

Exploring the Capabilities of English and Hungarian Search Engines for Various Queries

ERZSÉBET TÓTH

Library and Applied Information Science Group, Institute of Mathematics and Informatics,
College of Nyíregyháza, Nyíregyháza, Hungary

This paper presents a test that examined the linguistic capabilities of English and Hungarian search engines. Three English-language search engines were tested: Google, AltaVista and AlltheWeb. For comparison, five significant Hungarian search engines were considered: Heuréka, Origo-Vizsla, Kurzor, Góliát and Altavizsla. The analysis was based on the following aspects: stemming, handling of stopwords and diacritics, truncation and searching for synonyms. The results

indicate that while the Hungarian search engines are motivated to take into consideration the specific characteristics of the Hungarian language, on the whole the English-language search engines ignore the special characteristics of the Hungarian language. In the cases examined none of the general search engines handled diacritics well; that caused their resulting weaknesses in finding information relevant for Hungarian users.

Aim and motivation of the test

In this paper are reported the findings of a test relating to Bar-Ilan and Gutman's analysis conducted in November 2002. Primarily they examined from a morphological aspect the interpretation of queries in Russian, French, Hungarian and Jewish languages submitted to search engines based in English-speaking countries (referred to in the rest of this paper as 'English search engines') and local search engines. During their research they concentrated on how search engines took into account the specific characteristics of the selected languages and how effective they were in responding to non-English queries. In their findings they published the number of hits retrieved by the search engines along with the submitted queries (Bar-Ilan and Gutman 2005).

On the basis of this preliminary literature, a test was conducted between July and September 2005 to explore the linguistic capabilities of English and Hungarian search engines. English search engines were analysed on queries in both English and

Hungarian. However Hungarian search engines were examined only on the Hungarian queries because, except for Origo-Vizsla and Altavizsla, they mainly index only Hungarian Web pages. The main focus was on whether these search engines made any efforts to interpret these queries more precisely or whether this would remain a task to be solved further on in the search process. The findings of this test have supported the statements of Bar-Ilan and Gutman, that Hungarian search engines take into consideration to some extent the specific characteristics of the Hungarian language. In contrast the English search engines show serious inadequacies in this field because they search for exact word forms only and apply a simple pattern matching to queries.

Background information

Hungarian belongs to the Finno-Ugric branch of the Uralic languages. Its Latin alphabet contains fourteen vowels altogether. Among them, five pairs of vowels are just short and long counterparts of

the same sound. There are two other pairs (e-é, a-á) where each character represents a different sound. In Hungarian the use of diacritics is very special because the meaning of the same word form is completely different when omitting them. Two examples in the test reflected this case. *Kertem* means “my garden”, but if we use a diacritic, *kértem* has a different meaning “I asked”. In another case *alma* means “apple” but its diacritical version, *álma* means “his dream”. So it is an important user requirement for the search engines to take into consideration the exact form of the term, or otherwise a lot of irrelevant hits will be found.

In Hungarian articles do not change according to number, person, gender and case. The indefinite article is *egy*, while the definite article has two forms *a* and *az*. The first form is used before consonants and the second one before vowels, which is similar to the use of the indefinite articles *a* and *an* in English.

There is a complex case system in Hungarian including 16 to 24 distinct forms (depending on assumptions about the exact number of case suffixes). Suffixes represent cases and they always come after the suffix for plural and possession. Thus three inflectional suffixes may be connected to the word stem. The most common problem for stemming is that these suffixes may change the basic form of the word e.g. *bagoly-baglyok* (=owl-owls), *sátor-sátram* (=tent-my tent), *kő-kövek* (=stone-stones). Furthermore, it is difficult for the search engines to recognize the related word forms and to handle them together. Automatic truncation algorithms also have to cope with the problem of finding the suffixed stem forms of the query, which seems to be a difficult requirement to fulfil. If search engines fail to retrieve these stem forms, users can easily miss relevant Web pages concerning the topic. We can see this problem later through the example of the *májmetely* (=liver fluke) query.

In verb conjugation, personal suffixes are added to the stem in all tenses (e.g. *írom, írod, írta*=I write, you write, he wrote). Verbal particles appear as prefixes. Their function is to mark direction and aspect, and to make verbs transitive (e.g. *lemegy*=go down, *kimegy*=go out, *megeszi*=eat up). In several cases verbs have different meanings depending on which particle is attached to them (e.g. *megrendez* – arrange, *berendez* – furnish, *átrendez* – rearrange). These prefixes follow the base verb in the negative and imperative. Nouns can be created from most

verbs with or without verbal particles (e.g. *lát* – *látás* (=see – sight), *kilát* – *kilátás* (see out – view). Verbal particles express different meanings and the nouns created from them have the same (or nearly the same) meaning (e.g. *nézés* – *megnézés*, looking; *számolás* – *kiszámolás*; calculation). It would be useful for the user to find both forms of the same noun, but in general he only needs to retrieve the word with its prefix (Megyesi 1998; Bar-Ilan and Gutman 2005).

Methodology

A set of search terms was carefully constructed for testing. Then trial searches were run on each search engine to check if the results from the selected terms would reflect clearly the issue analysed and if they would correspond to the research goals. We selected those search terms that emphasized the linguistic difficulties of these languages and we also relied on our previous observations. Proper names were not chosen, since they do not change their morphological form in free text. We reviewed the help files of each search engine to have more information about their relevant features and capabilities. The first 100 hits were examined on the queries. The only exception was the analysis of truncation in Hungarian where a disease name called *májmetely* (=liver fluke) was entered as a query that provided a limited set of hits. In the case of truncation we recorded the number of results retrieved by the search engines. Primarily the abstracts of the results were used for analysis, but the contents of the Web pages were also checked where necessary. The searches were executed between July and September 2005. We did not evaluate the relevance of the results because we focused on the linguistic capabilities of the search engines.

Altogether three English search engines were analysed: Google [1], AltaVista [2] and AlltheWeb. [3] These leading general search engines were selected because they enabled users to search for Hungarian Web pages. In addition to this, Google provides a local version for Hungarian users. [4] As a test case, five significant Hungarian search engines were considered: Heuréka [5], Origo-Vizsla [6], Kurzor [7], Góliát [8], and Altavizsla. [9] (see Figures 1–5). Among them only in Altavizsla was help documentation not available. In the selection of the search engines an important requirement

Figure 1. Hungarian search engine, Heuréka (<http://www.heureka.hu>).

was that they should be stable and retrieve an appropriate number of hits to the queries within a short response time.

Aspects for the analysis

For testing, the following aspects served as a basis: stemming, handling of stopwords and diacritics, truncation and searching for synonyms. Stemming was examined to determine if the search engine retrieved the plural form of a search phrase or not, i.e. was it able to recognize the plural form of a query. The queries entered were as follows: *dog-dogs* (in English), *ház-házak* (=house-houses), *kocsi-kocsik* (=car-cars), *kutya-kutyák* (=dog-dogs). In the first two Hungarian examples the final vowel of the stem does not change in the plural form compared to the singular form. However, in the third Hungarian example it does change in the plural

form compared to the singular form. A noun phrase with a privative suffix was deliberately chosen for studying stemming in Hungarian, which was the *tisztességtelen* (=dishonest) phrase. Using this phrase we could monitor if the search engine retrieved any other word forms of this complex phrase and applied stemming to it.

In the case of stopwords, whether the search phrase appeared with definite and indefinite articles in the results was checked to determine whether the search engine searched separately for the articles entered or not. The outcome of this question was obvious because in most cases the search terms with the article were highlighted in the abstracts when the search included the articles. In the other case, when the articles were excluded from the searches, only the search term alone was highlighted from the abstracts. Thus the following queries were analysed in English: *a*

Figure 2. Hungarian search engine, Origo-Vizsla (http://www.origo.hu).

The screenshot shows the Origo-Vizsla search engine homepage. At the top, there's a search bar with the text "[vizsla24] : keresés az interneten" and a "KERESÉS" button. Below the search bar, there are several sections: "Internet-előfizetés", "freemail" login, and a "Piros lett az Andrássy út" news article. The page also features a sidebar with navigation links like "Itthon", "Választás 2006", "Nagyvilág", "Üzleti Negyed", "Sport", "Szórakozás", "Zene", "Autó", "Ingatlan", "Techbázis", "Tudomány", "Női Lapozó", "Babázó", "Egészség", "Filmklub", "Szex", "Otthon - design", "Programajánló", "Időjárás", "Szolgáltatások", "Fórum", "Klikkbank", "Utazás", "Állás", "Apróhirdetés", "Jármű apró", "Ingatlan apró", "Szoftverbázis", "Tévéműsor", "Moziműsor", "Vásárlás", "Biztosítás", "Chat.hu", "Társkereső", "Fotóidőkezelés", "Tudakozók", and "[o]-mobil".

dog, an aunt and the car. Concerning Hungarian the following terms were entered: *a ház* (=the house), *az ember* (=the man), *egy kocsi* (=a car). Regarding stopwords, frequent question words and numerals were not considered.

Diacritics were only taken into account in Hungarian since it was not a relevant aspect to analyse in English. Here two search phrases were used, namely *kertem* (=my garden) and *alma* (=apple). The question to be addressed here was whether the search engine found the diacritical versions of these queries among its hits or not.

Among Hungarian search engines only *Heuréka* [10] claimed to be able to implement truncation. In the help documentation of the other Hungarian search engines there was no information about truncation. For truncation an asterisk (*) was utilized after the search term in each case. In English the query *Olympi** was entered to find all those

Web pages that are about the Olympic Games containing the following phrases: Olympic, Olympics, Olympia, Olympian, etc.

In Hungarian a limited set of hits was chosen to find more easily the suffixed stem forms of the query and to check their real appearance within the set of hits. By entering the query *májmetely** the suffixed stem forms of *májmetely* (=liver fluke) disease were retrieved. In addition the sets of hits retrieved by the *májmetely** query through all search engines were examined to check if they really included the suffixed stem forms of the query. The final result was that these sets of hits did not contain any suffixed stem forms of the query. Thus it was necessary to look for the suffixed stem forms of this query one by one, manually, in order to check if there was any stem form, which should have been retrieved by search engines but had not been found.

Figure 3. Hungarian search engine, Kurzor (<http://www.kurzor.hu>).

Finally, whether the search engines retrieved the synonyms of a query or not was tested. How the synonyms were visible in the hits was also studied, for example if they were highlighted from the abstracts, if they appeared together with the query in the abstracts, or if they were available alone. The following English queries were examined from this aspect: *car*, *glasses*. In Hungarian the synonyms of *kutya* (=dog) and *vetélkedő* (=contest) were searched for.

During testing an error was found in AltaVista and AlltheWeb. When the search was limited exclusively to Hungarian Web pages, then English, Spanish and French pages also appeared among the first 100 hits on the *alma* (=apple) query. It appears that the language filter on results did not work properly in these two search engines. The same error did not occur in the case of Google and the Hungarian search engines.

Language solutions found in English search engines

In the following section the linguistic capabilities of English and Hungarian search engines will be summarized in light of the results from searching to test the aspects discussed earlier in this paper.

Google and AlltheWeb did not apply stemming to English and Hungarian queries. None of the English search engines carried out stemming to the *tisztességtelen* (=dishonest) query. Rather, they just retrieved the exact word form of the search term. Stemming in English operated appropriately in AltaVista, so it recognized automatically the plural form of the *dog* phrase. However stemming in Hungarian did not work at all in AltaVista.

Google seemed to be the best in handling of stop-words properly as compared to the other search

Figure 4. Hungarian search engine, Góliát (<http://www.goliat.hu>).

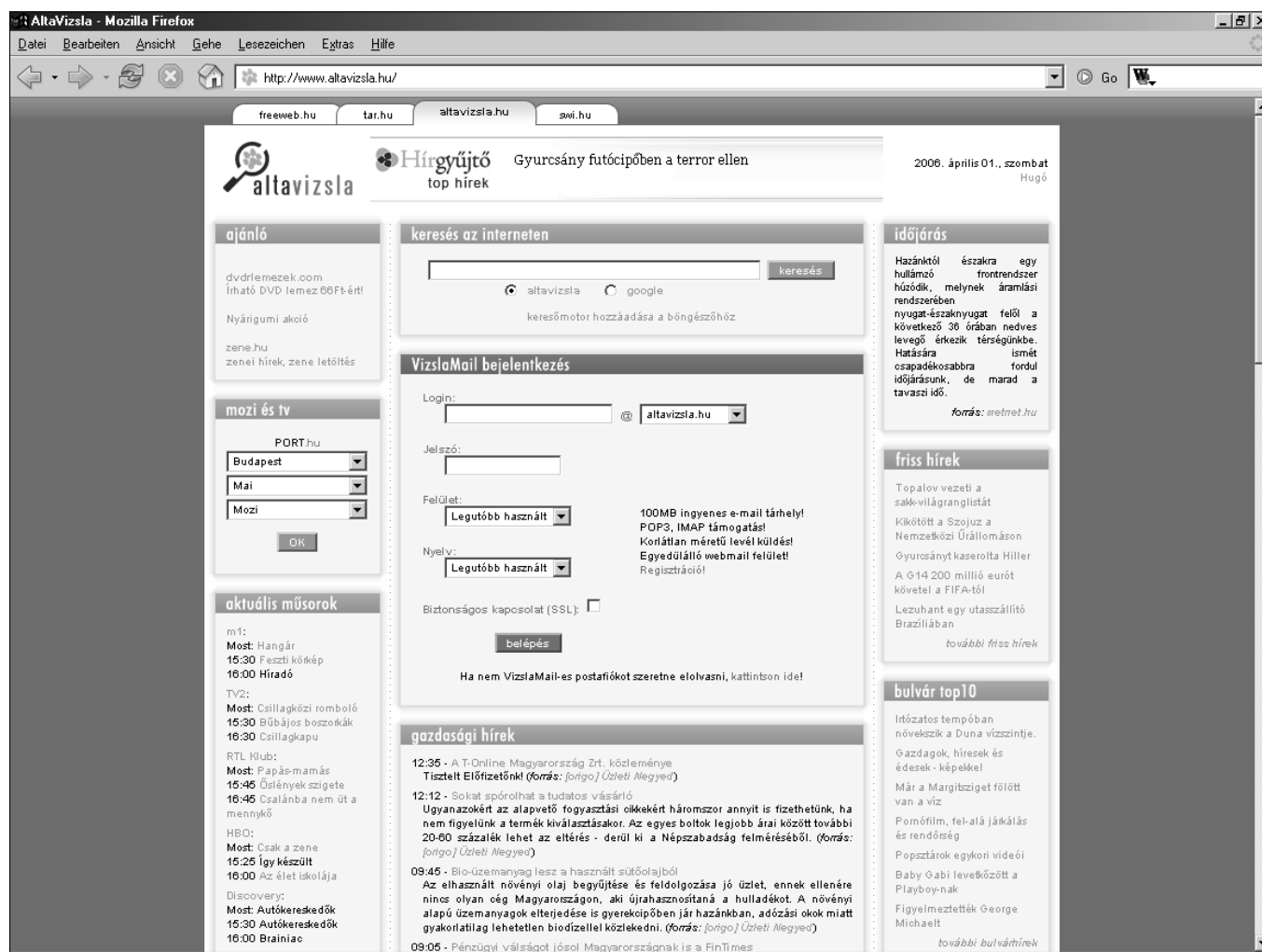
engines because it omitted the articles *a* and *an* from queries in English. Besides this, it excluded the definite article, but it did not exclude articles *az* or *egy* from queries in Hungarian. AltaVista and AlltheWeb did not omit definite and indefinite articles on queries in English or Hungarian. So they are evidently deficient in this respect.

Diacritics were not handled properly in the English search engines because they retrieved Web pages with *kértem* (=I asked) in response to the query *kertem* (=my garden), and Web pages with *álma* (=his/her dream) in response to the query *alma* (=apple).

All of them applied truncation efficiently to English queries. On the one hand they found those longer phrases where the *Olympi** term appeared. On the other hand they retrieved those Web pages where this truncated term appeared together with its other longer versions. Google found the *Olym-*

*pi** term in the following longer phrases: *Olympic*, *Olympia*, *Olympiad*. Among its hits the *Olympi** term appeared together with the following longer versions: *Olympics*, *Olympian*. AltaVista retrieved the *Olympi** term in the following longer phrases: *Olympic*, *Olympics*, *Olympia*, *Olympiad*, *Olympian*. In its hits the *Olympi** term appeared together with the following longer versions: *Olympia*, *Olympic*, *Olympics*. AlltheWeb explored the *Olympi** term in the following longer phrases: *Olympic*, *Olympics*, *Olympiad*, *Olympia*, *Olympian*. In its results the *Olympi** term appeared together with the following longer versions: *Olympic*, *Olympics*, *Olympiad*, *Olympia*, *Olympian*.

Truncation in Hungarian did not work at all in these English search engines; because it could be realized in each case that the sets of hits retrieved to the query *májmétely** did not comprise the suffixed stem forms of the query, it was necessary to

Figure 5. Hungarian search engine, Altavizsla (<http://www.altavizsla.hu>).

look for them one by one, manually. The following suffixed stem forms were explored in these search tools: *májmetelynek* (“-nek” is a dative or a genitive case suffix), *májmetelyt* (“-t” is an accusative case suffix), *májmetelyek* (“-k” is a plural suffix), *májmetelyről* (“-ről” is a prepositional suffix). The number of hits retrieved to the truncated and untruncated forms of the query *májmetely* (=liver fluke), the suffixed stem forms along with their number of hits are presented in Table 1.

We can see that only in AltaVista is there a slight difference in the number of results obtained for the truncated and untruncated versions of the query. In Google the number of hits retrieved to the truncated form of the query is fewer than that of its untruncated version. In AlltheWeb the same number of hits was found on the truncated and untruncated forms of the query. These last two cases suggest that truncation does not work at all

Table 1. Number of hits retrieved for the various forms of the *májmetely* search.

| | Google | AltaVista | AlltheWeb |
|---------------------|--------|-----------|-----------|
| Truncation | | | |
| <i>májmetely*</i> | 106 | 100 | 62 |
| <i>májmetely</i> | 140 | 94 | 62 |
| <i>májmetelyek</i> | 8 | 6 | 1 |
| <i>májmetelyt</i> | 6 | 5 | 4 |
| <i>májmetelynek</i> | 1 | 1 | 1 |
| <i>májmetelyről</i> | 3 | | |

because the truncated form of the query should have resulted in a larger set of hits than that of its untruncated form. We can also realize that all of the search engines failed to retrieve a small set of hits concerning the topic. These missed hits were explored by means of the suffixed stem forms.

Each of the search engines searched for the synonyms in English, but in the implementation of

Table 2. Summary of results for the searches 'car' and 'glasses'

| | Google | AltaVista | AlltheWeb |
|---------|------------------------|---------------------|---------------------|
| Results | e.g. <i>~car</i> – | e.g. <i>car</i> – | e.g. <i>car</i> – |
| re- | automobile (US), | automobile (US), | automobile (US), |
| trieved | automotive (US), | automotive (US), | automotive (US), |
| | auto (US), | auto (US), | auto (US), |
| | motor (not a | vehicle (not a | vehicle (not a |
| | synonym!), | synonym!); | synonym!); |
| | vehicle (not a | e.g. <i>glasses</i> | e.g. <i>glasses</i> |
| | synonym!), | – spectacles, | – eyeglasses, |
| | racing (not a | eyeglasses, | glassware, |
| | synonym!); | glassware, | sun-glasses (not |
| | e.g. <i>~glasses</i> – | sun-glasses (not | a synonym!), |
| | eyeglasses, | a synonym!), | reading- |
| | glassware, | reading- | glasses (not a |
| | sun-glasses (not | glasses (not a | synonym!), |
| | a synonym!), | synonym!), | goggles (not a |
| | goggles (not a | goggles (not a | synonym!). |
| | synonym!). | synonym!). | |

this function there were mistakes deriving from the wrong interpretations of the language. Table 2 summarizes the search results retrieved by the search engines to the queries *car* and *glasses*.

The following synonyms of the 'car' term appeared in hits: *automobile*, *automotive*, *auto*. In addition to this the terms *motor*, *racing* and *vehicle* also appeared as synonyms, but they could not be considered to be synonyms of the term *car* (although *motor* might be used as a colloquialism, and *vehicle* is a broader term). The following synonyms for the 'glasses' search were retrieved: *eyeglasses* and *spectacles*. In addition to this, the search engines also retrieved the terms *sunglasses*, *reading-glasses* and *goggles*. These were rather types of glasses and could not be regarded as real synonyms for the term *glasses*. Each search engine retrieved *glassware*, meaning "articles made of glass," as a synonym of the singular form of the query *glasses*. However, the singular form of the term *glass* usually means: "a hard brittle, usually transparent substance", "a drinking vessel made of glass", or "vessels and articles made of glass".

We can see from the Table 2 that all of the search engines retrieved three American synonyms for the 'car' query. However Google explored three terms that were not regarded as synonyms. AltaVista and Alltheweb found one phrase separately that might not be considered as synonym. In the other case AltaVista explored two synonyms in response to the 'glasses' phrase and three terms that could not be regarded as synonyms besides the

glassware. Google and Alltheweb retrieved term only one synonym in addition to the term *glassware*. Google found two phrases that were not considered as synonyms. AlltheWeb explored three terms might not be regarded as synonyms.

In Google we can look for the synonyms of a search phrase by using a tilde character (~) before the term. Here synonyms were highlighted from the abstracts and they mostly appeared together with the query. In contrast to Google, in AltaVista and AlltheWeb, synonyms were not highlighted from the abstracts, and they always appeared together with the search phrase. There was a search for the synonyms of the terms *kutya* (=dog) and *vetélkedő* (=contest) in the search engines, but none of them found valuable synonyms for the Hungarian queries. These search services just searched for the terms in their exact form and did not consider their Hungarian synonyms.

Language solutions found in Hungarian search engines

Among the Hungarian search engines we can only suppose that Heuréka applied stemming to queries because it retrieved the following word forms in response to the query *tisztességtelen* (=dishonest): *tisztesség* (=honesty), *tisztességes* (=honest), *tisztességért* (=for the honesty), *tisztességgel* (with the honesty), *tisztességtelen* (=dishonest), *tisztességtelenül* (=dishonestly). In Heuréka, the plural form of the search terms *ház* (=house), *kocsi* (=car), *kutya* (=dog) always appeared together with its singular form in the abstracts. The search phrases were not highlighted in the abstracts, so it was difficult to decide if Heuréka searched for the plural form of the query or not. Stemming did not operate in the case of the other search engines, which indicates that they did not search for the plural form of the query. Besides this none of them applied stemming to the query *tisztességtelen* (=dishonest) and they looked for the exact form only.

The handling of stopwords shows a quite divided picture among the search engines because Origo-Vizsla and Kurzor do not exclude definite and indefinite articles from searches. We cannot determine clearly whether definite and indefinite articles are omitted by Heuréka because the search terms with the articles were not highlighted from the abstracts. It could be also observed in the abstracts that the search terms appeared many times

Table 3. Number of hits retrieved for the truncated and untruncated form of *májmetely*.

| | Heuréka | Origo-Vizsla | Kurzor | Góliát | Alta-vizsla |
|--------------|---------|--------------|--------|--------|-------------|
| Truncation | | | | | |
| májmetely* | 0 | 35 | 30 | 0 | 0 |
| májmetely | 36 | 35 | 30 | 30 | 30 |
| májmetelyek | 1 | 3 | 1 | | |
| májmetelyt | | 1 | 2 | | |
| májmetelynek | | 2 | | | |

without the articles given in the queries. From this we could conclude that Heuréka tried to omit articles in the searches. Góliát and Altavizsla omitted the definite articles *a* and *az*, but they did not omit the indefinite article *egy*. The problem of highlighting search terms and articles emerged only in Heuréka, but it was not a problem in the case of the other search engines.

Regarding diacritics we can say that the majority of the Hungarian search services have successfully overcome the challenge. The only exception is Origo-Vizsla because it cannot handle diacritics appropriately. Thus it retrieved Web pages with *kértem* (=I asked) in response to the query *kertem* (=my garden) and Web pages with *álma* (=his/her dream) in response to the *alma* (=apple) query. In Heuréka there are two types of search options. One of them is the exact word form option that provides the proper handling of diacritics. However this functionality does not work at all in the case of the other option, when the system is allowed to add diacritics automatically. Kurzor, Góliát and Altavizsla handled diacritics efficiently.

None of the Hungarian search engines applied truncation to queries. During testing it happened several times that there were no hits retrieved to the truncated form of the query *májmetely* (=liver fluke). The same negative result was received in Heuréka, Góliát and Altavizsla. A reasonable explanation might be that the search engine was not able to interpret truncation as a retrieval operation. Here was another case when the same number of hits was retrieved for the truncated and untruncated forms of the query. It reflects the fact that truncation did not work at all, because as a default the truncated form of the query resulted in a larger set of hits than that of its untruncated form. This was the case in Origo-Vizsla and Kurzor. Table 3 shows detailed information about the number of hits retrieved to the truncated and untruncated forms of the query *májmetely* (=liver fluke) and the

suffixed stem forms along with their number of hits.

In Góliát and Altavizsla, we did not retrieve any hits to the suffixed stem forms of the query. These services are likely not to index those Web pages that would be relevant to this type of query. Heuréka, Origo-Vizsla and Kurzor missed only a few hits relating to the topic.

Origo-Vizsla and Heuréka were successful in finding synonyms for the queries entered. The other search engines failed to search for the synonyms of a query. In these two search engines the following synonyms were retrieved by the query *kutya* (=dog): *eb* (=dog), *öleb* (=lap dog). The following synonyms were searched in response to the query *vetélkedő* (=contest): *kvíz* (=quiz), *kvízzjáték* (=quiz show/game), *verseny* (=contest), *agytorna* (=mental exercise). In these search engines, the synonyms were not highlighted in the abstracts, thus making this type of response to queries less obvious. In most cases the synonyms appeared together with the query in the abstracts and they were sometimes available alone. Among Hungarian search engines only Heuréka has a built-in thesaurus that enables further searches to be conducted relating to a search topic.

Conclusions

Results of this study indicate that English search engines handle queries for English terms much better than they handle Hungarian terms. In these search tools truncation and searching for synonyms works properly in English but it is problematic in Hungarian. They show inadequacies in the same fields, for example in not handling diacritics well, which is an important user requirement regarding queries for Hungarian terms. This deficiency reflects their weaknesses in finding relevant information for Hungarian users. Google omits definite and indefinite articles in the Eng-

lish queries, but it has not solved this issue yet for Hungarian queries. AltaVista is good at finding the plural form of the English queries, so this is the only service where stemming works properly in English. However none of the English services coped with this problem in the Hungarian language. We can say that Google and AltaVista are equal in their performance of interpreting queries, followed by AlltheWeb. On the basis of these findings, we can conclude that more emphasis should be placed on stemming, handling of stopwords and diacritics in the future development of these search engines.

The majority of the Hungarian search engines handle diacritics efficiently, and in this way they meet an essential user requirement. If we evaluate the linguistic capabilities of Hungarian search engines we can say that Heuréka provided the best performance in fields such as stemming and searching for synonyms. There were two other fields where its performance was acceptable, namely handling of stopwords and diacritics. The next best performance was achieved by Góliát and Altavizsla, which showed very similar performance in the tests. They handled diacritics precisely, but they need to improve their functionality in dealing with stopwords because they are not yet perfect. In this ranking Origo-Vizsla and Kurzor are last with a similar level of performance. Origo-Vizsla shows inadequacies in several fields such as stemming, truncation, handling of diacritics and stopwords. Kurzor has deficiencies in the fol-

lowing fields: stemming, handling of stopwords, truncation and searching for synonyms.

The results of these tests have established that almost every Hungarian search tool has to improve its performance in the field of truncation and stemming, whilst the English search engines need to consider how best to search foreign language Web sites.

Notes

1. Google - URL: <http://www.google.com>
2. AltaVista - URL: <http://www.altavista.com>
3. AlltheWeb - URL: <http://www.alltheweb.com>
4. Google - Hungarian version URL: <http://www.google.com/intl/hu/>
5. Heuréka - URL: <http://www.heureka.hu>
6. Origo-Vizsla - URL <http://www.origo.hu>
7. Kurzor - URL <http://www.kurzor.hu>
8. Góliát - URL: <http://www.goliat.hu>
9. Altavizsla - URL: <http://www.altavizsla.hu/>
10. Heuréka - Help URL: <http://www.heureka.hu>

References

- Bar-Ilan, J. and T. Gutman. 2005. How do search engines respond to some non-English queries? *Journal of Information Science* 31(1): 13–28.
- Megyesi, B. 1998. Brill's rule based part of speech tagger for Hungarian. In: Master's course in computational linguistics. Computational Linguistics Department of Linguistics, Stockholm University URL: <http://stp.ling.uu.se/~bea/Duppsats.pdf> [viewed February 1, 2006]

Editorial history:

paper received December 2005;

final version received 21 February 2006;

accepted 27 February 2006