Comparative study between expert and non expert biomedical writings: their morphology and semantics

Jolanta CHMIELIK[a] and Natalia GRABAR[a,b,1]
[a] *INSERM, UMR_S 872, eq.20 Paris, F-75006 France; Université Paris Descartes, Paris, F-75006 France;* [b] *HEGP AP-HP, Paris, France*

Abstract. *The amount of health information on the Internet is constantly growing but little is done for detecting the technicality level of these documents and guiding of users towards documents which are appropriate to their expertise level. The objective of our work is to propose clues for the automatic distinction between expert and non expert medical documents. More precisely, we propose to study their morphological and semantic levels. We apply NLP tools, which provide access to the morpho-semantic content of documents. The work is done with French documents within three medical fields (cardiology, pneumology, diabetes). Our experiments and results highlight the fact that this level can indeed provide clues for the distinction of the technicality of documents, and that they appear to be significant and stable across the studied medical fields.*

Keywords. Natural Language Processing, Access to Information, Language, Documentation, Medical informatics

1. Introduction

A recent analysis of Internet use [1] shows that above 80% of user queries are related to medical topics. This situation requires high quality health information, and initiatives like CISMeF [2] and HON [3] contribute to the objectives of improving quality and ethical transparency of webpages. Besides, health documents also present heterogeneity as for their technicality, expressed for instance through the use of specialized terms and words: expert documents (created and used by health experts) coexist with non expert documents (created for non expert users), as well as didactic documents (created for medical students). According to their technicality, documents can be more or less difficult to understand, especially to non expert users. This heterogeneity is not transparent, while it should be clearly indicated: it has been observed that heavy technicality may have a negative effect on understanding health information by non expert users and on their communication with caregivers [4,5]. In order to guide non expert users towards suitable documents, search engines should make the distinction between expert and non expert documents. Currently, such distinction is based on manual categorization [2-3,6], but with the increasing amount of information it is necessary to perform it automatically. Definition of suitable criteria is an important step towards the automatic categorization. Thus, readability formulae [7-9] take into

[1] Corresponding Author: Natalia Grabar (natalia.grabar@spim.jussieu.fr)

account criteria such as the mean length of words or sentences (longer words are assumed to be heavier to understand). The formulas can be combined with medical terminologies to untangle the medical dimension [10], because short medical words can also be difficult to understand. Widely applied machine learning algorithms are based on different features (n-grams of characters [11], manually [12] or automatically [13] defined weight of terms, stylistic [14] or discursive [15] criteria, lexicon [16]). Recently, combination of features has been addressed [17-19]. However, detailed studies of expert *vs*. non expert medical discourses [20-22] remain rare. In our work, we are interested in studying the level of morpho-semantics, which has not been studied up to now. Our objective is to propose a morpho-semantic description of expert, didactic and non expert health documents in order to perform an automatic distinction between these discourses.

2. Material and Methods

Building corpora. Our corpora are collected through the CISMeF portal [2]. We exploit three types of information assigned to each indexed resource: (1) MeSH key-words (cardiology, pneumology and diabetes); (2) discourse of documents: expert, didactic and non expert; (3) URLs of documents. Documents were downloaded with *wget* tool. HTML and XML documents were converted in text format. We built three corpora (*cardiology, pneumology and diabetes*) each of them consisting of three parts (*expert, didactic and non expert*). Table 1 indicates the sizes of these corpora: number of documents and number of occurrences (words). It can be seen that their sizes are uneven. This material provides fundamental characterization of documents and constitutes the reference data.

Table 1: Characteristics of corpora

Specialties	*Number of documents*			*Number of occurrences*		
	Expert	*didactic*	*non expert*	*expert*	*didactic*	*non expert*
Cardiology	1 583	205	143	942 409	449 765	157 382
Pneumology	742	127	134	600 524	213 379	96 559
Diabetes	213	23	52	181 039	44 847	29 817

Accessing the morpho-semantic level. We apply Natural Language Processing (NLP) tools were applied in order to reach the morpho-semantic level of documents. TreeTagger [23] assigns to each word its most probable morpho-syntactic tag and proposes its lemma: *angioblastiques* is tagged as adjective (ADJ) and lemmatized to *angioblastique/ADJ*. Flemm [24] checks out the lemmatization, corrects and enriches it. Finally, DériF [25] performs the morpho-semantic analysis of lemmas. Two types of information this provides were used, *ie.* for *angioblastique/ADJ*: (1) Decomposition into morphological tree *[[angi N*] [blast N*] ique ADJ]* and (2) representation of the meaning in natural language: *"Qui est en relation avec cellule embryonnaire et vaisseau" ("Which is related to embryonic cell and vessel").* The following steps of the method exploit this morpho-semantic analysis.

Preparing the morpho-semantic material. For each specialty and discourse, bases that are the most productive were selected, *ie.*, those having large morphological families. For instance, base *cardio* constructs up to 57 lexemes in *cardiology* corpora (its morphological family contains 57 lexemes). If a complex lexeme (such as *angioblastique/ADJ [[angi N*] [blast N*] ique ADJ]* contains more than one base, this lexeme belongs to all the corresponding families (*angi* and *blast* in this example).

Studying and contrasting discourses. In order to detect salient morpho-semantic features a contrastive analysis was performed through the study of their productivity and frequency within the corpora. Productivity is studied through the number of lexemes constructed by each base: the more a base is productive the larger its family is. Two values were taken into account when measuring the productivity: raw and normalized. Raw values corresponded exactly to the number of constructed lexemes. Normalized values were normalized by the size of corresponding corpora. Raw and normalized frequencies of bases in corpora were studied on lemmatized material.

3. Results and Discussion

Selected morphological material. Bases selected for this study belong to two morphological categories: (1) suppletive bases, which do not appear independently in the language but are combined with other morphological units, such as *gastr(o)* realized through *gastrique* and, (2) autonomous bases, such as *bronches (bronchus)* and *bactérie (bacterium)*. On the basis of their productivity, we selected 38 suppletive and 7 autonomous bases. During this step, some limits of the morpho-semantic analysis by DériF were discovered: (1) reference lexicon missed some medical words which remained not analyzed, (2) some affixes (*ie, -eux*) were not yet implemented, (3) erroneous morpho-semantic analysis, for instance *gymnasium* were analyzed as *"enzyme of the nude",* (4) notation of bases identified by Dérif: thus, four bases (*hém(o), héma, hémat(o)* and *èm*) had different graphical forms although they convey identical meaning *"relative to blood*". This last fact can be awkward for automatic approaches, *ie.* for grouping lexemes within families. To resolve this situation, it is possible to either establish the equivalence between bases or to take into account their semantics (for instance, the four "blood" bases have the same or similar semantics *"related to blood, linked to blood").* It was decided to apply the first solution. Thus, 46 morphological families were built which contain 2,295 lexemes. Most of them occur in all corpora, but some families remain specific to some specialties.

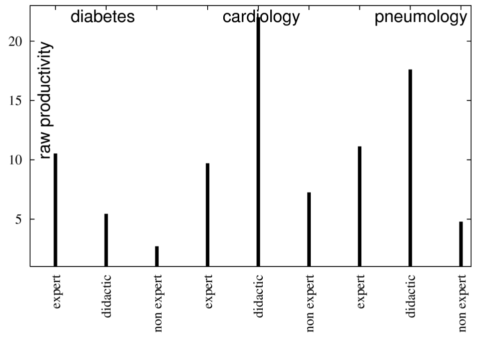

Figure 1a. Raw productivity of bases.

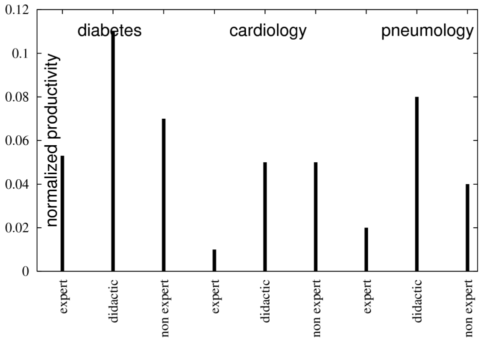

Figure 1b. Normalized productivity of bases.

Productivity of morphological bases. Productivity of each base corresponds to the size of its family, *ie.*, the number of lexemes it produces in a corpus. Figure 1a indicates mean raw sizes of the studied families and figure 1b the corresponding normalized values. It was clearly seen that productivity varies according to specialties and discourses. As the diabetes corpora are small, less attention was paid to these results. It was also observed that, among all the discourses, didactic corpora show the highest productivity. Indeed these documents provide precise and detailed medical information and usually have to introduce a great variety of medical notions. Otherwise, when expert and non expert corpora were compared, normalized values of productivity were higher in non expert corpora (fig. 1b), while the raw values were higher in expert corpora (fig. 1a). Indeed, expert corpora were six times larger than expert corpora, and this automatically decreases their normalized values. Analysis of corpora with more comparable sizes would certainly lead to different results. It is interesting to note that previous work [10] has also observed that the vocabulary in expert corpora is richer as compared to non expert corpora, but the authors studied only raw values.

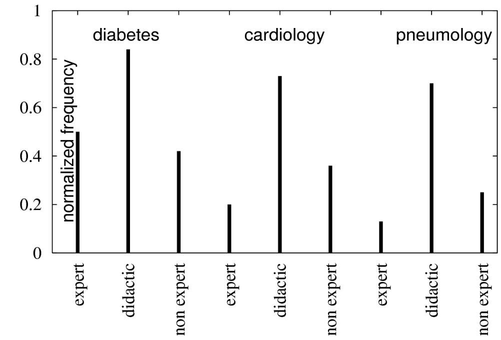

Figure 2. Frequencies of bases in corpora, normalized values.

Frequency of morphological bases. Figure 2 presents normalized frequencies of morphological bases (through their lexemes) within corpora. Didactic corpora show the highest frequencies in addition to the highest productivity observed in previous paragraph. Non expert corpora are in the second position, and expert in the third position: expert documents often address precise questions and seem to use smaller vocabulary.

4. Results and Discussion

In this work, we proposed a detailed analysis of the morpho-semantic level of expert, didactic and non expert French health documents in order to prepare their automatic distinction. We studied productivity and frequency of 46 morphological bases *(ie., angi-, blast-, cardio-).* We observed several morpho-semantic characteristics specific to discourses: didactic documents show higher frequency of bases (repetition reinforces learning) and their bigger productivity (bigger variety of vocabulary) and can be easily differentiated from the two other discourses. As for the comparison of expert and non expert documents, we observed that: (1) frequency is higher in non expert corpora; (2) raw productivity of bases is higher in expert corpora, while (3) normalized productivity is higher in non expert documents. The main perspective of our work consists of the use of the morphological criteria for automatic distinction of medical discourses within specialized portals: the obtained results suggest this would be possible. Otherwise, this study can be extended to a larger set of morphological material (currently, only 46 bases were studied) and of specialties, and could exploit documents from other sources *(ie.*, Santé Canada). Moreover, this approach can be applied to other languages as far as suitable NLP tools exist. Finally, it would be interesting to better exploit the productivity, which seems to be specific to discourses and genres [26].

References

[1] Fox S. (2006). On line Health Search 2006. Most Internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find. Technical report, Pew Internet & American Life Project, Washington DC

[2] www.chu-rouen.fr/cismef: Catalogue et index de sites médicaux francophones. (Last accessed on 24/04/09)

[3] www.hon.ch: Health on the Net Foundation. (Last accessed on 24/04/09)

[4] AMA (1999). Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. JAMA, 281(6), 552-7

[5] McCray A. (2005). Promoting health literacy. Journal of American Medical Informatics Association, 12, 152-163

[6] www.google.com/coop: access on inscription (Last accessed on 24/04/09)

[7] Flesch R. (1948). A new readability yardstick. Journal of Applied Psychology, 23, 221-233

[8] Gunning R. (1973). The art of clear writing. New York, NY : McGraw Hill

[9] Björnsson & Härd af Segerstad, Segezstad B. (1979). Lixpä franska och tio andra spräk. Stock holm:Pedagogiskt centrum, Stockholms skolförvaltning

[10] Kokkinakis D, Gronostaj MT. (2006). Comparing lay and professional language in cardiovascular disorders corpora. In WSEAS Transactions on BIOLOGY and BIOMEDICINE, pp. 429-437

[11] Poprat M., Marko K. & Hahn U. (2006). A language classifier that automatically divides medical documents for experts and health care consumers. In MIE 2006, pp. 503-508

[12] Zheng W, Millos E., Watters C. (2002). Filtering for medical news items using a machine learning approach. In AMIA, pp. 949-53

[13] Borst A, Gaudinat A, Boyer C, Grabar N. Lexically based distinction of readability levels of health documents. MIE 2008

[14] Grabar N, Krivine S, Jaulent MC. (2007). Classification of health web pages as expert and non expert with a reduced set of cross-language features. In AMIA 2007, pp. 284-8

[15] Goeuriot L, Grabar N, Daille B. (2007). Caractérisation des discours scientifique et vulgarisé en français, japonais et russe. TALN 2007

[16] Miller T, Leroy G, Chatterjee S., Fan J., Thoms B. (2007). A classifier to evaluate language specificity of medical documents. In HICSS, pp. 134-140

[17] Wang Y. (2006). Automatic recognition of text difficulty from consumers health information. In IEEE, Ed., Computer-Based Medical Systems, pp. 131-136

[18] Zeng-Treiler Q., Kim H., Goryachev S., Keselman A., Slaugther L. & Smith C. A. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. In MEDINFO, pp. 1117-1121

[19] G. Leroy; S. Helmreich; J. Cowie; T. Miller; W. Zheng. 2008. Evaluating Online Health Information: Beyond Readability Formulas. In AMIA 2008, pp. 394-8

[20] www.consumerhealthvocab.org

[21] Deléger L, Zweigenbaum P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In AMIA 2008, pp. 146-50

[22] Zeng Q., Tse T. (2006). Exploring and developing consumer health vocabularies. JAMIA, 13, 24-29

[23] Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, pp. 44-49

[24] Namer F. (2000). Flemm: Un analyseur Flexionnel de Français à base de règles. Traitement automatique des Langues, 41(2), pp.523-47

[25] Namer F. (2003). Automatiser l'analyse morphosémantique non affixale: le système DériF. Cahiers de Grammaire, 28, 31-48.

[26] Baayen H. (1994). Derivational productivity and text typology. Journal of quantitative linguistics, 1(1), pp. 16-34