Language Identification: Examining the Issues

Penelope Sibun[†] and Jeffrey C. Reynar[‡]
[†]The Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201
sibun@ils.nwu.edu
[‡]Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
jcreynar@unagi.cis.upenn.edu

Abstract

We examine the use of a simple technique for identifying the language of either an online text or a hardcopy document that has been converted to a character-shape based representation after scanning. Such a text may be as brief as a single line in length. There are several important issues in constructing an accurate, robust, and fast system to identify Roman-alphabet languages. These include the features to be used, the size of the input text, and the type of analysis used for identification. We present a working system that performs well on brief texts and that requires small quantitaties of language-tagged training material.

1 Introduction

Language identification has become an increasingly important issue as more and more language data are making their way online. For books and other physical documents, a good librarian can glance at a page and identify the languages of the text thereon. But since there is much online text that may not pass immediately before a human's eyes, computational systems need to be built that can perform this task. In this paper, we identify some of the factors to be considered and some of the choices to be made. We then discuss the application of a simple, well-understood technique to the task of language identification. We argue that language identification is straightforward and admits the use of standard techniques requiring modest amounts of training data.

Some of the considerations that we will address are:

1. **Type of significant features.** Should we use characters, words, or n-grams of either? Should we use linguistic rules, such as those for morphology, orthography, or capitalization?

2. **Form of analysis.** What algorithm should we use to determine the language of a text? Some approaches may be manual and require people to process texts, while others may be completely computational. Some approaches may be suited to detecting sequences, such as the characters that make up short words, while others depend on frequency analyses. Some methods may work well in distinguishing a few languages but may not scale up well to a few dozen.

3. **Form of encoding.** What is the best representation to work from, for accuracy, robustness, and speed? Possibilities include a comprehensive character coding, a simplified character coding (formed,

for example, by removing accents), and a coarser shape-based representation, possibly derived from document images.

4. **Constituency of language pool.** What assumptions should we make about which languages need to be distinguished? Do we want to be able to discriminate every Roman-alphabet language or to focus on a subset of "important" languages?

5. **Form of input.** Will a system work with online character-coded text, online images, or both?

6. **Size of text.** Can we identify the language of a sentence or a phrase as easily as that of a page or a paragraph? We will assume first of all that there is a need for knowing the language of brief passages of text.[1] It may sometimes be the case, of course, that it is sufficient to know the main language of a text, disregarding any embedded foreign-language bits. But sophisticated linguistic processing of a specific passage requires knowledge of the particular languages involved. Systems that can identify brief texts well should also scale up to larger texts. In fact, if performance is satisfactory on small texts, the language of a larger text may be most rapidly identified by analyzing random samples from the document.

7. **Appropriateness of statistical method.** Does the method meet our criteria for the behavior that we expect from a language identification system? We expect the statistical model to have certain reliable properties: performance should not degrade when identifying longer passages of text; performance should not degrade when more training data are used for each language; and the frequency of features, rather than simply their rank order, should be accurately modeled.

We expand on the issues enumerated above in the next section.

[1] For example, knowing the language of a passage would useful to an information retrieval system. With such knowledge, it could generate independent indices for each of the languages present in multilingual documents.

2 Language Identification Issues

Automatically determining the language of a document has many potential uses. Librarians and others who work with multilingual documents but who do not know the language of each of the documents in their collection would be helped by reliable language determination. Many automated systems, such as optical character recognition, information retrieval, and speech synthesis would also benefit. A wide range of methods have been proposed.

Some work on language identification (e.g., Ingle (1976) and Newman (1987)) is in the form of guides for humans, such as librarians or translators. Because people are, in fact, good at this task, these guides cover a large number of languages and the full range of Roman-alphabet characters that can appear on a printed page. But, since we want to avoid human intervention, most of the systems we will discuss are computational. We will consider these systems in light of the issues we raised in the introduction.

A variety of features have been used for language identification. These include: the presence of particular characters (Mustonen, 1965; Newman, 1987; Ziegler, 1991); the presence of particular words (Ingle, 1976; Henrich, 1989; Kulikowski, 1991; Batchelder, 1992; Souter *et al.*, 1994); the presence of particular character n-grams (Henrich; 1989; Ziegler, 1991; Souter *et al.*, 1994); and particularly-shaped words from images (Nakayama and Spitz, 1993; Sibun and Spitz, 1994). The frequency of character n-grams was used by Beesley (1988), Henrich (1989), Cavnar and Trenkle (1994), Dunning (1994), Souter *et al.* (1994), and Damashek (1995).

A number of analytic techniques have been employed, ranging from completely manual, (Ingle, 1976; Newman, 1987), to semiautomatic (Kulikowski, 1991), to fully automatic. Batchelder (1992) trained a neural network to distinguish languages. Both Henrich (1989) and Ziegler (1991) incorporated a diversity of knowledge into expert systems. Mustonen (1965), Nakayama and Spitz (1993), and Sibun and Spitz (1994) employed forms of discriminant analysis. Beesley (1988) used language-modeling techniques originally developed for cryptanalysis. Markov models were used by

Dunning (1994). One of the methods developed by Souter and his colleagues (1994) tested for the presence of unique character sequences. Henrich (1989), Cavnar and Trenkle (1994), and Souter *et al.* (1994) built task-specific statistical models. Damashek (1995) used a model that computed dot-products of frequency vectors. We address the question of the appropriateness of some of these models below.

Whereas human-oriented techniques exploit the full range of character encodings, automatic methods are limited to standard character sets. Most systems only handle ASCII text, which lacks diacritics and many symbols. However, Ziegler's system (1991) exploited the Latin subset of Unicode (Unicode is designed to encode all writing systems). Beesley (1988) used ASCII text augmented with ASCII-encoded accent information. Language identification work at Fuji Xerox Palo Alto Laboratory has concentrated on document images (Nakayama and Spitz (1993) and Sibun and Spitz (1994)).

All the automated systems distinguish ten or fewer languages except Sibun and Spitz's (1994), which handles 23 languages. Cavnar and Trenkle's task (1994) was to determine a document's newsgroup membership, not its language. The documents tested belonged to one of fourteen newsgroups; each was written in one or more of eight languages. Damashek (1995) demonstrated a similarity-based clustering of 31 languages. Strictly speaking, he did not present any language identification results. He did, however, indicate that his system would be useful for language identification.

Nakayama and Spitz (1993) and Sibun and Spitz (1994) identify the language of a document image, while all of the other automated systems work from online character-coded text.

The results presented in previous papers may be adequate for the task of language identification. However, the technique we will describe not only is adequate for language identification but has the properties we deem most desirable. That is, it performs well on brief passages of text, requires minimal amounts of language-annotated training data, is easy to implement, and is based on a well-understood technique. In addition, its performance is comparable to previous systems and it discriminates a wide variety of Roman-alphabet languages. Systems that have performed best include: Cavnar and Trenkle's (1994) at 99.8 percent accuracy on longer articles, (for a slightly different task—identifying language-based Usenet newsgroups); Dunning's (1994), which achieves 99.9 percent accuracy discriminating two moderately-related languages, English and Spanish; Henrich's (1989), which yields 100 percent accuracy when discriminating English, French, and German, when using language-specific rules; and the system of Souter *et al.* (1994), which discriminates nine languages with 100 percent accuracy using character bigrams with test texts 200 characters in length.

Much previous research has not been explicit about the size of texts whose language is identifiable by the reported method. Only Henrich (1989), Kulikowski (1991), Ziegler (1991), Dunning (1994), and Souter *et al.* report results specific to texts of one line (approximately 80 characters) or less. Not surprisingly, in nearly all cases, accuracy is inversely proportional to text size. (Henrich's system (1989) is an exception: if the text is the size of a sentence or smaller, he switches from a statistical method to a knowledge-intensive method that yields high accuracy.)

Any statistical method must be chosen carefully to ensure that it is appropraite to the task. We have specified what is means for a statistical model to be approprate for language identification. For example, correct classification of documents in one language should not depend on the presence in the training set of documents in some other language. Sibun and Spitz (1994) report that when using linear discriminant analysis, the removal of one language may decrease accuracy across all remaining languages, which is not desirable behavior. Another concern is the precise modeling of frequency information, if the method is one that uses frequency of some set of features for language discrimination. Cavnar and Trenkle's (1994) out-of-place method compares rankings of the most frequent n-grams to assign a document to a class. However, this method discards information: two languages may have similar rankings, but different absolute frequencies. In such a case, Cavnar and Trenkle's method might fail to distinguish the two languages, whereas a method based on absolute frequencies could suceed.

In the next section, we present a method for language determination that we believe simply and successfully addresses the issues we have

raised here.

3 Method

Our approach to language identification uses relative entropy, a well-known information theoretic measure also known as Kullback Leibler distance. The relative entropy between two probability distributions reflects the amount of additional information necessary to encode the second distribution using an optimal code generated for the first distribution. Practically, it is a useful measure of the similarity between probability distributions. Relative entropy ranges from 0 to ∞ with the minimum generated when the two distributions are identical. Technically, the measure is not a distance because it is not symmetric. The equation for the relative entropy (taken from Cover and Thomas (1991)) is shown below. The conventions Cover and Thomas assume regarding 0 values are also used: $0 \cdot \log \frac{0}{q} = 0$ and $p \cdot \log \frac{p}{0} = \infty$.

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)}$$

Applying this technique to language identification is straightforward. A portion of the subset of a corpus associated with a particular language is chosen as a training set for that language. The selection methodology varies and is explained below. This is done for each of the languages being discriminated. From each training set, a probability distribution is generated by first counting particular events, such as character unigrams, bigrams, or trigrams. Smoothing ensures that events found in the test data but not seen in the training data do not cause the relative entropy score to become ∞: a count of 0.5 is added to events seen in at least one language in the training set, but not seen for a particular language.[2] Finally, these counts are converted to probabilities.

Test data, which are constrained to not overlap with the training data, are also selected at random. For each of the test sets, a probability distribution is generated in the manner described above, but without the smoothing step. A test set is assigned the language which minimizes the relative entropy between the probability distribution of the test set and the distribution associated with the training set for that language. Let p be the distribution associated with test set P, and q_i be the distribution associated with training set Q_i, whose language is $L(Q_i)$. Assume the training sets are numbered 1 through n. We assign $L(P) = L(Q_x)$ where $x = \arg \min_{1 \leq i \leq n} D(p \parallel q_i)$.

4 Data

The simple language identification technique described here was applied to two types of data: on-line character-coded data and data derived from document images.

Our character-coded data are drawn from the European Corpus Initiative CD-ROM. We used a subset of the first and second parts of the ECI corpus, which were the highest-quality portions of the collection. This subset contains documents in the following 18 Roman-alphabet languages: Albanian, Croatian, Danish, Dutch, English, Estonian, French, German, Italian, Latin, Lithuanian, Malay, Norwegian, Portugese, Serbian, Slovak, Spanish, and Turkish.[3] All of these data are encoded in ISO Latin-1, an 8-bit superset of ASCII which contains accented characters (e.g., è) and additional symbols (e.g., ¶). We used the ECI-provided filters to remove markup from the data so that statistics were gathered about only the textual portions of each document. Files that contained extensive markup after passing through these filters were discarded.

Our document image database, initially reported on in Sibun and Spitz (1994), cur-

[2] If an event encountered in the test data was not present in the training data for any of the languages being discriminated, it is ignored. This poorly approximates the results that could be achieved using a more intelligent smoothing technique, but since the space of trigrams (or bigrams, for that matter) is relatively large ($256 \cdot 256 \cdot 256 = 16$ million for an 8-bit character set such as ISO Latin-1) compared to the number actually encountered, the probability distributions generated from small training sets would be swamped by simple smoothing approaches and more complicated approaches would significantly decrease the system's speed and increase memory requirements. For a discussion of smoothing issues, see Church and Gale (1991).

[3] We were surprised to find Serbian in this set, since this language is typically written in a Cyrillic alphabet. Apparently, this text is in Romanized Cyrillic.

Characters	Character shape codes
bdfhklt#$&%A-Z0-9*ß	A
ce	e
çgpqy	g
àáâèéêiîñôùû	i
j	j
n	n
amorsuvwxz	x
;:	:
?!	!
’‘	’
,	,
.	.
=	=
-~	-
<>[](){}\/ |	|
_ (underscore)	_

Table 1: Characters and character shape codes.

rently contains 892 one page document images in the following 27 Roman-alphabet languages: Afrikaans, Catalan, Croatian, Czech, Danish, Dutch, English, Finnish, French, (Irish) Gaelic, (Scottish) Gaelic, German, Hungarian, Icelandic, Italian, Norwegian, Polish, Portugese, Rumanian, Slovak, Spanish, Swahili, Swedish, Tagalog, Turkish, Vietnamese, and Welsh. Data generated from these images comprise classifications of characters based on gross morphology (for example, all characters with ascenders are placed in one class); each character class is represented by an arbitrary ASCII character, the character shape code. See Table 1 for this mapping. Refer to Sibun and Spitz (1994) for more details on the character-shape coding process.

5 Results

Four types of experiments were conducted using the relative entropy technique outlined above on the data from the ECI corpus. In all of these, the training data and the test data were both selected randomly from the subcorpus for each language and were nonoverlapping. Results presented here reflect averages of 100 random selections of both a test set and a training set for each of the languages involved. The 18 ECI languages were discriminated using either character unigram statistics or character bigram statistics. Either 200 or 2000 lines of training data were used, and 1, 5, 10, or 20 lines of test data. (Blank lines were eliminated from the corpus.) Although we are measuring our test data by lines, we can say that one to five lines approximate a sentence, and five to 20 lines approximate a paragraph.

In the first type of experiment, the 18 ECI languages were discriminated using statistics about the character frequency of the entire range of the ISO Latin-1 alphabet. The results are shown in Table 2. As the quantity of test data increases with the amount of training data fixed, accuracy increases; with the amount of test data fixed, if the quantity of training data increases, accuracy also increases. In addition, as Table 3 shows, the standard deviation across trials with a particular set of parameters (i.e., amount of training data and amount of test data) generally decreases as more training data is used and as the move from unigram to bigram statistics is made. A confusion matrix for single line identification using 2000 lines of training data and character bigrams is shown in Table 4.

The results using this technique compare favorably with the other methods we are aware of, and the task is more difficult than discriminating just a few languages. However, one could

Statistic used	Lines of training data							
	200				2000			
	Lines of test data				Lines of test data			
	1	5	10	20	1	5	10	20
Unigrams	78.2	96.7	98.5	99.3	81.5	97.2	99.0	99.4
Bigrams	90.2	99.6	99.9	100.0	94.1	99.7	99.9	100.0

Table 2: Percent correct when discriminating 18 languages from the ECI corpus.

Statistic used	Lines of training data							
	200				2000			
	Lines of test data				Lines of test data			
	1	5	10	20	1	5	10	20
Unigrams	10.2	3.9	2.9	1.8	9.4	4.0	2.4	1.7
Bigrams	6.6	1.4	0.5	0.0	5.1	1.1	0.8	0.0

Table 3: Standard deviations of correctness across trials when discriminating 18 languages from the ECI corpus.

argue that the task is made easy by the ECI data, because different languages use different special characters, and sharp differences in character sets are driving the recognition process. To test this, we collapsed the digits into a single class, treated all white space as identical, and collapsed all remaining non-alphabetic characters into a single class. The results shown in Table 5 using this technique under the same conditions as previously described demonstrate that special characters contribute little to the performance of the language identifier.

In order to facilitate comparison with those methods that discriminate a small number of languages, we tested the relative entropy technique on the English, French, and German portions of the corpus. In one variation, the entire ISO Latin-1 alphabet was used. In the other, ISO Latin-1 was converted to ASCII by deaccenting characters and removing characters not found in the ASCII character set. This conversion makes our data similar to those of Henrich (1989), who discriminated these three languages from ASCII text. This conversion also simulates conventions used in writing French and German in ASCII-based forums, such as Usenet newsgroups, that are often used as sources for text data (cf. Cavnar and Trenkle (1994)). See Table 6 for results for a variety of conditions. These results are averages of 50 random test and training data selections, rather than 100 as in the 18-language cases described above. Two hundred lines of training data were used for these experiments.

Two types of experiments were conducted using the data from document images. In the first type, the relative entropy technique was used to discriminate 27 languages. 892 one-page documents in these languages were available, and due to the small number of documents in each language, training data were generated from all of the documents except the one whose language was being identified. Unigram, bigram, and trigram character statistics were used; see Table 7. These results compare favorably with those presented in Sibun and Spitz (1994), where overall accuracy was slightly better than 90 percent.

The remaining experiments conducted on data derived from document images were limited to discriminating English, French, and German and used the same training set generation methodology. Results are shown in Table 8.

6 Conclusion

We have examined some of the issues in the automatic identification of the language of a text and have presented results from our language identification system. While some recent sys-

	alb	cro	dan	dut	eng	est	fre	ger	ita	lat	lit	mal	nor	por	ser	slo	spa	tur
alb	98	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
cro	0	94	0	0	0	0	0	1	0	0	0	0	0	1	0	4	0	0
dan	0	0	93	0	0	0	0	0	0	0	0	1	5	0	0	1	0	0
dut	0	0	0	97	0	0	0	2	0	0	0	1	0	0	0	0	0	0
eng	0	0	0	0	97	1	0	0	0	1	0	1	0	0	0	0	0	0
est	0	0	1	0	0	88	0	1	1	0	2	2	0	0	4	1	0	0
fre	0	0	0	0	0	0	98	0	0	0	0	0	0	2	0	0	0	0
ger	0	0	1	0	0	2	1	94	0	0	0	0	0	2	0	0	0	0
ita	0	0	0	0	1	0	1	0	95	1	0	0	1	0	0	0	1	0
lat	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
lit	0	0	0	0	0	0	0	0	0	0	99	1	0	0	0	0	0	0
mal	0	0	0	0	0	2	0	0	0	0	1	97	0	0	0	0	0	0
nor	0	0	1	1	0	1	1	1	0	0	0	0	93	0	1	1	0	0
por	0	0	0	0	0	0	1	0	0	0	0	0	0	96	0	0	3	0
ser	0	7	0	0	0	0	0	0	0	0	0	0	1	1	89	2	0	0
slo	1	8	0	0	1	0	0	0	0	0	1	1	0	0	0	88	0	0
spa	0	0	0	3	2	0	8	1	1	3	0	0	0	2	0	1	79	0
tur	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Table 4: Confusion matrix for single line identification using 2000 lines of training data and bigram statistics. Rows, which sum to 100, are the correct language. Columns are the language identified.

Statistic used	Lines of training data							
	200				2000			
	Lines of test data				Lines of test data			
	1	5	10	20	1	5	10	20
Unigrams	77.0	95.5	97.9	99.0	75.7	97.2	98.2	99.3
Bigrams	91.0	99.0	99.5	99.7	90.7	99.5	99.7	99.8

Table 5: Percent correct when discriminating 18 languages from the ECI corpus using character classes.

Statistic used	ISO Latin-1			ASCII-Converted		
	Lines of test data			Lines of test data		
	1	5	10	1	5	10
Unigrams	93.7	100.0	99.7	92.7	99.3	99.3
Bigrams	98.3	100.0	100.0	98.7	100.0	99.3

Table 6: Percent correct when discriminating only English, French, and German from the ECI corpus.

Statistic	Accuracy
Unigrams	81.3
Bigrams	95.7
Trigrams	97.3

Table 7: Percent correct when discriminating 27 languages using data from document images.

Statistic	Accuracy
Unigrams	93.4
Bigrams	99.2
Trigrams	99.2

Table 8: Percent correct when discriminating English, French, and German using data from document images.

tems perform with comparable accuracy, our system uses a simple, well-understood technique that requires small quantities of language-tagged training data. As we have shown, our system is further distinguished by its ability to discriminate a large number of Roman-alphabet languages, by its ability to do so for very brief texts, and by its demonstrated applicability to both online text and character-shape coded text generated from images.

7 Future Work

In order to test our approach in a more realistic setting, we are integrating our system with one that uses layout information to extract from document images meaningful areas of text such as paragraphs and captions; see Ozaki (1995). Decomposing a document into logical units in this way allows us to accurately identify the language of brief passages of multilingual texts. In addition, we would like to explore more principled approaches to smoothing the training data to account for unseen events.

8 Acknowledgements

We would like to thank Masa Ozaki, David Hull, and Larry Spitz for valuable comments. The majority of this work was conducted while the authors were employed at the Fuji Xerox Palo Alto Laboratory.

References

Batchelder, Eleanor Olds. *A Learning Experience: Training an Artificial Neural Network to Discriminate Languages.* Unpublished Technical Report, 1992.

Beesley, Kenneth R. "Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text." In *Language at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*, 12-16 Oct 1988, pp. 47-54.

Cavnar, William B. and John M. Trenkle. "N-Gram Based Text Categorization." In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, 11-13 April 1994, pp. 161-169.

Church, Kenneth W. and William A. Gale. "A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams." *Computer Speech and Language*, Vol. 5, 1991, pp. 19-54.

Cover, Thomas and Joy A. Thomas. *Elements of Information Theory.* New York: John Wiley & Sons, Inc., 1991.

Damashek, Marc. "Gauging Similarity with n-Grams: Language-Independent Categorization of Text." *Science*, Vol. 267, 10 February, 1995.

Dunning, Ted. *Statistical Identification of Language.* CRL Technical Memo MCCS-94-273, 1994.

Henrich, Peter. "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*, Paris, Sept. 1989, pp. 220-223.

Ingle, Norman C. "A Language Identification Table." *The Incorporated Linguist*, Vol. 15 No. 4 pp. 98-101, 1976.

Kulikowski, Stan. *Using Short Words: A Language Identification Algorithm.* Unpublished Technical Report, 1991.

Mustonen, Seppo. "Multiple Discriminant Analysis in Linguistic Problems." *Statistical Methods in Linguistics*, No. 4, Skriptor Fack, Stockholm, 1965, pp. 37-44.

Nakayama, Takehiro and A. Lawrence Spitz. "European Language Determination from Image." In *Proceedings of the International Conference on Document Analysis and Recognition*, 20-22 Oct. 1993, pp. 159-162.

Newman, Patricia. "Foreign Language Identification: First Step in the Translation Process." In *Proceedings of the 28th Annual Conference of the American Translators Association*, 8-11 October 1987, pp. 509-516.

Ozaki, Masaharu. "Logical Tagging of Document Images by White Space Pattern Matching." In *Shape and Structure in Pattern Recognition.* Dov Dori and Alfred Bruckstein, editors. Singapore: World Scientific, 1995.

Sibun, Penelope and A. Lawrence Spitz. "Language Determination: Natural Language Processing from Scanned Document Images." In *Proceedings of the Fourth Applied Natural Language Processing Conference*, Stuttgart, Germany, pp. 15-21, 1994.

Souter, Clive, Gavin Churcher, Judith Hayes, John Hughes and Stephen Johnson. Natural Language Identification using Corpus-Based Models. In K. Lauridsen and O. Lauridsen (guest editors), *Hermes Journal of Linguistics*, vol. 13, pp. 183-203. Faculty of Modern Languages, Aarhus School of Business, Denmark, 1994.

Ziegler, Douglas-Val. *The Automatic Identification of Languages Using Linguistic Recognition Signals.* Dissertation, State University of New York at Buffalo, 1991.