Assessing the impact of English language skills and education level on PubMed searches by Dutch-speaking users

Klaar Vanopstal[1,2], Robert Vander Stichele[3], Godelieve Laureys[4], Joost Buysschaert[1]

[1] LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
klaar.vanopstal, joost.buysschaert@hogent.be
[2] Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium
[3] Faculty of Medicine and Health Sciences, Heymans Institute of Pharmacology, Ghent University
De Pintelaan 185, 9000 Gent, Belgium
robert.vanderstichele@ugent.be
[4] Faculty of Arts and Philosophy, Department of Nordic Studies, Ghent University
Rozier 44, 9000 Gent, Belgium
godelieve.laureys@ugent.be

Abstract

The aim of this study was to assess the retrieval effectiveness of nursing students in the Dutch-speaking part of Belgium. We tested two groups: students from the master Nursing and Midwifery training, and students of the bachelor Nursing training. The test consisted of five parts: first, the students completed an enquiry about their computer skills, experiences with PubMed and how they assessed their own language skills. Secondly, an introduction into the use of MeSH in PubMed was given, followed by a PubMed search. After the literature search, a second enquiry was completed in which the students were asked to give their opinion about the test. To conclude, an official language test was completed. The results of the PubMed search, i.e. a list of articles the students deemed relevant for a particular question, were compared to a gold standard. Precision, recall and F-score were calculated in order to evaluate the efficiency of the PubMed search. We used information from the search process, such as search term formulation and MeSH term selection to evaluate the search process and examined their relationship with the results of the language test and the level of education.

1. Introduction

The internet explosion puts information which was inaccessible to the previous generation of researchers at the fingertips of all internet users. It has become a challenge not to drown in this information flood, and efficiency in searching is therefore of vital importance.

With English being the lingua franca of science, the "new Latin" (Eisenberg, 1996), many non-English scientists may experience difficulties when conducting a literature search. A closer look at the language diversity (see figure 1) tells us that more than 78% of all publications in MEDLINE/PubMed are written in English. Only 0.2% of all articles included in MEDLINE are originally written in Dutch, which implies that Dutch users of MEDLINE/PubMed not only have to deal with an English interface, but also with English information.

Whereas the use of a common international language may create terminological continuity, there is still a language barrier to surmount for non-native speakers of English, especially since English tends to prefer the use of internationalisms, or words of Greco-Latin provenance, over vernacular terms.

However, difficult medical terminology might not be the only factor influencing the efficiency of cross-language information retrieval: a basic level of English knowledge including linguistic items other than domain-specific terminology is needed in order to select relevant information (Lankamp, 1989). Moreover, several sub-languages are needed for efficient bibliographic retrieval: the languages of informatics, documentation and biomedical sciences (Mouillet, 1999).

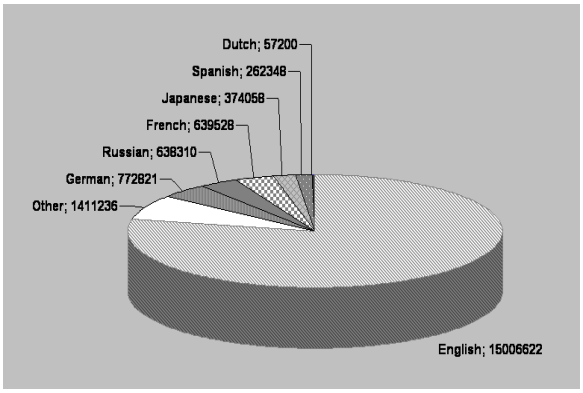

Figure 1: Linguistic diversity in MEDLINE/PubMed

This research focuses on the impact of several factors, including the English language -and not only terminology- barrier, and the level of education on the efficiency of literature searches by nursing students, specifically master and bachelor students of Nursing and Midwifery.

2. Methods

2.1. Background of test group

Vendel (1982) argues that medical knowledge plays a crucial role in the understanding of English medical literature. Next to the language aspect, this medical background adds an extra dimension to this study. Therefore, we opted for

two different groups of students: Dutch-speaking master and bachelor students. Both groups took the exact same test. The bachelor students were recruited at the Nursing Department of the University College Gent (n=31), the master students at the Department of Nursing and Midwifery at the University of Antwerp (n=23).
In the academic year 2008-2009, a total of 5,547 students (see table 1) registered for a bachelor training in one of the 13 institutions which offer this training in Flanders, the Dutch-speaking part of Belgium. 4,532 (82%) of them were female students, whereas only 1015 (18%) were male students. This distribution is comparable to the distribution male/female in our test sample, where we had 23 female (74%) and 8 male (26%) bachelor students. Analogously, the number of male and female master students is reflected in the test group of master students: we had 18 female (78%) and 5 male (22%) respondents, compared to 211 female (76%) and 66 male (24%) master students 5 Flemish institutions.

Bachelor	male	female
test group	26%	74%
total students	18%	82%
Master	male	female
test group	22%	78%
total students	24%	76%

Table 1: Representativity of the test groups

The respondents from the bachelor and master training took a compulsory course in the first year of their training in which they were initiated into the domain of research and where they learned to search for and understand specialist literature. In addition, the master students in our test group attended an additional programme on scientific research in their master training, which includes methodological principles of literature searching and systematic review and analysis of literature. One of the questions we will try to answer in this study is whether these master students actually perform better on the literature search test than the bachelor students, as their background in literature searching and scientific research is somewhat broader.

2.2. Test

This study deals with problems related to query formulation and to reading comprehension, which is crucial to the selection of relevant information. The test consisted of five parts. First, the students completed an enquiry which focused computer skills, familiarity and experiences with the search system, i.e. PubMed, and self-assessment of the English language skills. After completion of this questionnaire, an introduction into the use of MeSH (**Me**dical **S**ubject **H**eadings) in PubMed was given. PubMed is an interface to MEDLINE, which is the world's largest biomedical literature database created by the National Library of Medicine (NLM). MeSH is a controlled vocabulary also created by the NLM for the purpose of indexing journal articles and books in the biomedical sciences. PubMed users can consult this vocabulary to enhance their literature search.
In the third stage, the students conducted a literature search for a specific theme in nursing. This literature search was assessed in several different ways (cf. section 3.1). The students were asked to search for very specific information about fall prevention. The question was asked in Dutch, as the students' information needs normally also arise from a real-life situation in which the problem is formulated in Dutch. Subsequently, a second questionnaire was completed in order to see how the students experienced this test. For an objective assessment of the students' language skills, an official language test was completed (cf. section 3.2).

3. Evaluation

3.1. Evaluation of the search process

The students had fifteen minutes for their PubMed search. In these fifteen minutes, they had to go through the whole search process, from query formulation to relevance judgement. The result was a list of documents the students deemed relevant to the search question. These results were then compared with a gold standard. From this comparison, precision and recall rates and F-score were deduced. F-score is a harmonic mean of precision and recall in which both scores are weighted evenly:

$$F = 2 * (precision * recall)/(precision + recall) \quad (1)$$

Kagolovsky and Moehr (2003) define information retrieval as a science which has two links: one to computer science, and one to behavioural science. The interaction between user and information retrieval (IR) system is very important for this research, as IR comprises much more than just document retrieval using computers. Therefore, we analysed not only the results, but also the search process.
We used the Morae software (http://www.techsmith.com/morae.asp), a program specifically designed to record and analyse user-computer interaction, which allows a researcher to capture all operations executed by a user and to log tasks, markers, marker scores and add text notes. In the study configuration, we defined several tasks, including reading the search question (task 1), searching for relevant articles (task 2), final selection of articles (task 3), hesitations and/or errors (task 4), abstract/article view (tasks 5 and 15 to 23), and individual searches (tasks 7 to 14), which measure the time from search term formulation to article selection. Hesitations and errors can disputably be classified as a task, but this is the only way to mark events which occur over a period in time. This time span is important for our analysis, as it could indicate that the user is having problems with the search system, or that the user needs some time to think about the formulation of a search term. We also defined 26 different markers, the most important of which are "Search term formulation", "MeSH term selection", "PubMed search" and "Article selection". To each search term formulated and each MeSH term selected by the participants, a score was assigned: 0 for a "bad", 1 for a "medium" and 2 for a "good" search or MeSH term. Medium search terms include typographical errors such as *physiotherapy progroms* or *resiential care*, orthographical errors like *fysiotherapy* or *multifactoriel intervention*, and

search terms which are not completely relevant or which are not meaningful enough to include in the query. Examples of this kind of search term are *resident* or *clinical path*. Bad search terms include incorrect translations, such as *kine*, *kinesitherapi* and *kinestics* (instead of *physiotherapy*; translation for the Dutch word *kinesitherapie*), *movingexercises* or *residention nursinghome*. Search terms which are not relevant for this information search or too general to achieve relevant results (e.g. *therapist* or *housesettings*) are also considered as bad search terms.

3.2. Evaluation of the English language skills

The participants completed a freely available diagnostic language test, DIALANG (www.dialang.org). This test allows us to compare the English language skills of both test groups and to link the results to their performance on the literature search test. The test has been internationally validated and was developed by more than twenty major European institutions, with the support of the European Commission. It is based on the Common European Framework of Reference and is available for fourteen European languages, including English. The participants completed the reading and vocabulary test.

We can hypothesize that, in order to be able to select and understand relevant articles, users of PubMed should at least have B2 or C1 level on the reading and vocabulary test. With a B2 level in reading, people should be able to understand articles and reports about contemporary issues and most short stories and popular novels. A C1 level in reading means that the test person can understand long and complex factual and literary texts as well as differences in style and specialised language in articles and technical instructions. People with a B2 level in vocabulary should be able to write reports and essays, and people with a C1 level in vocabulary can write reports and essays about complex subjects.

4. Results

4.1. Language skills

The participants were tested for their English language skills in reading and vocabulary. They completed both parts of the DIALANG test, and their results were compared to their self-reported language skills.

		Bachelor	Master
		Count N%	Count N%
Score reading test	A1	3.2%	2.5%
	A2	9.7%	12.5%
	B1	35.5%	12,5%
	B2	38.7%	50.0%
	C1	9.7%	15.0%
	C2	3.2%	7.5%
Score vocabulary test	A1	0%	0%
	A2	3.2%	10.0%
	B1	12.9%	7.5%
	B2	67.7%	57.5%
	C1	9.7%	25.0%
	C2	6.5%	0%

Table 2: English language skills

We did not find a significant relation between the level of education (bachelor/master) and the scores on the language test. However, as can be observed in table 2, 65% of the master students have a B2 or C1 level for reading, whereas bachelor students scored somewhat lower: 48.4% obtained a B2 or C1 level. The scores for the vocabulary test were somewhat higher: 84.5% of the master students obtained a B2/C1 level, compared to 77.4% of the bachelor students obtained a B1/B2 level. These results correspond to their self-assessment results: we found a positive correlation between the self-assessment and the reading test scores (r_s=0.400; n=71; p=0.00) on the one hand, and between the self-assessment and the vocabulary test scores (r_s=0.346; n=71; p=0.00).

Master students also seem to estimate their language skills higher in the pre-test survey than the bachelor students who participated in this test. Table 3 presents the results of the Mann-Whitney test, which gave us a significant result (p-value = 0.003) and a z-score of -2.923.

	English language skills
Mann-Whitney U	381.000
Wilcoxon W	877.000
Z	-2.923
Asymp. Sig.(2-tailed)	.003
a	Grouping var.: level of education

Table 3: Self-reported English language skills

Another correlation which we investigated, was that between the language skills and the results of the literature search (F-score). Both reading and vocabulary tests correlate positively with the F-score (r_s=0.261; n=71; p=0.028 and r_s=0.258; n=71 and p=0.0298 respectively). This means that participants who have better English language skills perform better on the literature search task. Table 4 shows the F-scores per level of English language skills.

		F-score
		Mean
Score reading test	A1	.0361
	A2	.0234
	B1	.0495
	B2	.0683
	C1	.0753
	C2	.1197
Score vocabulary test	A1	.
	A2	.0521
	B1	.0210
	B2	.0575
	C1	.0885
	C2	.1517

Table 4: F-scores per level of English language skills

The relationship between the maximum time between inputs and the results on the language test is also interesting. Longer periods of inactivity (and thus a higher maximum time between inputs) might indicate that the test person was hesitating or unsure about the next step.

We found a non-significant negative correlation (reading test: r_s =-.226; n=71; p=.058 and vocabulary test: r_s =-.098; n=71; NS) between the scores on the language test and the maximum time between inputs, which means that higher scores on the reading and vocabulary tests often go together with a lower maximum time between inputs, or longer hesitations go together with low scores on the language test. These hesitations might be caused by uncertainty about the translation into or interpretation of the English language.
A non-significant negative correlation (r_s=-0.267; n=71; p=0.29) was observed between the participants' scores on the vocabulary test and the number of bad search terms formulated during the PubMed search. This means that participants who scored badly on the vocabulary test, tended to formulate a higher number of bad search terms.

4.2. Precision, recall and F-score

Precision, recall and F-scores were very low in both groups: the master nursing students had a mean precision of 29.97% and a recall of 4.42%, whereas the bachelor students achieved 37.58% precision and 2.69% recall. The F-scores for master and bachelor students were 7.22% and 4.85% respectively. The extremely low recall, precision and F-scores in both groups can partly be attributed to the limited time (15 minutes) the students had to search for relevant documents. Table 5 shows the average and maximum F-scores for both groups.

	Average F	Max. F
Bachelor	4.9%	18.7%
Master	7.2%	26.4%

Table 5: F-scores in both test groups

The differences between both groups were not significant, probably due to the very low scores.
After the test, the participants were asked what they thought of their search process and of the selection of articles they had made. We found a strong positive correlation between the participants' perception of how they performed on the test, and their overall F-scores, which means that they have a realistic view of their performances (table 6).

	r_s	n	p
good selection - F-score	.535	71	.000
found easily - F-score	.517	71	.000

Table 6: Correlations between F-score and search result satisfaction

4.3. Level of education

Apart from the correlation with precision, recall and F-score and the result on the language test, some observations can be made as to the level of education. When asked whether they use biomedical databases to search for -(bio)medical- information, all master students responded positively, as opposed to only 45% of the bachelor students. Master students tend to search for medical information in English more frequently than bachelor students do (Mann-Whitney U=245.00; z=-4.42; p =.000). When asked whether they find it more difficult to search for information in English, the bachelor students are less certain about their skills than master students (Mann-Whitney U=404.50; z=-2.54; p=.011). A bigger proportion of the bachelor students responded positively when they were asked whether they need more time to read English (medical) articles than to read Dutch articles (Mann-Whitney U=441.00; z=-2.27; p=.023)
There are also some differences between both groups with respect to their experience with PubMed. The master students received a more elaborate introduction into searching with PubMed than bachelor students did (Mann-Whitney U = 156.00; z = -5.97; p = .000), and -consequently?- they use this medium more often to search for medical information (Mann-Whitney U=78.00; z=6.38; p=.000). All the master students in the test group indicated that they knew what MeSH (Medical Subject Headings) is, and 77.5% sometimes use them to look for information in PubMed. 12,5% always use MeSH in PubMed. Only 3.2% of the bachelor students knew what MeSH was, and none of them had ever used this controlled vocabulary to search PubMed. When asked what their preferred language to look for medical information was, 82.5% of the master students expressed a preference for English, as opposed to only 9.7% of the bachelor students.
As mentioned above, the maximum time between inputs can be an indication of how long participants hesitated before taking the next step. We found a positive correlation between the level of education and the number of mouse clicks (t(42.289)= 5.496; p= .000). This might indicate that they are more proficient in searching PubMed (see also figure 2).

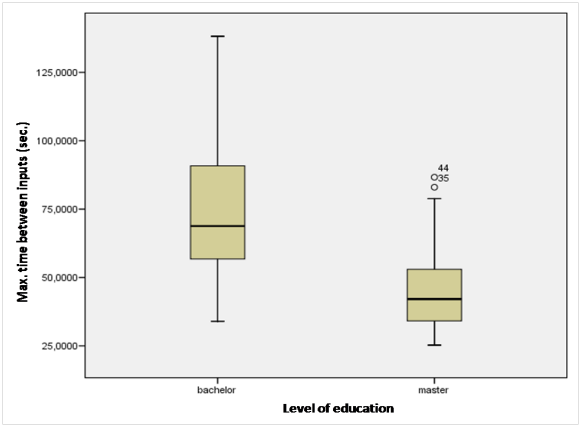

Figure 2: Maximum time between inputs for bachelor and master students

We used the Mann-Whitney test to analyse the relation between the level of education (bachelor/master) and the number of bad search terms formulated during the PubMed search. This test indicated a significant relation (Mann Whitney U=427.5; z=-2.261; p=.023) between the number of bad search terms and the level of education. In other words, bachelor students formulated more bad search terms than master students during their literature search in PubMed.

4.4. Search process

The students were asked to use MeSH terms to search for relevant information about a certain question. This means that they first entered a search term, and looked for most relevant MeSH term. Subsequently, they constructed a query by combining their MeSH terms with the Boolean operators AND, OR and NOT.

As opposed to the number of good, bad or medium search terms, the selection of incorrect MeSH terms proved to have a negative influence on the F-scores (r_s=-.259; n=71; p=.029), as did the selection of medium MeSH terms (r_s= -.144; n=71; NS). The selection of good MeSH terms correlates positively with the F-scores (r_s=-.140; n=71; NS). This is due to the fact that the students used MeSH terms to construct their queries, and not free text. If they entered a bad search term (e.g. *kinestherapy* for *physical therapy*), either a warning message appeared saying "The following term was not found in MeSH: kinestherapy. See Details. No items found.", or the MeSH terms suggested for the search term were not suitable for the search (e.g. the search term *multifactorial* yielded the following MeSH terms: *Multifactorial Inheritance*, *Causality*, *Nephrogenic Fibrosing Dermopathy*, *Typhlitis*, etc.). In this case, a new -and usually better- search term was formulated, and there was no impact on the search results. The selection of an incorrect MeSH term, however, did have an impact on the search results, as the MeSH terms were sent directly to the search box.

5. Conclusion

We found that English reading and vocabulary skills have an impact on the recall, precision and overall F-score of the search. Master students of Nursing and Midwifery did not achieve significantly better results on the language test or for the PubMed search than bachelor students of Nursing. However, their knowledge of the search system is better, which is reflected in their lower maximum time between inputs. The master students formulated a lower number of bad search terms than their colleagues from the bachelor training, but as they constructed their queries with MeSH terms, this did not influence their PubMed search. The Medical Subject Headings proved to be a useful language aid, as bad search terms yield a warning message and incite the user to formulate a better search term.

In our future research, we would like to conduct the same test at the Nottingham University Nursing School (UK), so that we have a test group (Belgian students) and a control group (British students). We will also ask an expert in bibliographic retrieval and an expert in the domain of the search question ("accidental falls in elderly") to perform the PubMed search, in order to compare their search techniques and results to those of our test group.

In the final stage of our research, we will conduct a similar test, but with a somewhat different set-up. A Dutch-speaking test group and control group will perform an internet search for some specific medical information, like the one the participants did in this test. However, the test group will get language support, whereas the control group will not. This language support will be provided in the form of Dutch translations of the MeSH.

6. References

A. Eisenberg. 1996. Using English as the international language of science. In D.C. Andrews, editor, *International Dimensions of Technical Communication*, pages 1–4. Society for Technical Communication, Arlington.

Y. Kagolovsky and J. R. Moehr. 2003. Terminological problems in information retrieval. *J Med Syst*, 27(5):399–408. 0148-5598 (Print) Journal Article Research Support, Non-U.S. Gov't Review.

R.E. Lankamp. 1989. *A Study on the Effect of Terminology on L2 Reading Comprehension. Should Specialist Terms in Medical Texts be avoided?* Proefschrift, Eindhoven University of Technology.

E. Mouillet. 1999. Language barriers and bibliographic retrieval effectiveness: use of MEDLINE by French-speaking end users. *Bull Med Libr Assoc*, 87(4):451–5. 0025-7338 (Print) Journal Article.

A.P.C. Vendel. 1982. Het lezen van vakliteratuur in een vreemde taal, achtergrond en een experiment. *Levende Talen*, 371:310–319.