# User queries collecting and analysis - first step for an automated query translation application development

Nicolas Fairon, Francoise Pasleau

## Introduction

For non native english speakers, it is often difficult to make an efficient Medline query. This is particularly true for people with no or little training with the use of Medline and for those who don't use it on a regular basis.

This paper describes the firsts steps of our broader project which aims to set up an interface for querying Medline in french natural language. The targeted population will be final-year classes' students of nursing schools and other medical-related schools.

## Objectives

Our objectives are to collect user written queries in order to build a corpus and analyze it. The harvest of queries is a decisive and important step to help us to decide how to build our software and to later assess it. For these reasons we must gather the most queries as possible, with the librarian medline search strategy associated. Once collected, these queries are analyzed by librarians and scientists for further use in natural language processing applications.

## Methods :

The students were contacted via their teachers to explain to them that they can benefit from a free bibliographic search undertaken by a professional librarian for their end-of-studies work. They also have been directly contacted with the use of poster explaining our service. The only condition was that this research is carried out only on basis of email which they sent. No example were given, nor restrictions or advices, in order to not influence them on their manner of writing these emails. Each email has then been treated by the same librarian, saving it in a relational database with the Medline search strategy undertaken associated. If need be, emails had been split up in subqueries, each treated separately.

We then examined the content of the queries, looking for medical concepts, mesh terms and subheadings. We made a distinction between expressed MeSH terms and subheadings and their variations (synonyms, etc) and the MeSH terms and subheadings derived from meaning of sentence.

## Results

Contrary to our expectations, only a few students, 27, responded to our request for bibliographic queries. These 27 answers lead to 84 distinct bibliographic searches. It has so been decided to complete our corpus with the use of medical queries written by university students (from medicine faculty) during teaching exercises the past 3 years. This gave us 83 new queries. As one can expect, these 2 groups have some distinct characteristics. There is a lot of spelling or typing mistakes in the queries from the emails of students and none or few in the medical questions of the university students. But it is overall the semantic content of the queries that differ. The questions from students of the medical faculty are more accurate, and an information extraction from the references is needed to answer them (this step is not and will not be completed in the immediate future). At the opposite, questions from bachelor students are more general, even too general sometimes.

After counting concepts, MeSH terms and Subheadings into the questions submitted by bachelor students, we found 212 medical concepts and 158 MeSH terms and subheadings clearly identified. Concepts not founds have no MeSH terms or are just badly expressed by the students. Only 38 (45,24%) queries have as many Mesh terms and subheadings as medical concepts. On the other side, 10 queries (11,9%) have no Mesh terms or subheadings matching their medical concepts content.

## Conclusions

The harvesting of medical queries from our users is not an easy task. It has to be pursued next year to collect more queries to enrich our corpus. We also have to process the queries collected from medical

faculty students and to sort efficiently the queries for the evaluation of our future application.

To avoid some problems with spelling and typing mistakes, it will be necessary to correct them in the corpus (in a distinct field of the database to preserve original queries). Further, it will still be possible to build a model taking in account these mistakes, but for now, it is not necessary as our project will only be a model.

Too general queries have to be separated from the others because Medline is not the best choice to retrieve information, a textbook is more useful for those case.