Variations in Use of Meta Tag Keywords by Web Pages in Different Languages
Timothy C. Craven
Journal of Information Science 2004; 30; 268
DOI: 10.1177/0165551504042805

The online version of this article can be found at:
http://jis.sagepub.com/cgi/content/abstract/30/3/268

Published by:
SAGE Publications
http://www.sagepublications.com

On behalf of:
cilip
Chartered Institute of Library and Information Professionals

Additional services and information for Journal of Information Science can be found at:

Email Alerts: http://jis.sagepub.com/cgi/alerts

Subscriptions: http://jis.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 16 articles hosted on the SAGE Journals Online and HighWire Press platforms):
http://jis.sagepub.com/cgi/content/refs/30/3/268
Variations in use of meta tag keywords by web pages in different languages

Timothy C. Craven
Faculty of Information and Media Studies, Middlesex College, The University of Western Ontario, Canada

Received 17 September 2003
Revised 11 November 2003

Abstract.
Sets of top-ranking pages in 19 languages returned by the Google search engine were downloaded and their titles and meta tag keywords analyzed. Results showed significant differences in proportion of pages with keywords depending on language; specifically, pages in Dutch, French, and German showed the highest proportions with keywords, while pages in Chinese and Korean showed the lowest proportions. Keywords were mostly in the languages of the pages, though on Chinese, Greek, Indonesian, and Turkish pages keywords in English or in English mixed with other languages predominated. The proportion of very long titles also varied significantly with language, with nearly 10% of titles on Russian pages exceeding 100 bytes, in contrast to less than 1% on Chinese, Finnish, Indonesian, and Polish pages. Both standard ASCII extensions and character entity references were used to code special characters in titles.

Keywords: world wide web; web pages; indexing; metadata; meta tags; keywords; titles; language variations; character sets; character coding; ASCII extensions; HTML character entity references

1. Introduction

1.1. Background

This paper constitutes a follow-up to a recent study of the use of meta tagged descriptions in Web pages in languages other than English [1]. It forms part of a series of research reports on how people and organizations summarize Web pages, especially how they summarize their own Web pages in descriptions in meta tags [2–6], though also to some extent how they summarize external Web pages [7]. A specific motivation on the part of the author for undertaking these studies was to apply an existing computerized abstractor’s assistant [8, 9] to summarizing Web pages. In addition to providing input specifically to the development of the computerized assistant, results of the studies may also have wider implications; for example, for advice to be given to Web page creators on the kinds of information to supply in meta tags or to search engine designers on the kind of information to be expected.

Earlier studies in the series examined, among other things, the degree to which words and phrases in the meta tag descriptions (more exactly, the CONTENT attributes of META tags with NAME attribute equal to ‘description’) duplicated words and phrases found in the titles and displayed text of the pages. Both titles and the first 200 words of displayed text appeared to represent potentially useful extractable elements to assist in description construction.

Results of the previous study most closely allied with the present one [1] showed significant differences in proportion of pages with descriptions and in lengths of descriptions depending on language; specifically, pages in major Western European languages showed
higher proportions with descriptions, while pages in
Chinese showed the lowest proportions. Descriptions
were mostly in the languages of the pages, though
English descriptions were provided on some non-
English pages. With few exceptions, coding schemes
adopted for diacritics and non-Roman characters were
standard.

Other aspects of the content of Web pages have been
studied by various other researchers; for example, page
layout of home pages [10]; characteristics of anchors
[11]; informetric measures [12]; links to e-journals and
their articles [13]; the feasibility of using implicit
structure in identifying differences automatically [14];
rates of change over time [15]. There has also been a
little other investigation into descriptions and key-
words in meta tags: Turner and Brackbill [16] have
reported results of a small experiment that showed that
addition of a description did not improve retrievability
of Web pages on Infoseek and Altavista; similar results
have been reported for these two search engines and
five others by Henshaw and Valauskas [17]; Drott [18]
has noted the extent to which descriptions and keyword
meta tags are used in the sites of 60 Fortune
Global 500 companies.

External descriptions of Web pages have been
examined also by Wheatley and Armstrong [19], and
Amitay [20] developed a tool called SnipIt to extract
descriptive passages with URLs from Web pages and
another tool called InCommonSense to select from
among these the ‘best’ descriptive passage for each
URL.

1.2. Purpose

‘For the most part, the domain of Internet studies is
very much dominated by the U.S.’ [21]. This has
meant, among other things, that investigations have
tended to look at practice in the English language
rather than by other major language communities.

A Global Reach survey in 2002 [22] indicated that
about 60% of Internet users were from non-English-
speaking zones. Speakers of non-English European
languages accounted for about 34%, with Spanish the
leading non-English European language, followed in
order by German, French, and Italian. Asian-language
zones accounted for about 26%, with Chinese in the
lead, followed closely by Japanese and more remotely
by Korean. English apparently still dominated in Web
content, however, at about 68%. Bilingual users may
prefer to access English-language sites [23]. Pages in
major Western European languages can be retrieved by
many Web search engines; search engines with
capacities for searching in either the traditional or
the simplified Chinese character set have been avail-
able for a number of years; less common languages,
especially those using non-Roman character sets, have
typically been supported by only a few search engines
[24].

It has long been recognised that one of the functions
of abstracts may be to ‘render the primary content of
the document in another language accessible in the
language of the abstract’ [25]. That background,
together with the growing participation of non-Eng-
lish-speaking users in the Internet, formed the chief
motivation for the previous study’s investigation of
whether meta tagged descriptions might also be
serving, among other things, to assist with language
barriers.

Moving beyond the emphasis on descriptions in the
previous study, the present study addresses the
following questions related to other, similar page
elements: to what extent Web pages in languages
other than English make use of meta tagged keywords,
and whether there are noticeable differences among
languages; how commonly English-language key-
words are supplied for non-English-language pages;
and whether there are substantial differences in the
length and language of page titles from language to
language.

Meta tagged keywords are now ignored by many Web
spiders, but they still matter to some smaller coun-
try-specific search engines [26]. Differences in adoption
from one language to another may therefore be justified
by differences in commonly preferred search engines
alone.

For multilingual sites, Dunlap [27] recommends
creating separate pages for each language, with key-
words in the language of the page. If this or similar
advice is followed generally by site creators, one would
not expect to see pages in which English keywords
were substantially mixed with keywords from other
languages. Dunlap’s examples show relatively straight-
forward translation of keyword sets from one language
to another, but bigmouthmedia [28] suggests the
desirability of further tailoring page features such as
keywords to fit the retrieval methods preferred by
search engines in specific countries.

Advice on length seems to be given less often for
keywords than it is for descriptions. Warnings against
irrelevant or repeated keywords are fairly standard [26,
29]. Estimates of a suitable number of keywords are 20
[30] or 25 to 30 [31], or no more than will fit into 1000
characters [32], or about the number of significant
words in the description [29].
2. Sample

As in the previous study, major world languages were selected using the counts given in Google’s main list (http://directory.google.com/Top/World). In addition to language, a country specification was also included: this was done primarily because Google did not accept a search in which no search term was entered; but it was also used, in the cases of Chinese, Portuguese, and Spanish, to compare results from different regions, and, in the case of Chinese, using different character sets. Although Catalan had a higher count than some of the selected languages, it was excluded because it was a ‘stateless’ language, scattered across several top-level-domains. Romanian, excluded through an oversight in the previous study, was included in the present study. For each language, one of the regions chosen was the one with the expected largest number of sites using its top-level-domain (TLD); hence, Germany for German, France for French, Brazil for Portuguese, and so on.

For each language and TLD pair the Google (http://www.google.com) ‘Advanced search’ was used to retrieve a set of Web pages in the target language. The language was selected from the ‘Language’ drop-down list. The two-letter top-level-domain code for the country was entered in the ‘with all the words’ field, preceded by a period, and occurrences were limited to the URL of the page.

The total approximate number of matching items reported by Google was noted. The first ten pages of 100 results for each pair were then saved and specially written software was used to extract the URLs of the references. A customized software package developed by the researcher was subsequently employed to access the pages and log data including title and any meta tags that they might contain.

As can be seen in Figure 1, reported match counts varied widely, the lowest being for Finnish, at 244,000 (up, however, from the count of 68,200 in the previous study), and the highest for German, at 7,210,000 (up from 5,610,000). Had other, less prominent, languages...
been included, the minimum might have been expected to be still lower.

Figure 1 also shows that accessibility of the pages listed by Google was quite good for most languages. This contrasts with the rather mediocre accessibility in the previous study. Although a possible factor is an improvement in Google’s currency in covering non-English pages, there is a change in technology used in the logging process that should not be overlooked: the previous study had employed an older HTTP client component (NEWT), while the present study used a different component (Indy); changes caused in the software by the use of the new component included support for HTTP 1.1, supplying of a fuller HTTP header that identified the client program as a version of Mozilla, and loss of the automatic 30-second timeout.

Poor accessibility may be caused by various factors, alone or in combination: slow server response, which in turn may be associated with the use of less advanced technology in some countries or with sites that have outgrown their predicted loads; Internet bottlenecks, such as may occur when following links between continents or when a server is connected with a low bandwidth line; discontinuation of Web service; large pages, which take too long to download; finicky servers, that test, for example, for use of particular browsers or require the latest version of HTTP. It may be noted that the changes introduced with the new HTTP component would tend to mitigate many of these factors.

A small proportion, 1–2%, of the 1000 for a typical set were eliminated during the extraction process because they did not use HTTP or were internal references to other Google pages.

Although there is a correlation between reported page count and accessibility of the first 1000 pages (0.43906), this appears to be due mostly to the two outliers Indonesian (.id) and Chinese (.hk simplified) and the distribution of points on the plot does not suggest a simple association, such as might be the result of the search engine’s ranking higher those pages from established, stable sites.

The accuracy of the Google search engine in actually returning pages in each requested language was tested by visual examination, using Internet Explorer, of every fifth HTML page that had been successfully accessed in the logging step. Pages were categorized as follows:

E. English only - no other language(s).
N. No English (except for the occasional English word incorporated into another language) - other language(s) only.
M. Mixture of English and non-English text (such as original language version and English translation).
U. Unknown; e.g. proper name only.
O. Element absent (e.g. no body element).
X. Page not found.

It seemed advisable to allow for some leeway in deciding the language in which a page was written, in view of the extent to which many languages borrow freely from other languages, especially from English, and especially for new and technical concepts, such as those relating to the Web. People may actually know or agree on the correct technical term in English more readily than they do in their native language [33].

The results were generally similar to those obtained in the previous study by examining the first 100 pages in each set. The proportion of English-only pages was above 5% only for the Chinese sets, Danish, Greek, Indonesian, Italian, Japanese, Romanian, and Swedish, and above 10% only for the Chinese sets other than .cn, Indonesian, and, marginally, Italian. Some caution is advisable in the analysis of results for these sets.

The categorization did not distinguish pages in the target language from pages in other languages other than English. Selective observation suggests that the latter were not particularly frequent, but the previous study had identified one particular area of higher incidence of this phenomenon: some pages in Catalan are retrieved in response to a request for pages in Spanish, one possible reason being the page creators’ mis-coding the language as Spanish in the appropriate meta tag.

Apart from accuracy, of course, representativeness of the samples must also be considered. The samples are considered to be representative of pages highly ranked by the Google search engine rather than of all Web pages in existence or available at the time the samples were taken. The number of links to a page is known to be an important factor in Google rankings, and it has been observed that the Google algorithm tends to give higher ranks to pages from established sources [34] [35]. It is not too surprising, therefore, that the sampled pages tended often to come from government departments, major corporations, and educational institutions. It was also observed that the URLs very often did not specify file names or subdirectories, but simply host and domain name, due perhaps in part to automatic URL shortening procedures. Personal pages are thus likely very under-represented in the sample sets.
3. Keywords

3.1. Presence

Figure 2 shows simply the proportion of pages accessed that contained at least one meta tag with the NAME attribute equal to 'keywords'. A small proportion of these actually contained no keywords; that is, the CONTENT attribute was null, sometimes because the page creator had incorrectly inserted a space after the 'CONTENT = ' in the HTML code and sometimes because no attempt had apparently been made to assign keywords. Conversely, a few cases of intended keywords were excluded from the count because they used the less common HTTP-EQUIV attribute and not the NAME attribute or because they misspelled 'keywords'.

It appears that European languages with speakers concentrated in mid latitudes (very roughly, around the 50\textdegree Parallel) (Dutch, French, German, Polish, Romanian, and Russian) lead, with quite strong showings for other European languages (including Turkish) and for Japanese, and with Chinese and Korean trailing. A chi-square test shows that the differences among languages in proportion of pages with keywords is highly significant ($p = 0.00000$). The proportions with descriptions closely track the keyword proportions and are similar to what was observed in the previous study.

There is a positive correlation (0.23085) between the reported Google page count and the proportion of pages with keywords. A similar correlation is observed for descriptions, as was the case in the previous study. Thus it might be hypothesized that the higher apparent use of keywords by certain languages was, at least in part, simply an artefact of the higher proportion of relatively high ranking pages in the sets for those languages. The correlation is not particularly strong, however; and a closer analysis of the data shows that about two-thirds of the value is due to a single language, German.

![Fig. 2. Proportion with keywords.](image-url)
3.2. Length

Within each language set, keywords values were divided into those up to 200 bytes in length and those more than 200 bytes in length. The 200-byte cut-off had been used in the earlier study because it was a commonly recommended maximum for descriptions and it worked well in the present study for keywords because it fell close to the observed medians. A chi-square test showed a highly significant difference by language in the proportion of keyword lists over 200 bytes (p = 0.00000). The highest proportions were for Dutch (68.1%), French (65.2%), Russian (64.8%), Italian (63.3%), and German (61.7%). Other languages scoring above 50% were Danish, Greek, Polish, Portuguese (both .br and .pt), Romanian, Spanish (both .es and .mx), and Turkish. The lowest proportions were for Chinese (.cn) (28.4%), Japanese (32.9%), Indonesian (35.6%), Chinese (.hk simplified) (36.0%), Chinese (.tw) (34.5%), and Korean (39.2%). Also scoring below 50% were Chinese (.hk traditional), Finnish, Norwegian, and Swedish.

A partial explanation for the variation in the length of keyword strings may lie in intrinsic differences between languages and writing systems. Most notably, the common coding of Chinese characters at two bytes each, with no spaces between the characters, provides, in the great majority of cases, shorter strings than would the use of the corresponding romanization. This explanation is clearly only partial, since it fails to account for the relative infrequency of longer keyword lists in the Indonesian set.

3.3. Language

Keyword sets were categorized by language or languages used. The categories used were the same as those noted for pages above.

As a consistency check, however, the researcher independently categorized the first 100 records for each language-region set. Overall consistency was acceptable, at 93.5%; consistency of less than 90% was observed only for Norwegian (83%), Romanian (87%), and Swedish (89%). Many inconsistencies in categorization hinged on subtle points; for example, one list of Norwegian keywords appears to be entirely in English, except for the one word ‘Internet’, which might be interpreted either as a Norwegian word or as a misspelling of the English word ‘Internet’. In some other cases, it was clear that the researcher had in fact erred and the assistant’s alternative choice was the correct one.

The results can be seen in Figure 3. Frequency of English-language keywords can to some extent be accounted for by the presence of English-language pages in all of the sets. On the other hand, other factors are clearly at work. In the Greek set, of the 29 pages with English keywords for which the page language was also determined, 19 pages were non-English, five were mixed, two indeterminate, and only two in English. The Chinese sets gave the following results: .tw, out of 13, six non-English, two mixed, one not retrieved, and four English; .cn, out of three, two non-English and one English; .hk simplified, out of six, four non-English, one mixed, and one English; .hk traditional, out of 13, seven non-English, three mixed, one indeterminate, and two English. For Indonesian, the results were, out of seven, four non-English, one not retrieved, and two English.

The broad observation is that most authors of pages in languages other than English prefer not to provide keywords in English only. Provision of keywords in both English and the page language, however, is quite common, especially for Greek and Turkish pages. By comparison, the earlier study had shown descriptions to be somewhat more predominantly English-only and somewhat less commonly mixed.

3.4. Character coding

Most major languages other than English, even when they employ the Roman alphabet, also include at least some diacritics or extra characters (Indonesian is an exception here, and Dutch requires diacritics only on occasional words). Diacritics and extra characters can pose problems in searching, especially if they are treated in different ways on different sites. In a study of five search engines for retrieving pages in Polish, Sroka [36] noted only one that correctly retrieved all occurrences of Polish words containing special characters or diacritics regardless of how they had been coded on the pages.

In the previous study, the practice for meta tagged descriptions in Western European languages appeared to be generally to employ the Latin-1 extension to ASCII. There were a few odd exceptions involving the use of HTML character entity references (such as &eacute;), e instead of the umlaut in German or the apostrophe instead of the grave accent in Italian, and even, in one case, a backslash followed by a three-digit number for an accented character.
Based on a sample of the first 100 keyword lists in each set, the general conclusion about character coding from the previous study seemed to be borne out also for keywords. A major exception was German, where HTML character entity references were found with equal frequency to Latin-1 extension characters. More than 10% of keyword sets used character entity references in Brazilian Portuguese, Finnish, Norwegian, and Swedish. Dutch keyword lists almost never required diacritics for Dutch words, but did use them occasionally for words from other languages; only once was a character entity reference used for this purpose. Romanian keyword lists also rarely marked diacritics.

Occasionally, both a coded and an uncoded form were given for the same word (for example, both ‘Attualità’ and ‘Attualita’; ‘Società’ and ‘Società’ at http://www.ansa.it/).

In a small number of cases, coding was mixed (‘bøker’ with ‘p&aring’ at http://www.bokkilden.no; ‘&ouml;vartning’ and ‘&locacute;rvaltning’ with ‘&ouml;rvaltning’, ‘&ouml;vartning’ with ‘&ouml;vartning’, ‘&ouml;koppsamordn&ouml;ning’ at http://www.statskontoret.se; ‘&ouml;janslar&yacute;’ and ‘haritas&yacute;’ with ‘kültür’ and ‘&ouml;bl&ouml;m’ at http://www.istanbul.com.tr/).

In the one instance where a Polish page used character entity references, these appeared not to be for Polish characters (‘podkr&ecirc;canie, podkr&ecirc;ca&eacute;' at http://www.tweak.pl/, apparently for ‘podkrćcanie, podkrćcać’, ‘turning, turn’). Two similar examples were found in Turkish, with &thorn; being used, not for  ş, but for a hooked lowercase S (ş), and &yacute;, not for  y, but for the undotted lowercase I (ı).

Major browsers always either render these character entities as their corresponding Western European forms, or, especially if a Central European or Turkish character set is selected, with a question mark.

Use of other coding schemes (notably UTF-8), was extremely unusual in all the languages using the Roman alphabet.

The previous study had concluded that descriptions should be constructed in the appropriate character set, rather than omitting diacritics or using character entity references or non-standard coding schemes. This advice likely remains satisfactory for keywords; but
the results noted, especially for German, suggest that some language communities may find the use of character entity references at least equally acceptable. Thus, Web crawlers should be prepared to recognize both conventions.

4. Titles

4.1. Presence

Not all pages retrieved had titles. Proportions missing titles were highest for Korean (7.3%), Spanish (.es) (7.3%), Chinese (.cn) (5.9%), and Spanish (.mx) (5.1%) and lowest for Russian (2.9%), Russian (2.5%), and Indonesian (2.4%).

4.2. Length

Within each language set, titles were divided into those up to 100 bytes in length and those more than 100 bytes in length, with the latter category being considered to be overly long since they would tend to overflow the caption space in a typical browser window. A chi-square test showed a highly significant difference by language in the proportion of overly long titles (p = 0.00000). The highest proportion was for Russian (9.5%), followed by French (6.1%), Dutch (5.2%), Polish (4.5%), Italian (3.8%), German (3.2%), Spanish (.es) (2.3%), Swedish (2.1%), and Turkish (2.0%), with the rest showing less than 2%.

As for keyword strings, so also for titles, a partial explanation for variation in length may lie in differences between languages and writing systems. Certain practices specific to Web-page creation, however, seem to be more prevalent among the speakers of some languages; for example, a number of the long Russian titles take the form of a brief actual title followed by a long list of what are obviously keywords. Some long titles may be artefacts of bad coding (e.g. inserting an angle bracket within the title, causing the parser to miss the title end tag), but examination of the long Russian titles generally demonstrates that they are intentional; for instance, the longest Russian title (http://www.wisp.ru/), at 2468 bytes, is clearly intentionally coded to include a huge list of keywords, ending with the following in English:

- business, cctv, commercial, complex, confidentiality, control, experience, financial, firm, firms, guard, information, invest, legal, links, market, monitoring, partnerships, penetrate, problems, professional, protection, reliable, risks, russian, safe, secrets, security, services, situation, solve, strategic, support, systems, tactic.

4.3. Language

Full language categorization of titles was again carried out by the research assistant, with the same consistency test being applied as in the case of keywords. Overall, consistency for titles was inadequate, at 72.3%. A high level of consistency (90% or greater) was achieved only for Chinese (.cn), Korean, and Russian. Consistency tended generally to be better for languages using non-Roman character sets, since character set used could be employed as a simple indicator of language. Many titles in the Roman alphabet were too short to categorize clearly, often consisting only of a corporate name or a host and domain name. In the most extreme case, 48.6% of the Danish pages were considered by the assistant to have titles whose language could not be determined.

Because of low reliability of categorization of titles where the target language uses the Roman alphabet, Figure 4 shows results only for languages that use non-Roman character sets. Even here, the proportion of titles where the language is indeterminate is often quite high, because the languages still may use the Roman alphabet for foreign company and trade names.

The proportion of mixed-language titles was low throughout; the highest values were for Chinese (.hk traditional) (11.0%), Chinese (.tw) (11.7%), Greek (7.4%), and Korean (6.7%). Mixed-language titles were actually very infrequent in the Russian set, in spite of the relative frequency there of overlong titles. It appears that page creators would consider the title, for the most part, to be too brief an element to accommodate multiple languages.

4.4. Character coding

In the titles in the present study, HTML character entity references seemed to be notably more common. Indeed, based on a sample of the first 100 items in each set, character entity references actually predominated in Mexican Spanish and Norwegian titles, and ran about equal to Latin-1 in French, German, and Italian titles. They formed a substantial minority in titles in Portuguese from both Portugal and Brazil, Spanish from Spain, and Swedish. Dutch (as expected) very rarely used special characters or diacritics; the same was also true of Romanian.
Polish and Turkish pages again overwhelmingly preferred their own language extensions rather than character entity references. One reason for this preference may be simply that some characters in those languages lack widely recognised HTML character entity codes; for example, the Turkish code &amp;#304; (used for the dotted capital İ, İ) is not found in the W3C’s list of character entity references in HTML 4 (http://www.w3.org/TR/REC-html40/entitie-s.html), even though recent versions of many browsers can handle it (even Lynx recognises it as a form of the letter İ). As further evidence of the difficulties that these language communities experience with character entity references, we may cite the mis-coding in keywords noted above.

Interesting cases are presented by titles in which the same word or phrase is repeated, once in Latin-1 and once using character references (‘guía’ and ‘gu&iacute;a’ at http://www.lycos.es/; ‘Universitä’ and ‘Universi-t&ouml;gt’ at http://www.unipv.it/; ‘T&iacute;ecnico’ and ‘T&eacute;cnico’ at http://www.ist.utl.pt/; ‘Vetenskapsr&amp;#229;det’ and ‘Vetenskapsr&amp;#229;det’ at http://www.vr.se/). Mixing the two coding schemes without repetition was also observed once (‘F%C3%B6rsta’ and ‘t&amp;aring;rg’ at http://www.expressen.se/). Use of other coding schemes (notably UTF-8), was again extremely unusual in all the languages using the Roman alphabet.

It is probably satisfactory general advice to tell page creators to construct titles in the appropriate character set. The results for titles given here, however, do suggest that many Western-European language communities may find the use of character entity references at least equally acceptable and that Web crawlers should be prepared to recognise both conventions.
5. Conclusion

5.1. Further research

As was already noted in the earlier study, other sampling methods might be considered in future research to overcome certain limitations of the approach taken here. A more aggressive procedure for eliminating duplicates, such as checking to see whether two host names share a common IP number, might produce cleaner results. A different search engine or more specific search specifications might be employed in order to obtain better representation of more obscure pages.

Additional languages could be included, but this action would probably be called for only in order to answer specific questions. Practices in other languages that require additional special characters might be examined. Do they tend to prefer local ASCII extensions, HTML character entity references, another scheme such as UTF-8, or a variety of different schemes?

Results showed significant differences in proportion of pages with keywords depending on language; specifically, pages in Western European languages showed higher proportions with keywords, while pages in Chinese showed the lowest proportions. It would be interesting to see whether this difference changes over time, say as support for viewing and searching using the Chinese writing system becomes more widespread.

Short titles especially proved difficult to assign reliably to language categories. A future study, therefore, might attempt to compile larger sets of pages with longer titles for categorization. Since most titles are short, however, the generalizability of results from such a study might be limited.

An alternative, automated, method of assigning language categories to keywords and other page elements could involve use of one or more spell checkers. Using a single English spell checker, page elements could be categorized by the extent to which they used common, correctly spelled English words, which could form a rough measure of the extent to which they were in English. Adding spell checkers in other languages would provide an additional filtering level; a page element that scored low on both English and the target language might be in a third language or contain many trade names or technical terms.

The variation in coding of diacritics and special characters for Roman-alphabet languages suggests yet another line of research that would combine elements of the work of Sroka [36] and that of Turner and Brackbill [16]: testing the extent to which different search engines deal appropriately with pages using different coding schemes. A corollary study might look at whether repeating the same keyword in plain Roman, using character entity references, and in Latin-1 improved retrievability, or actually degraded it because the repetition was detected and deemed to be word stuffing by the Web crawler [32].

5.2. Other implications

The difference in lengths of titles depending on language remains unexplained. As already noted, some languages do take more characters to say the same thing than other languages, but other factors seem also to be at work, as exemplified by the particularly long titles found on some Russian pages. Excessively long lists of keywords are rejected or truncated by some Web crawlers that otherwise make use of keyword lists; perhaps crawlers apply, or should apply, somewhat different standards for maximum length for different languages.

Though keywords were mostly in the languages of the pages, English keywords were often provided in addition to keywords in other languages. Page creators seem to use keywords for the purpose of overcoming the language barrier more often than they use descriptions; possible explanations include the greater ease of translating individual words and phrases in comparison to sentences and a more specific perception of keywords as something that will be picked up by search engines.

Robots scanning non-English-language Web pages for titles or similar elements, whether for research or as input to search engines, should generally be capable of dealing with both appropriate ASCII extensions and HTML character entity references, and possibly with other variations such as UTF-8. Creators of Web pages should be advised to construct titles, keywords, and descriptions in the appropriate character set for the language, or, alternatively, use correct character entity references.

Acknowledgements

Research reported in this article was supported in part by the University of Western Ontario Office of Research Services with funds provided by the Natural Sciences and Engineering Research Council of Canada.
The extensive contribution of research assistant Craig Huffman in data gathering and categorization is also acknowledged.

References


