How do search engines handle non-English queries? - A case study

Judit Bar-Ilan
The Hebrew University of Jerusalem
P.O, Box 1255
Jerusalem, 91904 Israel
972-2-6584663
judit@cc.huji.ac.il

Tatyana Gutman
Ex Libris Ltd
Malha Technology Park
Jerusalem, 91481, Israel
972-2-6798222 ext.184
tania.gutman@exlibris.co.il

ABSTRACT
In this paper, we explore the capabilities of search engines for non-English languages. As a test case, we examine four languages: Russian, French, Hungarian and Hebrew. For each of these languages we test three general search engines: AltaVista, FAST and Google and some local search engines. Our results indicate that in most cases the general search engines ignore the special characteristics of non-English languages, and sometimes they do not even handle diacritics well. These findings are rather disturbing, since for example Google is very popular in non-English speaking countries as well, and users are either not aware of what they miss when using search tools that do not take into account the structure and the special characteristics of the specific language or have no alternatives but to use these search engines.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval

General Terms
Measurement, Performance, Human Factors, Languages.

Keywords
Morphological analysis, diacritics, non-English languages, search engines, inflections, pre and postfixes

1. INTRODUCTION
The World Wide Web has become a major communication channel and information source. Even though English is the lingua franca of the Web, the use of other languages is clearly non-negligible. Data provided by Global Reach [10] indicate that only 36.5% of the Internet users are English speakers (as native/primary language). Among the non-English speaker (as native/primary tongue) users, 3.5% are reported as French speakers and 2.9% as Russian speakers, and the other two languages examined by us are spoken by less than 1% of Internet users. Other widely spoken languages are Chinese (10.9%), Japanese (9.7%), Spanish (7.2%), German (6.7%), Korean (4.5%), Italian (3.8%), Portuguese (2.9%) and Dutch (2%). On the other hand, when considering the distribution of Web pages by language the data is considerably different. Cyber Atlas [8] reports based on data from Vilaweb, that in the year 2000, 65.6% of the Web pages were written in English, 2.96% in French, 1.88% in Russian, 0.16% in Hungarian and 0.06% in Hebrew. Beside English the most popular languages on the Web were: Japanese (5.85%), German (5.77%) and Chinese (3.87%), followed by French (2.96%), Spanish (2.42%) and Russian (1.88%). The data provided by OCLC [30] showed that in the summer of 2001, 73% of the pages were written in English, 3% in French, and 1% in Russian (no data for Hungarian and Hebrew). The most popular languages after English were German (7%), Japanese (5%), followed by French, Spanish (3%) and Chinese (2%).

There are several possible explanations for the differences between the data on users and on Web pages: 1) A large number of pages created by non-English speakers are written in English: for non-English language sites it is customary to create an English language version as well, probably in order to be more widely accessible (as we said before English is the lingua franca of the Web). This claim is supported by Nunberg's findings [27]. 2) Not all users create Web pages and it may be that specific difficulties arise when creating pages in non-English languages (this is definitely true for Hebrew, which uses a non-Latin character set and is written from right to left). 3) There is a large number of multi-lingual pages, thus the percentages of pages by languages, do not add up to 100% (they exceed 100%). 4) Web statistics have to be handled with care, as the NUA Web site [26] states: "The art of estimating how many are online throughout the world is an inexact one at best." Data for the different statistics were collected and analyzed using diverse sampling techniques and methodologies, and because of the fast growth and the dynamic nature of the Web, data from the year 2000 is hardly comparable with data to 2002. See [18] for some methods to estimate English and non-English language use on the Web.

Even if the above numbers are not exact, it is clear that non-English language pages and users cannot be ignored. How can users looking for non-English information search the Web? They can access search engines and directories that cover countries or specific geographic regions or they can use the general search engines to search in their languages. In October 2002, among the leading general search engines, Google allowed users to restrict their searches to one of 35 languages, FAST (AlltheWeb) to one of 49 languages, AltaVista to one of 25 languages, MSNsearch

WWW 2003, May 20-24, 2003, Budapest, Hungary.

(Inktomi) to one of 15 languages and WiseNut to one of 25 languages (no language restriction for Teoma). Thus it is clear that the major search engines intend to provide answers to non-English language queries. Some of the search engines have local versions, for example Google had 34 local sites [11] and Google, AltaVista, FAST and Inktomi powered local search engines as well (sometimes based on their global index and sometimes on a local one) [37]. Google is not only the most successful search engine in the US (in terms of average and total time spent searching per visitor) [36], but the fourth top property (terminology used by Nielsen/NetRatings) in France as of September 2002 [24], the tenth in Israel as of April 2002 [25] in the Nielsen/NetRatings Survey, and according to the TIM/TNS survey, Google was the sixth most popular site in Israel (and the most popular search engine) as of May 2002 [38].

Users present queries, and the search engines retrieve results that match their queries. Sometimes "match" means just pattern matching - the sequence of symbols entered by the user appears somewhere in the document, but sometimes more sophisticated methods, based on morphological analysis of the language, are utilized (e.g. dealing with plurals, tenses, pre and postfixes). The major search engines are naturally geared toward English (since most of the Web pages are written in English).

The question that initiated the current research was, how well do the major search engines, that enable and encourage searches in non-English languages, handle queries in these languages and to what extent do they take into account the specific linguistic characteristics of them (e.g. inflections, diacritics, and prefixes)? We compare the performance of the general search engines to search engines that were specifically created to search in the local languages. Four languages were chosen for our case study: Russian, French, Hungarian and Hebrew. Although the languages were chosen as a convenience sample, they cover a wide range of potential obstacles for the search engines; each language belongs to a different language family (Slavic, Roman, Finno-Ugric and Semitic respectively); French and Hungarian use the Latin alphabet (with the addition of diacritic marks), Russian is written with Cyrillic letters, and Hebrew has its own alphabet, with the language being written from right to left.

In the next section we review the relevant literature and present background information on each language. In section 3 we describe our evaluation methodology, section 4 is dedicated to the results and section 5 concludes and summarizes our findings.

2. BACKGROUND INFORMATION

Our searches revealed that so far only a very small number of studies discussed non-English information retrieval from the Web. Sroka [34] described the capabilities of Web search engines for Polish information retrieval, his review included the general search engines Polski Infoseek and Polska AltaVista and some local search tools. The evaluation was based on ten queries (when the queries included diacritics the searches were conducted both with and without them), and the following criteria were examined: number of records, precision (out of the first ten hits), response time and overlap of records. There was no linguistic examination of the tools, except for the use of diacritics.

Mujoo et al. [23] described the implementation of two search engines for Indian languages. The paper discusses the architecture of these search engines (e.g. indexing, compressing, searching) along with the specific problems of the Indian languages written in Devanagari (several Indian languages use this script, e.g. Hindi and Sanskrit): morphological variants - variants of a word stem, presenting the same concept, but in different tenses or plurality; phonetic tolerance - phonetically equivalent characters, that can be used interchangeably, and font independence - the lack of a standard encoding. Two search engines were built trying to tackle language specific problems.

Some studies discussed the language specific features of non-English languages in the context of information retrieval, but not in the context of the Web (e.g. Arabic [1], Dutch, German and Italian [22], Finnish [2], French [33], Greek [17], Slovenian [31], and Swedish [15]).

The most relevant paper located by us was a paper by Moukdad and Large [21] that evaluated the capabilities of AltaVista for retrieving documents from full-text Arabic databases. The paper discusses the language specific characteristics of Arabic (many of which exist in Hebrew as well and will be discussed later). A small local Arabic full-text database was created, and was indexed by a free version of AltaVista. Altogether 560 searches were conducted. The queries were carefully constructed to highlight the special features of Arabic (mainly the frequent use of prefixes). They conclude that search engines designed for English do not work effectively with Arabic data, and the handling of prefixes necessitates the development of new information retrieval algorithms for this language.

A research area related to our current study is "cross-language information retrieval", defined in [28] as the "retrieval of documents based on explicit queries formulated by a human using natural language when the language in which the documents are expressed is not the same as the language in which the queries are expressed." Cross-lingual systems also have to take into account language specific features and characteristics. For an overview of cross-language information retrieval consult [12, 29].

English is a morphologically simple language and Harman [14] noted that there is no empirical evidence that stemming algorithms increase retrieval performance for English. Google does not apply stemming at all, i.e. when searching for information about dogs; one should specifically search for *dog* as well as *dogs*. Stemming in English is mainly useful for conflating singular and plural forms of nouns. This is not necessarily the case for other languages, in Russian and Hungarian declensions and cases are expressed through the usage of postfixes, which sometimes change the basic form of the word. In Hebrew the situation is further complicated, since the article, the conjunction and the prepositions are added as prefixes to the words. In French too, the article can be considered sometimes as a prefix of the word, although it is separated by an apostrophe.

Let us acquaint ourselves with the specific features of each of the languages examined by us. We shall only highlight the characteristics having impact on typical Web searches.

2.1 Specific characteristics of Russian

Russian belongs to the Slavic language family. Some Slavic languages use the Latin alphabet (e.g. Polish or Croatian), while others (e.g. Russian or Bulgarian) use the Cyrillic alphabet. Transliteration of our examples appears in brackets.

In the Russian language, nouns, pronouns and adjectives have 6 cases with different endings, 3 genders (masculine, feminine and neuter) and two numbers (singular and plural). The gender of common nouns is inherent and such nouns can be declined by

cases and numbers but not by genders (e.g. 'мальчик' [malchik] (a boy) will be always masculine, 'девочка' [devochka] (a girl) feminine, and 'солнце' [solntse] (sun) is neuter). However, some nouns describing title or occupation may have the same stem in both masculine and feminine genders: ученик [uchenik] (schoolboy)/ ученица [uchenitsa] (schoolgirl), король [korol] (king) / королева [koroleva] (queen) etc. Adjectives may be inflected in all genders and numbers. Usually applying a wild-card character allows retrieving all forms, but sometimes wild-card does not help: for irregular words when the stem changes in different forms (человек-люди [chelovek-ludi]; man/men, people), for vowel alternations in the stem (окно-окон [okno-okon]; a window (nominative)/ of the windows (plural genitive), or when the stem is too short (дом [dom]; home, house).

Retrieval of verbs is problematic as well. Vowel alternations and different stems are also frequent in the verbal forms (идти-шел [idti-shel]; go/he went). Russian verbs have perfect and imperfect aspects forms, for example: начинать – начать [nachinat'-nachat'] (to start). All are conjugated in three tenses, three persons in singular (different for three genders) and three persons in plural. The multiplicity of verb forms cannot even be covered by a wildcard (*); some verbs (with their adverbial participle, gerunds and participle phrases) may reach up to 250 different forms [40].

Russian words (nouns, verbs and other parts of speech) are very variable, and their meaning changes by adding to the same stem different prefixes or suffixes. Example of noun with suffixes: образ [obraz] (mode, image), образец [obrazets] (sample), образчик [obrazchik] (sample), образность [obraznost'] (figurativeness), образование [obrazovanie], (education/ formation). Thus the use of right truncation, suggested for taking care of inflections may introduce a lot of noise to the search results. Sometimes prefixes change the meaning only slightly, thus could all be retrieved as a result of a search (e.g. думать [dumat'], подумать [podumat'] (think), продумать [produmat'] (think over); передумать [peredumat'] (think better, or change one's mind); задумать [zadumat'] (think of, or decide); обдумать [obdumat'] (think over); придумать [pridumat'] (think up)), but search engines rarely apply left truncation.

For more extensive discussions of Russian grammar, the reader may consult, for example [7, 19, 20, 32]. Prompt's (http://translate.ru) free online translator, translates between several languages, including to/from Russian from/to English.

2.2 Specific characteristics of French

Diacritics, such as accents (grave accent, acute accent, circumflex) or cedilla, are inherent to the French language, and are a major source of problems for information retrieval. The most significant for the search might be accents, since accents may change the word meaning. For example: '[la] recherche' (search, research) is a noun, and 'recherché' is a participle verbal form used also as adjective (searched, or refined, exquisite).

Additional problems are caused by special form of plural, e.g. cheval - chevaux (a horse-horses). This kind of ending is very common in the French language, and it causes a retrieval problem that cannot be solved by a simple wildcard (*), since the stem may be too short to avoid "noise" in results.

Articles, pronouns, and prepositions (or articles compound with preposition) are considered "empty" words. They generally precede the word: le, la, de, du, de la, en, à, et, ou, des, etc. Singular definite articles are used with apostrophe before a vowel: l'art (the art), de l'amour (of the love), etc.

Other special characteristics are multiple verbal forms with many exceptions. Complex tenses often have different forms for masculine and feminine genders. For most verbs, the stem changes in participle, a verbal part of all composed tenses. Usually, a stem may contain only two characters: for example: lire (to read)- lu (participle of read); voir (see) – vu (participle); devoir (must) – dû (participle).

For more detailed discussions of French grammar, consult for example [16, 35]. Systran (http://www.systransoft.com/) is a free online translator between English and French.

2.3 Specific characteristics of Hungarian

Hungarian is a Finno-Ugric language. Although it uses the Latin alphabet there are a nine, accented vowels. The vowels, fourteen altogether come in pairs, five are just short and long versions of the same sound, while for two pairs (a - á, e - é) each character represents a different sound. If either the searcher or the search tool ignores the accents, different words are grouped together (e.g. kar - kár (arm, faculty - damage, pity); sor - sör (line - beer)).

Hungarian has a complex case system involving 16 to 24 distinct forms (depending on the assumptions about the exact number of case suffixes). Cases are represented as suffixes; these suffixes are added after the suffix for plural and possession. Thus three inflectional suffixes may be chained on the word stem. A possible source of problem is that the basic form of the word may change because of these suffixes (e.g. lámpa - lampák (lamp-lamps) or gyomor - gyomrom (stomach - my stomach), falu - falvak (village - villages)). This may cause problems for automatic stemming algorithms.

Verbs are conjugated, like in Russian; and personal suffixes are added to the verbs in all tenses (e.g. adom, adod, adta - I give, you give, he gave). Verbal particles are added as prefixes, they serve to mark direction, aspect and to make verbs transitive (e.g. ad - give, elad - sell, átad - hand over). These prefixes are separated from the base verb in negative and in imperative. It is possible to create nouns from most verbs with or without verb particles (e.g. ír - irat (write - document), leír – leírás (describe - specification)). Usually the verb particles express different senses, but occasionally these senses are very near, and the nouns created from them have the same (or nearly the same) meaning (e.g. javítás - kijavítás, repair (noun); bizonyítás - bebizonyítás, proof). In these cases it would be useful to search for both forms, while in general it is only necessary to retrieve the verb with its specific verb particle. This becomes more difficult in negated or imperative sentences, where the particle is separated, although the particle as a stand-alone word usually appears immediately after verb.

Our review of the Hungarian grammar was based on [18]. Intertran (http://www.tranexp.com:2000/Translate/result.shtml) is a free online translator between English and Hungarian, and the Computer and Automation Research Institute of the Hungarian Academy of Sciences (SZTAKI) has developed and published several online dictionaries (http://dict.sztaki.hu/index.jhtml).

2.4 Specific characteristics of Hebrew

Hebrew is a Semitic language. It uses the Hebrew alphabet, it is written from right-to-left and vowels are always omitted in writing (except for special cases like the Bible, poems and other

literary works and in books intended for beginning readers of Hebrew). Because of the omission of vowels, the number of homonyms (same spelling, but different meaning) are much higher in Hebrew than in English (e.g. מספר has at least six meanings 1)[mispar] number, 2) storyteller [mesaper], 3) from a book [misefer], 4) from a hairdresser [misapar], 5) numbered [muspar] or 6) coifed [mesupar] - all spelled the same, but the pronunciations are different). On the Internet at least four different encodings of Hebrew are used: Hebrew (ISO logical), Hebrew (ISO visual), Hebrew (Windows) and Hebrew (DOS).

Hebrew is a morphologically complex language, which uses word roots heavily. Many words, governed by different rules can be formed from a given root. Most prepositions, the definite article (ה - the) and some conjunctions (e.g. ו - and, ש - that) are prefixed to the words. Several layers of prefixes can be added. We counted more than 20 different combinations. Plurals and possessives (depending on the person) are added as postfixes. There are two genders in Hebrew male and female and each noun has a gender. Numbers and adjectives describing the noun appear in the same gender as the noun, and the related verbs are conjugated according to the gender of the noun. Verb conjugation is rather complex; a root can be conjugated according to several verb patterns (בנינים). There are seven verb patterns, when each verb pattern has a slightly different meaning (e.g. לבש [lavash] - to wear, הלביש [hilbish]- to get someone dressed, התלבש [hitlabesh]- to dress oneself, all from the root לבש [lavash]). Sometimes the meanings can be radically different, (e.g. השריש [hishrish] - to strike roots and שירש [shiresh] - uprooted, both from the same root שרש [sharash]).

The major problem for information retrieval are the prefixes, prefixes are so prevalent in the Hebrew language, that any search engine that does not strip these prefixes looses a lot of information. For example, when searching for information on a university (אוניברסיטה) [universita], pages on which the word university does not appear as a stand-alone word, but only with prefixes (e.g. האוניברסיטה [hauniversita] - the university or באוניברסיטה [bauniversita] - in the university) are potentially relevant to the topic, but will not be retrieved if only the exact form of the query term is searched. On the other hand, incorrect identification of the prefixes may also introduce noise to the search results (i.e. when a part of a word is mistakenly considered a prefix: מדבר could mean *desert* [midbar], *from a thing* [midavar] or *he talks* [medaber]. If the leading letter is stripped off, then the search will include all forms of the word *thing* and of the verb *talk* and hardly any results on the intended term *desert* would be retrieved.

For more detailed discussions of Hebrew grammar, consult for example [9, 39] in English or [41] in Hebrew. Morfix (http://milon.morfix.co.il) is a free online English-Hebrew / Hebrew-English dictionary.

3. METHODOLOGY

For each language we carefully selected a set of search terms and ran these queries on both the general and the local search engines. The queries were selected to emphasize the specific characteristics of each language. The searches were carried at the beginning of November 2002. When available, we consulted the help files of each tool to learn about their declared features and capabilities relevant to our searches.

The general search engines tested by us were: Google (the local versions were used instead of http://www.google.com), AlltheWeb (FAST) (http://www.alltheweb.com) and AltaVista (http://www.altavista.com). We searched AltaVista before it switched to its new format on November 12, 2002 (except for the Hebrew searches). Google only searches for terms in their exact form, and except omitting stopwords (like articles), it does not perform any morphological analysis in any language. AlltheWeb searches for exact form only, and AltaVista claims to be accent sensitive, where a query word without accents should retrieve all forms, and a query word with accents should retrieve the exact form only [3]. Google and AlltheWeb reportedly [4, 5] look for exact matches of the query term, i.e.a search for electricite (in French) will retrieve this form only, and not électricité or electricité, and a search for électricité not retrieve occurrences of electricité or electricite.

For Russian we picked three of the most popular local search tools: Yandex (http://www.yandex.ru), Rambler (http://www.Rambler.ru) and Aport (http://www.aport.ru). Yandex claims to search all terms in all grammatical forms. Rambler also searches for all declinations of a word, according to its help, and Aport claims to apply morphological treatment to regular Russian words, but not to rare words.

For French we consulted three language specific tools: Voila (http://www.voila.fr), AOL France (http://recherche.aol.fr/) and the French-Canadian portal (covers only a relatively small number of sites), La Toile de Quebec (http://www.toile.com/). According to [6], Voila is supposed to retrieve all forms of a word (with or without accent), regardless of the phrasing of the query. La Toile de Quebec is indifferent to accents. We were unable to locate any help files or discussion of the features of AOL France.

For Hungarian we accessed three Hungarian search engines: Origo-vizsla (http://www.origo.hu), Startlap (http://www.startlap.hu) and Heureka (http://www.heureka.hu/). We were able to locate reasonable documentation only for Heureka and Origo. In Heureka one may choose to enter the terms without diacritics (searches all forms), with diacritics (exact form) or one can allow the system to automatically add diacritics, according to the most prevalent possibility. For our examples we chose the exact form option. Heureka also allows right truncation (using *), Origo-vizsla allows truncation only after at least four characters, however it claims to be able to recognize word forms automatically, thus kutya (dog) will also search for kutyák (dogs). The help does not mention anything about the interpretation of diacritics.

The Hebrew language queries were submitted to Morfix (http://www.morfix.co.il), a Hebrew search engine with a built-in morphological analyzer, and to the most popular Israeli portal, Walla (http://www.walla.co.il). Morfix enables four types of searches: search for the exact form, for all morphological forms, for extended forms of the same root, and for the word and its synonyms. Walla is Israel's most popular portal, it indexes both Hebrew and English pages, and a query for a Hebrew term may retrieve non-Hebrew sites as well (all sites indexed by the portal have Hebrew summaries). When searching for sites indexed by Walla there are two options: to search for the exact form of the query word or to treat it as "partial word" (no explanation could be found, probably means that the string of letters s typed into the search box in a appear word, additional characters may appear both on the left and the right hand side of the string). We were unable to locate a help file or any other documentation for this

portal. In Walla one can also search for Web pages (a service powered by FAST), but this option is rather difficult to access, all searches are carried out first in the directory as the default.

4. RESULTS AND DISCUSSION

4.1 Russian

The results for Russian show that the major local search engines, apply morphological analysis to the queries (this was stated in their documentation as well), while the general search engines do not take into account the language specific characteristics of Russian, and essentially perform only a simple pattern-matching between the query terms and words in the documents. The results for Yandex, Rambler and Aport appear in Table 1. The results for the general search engines appear in Table 2. In spite of our expectations (based on the help) to retrieve the exact form only, AlltheWeb retrieved causal endings in our search for человек. Note the differences in the coverage of the search engines (the number of results reported by the search engines on our queries).

Table 1: Results for the Russian search tools

Query	Yandex	Rambler	Aport
Окно [okno] (a window)	1,761,038 pages, All forms retrieved	1,082,872 pages All forms retrieved	1749 pages All forms retrieved
Окон [okon] (of windows)	1,759,809 pages All forms retrieved Same first pages as in previous search	1,082,873 pages Same number of documents	1749 pages Same results
белый [belyi] (white)	2,381,600 pages Upper and lower case, all forms retrieved	1,473,232 pages	1445 pages Not case-sensitive
Белый [Belyi] (capital letter)	2,762,422 pages, Only upper case, all forms Cannot explain why more results than in previous search.	1,473,232 pages Same results, not case-sensitive	1932 pages Not case-sensitive, cannot explain why more results than in previous search
человек шел [chelovek shel] (man went)	279,542 pages человек: 34,967,771, шел: 9,325,768 All forms of both words	470,226 pages All forms	4000 pages All forms retrieved
люди идут [ludi idut] (men go)	300,994 pages люди: 42,134,627, идут: 10,039,246 Although all forms retrieved, results number and word statistics differ	470,226 pages All forms, Same number of documents	4000 pages All forms, same results
люди идут [ludi idut] (exact form)	22,763 pages !люди: 5690710, !идут: 738513 Only exact form retrieved	5,506 pages Only exact form	4000 pages for “!” normal search; 3000 pages for “ “ All forms, not exact search
начинать [nachinat’] (start, imperfect aspect)	10,891,949 pages начинать: 23341091 All forms of the same stem retrieved, not only verbs but nouns as well.	2,263,657 pages Many verbal forms	1920 pages All forms with/out suffixes, also nouns
начать [nachat’] (to start, perfect aspect)	10,896,081 pages начать: 23352435 Different results, but same first pages as in previous search, all forms	4,376,516 pages Different results, although many forms retrieved	1956 pages Many forms w/out suffixes, also nouns. Different results.

Table 2: Results for the general search engines on queries in Russian

Query	Google - pages in Russian	AlltheWeb - pages in Russian	AltaVista - pages in Russian
окно [okno] (a window)	525,000 pages Only exact form	1,651,720 pages Seemingly only exact form	383,891 results Only this form
окон [okon] (of windows)	176,000 pages Only exact form	1,633,407 pages Seemingly only exact form	136,895 results Only exact form
белый [belyi] (white)	467,000 pages Not case-sensitive	2,329,278 pages Case insensitive	428,080 results Case-insensitive
Белый [Belyi] (capital letter)	467,000 pages Same results, not case-sensitive	2,329,278 pages Case insensitive, same results	428,080 results Case-insensitive

Человек шел [chelovek shel] (man went)	271,000 pages Exact form for both words	198,254 pages Casual endings for 'человек' retrieved	185,341 results Only exact form
люди идут [ludi idut] (men go)	318,000 pages Exact form for both words	1,293,462 pages Only exact form	180,326 results Only exact form
люди идут [ludi idut] (exact form)	10,300 pages for " " Exact phrase retrieved	8,838 pages Exact phrase	5,424 results Exact phrase retrieved
начинать [nachinat'] (start, imperfect aspect)	202,000 pages Only exact form	3,818,732 pages Only exact form	162,186 results Only exact form
начать [nachat'] (to start, perfect aspect)	487,000 pages Only exact form	7,476,683 pages Only exact form	436,496 results Only exact form

4.2 French

The results for French were even more disappointing than for Russian: not only the general search engines ignore the language characteristics (accents, apostrophized articles and singulars/plurals), but also most of the language specific tools. The results for the language specific tools appear in Table 3. La Toile de Quebec is a portal; therefore its coverage is low. We have no explanation why for Voila, the number of results for l'électricité is higher than the number of results for électricité - the results for électricité should include pages with l'électricité. To our surprise, results for l'électricité included pages in which the word électricité appeared only without the apostrophized article.

In Table 4 we present the results for the general search engines. There are huge differences in the coverage, perhaps due to different interpretations of the search space, but apriori we thought that searching for French language pages in AlltheWeb, and for francophone pages in Google has the same meaning. The general coverage of AltaVista is lower than that of Google and AlltheWeb, hence the differences in the number of results.

A possible explanation for the incorrect handling of accents, is that the accents can either be written using the French character set, or by using the special html characters, e.g. é for é.

Table 3: Results for the French search tools

Query	Voila	AOL France	La Toile
electricite (electricity)	148,614 docs All forms, diacritics ignored	230,240 documents All forms, diacritics ignored	137 docs All forms, diacritics ignored
électricité	148,614 docs Same results	230,240 documents Same results	137 docs Same results
l'électricité	149,891 More results than without apostrophe	230,240 documents Same results as without apostrophized article	137 docs Same results as without apostrophized article
cheval (horse)	143,943 docs Results for sites (266 sites) include both forms, but results for pages include the exact form only	264 208 documents Only exact form	155 docs Only exact form
chevaux (horses)	97,565 docs Same results for sites, but not for pages	161 611 documents Singular and plural searched separately	143 docs Only exact form

Table 4: Results for the general search engines on queries in French

Query	www.google.fr (francophone pages)	AlltheWeb (in French)	AltaVista.com in French
electricite (electricity)	152,000 pages (includes forms with and without diacritics - in spite of what's stated in the help)	791,590 pages all forms	318, 758 pages Includes form with and without diacritics
électricité	149,000 pages Should be exact form only, seemingly this is the case	791,590 pages all forms, same results	260,904 pages Seemingly this form only
l'électricité	97,700 pages Supposed to ignore articles, but seemingly retrieves word+article.	257,332 pages word+article	165, 946 pages Apostrophized article retrieved as word
cheval (horse)	273,00 pages Exact form	1,173,587 docs This form only	255, 979 pages This form only
chevaux (horses)	193,000 pages Exact form	539,890 docs This form only	130,217 pages This form only

4.3 Hungarian

The Hungarian search tools, Origo-vizsla, Startlap and Heureka take into account to some extent the language-specific characteristics of Hungarian (see Table 5). Origo-vizsla claims to extend the search to various word forms, but even for the example suggested by it (dog-dogs), it reported slightly different number of results for the two searches. Startlap seems to over-extend and to include unrelated word forms as well (like zenekar (orchestra), when searching for kar (arm, faculty)).

It is easier to search only for the string entered by the user, but for morphologically complex languages, such as Hungarian, simple pattern matching is not appropriate. The right balance should be found between including too many word forms (and "noise") and over-restricting the retrieval. At this point of time, the general search engines (see Table 6) do not include word forms, a serious problem for Hungarian. AlltheWeb does not even differentiate between vowels with or without diacritics.

Table 5: Results for the Hungarian search tools

Query	**Origo-Vizsla (in the Hungarian Web)**	**Startlap**	**Heureka (in the Hungarian Web)**
kar (arm, faculty)	705,136 pages only kar as stand-alone word in first 100 results kar* not applicable (at least four characters preceeding * sign)	299,364 pages different forms, including **kar**eoke, or zene**kar** (orchestra)	21,782 pages kar* (right truncation) 72,425 pages
kár (damage, pity)	642,999 pages different form previous results	216,274 pages different forms in which the string appears, sometimes in the middle of a word	14,412 pages kár* 100,489 pages
kutya (dog)	395,947 pages kuty* 729,184 truncation is applicable here, according to the help, but top results are of low relevance or frequently changing pages (e.g. news sites)	235,166 pages	13,681 pages kuty* 21,526 pages
kutyák (dogs)	399,152 pages not the same number of results as before. The word **kutya** (dog) is also emphasized in the result summary	51,554 pages seemingly this tool interprets the search terms as *term*, i.e. it expands the search string in both direction	4399 pages Only for Origo-vizsla we received more results when searching for the plural form
falu (village)	410,000 pages	257,735 pages	15,767 pages
falvak (villages)	410,620 pages larger number of results	37,588 pages	3,983 pages
javítás (repair)	752,185 pages	136,067 pages	6956 pages
kijavítás (repair)	18,271 pages	98 pages	95 pages

Table 6: Results for the general search engines on queries in Hungarian

Query	**Google (search for pages written in Hungarian)**	**AlltheWeb (in Hungarian)**	**AltaVista (in Hungarian)**
kar (arm, faculty)	111,000 pages	138,026 pages includes **kár** as well, but not on top pages	30,945 pages
kár (damage, pity)	40,700 pages seemingly exact form only	142,266 pages top results contain **kar** instead of **kár** Not the same number of results as for the previous search	10,867 pages
kutya (dog)	41,500 pages	64,270 pages	8652 pages
kutyák (dogs)	15,600 pages	20,400 pages	14,322 pages more results than for singular
falu (village)	43,600 pages	49,548 pages	12,292 pages
falvak (villages)	12,000 pages	13,851 pages	3883 pages

javítás (repair)	18,600 pages	43,624 pages	4648 pages
kijavítás (repair)	298 pages seemingly (as stated in the help) the searches are for exact forms only	277 pages	74 results

4.4 Hebrew

The results for the Hebrew searches appear in Tables 7 and 8. The coverage of Hebrew language pages by the different search tools is rather variable. Morfix seems to cover only pages in the domain .il (Israel), while a non-negligible portion of Israeli sites are registered under other domains (an interesting example is the Israeli Postal Authority, http://www.postil.com), not to mention Hebrew pages created outside Israel.

In Morfix we utilized two out of the four existing options and carried out exact and morphological searches. In Walla we queried both the directory (sites) and the search engine powered by FAST (this tool is well hidden in the site). For site searches both the "whole word" and the "partial word" options were applied.

The major search engines search for the words in their exact form only. For the word *home / house* [bait] (בית) a large number of the occurrences are stand-alone, but even if this case, when Morfix retrieves all morphological forms instead of just the exact form, the number of results increases by 60%. The need to take into account prefixes and postfixes is even more emphasized for the word *university*. According to Morfix, it appears as a stand-alone word in texts only 3% of the time, in the remaining cases it either appears with the definite article (a prefix), as part of a possessive phrase (*university of* in a phrase like the*University of Tel Aviv*), with prepositions (e.g. *in the university*) or in some combination of prefixes and postfixes (e.g. *and that the university* – ושהאוניברסיטה [veshehauniversita]). In Google, when searching for the Boolean phrase *university* OR *the university* OR *university of* OR *in the university* OR *to the university* OR *in the university of* OR *from the university* OR *to the university of* OR *from the university of* (אוניברסיטה OR האוניברסיטה OR אוניברסיטת OR באוניברסיטת OR מהאוניברסיטה OR לאוניברסיטה OR באוניברסיטה OR מאוניברסיטת OR לאוניברסיטת) 56,500 results were retrieved, nearly eight times more than for the stand-alone word *university* (and the Boolean query, which is limited to at most 10 words have not covered all the possibilities). When searching for *university*, the user will simply enter the stand-alone word only, not thinking of all the combinations he misses this way; and even if he is aware of the need to search for additional forms of the word, he will not invest the effort to think of and to type in all the forms. An additional problem is that when searching for the stand-alone word, rather unprominent results appear at the top of the list, searches for *the university* or for *university of* give far better results (both in Google and in Morfix). Morfix is capable of retrieving all forms of a word but not in a consistent way, as the search for desert / he talks / from a thing [midbar/ medaber/ midavar] (מדבר) indicates (see last row of Table 7). When we examined the first 100 results for the morphological search for *university*, 54 of the pages were popular medical articles from the site doctors.co.il, where in the majority of the cases university was mentioned in the context of Prof. X from university Y; another 30 hits were outdated pages (from 1999-2001) from the calendar of events of the Hebrew University. Thus we judged at least 80% of the results as non-relevant. Retrieving all forms of a word is definitely a indispensable option for Hebrew, but this capability alone is not sufficient for producing high quality results.

Table 7: Results for the Hebrew search tools

Query	**Morfix (exact search)**	**Morfix (morphological search)**	**Walla (whole word)**	**Walla (partial word)**
אוניברסיטה [universita] (university)	1527 pages	51,862 pages	244 sites 3959 pages	285 sites
האוניברסיטה [hauniversita] (the university - prefix)	10,660 pages	51,862 pages	78 sites 81,663 pages	81 sites
באוניברסיטה [bauniversita] (in the university - prefix)	5813 pages	51,862 pages	28 sites 14,047 pages	31 sites
אוניברסיטת [universitat] (university of - postfix)	12532 pages	51,862 pages	136 sites 92,622 pages	378 sites for אוניברסיט (this should retrieve all prefixed and suffixed forms)
ושהאוניברסיטה – [veshehauniversita] (and that the university)	3 pages	51,862 pages	0 sites 2 pages	n/a
מהבית [mehabait] (from the home / from the house)	3521pages	587,020 pages	17 sites 62,856 pages	17 sites
בית [bait] (home / house)	363,490 pages	587,020 pages	More than 400 sites 804,047 pages	More than 400 sites

מדבר [midbar/ medaber/ midavar] (desert / he talks / from a thing)	16,921pages	19,433 pages Not all forms are included. Search for thing (דבר) [davar] retrieved 175,947 results and search for she talks (מדברת) [medaberet] retrieved 112074 results	43 sites 27,295 pages	80 sites

Table 8: Results for the general search engines on queries in Hebrew

Query	Google (http://www.google.co.il)	AlltheWeb (in Hebrew)	AltaVista (in Hebrew)
אוניברסיטה (university)	7100 pages	2061 pages	802 pages
האוניברסיטה (the university - prefix)	26,900 pages	30,261 pages	3328 pages
באוניברסיטה (in the university - prefix)	14,800 pages	5427 pages	1898 pages
אוניברסיטת (university of - postfix)	25,500 pages	36,412 pages	3800 pages
ושהאוניברסיטה (and that the university)	5 pages	1 pages	0 pages
מהבית (from the home / house)	12,200 pages	4161 pages	2286 pages
בית (home / house)	133,000 pages	391,064 pages	46,665 pages
מדבר (desert / he talks / from a thing)	30,500 pages	9337 pages	5425 pages

5. SUMMARY AND CONCLUSIONS

The World Wide Web is not the "Web of English": more than 50% of the users are not native English speakers, and estimates claim that about 1/3 of the Web pages are non-English pages (at least partially). Publishing Web pages is great, but even the greatest Web page needs to be found. The most prevalent source for locating Web pages on a given topic is the search tools. If these tools are not suited to search in the specific language, the page might never be found. Thus pages in non-English languages have a much larger chance of "being lost in Cyberspace".

Information retrieval research has been geared so far mainly towards English. English is a morphologically simple language. When extending search capabilities to non-English languages morphological variations have to be taken into account. English is spoken by hundreds of millions of people all over the world, while the native languages are spoken by much smaller populations. Probably, it is not worthwhile economically to develop good retrieval tools for these languages. However, if we want the Web to remain a place for everybody, regardless of the languages she speaks, an effort has to be made to provide these tools.

6. REFERENCES

[1] Abu-Salem, H., Al-Omari, M., and Evens, M. W. Stemming methodologies over individual query words for an Arabic IR system. Journal of the American Society for Information Science, 50 (1999), 524-529.

[2] Alkula, R. From plain character strings to meaningful words: producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval, 4 (2001), 195-208.

[3] Andrieu, O. AltaVista syntaxe. http://outils.abondance/av_syntaxe.html

[4] Andrieu, O. Google syntaxe. http://outils.abondance/goo_syntaxe.html

[5] Andrieu, O. All The Web syntaxe. http://outils.abondance/atw_syntaxe.html

[6] Andrieu, O. Voila syntaxe. http://outils.abondance.com/vl_syntaxe.html

[7] Beard, R. An on-line Russian reference grammar. http://departments.bucknell.edu/russian/language

[8] Cyber Atlas. Web pages by language. http://cyberatlas.internet.com/big_picture/demographics /article/0,1323,5901_408521,00.html (July, 2000)

[9] Glinert, L. Hebrew - An essential grammar. Routledge, London, 1994.

[10] Global Reach. Global Internet statistics (by language). http://www.global-reach.biz/globstats (September, 2002)

[11] Google. Language tools. http://www.google.com/language_tools (October, 2002)

[12] Grefenstette, G. Problems and approaches to cross language information retrieval. In Proceedings of the ASIS Annual Meeting, 35 (1998), 143-152.

[13] Grefenstette, G., and Nioche, J. Estimation of English and non-English Language Use on the WWW. In

Proceeding of RIAO 2000, 237-246. http://arxiv.org/ftp/cs/papers/0006/0006032.pdf

[14] Harman, D. How effective is suffixing? Journal of the American Society for Information Science, 42 (1991), 7-15.

[15] Hedlind, T., Pirkola, A., and Jarvelin, K. Aspects of Swedish morphology and semantics from the perspective of mono and cross-language information retrieval. Information Processing and Management, 37 (2001), 147-161.

[16] L'Huillier, M. Advanced French Grammar, Cambridge University Press, Cambridge, 1999

[17] Kalamboukis, T. Z. Suffix stripping with Modern Greek. Program, 29 (1995), 313-321.

[18] Megyesi, B. The Hungarian language - A short descriptive grammar. http://www.speech.kth.se/~bea/hungarian.pdf

[19] Miloslavsky, I. A short practical Russian grammar Russkii Iazyk, Moscow, 1988.

[20] Mitrevski, G. RWT Tutorial. http://www.auburn.edu/~mirtege/RWT/tutorials

[21] Moukhad, H., and Large, A. Information retrieval from full-text Arabic databases: Can search engines designed for English do the job? Libri 51 (2001), 63-74.

[22] Monz, C., and de Rijke, M. Shallow morphological analysis on monolingual information retrieval for Dutch, German and Italian. In Post-Conference Proceedings of the Cross Language Evaluation Forum Workshop (CLEF 2001). http://staff.science.uva.nl/~christof/ Papers/clef-2001-post.pdf

[23] Mujoo, A., Malviya, M. K., Moona, R., and Prahakar, T. A search engine for Indian languages. EC-Web 2000, Lecture Notes in Computer Science 1875 (2000), 349-358.

[24] Nielsen. Top 10 Web properties for the month of September 2002 - France. http://epm.netratings.com/fr/web/NRpublicreports.toppropertiesmonthly

[25] Nielsen. Top 10 Web properties for the month of April 2002 - Israel. http://epm.netratings.com/il/web/NRpublicreports.toppropertiesmonthly

[26] NUA Internet Surveys. How many online? http://www.nua.ie/surveys/how_many_online/index.html (September, 2002)

[27] Nunberg, G. Languages of the Wired World. Presentation at "The Politics and the Building of Modern Nations", Institut d'Etudes Politiques de Paris, October 2, 1998. http://www-csli.stanford.edu/~nunberg/WebPaper.html

[28] Oard, D. Cross-language information retrieval defined. http://raven.umd.edu/dlrg/clir/mlir_definition.html (1997).

[29] Oard, D., and Diekama, A. R. Cross language information retrieval. Annual Review of Information Science and Technology, 33 (1998), 223-256.

[30] OCLC Web Characterization Project. http://wcp.oclc.org/

[31] Popovic, M., and Willett, P. The effectiveness of stemming for natural language access to Slovene textual data. Journal of the American Society for Information Science, 43 (1922), 384-390.

[32] Pulkina, I.M A practical grammar with exercises: Russian. (For English speakers.) Russkij Yazyk, Moscow, 2002

[33] Savoy, J. A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50 (1992), 944-952.

[34] Sroka, M. (2000). Web search engines for Polish information retrieval: Questions of search capabilities and retrieval performance. International Information & Library Research, 32 (2000), 87-98.

[35] Solhaut. French grammar help online. http://www.geocities.com/solhaut

[36] Sullivan, D. Nielsen/NetRatings search engine ratings. http://www.searchenginewatch.com/reports/netratings.html (September, 2002)

[37] WebmasterWorld. European search engine chart. http://www.webmasterworld.com/forum18/544.htm (September, 2002)

[38] Walla (in Hebrew). The TIM survey: there are 1.7 million Internet users in Israel. http://news.walla.co.il/ts.cgi?tsscript=item&id=244719 (June, 2002).

[39] Wartski, I. Hebrew grammar and explanatory notes. The Linguaphone Institute, London, 1900.

[40] Yandex. http://company.yandex.ru/programs/dict/.

[41] Yelin, D. Dikduk HaLashon HaIvrit (Hebrew Grammar, in Hebrew). Jerusalem, 1970.