that the languages of metadata records might possibly be as varied as the languages of online search engine queries. This, and the fact that metadata title elements are often very short in length (Knudson, 2014), means that language identification of metadata title elements is a difficult task.

Research Design

Based on the literature review, I use four commonly and successfully implemented methods—Cavnar and Trenkle's (1994) *N*-gram profile, a modified version of Cavnar and Trenkle (1994), a vector-space model, and naïve Bayes—and develop them in the Perl programming language (https://www.perl.org/) for experimentation and analysis.

My study consists of the following steps: 1) Building a dataset of titles in each of the languages in the language identifier training data; 2) experimenting with approaches to language identification driven by the literature; 3) experimenting with an open-source language identifier; 4) analyzing and comparing the five different approaches with the performance measure of accuracy; and 5) proposing a language identification solution to digital metadata records based on the analysis. The research question to be answered is:

“Of the various approaches to automatic language identification, which one is most effective for the accurate language identification of metadata records, specifically, the title elements?”

Figure 2 illustrates the steps of the research design for this study.

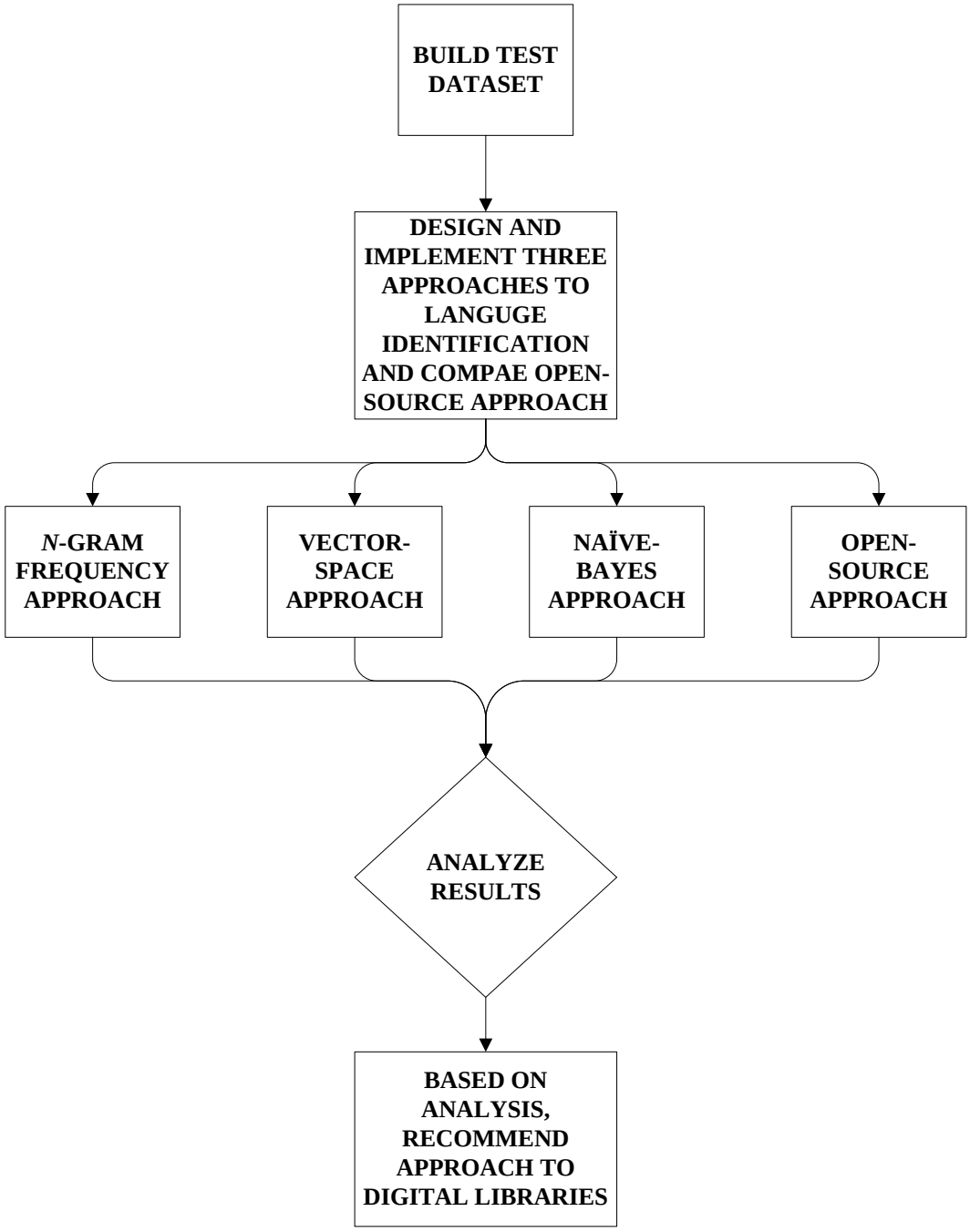

Figure 2. Research design

The following is a detailed explanation of each of these steps.

Data Set Construction for Testing

An ideal dataset for this study consists of titles of books, papers, etc. in each language to be identified by each of the language identification approaches. To build such a dataset, I collect and save titles verified to be in each of the 21 languages of the EuroParl corpus. The total number of titles for each language is not known at first, but preliminary research indicates that certain of the languages, e.g., Latvian and Slovene, have fewer titles available than others, such as Spanish and Italian. Once I collect a suitably-sized dataset, I subject the dataset to testing using each of the three language-identification approaches, along with one open-source approach (to be described below).

Training Data

I use the 21 languages of the EuroParl corpus for training in each of the in-house approaches to language identification. I identify these 21 languages, along with the size of the text files used for training, and have displayed them in Table 4 below.

Table 4

Europarl languages and size of text files for each

Language	Text File Size
Bulgarian	112.7 MB
Czech	98.2MB
Danish	295.5 MB
Dutch	327.4 MB
English	94.3 MB
Estonian	91.5 MB
Finnish	303.0 MB
French	346.9 MB
German	328.5 MB
Greek	390.6 MB
Hungarian	105.4 MB
Italian	325.6 MB
Latvian	96.3 MB
Lithuanian	92.4 MB
Polish	101.1 MB
Portuguese	328.1 MB
Romanian	66.4 MB
Slovak	97.3 MB
Slovene	85.2 MB
Spanish	324.9 MB
Swedish	285.9 MB

Each of the files in Table 4 will serve for training the 21 language models to be used in each of the three following approaches.

N-gram frequency profile and distance measure approach

The first approach to language identification of metadata records is quite similar to Cavnar and Trenkle's (1994) approach. While Cavnar and Trenkle used the top 100 to 400 occurring *N*-grams in their study, I use the top 100 to 500 occurring *N*-grams.

Cavnar and Trenkle (1994) also padded the *N*-grams for spaces, simply meaning that they added spaces to all *N*-grams to avoid including word boundaries in the store of *N*-grams. I also do this,

but I additionally construct the *N*-grams without padding with spaces, which simply means that word boundaries are included in the store of *N*-grams. The reasoning behind this is that the *N*-grams, with *N* = 1 to 5, which are common to a language, may very well be *N*-grams that include word boundaries. For example, in the top occurring 400 character *N*-grams of English, "e_th" is among them. This method worked well for Cavnar and Trenkle (1994), especially using the top 400 occurring *N*-grams. With this method, they accurately identified 99.8% of test materials when experimenting with only two languages. But in what cases would a cataloger need to select between only two languages for assigning a language value to a metadata record? The need exists to consider far more than two languages, especially in light of the vast and still-growing amounts of multilingual digital media. Although Cavnar and Trenkle (1994) achieved a high accuracy (99.8%), they only experimented with two languages. Because I consider many more languages, I cannot achieve as high a level of accuracy. Just like Cavnar and Trenkle (1994), an "out-of-place measure" is used to determine to which language a test text belongs. *Figure 3* illustrates this approach.

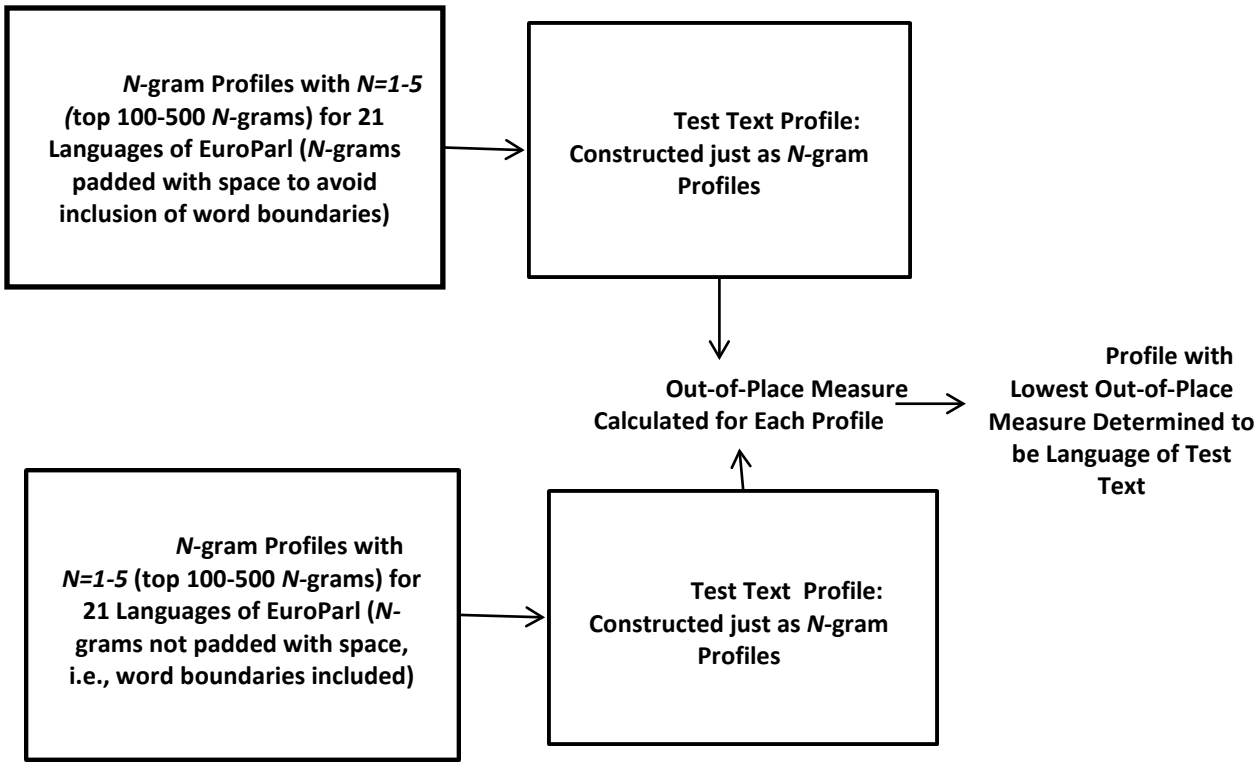

Figure 3. Two versions of out-of-place measure N-gram approach: padded with space and not padded with space

The out-of-place measure, again, is calculated by adding the differences between the rank orders of the *N*-grams in the language profiles and the *N*-grams' rank orders in the test profile. The language with the lowest out-of-place measure is then determined to be the most likely language of the test text.

The accuracy will consists of how many texts the program accurately classifies over the number of total texts. Texts which are misclassified will be reviewed and discussed in terms of possible reasons for the incorrect classification.

N-gram profile without padding for space

As noted above, Cavnar and Trenkle's approach will be slightly modified for another experiment using the same training and test data. Textcat is an open-source software which utilizes Cavnar

and Trenkle's (1994) method. A recent variation was proposed by Hornik et al. (2013) in their textcat implementation for R. They make valid observations about the traditional Cavnar and Trenkle (1994) approach, including the abundance of and redundancy of *N*-grams padded with space. For example, the bigrams and trigrams comprising the string “is” include:

_i

is

s_

_is

is_

s__

Hornik et al. (2013) pointed out the redundancy of the “s_” and “s__” bigram and trigram. They then proposed a “reduced” *N*-gram representations resulting in:

_i

s_

_is_

The model proposed by Hornik et al. (2013) determined to narrow the number of *N*-grams by eliminating the *N*-grams their model considered redundant or unnecessary. However, there is another method of ‘reducing’ the *N*-grams that they did not consider.

I run an experiment similar to Cavnar and Trenkle’s (1994) experiment with the following reductions of *N*-grams:

Space—represented as “_,” will be considered just as a linguistic character, i.e., a letter; Inter-word boundaries will be included in the *N*-grams, e.g., “s_t” will be a viable trigram;

The difference in this modified approach and the method implemented by Hornik et al. (2013) is that inter-word boundaries are included, under the evidence that Zipf's Law holds to the 676 possible trigrams for English inter-word trigrams. With letters, words, and even phrases/utterances in online human/chat-bot interactions (Wallace, n.d., para. 6), by extension, one can assume that sets of *N*-grams, e.g., inter-word trigrams, follow this distribution as well. A small experiment conducted to test this hypothesis is illustrated in *Figure 4* and *Figure 5* *Figure 5* below. Herman Melville's novel, Moby Dick, when analyzed for the distribution of inter-word character trigrams, demonstrates a clear Zipfian distribution of these trigrams. The text is stripped of punctuation, meaning that only inter-word trigrams made up of '[a-z]_[a-z]' are included.

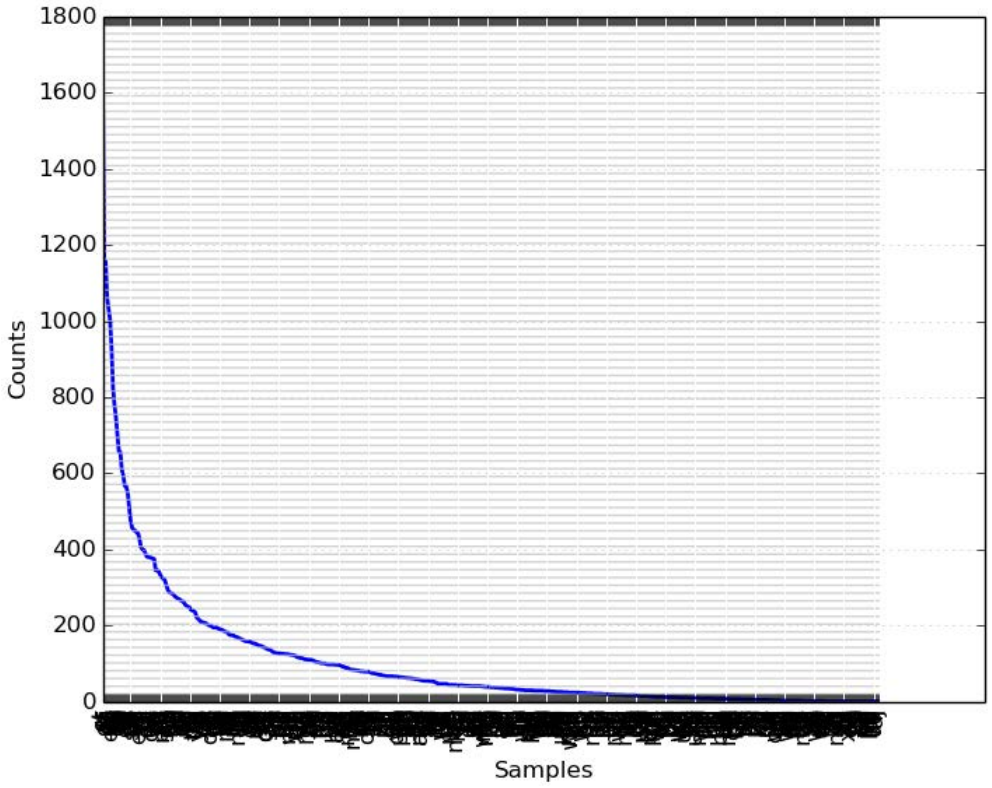

Figure 4. Distribution of inter-word character trigrams in Herman Melville's Moby Dick

Again, there are a possible 26 * 26, or 676 possibilities for inter-word character trigrams in English. Figure 5 illustrates statistics of the top 10 occurring and the top 50 occurring inter-word character trigrams within Figure 4's Zipfian distribution.

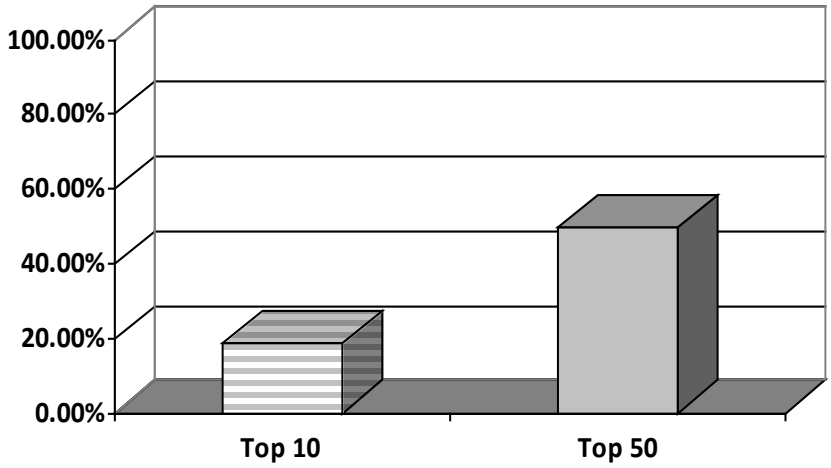

Figure 5. Top occurring inter-word character trigrams in Moby Dick; total of 528 unique trigrams and 54,590 cases

While the reduction is not identical to that of Hornik et al. (2013), it outperforms the *N*-gram frequency profile approach of Cavnar and Trenkle (1994). This is supported by the discriminatory power of inter-word character trigrams to distinguish one language from another. An interesting finding of this inquiry is that in English, the inter-word character trigram most common among most data is *e_t.* The following section illustrates the next approach for automatic language identification of titles for metadata records.

Vector Space Model and Cosine Similarity Approach

In the second approach, I implement a vector-space model, much like that used in information retrieval (IR), measuring the cosine similarity of the language models and test data to classify the language of the test data. While most IR work considers the "word" as the unit of language, the current study also consider *N*-grams, with *N* = 1 to 5, as well as individual words.

The vector-space model assigns a weight to each term in each document of a given collection. An effective and commonly used weighting schema is known as the term frequency-inverse document frequency, or TF-IDF (Salton & Buckley, 1988).

The term frequency, or TF, is a measure of frequency of the term in a given document. While the raw frequency of the term could and has been used to calculate *TF*, it is often normalized by dividing the raw frequency by the frequency of the highest occurring term. So, for example, if the term “adroit” appears twice in a document, and the highest occurring term in the same document is “the”, appearing 100 times, the *TF* of “adroit” would be calculated as: 2 / 100 = .02. Maximum normalization, which has proven effective, includes the following:

TF (normalized) = k + (1 – k) * (TF / TF (highest frequency term)),

where “k” is a constant optimal at values between .3 and .5.

The inverse document frequency, or IDF, was first proposed by Karen Spärck Jones (1972), and while it has little theoretical justification, it has proven quite effective for term weighting. The IDF is formulated as follows:

IDF = N_D / w ∈ d,

where N_D is the total number of documents in the collection, *w* is word, and *d* is the document in which the word appears.

These two measures, multiplied together, give each term in each document a weight, allowing for a vector to be built for each document, each component of which is the *TF-IDF* of each term within the document. A vector is then constructed using the query in traditional IR, and the vector of the query is compared to the vectors of the documents using the cosine similarity to determine which document(s) are most relevant. To determine the similarity of two documents,

the dot product of each document is first obtained. To illustrate, each document, including the query in IR, or each language and the test text in language identification, is assigned a vector, shown as:

$$\vec{a} = (a_1, a_2, a_3, \ldots) \text{ and } \vec{b} = (b_1, b_2, b_3, \ldots)$$

where a_n and b_n are the components of each vector, or in the case of IR, the *TF-IDF* of each term in each document. The *n* is the dimension of the vector. The dot product of the term vectors of two documents', a and b, is formulated as:

$$\vec{a} * \vec{b} = \sum_{i=1}^{n} a_1 b_1 = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n .$$

The cosine similarity is then calculated as:

$$\vec{a} * \vec{b} = || \vec{a} || \, || \vec{b} || \cos\theta.$$

The aim is to measure the distance between the two vectors with the understanding that the closer two vectors are to one another, the more related, or "relevant" in IR, the two documents should be. To illustrate, consider the following.

The vector-space model can be used identically for language identification. The only necessary change is to consider each language in the identifier as a document. To illustrate: consider the language English: ENG and the test document: TEST. Let us say ENG and TEST have the following vector values:

$\vec{ENG}$ = (3, 4, 10, …) and $\vec{TEST}$ = (1, 0, 2, …).

Therefore,

$\vec{ENG} * \vec{TEST}$ = 3 * 1 + 4 * 0 + 10 * 2 … = 23…

Again, the aim is to measure, using the cosine of the vectors' angles, how closely one vector projects into the other vector, as illustrated in *Figure 6*.

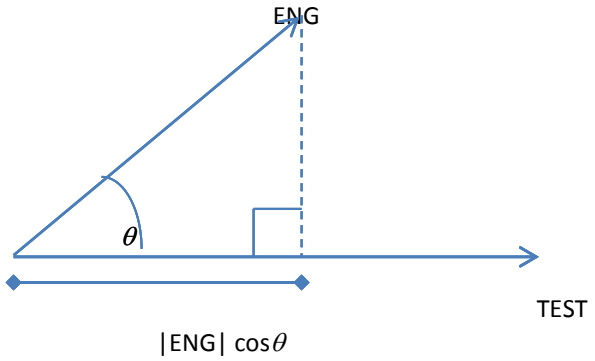

Figure 6. Cosine similarity matrix between hypothetical English and test text vectors.

This approach will compare each test text's vector with 21 vectors belonging to corresponding languages. The accuracy will be judged using a binary decision: Whether the test text was identified correctly, i.e., whether or not the test text belongs to the language which the program determined to be correct. Each misclassified test text will be viewed and discussed in terms of possible reasons for the misclassification.

Naïve-Bayes approach

The third approach utilizes a naïve-Bayes classification scheme. Naïve-Bayes is a supervised algorithm, supervised because a 'supervisor' predetermines the classes of training data. The following text illustrate this approach in terms of language identification.

A naïve-Bayes classifier applies Bayes' Theorem (from Bayesian Statistics) with 'naïve' independence assumptions regarding the features of classification. In other words, the assumption is that Feature A is entirely independent of Features B and C, Feature B is entirely independent of Features A and C, and so forth. This is seldom the case, but the approach still

proves successful in a number of classification tasks. The following demonstrates the naïve-Bayes classification approach by illustrating how I use the approach in this study for language identification.

The classes are the languages I use in this study. Therefore, there is an English class, a Slovak class, an Italian class, etc. The features are character *N*-grams, with *N=1-5*, and words. I calculate the probability that a test text *T* belongs to a language *L*, or *p(L|T).* By definition,

$$p(T \mid L) = \frac{p(T \cap L)}{p(L)}$$

and

$$p(L \mid T) = \frac{p(L \cap T)}{p(T)}$$

so the likelihood becomes:

$$p(L \mid T) = \frac{p(L)}{p(T)} p(T \mid L).$$

Using 21 languages $L_{i=21}$, the probability of a Language, L_i, given a test text *T*, consisting of *N*-grams *NG*, could be written as:

$$p(L_e \mid T) = \frac{p(L_i)}{p(T)} \prod_i p(NG_i \in T \mid L_i),$$

and it follows that the most likely language, *l* (of total Languages, *L)*, of a test text, *T,* consisting of *N*-grams, *ng* (of total *N*-grams, *NG*), is:

$$l = \arg\max_{l \in L} P(L_j) \prod_{ng \in NG} P(ng \mid l).$$

The three selected approaches have all resulted in high accuracy in the literature, some higher than others. For example, Cavnar and Trenkle (1994), reached accuracy levels of above 98%

using their *N*-gram frequency profile and distance measures. However, they only considered two languages in their research. The higher number of languages I involve in this study warrants investigating how these three commonly-used approaches fare in terms of accuracy when faced with far more languages than just a few.

Open-Source Approach

After developing the above three approaches, I compare these approaches with an open-source, freely available language identifier.

Selecting the open-source language identifier.

Many ope-source language identifiers are available for use, such as Python's guess-language 0.2 (https://pypi.python.org/pypi/guess-language/), a pure JavaScript Language Identification library (mazko.github.io/jsli/), Apache Tika's Class Language Identifier (tika.apache.org/1.2/api/org/apache/tika/language/LanguageIdentifier.html), and Perl's Lingua::Identify (search.cpan.org/~ambs/Lingua-Identify-0.56/lib/Lingua/Identify.pm), and TEXTCAT, which is an implementation of Cavnar and Trenkle's (1994) *N-Gram-Based Text Categorization* (odur.let.rug.nl~vannoord/TextCat/). The ultimate decision of which to use in this study depend on the languages included in these identifiers. I select the open-source language identifier which includes the highest number of languages in common with the Europarl corpus, since this contains the languages intended for use in training the above four approaches. Whichever open-source identifier is selected, the comparison with the above three approaches will be based only upon common languages, i.e., languages belonging to the EuroParl corpus, which will be used for training in the above three approaches.

Testing and analysis

The analysis of these four approaches and the various methods for language identification using each of the approaches consists of measuring the overall effectiveness of each approach, i.e., how many *title* elements each approach correctly identifies. The measure of the effectiveness of each approach is the accuracy—in the form of a percentage between zero and 100, with 100 being a perfect score.

The results of each approach are carefully analyzed with respect to the incorrect classifications, or instances in which a test text is misclassified as an incorrect language. The hypothesis is that misclassifications would mostly be of languages in the same family as the correct language, e.g., an instance of a Romanian string being classified as Portuguese or Italian, in that all three languages are Latin-based, or Romance languages.

Closing remarks

I evaluate each of the above methods in terms of accuracy of classification. I then suggest the superior method for accurate language identification of the "title" element of a metadata record. Keep in mind that the title element often consists of a small number of characters and/or words, and that accurate language identification of such a small number of characters/words is still a heavily researched area.

I manually construct the test data in all four approaches using "titles" known to be in each of the 21 languages belonging to the EuroParl corpus. I gather these titles from online sources such as the Project Gutenberg Online Book Catalog (http://www.gutenberg.org/catalog/). The titles are verified by knowledgeable persons in cases where I am unable to do so.

Before I report the results of my study, I am providing a chapter describing a pilot study funded by the IMLS that required the extraction of 2 million metadata records for research related to machine translation (MT) and multi-lingual information access (MLIA).

CHAPTER 4

PILOT STUDY OF THIS INVESTIGATION

A pilot study for the current research project was conducted and partially reported on in Knudson (2014). The following is a description of this pilot study.

Data and test collection

The research involved the translation of six elements of English metadata records into both Chinese and Spanish. The elements were: Title, creator, subject, publisher, description, and coverage. These elements are frequently used by searchers in information retrieval. A percentage of the records included title and other elements not written in English. After developing a language identifier to separate the non-English records, I tested the accuracy of the language identifier on a sample of the two million total records.

I randomly selected eight hundred records from the two million total records using a random number generator. I divided the eight hundred records into four files, each containing 200 records. One human then evaluated the records using a binary classification to determine whether the record was English or non-English. A second human then reviewed the first human's evaluation. This resulted in a gold standard collection which could be utilized to determine the accuracy of the language identifier. The collection consisted of 800 records: 241 manually classified as non-English and the remaining 559 classified as English.

The disagreement between the human evaluators was minimal: 1.4%. Most of the disagreements concerned metadata records in which only the title element was in a language other than English. Of course, the two libraries from where I extracted the records were American libraries, so the majority of the descriptive metadata was in English. This was no

surprise, in that librarians create metadata in the language(s) of the primary users of a collection.

The 1.4% disagreement related mostly to the "title" element and led to the justification for using only the title element for further experiments, while ignoring the remaining elements, all probably in English. The choice to only use the title element was also due to title being the highest accuracy—73.63%-- obtained by Knudson (2014) when considering all six elements in the classification. Consider the difficulties in identifying a language when using all six of the elements under consideration, all English, save one, in many cases. An automatic language identifier would be more likely to judge a metadata record with a two-word Turkish title element and a 65-word English description element as English than Turkish if the level of analysis were at the metadata record level rather than the metadata element level, i.e., elements were input into the language identifier as both belonging to one record, or test text. Yet if only the title element were input, the likelihood of accurate classification as Turkish would increase.

My decision to identify the language of only the title element was also due to cataloguer's needs when inputting a language element into metadata, as well as the human evaluators' tendency to judge a metadata record as non-English whose only non-English element is the title element, a phenomenon which may warrant further research. When cataloguers input a value for an object's language element, the title element provides an appropriate value, given that the language of the title was known. *Figure 7* and *Figure 8* show the distribution of the title elements, in terms of average number of characters (letters and spaces), and average number of words for the title elements, respectively. Recall that identification using shorter texts, as

Dunning (1994), Baldwin and Lui (2010), and others have stressed, makes language identification more difficult than when using larger texts.

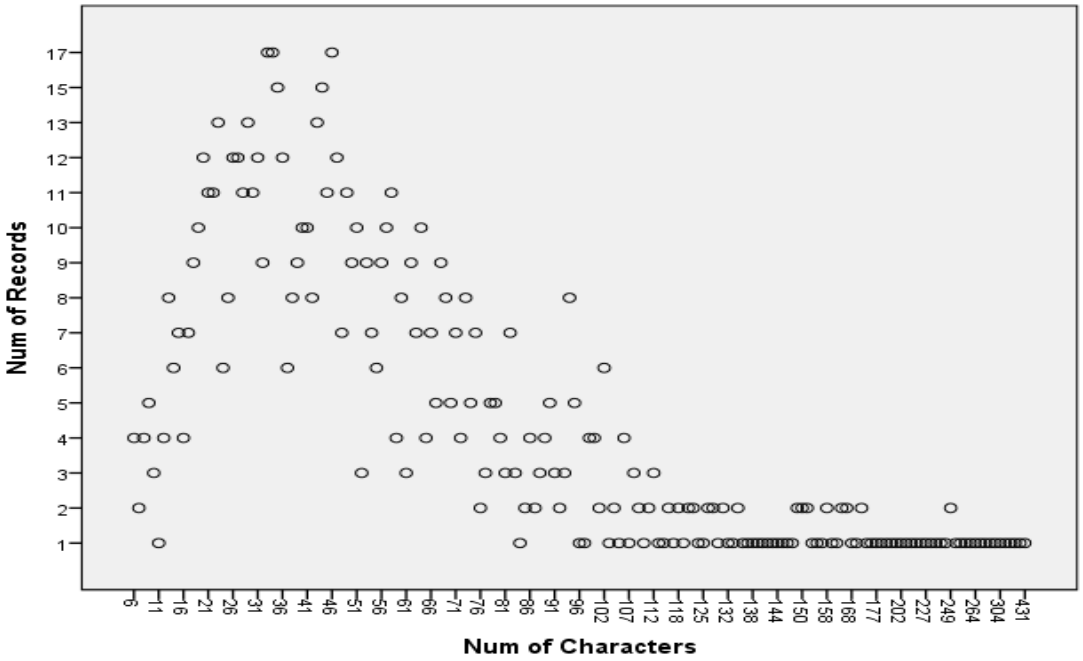

Figure 7. Number of characters per record. Average characters per record: 63.81125.

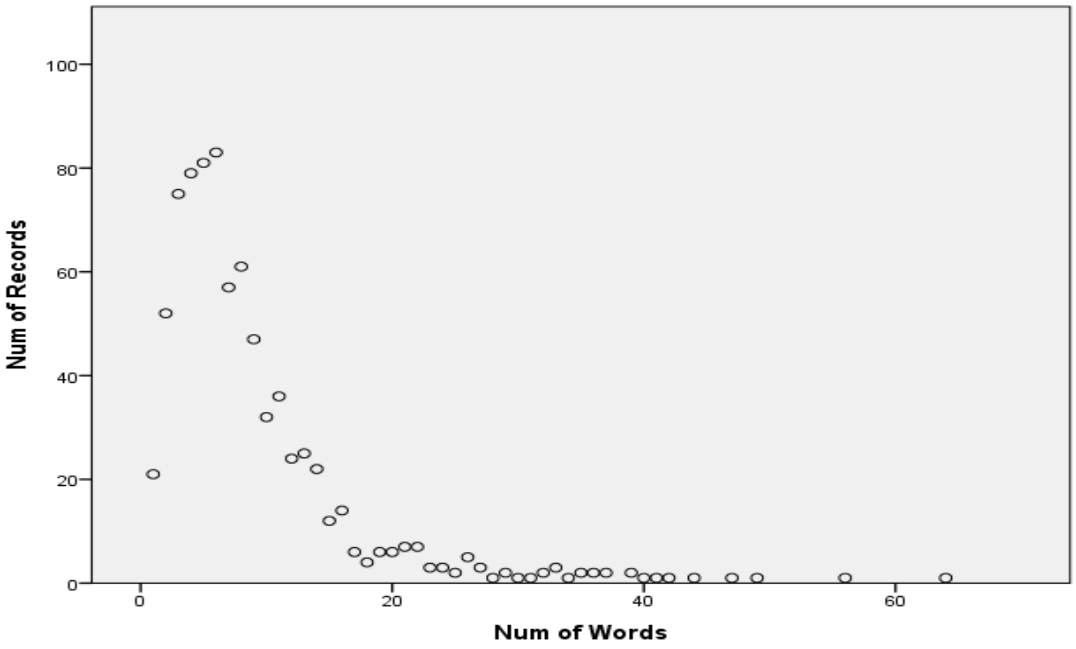

Figure 8. Number of words per record. Average words per record: 8.815.

Building the *N*-Gram Language Identifier

The following is a description of the method and data I utilized to build the *N*-gram language identifier.

Training data

For the training corpora, I selected the Europarl corpus (http://www.statmt.org/europarl/), which is comprised of political proceedings in 21 European languages, including: French, German, Italian, Latvian, Lithuanian, and Polish, among others. The files of the corpus differ in size. The smallest, Romanian, is 66.4 MB. The largest, Greek, is 390.6 MB. The Europarl Corpus is frequently used in machine translation research, and it is appropriate for language modeling for many tasks, including text prediction, fact mining, etc.

Building the language models

I parsed each of the Europarl files into lower-case and then parsed the files into character *N*-grams = 2, or "bigrams." The bigrams' counts were used to build a LM for each of the 21 languages. Each of the items in the LMs was an assignment of the conditional probabilities of the *N*-gram's occurring in the Europarl corpus. I normalized the items by dividing the number of occurrences by the total number of *N*-grams in the data. This probability was calculated as:

$$P(NG) = C(NG) / T(NG),$$

where "NG" is *N*-gram, "C" is count, and "T" is total.

This resulted in 21 character bigram LMs, containing the bigrams conditional probabilities. I then used the LMs for testing the 800 records of the test collection.

Testing the accuracy of the language identifier

In the first experiment with language identification, I used the aforementioned conditional probabilities of each of the *N*-grams to build the LMs. I checked each *N*-gram of the test data against each of the 21 LMs, I incremented a probability value for each of the 21 languages by the conditional probability of each *N*-gram, or:

$$P(L_i) = \sum_{NG \in L} P(NG \mid L_i),$$

for each *N*-gram in the test text, where "L" is language and "NG" is *N*-gram.

Table 5 lists the accuracy of this approach alongside the accuracy of Perl's Lingua::Identify program (an open-source language identification tool), for identifying the language of an entire record's (all six elements) language, and for identifying only the title element's language.

Table 5

Accuracy of conditional probabilities vs. Perl's Lingua::Identify for metadata records' language identification

Test text	Conditional probability	Lingua:identify
All 6 elements	68.9%	59.5%
Title element only	89.75%	84.75%

While this accuracy level may be desirable, the idea of a binary decision of English vs. non-English did not sit well with me. Scenarios came to mind in which the task would not consist of simply extracting only English records from a collection of metadata records for purposes of MT, but would rather be for obtaining a value to input into the "language" element in a metadata scheme. It was clear to me that a binary classification would not succeed for this task, so I developed a program that can accurately classify any of the 21 languages in the training data. Also, I consideedr various algorithms used for language identification commonly used in the literature in order to maximize the accuracy of the procedure. This pilot study, although it only

aimed to reach a binary decision, i.e., English or non-English, resulted in my current investigation, which furthers the investigation to include 21 languages and the accurate identification of these languages relating to metadata elements, specifically, the title element. By automating the language identification of these elements, catalogers who perhaps do not have linguistic training, may achieve more thorough cataloging.

CHAPTER 5

RESULTS AND DISCUSSION

Introduction

In this chapter, I include information about the test data, as well as the results of each language identification experiment I perform on the test data. I also include a discussion of the implications and lessons I learned through the experiments.

Test data: multilingual titles

The test data is comprised of titles in the 21 languages I consider. I gather these items from online searches of digital collections and online libraries. In cases where book titles are scant in online collections, I search in the Internet Movie Database (IMDB http://www.imdb.com) to yield supplemental titles for the dataset. The dataset I investigate is comprised of 82 titles in the 21 European languages. Appendix A contains these titles. Table 6 contains statistics of the average length of these titles in both characters and words.

Table 6

Title language statistics—average length in characters/words

Language	Number of Titles	Characters	Words
Bulgarian	3	28	2.66
Czech	3	26.33	4
Danish	3	32.33	5
Dutch	3	44	8
English	6	48.33	8.83
Estonian	4	17.25	2.25
Finnish	3	39.66	4.33
French	8	48	9.125
German	5	30.8	4.8
Greek	3	47.33	3.66
Hungarian	3	72.33	7.66
Italian	3	53.33	8.66
Latvian	3	19.33	2.33
Lithuanian	3	22.66	3.33
Polish	3	38.33	5.66
Portuguese	5	35.4	6
Romanian	6	27.66	4.33
Slovak	3	23	5
Slovene	3	24.66	5.33
Spanish	5	31.6	6.4
Swedish	3	53.66	8.33
Total Average:	3.90	36.77	5.79

A perfunctory glance at the statistics in Table 6 demonstrates the suitability of language identification as a problem for titles, being that the average length in words of all the titles in this dataset is 5.79. Remember that short-length texts complicate language identification, cf: Baldwin and Lui, 2010; Ceylan and Kim, 2009; and Dunning, 1994. In the following pages, I detail each of the proposed approaches to language identification of titles along with the results, measured in accuracy, of each approach.

N-Gram frequency profile and distance measure

In the following sections, I demonstrate the effectiveness of Cavnar and Trenkle's (1994) *N*-gram frequency profile and distance measure in identifying the language of titles in the test data collection. In the first section, I demonstrate the effectiveness of this approach following Cavnar and Trenkle's (1994) "padding with space." I follow this with the results of the modified version of this approach, i.e., considering the space as just another character, rather than adding redundant spaces at the end of words. This modified approach results in word boundaries being included in the *N*-grams. I justify this because trigrams that contain a space—for a word boundary, in their midst—are a finite set of trigrams. For example, in English, because there are 26 characters in the alphabet, the number of possible inter-word trigrams is 26 x 26, or 676. Because the Zipfian distribution holds true for these trigrams, there are higher numbers of a relatively few of these 676 inter-word trigrams in textual data.

Cavnar and Trenkle (1994)—"Padded with Blanks"

Cavnar and Trenkle (1994) explained that *N*-grams were constructed while padding with space. They illustrated this with the following:

"Thus, the word "TEXT" would be composed of the following *N*-grams:

bi-grams: _T, TE, EX, XT, T_

tri-grams: _TE, TEX, EXT, XT_, T_ _

quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _

In general, a string of length "k," padded with blanks, has k-1 bi-grams, k-1 tri-grams, k-1 quad-grams, and so on" (p. 162). This approach was implemented on the titles in the Table 6 above.

Table 7, below, includes the accuracy of this approach, following Cavnar and Trenkle's (1994) *N*-gram frequency profiles where *N*=1-5 and using the top 100 to 500 most frequently occurring *N*-grams.

Table 7

Title language identification accuracy using Cavnar and Trenkle's (1994) N-gram frequency profiles and distance measure

N-Gram Profile Size	Correctly Identified	Accurac y
100	50 of 81	61.73%
200	59 of 81	72.84%
300	63 of 81	77.77%
400	61 of 81	75.31%
500	63 of 81	77.77%

The average number of words in the titles was 5.79 (Table 6) which invariably has an effect on the accuracy of identification. When considering Cavnar and Trenkle's (1994) approach, it is worth noting that padding the strings with blanks adds considerable noise to the *N*-gram frequency profiles. For example, all strings that end in character "d" result in equal occurrences of bigrams, trigrams, and so on, with "d" as their first character, and trailing blanks. These equal occurrences motivate another experiment to lessen the noise created by these duplicates.

Cavnar and Trenkle (1994) modification—not "Padded With Blanks"

To illustrate the occurrence of duplicate values in the profiles of the above experiment, Cavnar and Trenkle (1994) conduct another experiment on the English training data to determine how many *N*-grams consisted of the character "d" followed by blanks. Within the top-occurring 500 *N*-grams in the English training data, there are 1,237,879 bigrams, trigrams, quadgrams, and pentagrams that begin with "d" and are followed by blanks. This means that of the *N*-grams belonging to English, there are 4 nearly identical *N*-grams in the *N*-gram frequency profile.

Consequently, Hornik et al. (2013) proposes a 'reduced' *N*-gram representation, to eliminate these redundancies. This experiment implements everything identically to the Cavnar and Trenkle (1994) experiment, with one exception—the "blank" is considered as just another character in the profiles. In effect, this approach nullifies the importance of word boundaries that Cavnar and Trenkle (1994) propose. While Hornik et al. (2013) take multiple steps to remedy redundancy in the *N*-gram profile, including the deletion of pre- and post-word blanks, the exclusion of any word boundary except for the padded blank, and *N*-grams where $k > 1$ yielding only *N*-grams $n <= k$, their reductions eliminate some *N*-grams useful in distinguishing a language. Rather than take such extreme measures as Hornik et al. (2013), this modified approach simply considers the blank, or '_' as another character, and in effect, includes inter-word *N*-grams as well as word-initial and word-final blanks, while eliminating redundant final-word blanks.

To illustrate, consider the following:

The text: 'TWO_DAYS_', in this modified approach, consists of the following:

Bigrams: TW, WO, O_, _D, DA, AY, YS, S_

Trigrams: TWO, WO_, *O_D*, _DA, DAY, AYS, YS_

Quadgrams: TWO_, WO_D, O_DA, _DAY, DAYS, AYS_

And so on.

Note that the italicized *N*-grams in the example do not simply pad the end of a word with blanks, but rather, they include the word boundaries in the *N*-gram profiles. The purpose of this modification is to generate robust models for classification, eliminating unnecessary duplicates from the *N*-gram profiles. This modification also considers the distribution of *N*-grams in a

language. For example, English has 26 characters in its alphabet. Therefore, the number of inter-word trigrams possible is represented as:

N_N, where N=26

Since the blank remains constant, the number of possible inter-word trigrams is "26 x 26," or 676 possible inter-word trigrams in English. As mentioned above, these inter-word trigrams follow a Zipfian distribution, meaning that the majority of inter-word trigrams that occur in English data belong to a quite small number of these possibilities. Remember that Zipf's Law stated that the most frequent occurring items in language constitute the majority of examples in that language.

This was the main reason behind modifying Cavnar and Trenkle's (1994) approach. Table 8 displays the accuracy of the classifier described here.

Table 8

Title language identification accuracy using modified (not "padded with blanks") Cavnar and Trenkle's (1994) N-gram frequency profiles and distance measure

N-Gram Profile Size	Correctly Identified	Accuracy
100	47 of 81	58.02%
200	67 of 81	82.72%
300	64 of 81	79.01%
400	68 of 81	83.95%
500	68 of 81	83.95%

The most effective *N*-gram profile sizes in this experiment are the top occurring 400 and 500 *N*-grams. The modified approach, i.e., not "padded with space," demonstrates a significant increase in the accuracy of the traditional approach utilized by Cavnar and Trenkle (1994)—an increase of 6.18%. These results are notable and justify the modification.

Vector-space model

While "tf" and "idf" (Salton & Buckley, 1988) are commonly used in information retrieval, this term-weighting scheme can also be utilized in language identification. Gottron and Lipka (2010) use a vector-space model, utilizing term-weighting, to identify the language of news headlines at an accuracy level of 54.68% for character *N*-grams with *N=1* and at an accuracy level of 75.37% for character *N*-grams with *N=5.* Note that the "term" here is a character or string of *N* characters. The following experiment utilizes the same approach, i.e., character *N*-grams with *N*=1-5, with the addition of unigram words. Table 9 below contains the results of this experiment.

Table 9

Title language identification using term-weighting (tf-idf) and vector-space model approach

N-Gram Size	Correctly Identified	Accuracy
Unigrams	3 of 81	3.70%
Bigrams	5 of 81	6.17%
Trigrams	8 of 81	9.88%
Quadgrams	33 of 81	40.74%
Pentagrams	50 of 81	61.73%
Unigram Words	40 of 81	49.38%

Naïve-Bayes

As already noted, Naïve-Bayes assumes independence of items in relation to their surrounding items. While this seems counterintuitive, especially with language items, it has been used with great success in text classification problems. Using character bigrams as the unit of analysis, 62 of 81 items in the test data are correctly identified, showing an accuracy of 76.54%.

Python's Open-Source Language Identification

In determining which open-source language identifier to use for comparison, the main factor affecting this decision is finding an identifier that allows for the same 21 languages used in the previous experiments. Open-source language identifiers are available, e.g., textCat (http://software.wise-guys.nl/libtextcat/), but most do not include all these 21 languages. This is solved using Cavar's (2011) Python language identifier tool—lid.py, which allows the user to build the language models on the fly using whatever corpora are available to him or her.

This software selects all trigrams from the training data for a language. Trigrams that include punctuation within them are deleted, e.g., the trigram 'r-w' is removed. The trigrams are then compared with the trigrams of an unknown text by calculating the difference between the trigrams' relative probabilities within the unknown text and the language models.

So, the string 'Words are' would contain the following trigrams:

Wor – 1

ord – 1

rds – 1

ds_ - 1

s_a – 1

_ar – 1

are – 1

And since there are seven total trigrams in this string, the relative probability of each trigram is:

Wor – 1/7

ord – 1/7

rds – 1/7

ds_ - 1/7

s_a – 1/7

_ar – 1/7

are – 1/7

Now consider the above as the language model, and consider an unknown string: 'are fun', which has the following trigrams and relative probabilities:

are – 1/5

re_ - 1/5

e_f – 1/5

_fu – 1/5

fun – 1/5

The distance between the language model and the unknown string is calculated as:

are 1/5 – 1/7

re_ 1/5 – 0 (since “re_” is not present in sample language model)

e_f 1/5 – 0

_fu 1/5 – 0

fun 1/5 – 0

In cases where the distance is negative, the Python LID solves this by inverting the negative number or squaring it. This approach, as noted above, allows language models to be built using textual data on hand, so the 21 language models are built using the same training data as in the previous experiments, i.e., the 21 europarl corpora. This is done to facilitate comparison

between the various approaches implemented and the open-source approach. Using this approach on the 81 titles in the test collection, 56 of 81 trigrams are correctly identified showing 69.14% accuracy.

Misclassifications in the highest scores of each approach

A common question concerns cases of misclassification in language identification. Often, misclassified texts are deemed to be in a language related or similar to the language of the content. For example, an unknown string in Spanish would more likely be misclassified as Italian or Portuguese than it would misclassified as Greek or Bulgarian.

Table 10 lists the accuracy of each of the above approaches, based on the language under consideration. In cases where multiple methods are used in an approach, i.e., *N*-gram frequency profile and distance measure, "modified" *N*-gram frequency profile and distance measure, and vector-space model, only the highest-scoring method is listed. In the two cases where there are ties in the methods' accuracy, the larger *N*-gram frequency profile is shown, i.e., 500 *N*-grams rather than 300 or 400.

Table 10

Misclassification of all highest-scoring methods in each approach

Language:	Approach:				
	N-Gram Freq Profile (500)	*N*-Gram Freq Profile—Not Padded with Blanks (500)	Vector-Space Model (Pentagrams)	Naïve-Bayes	Python's Trigram LID
Bulgarian	100%	100%	100%	100%	100%
Czech	33.33%	0%	33.33%	33.33%	66.6%
Danish	66.66%	66.66%	66.66%	66.66%	66.66%
Dutch	100%	100%	100%	100%	100%
English	66.6%	100%	66.66%	66.66%	66.66%
Estonian	50%	75%	100%	75%	0%
Finnish	66.6%	100%	100%	100%	100%
French	100%	100%	62.50%	100%	87.5%
German	80%	100%	80%	80%	80%
Greek	100%	100%	100%	100%	100%
Hungarian	100%	100%	66.66%	100%	100%
Italian	66.66%	100%	66.66%	66.66%	66.66%
Latvian	66.66%	100%	0%	100%	66.66%
Lithuanian	100%	33.33%	66.66%	66.66%	0%
Polish	100%	100%	100%	66.66%	66.66%
Portuguese	100%	100%	60%	100%	60%
Romanian	66.66%	66.66%	100%	33.33%	66.66%
Slovak	66.66%	66.66%	66.66%	33.33%	33.33%
Slovene	66.66%	66.66%	66.66%	100%	100%
Spanish	60%	80%	20%	60%	40%
Swedish	66.66%	66.66%	33.33%	66.66%	100%
Total Accuracy:	77.77%	83.95%	69.14%	76.54%	69.14%

Summary of Results

Notice the highest-scoring approach is the modified *N*-gram frequency profile and distance measure. Recall that this modification removes replicated blanks from the *N*-grams and considers inter-word *N*-grams of all sizes, i.e., *N* = 1-5. The misclassified texts are in most cases misclassified as being in a language closely related to the actual language, just as Hornik et al. (2013) demonstrate when they describe the clustering of related languages such as the Scandinavian languages. Of note is the identical accuracy of all approaches for the Danish classification. Each of the misclassifications for the Danish language, in all approaches, is classified as Swedish, which is, like Danish, a Scandinavian language.

In the following chapter, I discuss this chapter's results in more detail; explain the problems I faced along the way, the limitations, future research plans; and include a conclusion.

CHAPTER 6

FINDINGS AND CONCLUSION

Introduction

In this chapter, I cover the results of the study, findings of note reached during the study, difficulties encountered, misclassifications of language, limitations of the research, and plans for future research. I begin with an explanatory discussion of language identification, a few things worth noting about the investigation, and what I discover therein.

Automatic language identification

Automatic language identification is a classification task. Simply stated, that means that the object is to match an unknown language to a known language and to do it accurately. This requires that the collection of known languages contains the unknown language. Otherwise, a best guess is posited by the identifier. *Figure 9* is a simple illustration of this task.

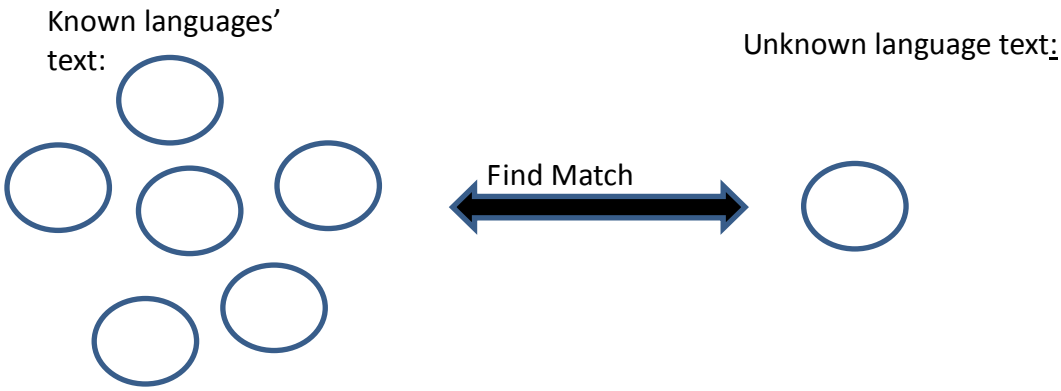

Figure 9. Simplified illustration of the task of automatic language identification.

There are several methods for performing this classification task. These approaches include, but are not limited to: Naïve Bayes, vector-space model, Markov modeling, *N*-gram frequency

profiles, and Rocchio's algorithm. The task remains the same, but the approach may differ. The current investigation utilizes four commonly used, tried, and tested approaches to this problem. During this investigation, I discover many things of note. The following section describes these findings.

Noteworthy findings

Reduced *N*-gram frequency profile vs. traditional

In my opinion, the most interesting finding in this study is the improvement in classification accuracy that is achieved simply by reducing the *N*-gram frequency profile and distance measure approach of Cavnar and Trenkle (1994) to exclude redundant blanks, or spaces, and to include blanks in cases where they occur in the midst of character *N*-grams. This reduction results in a 6.18% increase in total accuracy of title language identification.

While my *N*-gram frequency profile reduction is similar to that implemented by Hornik et al. (2013), my reduction does not eliminate mid-word, final *N*-grams, e.g., the quadgram "rpus**"** taken from the English word "corpus." Future investigations can determine the difference in accuracy of these two approaches, both of which reduce the *N*-grams, but one not quite to the same degree as the other.

Zipfian Distribution of Inter-Word Character Trigrams

Also of note is the fact that the inter-word trigrams, as expected, cleanly follow a Zipfian distribution, i.e., the most frequently occurring grams are few in number, while many single-occurring grams exist. As noted above, the most frequently occurring inter-word trigram in

English was found to be “e_t,” which indicates that words ending in “e” followed by words that begin with “t” are the most common inter-word character trigrams in the English language. While, in English, there are 676 possible inter-word character trigrams, seldom will a text contain all of these. For example, Herman Melville’s Moby Dick includes only 528 of these possible 676 character trigrams. With a little thought, it is not difficult to understand the absence of some of these trigrams. Consider the trigram “x_x” and its likelihood in English. A few grammatical examples might include “six xylophones,” “six xenophiles,” and so on. Instances of this trigram are far less common in a usual English text than, say, “e_t,” which is the most common.

Alphabets of test data items

Each of the approaches implemented here accurately identifies 100% of Bulgarian and Greek and Dutch titles. This is no surprise for the former two, since Bulgarian is comprised of Cyrillic characters and Greek is comprised of characters from the Greek alphabet. However, it is a big surprise for Dutch, which should be less easy to recognize than is German, because German has frequent umlauted vowels (ü, ö, ä), which Dutch does not. Both Bulgarian and Greek texts are unique in this study, i.e., none of the other 19 languages in these experiments uses these scripts, so distinguishing these two languages from the others does not pose a problem, as is shown by each approach’s perfect accuracy for both languages. The difficulty starts when attempting to distinguish languages which use the same set of characters. It would be relatively easy to distinguish between Simplified Chinese text and Bulgarian, since the languages do not share a common text.

Difficulties of investigation

Length of texts

It is well known that accurate language identification grows more difficult to achieve with short texts, cf: Baldwin and Lui, 2010; Ceylan and Kim, 2009; and Dunning, 1994. Again, it is much easier to accurately identify the language of an entire page of a book than a single word on a page. A few experiments not reported above demonstrated that longer texts are easier to accurately identify by all the approaches used. The accuracy of each approach when implemented on paragraphs taken from Wikipedia in each of the 21 languages accurately identifies 100% of these 21 sample paragraphs. Longer texts facilitate more accurate language identification, which is consistent with the literature. And as an unknown text gets shorter and shorter, accurately identifying the language becomes more complicated.

Titles are short in nature, as illustrated by the test data titles in this study. There are several titles among the 81 total that are comprised of only one word, which often impedes 100% accurate identification of them.

The difficulty with short texts can also be seen in the accurate identification of longer titles. For example, one of the English titles contains 16 words, and all of the approaches accurately identify this particular title. This clearly demonstrates the connection between text length and accurate language identification. However, titles are often composed of far fewer than 17 words.

Obtaining a dataset

Another difficulty of this investigation is the creation of the dataset. Ideally, a dataset can be created by collecting titles from metadata. However, as already noted, many metadata have null values for the language element, making them unsuitable for a gold standard without having working knowledge of all 21 languages, which would enable visual inspection through a manual check of each title. This problem led me to collect the titles through searches in online digital collections and movie databases.

Misclassifications

Languages that are closely related, e.g., Italian and Portuguese—both Italic languages, are the most common source of misclassification in this study. Table 11 groups these languages according to their misclassifications.

Table 11

Language families of 21 languages of test data

Family:	Baltic:	Germanic:	Hellenic:	Italic:	Slavic:	Uralic:
Languages:	Lithuanian Latvian	Danish Dutch English German Swedish	Greek	French Italian Portuguese Romanian Spanish	Bulgarian Czech Polish Slovak Slovene	Estonian Finnish Hungarian

Note: Although Bulgarian is a Slavic language, unlike the other members in this group, Bulgarian uses Cyrillic script rather than Latin script.

In the majority of misclassifications, the language identifiers estimate that the titles in question belong to another language in the same family. For example, when considering a Swedish title, the classifier misclassifies the title as Danish. Both of these languages are Germanic, and both are Scandinavian. Denmark exerted extreme influence over Sweden until 1525 when Gustav Vasa led a revolt against Denmark that led to a new Swedish government with intent to purify

the Swedish language of Danish influence (“Swedish”, 2015). This revolt makes this misclassification understandable.

Classification accuracy decreases when considering highly-related languages (Trieschnigg, et al., 2012). In all cases, save one, in which an English (Germanic) title is classified as Estonian (Uralic), the classifiers in this study classify the titles within the same language family in cases of misclassification.

Language relatedness is a common problem for accurate automatic language identification. Cases abound in which an English text is classified as German, etc. Methods of improving classification between related languages should exploit characteristics of each language that clearly distinguish it from its relatives.

The misclassifications in this study are considered one of the study’s limitations. Below is a discussion of some of the other limitations of this study.

Limitations of investigation

As noted above, misclassifications occurring between related languages can be considered a shortcoming of this investigation. This is not the only problem. The following section describes a few more limitations of this investigation.

Limited domain of training data

The data used for training in this investigation is of a political nature. It is comprised of freely available European Parliamentary Proceedings. While many metadata records might describe digital objects of a political nature, this is definitely not the case for all metadata records, the domain of which is limited only by the number of potential domains of digital objects. Metadata may describe objects of the musical, literary, or scientific domain, among others.

The European Parliamentary Proceedings are likely not comprised of discussions of musical works, literary works, or the like, which might seemingly limit the language modeling done on such corpora to a single domain. This limitation, however, is not considered a threat due to the nature of the language modeling performed, i.e., language modeling at the character level of language rather than at the word level. In all but one of the cases, i.e., the vector-space model approach to language identification, the language modeling is done at the *N*-gram level, not the word level. This means simply that the characteristics used to distinguish languages in this investigation are various-sized groups of character *N*-grams rather than words or tokens.

This limitation could become a serious problem when dealing with metadata from multiple domains. The metadata considered in the pilot study of this investigation contained many titles consisting of musical scores and performances, often in German or Italian. This leads me to believe that using multiple domains for data for training might alleviate this limitation.

The limited domain, although not a serious limitation due to the nature of language modeling instituted in this investigation, is not the only limitation. Below is a description of the limited languages I consider within this investigation.

Limited languages

As already discussed, the languages considered within this investigation are comprised solely of European languages—21 taken from the Europarl Corpus (http://www.statmt.org/europarl/). The decision to use these languages is mainly influenced by the availability of the corpora for training the LMs. These 21 corpora are freely available for research use, which facilitates building language models without collecting language data from multiple languages manually.

Also noted above, the number of languages potentially found within the title element of metadata records is limited only by the number of extant languages with a system of writing. While the language of titles in metadata likely follows a Zipfian distribution, i.e., there are far more English metadata titles than Tok Pisin metadata titles, this is currently not known to be the case.

This limitation is understandable if one realizes that this work is the first of its kind, i.e., the first to deal with language identification of metadata records. No other investigation to date has considered automatic language identification of title elements within metadata records. This investigation is intended to encourage work in the area of language identification of metadata records, which could greatly improve information retrieval and access within digital collections by allowing language a more prominent place in searches. The following section is a description of plans for the future.

Future research

The limitations above give insight into how this research might be refined and improved for future investigations. The following sections describe possible plans for future investigations.

Refining the reduced *N*-gram frequency profile approach

Cavnar and Trenkle's (1994) *N*-gram frequency profile approach has been commonly used for the problem of automatic language identification. However, as Hornik et al. (2013) pointed out, the *N*-gram modeling has not been questioned. Hornik et al. (2013) reduces the *N*-gram frequency profiles to exclude redundant items related to final-word *N*-grams and spaces.

I intend to investigate methods of further reducing the *N*-gram frequency profiles to eliminate even more redundant information. One such method might involve maintaining capital letters in

named entities, in effect better representing the distribution of *N*-grams within training data. For example, consider the text “Why does all America always act quickly.” In this text, there are four word-initial unigrams *“a”* if capitalization is removed during processing. However, it is possible that maintaining the capital “A” might result in a more discriminating frequency profile than one that did not include capital letters.

It is also possible that reducing the size of the *N*-grams might affect discriminatory power of the frequency profiles. For example, what if only the most frequently occurring *N*-grams with *N*=1-4 were used in place of *N*=1-5? This is another future plan I have for investigation and reporting. Another plan for future research involves expanding the domain of the training data. This is discussed below.

Utilizing multi-domain training data

Above it was stated that domain is of less importance in this investigation due to the nature of the language modeling implemented, i.e., character-level *N*-gram modeling. However, this may decrease classification accuracy when dealing with metadata titles from multiple domains, especially in cases where titles consist of words/characters from another language.

Consider an English metadata title comprised of “Sonata in C.” Sonata is an English word, but it was borrowed from Italian. A language identifier might classify this title as Italian rather than English, even when using character-level *N*-grams for the language modeling. This sort of misclassification may be restricted, even when using character-level *N*-grams, if the training data contains data of multiple domains.

The problem with training using multi-domain data is obtaining this kind of data. One solution might entail having humans evaluate metadata titles and classify the language manually.

However, having humans classify the language manually would require a higher level of linguistic expertise than most people possess. Another solution might be to use metadata records that contain a valid language value, but as described above, the language element often contains a null value in metadata records. Another limitation might be finding suitable numbers of multilingual metadata records that have a valid value in the language element from digital collections. Further complications might include attempting to obtain large enough numbers in various languages. This would be an arduous task.

Another plan for future research involves increasing the number of languages considered for identification of metadata title elements. This plan is described below.

Increasing the number of languages considered

Of the 21 languages used in this investigation, 19 used Latin script. Bulgarian consists of Cyrillic characters, and Greek consists of Greek characters. All other languages considered use Latin script. While language identification of Latin-based script languages is a well-researched field, according to Choong, Mikami, and Nagno (2011), the spawning of Asian, African, and other languages on the World Wide Web is an indicator that language identification research will continue.

While the number of languages available in digital media is increasing, this may not affect the current investigation. Most digital collections have metadata in the language of the collections' users. For example, an American digital collection will indubitably contain American English in its metadata records. There will rarely or possibly never be a case in which an American digital collection would contain metadata records whose title elements were in Thai, unless, of course, the title reflects the Thai language content with Latin script. The metadata records I used in the

pilot study to this investigation consisted of entirely Latin script, even in cases where the language of the title was Arabic or even Chinese.

This being said, there are still numerous languages that use Latin script and are not a part of the 21 languages I consider here. These could be added to the classifier to increase the total number of identifiable languages. Also, if language items outside the realm of metadata elements are considered for identification, e.g., search query terms, Web page contents, etc., the script of the languages would not be as great a concern, and would make it possible for more scripts to be included in the classifier.

There is also a growing number of languages emerging on the World Wide Web and in digital media. These include languages of various scripts, both Latin and non-Latin. Non-Latin scripts are outside the purview of language identification for American metadata records, but in domains outside of American metadata, these non-Latin scripts may be used for language identification.

Small amounts of textual data for training lead to less accuracy in language identification (Hughes, Baldwin, Bird, Nicholson, & MacKinlay, 2006; King & Abney, 2013). If trends continue in digital media and uncommon languages continue making a larger presence online, this growth will enable creation of more accurate, discriminating language classifiers capable of identifying more and more languages possible.

Creating an open-source language identifier for digital collections

An important goal for future research is to make a high-performance, multilingual language identifier available for use in the metadata of digital collections. This identifier would enable

digital collections to input values for language elements for metadata of newly acquired digital objects as well as metadata of existing digital objects in their collections.

The benefits of such a technology are two-fold. Firstly, such a technology enables more complete metadata to be generated and the update of existing metadata, i.e., supply a language element in metadata records which contains a null value in the language element's place. In essence, this would contribute to more thoroughly descriptive metadata than exists currently in digital collections. Secondly, supplying valid values for language elements in the metadata of digital collections would increase the precision/recall in information retrieval and information access. This would lead to more thorough retrieval and access especially in cases where language is an item of interest to a user. For example, if a user searches for Spanish objects in a multilingual digital collection, more items would be returned if the majority of the collection's metadata elements have a value in the language element's place, more thorough than a similar search in a collection with null values in the language element's place.

Conclusion

In this study, I address the problem of null values in language elements within metadata records in digital collections. I propose a viable solution to this problem for digital collections. This problem impedes information retrieval and access to digital collections where the language of the digital object is a concern.

My major intent in this study is to propose the delivery of a language identifier to digital collection curators for use on their existing and new metadata records. Again, this would enable more robust information access within digital collections.

In this study, I consider five approaches to automatic language identification on a multilingual test set that I create of book and movie titles. Each approach intends to accurately identify the languages of titles, modeled after the 21 europarl languages. The approaches I consider include the following: Cavnar and Trenkle's (1994) *N*-gram frequency profile and distance measure; reduced *N*-gram frequency profile and distance measure; vector-space model; naïve Bayes; and an open-source approach (Python's language ID).

The reduced *N*-gram frequency profile and distance measure approach outperform all others by over 6%. This approach accurately identifies 68 of 81 multilingual titles, a total accuracy of 83.95%. This finding demonstrates that the most suitable approach for implementation for digital collections is the reduced *N*-gram frequency profile approach.

The study concludes with plans for future research in this area. These include: refining of the best-performing approach; expanding the domain of the training data; increasing the number of languages considered; and creating an open-source program for language identification and offering it to digital collection curators for use on their metadata.

APPENDIX A

82 TEST TITLES IN 21 LANGUAGES

Bulgarian:

Епически песни

Олаф ван Гелдерн

Мислите в главите

CZECH:

Dvojník. Nétička Nezvánova a Malinký Hrdina

Zápisky z mrtvého domu

Marketa Lazarová

Danish:

Kort og sandfærdig Beretning om den vidtudraabte Besættelse udi Thisted

Haabløse Slægter

Min gamle Kammerat

Dutch:

Een feestelijk verbeeldingsspel in acht tooneelen

Het settlement Malakka en het sultanaat Perak De Aarde en haar Volken, 1908

Het Eiland Marken en Zijne Bewoners

English:

The Industrial Canal and Inner Harbor of New Orleans: History, Description and Economic aspects of Giant Facility Created to Encourage Industrial Expansion and Develop Commerce

Weather and Folk Lore of Peterborough and District

Life and adventures of Frank and Jesse James: The noted western outlaws

The Hunger Games

The Great Gatsby

Animal Farm

Estonian:

Nullpunkt

Risttuules

Hukkunud Alpinisti hotell

Libahunt: Draama wiies vaatuses

Finnish:

Väkevin: Kummallinen kertomus

Adlercreutzin sanansaattaja: Tapaus Revonlahden tappelusta v. 1808

Muistelmia matkoilta Venäjällä vuosina 1854-1858

French:

Aux portes de l'eternite -le siecle 3

Thérèse Desqueyroux

Les outils de la pensée

L'épée de Vérité 14 : Le Crépuscule des Prophéties

3 fois par jour: Premier tome

Douze ans de séjour dans la Haute-Éthiopie

Histoire des salons de Paris (Tome 3/6) Tableaux et portraits du grand monde sous Louis XVI, Le Directoire, le Consulat et l'Empire, la Restauration et le règne de Louis-Philippeler

Curiosités judiciaires et historiques du moyen âge. Procès contre les animaux

German:

Briefwechsel zwischen Abaelard und Heloise: mit der Lebensgeschichte Abaelards

Aus Kroatien: Skizzen und Erzählungen

Die Bekanntschaft auf der Reise

Der Kleine Prinz

Die Physiker

Greek:

Η Βιογραφία του στρατηγού Γεωργίου Καραϊσκάκη

Κύρου Ανάβασις Τόμος 1

Αττικαί ημέραι

Hungarian:

Takáts Sándor Szalai Barkóczy Krisztina 1671-1724 című könyvének ismertetése

Csongrádmegyei gyüjtés (Népköltési gyüjtemény 2. kötet)

Elegyes gyüjtések Magyarország és Erdély különbözo részeibol (Népköltési gyüjtemény 1. kötet)

Italian:

Dialogo sopra la generatione de venti, baleni, tuoni, fulgori, fiumi, laghi, valli et montagne

Costituzione della Repubblica Italiana e Statuti Costituzionali del Regno d'Italia

Danza macàbra

Latvian:

Jaukas Pasakkas unStahsti

Latviešu Avīzes

Pēterburgas Avīzes

LITHUANIAN:

Skamba Kankliah Ir Trimintai

Tadas Blinda. Pradzia

Mieganciu drugeliu tvirtove

Polish:

Jeden miesiąc życia: utwory prozą

Tajemnica Baskerville'ów: dziwne przygody Sherlocka Holmes

Szachy i Warcaby: Droga do mistrzostwa

Portuguese:

Descobrimento das Filippinas pelo navegador portuguez Fernão de Magalhães

Como e porque sou romancista

Os Filhos do Padre Anselmo

A emergência das migrações no feminino

A globalizaçăo no Mundo Antigo

Romanian:

Creierul, O Enigma Descifrata

In numele poporulur meu

Eminescu: Romanii din afara granitelor tarii

Eminescu: Sfantul pamant al Transilvaniei

Civilizatia romanilor in viziunea lui Eminescu

Poezii

Slovak:

Vsetko co mam rad

Sedím na konári a je mi dobre

Legenda o Lietajúcom Cypriánovi

SLOVENE:

Za narodov blagor: Komedija v štirih dejanjih

Zoran, il mio nipote scemo

Gremo mi po svoje

Spanish:

Historia de Venezuela, Tomo I

Reseña Veridica de la Revolución Filipina

Doña Clarines y Mañana de Sol

Andanzas de puck en el sueño de una noche de verano

Andanzas del impresor Zollinger

Swedish:

Allvarsord om allting och ingenting 31

Carl Wilhelm Scheele ett minnesblad på hundrade årsdagen af hans död 58

Himlauret eller det profetiska ordet: Hänvisningar. 3 öfversigtstabeller och 1 diagram 72

REFERENCES

About the World Digital Library: Missions and objectives. (n.d.) Retrieved from http://www.wdl.org/en/about/

Bade, D. (2002). *The creation and persistence of misinformation in shared library catalogs: Language and subject knowledge in a technological era.* Champaign, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Balahur, A., & Montoyo, A. (2008). A feature dependent method for opinion mining and classification. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering*, (pp. 1-7).

Baldwin, T., & Lui, M. (2010). Language identification: The long and the short of the matter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California, (pp. 229-237).

Batchelder, E.O. (1992). *A learning experience: Training an artificial neural network to discriminate languages.* Unpublished manuscript.

Beall, J. (2004). Dublin Core: An obituary. *Library Hi-Tech News 8*, (pp. 40-41).

Beesley, K. (1988). Language identifier: A computer program for automatic natural-language identification of on-line text. In *Language at Crossroads: Proceedings of the 29th Annual Conferenceof the American Translators Association*, (pp. 47–54).

Black, P. E. (2009). Zipf's law. In *Dictionary of Algorithms and Data Structures* [online], Pieterse, V., and P.E. Black., Eds. (accessed October 5, 2014) Available from: http://www.nist.gov/dads/HTML/zipfslaw.html.

Brown, R.D. (2012). Finding and identifying text in 900+ languages. *Digital Investigation, 9* ((supplemental), S34-S43.

Cavar, D. (2011). *LID: Language identification [Software].* Available from http://www.cavar.me/damir/LID/

Cavnar, W.B. & Trenkle, J.M. (1994). *N*-gram based text categorization. In *Proceedings of SDAIR-94. 3rd Annual Symposium on Document Analysis and Information Retrieval*. SDAIR-94, Las Vegas, NV, (pp. 161-175).

Ceylan, H. & Kim, Y. (2009). Language identification of search engine queries. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2,* (pp. 1066-1074). Morristown, NJ.

Choong, C.Y., Mikami, Y., & Nagano, R.L. (2011). Language identification of web pages based on improved *N*-gram algorithm. *International Journal of Computer Science Issues, 8*(3), (pp. 47-58).

Dalal, M.K., & Zaveri, M.A. (2011). Automatic text classification: A technical review. *International Journal of Computer Applications, 28*(2), (pp. 37-40).

Damashek, M. (1995). Gauging similarity with *N*-grams: Language-independent categorization of text. *Science, 267*, (pp. 843-848).

Deepamala, N., & Ramakanth Kumar, P. (2012). Language identification of Kannada language using *N*-gram. *International Journal of computer Applications, 46*(4), (pp. 24-28).

Dublin Core Metadata Initiative. (2014). *History of the Dublin Core metadata initiative.* Retrieved from dublincore.org/about/history/

Dunning, T. (1994). Statistical identification of language. Computing Research Laboratory

technical memo MCCS 94-273, New Mexico State University, Las Cruces,

New Mexico.

Durant K. T. & Smith, M.D. (2006). Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. WebKDD 2006, LNAI 4811, (pp. 187-206), Springer-Verlag Berlin Heidelberg.

Gottron, T., and Lipka, N. (2010). A comparison of language identification approaches on short, query-style texts. In Gurrin, C.; He, Y.; Kazai, G.; Kruschwitz, U.; Little, S.; Roelleke, T.; Rüger, S.; and van Rijsbergen, K., eds., *Advances in Information Retrieval, volume 5993 of Lecture Notes in Computer Science.* (pp. 611-614). Springer Berlin / Heidelberg.

Grefenstette, G. (1995). Comparing two language identification schemes. In *Proceedings of the Third International Conference on the Statistical Analysis of Textual Data (JADT '95),* (pp. 263-268).

Guenther, R. S. (2003). MODS: The metadata object description schema. *portal: Libraries and the Academy, 3*(1), (pp. 137-150).

Henrich, P. (1989). Language identification for the automatic grapheme-to-phoneme conversion of foreign words in a German text-to-speech system. In *Proceedings of Eurospeech 1989, European Speech Communication and Technology,* (pp. 220-223).

Hillman, D. (2005, November 7). *Using Dublin Core – the elements*. Retrieved from dublincore.org/documents/usageguide/elements.shtml

Hornik, K., Mair, P. Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2013). The textcat package for *N*-gram based text categorization in R. *Journal of Statistical Software, 52*(6).

Hughes, B., Baldwin, T., Bird, S., Nicholson, J., and MacKinlay, A. (2006). Reconsidering language identification for written language resources. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 485-488).

Ingle, N. C. (1976). A language identification table. *The Incorporated Linguist, 15*(4), (pp. 98-101).

Jones, K.S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), (pp. 11-21).

Jurafsky, D., & Martin, J. H. (2009). *N*-Grams. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd ed.). Upper Saddle River, N.J.: Prentice Hall.

Kim, S., Han, K., Rim, H., & Myaeng, S. H. (2006). Some effective techniques for naïve bayes text classification. *IEEE Transactions on Knowledge and Data Engineering, 18*(11), (pp. 1457-1466).

King, B. and Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the NAACL* (pp. 1110-1119).

King, B., Radev, D., & Abney, S. (2014). Experiments in sentence language identification with groups of similar languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial).*

Knudson, R. (2014, November). Language identifier for metadata records. Poster presented at the 77th Annual ASIS&T Conference, Seattle, WA.

Kulikowski, S. (1991). *Using short words: A language identification algorithm.* Unpublished manuscript.

Library fast facts. (n.d.). Retrieved from http://en.childrenslibrary.org/about/fastfacts.shtml

Liu, T., Chen, Z., Zhang, B., Ma, W., & Wu, G. (2004). Improving text classification using local latent semantic indexing. In *Proceedings of the 4th IEEE International Conference on Data Mining.* Brighton, UK., IEEE Computer Society Press, (pp. 162-169).

McNamee, P. (2005). Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges, 20*(3), (pp. 94–101).

Martins, B., & Silva, M.J. (2005). Language identification in web pages. In *Proceedings of the 2005 ACM Symposium on Applied Computing,* (pp. 764-768).

McCallum, S.H. (2004). An introduction to the Metadata Object Description Schema (MODS). *Library Hi Tech 22*(1), (pp. 82-88).

Missen, M.M., & Boughanem, M. (2009). Using WordNet's semantic relations for opinion detection in blogs. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval* (ECIR '09), Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy (Eds.). Springer-Verlag, Berlin, Heidelberg, (pp. 729-733).

Metadata Object Description Schema (MODS). (2013). *Top-level element: <language>.* Retrieved from http://www.loc.gov/standards/mods/userguide/language.html

Mustonen, S. (1965). Multiple discriminant analysis in linguistic problems. *Statistical Methods in Linguistics, 4*, (pp. 37-44).

Newman, P. (1987). Foreign language identification: First step in the translation process. In *Proceedings of the 28th Annual Conference of the American Translators Association,* (pp. 509-516).

Polpinij, J., & Ghose, A. K. (2008). An ontology-based sentiment classification methodology for online consumer reviews. In *Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, (pp. 518-524).

Prager, J.M. (1999). Linguini: Language identification for multilingual documents. *Journal of Management Information Systems, 16*(3): (pp. 71–101).

Reitz, J.M. (2004). *ODLIS: Online dictionary for library and information science.* Retrieved from www.abc-clio.com/ODLIS/odlis_m.aspx

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: An International Journal, 24*(5), 513-523.

Spink, A., Wolfram, D., Jansen, M.B.J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), pp. 226-234).

Swedish. (2015). Retrieved from: http://www.alsintl.com/resources/languages/Swedish/

Trieschnigg, D., Hiemstra, D., Theune, M., Jong, F., & Meder, T. (2012). An exploration of language identification techniques for the Dutch folktale database. In *Adaptation of Language Resources and Tools for Processing Cultural Heritage workshop (LREC).*

Wallace, R. S. (n.d.). Zipf's Law (A.L.I.C.E. AI Foundation). Retrieved March 5, 2015, from http://www.alicebot.org/articles/wallace/zipf.html

Zeng, M.L., & Qin, J. (2008). *Metadata.* New York, New York. Neal-Schuman Publishers, Inc.

Zhao, L., & Li, C. (2009). Ontology based opinion mining for movie reviews. *Knowledge Science, Engineering, and Management: Lecture Notes in Computer Science*, 5914, (pp. 204-214).