

# A Study of the Adequacy of User and Indexing Vocabularies in Natural Language Queries to a MeSH-Indexed Health Gateway

N. Grabar, M.Sc.,<sup>1</sup> P. Zweigenbaum, Ph.D.,<sup>1</sup> L. Soualmia, M.Sc.,<sup>2,3</sup> S.J. Darmoni, M.D., Ph.D.<sup>2</sup>

{ngr,pz}@biomath.jussieu.fr, {lina.soualmia,stefan.darmoni}@chu-rouen.fr

<sup>1</sup> DIAM — STIM, DSI, Paris Hospitals, Dpt Biomathématiques, U. Paris 6, Paris, France

<sup>2</sup> Computer and Networks Department, Rouen University Hospital, Rouen, France

<sup>3</sup> Perception, Information and Systems Lab, National Institute of Applied Sciences, Rouen

## BACKGROUND

We examine an enabling condition for natural language access to medical knowledge resources (Medline, CISMef) indexed with controlled vocabularies (MeSH): are the words in user queries comparable with those of the index terms? To which extent do morphological word variants<sup>1</sup> help in this mapping?

## MATERIAL AND METHODS

The *queries* studied come from the ‘simple search’ interface of Doc’CISMef (doccismef.chu-rouen.fr, Sep. 2000 to Jan. 2001): 21,112 unique words (131,570 occurrences). The *target terms* indexing CISMef come from the French MeSH: 21,475 unique words (58,912 occurrences). *Morphological knowledge* was reused from previous work and extended: 308,847 words pairs for lemmatization (*abdominaux* → *abdominal*) compiled from general dictionaries and medical corpora and 1,041 words pairs for stemming (*abdominal* → *abdomen*) compiled from medical terminologies.

The method consists in tokenizing source (queries) and target (MeSH) terms into words and in matching the resulting vocabularies after successive normalizations. Both vocabularies were submitted to the same types of normalizations (character- and linguistic-level) and then compared. *Character normalizations* consist in *lowercase conversion* and *deaccenting*. *Morpholexical normalizations* apply lemmatizations based on rules and on *lemma-form* pairs and ‘stemming’ based on *base-derived form* pairs. *Spelling correction* was applied to the remaining words using the Unix `ispell` tool, with the MeSH words as reference dictionary.

## RESULTS AND DISCUSSION

In the original word sets, about one half of source occurrences and only one sixth of unique words matched target CISMef MeSH words. This is a low proportion, which means that without normalization, about one half of user queries would obtain lower-quality results (in Doc’CISMef, queries with terms outside the MeSH generate a full text search on Doc’CISMef records). The largest improvement is obtained by

character-level normalizations (matching of 50.8% of the unique words and 84.9% of the occurrences). The contribution of morphological knowledge is lower (increase to 56.8% of the unique words and 86.8% of occurrences), and requires more initial resources. Here as in many problem-solving situations, beyond a certain point, the additional effort needed grows while the corresponding improvements decrease. However, the contribution of morphological processing is higher in terms of unique words than occurrences; it is therefore useful in order to allow users to submit a larger variety of queries. After spell checking, up to 65.5% of unique words (89.3% of occurrences) are recognized. This evaluation shows the importance of ‘low-level’ processing to match user and indexing vocabularies. It also shows the relevance and limits of morphological knowledge in this task; more information on these experiments is provided in Grabar *et al.*<sup>2</sup> The proportion of remaining unknown words points out the necessity to apply other methods, such as using synonyms<sup>3</sup> or distributional similarities,<sup>4</sup> or to backoff to full-text search. Finally, this method provides a useful measure of the adequacy of users and indexing vocabularies in a natural language query interface to a Health gateway. It should become one of the metrics used for the routine follow-up of Doc’CISMef.

## REFERENCES

1. Jacquemin C and Tzoukermann E. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In: Strzalkowski T, ed, *Natural language information retrieval*. Kluwer Academic Publishers, Dordrecht & Boston, 1999:25–74.
2. Grabar N, Zweigenbaum P, Soualmia L, and Darmoni SJ. Les utilisateurs de Doc’CISMef peuvent-ils trouver ce qu’ils cherchent ? Une étude de l’adéquation du vocabulaire des requêtes des utilisateurs au MeSH. In: JFIM 2002, Québec. 2002.
3. Pouliquen B, Delamarre D, and Le Beux P. Indexation de textes médicaux par extraction de concepts, et ses utilisations. In: JADT, St-Malo. 2002:617–28.
4. Xu J and Croft BW. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 1998;16(1):61–81.