guage distance or confusability can significantly affect the results (section-7), we divided the language-encodings in four categories based on the distances among languages: (a) unrelated, (b) less related, (c) related and (d) mixed. Our evaluation is on these categories (Table 2). We experimented with and without using UNGs, but there was not much difference, which shows that either UNGs have no effect on the precision, or a more effective way of using them has to be found.

Since it is very difficult to get test data (for evaluation) under the high diversity assumption where not only the language-encodings of the document, but the language-encodings of each word or segment are known, we generated such data from monolingual documents by mixing words from documents in different language-encodings randomly, but in definite proportions and preserving the sequential order. The two proportions we used for evaluation were 50%-50% and 80%-20%. The final precision was averaged over documents containing language-encodings in these proportions. The maximum document size was kept at 1000 words. Since the proportion affects precision (though not uniformly), we also report the results for two different proportions. In cases where one language is in much less proportion than the other (80%-20%), the performance was lower in most cases.

For language enumeration, we performed evaluation for the cases when both the language-encodings are correctly identified and also when two out of three are correctly identified. The results are presented in Table 3. In all, we tested on 8498 documents, out of which two out of three were correctly identified in 8175 (96.20%) documents. Both were identified correctly in 7388 documents (86.94%). The separate results for the two proportions (50%-50% and 80%-20%) are shown in Table 3.

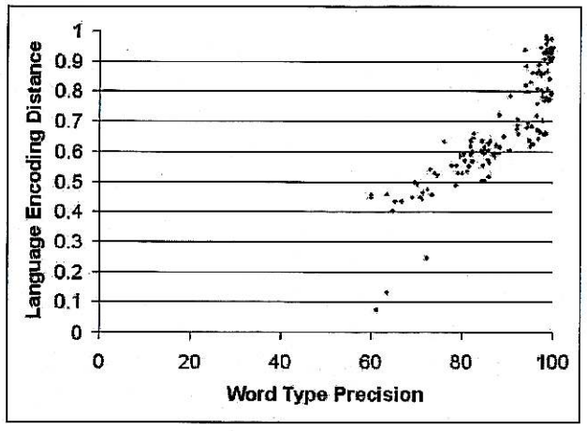

Figure 3. Precision vs. Language-Encoding Distance

For word identification, we have calculated the precision for type as well as token. We considered two cases here, one when the languages in the document are known in advance and the other when they are not. The second case gives the combined precision for language enumeration and word identification. The total number of types tested on (in all documents) was 2447424, out of which 2225079 (90.92%) were correctly tagged. The total number of tokens tested on was 3976754, out of which 3451698 (86.80%) were correctly tagged. The results are shown in Table 3. The results show that word identification for unknown languages is better when we take the top 2 instead of the top three language-encodings during the enumeration stage. This is because the top 3 usually include a language-encoding similar to the one dominant in the document (in terms of proportion) and this makes word identification difficult.

We tried to study the relationship between precision and language-encoding distance as defined in section 7. The precision has been plotted against normalized distance in Figure 3. It clearly shows that the problem is harder for closer language-encodings.

9. Future Directions

There can be irresolvable ambiguity in identification because the same word may belong to many languages and those languages may be using the same encoding. This makes segment identification very difficult. We plan to explore ways to overcome this difficulty. For segment token identification, we can also take the context into account. One way to do this will be to find out places where sudden drops in n-gram based probabilities of word sequences occurs. We also plan to modify our system so that it can make use of any available prior information about the scripts and the encodings (*e.g.*, when the documents are in Unicode), the languages, the charsets and the fonts (Shusha font maps to Hindi with Shusha Phonetic encoding) from the tags in web pages, etc.

10. Conclusion

We explored the problem of language and encoding identification in multilingual documents, including crucial assumptions such as global and local diversity assumption etc. We divided the problem into three parts: (a) monolingual identification, (b) language enumeration and (c) segment identification. A method for (b) and (c) was presented. We performed a fairly extensive evaluation on bilingual documents. Enumeration precision was calculated for the case where two out of three language-encodings were correctly identified (96.20%) and also for the case when both were correctly idetified (86.94%). Word token and type precision was calculated when the language-encodings in the document were known (86.80% and 90.91%) and also when they were not known (76.82% and 80.73). The results are promising, but have scope for improvement. We also discussed language distance in the context of electronic documents and showed that it has to be defined differently for electronic text. We found the precision to be higher for distant language-encodings.