

# Do search engines understand Greek or user requests “sound Greek” to them?

Fotis Lazarinis

*Department of Technology Education & Digital Systems*

*University of Piraeus*

*80 Karaoli & Dimitriou, 185 34 Piraeus, Greece*

*lazarinf@teimes.gr*

## Abstract

*This paper presents the outcomes of initial Greek Web searching experimentation. The effects of localization support and standard Information Retrieval techniques such as term normalization, stopword removal and simple stemming are studied in international and local search engines. Finally, evaluation points and conclusions are discussed.*

## 1. Introduction

The Web has rapidly gained popularity and has become one of the most widely used services of the Internet. Its friendly interface and its hypermedia features attract a significant number of users. Finding information that satisfies a particular user need is one of the most common and important operations in the WWW. Data are dispensed in a measureless number of locations and so utilization of a search engine is necessary.

Although international search engines like Google and Yahoo are preferred over the local ones, as they employ better searching mechanisms and interfaces, they do not really value other spoken languages than English. Especially in languages like Greek which has inclinations and intonation, it seems that the majority of the international search engines have no internal (indexing) or external (interface) localization support. Thus the user has to devise alternative ways so as to discover the desired information and to adapt themselves to the search engine's interface.

This paper reports the results of initial experimentation in Greek Web searching. The effect of localization support, upper or lower case queries, stopword removal and simple stemming is studied and evaluation points are presented. The conclusions could be readily adapted to other spoken languages with similar characteristics to the Greek language.

## 2. Experimentation and evaluation

Interface simplicity and adaptation is maybe the most important issue which influences user satisfaction and acceptance of Web sites and thus search engines [1, 2].

User acceptance factor is obviously increased when a search engine changes the language and maybe its appearance to satisfy its diversified user basis. This is significant especially to novice users.

Stopword removal, stemming and capitalization or more generally normalization of index and query terms are amongst the oldest and most widely used IR techniques [3]. All academic systems support them. Commercial search engines, like Google, explicitly state that they remove stopwords, while capitalization support is easily inferred. Stemming seems to not be supported though. This may be due to the fact that WWW document collection is so huge and diverse that stemming would significantly increase recall and possibly reduce precision. However simple stemming, like final sigma removal which will be presented later in the paper, may play an important role when seeking information in the Web using Greek query terms.

These four issues were examined with respect to the Greek language. For conducting our assessment we used most of the predominately known worldwide .com search engines: Google, Yahoo, MSN, AOL, Ask, Altavista. The .com search engines were selected based on their popularity [4]. Also, for comparison reasons, we considered using some native Greek search engines: In ([www.in.gr](http://www.in.gr)), Pathfinder ([www.pathfinder.gr](http://www.pathfinder.gr)) and Phantis ([www.phantis.gr](http://www.phantis.gr)).

### 2.1. Interface issues

Ten users participated in the interface related experiment and they also constructed some sample queries for the subsequent experiments. Users had varying degrees of computer usage expertise. We needed end users with technical expertise and obviously increased demands over the utilization of web searchers. On the other hand we should measure the difficulties and listen to the people who have just been introduced to search engines. This combination of needs reflect real everyday needs of web “surfers”.

The following sub-issues extracted from a more complete evaluation study of user effort when searching the Greek Web space utilizing international search engines [5]. Here we extend (with more users and search

engines) and present only the issues connected with whether search engines really value other spoken languages than English, like Greek, or not.

**2.1.1. Localization support.** The first issue in our study was the importance of a localized interface. All the participants (100%) rated this feature as highly important as many users have basic or no knowledge of English. Although search engines have uncomplicated and minimalist interfaces their adaptation to the local language is essential as users could easily comprehend the available options.

From the .com ones only Google automatically detects local settings and adapts to Greek. Altavista allows manual selection of the presentation language with a limited number of language choices and setup instructions in English. Also if you select another language, search is automatically confined to this country's websites (this must be altered manually again).

**2.1.2. Searching capability.** In this task users were asked to search using queries with all terms in Greek. All search engines but AOL and Ask were capable of running the queries and retrieving possibly relevant documents. AOL pops-up a new Window when a user requests some information but it cannot correctly pass the Greek terms from the one window to the other. So no results are returned. However, when requests typed directly to the popped-up window then queries are run but presentation of the rank is problematic again.

Ask does not retrieve any results, meaning that indexing of Greek documents is not supported. For example zero documents retrieved in all five queries of section 2.2. For these reasons AOL and Ask left out of the subsequent tests.

**2.1.3. Output presentation.** An important point made by the participants is that some of the search engines rank English web pages first, although search requests were in Greek. For example in the query "Ολυμπιακοί αγώνες στην Αθήνα" (Olympic Games in Athens) Yahoo, MSN and Altavista ranked some English pages first. This depends on the internal indexing and ranking algorithm but it is one of the points that increase user effort because one has to scroll down to the list of pages to find the Greek ones.

## 2.2. Term normalization, Stemming, Stopwords

Trying to realize how term normalization, stemming and stopwords affect retrieval we run some sample queries. We used 5 queries (table 1) suggested by the participants of the previous test. They were typed in lower case sentence form with accent marks leaving the default options of each search engine. A modified version

of Recall and Precision [6] are used for comparing the results of the sample queries. Recall refers to the number of retrieved pages, as indicated by search engines, while precision (relevance) was measured in the first 10 results.

Table 1. Sample queries.

No	Queries in Greek	Queries in English
Q1	Μορφές ρύπανσης περιβάλλοντος	Environmental pollution forms
Q2	Εθνική πινακοθήκη Αθηνών	National Art Gallery of Athens
Q3	Προβλήματα υγείας από τα κινητά τηλέφωνα	Health problems caused by mobile phones
Q4	Συνέδριο πληροφορικής 2005	Informatics conference 2005
Q5	Τεστ για την πιστοποίηση των εκπαιδευτικών	Tests for educators' certification

Table 2 presents the number of recalled pages for each query. From table 2 we realize that In and Pathfinder share the same index and employ exactly the same ranking procedure. The result set was identical both in quantity and order. Their only difference was in output presentation. Altavista and Yahoo had almost the same number of results, ranked slightly differently though.

Table 2. Recall in lower case queries.

	Q1	Q2	Q 3	Q4	Q5
Google	867	3400	805	15500	252
Yahoo	820	933	527	11200	186
MSN	1357	1537	542	6486	272
Altavista	821	939	515	11400	191
In	251	343	67	689	49
Pathfinder	251	343	67	689	49
Phantis	33	63	22	88	6

In all cases the international search engines returned more results than the native Greek local engines. However, as seen in table 3, relevance of the first 10 results is almost identical in all cases, except Phantis, which maintains either a small index or employs a crude ranking algorithm. Query 4 retrieves so many results because it contains the number (year) 2005. So, documents which contain one of the terms and the number 2005 are retrieved, increasing recall significantly.

Table 3. Precision of the top 10 results.

	Q1	Q2	Q 3	Q4	Q5
Google	5	7	9	8	8
Yahoo	5	7	8	7	8
MSN	4	7	8	6	7
Altavista	5	7	8	7	8
In	5	7	8	6	8
Pathfinder	5	7	8	6	8
Phantis	2	2	2	1	0

We confined the relevance judgment to only the first ten results so to limit the required time and because the first ten results are those with the highest probability to be visited. Relevance was judged upon having visited and inspected each page. The web locations visited had to be from a different domain. So if two consecutive pages were on the same server only one of them was visited.

An interesting point to make is that although recall differs substantially among search engines precision is almost the same in all cases. Another point of attention is that the third query shows the maximum precision. This is because in this case terms are more normalized, compared to the other queries. This means that they are in the first singular or plural form which is the usual case in words appearing in headings or sub-headings. Consequently a better retrieval performance is exhibited. But, as we will see in section 2.2.3, it contains stopwords which when removed precision is positively affected and reaches 10/10.

**2.2.1. Term normalization.** We then re-run the same queries but this time in capital letters with no accent marks. Recall (table 4) was dramatically diminished in most of the worldwide search enabling sites while it was left unaffected in two of the three domestic ones (In and Pathfinder). Precision was negatively affected as well (table 5), compared to results presented in table 3.

Table 4. Recall in upper case queries.

	Q1	Q2	Q 3	Q4	Q5
Google	22	3400	41	673	252
Yahoo	18	229	2	116	8
MSN	10	233	2	379	10
Altavista	18	239	2	117	9
In	251	343	67	689	49
Pathfinder	251	343	67	689	49
Phantis	4	63	3	14	6

These observations are true for Yahoo, MSN and Altavista. Google and Phantis exhibit a somehow unusual behavior. In queries 2 and 5 Google and Phantis retrieve the same number of documents in the same order and have the same precision therefore. Upper case queries 1, 3 and 4 recall only a few documents compared to the equivalent lower case queries. Correlation between results is low and precision differs.

Trying to understand what triggers this inconsistency we concluded that it relates to the final sigma existing in some terms of queries 1, 3 and 4. The Greek capital sigma is  $\Sigma$  but lower case sigma is  $\sigma$  when it appears inside a word and  $\varsigma$  at the end of the word. Phantis presents the normalized form of the query along with the result set. Indeed it turns out that words ending in capital  $\Sigma$  are transformed to words with the wrong form of sigma, e.g. “ΜΟΡΦΕΣ” (forms) should change to “μορφες” but it

changes to “μορφεσ”.

These observations are at least worrying. What would happen if a searcher were to choose to search only in capital letters or without accent marks? Their quest would simply fail in most of the cases leading novice users to stop their search. In English search there is no differentiation between capital and lower letters. The result sets are identical in both cases so user effort and required “user Web intelligence” is unquestionably less.

Table 5. Precision of the top 10 results.

	Q1	Q2	Q 3	Q4	Q5
Google	4	7	3	10	8
Yahoo	3	8	0	5	7
MSN	3	6	0	7	7
Altavista	3	8	0	5	7
In	5	7	8	6	8
Pathfinder	5	7	8	6	8
Phantis	0	2	0	0	0

Wrapping up this experiment one can argue that in Greek Web searching the same query should be run both in lower and in capital letters, so as to improve the performance of the search. Sites where there are no accent marks or contain intonation errors will not be retrieved unless variations of the query terms are used. Greek search engines are superior at this point and make information hunting easier and more effective. From the international search engines only Google has recognized these differences and try to improve its searching mechanism.

**2.2.2. Stemming.** Another factor that influences searching relates to the suffixes of the user request words. For example the phrases “Εθνική πινακοθήκη Αθηνών” or “Εθνική πινακοθήκη Αθήνας” or “Εθνική πινακοθήκη Αθήνα” all mean “National Art Gallery of Athens”. So while they are different they describe exactly the same information need. Each variation retrieves quite different number of pages. For example Google returned 3400, 722 and 5420 web pages respectively. Precision is different in these three cases as well, and correlation between results is less than 50% in the first ten results.

One could argue that such a difference is rational and acceptable as the queries differ. If we consider these queries solely from a technical point of view then this argument is right. However if the information needed is in the center of the discussion then these subtle differences in queries which merely differ in one ending should have recalled the same web pages. Stemming is an important feature of retrieval systems [3] (p. 167) and its application should be at least studied in spoken languages which have conjugations of nouns and verbs, like in Greek. Google partially supports conjugation of English verbs.

**2.2.3. Stopwords.** Google and other international search engines remove English stopwords so as to not influence retrieval. For instance users are informed that the word *of* is an ordinary term and is not used in the query “National Art Gallery of Athens”. Removal of stopwords [3] (p. 167) is an essential part of typical retrieval systems.

We re-run, in Google, queries 3 and 5 removing the ordinary words. Queries were in lower case and with accent marks so results should be compared with tables 2 and 3. Query 3 recalled 839 pages and precision equals 10 in the first 10 ranked documents. Similarly for the fifth query Google retrieved 275 documents and precision raised from 8 (table 2) to 10. As realized, recall was left unaffected but precision increased by 10% and by 20% respectively. This means that ranking is affected when stopwords are removed. However more intense tests are required to construct a stopword list and to see how retrieval is affected by Greek stopwords

## 4. Conclusions

This paper presents a study regarding utilization of search engines using Greek terms. The issues inspected were the localization support of international search engines and the effect of stopword removal, capitalization and stemming of query terms. Our analysis participants identified as highly important the adaptation of search engines to local settings. Most of the international search engines do not automatically adapt their interface to other spoken languages than English and some of them do not even support other spoken languages. At least these are true for Greek.

In order to get an estimate of the internal features of search engines that support Greek, we run some sample queries. International search engines recalled more pages than the local ones and they had a small positive difference in precision as well. However they are case sensitive, apart from Google, hindering retrieval of web pages which contain the query terms in a slightly different form to the requested one. Even if the first letter of a word is a capital letter the results will be different than when the word is typed entirely in lower case.

Endings and stopwords are not removed automatically, thus affecting negatively recall of relevant pages. Stopwords are removed from English queries making information hunting easier, looking at it from a user’s perspective. Terms are not stemmed though even in English. However in a language with inclinations, like Greek, simple stemming seems to play an important role in retrieval assisting end users. In any case more intensive tests are needed to realize how endings, stopwords and capitalization affect retrieval.

Trying to answer the question posed in the article’s title it can be definitely argued that international search enabling sites do not value the Greek language and possibly other languages with unusual alphabets. Google is the only one which differs than the others and seems to be in a process of adapting to and assimilating the additional characteristics.

## 5. References

- [1] J. Nielsen, R. Molich, C. Snyder, S. Farrel, *Search: 29 Design Guidelines for Usable Search* <http://www.nngroup.com/reports/ecommerce/search.html>, 2000.
- [2] Carpineto, C. et al., “Evaluating search features in public administration websites”, *Euroweb2001 Conference*, 2001, 167-184.
- [3] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison Wesley, ACM Press, New York, 1999.
- [4] D. Sullivan, *Nielsen NetRatings: Search Engine Ratings* <http://searchenginewatch.com/reports/article.php/2156451>, 2005.
- [5] Lazarinis, F., “Evaluating user effort in Greek web searching”, *10<sup>th</sup> PanHellenic Conference in Informatics*, University of Thessaly, Volos, Greece, 2005 (to appear)
- [6] S. E. Robertson, “The Parameter Description of Retrieval Systems: Overall Measures”, *Journal of Documentation*, 1969, 25, 93-107.