

Re: check for non-english

Source: <http://sci.tech-archive.net/Archive/sci.lang/2006-07/msg00923.html>

- *From:* "Dylan Sung" <dylanwhs.tsktsktsk@xxxxxxxxxxxxxxxx>
 - *Date:* Mon, 17 Jul 2006 12:19:21 +0100
-

<todaysmulan@xxxxxxxxxxxxxxxx> wrote in message
<news:1153098675.133026.198010@xx>

how does a translation machine check for chinese korean japanese and other weird languages ?

By looking at what code the computer encounters in the document. Chinese from mainland China use GB encoding, whilst Chinese from Taiwan uses Big5. Japanese can be found encoded in JIS, Shift-JIS and other encodings, and Korean has its own too. The characters used in each set of encodings is slightly different, and from any text the character codes fall into certain ranges which can be used as a guess as to what encoding, and hence language it comes from. However, codings like GB, JIS and Korean EUC inhabit the same code ranges, which makes it difficult for a human to know what language it is in, unless he looks at the character display. This requires some knowledge of the languages concerned. If the displayed text is gibberish, it is very likely that the wrong encoding was selected. Therefore, any machine translation of that text using that would be wrong. This is why online translation software requires you to select input and target output languages.

Dyl.

.