Current research issues and trends in non-English Web searching

Fotis Lazarinis

Technological Educational Institute of Mesolonghi, Greece

lazarinf@teimes.gr

Jesús Vilares

Department of Computer Science, University of A Coruña, Spain

jvilares@udc.es

John Tait

Information Retrieval Facility, Vienna, Austria

john.tait@ir-facility.org

Efthimis N. Efthimiadis

The Information School, University of Washington, USA

efthimis@u.washington.edu

Abstract

With increasingly higher numbers of non-English language web searchers the problems of efficient handling of non-English Web documents and user queries are becoming major issues for search engines. The main aim of this review paper

is to make researchers aware of the existing problems in monolingual non-English Web retrieval by providing an overview of open issues. A significant number of papers are reviewed and the research issues investigated in these studies are categorized in order to identify the research questions and solutions proposed in these papers. Further research is proposed at the end of each section.

Keywords: Non-English retrieval, Web searching, Query log analysis, Segmentation, Indexing, Stopwords, Stemming, Lemmatization, Language identification, Encoding handling

Introduction

Search engines are essential tools for finding and exploring information from Web pages and other specialized Web information systems, e.g. e-commerce sites. Search engines originate from traditional Information Retrieval (IR) tools but they take many forms reflecting the dynamism of the web. Traditional IR systems (Baeza-Yates and Ribeiro-Neto 1999) typically operate on closed corpora. However, search engines have to regularly crawl the Web to index millions of constantly changing hypertext documents containing information in a variety of languages and media formats. Further, the search engine services are available globally to every user with Internet access, users who have different computer handling abilities, cultural backgrounds, education, aims and, most importantly, who speak different languages. However, in two recent iNEWS (improving Non-English Web Searching) workshops, a theme that emerged was that search engines ignore the intricacies of non-English natural languages and this results in lower accuracy (Lazarinis *et al.* 2007; 2008).

Thus, the main aim of this article is to make researchers aware of the existing problems in non-English Web retrieval by providing insights into the open research issues, and by focusing on monolingual search, not cross- or multi-

lingual searching. We review research studies on non-English Web searching and attempt to categorize the problems identified in the literature. The rest of the paper is structured as follows. Initially, studies discussing issues arising during the pre-processing and indexing of non-English Web texts are discussed. Then, studies related to various aspects of Web searching are presented by language: Arabic, Slavic, German, Greek, Italian, Iberian, Asian, and finally studies with more than one language investigated. The next section reviews the limited number of research studies on query log analysis of non-English queries and presents their main findings. The last section presents the main conclusions from this review.

Indexing

Search engines crawl the Web and fetch documents which are then indexed and included in their databases. Indexing of the fetched Web documents is a complex procedure which requires, among other specialized routines, identification of the language of the document, pre-processing of the texts, tokenization, stopword identification, stemming, and uniform handling of the morphological variances of the tokens.

Language identification

Language identification in Web pages is an important issue which influences the subsequent services of the search engines. There are well-established mechanisms for the automatic identification of the language of a document based on the content of the text (Dunning 1994; Grefenstette 1995), although the algorithms must be adjusted to specific characteristics of Web texts (Martins and Silva 2005). Macdonald *et al.* (2007) mention that one of the problems they faced in adapting the Terrier IR system (Ounis *et al.* 2006) to index non-English texts was the identification of the language of the documents. They employed the language identification tool TextCat (Cavnar and Trenkle 1994), combined with evidence from the URL and the HTML of each document. In the case of Sigurbjörnsson *et*

al. (2006), a specialized tool was used to determine the language of the Web pages used in the construction of a multilingual Web corpus.

Moreover, the existence of different dialects or closely-related languages makes this task even harder. This is the case of Indonesian, for example. The issues in designing and developing a search engine for this language are reported in (Vega and Bressan 2001), which describes a language identification algorithm for Indonesian text documents which is comparatively complex because several hundreds of regional languages and dialects co-exist in this country.

Finally, it should be noted that documents written in several languages at a time are sometimes found. In this case, language identification algorithms should deal with such a multilingual content by both identifying the languages present in the document and identifying the location of a language shift (Artemenko *et al.* 2006).

Encoding handling

Another issue that should be taken into account during indexing is the existence of different encodings for the documents. This is particularly relevant in the case of Asian languages since it causes problems in language identification (Pingali *et al.* 2006; Chau *et al.* 2007; Macdonald *et al.* 2007).. Pingali *et al.* (2006), for example, mention that more than 95% of Indian language content on the web is not searchable and multiple encodings of web pages are specifically identified as a major cause. Most of these encodings are proprietary and hence need some kind of standardization for making the content accessible via a search engine. Moreover, Indian language words also have standardization issues in spelling, hereby resulting in multiple spelling variants for the same word with the consequent difficulties this presents for search systems. Pingali and colleagues present WebKhoj, a search engine which is capable of searching multi-script and multi-encoded Indian language content on the web. Their focused crawler, which is embedded with the necessary knowledge, is able to handle efficiently several scripts and transcoded Indian texts.

Greek is another good example of this kind of problem. Lazarinis (2007b), for example, notes that systems need to take account of several non-Greek punctuation marks which are today often used in Greek texts. The use of Latin upper case characters in Greek words was also observed. For example, the word *ABAKAΣ* (abacus) seems that is a term encoded in the Greek alphabet. However, when this word was transformed to lower case, then instead of the Greek word *αβακας* the semi-Greek-semi-Latin term *abakaς* appeared. These terms were mostly in capital letters, because several Latin capital letters are identical to the respective Greek capital letters. Search engines need to take account of this to avoid such words being treated as unique terms unrelated to their lower case counterpart.

Moreover, different encodings may cause problems not only with the content of the documents, but with their filenames too, as is shown in Lazarinis and Efthimiadis (2008). This work about web image retrieval is studied show that image filenames are encoded in Latin scripts even in non-Latin languages, like Greek and Russian, thus creating false coordination problems. For example, the Polish query *"pies"* (dog) was falsely taken as the plural form of the English word *"pie"* and therefore no relevant canine images were retrieved. In addition, it was found that the absence of diacritics causes fewer relevant images to be retrieved.

Pre-processing and text segmentation

One of the most important tasks in any text processing system is the accurate pre-processing and segmentation of the input text. In the case of text segmentation, for example, such a task consists of dividing a text into linguistically meaningful units, which will be the fundamental units passed to further processing stages, such as information retrieval systems (Palmer 2000). However, these tasks are frequently approached in a naive way, as in the case of search engines, which often rely on very simple algorithms similar to those used in programming language compilers (Aho *et al.* 1986). These algorithms tokenize the text by

taking into account only the blanks and the punctuation marks, which may be enough for a program written in C, but not for human texts. Basically, the main problem of these approaches is the fact that the spelling concept of 'word' does not always coincide with the linguistic reality, as in the case of compound words, multiword expressions, etc. (Graña *et al.* 2002).

The problems of Spanish pre-processing and segmentation have been studied in-depth by Graña *et al.* (2002). Their work presents a linguistically-based pre-processing-segmenter system able to deal successfully with complex phenomena, such as multiword expressions, contractions, enclitic pronouns attached to verbs, and even segmentation ambiguities. The system was originally designed for Natural Language Processing (NLP) applications for Galician, a Romance language closely related to Portuguese and which shares official status with Spanish in Galicia, Northwest Spain. However, the general architecture of their pre-processing-segmenter was designed to be easily adapted for other languages, and a version for Spanish was later built. It has also been optimised for IR applications (Barcala *et al.* 2002).

Word boundary identification is even harder in Asian languages such as Chinese (Chen and Liu 1992; Yang *et al.* 2000; Foo and Li 2004), for example, since words are not delimited by blanks. Foo and Li (2004) conducted experiments to study the impact of Chinese word segmentation and its effect on IR. Four automatic character-based segmentation approaches and a manual one were used to index and evaluate the accuracy of these approaches. The experiments revealed that the segmentation approach had an effect on IR effectiveness. Accuracy varied from 0.34 to 0.47 based on the segmentation method. Better results could be achieved by using the same method for query and document processing, which increased the probability of matching queries to documents.

Compound words are another major problem in text segmentation. In languages like Dutch compound terms appear regularly both in texts and in user queries, and

a number of techniques ranging from dictionary-based approaches to statistical models have been proposed for decompounding these terms for both indexing and retrieval (Pohlmann and Kraaij 1997; Hollink *et al.* 2004; De Vries 2001). Pohlmann and Kraaij (1997) showed that when a query is expanded with the constituents of compounds already occurring in it and new compounds are added to the query by combining query terms, recall improves while precision does not deteriorate. The case of German language is similar (Goldsmith and Reutter 1999), where carefully designed decompounding of words increases significantly the performance of retrieval systems (Braschler and Ripplinger 2004). Monz and de Rijke (2002) show that compound splitting leads to improvements in mono-lingual retrieval performance for Dutch and German. Hedlund (2002) also reports on the effectiveness of compound splitting for Swedish.

Another research study looked into the impact of decompounding on monolingual and bilingual retrieval of English, Finnish, German and Swedish queries (Airio 2006). The authors reported a varied increase in precision in their runs. For example, the application of lemmatization and decompounding together resulted in 62.9% increase in precision in the retrieval of Finnish documents. The study argues that if no compound splitting is performed during the indexing phase in these non-English languages, only the full compound will be in the index, not its parts. This will cause some queries to fail as the queries may include only parts of the compounds. Similar issues about Swedish are discussed in Ahlgren and Kekäläinen (2006). Trial runs with Hummingbird retrieval system suggested that Hungarian would also benefit from decompounding (Tomlinson 2006a). Vega and Bressan (2001) discuss some issues on handling the boundaries between repeated Indonesian words.

Eguchi and Croft (2009-current issue) use a structured query approach using word-based units to capture compound words, as well as more general phrases, in a query. The paper discusses problems, such as compound words and

segmentation that appear in Japanese information retrieval and some research efforts to address these problems.

These studies show that pre-processing and tokenization should also be taken into account for both English and non-English retrieval. Search engines should be aware of the morphology of non-English languages and adjust their algorithms accordingly when necessary. Moreover, the user queries should be thoroughly studied to reveal how users express their information needs in these languages.

Finally, although the use of words as the processing unit is dominant, some recent works have studied an alternative proposal based on the use of character n-grams instead of words for indexing and retrieval purposes (McNamee and Mayfield 2004; Otero *et al.* 2008; Savoy 2003). Such an approach has multiple advantages, particularly for non-English languages, since the use of character n-grams allows partial matching, avoiding the need for word normalization, and also deals with misspelled and out-of-vocabulary words. Moreover, since such a solution does not rely on language-specific processing, it can be used with languages of very different natures even when linguistic information and resources are scarce or unavailable.

Stopwords

Typically during the indexing phase stopwords are identified and they are either removed, at least in typical IR systems, or noted as stopwords in order not to influence significantly the subsequent search. Although, stopword lists have existed for English for decades now (Fox 1990; Frakes and Baeza-Yates 1992), such lists are not available in many non-English languages and their effects in Web searching have not been extensively studied in many cases.

Savoy (1999) analyzed the construction process of a stopword list for the French language. This was created semi-automatically based on term frequency and on careful manual elimination of certain words from the list.

Chen and Gey (2002) developed an Arabic stopword list consisting of Arabic pronouns, prepositions, and other non-significant terms that were found in an elementary Arabic textbook. They also added some Arabic words translated from an English stopword list.

Chinese stopword identification is discussed in Zou *et al.* (2006). As mentioned earlier, Chinese text tokenization is more difficult than in many other languages since the word boundaries are not well defined. Zou and colleagues employ a segmentation algorithm first and then they build a statistical model for engineering the stopword list. This statistical model is primarily based on calculating the term frequencies of the words in a given collection. The frequencies are normalized based on document length and then the probability of a word being a stopword is calculated.

Lazarinis (2007b) discusses the construction of a stopword list for Greek, and found that the elimination of stopwords from user queries improves the precision of search engine results. The absence of freely available Greek text collections to work with is noted.

Finally, some researchers have recently addressed the problem of developing an automatic and language-independent way to generate stopword lists taking as input a document collection similar to (or equal) the one to be indexed (Blanco and Barreiro 2007; Makrehchi and Kamel 2008; Lo *et al.* 2005). This solution allows the construction of new lists in a much easier and faster way, giving the possibility of building lists for languages with a lack of resources available, or the development of specialized stopword lists for specialized collections (e.g., technical documents).

Conflation

Stemming is the process of reducing a word to its stem or root form. It is essentially a recall enhancing (and therefore precision damaging) technique

allowing documents in which a term is expressed using a different morphological from the query to be found. For web search, in which precision tends to be more important than recall, its use is therefore questionable. However it needs to be covered here partially to because of its historical importance in IR and partially because for some languages the morphological complexity and relatively small collections make it more useful.

The best known stemmer for the English language is the rule-based algorithmic stemmer of Porter (1980). The main advantage of stemming is the increase in recall as its application allows the retrieval of most of the morphological variants of the query terms. The construction of stemmers for non-English languages is more difficult than for English due to the relative morphologic simplicity of English, particularly at the inflectional level (Jurafsky and Martin 2000; Arampatzis *et al.* 2000). Stemming performance depends on the morphological nature of its language, often showing problems in languages with a complex morphology or with many irregularities (Arampatzis *et al.* 2000, Figuerola *et al.* 2001).

Alemayehu and Willett (2003) studied the effectiveness of stemming for information retrieval in Amharic. Amharic, which is spoken in Ethiopia, is the second most spoken Semitic language in the World (after Arabic), and has a very rich morphology. This means that systems for searching Amharic text databases can be effective in operation only if full account is taken of the many word variants that may occur.

Stemmers have also been reported for a wide range of languages, including Arabic (Al-Kharashi and Evens 1994; Chen and Gey 2002), French (Savoy 1999), Greek (Kalamboukis 1995), Latin (Schinke *et al.* 1998), Malay (Ahmad *et al.* 1996), Slovene (Popovic and Willett 1992) and Turkish (Solak and Oflazer 1993; Ekmekçioglu and Willett 2000). The main conclusion from these papers is that stemming improves recall, and that the construction of effective stemmers

requires a thorough understanding of the inflectional morphology and the irregularities of the specific languages. Although stemming has been reported as being beneficial for standard IR systems its effect in non-English Web searching is still an open research issue.

Savoy (2007) argues that a general light stemmer can be quite effective for Bulgarian Web searching, producing significantly better mean average precision (MAP) than an approach not applying stemming. In a short study on the effects of stemming in Greek Web searching (Lazarinis 2007c) it was shown that the application of a light stemmer which removes specific endings from Greek nouns improves proportion of the top 10 retrieved documents which are actually relevant. The same was also suggested in Web searching experiments with Greek queries in CLEF 2005 (Tomlinson 2006b).

For German it has been shown that for short queries stemming may enhance mean average precision by 23%, compared to 11% for longer queries (Braschler and Ripplinger 2004). The Indonesian language is also a morphologically rich language (Vega and Bressan 2001). There are around 35 standard affixes (prefixes, suffixes, circumfixes and some infixes). Affixes can virtually be attached to any word and they can be iteratively combined. The authors refer to the need to apply stemming to both queries and index terms in Indonesian to increase the performance of their Web search engine.

Another interesting proposal is that of (Nunzio *et al.* 2004), which studies the automatic generation of stemmers employing probabilistic models. This work, which was presented in CLEF 2003, successfully tested their proposal for five languages: Dutch, French, German, Italian and Spanish. On the other side, Xu and Croft (1998) propose to improve the behaviour of a stemmer for a corpus or language by refining the equivalence classes it generates. For this purpose, statistics of corpus-based word variant occurrences are used. Since it is a statistical approach, there is no need for a human expert.

In some studies, the effectiveness of stemming has been criticised in retrieval even in static corpora (Harman 1991; Hull 1996). Generally stemming is a language dependent recall enhancing technique. Erroneous stemming may damage precision and there are at least two forms of error which affect stemming in non-English web search: (*i.)* the application of term conflation in two semantically distinct search or index terms which at the end are reduced to the same stem and (*ii.)* the erroneous application of a stemmer to search terms which are actually in a language other than the one for which the stemmer is designed.

An alternative to stemming, which is also proposed in some of the studies discussed above, is to apply lemmatization to query and index terms. Lemmatization involves the reduction of words to their respective headwords (i.e. lemmas). Lemmatization always produces complete words. In linguistic dictionaries, for example, every entry corresponds to a lemma that defines a set of words with the same lexical root. In contrast, stemming may produce forms which are not linguistically acceptable in themselves (e.g. "irritant" to "irrit" in the Porter stemmer). Lemmatization has been shown to be important in local Web site searching (Lazarinis 2007d).

Hollink *et al.* (2004) investigate the impact on retrieval effectiveness of stemming and lemmatization in retrieval for a number of non-English European languages (i.e. Dutch, Finnish, French, German, Italian, Spanish, Swedish). In some cases the lemmatizer performs better than the stemmer but the results cannot be considered conclusive because of the limited number of queries and the static nature of the document collection. Context sensitive stemming for Web search could possibly enhance the retrieval performance in non-English queries as well (Peng *et al.* 2007).

Knowledge-poor methods for tackling person name matching and lemmatization in Polish, a highly inflectional language with a complex personal name declension paradigm is discussed in Piskorski *et al.* (2009-current issue). Their method

applies mainly well-established string distance metrics for automatically acquiring simple suffix-based lemmatization patterns. The evaluation showed that achieving lemmatization accuracy figures greater than 90% seems to be difficult, whereas combining string distance metrics with suffix-based patterns results in 97,6-99% accuracy for the name matching task.

The successful application of lemmatization for text conflation in Spanish retrieval is described in Vilares *et al.* (2008; 2003). This work looks at managing the highly complex inflectional morphology of Spanish. As an example, in the case of verbs, 3 regular and more than 30 irregular groups have been identified, with 118 inflected forms for each verb; in the case of nouns and adjectives 20 variation groups for gender and 10 for number are found (Vilares *et al.* 1996).

Google supports lemmatization and even retrieval of semantically-related terms in the case of English. An interesting research path would be the development and utilization of lemmatizers for non-English languages. These tools need to take into account the inflectional morphology of each specific language, their irregularities and even their segmentation characteristics. This is the case, for example, of the work developed for both Spanish and Galician with MrTagoo tagger-lemmatizer (Graña *et al.* 2001; Graña *et al.* 2002). Further, the forms of the user queries should be studied in non-English languages to realize how users type their queries; whether, for example, they use the same terms in various inclinations with various endings.

We can conclude from these works that the effect of both stemmers and lemmatizers still needs to be further investigated in the case of non-English Web retrieval.

However, since lemmatization is restricted to inflectional variation, some researchers have gone further and faced the problems of derivational variation (i.e., words related through derivational relations, such as *derive* and *derivation*). In these works derivational morphology is applied in order to obtain the words

derivationally-related to the original term, either for conflation or expansion purposes (Arampatzis *et al.* 2000). Many of these papers are focused on Romance languages, such as Spanish (Vilares *et al.* 2001; 2008), French (Tzoukermann *et al.* 1997) or Portuguese (Gonzalez *et al.* 2005). Although some of them are used for single-term-based retrieval (Vilares *et al.* 2001; 2003), most of them are focused on multi-word terms. This is due to the fact that these derivational mechanisms are quite sensitive to over-generation, whose effects are reduced in the case of multi-word terms because of the existence of a partial context (the multi-word term itself), which allows to partially disambiguate the derivationally-related term implicitly. However, the application of context information may overcome this problem, as suggested by Moreau *et al.* (2007), by using analogy-based machine learning for identifying derivationally-related terms to be used in query expansion.

Searching

The previous sections have discussed a number of non-English retrieval studies related to tasks primarily performed during the indexing phase: language identification, word segmentation, stopword removal, stemming and lemmatization. All these tasks influence the subsequent performance of the search engines. In this section we review papers related to the performance of search engines in non-English queries during the search and retrieval phase.

Arabic

In Moukdad (2004) the performance of general and Arabic search engines were compared based on their ability to retrieve morphologically related Arabic terms. The authors ran a limited number of single term queries which were in fact morphological variants of the same queries. For example, they used the Romanized queries *jamct* (university), *aljamct* (the university) and *baljamct* (in the university). The queries were submitted in three general (AlltheWeb,

AltaVista and Google) and in three Arabic (Al bahhar, Ayna and Morfix) search engines. The morphologically varied query terms were carefully selected to emphasize the specific characteristics of Arabic that differentiate it from English. The findings of this study show that although worldwide search engines have greater coverage, local search engines were able to retrieve pages containing the morphological variants of the query terms. The morphology of the Arabic query terms and how it influences the retrieval of documents is discussed also in (Darwish and Oard 2007). In this work, adaptations of existing Arabic morphological analysis techniques are presented to make them suitable for the requirements of IR applications by leveraging corpus statistics. A framework to enhance the retrieval effectiveness of search engines to search for diacritic and diacritic-less Arabic text through query expansion techniques is proposed in (Hammo 2009-current issue). A rule-based stemmer and a semantic relational database compiled in an experimental thesaurus were used for the query expansion. The research concludes that query expansion for searching Arabic text is promising and it is likely that the efficiency can be further improved by advanced natural language processing tools.

Slavic Languages

Bulgarian retrieval and the difficulties derived from its morphology are presented in Savoy (2007). The author worked on the collection which was made available during the 2005 and 2006 CLEF evaluation campaigns (Peters *et al.* 2006). As a Slavic language, Bulgarian has a rich morphology and includes the use of suffixes to denote the definite article (*the*). Using 99 queries, the study experimented with stopword removal, stemming and light decompounding. Specific queries which cause precision to increase or to drop, alternative stemmers and stopword lists were examined. In general, their experiments showed that the combination of the above IR techniques increases the mean average precision across all the submitted

queries. Similar experiments and results for the Hungarian language are reported in Savoy (2008).

Polish supporting search engines were examined in Sroka (2000). Polish versions of English language search engines and homegrown Polish search engines were assessed. The searching capability and retrieval performance were considered. Main emphasis was given to the precision criterion, which was based on relevance judgments for the first 10 matches from each search engine. Of the five search engines evaluated, Polski Infoseek and Onet.pl had the best precision scores, and Polski Infoseek turned out to be the fastest Web search engine. In a more recent paper the effectiveness of retrieval for Polish queries with Diacritics is tested (Chorós 2005). In the Polish language there are several local characters with diacritic symbols, such as: *ćçńółśżź*. Chorós submitted a number of queries with and without the diacritics in major and local search engines and found that search engines retrieve different results when diacritics are not used. It is also mentioned that several users do not type the letters with diacritics in their Web queries or in pages. So search engines should take this into consideration to increase their precision. This was also suggested for Greek (Lazarinis 2007a).

German

German Web searching is reviewed in Lewandowski (2008a). The purpose of this study was to compare five major Web search engines (Google, Yahoo, MSN, Ask.com and Seekport) for their retrieval effectiveness, taking into account not only the results but also the descriptions of the results. The study employs real user provided queries and the results are judged by the persons posing the original queries. The overall conclusion is that the major search engines exhibit comparable performance in terms of accuracy among the top ten results. In Lewandowski (2008b) the ability of major search engines to distinguish between German and English-language documents is tested. 50 queries, using words common in German and in English, were posed to the engines. The advanced

search option of language restriction was used, once in German and once in English. The top 20 results per engine in each language were investigated. The study found that while none of the search engines faces problems in providing results in the language of the interface that is used, both Google and MSN face problems when the results are restricted to a foreign language. The searching behaviour of German users is also investigated in Machill *et al.* (2004).

Greek

Greek is a morphologically complex language based on a non-Latin alphabet. Further, diacritics are used with lower case vowels. Lazarinis (2007a) studied a number of factors which influence Greek Web retrieval. With the aid of real users and a number of user provided queries the capabilities of search engines were evaluated. Initially, users indicated that they prefer search engines with simple interfaces and localized services. The user provided queries contained diacritics and words of low significance (e.g. prepositions). For example, the query *ευρωπαϊκό δικαστήριο* (european court) contains two different types of diacritics, i.e. accents and diaeresis. The queries were submitted in different forms in seven international search engines (AlltheWeb, AltaVista, AOL, Ask, Google, MSN and Yahoo) and four native Greek engines (Anazitisis.gr, In.gr, Pathfinder.gr and Robby.gr). The results showed that although international search engines have a higher coverage than the domestic ones, they fail to handle uniformly queries in upper or lower case and queries with or without the diacritics. Further, it seemed that most search engines treated stopwords as important search terms. This was supported by the fact that their manual removal from user queries improved precision. Another finding is the inability of some search engines to retrieve any result with Greek queries. Similar factors, i.e. existence of stopwords, diacritics and upper or lower query versions, influence Web image retrieval and product searching using Greek queries (Lazarinis 2008a; Lazarinis 2007d).

Efthimiadis *et al.* (2008; 2009-current issue) used a different approach to evaluate the effectiveness of search engines in Greek. They conducted a series of homepage finding evaluations using 309 Greek navigational queries for known Greek organizations. The queries submitted to five global search engines (A9, AltaVista, Google, MSN Search and Yahoo) and five Greek engines (Anazitisi, Ano-Kato, Phantis, Trinity and Visto) in 2004 and 2006. Searches were performed using the Greek, and English or transliterated name of each organization. The analysis showed that the global search engines ignored the characteristics of the Greek language, hence treating semantically similar Greek queries differently. Despite this finding the performance of the global search engines outperforms that of the Greek engines.

Italian

Italian Web searching was studied in (Lazarinis, 2008c). Using an approach similar to the above Greek experiments, the effectiveness of native Italian (Virgilio.it and Libero.it) and international search engines (Google, Yahoo, MSN, AOL, ASK, AlltheWeb) was tested with a small number of Italian queries. Some Italian terms had diacritics and some were in plural. Although the international search engines handled Italian queries better than the Greek scripts it was shown that the native Italian engines were inferior to the international ones. In addition, both local and major search engines handled inflectional term variations as different terms, producing quite different results. Stemming could improve Italian Web searching (Monz & de Rijke 2002).

Iberian Languages

Guzman *et al.* (2009-current issue) study the use of the Web as a Spanish linguistic resource for text classification. They retrieved their initial data using Google and they were able to develop a self-training method, which makes use of the Web as a lexical support resource.

A Portuguese question answering searching system is presented in (Amaral *et al.* 2004). The goal of their search engine is to find a sentence in the collections that answers a question in natural language. Although the aim of this study is different from standard Web searching, issues related to the morphology of the query terms and the inflectional morphology of the Portuguese language are discussed. These issues cause the precision of the Portuguese question answering tool to decrease.

EusBila is a search service for Basque that relies on the APIs of search engines, undertaking a lemma-based and language-filtered search by means of morphological query expansion and language-filtering words (Leturia *et al.* 2007). The authors argue that using standard search engines to query in a minority and agglutinative language like Basque is unsatisfactory in terms of precision. EusBila uses the indexes of other search engines and limits the results in Basque by using language-filtering words.

Asian Languages

Bitirim *et al.* (2002) evaluated Turkish search engines with respect to precision, normalized recall, coverage and novelty ratios. Seventeen queries were defined for the Arabul, Arama, Netbul and Superonline search engines. These queries were carefully selected to assess the capability of a search engine for handling broad or narrow topic subjects, exclusion of particular information, identifying and indexing Turkish characters, retrieval of hub/authority pages, stemming of Turkish words and the correct interpretation of Boolean operators. It was found that the morphology of the queries and the inflections of the query terms influence the retrieval of Turkish Web pages. Similar tests including worldwide search engines such as Google and Yahoo are repeated in a more recent study (Demirci *et al.* 2007). The results show that although the major search engines perform better than the local search engines they still need a lot of improvements in order to handle the Turkish queries effectively.

Chinese retrieval is studied in Moukdad and Cui (2005). This research article explored the characteristics of the Chinese language and how queries in this language are handled by different search engines. Queries were entered in two major search engines (Google and AlltheWeb) and two search engines developed for Chinese (Sohu and Baidu). Criteria such as handling word segmentation, number of retrieved documents, and correct display and identification of Chinese characters were used to examine how the search engines handled the queries. The results showed that the performance of the two major search engines was inferior compared with the search engines developed for Chinese. The capabilities of three search engines in Chinese queries were evaluated in (Long *et al.* 2007). 270 participants evaluated 655 queries extracted from a query log, focusing on the relevance of the top ten results. The paper does not address any issue related to how queries are typed or handled by the search engines. The participants' assessments focus on the relevance of the top 10 results. This short paper concludes that the accuracy of the three search engines is similar but overall the relevance is a subjective issue depending on the goals of the users.

In Tongchim *et al.* (2007) web search performance was evaluated using queries written in Thai. The queries were submitted to SiamGURU, Sansarn, Google, Yahoo, MSN, AltaVista, AlltheWeb and a number of meta-search engines. The first two search engines are Thai-focused engines. The authors used 56 Thai queries in their evaluation. The length of the queries ranged between one and four words. The binary (relevant/not relevant) judgments were performed by seven judges. The aim of the study was to test the accuracy of the produced results across different search engines. Google had the highest mean average precision for Thai queries. Unfortunately, this study does not analyse the factors which reduce precision for the other search engines.

The positive effects of stemming and spelling correction on retrieving Malay texts are discussed in Bakar *et al.* (2000) and Saian and Ku-Mahamud (2004).

Classification of Amharic texts compiled from the Web is discussed in Asker *et al.* (2009-current issue). The effect of operations like stemming or part-of-speech tagging on text classification was also investigated. The experiments indicated that stemming plays a less important role than expected for text classification performance for a highly inflected language like Amharic. In addition, written languages that do not use a standardised representation require a lot of time and effort in order to create a uniformly represented text corpus.

Evaluation of Multiple Languages

Bar-Ilan and Gutman (2005) explored how search engines respond to queries in four non-English languages: Russian, French, Hungarian and Hebrew. For each of the languages they searched using three global search engines (AltaVista, FAST and Google), and in local search engines. The local engines were the Russian Yandex, Rambler and Aport; the French Voila, AOL France and La Toile de Quebec; the Hungarian Origo-vizsla, Startlap and Heureka; and the Hebrew Morfix and Walla. For each of the four languages the authors developed queries that emphasized specific linguistic characteristics of that language. The top ten results of each search were evaluated not for relevance, but for whether the exact word form or a morphological variant of the query was retrieved. They found that the search engines ignored the special language characteristics and did not handle diacritics well.

The effect of multilingual queries for homepage finding is studied in Blanco and Lioma (2009-current issue), where the aim of their Web retrieval system is to return a single document, namely the homepage described in the query. The authors submitted 766 queries in 35 different languages in four major search engines (Ask, Google, Microsoft Live Search and Yahoo). The queries were names of football teams which compete in their national premier league in 2008 according to FIFA, and which also have a homepage on the Web. Teams without

a homepage were excluded. Queries were submitted in the script of each language and in Latin script. The authors found that in some cases Latinized versions of the queries retrieved better results. A possible explanation for this finding is the nature of the queries which refer to names of teams which are often written in Latinized forms in international games and thus in newspaper articles commenting these games. The study also reports that the local domain search engine (e.g., google.es for Spanish) has better average precision than the global .com interface.

In Lazarinis and Efthimiadis (2008) the effectiveness of Google and Yahoo in image retrieval is studied. Five one-word queries, such as *dog, flower*, were submitted to Google and Yahoo in eleven languages (Croatian, English, French, German, Greek, Italian, Norwegian, Polish, Russian, Spanish and Turkish). The queries were submitted in various modes, e.g., upper and lower case, singular and plural, and with and without diacritics. One of the main findings of this study is that the localized search interfaces help disambiguate the query and retrieve results relevant to the language of the query.

The information-seeking behaviour of non-English Web users is studied in Berendt and Kralisch, (2009-current issue). The study established that content and link creation behaviour leads to an underrepresentation of non-English languages in the Web. It also provides evidence that link-following behaviour leads to an under-utilization of non-English content. Based on a number of experiments the authors conclude the general desirability of more translation and better language tools by non-English users. Another conclusion is that the behaviour of non-English searchers is influenced by the English language skills of the non-English users.

Query log analysis

Logs of Web queries are good resources that record user search histories and can be utilized in reaching useful conclusions about the user behaviour during

searching. By analyzing query logs, useful statistics about the search topics and the morphology of the queries could be obtained. The derived information could be then used in improving search engines. Silverstein *et al.* (1999) were the first to analyze a large Web query log of AltaVista. This study provides statistics about the topics, the number of terms per query, and the duplication of queries. A similar approach followed in (Spink *et al.* 2001). These studies, however, do not take into consideration the natural language of the submitted queries in their statistical analysis which we believe is an important factor in understanding the query formation process and the user searching patterns.

Jansen and Spink (2005) studied the trends in Web searching characteristics by European users of the AlltheWeb search engine. The study reports statistics about the query length, the session duration and the language of the users. The average query length and the mean number of queries per session differ among the European languages. However, the number of queries per language varies significantly and no stable conclusions could be reached for some languages.

A Greek query log of 2.5 million queries from AltaVista is analyzed in Efthimiadis (2008). The majority are one- or two-word queries (53%), three word queries account for 21%, and four-word queries for 12.6%. Phrase searching using double quotes accounted for 14.84% while a small number of queries contained logical operators (0.9%). Most queries were expressed in a Latin form (93.4%) rather than Greek (6.56%). This finding is not surprising given the inadequate treatment of the Greek language by search engines and that one in three navigational queries fail (Efthimiadis *et al.*, 2009-current issue).

In another analysis of a smaller Greek query log, the user search strings of a number of academic departments, the queries were grammatically and morphologically analyzed (Lazarinis 2008b). This log was studied mainly in terms of the following factors: query length, capitalization, accentuation, lemmatized form and the existence of stopwords. The statistical analysis showed

that the majority of the approximately 5,000 Greek queries contain 2 or 3 terms and that, although queries appear mostly in lower case, a significant number of queries are typed in upper case or in title case. Queries are usually in non lemmatized form and about 1 out of 4 of the queries contain words of low discriminatory value. Diacritics are often omitted and a number of typographic errors were identified. Further, a number of queries were Latinized.

Three months of search query logs of Timway, a Chinese search engine based in Hong Kong, were collected and analyzed in (Chau *et al.* 2007). Metrics on sessions, queries, search topics, and character usage are reported. Their analysis suggests that some characteristics identified in the search log, such as search topics and the mean number of queries per sessions, are similar to those in English search engines; however, other characteristics, such as the use of operators in query formulation, are significantly different. The analysis also shows that only a very small number of unique Chinese characters are used in search queries. In all the Chinese search queries, there are only 7303 unique Chinese characters in total, which is much lower than the number of unique terms in English queries. One reason is that Chinese characters are generally bounded to a closed class. New characters are seldom created.

Baeza-Yates *et al.* (2007) studied the characteristics of search queries on mobile phones in Japan, comparing them with previous results of generic Japanese queries and mobile search queries in the USA. The study analyzed the queries based also on their scripts (Kanji, Hiragana, Katakana and Romaji). This preliminary study confirms the results on the most popular topics of the previous studies with English queries but also indicates that the query length and the topics may vary according to the used script.

Brill *et al.* (2001) implemented the mining of Katakana-English terms pairs and phrases along with their English counterparts from non-aligned monolingual Web

search engine query logs. The data obtained could be used for enhancing Web searching by appropriately expanding user queries.

The query logs of a major Korean Web search engine, NAVER, were analyzed to track the information-seeking behaviour of Korean Web users (Parka 2005). These transaction logs include more than 40 million queries collected over 1 week. The results of this study show that users behave in a simple way: they type in short queries with a few query terms, seldom use advanced features, and view few result pages. Based on the statistics provided in the study, Korean users submit mostly one-word queries and the mean queries per session for NAVER users is lower than those of users from other regions. In several occasions users use stopwords in their queries which influences the performance of the search engines.

Lewandowski (2006) investigated the topics of searches in German web search engines and the query types used. Based on the query types identified by Broder (2002) and the classification of search topics developed in Spink *et al.* (2001), 1500 queries from German search engines Fireball, Seekport and Metager are assigned to a topic category and to a query type. The findings of the study corroborate the results of previous studies on the analysis of English Web queries.

In this section we have presented research which shows that there is significant variation between queries formulated by searchers from different countries. Dimensions of variation include query length, morphology, inflections and script. In one case correlation between the query script and search topic is noted. Such variation and correlation must be taken into account if search engines are to fully exploit the very limited information about the user needs expressed in the query.

Conclusions

Over 100 papers related to non-English Web retrieval have been reviewed. The papers examined and the techniques presented concern various European, Asian,

and African languages, and different scripts, such as Latin, Greek, Cyrillic and Chinese ideograms. The papers also discuss problems related to the dialects spoken in some countries, e.g., Indonesia.

Initially the paper discussed the issues arising during the indexing of non-English texts. Language identification and encoding is a problem especially in Asian languages, where several ideograms and dialects exist. Evidence from the header and the text of a webpage should be combined in order to efficiently cope with this problem.

Text segmentation is more complicated in non-English languages as word boundaries are harder to identify in many languages; Latin characters are mixed with non-Latin letters; compound words are more common; local punctuation marks are used interchangeably with English punctuation marks and numerical metrics. Language specific techniques have been proposed for boosting the performance of text processing and indexing tools. Language independent approaches based on character n-grams instead of words for indexing and retrieval purposes are promising and generic. Compound words occur very frequently in languages like German, Dutch and Swedish. Effective decompounding is very important in the indexing and retrieval process. It was shown that when queries are expanded with the constituents of compounds recall improves.

Segmentation of Chinese texts is a more difficult task than for European languages as there are hundreds of ideograms and in addition the boundaries and grouping of ideograms are not clear. This influences stopword identification as well. Stopword elimination from user queries improves the accuracy of the retrieved document set. However, their effect in Web searching has not been extensively studied for most of the non-English languages. Stopword lists are not available for many non-English languages and their construction is more demanding because of the absence of free linguistic resources for researchers to work with.

One of the main issues emerging from the non-English Web searching studies is that some search engines do not handle semantically identical queries in a uniform way. For example, queries in upper case are handled differently from lower case queries and the omission of diacritics from query terms produces different results compared to the same query terms with diacritics. These problems influence also Web image retrieval and the searching of e-commerce sites. In highly inflectional languages user queries can be expressed in various declensions and with different endings. Search engines do not always take these language intricacies into account and this compromises search result quality. In some cases, e.g., Chinese, the local search engines, outperform the major international players while in smaller countries, e.g., Greece, local search engines have a low coverage of the local Web pages possibly masking this effect. In many studies it was shown that standardised IR techniques such as stemming, stopword removal and effective decompounding increase the accuracy of the Web retrieval tools.

The page encoding influences the indexing and thus the retrieval of Web texts. For Asian languages, where several encodings and dialects exist, this is a major problem. Other Web searching studies discussed the English character of the internal characteristics of Web pages, such as links and filenames and how these influence the retrieval of documents. Several words have different meanings across European languages and therefore effective ways of disambiguating user queries are needed. Localized interfaces are important for search engines to increase their user bases internationally.

The query log analysis studies indicate that there are certain differences among English and non-English queries. The patterns of query length and morphological expression differ between languages. Mixed English and non-English queries are often issued by users and they need to be handled effectively. However, the studies on non-English query log analysis are limited and therefore more research

is needed in order to understand the ways in which query expression varies between different languages.

Overall, it can be argued that processing and searching of non-English text pose additional difficulties not faced in English texts. Several techniques have been proposed and tested for non-English Web searching, some of which have been proven quite successful. Nonetheless, much work remains to be done for search engines in order to reach the same levels of effectiveness in non-English language queries as with English language searching.

Although this work has shown that language is a major factor to be taken into account in web search, research community has started to study the influence of cultural factors in the use of web searchers. This is the case of Mandl & de la Cruz (2009), where the influence of the culture of the user, in addition to the language, in the process of evaluation of web searches is studied.

Finally, we should make notice a great problem common to many non-English languages, the lack of freely available resources for IR research and evaluation. Fortunately, following the example of the Text REtrieval Conference (TREC, http://trec.nist.gov) in the case of English, initiatives such as the Cross-Language Evaluation Forum (CLEF, http://www.clef-campaign.org) for European languages, the NII Test Collection for IR Systems Project (NTCIR, http://research.nii.ac.jp/ntcir/) for Asian languages and the Forum for Information Retrieval Evaluation (FIRE, http://www.isical.ac.in/~clia/) in the case of Indian languages, have emerged during the last years.[1] They allow researchers to have access to such resources by providing them with document collections (mainly formed by news articles), a test suite suite of queries –both in the required language–, and the corresponding set of relevance judgements. However, although

[1] During the years, their activity has also extended to other language out of their main sphere, as in the case of Arabic in TREC or Persian in CLEF. Moreover, they have also extended their scope to more specialized tasks such as speech retrieval or geographical retrieval, and to other information processing tasks such as question answering.

these initiatives have proven invaluable for non-English IR research, the resource availability is still more limited than that for English, and there are many languages without any available evaluation corpora.

Acknowledgements

The authors wish to thank Prof. Thomas Mandl and Prof. Arjen P. de Vries for their helpful comments and suggestions. The authors also acknowledge the assistance of Jennifer Rohan in compiling part of the bibliography and the University of Washington Information School for resources.

Prof. Vilares' research has been partially funded by the Spanish Government and FEDER (through project HUM2007-66607-C04-03) and the Galician Autonomous Government (through the "*Galician Network for NLP and IR*", *"Human Resources Program"* grants, and projects PGIDIT07SIN005206PR, INCITE08E1R104022ES and PGIDIT05PXIC30501PN).

References

Ahlgren, P., & Kekäläinen, J. (2006). Swedish full text retrieval: Effectiveness of different combinations of indexing strategies with query terms. *Information Retrieval*, 9(6), 681–697. DOI 10.1007/s10791-006-9009-1.

Ahmad, F., Yusoff, M., & Sembok, T.M.T. (1996). Experiments with a stemming algorithm for Malay words. *Journal of the American Society for Information Science*, 47(12), 909-18.

Aho, A.V., Sethi, R., & Ullman, J.D. (1986). *Compilers: Principles, Techniques and Tools*. Addison-Wesley.

Airio, E. (2006). Word normalization and decompounding in mono- and bilingual IR. *Information Retrieval*, 9(3), 249-271. DOI: 10.1007/s10791-006-0884-2.

Alemayehu, N., & Willett, P. (2003). The effectiveness of stemming for information retrieval in Amharic. *Program: electronic library and information systems*, 37(4), 254-259.

Al-Kharashi, I.A., & Evens, M.W. (1994). Comparing words, stems and roots as index terms in an Arabic information retrieval system. *Journal of the American Society for Information Science*, 45(8), 548-60.

Amaral, C., Laurent, D., Martins, A., Mendes, A., & Pinto, C. (2004). Design & Implementation of a Semantic Search Engine for Portuguese. *Proceedings of the Fourth Conference on Language Resources and Evaluation*.

Arampatzis, A., van der Weide, Th.P., van Bommel, P., & Koster, C.H.A. (2000). Linguistically motivated information retrieval. In *Encyclopedia of Library and Information Science*, vol. 69, pp. 201-222. Marcel Dekker.

Artemenko, O., Mandl, T., Shramko, M., Womser-Hacker, C. (2006). Evaluation of a language identification system for mono- and multilingual text documents. Proceedings of the 2006 ACM symposium on Applied computing, pp. 859-860. ACM. DOI: http://doi.acm.org/10.1145/1141277.1141473.

Asker, L., Argaw, A., Gambäck, B., Asfeha, S. E., & Habte L. N. (2009-current issue). Classifying Amharic Webnews, Information Retrieval

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* Reading, MA: Addison Wesley, ACM Press.

Baeza-Yates, R., Dupret, G., & Velasco, J. (2007). A study of mobile search queries in japan. In: Amitay, E., Murray, C. G., Teevan, J. (Eds.). *Query Log Analysis: Social*

And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007).

Bakar, Z. A., Sembok, T. M., & Yusoff, M. (2000). An Evaluation of Retrieval Effectiveness Using Spelling-Correction and String-Similarity Matching Methods on Malay Texts. *Journal of the American Society for Information Science*, 51(8), 691–706.

Barcala, F.M., Vilares, J., Alonso, M.A., Graña J., & Vilares, M. (2002). Tokenization and Proper Noun Recognition for Information Retrieval. In *Proceedings of Thirteen International Workshop on Database and Expert Systems Applications*, pp. 246-250.

Bar-Ilan, J., & Gutman, T. (2005). How do search engines respond to some non-English queries? *Journal of Information Science*, 31(1), 13-28. DOI: 10.1177/0165551505049255.

Berendt, B., & Kralisch, A. (2009-current issue). A user-centric approach to identifying best deployment strategies for language tools: The impact of content and access language on Web user behaviour and attitudes. Information Retrieval

Bitirim, Y., Tonta, Y., & Sever, H. (2002). Information Retrieval Effectiveness of Turkish Search Engines. In *Advances in Information Systems*, vol. 2457 of *Lecture Notes in Computer Science*, pp. 93-103.

Blanco, R, & Lioma, C. (2009-current issue). Mixed Monolingual Homepage Finding in 35 Languages: The Role of Language Script and Search Domain. Information Retrieval

Blanco, R., & Barreiro, A. (2007). Static Pruning of Terms in Inverted Files. In *Advances in Information Retrieval*, vol. 4425 of *Lecture Notes in Computer Science*, pp. 64-75.

Braschler, M., & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval? *Journal of Information Retrieval*, 7(3-4), 291-316. DOI: 10.1023/B:INRT.0000011208.60754.a1.

Brill, E., Kacmarcik, G., & Brockett, C. (2001). Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Log. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pp. 393-399.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2), 3-10.

Cavnar, W.B., & Trenkle, J.M. (1994). N-Gram-Based Text Categorization. *3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-176. Las Vegas, Nevada, USA.

Chau, M., Fang, X., & Yang, C. (2007). Web Searching in Chinese: A study of a Search Engine in Hong Kong. *Journal of the American Society for Information Science*, 58(7), 1044-1054. DOI: 10.1002/asi.20592.

Chen, A., & Gey, F. (2002). Building an Arabic Stemmer for Information Retrieval. In *TREC 2002*. Gaithersburg: NIST, pp. 631-639.

Chen, K., & Liu, S. (1992). Word identification for Mandarin Chinese sentences. *Proceedings of the 14th conference on Computational linguistics*, 101 - 107, 10.3115/992066.992085.

Chorós K. (2005). Testing the Effectiveness of Retrieval to Queries Using Polish Words with Diacritics. In *AWIC 2005*, vol. 3528 of *Lecture Notes in Artificial Intelligence*, pp. 101–106.

Darwish, K., & Oard, D. (2007). Adapting Morphology for Arabic Information Retrieval, In *Adapting Morphology for Arabic Information Retrieval*, pp. 245-262, Springer, 978-1-4020-6045-8.

De Vries, A.P. (2001). A Poor Man's Approach to CLEF. In *Cross-Language Information Retrieval and Evaluation*, vol. 2069 of *Lecture Notes in Computer Science*, pp. 149-155.

Demirci, R., Kismir, V., & Bitirim, Y. (2007). An Evaluation of Popular Search Engines on Finding Turkish Documents. *2nd IEEE International Conference on Internet and Web Applications and Services (ICIW'07)*. DOI: http://doi.ieeecomputersociety.org/10.1109/ICIW.2007.15.

Di Nunzio, G.M., Ferro, N., Melucci, M., & Orio, N. (2004). Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In *Comparative Evaluation of Multilingual Information Access Systems*, vol. 3237 of *Lecture Notes in Computer Science*, pp. 220-235.

Dunning, T. (1994). Statistical Identification of Language. *Technical Report MCCS*, 94-273, New Mexico State University, New Mexico.

Efthimiadis, E.N. (2008). How do Greeks search the web?: a query log analysis study. In *Proceeding of the 2nd ACM Workshop on Improving Non English Web Searching* (Napa Valley, California, USA, October 30 - 30, 2008). iNEWS '08. ACM, New York, NY, 81-84. DOI= http://doi.acm.org/10.1145/1460027.1460041

Efthimiadis, E.N., Malevris, N., Kousaridas, A., Lepeniotou, A., & Loutas, N. (2008). An Evaluation of How Search Engines Respond to Greek Language Queries. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. DOI: http://doi.ieeecomputersociety.org/10.1109/HICSS.2008.52.

Efthimiadis, E.N., Malevris, N., Kousaridas, A., Lepeniotou, A., & Loutas, N., (2009-current issue). Non-English Web Search: An Evaluation of Indexing and Searching the Greek Web. Information Retrieval

Eguchi, K., & Croft, B. (2009-current issue). Query Structuring and Expansion with Two-stage Term Dependence for Japanese Web Retrieval. Information Retrieval

Ekmekçioglu, Ç, & Willett, P. (2000). Effectiveness of stemming for Turkish text retrieval. *Program*, 34(2), 195–200.

Figuerola, C.G., Gómez, R., Zazo-Rodríguez, A.F., & Alonso-Berrocal, J.L. (2001). Stemming in Spanish: A first approach to its impact on information retrieval. In *Working Notes for the CLEF 2001 Workshop*.

Foo, S. & Li, H. (2004). Chinese word segmentation and its effect on information retrieval. *Information Processing & Management*, 40 (1), 161-190.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

Frakes, W., & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice Hall.

Goldsmith, J., & Reutter, T. (1999). Automatic Collection and Analysis of German Compounds. In Busa, F., Mani, I., Saint-Dizier, P. (Eds.), *The Computational Treatment of Nominals: Proceedings of the Workshop COLING-ACL '98*, Montreal, pp. 61-69.

Gonzalez, M., de Lima, V.L.S., & de Lima, J.V. (2005). Binary Lexical Relations for Text Representation in Information Retrieval. In *Natural Language Processing and Information Systems*, vol. 3513 of *Lecture Notes in Computer Science*, pp. 21-31.

Graña, J., Barcala, F.M., & Vilares, J. (2002). Formal Methods of Tokenization for Part-of-Speech Tagging. In *Computational Linguistics and Intelligent Text Processing*, vol. 2276 of *Lecture Notes in Computer Science*, pp. 240-249.

Graña, J., Chappelier, J.C., & Vilares, M. (2001). Integrating external dictionaries into stochastic part-of-speech taggers. In *Proceedings of EuroConference Recent Advances in Natural Language Processing (RANLP 2001)*, pp. 122–128.

Grefenstette, G. (1995). Comparing two language identification schemes. *3rd International Conference on the Statistical Analysis of Textual Data (JADT'95)*, Rome, pp. 263-268.

Guzman, R., Montes-y-Gómez, M., Rosso, P., & Villaseñor-Pineda, L. (2009-current issue). Using the Spanish Web for Self-training Text Classification Tasks. Information Retrieval

Hammo, B. H. (2009-current issue). Towards Enhancing Retrieval Effectiveness of Search Engines for Diacritisized Arabic Documents. Information Retrieval

Harman, D. (1991). How effective is suffixing?. *Journal of the American Society for Information Science*, 42(1), 7–15.

Hedlund, T. (2002). Compounds in dictionary-based cross-language information retrieval. *Information Research*, 7(2). Available at http://InformationR.net/ir/7-2/paper128.html

Hollink, V., Kamps, J., Monz, C., & de Rijke, M. (2004). Monolingual Document Retrieval for European Languages. *Information Retrieval*, 7(1-2), 33-52. 10.1023/B:INRT.0000009439.19151.4c.

Hull, D. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70–84.

Jansen, B., & Spink, A. (2005). An analysis of web searching by European AlltheWeb.com users. *Information Processing and Management*, 41(2), 361-381. DOI: 10.1016/S0306-4573(03)00067-0.

Jurafsky, D., & Martin, J.H. (2000). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kalamboukis, T.Z. (1995). Suffix stripping with modern Greek. *Program*, 29(4), 313-321.

Lazarinis, F. (2007a). Web retrieval systems and the Greek language: Do they have an understanding?. *Journal of Information Science*, 33(5), 622-636. DOI: 10.1177/0165551506070739.

Lazarinis, F. (2007b). Engineering and utilizing a stopword list in Greek Web retrieval. *Journal of the American Society for Information Science and Technology*, 58(11), 1645-1652. DOI: 10.1002/asi.20648.

Lazarinis, F. (2007c). Lemmatization and stopword elimination in Greek Web searching. *ACM EATIS 2007,* ACM Digital Library. DOI: http://doi.acm.org/10.1145/1352694.1352757.

Lazarinis, F. (2007d). Evaluating the searching capabilities of e-commerce web sites in a non-English language: A Greek case study. *Online Information Review*, 31(6), 881-891. DOI: 10.1108/14684520710841829.

Lazarinis, F. (2008a). Improving concept based Web image retrieval by mixing semantically similar Greek queries. *Program: Electronic Library and Information Systems*, 42(1), 56-67. DOI: 10.1108/0033033081085159.

Lazarinis, F. (2008b). Retrieving non-Latin information in a Latin Web: the case of Greek, In M. Song, Y-F B. Wu (Eds.), *Handbook of Research on Text and Web Mining Technologies*, IDEA Publishing, pp. 530-545.

Lazarinis, F. and Efthimiadis, E. N. (2008). Measuring search engine quality in image queries in 10 non-English languages: an exploratory study. In *Proceeding of the 2nd ACM Workshop on Improving Non English Web Searching* (Napa Valley, California, USA, October 30 - 30, 2008). iNEWS '08. ACM, New York, NY, 89-92. DOI= http://doi.acm.org/10.1145/1460027.1460043.

Lazarinis, F., (2008c). Towards a model for evaluating web retrieval systems in non English queries. In: Coral Calero, Maria Angeles Moraga, Mario Piattini (eds), *Handbook of Research on Web Information Systems Quality,* Idea Group Inc., USA. pp. 510-527

Lazarinis, F., Vilares, J., & Tait, J. (2007). Improving non-English web searching. *ACM SIGIR Forum*, 41(2), 72-76.

Lazarinis, F., Efthimiadis, E.N., Vilares, J., & Tait, J. (2008). Improving non-English web searching (iNEWS08). Proceedings of ACM-*CIKM Workshop*.

Leturia I., Gurrutxaga, A., Areta, N., Alegria, I., & Ezeiza A. (2007). EusBila, a search service designed for the agglutinative nature of Basque, In F. Lazarinis, J. Vilares, J. Tait (Eds.), *Improving Non-English Web Searching (iNEWS07). SIGIR07 Workshop*, pp. 47-54.

Lewandowski, D. (2006). Query types and search topics of German Web search engine users. *Information Services and Use*, 26(4), 261-270.

Lewandowski, D. (2008a). The Retrieval Effectiveness of Web Search Engines: Considering Results Descriptions. *Journal of Documentation*, 64 (in press).

Lewandowski, D. (2008b). Problems with the use of Web search engines to find results in foreign languages. *Online Information Review*, 32(5), 668 – 672.

Lo, R.T.W., He, B., & Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. In *Proceedings of 5th Dutch-Belgian Information Retrieval Workshop (DIR'05)*.

Long, H., Lv, B., Zhao, T., & Liu, Y. (2007). Evaluate and Compare Chinese Internet Search Engines Based on Users' Experience. *Proceedings of IEEE Wireless Communications, Networking and Mobile Computing Conference (WiCom 2007)*, 6134 – 6137. DOI: 10.1109/WICOM.2007.1504.

Macdonald C., Lioma, C., & Ounis, I. (2007). Terrier takes on the non-English Web, In F. Lazarinis, J. Vilares, J. Tait (Eds.), *Improving Non-English Web Searching (iNEWS07). ACM SIGIR07 Workshop*, pp. 21-28.

Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2004). Navigating the Internet: A Study of German-Language Search Engines. *European Journal of Communication*, 19(3), 321–347. DOI: 10.1177/0267323104045258.

Makrehchi, M., & Kamel, M.S. (2008). Automatic Extraction of Domain-Specific Stopwords from Labeled Documents. In *Advances in Information Retrieval*, vol. 4956 of *Lecture Notes in Computer Science*, pp. 222-233.

Mandl, T., & de la Cruz, T. (2009). International differences in web page evaluation guidelines, *Int. J. Intercultural Information Management*, to appear

Martins, B., & Silva, M. J. (2005). Language identification in web pages. *In SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, New York, NY, USA, 2005. ACM Press, pp. 764–768.

McNamee, P., & Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2), 73-97.

Monz, C., & de Rijke, M. (2002). Shallow morphological analysis in monolingual retrieval for Dutch, German, and Italian. In *Accessing Multilingual Information Repositories*, vol. 2406 of *Lecture Notes in Computer Science*, pp. 262-277.

Moreau, F., Claveau, V. & Sébillot, P. (2007). Automatic Morphological Query Expansion Using Analogy-Based Machine Learning. In *Advances in Information Retrieval*, vol. 4425 of *Lecture Notes in Computer Science*, pp 222-233.

Moukdad, H. (2004). Lost in cyberspace: how do search engines handle Arabic queries? In: *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*, Winnipeg. Available at: www.cais-acsi.ca/proceedings/2004/moukdad_2004.pdf (accessed 31 July 2006).

Moukdad, H., & Cui, H. (2005). How do search engines handle Chinese queries? *Webology*, 2(3), article 17. Available at: http://www.Webology.ir/2005/v2n3/a17.html.

Otero, J., Vilares, J. & Vilares, M. (2008). Corrupted queries in Spanish text retrieval: error correction vs. n-grams. In *Workshop Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008): 2nd ACM Workshop on Improving Non-English Web Searching (iNEWS'08)*, pp. 39-46. ACM. DOI: http://doi.acm.org/10.1145/1460027.1460034.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. *In Proceedings of OSIR 2006*.

Palmer, D.D. (2000). Tokenisation and Sentence Segmentation, chapter 2. In R. Dale, H. Moisi and H. Somers (Eds.). *Handbook of Natural Language Processing*. Marcel Dekker.

Parka, S., Leeb, J., & Bae, H. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*, 27(2), 203-221. DOI: 10.1016/j.lisr.2005.01.013.

Peng, F., Ahmed, N., Li, X., & Lu, Y. (2007). Context sensitive stemming for web search. *Proceedings of the 30th ACM SIGIR conference*, pp. 639 – 646.

Peters, C., Gey, F. C., Gonzalo, J., Muller, H., Jones, G. J. F., Kluck, M., Magnini, B., & de Rijke, M. (2006). *Accessing Multilingual Information Repositories*, vol 4022 of *Lecture Notes in Computer Science*. Spinger-Verlag.

Pingali, P., Jagarlamudi, J., & Varma, V. (2006). WebKhoj: Indian language IR from multiple character encodings. *Proceedings of the 15th international conference on World Wide Web*, pp. 801 – 809.

Piskorski, J. Wieloch, K., & Sydow, M., (2009-current issue). On Knowledge-poor Methods for Person Name Matching and Lemmatization for Highly Inflectional Languages. Information Retrieval

Pohlmann, R., & Kraaij, W. (1997). The effect of syntactic phrase indexing on retrieval performance for Dutch texts. *Proceedings of RIAO 97*, pp. 176–187.

Popovic, M., & Willett, P. (1992). The effectiveness of stemming for natural language access to Slovene textual data, *Journal of the American Society for Information Science*, 43(5), 384-90.

Porter, M. (1980). An algorithm for Suffix Stripping. *Program*, 14(3), 130-137.

Robertson, S. E. (1969). The Parameter Description of Retrieval Systems: Overall Measures. *Journal of Documentation*, 25, 93-107.

Saian, R., & Ku-Mahamud, K. R. (2004). Searching Malay text using stemming algorithm, JICT, 3(2), 107-117. Available at http://jict.uum.edu.my.

Salton, G., & McGill, M., (1983). *Introduction to modern information retrieval*. McGraw Hill.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10), 944 – 952.

Savoy, J. (2003). Cross-Language Information Retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management*, 39, 75-115.

Savoy, J. (2007). Searching strategies for the Bulgarian language. *Information Retrieval*, 10(6), 509–529. DOI: 10.1007/s10791-007-9033-9.

Savoy, J. (2008). Searching strategies for the Hungarian language. *Information Processing and Management*, 44(1), 310–324. DOI: 10.1016/j.ipm.2007.01.022.

Sigurbjörnsson, B., Kamps, J., & de Rijke, M. (2006). EuroGOV: Engineering a Multilingual Web Corpus. In *Accessing Multilingual Information Repositories*, vol. 4022 of *Lecture Notes in Computer Science*, pp. 825-836.

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum*, 33(1), 6-12.

Solak, A., & Oflazer, K. (1993). Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing*, 8(3), 113-30.

Spink, A., Wolfram, D., Jansen, B. J. & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 52(3), 226-234.

Sroka, M. (2000). Web search engines for Polish information retrieval: questions of search capabilities and retrieval performance. *International Information & Library Research*, 32(2), 87–98.

Tomlinson, S. (2006a). Bulgarian and Hungarian Experiments with Hummingbird SearchServer at CLEF 2005. In *Accessing Multilingual Information Repositories*, vol. 4022 of *Lecture Notes in Computer Science*, pp. 194-203.

Tomlinson, S. (2006b). Danish and Greek Web Search Experiments with Hummingbird SearchServer at CLEF 2005. In *Accessing Multilingual Information Repositories*, vol. 4022 of *Lecture Notes in Computer Science*, pp. 846-855

Tongchim, S., Sornlertlamvanich, V., & Isahara, H. (2007). Improving Search Performance: A Lesson Learned from Evaluating Search Engines Using Thai Queries. IEICE TRANS. INF. & SYST., E90–D(10), 1557-1564. DOI: 10.1093/ietisy/e90–d.10.1557.

Tzoukermann, E., Klavans, J., & Jacquemin, C. (1997). Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsing. In *Proceedings of the 20th ACM SIGIR Conference (SIGIR'97)*.

Vega V., & Bressan, S. (2001). Indexing the Indonesian Web: Language Identification and Miscellaneous Issues. *10th International World Wide Web Conference*, http://www10.org/cdrom/posters/p1044/index.htm.

Vilares, J., Alonso, M.A., & Vilares, M. (2008). Extraction of Complex Index Terms in Non-English IR: A Shallow Parsing Based Approach. *Information Processing and Management*, 44(4), 1517-1537.

Vilares, J., Alonso, M.A., Ribadas, F.J., & Vilares, M. (2003). COLE experiments at CLEF 2002 Spanish monolingual track. In *Advances in Cross-Language Information Retrieval*, vol. 2785 of *Lecture Notes in Computer Science*, pp. 265-278.

Vilares, J., Cabrero, D., & Alonso, M.A. (2001). Applying Productive Derivational Morphology to Term Indexing of Spanish Texts. In *Computational Linguistics and Intelligent Text Processing*, vol. 2004 of *Lecture Notes in Computer Science*, pp. 336-348.

Vilares, M., Graña, J., & Alvariño, P. (1996). Finite-State Morphology and Formal Verification. *Journal of Natural Language Engineering*, 2(4), 303-304.

Xu, J., & Croft, W.B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, 16(1), 61-81.

Yang, C., Luk, J., Yung, S., & Yen, J. (2000). Combination and boundary detection approaches on Chinese indexing. *Journal of the American Society for Information Science*, 51(4), 340 – 351.

Zou, F., Wang, F. L., Deng, X., & Han, S. (2006). Automatic Identification of Chinese Stop Words. *Research on Computing Science: Special issue on Advances in Natural Language Processing*, 18, 151-162.