choose exponentially decreasing weights for ranks, the score difference between a class occurring at rank 1 a few times (English) and a class occurring mostly at rank 2 for a large number of words (Marathi) will be wide enough to allow correct discrimination of true language classes based on the cumulative scores.

3. After this has been done for all the words, a cumulative score for each language class is computed as the sum of the weights of all the words for which this class has been assigned.

4. Steps 1-3 are repeated after reducing the value of K, till we get the required number of m classes, say $m = 2$ or $m = 3$.

In the language enumeration stage, we do not necessarily have to select exactly the number of languages which are supposed to be in the document if our final purpose is to identify the language-encodings of segments or words. For example, if the document is known to be bilingual, we can select the top three language-encodings. This will ensure that in cases of errors in enumeration, we do not miss out on one of the correct languages in the document while identifying the segments or words. However, our results show that this logic may not apply in practice. Also, during this stage (*i.e.*, enumeration), we do not consider small words (less than 6 bytes), though we do consider them while identifying the language-encoding of word types and tokens.

There is another technique that our system uses optionally for language enumeration. This is based on preparing the list of unique n-grams (UNGs) for each language-encoding class. The accumulated scores of each class are multiplied by the normalized count of unique n-gram matches between the n-gram model of the class and the n-gram model of the test data.

6.2. Word Type Identification

Once the best possible m language-encodings have been identified for the document, we can simply use the monolingual identifier to tag the language-encoding of each word type. The important point is that we only have to discriminate between m classes and m will be usually only 2 or 3.

6.3. Word Token Identification

In the current work, we assign the language-encoding class of a word token to be the same as that of the word type of which it is an instance. In other words, we are not taking the context of the token into account. We plan to explore how context can be used to improve token identification.

7. Language Distances and Confusability

Intuitively, the difficulty of identifying the correct language out of two possible candidates should be more if the two candidates are closer, *i.e.*, if the languages are related.

This fact can also be stated in terms of the linguistic notion of *language distance* or *divergence*, sometimes also called the *genetic distance* between languages. Earlier attempts at this were based on comparing a list of (say, 200) words (Swadesh 1952). For purposes of information extraction, linguistic distances were adapted from the Dyen *et al.* (1992), who also used Swadesh like lists of words. In general, the language-language distance can also be calculated if the distribution of words is known, by using a distributional similarity measure like relative entropy. Nerbonne and Heringa (1997) measured dialect distance phonetically. Ellison and Kirby (2006) recently described an attempt at building genetic language taxonomies using a measure based on language internal similarities within the forms.

From our point of view, there is one important element missing from these methods of finding language distance: the encoding of the document. If the text in different languages is encoded differently and there is no converter available to convert between all possible pairs of language-encodings, then we can't directly work on word lists or distributions of words. What we need is a measure of distance between two language-encodings which gives us a quantitative measure of the confusability of two language-encodings. We have used a simple method for this. We are preparing byte based n-gram models from the training data for all language-encodings being considered. If we compare these models among themselves using a distributional similarity measure such as symmetric or mutual cross entropy, we will get an estimate of language-encoding distance. We can then study the relationship between identification precision and language-encoding distance (Figure 3).

	Enumeration Precision		Segment Identification Precision			
	(2 out of 2)	(2 out of 3)	Word Type Precision		Word Token Precision	
			Unknown	Known	Unknown	Known
Related	88.92%	97.87%	78.15%	85.85%	74.45%	81.96%
Less Related	87.55%	94.51%	81.66%	91.93%	77.75%	88.10%
Unrelated	85.51%	95.89%	81.97%	93.68%	77.86%	89.16%
Mixed	86.93%	96.19%	80.73%	90.91%	76.82%	86.80%

Table 3. Precision for Language Enumeration and Segment Identification

	Languages Unknown		Languages Known	
	50-50	80-20	50-50	80-20
Enumeration	93.59	83.53	-	-
Token	81.06	74.75	86.82	86.75
Type	85.43	78.24	90.96	90.87

Table 4. Precision for Two Different Ratios of Languages in a Document

8. Evaluation

We tested our method on bilingual documents with a fairly high level of diversity. We evaluated language enumeration as well as segment (word) identification. Since lan-