# The Contribution of Morphological Knowledge to French MeSH Mapping for Information Retrieval

**P. Zweigenbaum, Ph.D.,**[1] **S.J. Darmoni, M.D., Ph.D.,**[2] **N. Grabar, M.Sc.,**[1]

[1] DIAM — Service d'Informatique Médicale, DSI, Assistance Publique – Paris Hospitals &
Département de Biomathématiques, Université Paris 6, Paris, France
`{pz,ngr}@biomath.jussieu.fr` `http://www.biomath.jussieu.fr/`

[2] Computer and networks department, Rouen University Hospital, France &
Perception, Information and Systems Lab, National Institute of Applied Sciences, Rouen, France
`stefan.darmoni@chu-rouen.fr`

*Approximate term matching is a key technique for many language processing tasks, such as information retrieval. The UMLS provides tools and lexical knowledge bases for implementing linguistically sound approximate matching of English medical terms. We describe here the design of lexical knowledge bases for performing approximate matching on French medical terms, and the initial evaluation of their contribution to an information retrieval task: access to the MeSH-indexed directory of French-language medical Internet resources (Doc'CISMeF). The observed trend is in favor of the use of morphological knowledge as a moderate but effective factor for improving query to term mapping capabilities.*

## INTRODUCTION

Term matching is one of the important techniques for natural-language-based medical information processing. It is useful for coding medical information into controlled vocabularies such as the International Classification of Diseases or SNOMED[1]. It is also important for Information Retrieval from natural language queries, to map these queries either to controlled index terms such as those of the Medical Subject Headings, or to the content of full text documents such as article abstracts or Web pages (as for instance in NLM's PubMed and Gateway servers to Medline).

Several types of techniques have been designed to perform "approximate" term matching, *i.e.*, to identify target terms that are close to, but not identical to, query terms. String-based techniques consider words and expessions as strings of characters; they can cope with character-level differences such as typos, and can be implemented very efficiently[2]. Stemming is an algorithmic technique for reducing a word to its "stem"[3], removing common affixes, so that words that belong to the same morphological family (*e.g.*, {*probability*, *probabilistic*}) are considered identical for matching purposes. Truncation is a simple approximation of stemming: the user directly specifies a string, which shall match words that start (or end) with this string. The use of linguistic knowledge can provide a more accurate account of morphological variation[4]: inflection deals with the grammatical variation of a single word (number, gender, etc.), derivation adds affixes to a base word form to produce new words (*e.g.*, *infection*, *infectious*) and compounding combines several radicals to obtain complex words (*e.g.*, *hypercalcemia*). Appropriate parsers can deal with syntactic variation, *e.g.*, inserting modifiers and conjunctions, or changing word order and part-of-speech[5]. Semantic techniques often refer to the substitution of non-morphologically-related synonyms (*e.g.*, {*heart*, *cardiac*}), or to the use of links to related notions such as hyponymy or meronymy (*e.g.*, {*myocardium*, *heart*}), typically drawn from structured terminologies. Statistical methods can help identify semantically related words, *e.g.*, by exploiting word co-occurrence in large text collections, as done by the Ovid system (`gateway.ovid.com`). Concept-based matching either is another name for semantic techniques, or assumes a symbolic representation of information on which matching is performed[6].

We focus here on morphology-based techniques for French. It has been shown that stemming techniques, while widely used for the English language in Information Retrieval tasks, are not sufficient for languages with a richer morphology such as French[5]; hence the lack, to the authors' knowledge, of a port of any of the common English stemmers to French. Linguistically-grounded morphological knowledge and techniques are therefore needed. Some work has been done on morphological techniques for French medical language processing[7,8]. Researchers at Xerox Research Center Europe have evaluated the impact of morphological and syntactic techniques on an Information Retrieval task on a French newspaper corpus[9]. They showed that inflectional knowledge brings a significant increase in average precision, whereas derivational knowledge brings a non-significant additional improvement. However, on the one hand, no morphological knowledge base is yet publicly available for French; inflectional knowledge has been available for some time, but derivational knowledge is still a rarer resource. On the other hand, the effective contribution of morphological knowledge to a medical language processing task remains to be assessed.

We present here such knowledge and techniques, and an experiment to evaluate their impact on an Information Retrieval task: matching natural language queries to a controlled vocabulary (basically, French MeSH terms) used to index French web pages (the Doc'CISMeF directory[10]). This experiment relies on a log of actual queries to Doc'CISMeF. Its goal is to assess the differential contribution of inflectional and derivational knowledge to term matching in this context.

## BACKGROUND AND MATERIAL

Doc'CISMeF (D'C, `doccismef.chu-rouen.fr`) is a generic search tool based on an information structure model which encapsulates the MeSH thesaurus. To index resources, D'C uses four levels of hierarchy in its information structure model:[10] "meta-term", keyword, subheading, and resource type. Two levels of searches are currently available: simple search where the end-user can input a single term or expression. If this term belongs to the D'C information structure model, it is "exploded" (substituted with the set of its descendants in that structure). If not, a full-text search is performed on the fields of the D'C pages. In the advanced search, complex searches are possible combining Boolean operators with meta-terms, keywords, subheadings and resource types.

Doc'CISMeF was launched in June 2000. It has received since then a steadily increasing number of queries per day – in January 2001, an average of over 1,200 a day from 400 unique users. We extracted from the http server log all queries sent to the Doc'CISMeF search "servlet". We discarded: $(i)$ the first two months of operation, in order to eliminate potential startup effects; $(ii)$ all queries sent from within the CISMeF team and more widely from the Rouen University Hospital, where users have received a specific training about the MeSH thesaurus (a monthly 2 hour training session), and from the Rouen INSA engineering school, which is linked to the CISMeF team; $(iii)$ all queries performed through the "advanced" interface; and $(iv)$ all empty queries. The number of occurrences of each query is irrelevant, since for technical reasons, one query may result in several lines in the log. Our corpus of queries totals 27,029 queries; among these, the August queries (2,389) were used to debug the system.

The target terms are those used for indexing in the Doc'CISMeF French medical directory: the French MeSH[11] (19,971 terms and 83 qualifiers), augmented with 38 metaterms and 101 resource types, including some accented variants.

Morphological knowledge was automatically derived from SNOMED and ICD-10 in previous work[8]. It includes: $(i)$ pairs of morphologically related words (*e.g.*, {*abdomen*, *abdominal*}, {*abdominal*, *abdominale*}); $(ii)$ morphological rules (*e.g.*, *en*|*inal*); $(iii)$ words and rules tagged with part-of-speech information (*e.g.*, {*muscle/NN*, *musculaire/ADJ*}); $(iv)$ pairs and rules with "lemmatized" words: each word is replaced with its uninflected form (*e.g.*, singular masculine for adjectives; this is the case for the tagged pair above).

## METHODS

### Design of the Lexical Knowledge Bases

We collected from the above data (lemmatized, tagged word forms) a set of pairs {*lemma, inflected form*}, from which we further derived and manually completed inflection paradigms for the words in ICD-10 and $\mu G$. 2906 unique inflection pairs, corresponding to 1224 families (*e.g., apical, apicale, apicaux, apicales*) and 4125 different word forms, were collected. We shall consider that, for term matching purposes, any of the forms in a family is equivalent to the others.

For derivational knowledge, we started from the tagged and lemmatized word pairs. Derivation pairs are generally substitutable for Information Retrieval. Compound pairs are more complex to use, and have been reserved for further investigation. To separate compounds from derivations, we used the following heuristic: most pairs with two differents parts of speech are derivations (*e.g.*, from noun to adjective), whereas most pairs with identical parts of speech are compounds ({*lymphe/NN, lymphoblaste/NN*}). This rough division was then adjusted manually. We finally collected 1042 derivation pairs (794 distinct families, *e.g., aorte, aortique, aorto*) for 1759 lemmas. When merged, inflection and derivation knowledge involve 1600 families and 5462 word forms (*e.g., immun, immune, immunes, immunisation, immunité, immuno, immuns*). Note that these lexical knowledge bases were not specifically prepared for the Doc'CISMeF (MeSH) vocabulary. This will be performed in a forthcoming experiment.

In the usual Information Retrieval paradigm, some words are considered as content-bearing (nouns, adjectives, etc.) whereas the others, called "stop words", are considered as void. We consider as content words all the words in our training vocabularies (ICD-10 and $\mu G$) that were tagged as Noun, Proper noun, Adjective, Abreviation and Prefix (such as *adeno*). The rest make up our list of stop words: determiners, prepositions, adverbs, verbs, pronouns, conjunctions, except a few verbs that we deemed had actual content here (*e.g., oxydant*). This yielded 190 stop words.

### Approximate Term Matching

A query is first segmented into words according to whitespace and punctuation, which are filtered out. All words are transformed into lower case. This results in a sequence of word forms. Stop words are then removed. The remaining are considered "content" words, and may include word forms that are unknown in the target vocabulary.

The next steps ("query expansion") add "equivalent" word forms to each content word if appropriate. These word forms may include reaccentuated or disaccentuated forms if they exist in the target vocabulary; and other inflected forms or derived words depending on the morphological knowledge provided. The result for each input query word is a disjunction of "equivalent" word forms (*e.g., muscle/musculaire/musculaires*); and the result for a query is a sequence of such disjunctions (*e.g., (personnes/personne) AND (agees/age/agee)*).

Each target term is segmented and lower-cased as a query, but not further processed. It is then handled as a "set" of words – *i.e.*, word order and repetition are not significant. Given a query, target terms are ranked first if they (in the specified order): $(i)$ satisfy the largest number of disjunctions (contain the largest number of query words); $(ii)$ have the smallest number of extra words; $(iii)$ contain the largest number of exact words forms from the original query (resort less to "equivalents"); $(iv)$ contain words closer to the beginning of the query; in case of a tie, the final decision is alphabetic order. A "greedy" algorithm is used to successively select target terms in order to "cover" the words of the query.

The algorithm was implemented within an existing term matching program[12] written in Perl5; matching speed is reasonable for testing purposes at about 600 queries/mn on an HP-UX machine.

### Experiments

The matching algorithm was run on all queries of each month in the Doc'CISMeF log corpus. The following strategies were tested: $(i)$ exact match: string identity (baseline); $(ii)$ punctuation and order variants; $(iii)$ inflection equivalents; $(iv)$ inflection + derivation equivalents. For each strategy, we counted the number of answers returned by the algorithm: each answer is a target term that matches all or a part of the query words.

## RESULTS

Table 1 shows a human, qualitative evaluation of results of the non-morphological strategy "order". It was performed on a set of 58 queries with two words or more, that were not exact target terms and had no spelling errors. Rating was performed on a 4-value scale: from 0 (very bad) to 3 (very good). It was obtained by consensus among three people of the CISMeF team (one medical informatician and two medical librarians).

Table 1: Number of queries with $r$-rated answers.

| $r$ = | 0 | 1 | 2 | 3 | mean ± sd |
|---|---|---|---|---|---|
| order | 13 | 11 | 13 | 21 | 1.72 ± 1.18 |

Table 2 shows the results of a quantitative evaluation of the changes in answers to queries when the strategy varies. They were computed over a total of 6469 queries (month 09/2000). The examination of results for other months shows a similar pattern, so that we can focus on this one. 1198 queries (19%) are exact MeSH terms.

Table 2: Number of queries (total 6469) with $n$ answers depending on strategy, and total number of different ($\neq$) answers when going from one strategy to the next.

| $n$ = | 0 | 1 | 2 | 3 | 4 | $\neq$ |
|---|---|---|---|---|---|---|
| exact | | 1198 | | | | |
| order | 1471 | 3696 | 1092 | 181 | 27 | |
| inflection | 1466 | 3703 | 1088 | 183 | 27 | 31 |
| infl+deriv | 1460 | 3716 | 1076 | 187 | 28 | 127 |

As a preliminary evaluation of the contribution of inflectional knowledge, we examined the first 15 queries of the 31 whose results were changed by the use of inflections. They were manually rated on a 0-3 scale as above; the average went up from .53 to 1.07. While some change for a good answer (*e.g., adenome prostatique*, that obtained only *adenome*, now additionally maps to *maladies prostatiques*), others only get more noise (*e.g., hématome intrarachidien* goes from *hematome* to an additional noisy *injection intrarachidienne*). This increase both in recall and in noise requires further investigation.

We also examined the first 10 queries of the 127 whose results were changed by the use of derivations. The average increased from 1.0 to 1.9. The trend here is more positive than for inflection, but will also need a more detailed evaluation, which is under way. Examples of queries with improved answers are *abces hypophysaire*, where the answer goes from (*abces* + *nanisme hypophysaire*) to (*abces* + *hypophyse*), or *dentaire* : (*email dentaire*) → (*dent*). A reduction in quality can occur in examples such as *colites inflammatoires* : (*colite* + *intestin, maladies inflammatoires*) → (*colite* + *inflammation*), which is judged slightly less relevant by our experts (although the initial better result is somewhat obtained by chance).

## DISCUSSION AND CONCLUSION

In the settings of these experiments, morphological knowledge has a moderate but actual impact on query results. This confirms what was observed by the Xerox team[9], but the respective impact of inflection and derivation may here be different. According to our preliminary evaluations, both inflection and derivation have a positive impact on matching results, slightly stronger for derivation. This is an encouragement to continue the construction of more complete and accurate derivational resources.

The presence of English words both in what we considered as the French target terms and in some of the queries lead to some erroneous morphological equivalents. The target terms should be cleaned before further experiments take place.

This study has several limitations. First, human evaluation needs to be performed on a larger scale – this is under way and should be completed shortly. The off-line assessment does not evaluate end user satisfaction in an actual information search situation: are the results useful, do the proposed target terms lead to the information needed? This is planned for subsequent investigation when this prototype is integrated and can be used on the Doc'CISMeF site. This work does not compare the proposed method and knowledge with others: other morphological methods, related terms, etc. This would require a common benchmark to be set up: a common test collection, with queries, target terms and gold standard answers. If other parties are interested, this could be the subject of further work.

Let us stress that this experiment mainly involves a single technique: morphology-based query expansion. It is meant as a methodological test for that technique in isolation. It is only one block in a series of complementary techniques, currently in construction, that are to work in cascade to match queries to target terms.

Finally, our implementation of this technique can be improved in several ways. One of them is the handling of stop words: they are currently completely ignored in queries. However, they could intervene in the ranking algorithm: when several target terms contain the same number of content words of the query, those that also contain stop words of the query should be given priority. Another path for improvement consists in expanding and adapting our morphological knowledge over the target vocabulary, and in training the term matching algorithm on the log queries.

## References

1. Wingert F, Rothwell D, and Côté RA. Automated indexing into SNOMED and ICD. In: Scherrer JR, Côté RA, and Mandil SH, eds, *Computerised Natural Medical Language Processing for Knowledge Engineering*. North-Holland, Amsterdam, 1989:201–39.

2. Lovis C and Baud R. Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language. *J Am Med Inform Assoc* 2000;7(4):378–91.

3. Porter MF. An algorithm for suffix stripping. *Program* 1980;14:130–7.

4. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235–9.

5. Jacquemin C and Tzoukermann E. NLP for term variant extraction: Synergy between morphology, lexicon, and syntax. In: Strzalkowski T, ed, *Natural Language Processing and Information Retrieval*. Kluwer, Boston, Mass, 1997.

6. Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, and Boisvieux JF. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. *J Am Med Inform Assoc* 1997;4(suppl):590–4.

7. Lovis C, Baud R, Rassinoux AM, Michel PA, and Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. *Artif Intell Med* 1998;14:201–14.

8. Grabar N and Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. *J Am Med Inform Assoc* 2000;7(suppl):310–4.

9. Gaussier E, Grefenstette G, Hull D, and Roux C. Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues* 2000;41(2):473–93.

10. Darmoni SJ, Leroy JP, Thirion B, et al. CISMeF: a structured health resource guide. *Methods Inf Med* 2000;39(1):30–5.

11. Institut National de la Santé et de la Recherche Médicale, Paris. Thésaurus Biomédical Français/Anglais, 2000.

12. Blanquet A and Zweigenbaum P. A lexical method for assisted extraction and coding of ICD-10 diagnoses from free text patient discharge summaries. *J Am Med Inform Assoc* 1999;6(suppl).