Incorporation of two terminology projects into a system for information retrieval using NLP for term expansion

Klaar Vanopstal and Kurt Van Wiele
Language and Translation Technology Team (LT3)
University College Ghent
Groot-Brittanniëlaan 45
9000 Gent, Belgium
klaar.vanopstal@hogent.be, kurt.vanwiele@hogent.be

Abstract

In this paper, we will discuss two medical terminology projects at the University College of Ghent, Faculty of translation studies, and the benefits of combining them to provide Dutch professionals and laymen with better access to information in biomedical databases. Our first project, the *MeSH Termbase Project* (MTB) is aimed at health care professionals, medical translators and also patients in need of language support. The main aim of our second project, the *Multilingual Glossary of Technical and Popular Medical Terms*, is the simplification of the terminology used in patient information leaflets.

Keywords: medical terminology, morphology, MeSH, patient information leaflets

1. Introduction

With the information explosion, access to medical information is both facilitated and hindered. Dutch-speaking patients who perform searches with lay terminology are often overwhelmed by the abundance of -sometimes unreliable- information, for instance on blogs, forums and self diagnosis sites. On other occasions, their searches may be inefficient and sometimes yield few or no relevant results, because they do not know the essential medical terminology, or because of to the English language barrier they are confronted with.

Suppose a Dutch-speaking patient with a blood coagulation disorder wants to search for some more information about his condition. A feasible task in the internet era, one might think. He googles for the term *samenklontering van bloedplaatjes* («coagulation of blood platelets»), and gets 688 hits, the first of which is about scleroderma and the adverse effects of its treatments. The second link leads him to an enumeration of medicines, where *Plavix* is listed as a treatment for people who underwent a «balloon dilatation» (*ballondilatatie*) and «stent implantation» (*plaatsing van stent*). This information is very unlikely to make the patient feel any wiser about his condition. So he tries PubMed. This means that he will have to translate the term into English first. The Van Dale dictionary (Martin, 2006), a leading Dutch dictionary, gives two possible translations for *bloedplaatje*: «(blood) platelet» and «thrombocyte». As thrombocyte sounds much more scientific, he opts for this translation. The same dictionary gives «coagulate» as a translation for *samenklonteren*. Combined, this gives «thrombocyte coagulation» as a "scientific" translation for *samenklontering van bloedplaatjes*. When entered as a query in PubMed, this term yields the message "Quoted phrase not found". By this time, the patient will feel at least a little frustrated. If he had known that the correct MeSH term was «blood coagulation disorder», he

would at least have been one step further in his search. Nevertheless, he will still feel confused by the overwhelming amount of -expert- information.

Furthermore, patients may feel distressed when they read patient package inserts, which are often translated from English. Unlike Dutch, English has a strong tendency towards using Latin-based terminology (Zethsen, 2004:125), even in non-expert communication. These Latin-based terms are often copied indiscriminately in the Dutch translation, making the information less accessible to the average reader. Askehave and Zethsen (2000) describe this as a problem of *inter-linguistic translation*. Another type of translation mentioned by Askehave and Zethsen (2000), *inter-generic translation,* may also cause problems. Patient information leaflets are usually translated from scientific leaflets (the SPCs). During the translation process, register is often disrespected, and expert language may be transferred directly into the patient information. The following extract demonstrates the sometimes incomprehensible wordings used in medical information:

> Na de toediening van **zoledroninezuur** werden nierstoornissen waargenomen, met name bij patiënten met een reeds bestaande nierstoornis of bijkomende risicofactoren (bijv. **oncologiepatiënten** met chemotherapie, gelijktijdige **nefrotoxische** geneesmiddelen, ernstige **dehydratatie**, enz.); dit wordt beschouwd als een **klasse-effect**. Alhoewel niet waargenomen tijdens klinische studies naar de botziekte van Paget zijn andere klasse-effecten **iritis**, **uveïtis**, **episcleritis**, **conjunctivitis** en **osteonecrose** van de kaakbeenderen.

> (Renal dysfunction has also been observed following the administration of **zoledronic acid**, especially in patients with pre-existing renal compromise or additional risk factors (e.g **oncology patients** with chemotherapy, concomitant **nephrotoxic** medications, severe **dehydration**, etc), and is known to be a **class effect**. Although not observed in the clinical studies in Paget's disease, other class effects include **iritis**/ **uveitis**/ **episcleritis**/ **conjunctivitis** and **osteonecrosis** of the jaw.)

The projects described in this paper are aimed at providing language support to Dutch-speaking lay people, medical translators and professionals whose knowledge of the English language is insufficient to perform their searches confidently in English.

2. MeSH Termbase project (MTB)

Since 1987, students at the Faculty of Translation Studies have been working on an English-Dutch termbase (Buysschaert, 2006) based on the Medical Subject Headings (MeSH). The aim of this project is twofold: first, to build a MeSH-based bilingual thesaurus for translators, terminologists and other users interested in bilingual biomedical information, and second, to provide language support in biomedical information or document retrieval. Until now, one fifth of all concepts in this rich controlled vocabulary have been translated into Dutch, and arguably the most important part, chapter C is nearing completion. For each of these terms, a thorough terminological study was performed. The resulting terminological records include grammatical and morphological information (word class, inflection), definitions and contexts, reliability of the translation, semantic information (related terms, broader and narrower terms, (quasi-)synonyms), international terms, etc. These data may be of interest to translators, terminologists, medical staff and laymen in search of biomedical information. A summary of these terminological records has been stored in a database,

which was then linked to the MeSH tree. A demo version of this tree can be found at http://veto.hogent.be/mesh. The termbase was intended to serve multiple purposes, such as indexing, document retrieval and NLP processing. Moreover, this termbase is a particularly useful instrument for medical translators, which will save them time and effort when researching terms.

3. Peculiarities of the language of medicine

The language of science, and more specifically the language of medicine, is often criticized as "opaque" (Maylath, 1997:32) and "unemotional" (Kronick, 2007:73). Jolly (1987:1) labels it as "med-speak", a sublanguage characterized by "obscure, highly technical terms that cloak the reality of the patient's condition". Mintz (1992:18) explains this discrepancy between the language of doctors, which he designates as "medicalese" and that of patients as a strategy to "avoid confronting the intensely personal and upsetting implications of the disease". Medicalese creates a distance between doctor and patient, allowing the doctor to block out the emotional aspect of his or her work.

3.1. The morphology of English vs. Dutch medical terminology

3.1.1. General observations

For centuries, Greek and -especially- Latin were the languages of science. The style and language of scholarly written texts has not affected the vernacular production in Dutch-speaking countries as radically as in English-speaking countries. As a result, two parallel Dutch languages were created: one for professionals, and one for the common man.This phenomenon is described by Corson (1985:35) as *lexical apartheid*. Admittedly, Latinates have some advantages over vernacular terms: they create terminological continuity and consequently increase the efficiency of medical communication. Moreover, the use of Greek and Latin affixes is an extremely productive word formation technique. This means that not all Greco-Latinates in the medical language are of ancient provenance: many new terms for new concepts in the rapidly growing science of medicine of the past century were based on Greek or Latin. These neologisms, also called neoclassical compounds, resemble the ancient Greek or Latin words to such a degree that it is almost impossible to make a distinction between the borrowings and neologisms (Banay, 1948:2). However, these advantages, i.e. continuity and effective communication only apply to professionals, whereas lay people who want access to medical information are somewhat neglected in this respect. Moreover, as observed above, doctors use these technical, foreign sounding terms in order to avoid confrontation with the emotional side of their job. This builds up an emotional barrier between doctor and patient and leaves the patient overwhelmed (Mintz, 1992:230).
    Although English Greco-Latinates are also considered as being more difficult than Anglo-Saxon words, English has a stronger borrowing tradition than Dutch. Banay (Banay, 1948:18) rightfully claims that "English is to some extent Germanic in form and part of its vocabulary is Germanic, but a considerable section is of Latin ancestry". Consequently, Latin- or Greek-based terms are accepted more easily in English and –

unlike in Dutch- there tends to be no difference between expert terms and non–expert terms. Some examples:

- «dental arch» («dental» is used by experts as well as lay people) vs. *tandboog* (*tand-* is a non-expert Dutch alternative for the expert term *dentaal*.)
- «spinal disorder» («spinal» can be considered as both an expert and a non-expert term in English) vs. *aandoening van de wervelkolom* (*wervel-* in compounds is a more popular term for the expert term *spinaal*)

3.1.2. MeSH experiment

A short contrastive analysis of the morphology of English and Dutch medical language was performed in order to find illustrate the nature and extent of the language barrier lay people are confronted with when searching for medical information. For this analysis, two different experiments were performed. The first experiment includes a manual annotation of 500 MeSH terms, which were randomly selected. Six different labels were used:

- **Vern**: vernacular forms, morphologically "pure" Dutch or English forms. e.g. *snake bite* (En), *slangenbeet* (Du);
- **LG**: Latin or Greek terms which have not undergone any adaptation to the vernacular language. e.g. *myotonia atrophica;*
- **En**: Only applies to the analysis of the Dutch terms. English terms used in Dutch without morphological adaptation. e.g. *lumpy skin disease;*
- **Int**: internationalisms, based on Latin or Greek but with morphological adaptations. e.g.: **intra**<u>hepat</u>**ic** *cholestasis* (En), **intra**<u>hepat</u>**ische** *cholestase* (Du), where the prefixes (intra-), suffixes (-ic, -ische, -is, -ase) and confixes (hepat-) indicate the Greco-Latin provenance of the terms. These terms appear in several languages, similar in form and etymology;
- **Par**: paraphrased terms (multi-word terms). e.g. *aandoening **van** de cervix uteri* (*cervix disease*). These terms are combinations of Dutch words and internationalisms linked by a preposition;
- **Other**: terms which do not have Germanic or Romance origins. e.g. *lichen planus* (Turkish);

The Oxford English Dictionary was consulted for etymological information, together with a list of prefixes, suffixes and confixes (Banay, 1948). As the experiments evolved, new affixes were added to this list.

	Vern	**LG**	**En**	**Int**	**Par**	**Other**
Dutch terms	171	33	9	264	22	1
	34,2%	6,6%	1,8%	52,8%	4,4%	0,2%

Table 1 - Analysis of the origins of 500 Dutch equivalents of MeSH terms in MTB

From the table above, it can be concluded that only one third (34,2%) of the Dutch MeSH terms are "pure" Dutch terms, with a typically Dutch morphology. The rest, i.e. more than 65%, are "foreign" terms, with different degrees of adaptation. On a continuum line showing the degrees of adaptation of terms in Dutch (cf. figure 1),

Latin/Greek and English terms are situated at one end, and pure Dutch terms on the other. In between are the internationalisms: loan words (33,3%), showing a low degree of adaptation, and hybrid terms (19,5%), i.e. a combination of loan word and vernacular word, thus showing a slightly higher degree of adaptation. These hybrid terms are also called semi-neoclassical terms. Paraphrases can also be situated on the right-hand side of the continuum, as they consist of a Dutch part, combined with an internationalism.
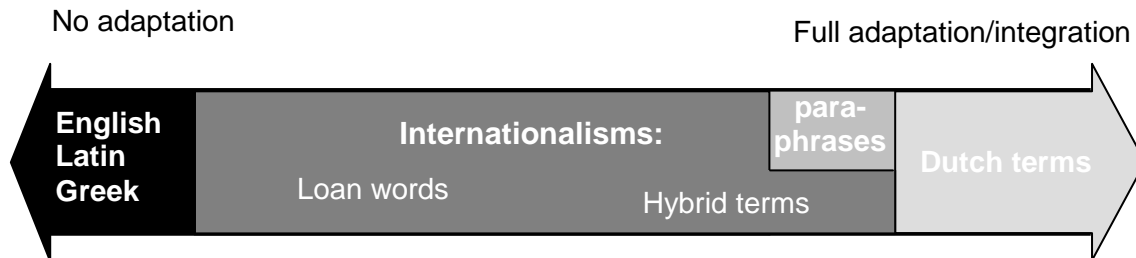
No adaptation   Full adaptation/integration

Figure 1 – Adaptation/integration of medical terms into the Dutch language

A comparison with the English MeSH terms (cf table 2) shows that English has even less "purely vernacular" medical terms (24%) than Dutch. The majority are Greco-Latinates, terms which carry traces of Greek or Latin origin. The number of internationalisms is even higher than in Dutch (68,4%). This may be explained by the historical evolution, i.e. the strong borrowing traditions of the English language, and the significance of Latin and Greek as the languages of science in general (cf 3.1.1).

	Vern	**LG**	**En**	**Int**	**Par**	**Other**
English	120	36		343	0	1
	24%	7,2%		68,4%	0%	0,2%

Table 2 - Analysis of the origins of 500 English MeSH terms

For most of the internationalisms, equivalence has been established between the terms in both languages. The morphological similarities between English and Dutch can be seen as a validation of the equivalence between the terms, yet only on the denotational level. In many cases, this equivalence is found only on the denotational level, whereas the connotational aspect of the terms, namely the register, is not always respected. The differences in term formation, especially in borrowing traditions, make inter-linguistic and inter-generic translation of medical texts, including patient information leaflets, a complex and delicate task. Both experiments show that a large number of medical terms have a foreign aspect. The direct transfer of these foreign sounding expert terms into Dutch may undermine the comprehensibility of medical language (Zethsen, 2004:126) and thus form a language barrier for ordinary people who are searching for medical information.

From a morphological point of view, the most interesting category in medical terminology is clearly that of the internationalisms, both in Dutch and in English. For a more detailed study of their morphological characteristics, a Perl-script was written to detect Latin and Greek affixes. We first separated all multi-word terms into their different components. Next, the unique terms were filtered, which resulted in 722 English and 509 Dutch words, which were then compared to a list of prefixes and suffixes using a Perl script. For the purpose of this analysis, confixes in initial position

were considered as prefixes, and confixes in final position as suffixes. Table 3 shows the results of this comparison.

	prefixes	suffixes	unique terms
English	459	459	596
Dutch	306	323	401

Table 3 - Morphological analysis based on prefixes and suffixes

On a total of 722 English internationalisms, 459 Greco-Latinate prefixes (64%) were counted, and 459 suffixes (64%). The figures for Dutch are comparable: 306 Dutch internationalisms (60%) have a Greco-Latinate prefix, and 323 (63%) have a Greco-Latinate suffix. In total, 83% of the English internationalisms and 79% of the Dutch terms could be detected, based purely on their prefixes and suffixes. This means that detection based on affixes is a relatively reliable technique to isolate foreign terms, which in a next step could be replaced with more popular terms (cf. section 6).

4. The Multilingual Glossary of Popular and Technical Medical Terms

It is generally known that patient information leaflets are not accessible to lay people. One of the main problems is the abundance of specialist terms used in such information leaflets. It should not come as a surprise that only a small minority of people - especially those who are dealing with medical terminology on a day to day basis - are acquainted with medical terms as *xeroftalmie* («xerophthalmia») and *steriliteit* («sterility» referring to "the inability to produce offspring").
The presence of specialist terms in Dutch patient information leaflets can, however, easily be explained. First of all, it should be noted that the Dutch drug market is rather small in comparison to, for example, the English market. As a consequence, a great deal of Dutch information leaflets are not more than mere translations of their English counterparts, which brings along inter-linguistic translation problems, as mentioned above. The fact that medical translators do not replace technical medical terms as "sterility" (which is both the technical and the popular term in English) with *"onvruchtbaarheid"* (the popular equivalent for the Dutch technical term *steriliteit*) in the Dutch translation should be attributed to inaccuracy rather than inexperience.
Besides inter-linguistic translation problems, inter-generic translation problems (cf. supra) should also be mentioned as an explanation for the multitude of technical terms which can be found in Dutch patient-oriented leaflets. Originally, leaflets were enclosed in the drug package to provide information for the general practitioner rather than for the patient. Or to put it differently, they were written for specialists by specialists. When "translating" these technical leaflets, the medical translator is often unaware of the existence of the popular terms to replace the specialist terms with. And even if the translator is aware of their existence, it will often cost him a great deal of time to retrieve the popular equivalent.

Nevertheless, patient information leaflets contain a wealth of information not only for the medical specialist, but for the patient as well. Think for example of therapeutic indications, contraindications, method and route(s) of administration and side effects. This, combined with the fact that GPs nowadays often do not have the time to read all

information leaflets, seems an absolute must to make leaflets more legible to the general public, so as to avoid, for instance, improper use of medication or to help the patient to diagnose possible side effects. It is therefore advisable to use "*droge ogen*" («dry eyes») instead of "*xeroftalmie*" («xerophthalmia») in the future.

The problems described above were, moreover, the initial impetus for the creation of the *Multilingual Glossary of Technical and Popular Medical Terms* in 1995. Belgium had already been a pioneer in 1984 by introducing a law that stated that "package leaflets have to be written in such a way that they are legible for adults with the educational level of a sixteen-year-old" (compulsory school age in Belgium was at the time set at sixteen; later on it was increased to eighteen years of age). Two years later the *Drietalig vocabularium van wetenschappelijke en populaire medische termen* (Trilingual vocabulary of scientific and popular medical terms) was issued under the authority of the Belgian Ministry of Health. This vocabulary, containing 1,400 scientific medical terms and their popular equivalents in the three official languages spoken in Belgium (Dutch, French and German), was edited by P. Van Hauwermeiren, the former Dean of University College Ghent. Belgium was, furthermore, together with Switzerland the first country to introduce patient information leaflets. In that respect, approximately 6,000 technical leaflets were replaced with patient information leaflets in Belgium between 1988 and 1992.

Following the Belgian example, Europe decided that henceforth drug packages had to contain a comprehensible -patient- information leaflet. In 1992 the then EEC issued Directive 92/27/EEC which would not come into effect until 1994. Article 8 of this Directive stipulates that

> "the package leaflet must be written in clear and understandable terms for the patient and be clearly legible in the official language or languages of the Member State where the medicinal product is placed on the market. This provision does not prevent the package leaflet being printed in several languages, provided that the same information is given in all the languages used".

The fact that the then twelve member states of the EEC were obliged to comply with this directive, created the need for a multilingual glossary in the nine languages spoken in the EEC at that moment (EN, NL, FR, DE, ES, PT, IT, EL & DA). This was the immediate cause for setting up the *Multilingual Glossary of Technical and Popular Medical Terms* in 1993, which was completed two years later in 1995.

For this project, the European Commission appealed to the Heymans Institute for Pharmacology of University Ghent given their expertise in this field, as they had been entrusted with the evaluation of the Belgian patient information leaflets project in 1988 (cf. above). The Heymans Institute, however, needed the cooperation of an institution specialized in translating disciplines. Preference was given to University College Ghent because of two reasons. Firstly, because the major European languages are taught at University College Ghent, namely English, French, German, Spanish and Dutch. Secondly, the institution had acquired some experience in medical terminography with the publication of a trilingual medical glossary under the authority of the Belgian Ministry of Health (cf. above).

The result of this project was a multilingual glossary containing 1,830 technical medical terms and their popular equivalents in the nine European languages mentioned earlier.

Moreover, it was decided to add an English definition to each of the 1,830 English technical terms so as to avoid polysemy-related problems, since it had soon become clear from practice that one and the same technical term can often refer to several concepts. The result was put on the web (http://users.ugent.be/~rvdstich/eugloss/welcome.html) at the disposal of the general public, where it can still be consulted. In view of the continuing EU interest in the information leaflet problems (think for example of the creation of EMeA, the European Medicines Agency, in 1995), it was recently resolved to commence with an update of the glossary since much of the information in it has become out of date over the last few years. This update is, however, yet to be completed.

5. Combinatory approach to both projects

The integration of both projects into search systems may solve the problems of inter-linguistic and inter-generic translation and provide a useful basis for language support in biomedical information retrieval. By linking the Dutch MeSH vocabulary to the English index terms of search engines, the efficiency and chance of success of searches performed by Dutch-speaking users may be increased considerably. A similar system has been designed and tested by Plovnick and Zeng (2004). They mapped consumer-generated terminology to the UMLS Metathesaurus terminology. Reformulation of patient queries improved their search results in 42% of the cases.

Moreover, the integration of the Dutch popular terms can make the system and the information it contains more accessible to the general public. As the MeSH controlled vocabulary is used for indexing in numerous biomedical databases, the Dutch termbase could be used in many databases.

In this particular case, the vocabulary will be integrated into CEBAM's (Centre for Evidence-Based Medicine) Virtual Medical Library (http://www.iscientia.net/federate/cgi-bin/query-meta.exe?v:frame=form&fr, cf. figure 3). This library has its own search engine using a federated search system (cf. figure 2) which cooperates with databases such as PubMed, BCFI Folia, NHS Guidelines Finder, Cochrane Library, Journals@Ovid Full Text etc.
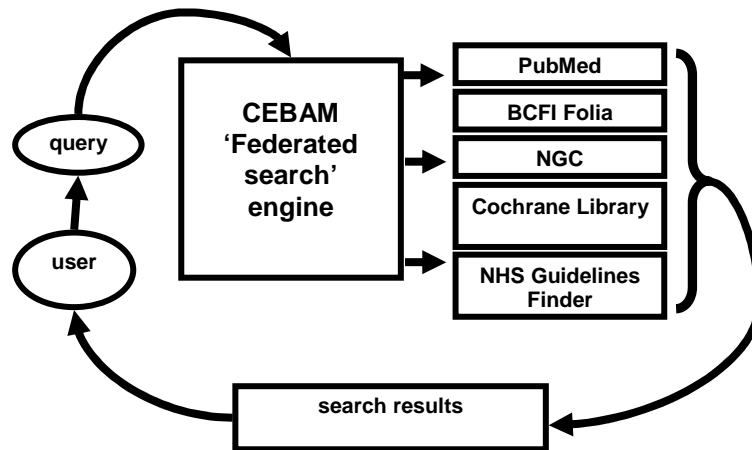
Figure 2 – Federated search

Once the MeSH termbase has been incorporated into this retrieval system, a Dutch query will be linked automatically to the English MeSH term and documents indexed with these English descriptors will easily be retrieved.

> e.g.: query= "miltvuur" [Dutch term] ➜ "anthrax" [translated automatically; MeSH descriptor] ➜ search results:
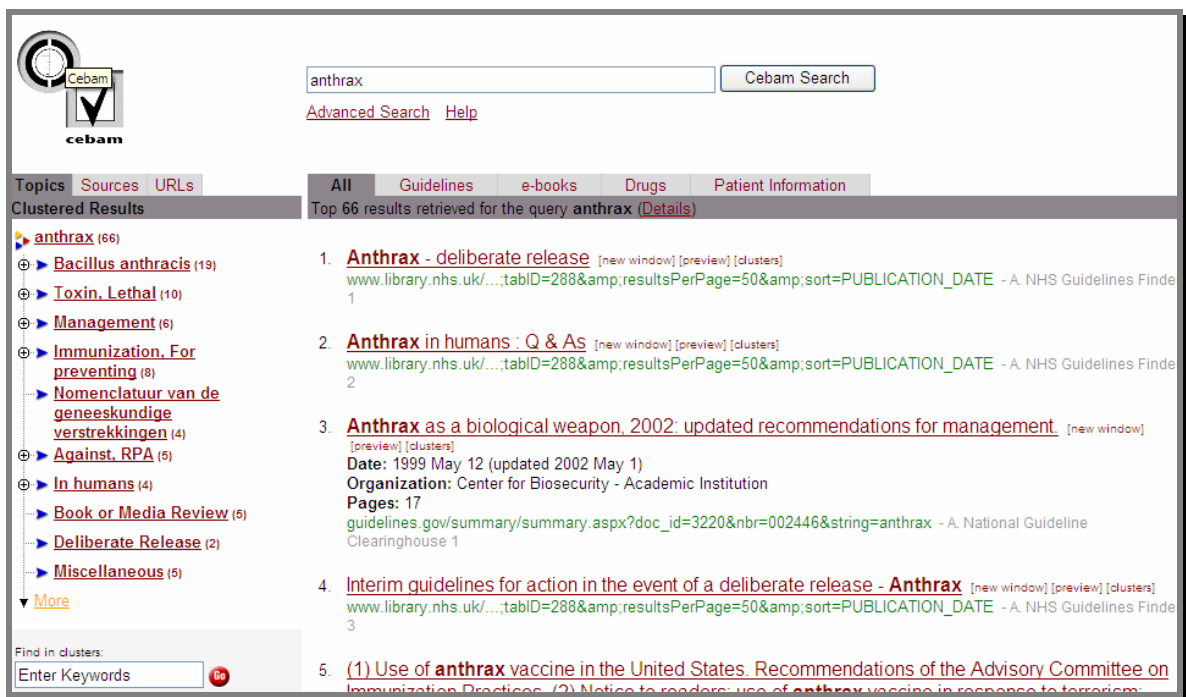

Figure 3 – Screenshot CEBAM federated search engine

The Dutch term *miltvuur* was found in the MeSH Termbase as a translation for the English MeSH term «anthrax». As both terms are linked to one another, the search engine will automatically retrieve documents for the query «anthrax» from all five databases in the system.

The popular terms in the *Multilingual Glossary of Technical and Popular Medical Terms* can be used as language support for lay people, and in combination with -the synonyms in- the MeSH Termbase they enable term expansion. A Dutch popular entry term will automatically lead to a Dutch technical term, and subsequently to the corresponding Dutch and English MeSH descriptors. For example, when a user enters "verwijding van de bloedvaten" (Dutch popular term, in the Multilingual Glossary) as a query, this term will be automatically linked to the term "teleangiëctasie" (Dutch technical term, in the Multilingual Glossary as well as in the Dutch MeSH Termbase) and from there to the English terms "telangiectasis" (English MeSH descriptor, in MTB), "telangiectasia" (English synonym, in MTB) and "broken veins" (English popular term, in Multilingual Glossary). Term expansion thus allows the system to perform a full text search on the terms "telangiectasia" and "broken veins" and an indexed search on "telangiectasis" for a Dutch query "verwijding van de bloedvaten".

6. Further research

The merger of the projects described above is still in a theoretical phase: the linguistic components -a partial translation of the MeSH thesaurus and a limited multilingual glossary with popular and technical medical terms- are available, but the practical application is still in a construction phase.

An application has been submitted for a new project (ABOP, Automatische Bijsluiter Optimalisator) in cooperation with the University College of Antwerp. The aim of this project is to develop a system which adapts patient information leaflets in a (semi-) automatic way to the level of non-professional users. The *Multilingual Glossary of Technical and Popular Medical Terms* will be expanded using lexicon-based and learning-based term extraction. A preliminary study (Hoste, Vanopstal and Lefever, 2007) has been performed on the automatic detection of scientific terms, combining these two approaches. The next step will be the automatic replacement of scientific terms by their popular counterparts.

As to the MTB, the translation will be elaborated and the termbase, which now has the structure of a dictionary, will be converted into a thesaurus which meets the ISO 25964 standard concerning structured vocabularies for information retrieval.

7. Conclusion

In this paper, we have presented two medical terminology projects with a similar objective: making medical information more accessible. We have demonstrated that medical terminology has specific morphological characteristics and that the differences in term formation between English and Dutch may cause errors in both inter-linguistic and inter-generic translation. The problem of inter-generic translation can be solved with the integration of the popular terms from the *Multilingual Glossary of Technical and Popular Medical Terms*.

With these projects, we hope to break down the language barrier and to provide lay people, linguists, translators and medical staff with language support for biomedical information retrieval.

8. References

Ahlfeldt, H., L. Borin, P. Daumke, N. Grabar, C. Hallett, D. Hardcastle, D. Kokkinakis, C. Mancini, K. Markó, M. Merkel, C. Pietsch, R. Power, D. Scott, A. Silvervarg, M. T. Gronostaj, S. Williams and A. Willis (2006). Literature Review on Patient-Friendly documentation systems. Technical Report 2006/04., Computing Department, The Open University

Askehave, I. and K. K. Zethsen (2000). Inter-generic and Inter-linguistic translation of patient package inserts. 12th European Symposium on Language for Special Purposes, Tübingen: Gunter Narr. 882.

Banay, G. L. (1948). "An Introduction to Medical Terminology I. Greek and Latin Derivations." Bull Med Libr Assoc **36**(1): 1.

Buysschaert, J. (2006). The development of a MeSH-based biomedical termbase at Hogeschool Gent. LREC 2006 Satellite Workshop W08. Acquiring and representing multilingual, specialized lexicons: the case of biomedicine, Genova. 39.

Corson, D. (1985). The Lexical Bar. New York: Pergamon Press.

Hoste, V., K. Vanopstal and E. Lefever (2007). The Automatic Detection of Scientific Terms in Patient Information. RANLP, Borovets, Bulgaria. 41.

Jolly, P. (1987). The Language of Medicine: The Ethics of Obscurity. Conference on College Composition and Communication. Atlanta.

Kronick, R. (2007). "Medically speaking " CMAJ **176**(1): 73.

Martin, W. (2006). Van Dale groot woordenboek Engels-Nederlands, Nederlands-Engels [CD-ROM]. Utrecht ; Antwerpen Van Dale Lexicografie.

Maylath, B. (1997). "Why Do They Get It When I Say "Gingivitis" But Not When I Say "Gum Swelling"?" New directions for teaching and learning **1997**(70): 29.

Mintz, D. (1992). "What's in a word: The distancing function of language in medicine " The Journal of Medical Humanities **13**(4): 223.

Plovnick, R. M. and Q. Zeng (2004). "Reformulation of consumer health queries with professional terminology: a pilot study." J Med Internet Res **6**(3): e27.

Zethsen, K. K. (2004). "Latin-based terms. True or false friends?" Target **16**(1): 125.