	German	French	Spanish	Chinese (Traditional)	Chinese (Simplified)	Japanese
Slogan	1.30	1.33	1.06	0.19	0.01	0.10
Piece	1.23	1.52	1.32	1.22	1.15	1.48
Peace	1.53	1.65	1.44	1.23	1.15	1.85
Town	1.61	1.53	1.49	1.36	1.40	1.56
Trouser	0.10	0.06	0.76	0.02	0.02	0.60
Clutter	0.32	0.08	0.05	0.03	0.08	0.07
Down	4.48	1.67	1.44	1.28	1.5	1.64

Table 1. Multilingual Pages on the Web: These statistics indirectly indicate the number of bilingual pages on the Web. The numbers (in millions) are actually the number of results returned by Google when an English word was searched among pages of some other languages. The English words searched were deliberately selected to be of different origins: Latin, Celtic, Germanic etc. This was done to take care of the cognate words factor. The words are also diverse in terms of their frequency of occurrence in English.

In this paper, we will discuss the problem of multilingual language identification and consider different scenarios and the assumptions they imply or require. The solution to the problem will depend on these assumptions. We also show that the problem can be divided into three parts and these parts can be solved separately. The first part is monolingual identification. Many methods with very high precision are available for this part. The second part is **language enumeration**, *i.e.*, finding out what languages are present in the document. The third part is **segment identification**, *i.e.*, identifying the language of segments of text in the document. If the segments are assumed to be single words, we can further divide the problem into **word type identification** and **word token identification**. In this first work on formulating the problem of multilingual language identification and solving it in a systematic way, we propose a method to solve the language enumeration and segment identification problems under one of the most likely scenarios. We have evaluated these methods fairly extensively. The results achieved are highly encouraging. We also consider the relationship between precision (of identification) and the distance between language-encodings. Throughout this paper, *identification* means *language and encoding identification*, unless stated otherwise.

2. Assumptions

The solution to the multilingual identification problem completely depends on the assumptions we make. One or two of these assumptions may be unavoidable to make the problem tractable. Some of these assumptions have been given in below.

1. **Diversity Assumption**: The accuracy of a language identifier depends on the number of languages from which the identifier has to select one. This reflects the coverage of the identifier in terms of linguistic diversity, which implies an assumption about linguistic diversity. There are two kinds of *diversity assumptions*, both of which can be applicable at the same time.

   (a) **Global Diversity Assumption**: This is about how many languages are assumed to be in the world. In practical terms, this is reflected in the number of languages for which the system has been trained.

   (b) **Local Diversity Assumption**: For a particular user or for a particular context, the number of possible and **relevant** languages may be less than the number for which the system has been trained. For example, a user may only be interested in the documents in European languages, even though the system has been trained for languages from around the world. In such a case, a *local diversity assumption* is likely to increase the accuracy and speed of the identifier.

2. **Limited Ambiguity Assumption**: Multilingual documents have text in more than one language, but if we do not assume a small limit to this number, the problem may not be tractable, unless we assume large segment sizes. All the algorithms for monolingual identification work well only when the test data size is sufficient, *e.g.*, 100 characters. Thus, to make the problem solvable, we will make the *limited ambiguity assumption*, viz., that the number of languages to be disambiguated *for a segment* is a very small number. In our experiments, we have assumed this number to be either two or three, which means that the multilingual documents can be either bilingual or trilingual. Unlike the *diversity assumption*, this assumption is about the possible languages in a document, not in the world. Therefore, it applies only to a multilingual identifier, not to a monolingual identifier.

3. **Language Switching Assumption**: Another assumption that applies only to a multilingual identifier is the *language switching assumption*. This specifies how frequently or where a shift from one language to another can occur in a document. There are two such assumptions, only one of which can apply at a time.

   (a) **Long Sequence Assumption**: This assumption says that the minimum segment size in any language is large enough for a monolingual identifier to identify its language accurately. If we make this assumption, the problem of segment identification actually becomes a problem of identifying where language shift occurs and from which language to which language. This is, of course, a less realistic assumption.

   (b) **Isolated Word Assumption**: The more realistic assumption is that every word in the document can be in a different language, subject to the *limited ambiguity assumption*, *i.e.*, language switch can occur at any word boundary. Our experiments have been conducted under this assumption. The problem in such a case is to identify the language of every word, as every word is a segment. In one sense, this is a simpler problem because we do not need to identify the boundaries of the segments. However, since the segment size can be as small as a word of one character, the precision is likely to be low.