Collecting and Analyzing User's Queries

- a first step in developing an application aimed at the automatic translation of queries written in French natural language

Nicolas Fairon and Françoise Pasleau
Life Sciences Library, University of Liege, Belgium

Introduction

Writing an efficient Medline query remains a difficult exercise for users with little or no training, or for those who do not search the database on a regular basis. Language is an additional barrier for non-native English speakers. For these reasons, we engaged in a Ph.D. research program aimed at providing our users with a new web-based interface able to search Medline in French natural language.

The development of such a computer system relies mainly on MeSH (Medical Subject Headings), an important thesaurus part of the UMLS metathesaurus. Metathesaurus is a multi-lingual vocabulary database that links alternative names, synonyms or phrases that designate medical and health-related concepts. It is also categorized in order to show the relationships between the different concepts. Ultimately the French interface will be linked to the PubMed or Ovid search engines.

The first research step consisted of collecting research material. An electronic corpus of queries was built, which will be used to develop and to test our future application.
This poster describes how we selected two target populations of Bachelor's students from the Medical Faculty and from different medical nursing schools and other paramedical schools. Harvested queries were further analyzed and characterized according to their syntax as well as to their lexical and scientific content.

Materials and Methods

1. Harvesting Medline Queries

2nd year Bachelor's
Medicine & related disciplines
IT literacy course
(printed French written group work)

3rd year Bachelor's
Nursing schools & other paramedical schools
Bibliographic research for end-of-studies work
(French written questions sent by e-mail)

2. Building a corpus of queries

- Queries collected are stored in a database
- Each query is linked to its original group
- Queries writen in French natural language are linked to their resulting Medline query
- Additional fields are added to characterize query complexity

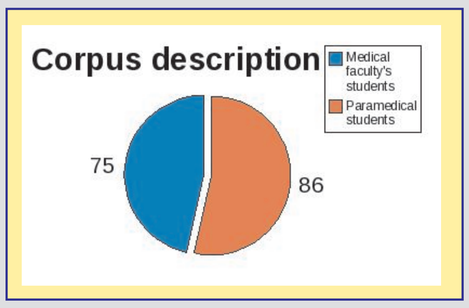
3. Filtering

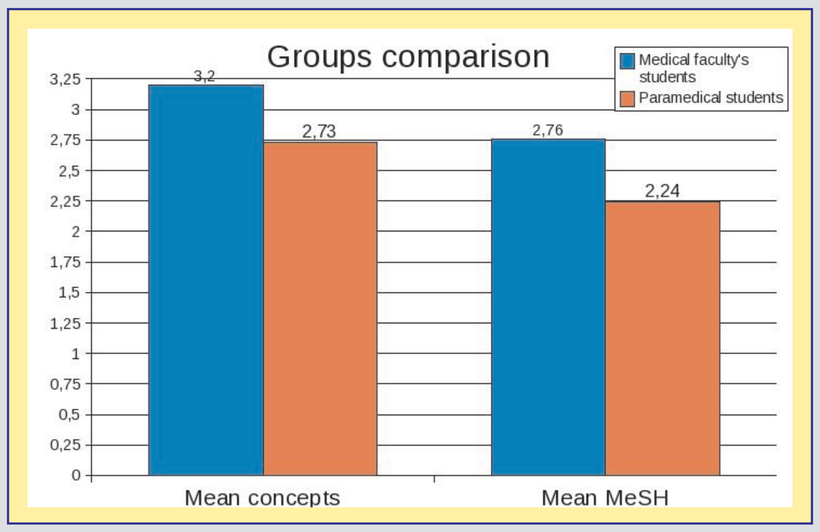
- **Discarding** queries with too little or no medical content
- **Clearing** queries of unnecessary sentences (Polite phrases, letter endings, noise, ...)

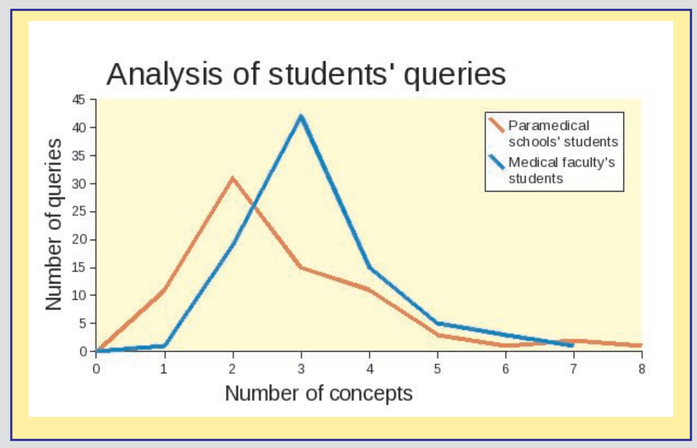
4. Analyzing queries

- **Authors'** level of education
- **Content:** manual listing of biomedical concepts
- **Complexity:** counting biomedical concepts, headings, subheadings
- **MeSH database** was used to search and record all the relevant headings & subheadings (including related terms from UMLS)
- **Free language terms:** for which no MeSH terms are found

Results







Description of the corpus according to authors category.
We received a total of 171 queries, 10 of them were discarded according exclusion criteria (See Step 3 : Filtering). This graph shows the repartition of the remaining queries according to authors' education level (See Step 1 : Harvesting Medline Queries).

Groups comparison
We computed the average numbers of biomedical concepts and MeSH terms included in the queries of both groups. A statistical difference is observed between the 2 categories of students. Second year Bachelor's students who are taught in IT Literacy are able to write queries with a richer scientific content.

Analysis of students' queries
This graph shows the quantitative analysis of queries according to the number of included concepts. Most of the questions submited by students working on their end-of-studies work contain more than two different concepts.

Conclusions

Each of our two populations of students was shown to express queries differently, as expected. The goals of students were different and the Medical Faculty students had already been introduced to Medline during their studies; this could explain the higher prevalence of MeSH terms in their queries.

There are a lot of tasks remaining in order to complete our PhD research:
- To continue collecting queries: the bigger the corpus, the better it will be. Despite the 161 queries already collected, we need more queries in order to split our corpus into a test corpus and an evaluation corpus.
- To sort results by complexity in order to facilitate the development of our application
- To sort queries into morphosyntactic families
- To correct syntactic mistakes (due to typing, grammatical errors, ...)
- To exclude too general queries for which consultation of a textbook would be more appropriate

After performing this first step, it is evident that the task is difficult in regard to the complexity of the many requests and the complexity of the informatic tool to be developed in order to build a web-interface to search Medline in French natural language. Nevertheless, it is an interesting challenge in order to provide our users with a new tool to improve their bibliographic searches.