Journal of the American Society for Information Science and Technology

Measuring the value of health query translation: An analysis by user language proficiency

Journal:	*Journal of the American Society for Information Science and Technology*
Manuscript ID:	JASIST-2012-05-0233.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Lopes, Carla; University of Porto, Informatics Engineering Ribeiro, Cristina; University of Porto/INESC-Porto, Informatics Engineering
Keywords:	

SCHOLARONE™
Manuscripts
For Review

Measuring the value of health query translation: An analysis by user language proficiency

Carla Teixeira Lopes[1*]
Cristina Ribeiro[1,2]

{ctl,mcr}@fe.up.pt

[1]DEI, Faculdade de Engenharia, Universidade do Porto
[2]INESC TEC
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

For Peer Review

Abstract

English is by far the most used language on the Web. In some domains the existence of fewer content in the users' native language may not be problematic and even help to cope with the information overload. Yet, in domains like health, where information quality is critical, a larger quantity of information may mean an easier access to higher quality content. Query translation may be a good strategy to access content in other languages but the presence of medical terms in health queries makes the translation process more difficult, even for users with very good language proficiencies. In this study we evaluate how translating a health query affects users with different language proficiencies. We chose English as the non-native language because it is a widely spoken language and it is the most used language on the Web. Our findings suggest that non-English speaking users having at least elementary English proficiency can benefit from a system that suggests English alternatives for their queries or automatically retrieves English content from a non-English query. This awareness of the user profile results in higher precision, more accurate medical knowledge and access to high-quality content. Moreover, the suggestions of English translated queries may also trigger new health search strategies.

Keywords: query formulation, personalization, health, consumer health information, medical vocabulary translation, user study

*Corresponding author.

1 Introduction

The use of the Web to search for health information by patients, relatives and friends is growing. A 2009's study revealed that 61% of the American adults and 83% of the Internet users do it (Fox & Jones, 2009). A more recent survey found that, in 2010, this figure rose to 88% of the US online population, the largest year-over-year increase (Petrock, 2010).

An obstacle that users commonly face in consumer health retrieval is the lack of content in their native language (Cline & Haynes, 2001). In 2000, Grefenstette and Nioche estimated the ratio of web content in several non-English European languages in relation to English content. Through these values, lower than 7% in all languages, we can see that English was in 2000, by far, the most used language on the Web, compared with European languages. Presently, the reality has not changed much and German, the second most popular language, is only 11.5% of the English content (W3Techs, 2012). Health is no exception and, in this domain, English is even considered the *lingua franca* (W. Hersh, 2008).

In the health domain, where information quality is critical, a larger quantity of information may mean an easier access to higher quality content. We think this can be explored by web search engines to provide native speakers of languages with less presence on the Web a better service in health searches. This could be done through the suggestion of alternative English queries or, with less user involvement, through the inclusion of English content in the set of retrieved documents. About these strategies we emphasize that, while the translation of commonsense terms may be simple to some users, on health matters translations are often not obvious, frequently requiring domain knowledge to translate medical terms to the correct terms in the target language. Moreover, since English documents are not accessible to every user having another primary language, the adopted strategy has to be personalized to users' English proficiency.

In this study we evaluate the effect of translating query terms from the users' native language to the English language, in users with different English proficiencies. Users' native language is Portuguese, a language with a smaller Web presence and a large number of native speakers. In 2000, Grefenstette and Nioche estimated that Portuguese content was only 2.4% of the English Web content. Nowadays, this proportion has raised but it is still only 3.8% (W3Techs, 2012). This accentuates the importance of promoting access to English content in portuguese health searches. On the other hand, Portuguese is one of the most spoken languages in the world with about 178 million native speakers (Lewis, 2009) meaning that our results can be generalized to a substantial number of users.

The research presented in this paper is different from previous work in several ways. First, unlike previous studies, it considers user characteristics, namely his language proficiency, while studying the impact of a query processing technique. Second, the evaluation is focused not only on users' relevance assessments but also on documents' type, language, comprehension and readability, users' motivational relevance and, more importantly, on the quality of the medical knowledge that emerges from the search session. As defined by Saracevic

For Peer Review

(1996), the motivational relevance relates the users goals and motivations with the information objects and it is expressed by the users feeling of success and his satisfaction. Third, it is one of the first works exploring the impact of query language translations in consumer health information retrieval.

The remainder of this paper is structured as follows. In Section 2 we do a literature review on cross-language health Information Retrieval (IR). We describe our study in Section 3. Sections 4, 5 and 6 report our findings and Section 7 discusses the results and presents their implications. We conclude in Section 8.

2 Cross-language Health IR

In Cross-Language Information Retrieval (CLIR) users can formulate a query in their most fluent language and retrieve documents in different languages. This can be accomplished through query or document translation. In the healthcare domain, the amount of work on CLIR has been small and most of the work focuses on the development of multilingual resources to assist it (W. Hersh, 2008).

As we will point out, previous studies are focused on cross-language health IR but ignore the diversity of users and their inherent characteristics and specificities. The majority of the works explore the multilingual characteristics of the Unified Medical Language System (UMLS) Metathesaurus. Only the works from Pirkola (1998), Rosemblat, Gemoets, Browne, and Tse (2003) do not use this aggregation of knowledge sources maintained by the U.S. National Library of Medicine. Pirkola studied the effects of query structure and three query translation methods using a general and a medical dictionary. The evaluation of the methods used TREC's health related topics and showed that structured queries translated with medical and general dictionaries are almost as good as the original English queries. Rosemblat et al. compared query and document translation for CLIR using machine translation and a subset of queries from the ClinicalTrials.gov website, concluding that query translation outperforms document translation in terms of retrieval performance.

To internationalize SAPHIRE, a system that extracts concepts from documents and queries, W. R. Hersh and Donohoe (1998) used the six languages available in the 1998 Metathesaurus after mapping the text in the documents to the concepts in the thesaurus. Its performance was evaluated on German terms and showed that additional work was still necessary to handle plural and suffix variants. Another experiment with German was made by Volk et al. (2002). The authors annotated documents and queries with linguistic information that included the identification of medical terms and semantic relations between them. This study's results showed that linguistic processing is essential to a good performance in the German language. Moreover, authors concluded that semantic information increases performance in retrieval.

Eichmann, Ruiz, and Srinivasan (1998) used the UMLS to translate Spanish and French queries into English using several strategies (full match, partial

For Peer Review

match, dictionary based and simple addition of Spanish/French query words). To analyze the retrieval effectiveness of the translated queries, the authors used OHSUMED and concluded that, for Spanish, the results are similar to the ones reported on the CLIR literature and, for French, the results are worse. A French/English CLIR system was also proposed by Tran, Garcelon, Burgun, and Le Beux (2004) to support the retrieval of English documents with French queries. Authors used a method to translate queries that mixes an hybrid machine translation method and a translation method based on the UMLS. Using Google and PubMed to predict the accuracy of their translations, authors showed that the hybrid method is better than the machine translation and thesaurus-based methods alone. Also in French and more focused on consumer health information retrieval, is the work from Névéol, Pereira, Soualmia, Thirion, and Darmoni (2006). Here, the UMLS is used to translate the lay terms of a French medical catalogue to MedlinePlus English topics. Authors suggest that this can be explored in the future to translate patient queries.

Lu, Lin, Chan, and Chen (2005) have worked on Chinese-English health CLIR. Using Web-based term translation, they proposed a semi-automatic approach to construct a Chinese-English MeSH.

The works described in this section evidence the lack of studies exploring the impact of translating queries in health information retrieval. This reality is even more extreme in consumer health information retrieval. Moreover, our study is the first to detail the analysis of the effect of query translation according to characteristics of users and documents.

3 Case Study

In this section we discuss the methodology followed in our research. We start by exposing our research questions and then describe the experiment settings.

3.1 Research questions

One major research question drove our research and two secondary ones emerged during the study design. The prime question being investigated in this study is:

What is the impact of translating a health query to English, in users with different levels of English proficiency?

The impact will be analyzed in six perspectives: documents' characteristics; precision; medical accuracy; documents' comprehension; documents' readability and motivational relevance. In the first perspective, the analysis will only be done generally, that is, without considering users' English proficiency because this characteristic does not affect documents' characteristics.

The two other research questions are secondary because we can envision other experiment settings better suited to explore them. However, we feel the

For Peer Review

current study allows a superficial analysis that can help raise research hypothesis for future studies. These questions are:

1. Does the access to English content affect users' query reformulation behavior? Is this similar in all English proficiency levels?
2. Is it possible to predict English proficiency through users' search habits?

3.2 Experiment Settings

To accomplish our research goals we conducted a laboratory user study with the following settings.

3.2.1 Information situations

We defined 8 health information situations that act as the platform against which relevance is judged. For each of them we also defined 4 search queries, 2 in English and 2 in Portuguese, the users' native language.

The information situations were defined based on questions submitted to the health category of the Yahoo! Answers service. To select the questions from the Yahoo! Answers we had the following considerations. Since most of the health Web searches are about diseases (Fox, 2006), we decided to focus on questions about treatments to a symptom/disease. Moreover, we also guaranteed that each query had at least 30 results in both search systems used in this study. Every information situation is associated with only one clinical question type (treatment) to minimize the influence of this variable on the experiment. The defined health information situations are:

1. About 3 days ago, I started having a burning feeling every time I urinated. How should I treat this?
2. For the past 5 days my head has been very itchy and I don't have lice. What can I do to stop the itching?
3. I have high uric acid (8.0 mg/dL) with reference units 3.6 - 7.7. How can I lower my uric acid level?
4. I am suffering with an inflammation on my lips and mouth area for more than a year. I have difficulties eating. What can I do to treat it?
5. My father got bit by a dog and is in the hospital with a bone infection. How is this treated?
6. I frequently get heartburn even when I stay away from spicy stuff. What can I do to prevent it?
7. I have been noticing lots of hair coming out from my head. Usually I only comb my hair once a day. What can I do to stop losing my hair?

For Peer Review

8. I'm on the computer all day so I type a lot and use the mouse. My right pointing finger is starting to give me some joint pain. How I can treat my finger?

Information situations were initially formulated in English and afterwards translated to Portuguese by the researchers. They were communicated to users in Portuguese. Queries were built concatenating the symptom or disease with the word 'treatment'. The terms' translation between languages was supported by the Multilingual Glossary of technical and popular medical terms in nine European Languages (Stichele, 1995).

3.2.2 Retrieval systems

We have used Google as a *black-box* search engine with two different collections. The first is associated with Google entire index and the second is the set of pages indexed by Google that are HONcode certified. To simplify, we consider that these two different collections lead to what we call two retrieval systems. The HONcode certification is proposed by the Health On the Net Foundation (HON) to help assess the accuracy of health content and the credibility of the publishers. We have used the Google custom search built by the HON to restrict the Google collection to HONcode certified sites. Currently, this collection contains more than 1 million pages and 52% of the sites are in English (Baujard, Boyer, & Geissbühler, 2010). For each query, we automatically collected the top-30 results from each retrieval system. To reduce the risk of Google learning from the previous submitted queries, we ensured that returned links were never clicked. Additionally, to prevent changes in the search engine or in the HON collection, we submitted all queries within a very short time span.

In a previous study (Lopes & Ribeiro, 2011) we compared the performance of the two systems mentioned above. We found that the system that works with Google entire index, including certified and non-certified documents, has a better performance in all aspects but medical accuracy. Users assess relevance higher in this system, understand better its documents, feel more satisfied after the search sessions and the documents it retrieves are less difficult to read. We also found that the medical accuracy of the knowledge acquired after a search session is in risk if users do not understand documents or if the session has a few documents with unreliable information.

3.2.3 Assessment tasks

A query run on a retrieval system leads to an assessment task that a user can execute. Since we defined 8 information situations with 4 queries each, and we use 2 retrieval systems, a total of 64 different assessment tasks exist in our user study.

Each user was assigned a set of 8 different assessment tasks. In the assignment of tasks to users we applied a Latin-square like procedure so that all users assessed the relevance: (1) of all information situations, but only once each; (2) of queries of both languages the same number of times; and (3) in all the

For Peer Review

retrieval systems the same number of times. We have also permuted the order of assessment tasks to avoid possible bias of relevance assessments owing to human behavior. We also guaranteed that each iteration of relevance assessments contained, in the same number of times, (4) queries of both languages and (5) tasks in both retrieval systems. In Table 1 we present the tasks assigned to a subset of 16 users.

INSERT TABLE 1 HERE

3.2.4 Search Procedure

Users started answering a quiz designed to evaluate their English proficiency with questions of a test available on the Web[1]. Users were asked to answer, in less than 20 minutes, 26 multiple-choice questions, 8 questions from the *English Grammar I* category, 8 from the *English Grammar II*, 5 from the *English Vocabulary* and 5 from the *English Reading comprehension*.

Then, they answered a questionnaire that intends to collect user characteristics like age, gender, self-evaluation of health condition, web search habits, health search habits and the queries they would use for the information need triggered by the information situation. After this questionnaire, users enrolled in a sequence of 8 assessment tasks in which users judged the top-30 URL collected by the researchers for each task. Afterwards, users had to answer a post-search questionnaire.

For each URL, the user had to indicate the type of the document (webpage, pdf, ppt, doc or other), the language of the document (Portuguese, English or other), the relevance of the document to the information situation considering the user's own context, and the extent to which the document was understood. Users were allowed to follow links to the internal pages of the URL's site if they felt it made sense. Situations in which this was expected include pages with scientific papers' abstracts in which the access to the full-paper was only one-click away or pages in which content was deliberately separated in several pages with access through a 'previous-next' menu. Users were instructed to report the URL where they followed hyperlinks.

Relevance and comprehension were assessed in a 3-value scale to convey more realism to the experiment (Borlund, 2003). Since we wanted the relevance judgments to represent the value of the information objects for each particular user, we instructed them to judge relevance in accordance to the definition of situational relevance. According to Saracevic (1996), situational relevance is "the relation between the task at hand and the retrieved documents, being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty". To assess relevance, the three options pertaining the usefulness of the URL to the resolution of the problem, were 'not relevant', 'partially relevant' and 'totally relevant', denoted by 0, 1 and 2, respectively. For comprehension, the three options were 'I did

[1] `http://www.transparent.com/learn-english/proficiency-test.html` (Archived by WebCite at `http://www.webcitation.org/5ym7JFqw1`)

For Peer Review

not understand the document's content', 'I partially understood the document's content' and 'I understood the document', denoted by 0, 1 and 2, respectively.

Users were also instructed to report the situations in which there was an error loading the URL and the situations in which the page loaded but it had no content (e.g.: restricted access). In the post-search questionnaire users were asked to (1) evaluate the search task's completion status, (2) indicate two additional queries for the information need triggered by the information situation, and (3) indicate treatments for the condition mentioned in the information situation. The first question in this questionnaire was used to evaluate the motivational relevance.

3.2.5 Readability assessment

Documents' readability was computed using the Simple Measure of Gobbledygook (SMOG) metric in Equation 1. Since this metric has been recommended as a measure of readability in consumer-oriented healthcare documents (Fitzsimmons, Michael, Hulley, & Scott, 2010), we also chose it in our analysis.

$$SMOG = 1.043\sqrt{30\frac{\#polysyllables}{\#sentences}} + 3.1291 \quad (1)$$

The computation of the readability metric was performed in three stages. We started by extracting the main content of the documents, excluding components like menus, advertising, footers and headers. Then, we excluded the HTML tags from the document generated in the first phase to obtain a text document with the main content of the original one. This document is the input of the third stage that computes SMOG.

We could not compute the SMOG metric in 8.7% of the URL for one of the following reasons: error loading the original URL (0.8%), restricted access to content in the original document (0.4%), main content with no text (6.5%) and errors during the extraction of the main content (1%).

Although the use of SMOG in languages like Portuguese lacks statistical validation, we still decided to use it in Portuguese documents due to its applicability to health content and to the lack of a validated tool to measure Portuguese readability.

3.2.6 Medical accuracy assessment

Session's medical accuracy was obtained through the proportion of health certified web pages in the session and through a medical evaluation of users' answers. In the post-search questionnaire, users had to write an answer to the information situation that triggered the assessment task. Answers were evaluated by a medical doctor in relation to the correct and incorrect content they possessed. The answer's correctness was evaluated in a scale of 0 (inappropriate answer) to 2 (appropriate answer). The 1 was used for answers with *some value*. In terms of answer's incorrectness, user's answer was classified with 0 (all or almost all content is incorrect), 1 (some incorrect content) or 2 (no incorrect content).

For Peer Review

To measure the reliability of the ratings, a second medical doctor evaluated 30% of the answers and we estimated the inter-rater reliability through the weighted Cohen's Kappa, an adaptation of Cohen's Kappa to ordinal scales that treats disagreements differently. The measured weighted Cohen's Kappa, with squared weights, for the correctness ratings is 0.68 (95% CI: [0.54, 0.77]), indicating a substantial agreement. For the incorrectness ratings, this measure is 0.7 (95% CI: [0.48, 0.84]), also pointing a substantial agreement. These inter-rater reliability results assure the quality of the initial ratings.

From the answer's correctness and incorrectness we computed a third variable to which we called *medical accuracy* that corresponds to their sum. The *medical accuracy* varies therefore between 0 (lowest accuracy) and 4 (highest accuracy).

3.2.7 Summary of context features

In Table 1 we summarize context features used in this study. We group features into categories and, for each of them, we present its definition, measure scale and data collection methods. All the features have already been discussed with greater detail in the previous sections.

3.2.8 Users

Forty undergraduate students participated in this study (25 females; 15 males) with a mean age of 22.25 years (SD = 6.42). Although 4 students have non-Portuguese nationalities, all of them have Portuguese as their native language. The evaluation of the English proficiency quiz was done in a 0 to 100 scale and students' average classification was 73.94 (SD=18.54). Hierarchical clustering was used to identify low English proficiency (EP1) (n=8), elementary English proficiency (EP2) (n=21), and good English proficiency (EP3) (n=11) groups.

In a scale of 1 (Not healthy) to 5 (Very healthy), 77.5% answered 4 or 5 revealing a sample of healthy users. The mean number of years users have been searching the Web is 8.55 (SD = 2.17), most of the users (60%) do more than one search per day (5 in a 1-5 scale) and 70% say they find what they want almost all the time (4 in a 1-5 scale).

A small proportion of users (20%) has never conducted a health search on the Web. The majority of the remaining users say they perform one health search per month (50%). Although the mode and median of the health search frequency decreases as health status improves, which might lead to the conclusion that less healthy people search more about health issues on the Web, the difference of medians of health search frequency between health status levels is not significant (KW $\chi^2(2) = 3.1$, p=0.22).

In health searches, users feel less successful than in general searches, being mostly divided between "I sometimes find what I am looking for" (3 in a 1-5 scale) - 41% and "I frequently find what I am looking for" (4 in a 1-5 scale) - 47%. A large number of users do their health searches always (63%) or almost always (31%) in Portuguese. The use of the English language to express health

For Peer Review

Table 1: Summary of context features used in this study.

Category	Context feature	Definition	Scale	Collection method
User	English Proficiency	Users' English skills.	Ordinal: low, elementary and good English proficiency.	Assessed through an English proficiency test graded from 0 to 100. Grouped later through hierarchical clustering.
User & Document	Relevance	It relates the task and the retrieved documents, "being inferred by criteria like usefulness in decision making, appropriateness of information in resolution of a problem and reduction of uncertainty" (Saracevic, 1996).	Ordinal: from 0 (not relevant) to 2 (totally relevant).	Users' judgment pertaining the usefulness of the document to the resolution of the problem. Part of their assessment task.
	Comprehension	Users' understanding of the document.	Ordinal: from 0 (not understood) to 2 (totally understood).	Users' judgment. Part of their assessment task.
Document	Readability	The degree to which the text contained in the document is easily read.	Rational.	Automatically computed through the Simple Measure of Gobbledygook readability measure.
	Type of document	File type of the main content found in the URL being assessed.	Nominal: webpage, pdf, ppt, doc or other.	Identified by users and manually validated by authors when inconsistencies were found.
	HONcode certification	Is the document certified by the Health on the Net Foundation?	Nominal: yes or no.	Positive if it is in the set of retrieved documents of both systems. Automatic extraction.
User & Task	Answer's medical accuracy	The degree to which the answer that users give after each search task contains the adequate quantity of correct information and no incorrect information.	Varies between 0 (least accurate) and 4 (most accurate).	Computed from the medical evaluation of users' answers in terms of their correct and incorrect contents.
	Motivational relevance	It relates the users goals and motivations with the information objects. It is expressed by the users feeling of success and his satisfaction (Saracevic, 1996).	Ordinal scale of 1 (totally disagree) to 5 (totally agree).	Obtained through users' agreement with the following claim "I believe I succeeded in this search task" after the search task.

For Peer Review

queries is less consensual: 22% answered they never use it, 34% almost never and 31% expressed they use it sometimes. A large majority of the users never use languages other than Portuguese and English (88%).

In an open question, users were asked about their difficulties when performing health searches on the Web. Two of the major issues identified in the set of answers are the small amount of documents in Portuguese (8%) and the difficulties of translating medical terms to English (4%).

4 Query translation effects

The impact of queries' language translation is analyzed by documents' characteristics, including their comprehension and readability, by the medical accuracy of the answers and by users' motivational relevance. In the analysis that follows we consider the users' assessments made in both retrieval systems. We will use a * to mark significant results at $\alpha = 0.05$ and a ** to mark significant results at $\alpha = 0.01$.

4.1 Documents' characteristics

We manually checked all the URL in which there were discrepancies in users' assessments regarding the document type, its language and the cases in which users reported errors on HTTP loading or on access to content.

Portuguese queries led to more HTTP errors (1.4% of all URL retrieved through queries in this language) than English queries (0.2%). This difference is statistically significant at $\alpha = 0.01$ ($\chi^2(1) = 5.7$). In terms of URL without access to content, the proportion was similar in both languages (0.4%).

English queries returned 100% of documents in the English language while Portuguese queries returned 93% of Portuguese documents and 7% of Spanish documents. Spanish documents were retrieved due to the similarity of some terms between the two languages. They were retrieved in the queries *disúria tratamento* and *hiperuricemia tratamento*. The Spanish translations of these queries are *disuria tratamiento* and *hiperuricemia tratamiento* and the differences lay in the accentuation of one word and an additional character in the word *tratamento*. The first difference is often ignored by search engines and the second may be considered a typographical error.

In terms of document type, as expected, both languages retrieved mostly webpages (96.3% in English queries and 85.8% in Portuguese ones) and the proportion of webpages in English queries is significantly higher than in Portuguese ones ($\chi^2(1) = 54.2$, $p < 0.01$**). English queries retrieved less pdf documents (3.7% against 13.1%; $\chi^2(1) = 47.7$, $p < 0.01$**) and did not retrieve powerpoint (0.6% in Portuguese queries) or word documents (0.5% in Portuguese queries). This seems to indicate that Portuguese health web documents, when compared with English documents, have less documents built specifically for the dissemination of health information on the Web.

For Peer Review

4.2 Precision

We use Graded Average Precision (GAP) and Graded Precision (gP) measures to evaluate and compare precision. These measures were recently proposed by Robertson, Kanoulas, and Yilmaz (2010) and are based on a probabilistic model that generalizes precision and average precision to the case of multi-graded relevance. Behind these measures is a model in which the user has a binary view of relevance even when using a non-binary scale of relevance. Here, each point in the scale of relevance has a probability g_i of being the grade from which the user starts considering the documents relevant. Robertson et al. (2010) give details on the definition and calculation of these measures. Since the proponents of these measures found that an equally balanced g_1 and g_2, i.e., $g_1 = g_2 = 0.5$ made GAP more informative than normalized discounted cumulative gain (nDCG) and average precision, we will use these threshold probabilities.

We start by a global analysis that does not consider users' English proficiency. In Table 2 we present, for each precision measure and language, the mean and standard deviation. As can be seen, English queries have a higher precision and lower dispersion in all measures. As also shown in Table 2, all the differences are statistically significant at $\alpha = 0.01$.

Table 2: Mean and standard deviation of GAP, gP10 and gP5 by language. Statistical differences between languages in each measure.

	EN		PT		$\mu_{en} > \mu_{pt}$?	
	$\overline{x}$	s	$\overline{x}$	s	test value	p value
GAP	0.73	0.16	0.61	0.24	$t(281.3) = 5.3$	p=0.00**
$gP10$	0.77	0.26	0.58	0.33	$t(307) = 6.5$	p=0.00**
$gP5$	0.69	0.26	0.48	0.31	$t(301) = 5.5$	p=0.00**

An analysis by level of English proficiency shows that English queries have, consistently, a higher GAP in all levels of English proficiency. Furthermore, English queries also have lower dispersion in all English proficiency levels. In each level of English proficiency, we also tested whether the differences between languages were significant. In Table 3 we can see that, excluding one case, with the three measures used in this study, English queries had a significantly higher precision at $\alpha = 0.01$ in all levels of English proficiency. In the exception, this difference is significant at $\alpha = 0.05$.

Using the Kruskal-Wallis test, we also investigated, in each language, if there were significant differences in the mean GAP/gP5/gP10 between levels of English proficiency. In both languages, we did not find any significant differences between the three levels of proficiency.

For Peer Review

Table 3: GAP, gP5 and gP10 statistical differences in levels of English proficiency.

	Low EP	Elementary EP	Good EP
$\mu_{GAP_{en}} > \mu_{GAP_{pt}}$	t(53.74)=2.45	t(144.7)=3.82	t(78.72)=2.79
	p=0.01**	p=0.00**	p=0.00**
$\mu_{gP10_{en}} > \mu_{gP10_{pt}}$	t(59.12)=2.86	t(158.78)=4.83	t(84.51)=3.31
	p=0.00**	p=0.00**	p=0.00**
$\mu_{gP5_{en}} > \mu_{gP5_{pt}}$	t(58.22)=3.82	t(155.6)=3.61	t(82.65)=2.34
	p=0.00**	p=0.00**	p=0.01*

4.3 Documents' comprehension

On a general perspective, users considered the documents easy to read because the median of comprehension is 2 (in a scale of 0-2). The median is the same in English and Portuguese queries.

An analysis by English proficiency (Figure 1) shows that, in users with low English proficiency, the proportion of documents in which the users understood the document (2 in the comprehension scale) is higher in documents retrieved with Portuguese queries (41.6%) than in English queries (24.6%). In elementary English proficiency users, Portuguese queries still have an higher degree of comprehension, although the difference was smaller than in low proficiency users. In the two groups of users mentioned above, the median of comprehension of documents retrieved with English queries is significantly lower than with Portuguese queries at $\alpha = 0.01$ (W=387138 in low proficiency users and W=2872800 in elementary proficiency users). In good English proficiency users, English queries have slightly higher comprehension scores but this difference is not significant.

INSERT FIGURE 1 HERE

Since Portuguese queries retrieved both Portuguese and Spanish documents, we also analyzed how Spanish documents affected the reality described above. As mentioned before, Spanish documents have lower comprehension and, in accordance with this, we noticed that Spanish documents made the above differences get smaller in the low and elementary proficiency users and higher in the good proficiency group. In other words, if we made the above comparisons considering only English and Portuguese documents, in low and elementary English proficiency users, the comprehension difference between Portuguese and English documents would be clearer and, in good English proficiency users, less clear.

In English queries, users with a higher level of English proficiency tend to evaluate the documents' comprehension higher than users in lower levels. We have applied the Kruskal-Wallis test and verified there are statistically significant differences in documents' comprehension between the three groups of users at $\alpha = 0.01$ (KW $\chi^2(2) = 431.4$). Further analysis with the Mann-Whitney

For Peer Review

test and the Bonferroni correction indicates that differences are significant, at $\alpha = 0.01$, between all levels of proficiency (Table 4). This agrees with what was expected.

Table 4: Differences of comprehension between levels of English Proficiency. R_i is the median of the comprehension in the proficiency level i.

	$R_1 < R_2$	$R_1 < R_3$	$R_2 < R_3$
EN	$W = 770864.5$	$W = 362368.5$	$W = 1517136$
	$p < 0.01/3$**	$p < 0.01/3$**	$p < 0.01/3$**

4.4 Documents' readability

As previously explained, documents readability was automatically evaluated using the SMOG metric. Higher SMOG scores indicate documents with lower readability. We excluded from the readability analysis all documents to which we could not compute the SMOG readability metric; the assessments in which the user mentioned he could not access the content of the URL, although we did have access to it and did calculate the SMOG metric; and an English document that was a severe outlier with a SMOG of 78.9 while the mean SMOG for English documents is 6.2. After a manual analysis we verified that this severe outlier was a document containing a set of sentences with the word inflammation with no logical sense. It looked like a work in progress document that was, accidentally, put online.

Different languages have different characteristics and therefore are associated with readability metrics of different magnitude. We empirically expect Portuguese documents to have a higher SMOG than English documents. Since readability measures are language dependent and Portuguese queries retrieved both Portuguese and Spanish documents, we will do this analysis by document language and not by the query's language. Also, and for the same reasons, we will only be able to compare the SMOG metric in documents of the same language. Our results show that Portuguese documents have a higher average SMOG (10% trimmed mean of 8.22) than English ones (10% trimmed mean of 6.19). Since SMOG is based on the number of polysyllables (words of 3 or more syllables), this indicates that Portuguese has more polysyllables than English, a statistically significant difference at $\alpha = 0.01$ (t(8210.7)=-35.12).

An analysis of the mean SMOG distribution by documents' comprehension shows that, in Portuguese and English documents, the degree of comprehension increases as SMOG decreases, this is, as the text becomes simpler. This is shown on Table 5. As expected, this difference is stronger in Portuguese documents where all users have similar competencies while, in English, different proficiencies may affect users' comprehension even in the simplest document. We only detected statistical differences in the mean SMOG between levels of comprehension in the Portuguese documents (KW $\chi^2(2) = 99.8$, $p < 0.01$**).

For Peer Review

A pairwise comparison showed that only documents *fully understood* by users have a significantly lower mean SMOG than the other documents (Table 6).

Table 5: Mean and standard deviation of SMOG by language and level of comprehension.

	Not understood		Partially understood		Totally understood	
	$\overline{x}$	s	$\overline{x}$	s	$\overline{x}$	s
EN	6.57	2.38	6.44	2.72	6.38	2.58
PT	8.65	2.24	8.40	1.89	8.10	2.20

Table 6: SMOG differences between comprehension levels. S_i is the mean SMOG for comprehension level i.

	$S_0 > S_1$	$S_0 > S_2$	$S_1 > S_2$
PT	$W = 193599$	$W = 446717$	$W = 1991804$
	$p = 0.02$	$p < 0.01/3^{**}$	$p < 0.01/3^{**}$

In English documents, we did not find significant differences between levels of comprehension in each level of English proficiency.

In Figure 2, we present the distribution of the SMOG mean by level of relevance assessment and language of the document. As can be seen, Portuguese documents that are easier to read (lower SMOG) have higher relevance scores but this is only a tendency because the differences are not significant. Surprisingly, in English documents, there is an opposite trend that is significant (KW $\chi^2(2) = 28.4$, $p < 0.01^{**}$). A pairwise comparison showed that totally relevant documents have a higher SMOG mean than *non-relevant* ($W = 857138.5$, $p < 0.01/3^{**}$) or *partially relevant* ($W = 1060116$, $p < 0.01/3^{**}$) documents. This indicates that, in English documents, readability is not a major factor affecting relevance assessments.

INSERT FIGURE 2 HERE

As we suspect this reality may differ in users with different proficiencies, we checked if there were significant differences in the SMOG mean between documents with different relevances, in each level of English proficiency. In users with low English proficiency, we found no significant differences ($F(2) = 2.12$, $p = 0.12$). In users with different English proficiencies, where significant differences were found (KW $\chi^2(2) = 13.86$, $p < 0.01^{**}$ in elementary proficiency and KW $\chi^2(2) = 12.1$, $p < 0.01^{**}$ in good proficiency), we run a set of pairwise comparison tests (Table 7). In elementary and good English proficiency users, the trend depicted in Figure 2 still applies, i.e., documents classified as totally relevant have a higher SMOG mean than documents assessed as *non-relevant* or *partially relevant*. This allows us to update our previous conclusion to the

For Peer Review

following: in elementary and good English proficiency, English documents' readability is not a major factor affecting relevance assessments.

Table 7: Differences of the mean SMOG between relevance scores in different levels of English Proficiency in English documents.

English Proficiency	$Rel_0 \neq Rel_1$	$Rel_0 < Rel_2$	$Rel_1 < Rel_2$
Elementary	$W = 244990.5$	$W = 241057.5$	$W = 302445.5$
	$p = 0.42$ (>)	$p < 0.01/3$**	$p < 0.01/3$**
Good	$W = 75703.5$	$W = 67203.5$	$W = 68625.5$
	$p = 0.13$ (<)	$p < 0.01/3$**	$p < 0.05/3$**

4.5 Medical accuracy

Not considering English proficiency, we can see in Figure 3 and Figure 4 that the query language does not affect the distribution of *medical accuracy* and *answer correctness*. In fact, in terms of these two variables, the proportion of answers in each level of classification is very similar in both languages. In terms of incorrectness, as shown in Figure 5, Portuguese queries show a better performance with a larger number of answers with *no incorrect content*. Yet, there are no statistical differences between the medians of incorrect content in both languages ($W = 12124.5$, p=0.19). Comparing the answers in terms of correctness and incorrectness, we can see that it is more probable to find an answer with *no incorrect content* than with *appropriate content*.

INSERT FIGURE 3 HERE

INSERT FIGURE 4 AND FIGURE 5 SIDE BY SIDE HERE

Considering the English proficiency and the answer correctness, English queries had the best performance on elementary English proficiency users where only 35% had an inadequate answer in terms of medical content. This group is followed by the good English proficiency users where 52% of the answers were considered *with some value*. The proportion of inadequate answers in the low proficiency users was high (56% in Portuguese queries and 47% in English queries). Contrary to our expectations, in this group of users, English queries resulted in more adequate answers in terms of medical content than Portuguese queries. All these differences only show a general tendency since differences are not statistically significant, neither between levels of proficiency in each language, neither between languages in each level of proficiency.

In terms of answer's incorrectness, the distribution is similar in the 3 different levels of English proficiency and it is also similar to what was described previously when the English proficiency of the users was not being considered. In all levels of English proficiency, Portuguese queries result, in median, in answers with less incorrect content than English queries. Similarly to what happens in answer correctness, the best scenario, i.e., the scenario with less incorrect medi-

For Peer Review

cal content, happens with English queries in the elementary English proficiency users.

In the medical accuracy analysis, in low English proficiency users, Portuguese queries result in more accurate knowledge than English queries (Figure 6). This matches our initial expectations since their comprehension in English content may be limited by their English proficiency. English queries result in more accurate answers in elementary and good proficiency users with less dispersion in elementary proficiency. Like in answer correctness, these differences are not statistically significant.

INSERT FIGURE 6 HERE

We also noticed that English queries led to more HONcode certified pages (60.8%) than Portuguese ones (52.5%), a significant difference ($\chi^2(1) = 89.7$, p<0.01**). An analysis by document language instead of query language shows that 61.1% of the retrieved English documents are HONcode certified against 52% of the Portuguese documents, a statistically significant difference ($\chi^2(1) = 77.6$, p<0.01/3**). Note that the reported percentages are above 50% because half of the documents are retrieved by the HON retrieval system and are certainly HONcode certified. Also note that, since we used the Latin square procedure described in Section 3.2.3 and since an equal number of documents was assessed in each task, we assure no biases are introduced in this analysis. In fact, the number of documents assessed with English queries in one system is equal to the number of documents assessed with English queries in the other system that is equal to the number of documents assessed with Portuguese queries in either system.

4.6 Motivational relevance

To evaluate motivational relevance, we asked users to classify the search task completion status in a scale of 1 (completely unsatisfied) to 5 (completely satisfied) in the post-search questionnaire. Without considering the English proficiency of the users, we verified that the distributions of motivational relevance in English and Portuguese queries is very similar in terms of central tendency and dispersion. Both distributions have 4 as median. The main difference between them lays in the number of outliers in the first level of motivational relevance which is greater in Portuguese queries. In other words, with Portuguese queries, users feel *completely unsatisfied* more frequently. The analysis by English proficiency shows that, in low English proficiency users, the median level of search task's satisfaction is lower (3) than in users with higher proficiency (4 in both elementary and good proficiency levels). Only the former type of users is *completely unsatisfied* in tasks with English queries. Moreover, low English proficiency users are the only group that is never *completely satisfied* with English queries.

Hypothesis tests allowed us to conclude that differences between languages in each level of English proficiency are not significant. However, significant differences were found in the motivational relevance median between the three levels of English proficiency ($KW\chi^2(2) = 9.93$, p=0.00** in English queries

For Peer Review

and $KW\chi^2(2) = 6.41$, p=0.04* in Portuguese queries). To investigate the exact location of the differences, we did a pairwise comparison with the Bonferroni correction (Table 8). We found evidence to conclude, at different significance levels, that low proficiency users feel less satisfied than elementary and good English proficiency users with English queries.

Table 8: Differences of Motivational Relevance (MR) between levels of English Proficiency.

	$MR_1 < MR_2$	$MR_1 < MR_3$	$MR_2 \neq MR_3$
EN	$W = 884$	$Z = -2.34$	$W = 1928$
	$p < 0.01/3$**	$p < 0.05/3$*	$p = 0.66$
PT	$W = 968$	$Z = -1.41$	$W = 2056$
	$p < 0.05/3$*	$p = 0.08$	$p = 0.26$

5 Query formulation behavior

In the initial questionnaire, users were asked to introduce a query they would use for the information needs triggered by the information situations. Then, after each assessment task, users were asked to write two additional queries. Each of these queries was manually analyzed in terms of number of terms, language and the existence of syntactic errors.

Regardless of their English proficiency, all users formulated an initial query in Portuguese. The mean number of terms was 4.13 (SD = 1.83). In this initial query, we found 4 queries (1.25%) with syntactic errors.

When users have completed an assessment task having a Portuguese query, the subsequent queries were mainly in Portuguese. In users with low and good English proficiency, 100% of the subsequent queries were Portuguese and, in elementary proficiency users, this proportion downs to 99% in the 2nd query and 98% in the 3rd query.

After English tasks, as expected, users formulate more frequently English subsequent queries. In a global perspective, 3.8% of the English tasks had both subsequent queries in English and, in 56% of the tasks, one of the queries was in English. In Table 9 we present an analysis by English proficiency in which we cannot detect an association between English user proficiency and the use of English to formulate queries. With respect to errors, 3.9% of subsequent Portuguese queries had syntactic errors against 4.3% of the English queries.

6 English proficiency prediction

In this section we explore the associations between users' habits with respect to searches in English and their proficiency in this language. If a relationship is

For Peer Review

Table 9: Post-search queries in English after an English assessment task by user proficiency.

#English queries	Low EP	Elemen. EP	Good EP
2	6%	3%	4%
1	55%	57%	55%
0	39%	40%	41%

found, hypothesis can emerge and be studied in future studies using the search logs to predict language proficiency through past queries in English.

In the initial questionnaire, users were asked about how often they conducted their health searches in Portuguese and in English. Since only users that had previously conducted a health search have answered this question, we regret we did not ask about their general behavior instead of focusing only on health searches. This answer was given in a scale of 1 (never) to 5 (always) in each language. Excluding users that did not answer, almost all said they search in Portuguese *always* or *almost always*. In low and good English proficiency, 100% of the users chose these one of these two options and, in elementary proficiency users, 4.8% also answered *sometimes*.

When inquired about their rate of English health searches, a large proportion of answers were concentrated on the opposite side of the scale. The majority (62.5%) of the low English proficiency users said that *never* or *almost never* did it, while in elementary proficiency users this proportion downs to 56.3% and in the good English proficiency it downs even further to 50%. This shows a general tendency but there are no significant differences in the median between users of different English proficiencies (KW $\chi^2(2) = 1.2$, p=0.5).

7 Discussion and implications

Through a user study we have investigated, in several perspectives, the impact of translating queries to English, in health IR search tasks done by users with different English proficiencies. We have also analyzed query formulation and reformulation behavior according to users' English proficiency and the main language of the previously assessed documents. Finally, we have explored the existing relations between search habits and English proficiency.

As a result of the analysis of the documents in the search results, we can provide evidence that the quality of health web information is better in English than in Portuguese. First, English queries retrieve a smaller proportion of pages with enduring HTTP errors. Second, when compared with Portuguese queries, English queries retrieved a larger proportion of webpages (96.3%) and a smaller diversity of document types (only webpages and *pdf*). Since we expect content built for the dissemination of consumer health information to be in a webpage format, we conclude that English queries retrieve more documents built specif-

For Peer Review

ically for the dissemination of health information on the Web. Finally, we also found that English pages have a significantly higher proportion of HONcode certified pages. This shows that the use of an English query results in a higher probability of retrieving certified content. This result on document characteristics is independent of any user features and reinforces one of the assumptions of this study, namely that the larger quantity of information in English may mean an easier access to higher quality content.

In all levels of English proficiency, English queries have a significantly higher precision independently of the used measure. Since we also found that low proficiency users feel less satisfied than elementary and good English proficiency users with English queries, we conjecture that low proficiency users assess topic relevance, that is, "the relation between the querys topic and the documents topic" (Saracevic, 1996) instead of situational relevance. We think these users' proficiency is sufficient for them to identify if a document is about a certain topic but is not enough for them to understand the main message, what explains the lower satisfaction rates. This is also consistent with the results we have found in terms of medical accuracy.

Since all users have Portuguese as their native language we already expected users to evaluate the comprehension of Portuguese documents higher than the comprehension of English documents, and this was confirmed for the low and elementary proficiency users. In good proficiency users, the difference between both languages is not significant. This makes sense since their English proficiency is closer to the overall Portuguese proficiency. As expected, in English documents, the comprehension increases with the English proficiency level.

In terms of readability we found that comprehension increases as the documents become easier to read. Since we only detected significant comprehension differences by documents' readability in Portuguese documents, we suspect this is a factor that only comes into play if the language proficiency is guaranteed. The relation between relevance assessments and documents' readability shows that, in English documents and elementary and good English proficiency users, totally relevant documents have a significantly lower readability than partially and non-relevant documents. Either this means that, in these users and language, the presence of more scientific terminology boosts relevance or that readability is not one of the major factors determining the documents' relevance.

English queries tend to result in more accurate answers in elementary and good proficiency users, with less dispersion in elementary proficiency. In low English proficiency users, Portuguese queries tend to result in more accurate knowledge than English queries. This probably happens because their comprehension of English documents is limited by their proficiency.

None of the users formulated an initial query in English and few formulated the subsequent queries in this language. After performing English tasks, users formulate subsequent queries in English more frequently, regardless of their English proficiency level. This indicates that suggesting alternative English queries or even incorporating English documents in the answer set may also be a good way to trigger ideas on how to express the information need. The detection of more syntactic errors in English than in Portuguese may indicate

For Peer Review

that English terms for health concepts may not always be known or recalled. This adds value to suggestions of English translations of health concepts.

We believe English proficiency may be inferred from past search behaviors. In users with higher proficiency, we found an increased tendency to use English queries but, since differences are not significant, this is not conclusive. We think a specific study has to be done with this goal.

As shown in the previous paragraphs, English queries consistently have better results in users with at least elementary English proficiency. On the other hand, Portuguese queries behave better in users with low English proficiency, resulting in more accurate answers and a higher overall satisfaction with the search task. Together with the higher quality of English health web content, these findings confirmed our initial expectations and show that English content may and should be used to help users with other native languages and enough English proficiency. Existing Web search engines may use these conclusions to define personalized strategies that help users access English content when they formulate queries in their native language. These strategies may involve the user in the process or can be totally automatic. In the first case, alternative English queries can be suggested to the user who determines if he wants to use them or not. The alternative query may simply be a translation of the original query to English or may also include other variants of the English query through the inclusion/replacement of synonyms. Totally automatic strategies may be implemented through the inclusion of English content in the result set of the query in the users' native language. Merging the result sets of the original query and of its English translation may be a good strategy to do it. Although more important in the totally automatic strategies, personalization is also essential in the other strategies to avoid unnecessary distractions, users' waste of time and the overload of the search interface. Even though this study does not allow the generalization of these results to languages other than Portuguese, we believe the conclusions of this study would still apply in several languages with small presence on the Web and we would like to confirm it in future studies.

8 Conclusion and Future Work

English is by far the most used language on the Web and has, therefore, a larger proportion of English health content. We observed that English health content have a larger proportion of health-certified documents, are more suited to disseminate health information and are associated with less HTTP errors. For these reasons, we are convinced this can be explored to provide a better service to non-English speaking users. Difficulties expressed by users on health searches strengthen our conviction. Yet, we are aware that the approach has to be personalized to users' English proficiency. Results suggest that translation approaches should be used only on users with at least elementary English proficiency. As revealed by this study, despite the higher precision of English queries in low English proficiency users, these users have a lower degree of comprehension of English documents, obtain a less accurate knowledge through

For Peer Review

English queries and feel less satisfied in the tasks with this type of queries. Although some of these results are not surprising, we consider important to have an empirical demonstration of these facts. We also found that the readability of documents should be a criterion for ranking, specially if the user is proficient in the documents' language. Moreover, we found that more complex terminology may inspire confidence in the retrieved document but this conclusion has to be further explored. These findings suggest that a cross-lingual assistance personalized to the users' English proficiency could improve non-English consumer health retrieval and could be helpful in an educational sense, enabling non-English speaking users to learn English medical terminology. Moreover, it may also be helpful to trigger new search strategies and to help the user construct queries that give access to documents that may not be reached otherwise. Future work involves studying how English proficiency may be automatically inferred by search logs. We would also like to repeat this study with non web-dominant languages other than Portuguese to verify if our results are generalizable to other languages. Also interesting would be to analyze if the same conclusions would emerge in a study that involves two web-dominant languages like English and Mandarin, that is, to know if English translations would still be useful to Chinese users. Finally, we would also like to study if, in low English proficiency users, query translation complemented with the machine translation of the retrieved documents would be a good strategy to give them access to the higher quality and quantity of English content.

9 Acknowledgments

Thanks to Fundação para a Ciência e a Tecnologia for partially funding this work under the grant SFRH/BD/40982/2007. Thanks to Dagmara Paiva, M.D., and to Michael Luís, M.D., for their contribution on the evaluation of the medical accuracy of users' answers.

References

Baujard, V., Boyer, C., & Geissbühler, A. (2010). Evolution of Health Web certification. In *23rd annual days of the swiss society of medical informatic.*

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3).

Cline, R. J. W., & Haynes, K. M. (2001, December 1). Consumer health information seeking on the Internet: the state of the art. *Health education research*, *16*(6), 671–692.

Eichmann, D., Ruiz, M. E., & Srinivasan, P. (1998). Cross-language information retrieval with the UMLS metathesaurus. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 72–80). New York, NY, USA: ACM.

Fitzsimmons, P. R., Michael, B. D., Hulley, J. L., & Scott, G. O. (2010, December). A readability assessment of online Parkinson's disease information.

For Peer Review

The Journal of the Royal College of Physicians of Edinburgh, *40*(4), 292–296.

Fox, S. (2006, October 29). *Online Health Search 2006* (Tech. Rep.). Pew Internet & American Life Project.

Fox, S., & Jones, S. (2009, June 11). *The Social Life of Health Information* (Tech. Rep.). Pew Internet & American Life Project.

Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English language use on the WWW. In *Proceedings of riao*.

Hersh, W. (2008). *Information Retrieval: A Health and Biomedical Perspective (Health Informatics)* (3rd ed.). New York, NY, USA: Springer.

Hersh, W. R., & Donohoe, L. C. (1998). SAPHIRE International: a tool for cross-language information retrieval. In *Amia annual symposium proceedings* (pp. 673–677).

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the World* (16 ed.). Dallas, Texas: SIL International.

Lopes, C. T., & Ribeiro, C. (2011). Data Certification Impact on Health Information Retrieval Information Quality in e-Health. In A. Holzinger & K.-M. Simonic (Eds.), *Usab 2011 - information quality in ehealth* (Vol. 7058, pp. 31–42). Berlin, Heidelberg: Springer Berlin / Heidelberg.

Lu, W.-H., Lin, S.-J., Chan, Y.-C., & Chen, K.-H. (2005). Semi-automatic construction of the Chinese-English MeSH using Web-based term translation method. *AMIA Annual Symposium proceedings*, 475–479.

Névéol, A., Pereira, S., Soualmia, L. F., Thirion, B., & Darmoni, S. J. (2006). A method of cross-lingual consumer health information retrieval. *Studies in health technology and informatics*, *124*, 601–608.

Petrock, V. (2010, August 10). *Cyberchondriacs Becoming Empowered Health Information Seekers*. Available from: http://www.emarketer.com/blog/index.php/cyberchondriacs-empowered-health-seekers/ [cited 2011-05-11] (Archived by WebCite at http://www.webcitation.org/5ym7xLoYp).

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 55–63). New York, NY, USA: ACM.

Robertson, S. E., Kanoulas, E., & Yilmaz, E. (2010, July). Extending average precision to graded relevance judgments. In *Proceeding of the 33rd international acm sigir conference on research and development in information retrieval* (pp. 603–610). New York, NY, USA: ACM.

Rosemblat, G., Gemoets, D., Browne, A. C., & Tse, T. (2003). Machine translation-supported cross-language information retrieval for a consumer health resource. *AMIA Annual Symposium proceedings*, 564–568.

Saracevic, T. (1996, October). Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (colis 2)* (pp. 201–218).

Stichele, R. V. (1995, December). *Multilingual Glossary of technical and pop-*

For Peer Review

ular medical terms in nine European Languages (Tech. Rep.). Heymans Institute of Pharmacology, University of Gent.

Tran, T. D., Garcelon, N., Burgun, A., & Le Beux, P. (2004). Experiments in cross-language medical information retrieval using a mixing translation module. *Studies in Health Technology and Informatics*, *107*(Pt 2), 946–949.

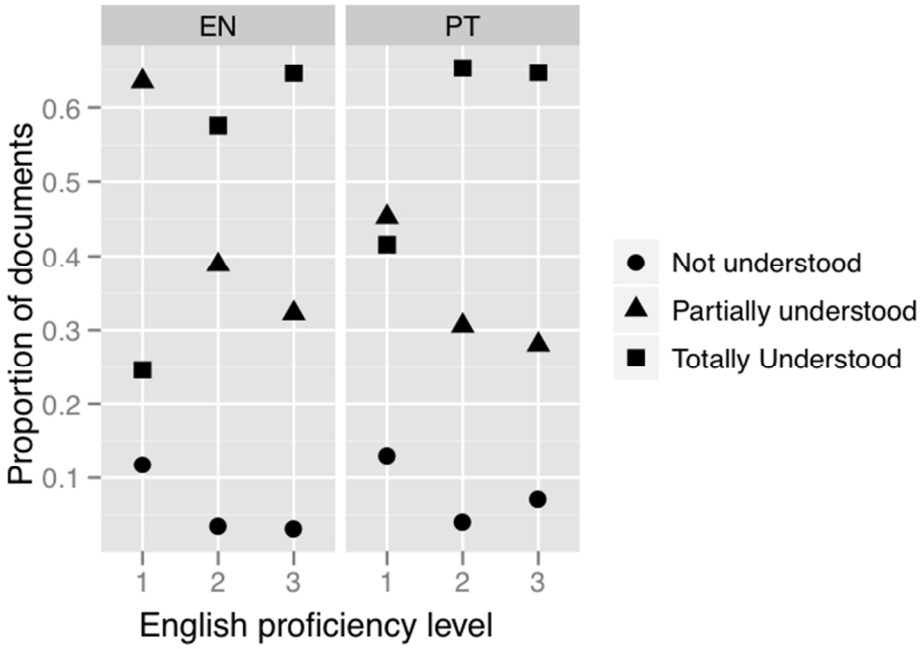
Volk, M., Ripplinger, B., Vintar, S., Buitelaar, P., Raileanu, D., & Sacaleanu, B. (2002, December 4). Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, *67*(1-3), 97–112.

W3Techs. (2012, July 23). *Usage of content languages for websites.* Web Technology Surveys.

For Peer Review

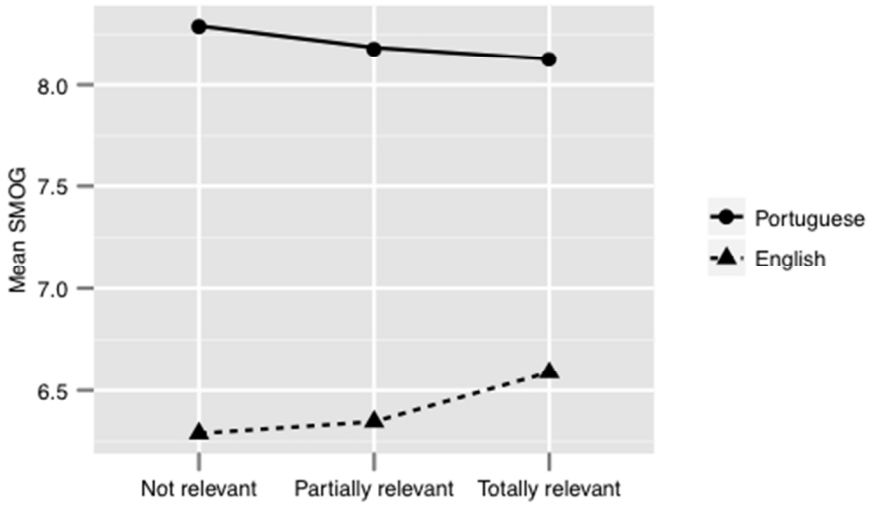
user	#Iteration							
	1	2	3	4	5	6	7	8
1	1e	2e	3p	4p	5p	6p	8e	7e
2	1e	2e	4p	3p	5p	6p	7e	8e
3	2p	1p	3e	4e	6e	5e	8p	7p
4	2p	1p	4e	3e	6e	5e	7p	8p
5	3e	4e	2p	1p	7p	8p	5e	6e
6	3e	4e	1p	2p	7p	8p	6e	5e
7	4p	3p	2e	1e	8e	7e	5p	6p
8	4p	3p	1e	2e	8e	7e	6p	5p
9	5p	6p	8e	7e	1e	2e	3p	4p
10	5p	6p	7e	8e	1e	2e	4p	3p
11	6e	5e	8p	7p	2p	1p	3e	4e
12	6e	5e	7p	8p	2p	1p	4e	3e
13	7p	8p	5e	6e	3e	4e	2p	1p
14	7p	8p	6e	5e	3e	4e	1p	2p
15	8e	7e	5p	6p	4p	3p	2e	1e
16	8e	7e	6p	5p	4p	3p	1e	2e

Latin square procedure followed in assessment task assignment. Cells' background color define the retrieval system, [1-8] defines the information situation and [p, e] the queries' language. This shows the task assignment for a subset of 16 users.
113x122mm (150 x 150 DPI)



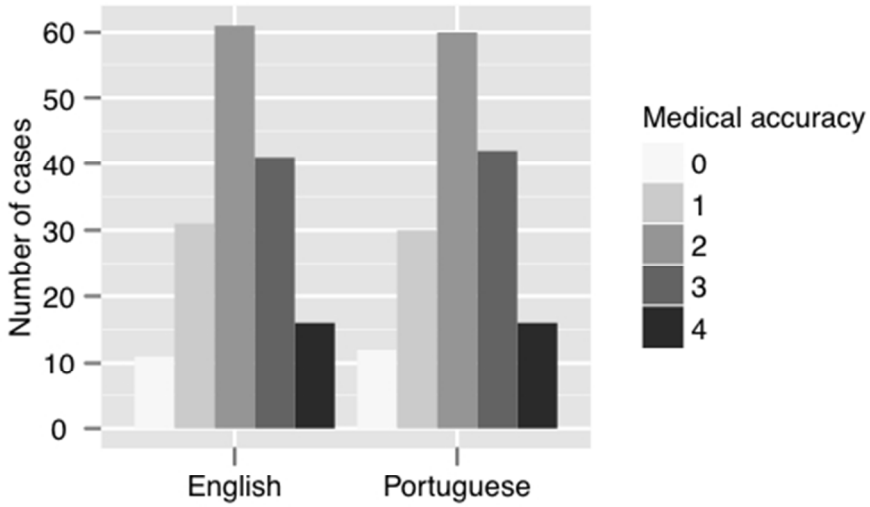
Proportion of documents by English proficiency level (low-1; elementary-2; good-3), query language and users' comprehension.
127x88mm (150 x 150 DPI)

For Review



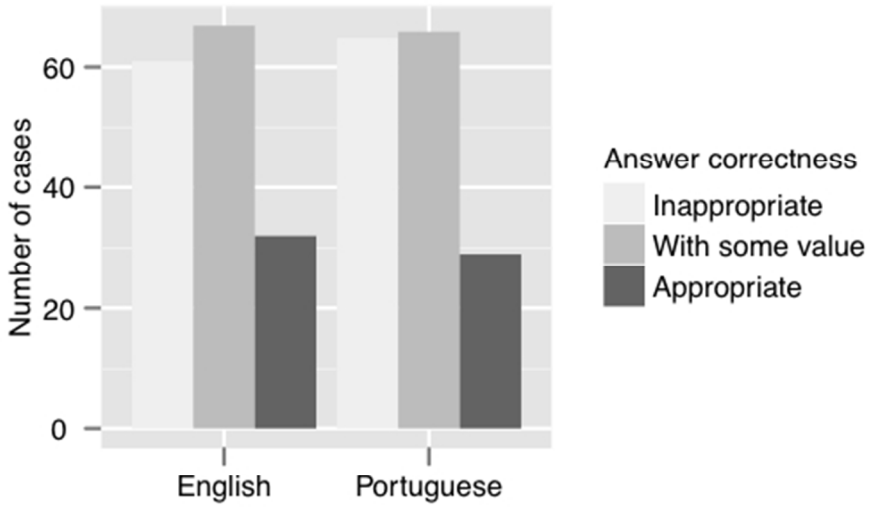
Mean SMOG by documents' relevance in each language.
97x61mm (150 x 150 DPI)

For Review



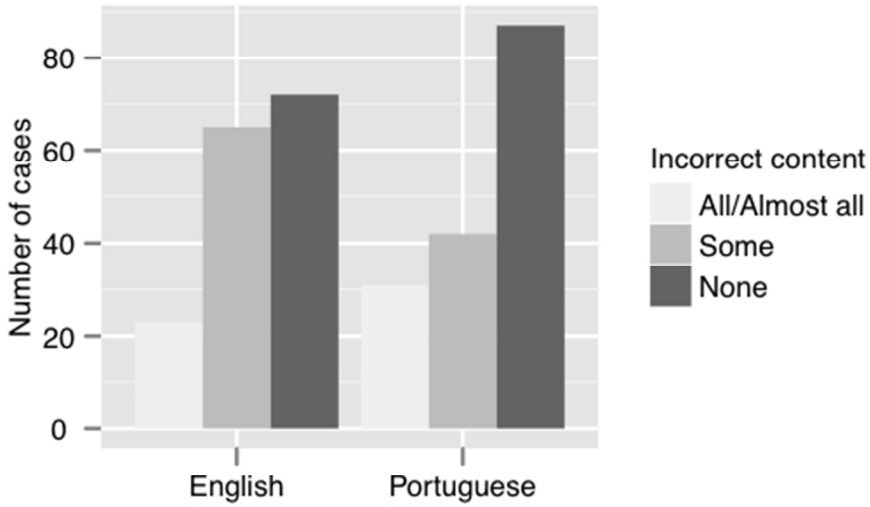
Answers' medical accuracy by query language.
97x61mm (150 x 150 DPI)

For Review



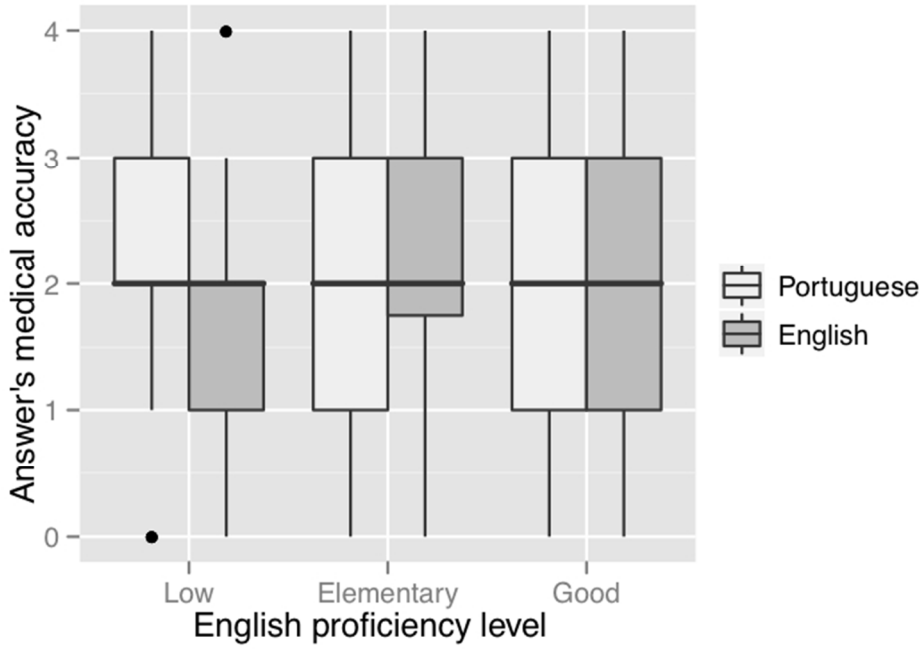
Answers' correctness by query language.
97x61mm (150 x 150 DPI)

For Review



Answers' incorrectness by query language.
97x61mm (150 x 150 DPI)

For Review



Medical accuracy boxplots by English proficiency and query language.
127x88mm (150 x 150 DPI)

Review