For Guust

Kaftinformatie: © Bart Deygers

ISBN: 978-94-9179-107-9

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of enige andere manier, zonder voorafgaande toestemming van de uitgever.

UNIVERSITEIT GENT

Impact of language skills and system experience on medical information retrieval

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Taalkunde aan de Universiteit Gent te verdedigen door

Klaar Vanopstal

Promotoren	Prof. dr. Godelieve Laureys	
	Prof. dr. Joost Buysschaert	
	Prof. dr. Robert Vander Stichele	
Decaan	Prof. dr. Marc Boone	
Rector	Prof. dr. Paul Van Cauwenberge	Gent, 2013

Acknowledgements

This PhD has been a truly challenging experience, and it would not have been possible without the support and guidance of many people. Finishing a PhD gives you the unique opportunity to thank the people you are surrounded by, and I would therefore like to express my sincere appreciation to those who have supported me in one way or another during this six-year endeavor.

At this moment of accomplishment, first of all, I would like to address a word of thanks to my supervisors, Prof. Dr. Robert – Bob - Vander Stichele, Prof. Dr. Joost Buysschaert, and Prof. Dr. Godelieve Laureys. Thank you, Bob, for meticulously reading my drafts and putting my ideas to the test in the many animated discussions we had. They made me grow as a researcher and as a person. I warmly thank Joost Buysschaert, who has been my mentor for the past six years, for his relentless support, encouragement and understanding, and for believing in me. I am also grateful to Godelieve Laureys for her guidance, time and advice.

I feel very fortunate to have been surrounded by such fantastic colleagues in the LT3 team. I greatly enjoyed their company and the exquisite birthday treats and would like to thank them all for creating such an inspiring working atmosphere. I extend my sincere gratitude to Véronique for introducing me into the world of language technology, for being such an excellent coach, and for giving me the opportunity to both learn and teach.

I warmly thank my colleagues in the office across the hallway, Kathelijne, Marjan, Joke, Lieve, and especially Els, for helping me so patiently through my struggles with Perl. Heartfelt thanks go to Orphée, whose down-to-earthiness in combination with her sense of humor and borderless enthusiasm at times helped me put things back into perspective. Although only briefly, I enjoyed working with Bart on acronym detection and resolution. His eagerness to learn is infectious and inspiring. I would also like to thank the colleagues with whom I share(d) not only an office but also laughs and friendship: Peter, Sarah, Gwendolijn, Dries, Philip, Bram, Geert, Sofie and Stijn. I would

like to single out Isabelle – Bebbel, the best housemate ever - a good friend and a much appreciated colleague, gifted with a sixth sense when it comes to detecting one's – especially troubled – moods.

I am very much endebted to Koen for his time and statistical support, to Gitte for helping me with the layout and cover of this book and to Bart for providing the cover photograph.

I also want to thank some colleagues-slash-friends outside the LT3 team, and of course many other friends for their unwavering support and encouragement, and for the laughter that we shared. At the inevitable risk of forgetting someone, I would especially like to refer to Bernard, Sofie, Michael, Sabine, Jasper, Liesbeth and Kevin, Natje, Claire, Sofie and Bert, Bart and Sofie, Dietger, Koen and Evelien, Wim and Eva.

A word of thanks is also in place for my brother-in-law, "jonkel Tim", for being such a good neighbor, and for taking care of Guust whenever I had to work late and couldn't make it home in time. A big thank you to my other in-laws as well for their support.

Finally, I would like to acknowledge the people who mean the world to me, my family. A special thanks to my parents for being my very first and most important teachers, and for always encouraging me to pursue my intellectual and other interests. Sebbe and Lien, dreamer versus pragmatic, you always succeeded in letting me see a different side of things. To my son Guust: your birth challenged and inspired me at the same time. You taught me real happiness and gave me something extra to look forward to every day. It is hard to overstate my gratitude to my husband, Bart. Thank you for your relentless support, encouragement, patience and understanding. I admire your uncanny ability to stay calm, even when I'm not. Thank you also for taking care of us with so much love. Thank you.

Table of contents

0

Introduction

1. Preamble

This manuscript is based on a collection of four articles published in or submitted to international, peer-reviewed journals. The central theme in these papers is medical terminology in information retrieval. Each of the publications will be presented as a separate chapter in this dissertation:

Chapter I 2011 Vocabularies and retrieval tools in biomedicine: disentangling the terminological knot. *Journal of Medical Systems 35 (4): 527-543*

Chapter II 2011 PubMed searches by Dutch-speaking nursing students : the impact of language and system experience. *Journal of the American Society for Information Science and Technology, 63 (8):1538–1552*

Chapter III 2013 Lost in PubMed. Factors influencing the success of medical information retrieval. *Expert Systems with Application, 40 (10): 4106-4114*

Chapter IV 2013 Query formulation and relevance judgment in native and non-native English-speaking PubMed users. *Journal of the American Medical Informatics Association* (submitted)

As each of these chapters was published in or submitted to separate international journals, there is inevitable overlap in those parts that explain the set-up of the experiment. This is especially the case in the introductory sections of Chapters II, III and IV. In order not to add to this overlap, this general introduction will be kept concise, and will be limited to an overview of the research questions and short descriptions of methodology for each part in this thesis.

The first part (Chapter I) presents a theoretical study of vocabularies for medical information retrieval, and the way they are defined in the literature. The starting point of this study was MeSH (Medical Subject Headings), a vocabulary used to index and retrieve information. This vocabulary will be used in the retrieval experiment in part 2.

The second part (chapters II, III and IV) elaborates on medical information retrieval and the difficulties nursing students experience when they search for medical information in PubMed/MEDLINE.

2. Research questions and methods

2.1. Part 1: the terminology of medical information retrieval

In view of the other studies conducted within the framework of this dissertation, the Medical Subject Headings (MeSH) - and thesauri or controlled vocabularies in general - were of particular interest. The National Library of Medicine (NLM), who created and maintain the MeSH, describe the vocabulary as follows: "*MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed*"[1]. Apparently, the MeSH is both a controlled vocabulary, and a thesaurus.

The literature gives a number of diverging definitions for the types of vocabulary that can be used in information retrieval, viz. thesauri, controlled vocabularies, but also ontologies, taxonomies, glossaries and topic maps. The main aim of the first study in this dissertation was to provide an overview of the usage of these terms, and to find a consensus definition. Secondly, we wanted to examine some of the existing vocabularies in the domain of medicine for their compatibility with these definitions.

Research questions to be answered in this part were:

1. Which definitions of *glossary, taxonomy, controlled vocabulary, thesaurus, ontology* and *topic maps* can be found in the literature? Are they consistent?
2. What causes inconsistencies in the use of these terms?
3. Is it possible to formulate a domain-independent definition for thesauri and controlled vocabularies? How do the Medical Subject Headings relate to this definition?

In order to answer research questions 1 and 2, we built a corpus of definitions based on a comprehensive literature study. We compared the definitions in this corpus and tried

[1] http://www.ncbi.nlm.nih.gov/mesh

to make a classification on the basis of the domains they were used in. This classification led to clearer definitions across several dimensions - linguistics, knowledge management and bibliographic retrieval. In the second part of this study, we tested some of the major existing medical vocabularies for their compatibility with these definitions.

2.2. Part 2: the role of terminology in medical literature searching

The Internet explosion puts information that was inaccessible to the previous generation of researchers at the fingertips of current researchers. Moreover, the massive availability of medical information is further boosted by the growing number of biomedical journals (Dogan et al., 2009). However, when more threatens to become less, well-designed search tools and the skills to use them efficiently are crucial for people working in the medical field in order to keep abreast of the biomedical literature.

Next to searching skills and tools, a fair level of English language skills is required, as English is the lingua franca of medicine, and of science in general. English "is understood, or due to numerous reasons, is desired to be understood by almost every individual and every nation on the globe who want to enjoy access to the latest developments, whatever field of study it may be" (Abdullah & Chaudhary, 2012). This adds an extra level of complexity to information retrieval for non-native speakers of English. The Dutch-speaking participants in our test were all speakers of English as a Foreign Language (EFL).

The research questions to be answered in this part were:

1. Do English language skills in Dutch-speaking users of PubMed affect the efficiency of their literature searches? (Chapter II)

2. How can we distinguish between best and worst performers? Can their characteristics be linked to the errors they made? (Chapter III)

3. To what extent do language skills and searching skills in native and non-native speakers of English contribute to the outcome of literature searches in PubMed? (Chapter IV)

In order to answer these research questions, we conducted a retrieval experiment with four types of respondents:

- Dutch-speaking bachelor's nursing students (Nursing Department at University College Ghent)
- Dutch-speaking master's nursing students (Nursing and Midwifery Department at the University of Antwerp)
- native English-speaking bachelor's nursing students (School of Nursing at the University of Nottingham)
- native English-speaking master's nursing students (School of Nursing at the University of Nottingham)

The test participants were given a pre-formulated question that represented the information need in this experiment. They had to find as many citations as possible in PubMed that answered all aspects of this information need. Screen recordings and keystroke logging allowed us to study the search process in detail. The outcome of the searches was studied in terms of – different types of – recall, and precision.

References

Abdullah, Sayeh S., & Chaudhary, Mohammad Latif (2012, 26-27 December 2012). *English as a Global Lingua Franca.* Paper presented at the International Conference on Education, Applied Sciences and Management (ICEASM'2012) Dubai.

Dogan, R. I., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database (Oxford), 2009*, bap018. doi: 10.1093/database/bap018

1

The terminology of information retrieval

> There is no greater impediment to the advancement of knowledge than the ambiguity of words.
>
> Thomas Reid, 18th century philosopher

Chapter I: Vocabularies and retrieval tools in biomedicine: disentangling the terminological knot

Abstract

Terms like “thesaurus”, “taxonomy”, “classification”, “glossary”, “ontology” and “controlled vocabulary” can be used in diverse contexts, causing confusion and vagueness about their denotation. Is a thesaurus a tool to enrich a writer’s style or an indexing tool used in bibliographic retrieval? Or can it be both? A literature study was to clear the confusion, but rather than giving us consensus definitions, it provided us with conflicting descriptions. We classified these definitions into three domains: linguistics, knowledge management and bibliographic retrieval. The scope of the terms is therefore highly dependent on the context. We propose one definition per term, per context.

In addition to this intra-conceptual confusion, there is also inter-conceptual vagueness. This leads to the introduction of misnomers, like “ontology” in the *Gene Ontology*. We examined some important (bio)medical systems for their compatibility with the definitions proposed in the first part of this paper. To conclude, an overview of these systems and their classification into the three domains is given.

Keywords: information retrieval; medical terminology; medical coding systems; taxonomy; thesaurus; ontology; controlled vocabulary; classification

1. Introduction

Terms such as thesaurus, taxonomy, ontology and controlled vocabulary, and even glossary, dictionary and lexicon at first sight seem to be unambiguous terms. However, they are used in different ways in different contexts, causing continual confusion. Moreover, the distinction between the terms themselves is not always straightforward.

A look at the information about the term *'death'* in three different thesauri (see Table 1), tells us that not all thesauri give the same kind of information:

Table 1: The word "death" in several thesauri

Unesco thesaurus	Roget's II	ICPC2-ICD10 thesaurus
Death [93]	**Death**	**Death**
Terme français: Mort Término español: Muerte Русский термин : Смерть **MT** 2.70 Biology **UF** Causes of death **BT** Life cycle [77] **RT** Ageing [88] **RT** Birth rate [91] **RT** Euthanasia [24] **RT** Homicide [24] **RT** Mortality [242] **RT** Suicide [29]	***See also*** 2 (non-existence); 62 (end); 32 (killing); 325 (burial). **n.** *death*, mortality, fatality, casualty, losses, death toll; *extinction*, decease, departure, exit, demise, release; *natural death*, accidental death, cot death, stillbirth, miscarriage, brain death, abortion; *unnatural death*, [...] **adj.** *dying*, moribund, half-dead, not long for this world, done for, slipping away, in extremis; *dead*, [...] **vb.** *Die*, perish, expire, pass over/away, fall asleep, give up the ghost, depart this life, croak *(colloq.)*, peg out *(colloq.)*, pop one's clogs *(colloq.)*, [...]	**ICD10** : R99 Other ill-defined and unspecified causes of mortality **ICPC**: A96 Death

The *Unesco Thesaurus* (University of London Computer Centre (ULCC), 2003) includes information such as narrower terms (NT), broader terms (BT), related terms (RT), other language equivalents (SP, FR) and related terms (RT). In *Roget's Thesaurus* (Roget, 1995), by contrast, other information is given: function, derivations, and related terms. The ICPC2-ICD10 thesaurus, a system used for medical classification which links concepts of ICPC2 to ICD-10 concepts, gives the classification codes R99 (ICD-10) and A96 (ICPC) for

'death'. It is clear that these thesauri differ considerably in their structure and scope. Does this mean that for one of them, the denomination "thesaurus" is not - or less- apt?

The main problem is that the terms taxonomy, classification, thesaurus, ontology and controlled vocabulary are used in many different contexts, including linguistics, bibliographic information retrieval (IR) and knowledge management, including medical coding. As Kagolovsky and Moehr (2003) point out, "information retrieval" has no common definition, due to the different research backgrounds of the authors who use the term. Kagolovsky and Moehr propose the following definition, citing Harter and Hert (1997): a system that "retrieves documents, or references to them, rather than data". This definition corresponds to what we will call in this paper bibliographic retrieval. Medical registration systems, on the other hand, are established in the first place to represent and store information –rather than documents- and in the second place to later retrieve and re-use that information.

The first section of this paper gives an overview of the different fields in which the terms "glossary", "lexicon", "dictionary", "taxonomy", "classification", "thesaurus", "ontology" and "controlled vocabulary" can be used. On the basis of these observations, definitions will be suggested and recommendations made for a more consistent and unambiguous use of the relevant terminology. In the second section, these insights will be applied to the biomedical domain, where these issues are particularly relevant. To conclude, an overview (part 3) of the existing tools in the three dimensions (linguistics, knowledge management -including medical coding- and bibliographic retrieval) is presented.

2. Domains of application of the terms

As mentioned above, terms such as taxonomy, thesaurus, ontology, controlled vocabulary etc. can be defined in various ways depending on the domain of application. We will discuss three domains, namely linguistics, knowledge management -including medical coding systems- and bibliographic retrieval.

There are several linguistic tools which can help to find the right terms, or to find an explanation or definition for a certain term, viz. dictionaries, lexicons, glossaries,

thesauri and controlled vocabularies. These systems (can) have a purely linguistic function. However, thesauri and controlled vocabularies can also be used for the retrieval of documents or data.

A second domain which will be discussed here, is that of the storage and retrieval of knowledge. We especially focus on medical coding systems, such as ICPC and ICD. Medical coding systems can be described as classifications or nomenclatures of health- and medicine-related phenomena. These concepts are structured and usually given a code which indicates the place of the concept in the nomenclature, as can be seen in figure 1.

(K35-K38) Diseases of appendix

- (K35.) Acute appendicitis
  - (K35.0) Acute appendicitis with generalized peritonitis
  - (K35.1) Acute appendicitis with peritoneal abscess
  - (K35.9) Acute appendicitis, unspecified
- (K36.) Other appendicitis
- (K37.) Unspecified appendicitis
- (K38.) Other diseases of appendix
  - (K38.0) Hyperplasia of appendix
  - (K38.1) Appendicular concretions
    - Faecalith
    - Stercolith
  - (K38.2) Diverticulum of appendix
  - (K38.3) Fistula of appendix
  - (K38.8) Other specified diseases of appendix
    - Intussusception of appendix
  - (K38.9) Disease of appendix, unspecified

Figure 1: Extract of the ICD10 classification: "diseases of appendix"

Bibliographic retrieval can be defined as the science of searching a database for journal or magazine articles, containing citations, abstracts and often full texts or links to the full texts. The underlying structures to search for articles in databases include taxonomies, thesauri, ontologies, controlled vocabularies and topic maps.

2.1. Linguistics

2.1.1. Glossaries, dictionaries and lexicons

The term 'glossary' originates from the Latin word *glossarium*, a collection of glosses. 'Gloss', in its turn, originates from the Greek word *glossa* (γλῶσσα) which denotes the explanation of a specialized expression or difficult word. Hence, 'glossary' can be defined as a list of terms in a particular field of knowledge, with definitions or explanations.

Glossaries are usually arranged alphabetically. The terms in monolingual glossaries usually refer to LSP (Language for Specific Purposes) and are furnished with definitions. These definitions generally apply to one domain only, and thus rarely include variant meanings. In practice, however, these definitions are often omitted in multilingual glossaries.

Glossaries can be integrated into a book or a website, but they can also be stand-alone lists. They can be used as, but are not, per se, controlled vocabularies (see 2.1.3.). They can be monolingual (e.g. Wikipedia's *Glossary of medical terms related to communications disorders*[2] or the Dutch *RIZIV* glossary[3]), bilingual (e.g. the *TERMISTI glossaries* of *abortion*[4] and *autism*[5] *terms*) or multilingual (e.g. *Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages* [6]).

The term glossary is used interchangeably with lexicon and dictionary. This presumed equivalence, however, leads to a blurring of the conceptual boundaries of the terms. Ananiadou (2006) defines 'lexicon' as a list containing "the lexical elements (either as full forms or as canonical base forms), together with additional linguistic information about them, which is required for further morphological, syntactic, and semantic processing." She adds that lexicons are not fully standardized, which allows their

[2] http://en.wikipedia.org/wiki/Glossary_of_medical_terms_related_to_communications_disorders
[3] http://www.riziv.fgov.be/nl/glossary.htm
[4] http://www.termisti.refer.org/data/ivg/index.htm
[5] http://www.termisti.refer.org/data/autisme/frame.html
[6] http://users.ugent.be/~rvdstich/eugloss/welcome.html

makers to model them so that they best suit their own purposes. We will adopt Ananiadou's definition.

Dictionaries, both monolingual and multilingual, can refer to general language or to a specialized terminology. They often give limited morphological and grammatical information (e.g. gender, part of speech, plural form) and sometimes also a phonetic transcription, next to a definition. Bi- and multilingual general language dictionaries provide a translation -or several translations used in different contexts-, collocations and idiomatic expressions. Conversely, specialized multilingual dictionaries usually offer translations with very little further information. An example from the *Wörterbuch für Industrie und Technik* (French-English/ English-French) (CILF, 1993):

Reprofilage n.m. **Neuprofilierung** n.f. Bâtiments et travaux publics[7]

In summary, the boundaries between the terms glossary, lexicon and dictionary have blurred to some extent. However, we define 'glossary' as "a list of words or terms with their explanations", 'lexicon' as "a list of words or terms, together with linguistic information about them" and 'dictionary' as "a list of words or terms with limited linguistic information, usually a definition, and, in the case of bi- or multilingual dictionaries, one or more translations".

2.1.2. Thesauri

The word 'thesaurus' is derived from the ancient Greek 'thesauros' (θησαυρός), or 'treasure'. In the 16th century, its meaning was narrowed to 'treasure of words', like a dictionary or an encyclopedia. The word 'thesaurus' fell into disuse for some time, but revived with the release of Roget's *Thesaurus of English Words and Phrases* in the 19th century. Roget adopted an onomasiological approach -providing the word for a given idea- in his thesaurus, whereas most dictionaries were, and still are, characterized by a semasiological approach, i.e. they describe the referential meaning denoted by words.

[7] The first colum refers to the French term, the second to the German translation and the third column refers to the corresponding domain.

Roget did not organize his thesaurus alphabetically, but systematically, i.e. according to ideas or concepts.

> *The purpose of an ordinary dictionary is simply to explain the meaning of words; and the problem of which it professes to furnish the solution may be stated thus:—The word being given, to find its signification, or the idea it is intended to convey. The object aimed at in the present undertaking* [Roget's Thesaurus] *is exactly the converse of this: namely,—The idea being given, to find the word, or words, by which that idea may be most fitly and aptly expressed.* (Mawson, 1922)

A thesaurus can thus be a purely linguistic tool, which provides a standard language of a particular field of knowledge and contains information about nuances of concepts. This type of thesaurus is referred to by Kilgarriff and Yallop (2000) as the 'Roget-style thesaurus'. Its objective is to improve the effectiveness of communication: the relationships outlined in the thesaurus help to fine-tune style or to obviate misunderstandings.

Later, in the mid-twentieth century, the term experienced another shift in meaning, adopting the information retrieval aspect (see infra).

2.1.3. Controlled vocabulary

A controlled vocabulary is a set of terms which provides a standard language for a specific domain. It consists of two types of terms: preferred terms, which are designed to control a domain-specific language, and non-preferred terms used as "access vocabulary", "lead-in" or "entry" terms. The use of preferred and non-preferred terms is illustrated by Wodtke (2002):

> *In our restaurant we had the preferred term, "first course", and all the terms our patron might use, "starter, first course, hors d'oeuvres, appetizer", neatly tucked into our head. So if a patron wanted an appetizer of smoked salmon, we would write in the check "first course: smoked salmon".*

A controlled vocabulary can be used as a prescriptive terminology, as a means to ensure language hygiene and/or consistency in the use of terminology. The *Plain English Campaign*[8] is an independent British organization which helps businesses, local governments and government departments to improve their communication by

[8] http://www.plainenglish.co.uk/

providing editing services, training courses and glossaries. They also published a controlled vocabulary, *The A to Z of alternative words*, which is a list of words with their simpler alternatives designed for writers of all text types to ensure readability.

2.2. Knowledge management and medical coding

2.2.1. Taxonomies and classifications

A literature search for the term *taxonomy* proves that Garshol (2004) is right in saying that the term has been "used and abused to the point that when something is referred to as a taxonomy it can be just about anything" and that the basic denominator is that of an "abstract [hierarchical] structure".

Taxonomy is derived from the Greek words *taxis (τάξις)*, 'order' and *nomos (νόμος)*, 'rules, law' and is often described as "the science of classification of organisms" (Davis & Heywood, 1963). However, the term taxonomy can also be defined in terms of its structural characteristics: "a taxonomy provides a classification structure that adds the power of inheritance of meaning from generalized taxa to specialized taxa" (ISO/IEC_11179-2, 2005). This inheritance implies that subclasses take over characteristics of their ancestor classes. Agro (2004) and Beck (2002) also use the term in the sense of a hierarchical structure which represents (a part of) reality. Dictionaries such as Oxford English Dictionary and Merriam-Webster and other reference works such as WordNet and Roget's Thesaurus differentiate between the two meanings, i.e. taxonomy as a science and taxonomy as a hierarchical representation of reality. Sterkenburg (2003) combines both meanings in his definition: "study of the theory, practice and rules of classification of terms, objects and concepts".

The term taxonomy originated in biology, where it referred to the classification of the names of organisms. It was the Swedish scholar Carolus Linnaeus who combined the loose principles of the existing taxonomies into the 'Linnaean taxonomy' (*Systema Naturae* 1735). In this hierarchical classification, nature was divided into kingdoms, phyla (for animals) and divisions (for plants), classes, orders, families, genera and species. In the figure below (figure 2), modern humans (homo sapiens) are defined according to the Linnaean taxonomy.

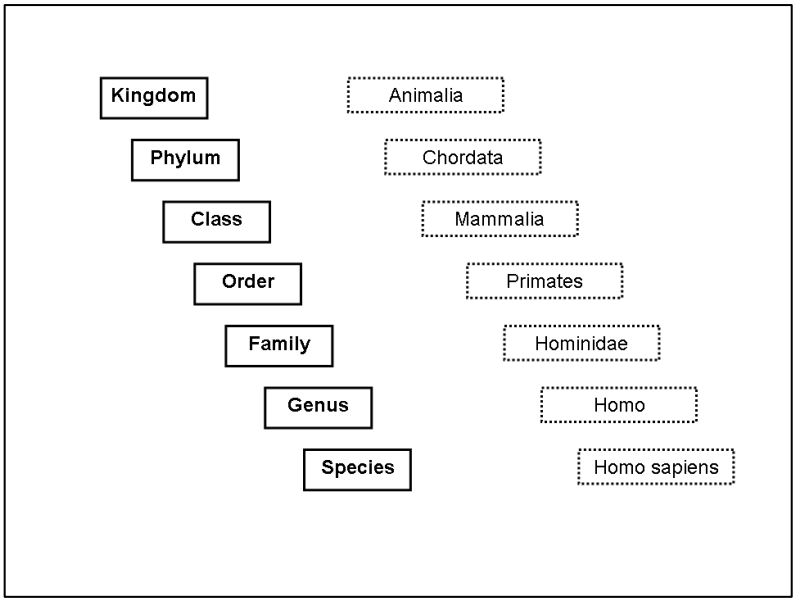


Figure 2: Modern humans in the Linnaean taxonomy

Linnaeus' taxonomy, which is now called the alpha taxonomy or classical taxonomy is still a model for biological classifications.

The designations "taxonomy" and "classification" are used interchangeably, whereas they are not completely synonymous. Agro (2004) and Van Rees (2003) argue that taxonomies distinguish themselves from classifications in that they group concepts according to essential, internal attributes, i.e. according to relationships between the concepts. Taxonomies, unlike classifications, are created from the bottom up, are based on actual content and guide users through a body of information. A classification, on the other hand, is a grouping of concepts according to arbitrary, external attributes (Van Rees, 2003). These external attributes can be color, shape, geography, size, usability, etc. Classifications are created from the top down and are based on conceptual frameworks (Agro, 2004; Van Rees, 2003). Table 2 summarizes the characteristics of taxonomies versus classifications according to Agro and Van Rees.

Table 2: Taxonomy versus classification according to Agro and Van Rees

Taxonomy	Classification
grouping of concepts according to essential, internal attributes	grouping of concepts according to arbitrary, external attributes
created from the bottom up	created from the top down
based on actual content	based on conceptual frameworks
created by a multidisciplinary team	created by domain experts
flexible, dynamic	static

Cann (1997), however, uses other criteria to define the concepts of classification and taxonomy. He describes special versus general, analytical versus documentary and enumerative versus faceted classifications. Firstly, a classification describes either general knowledge, e.g. the Universal Decimal Classification (UDC) or a specific knowledge domain, e.g. the International Classification of Diseases (ICD). Secondly, a classification can be analytical or documentary. Analytical implies that physical phenomena are systematized into an understandable scheme. Cann (1997) also designates this type of classification as “taxonomies”. In his opinion, “taxonomy” and “classification” are not, as argued by Agro and Van Rees, co-hyponyms, rather “taxonomy” is hyponymous to “classification”, or a taxonomy is a 'kind of' classification. Documentary classifications are used as information management and retrieval tools (e.g. UDC). Thirdly, classifications can be either enumerative or faceted. An enumerative classification lists certain classes and all their subclasses of interest (Cann, 1997), is created from the top down and allows for compound subjects. This type of classification is often called hierarchical, which is a common misunderstanding, as faceted classifications can also have a hierarchical structure. Faceted classifications are created from the bottom up and do not provide “ready-made class numbers for compound and complex subjects” (Indira Gandhi National Open University, 2006). In enumerative classifications, there is usually only one path the user can follow to find his subject, i.e. from a broad category to the specific concept. In faceted classifications, the concepts are organized into classes according to several principles of division. An

example of a faceted classification can be found in Springerlink's[9] organization of documents, where documents can be retrieved using different principles of division. The collection can be searched by the facets "content type", "featured library" or "subject collection".

Cann's view (see figure 3) seems to be more solid and logical. Here, a classification is considered as a hypernym for all types of concept categorization. However, Cann still overlooks the fact that analytical classifications, or taxonomies, have also come to play a role in information retrieval, i.e. they have adopted the function of documentary classifications.

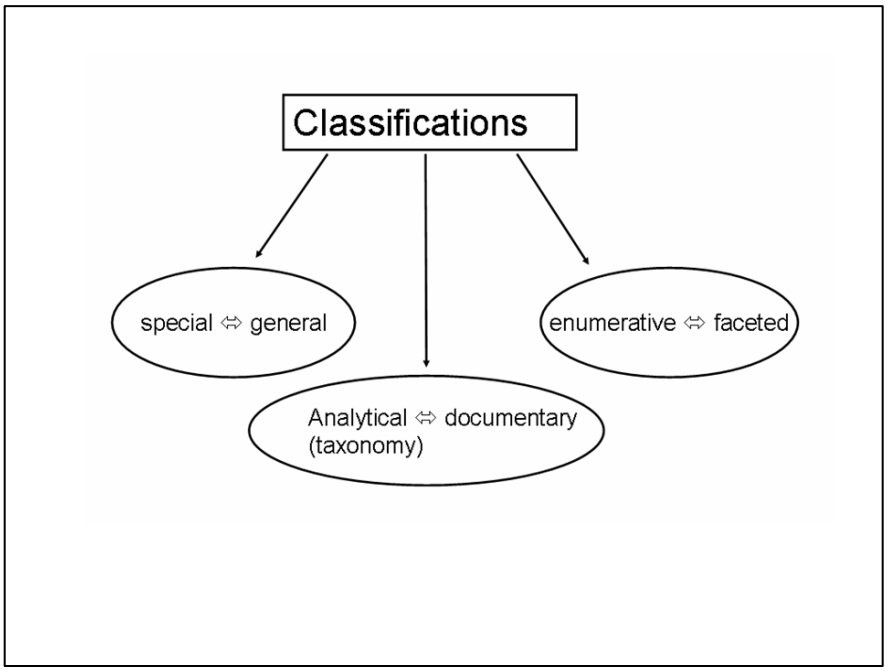


Figure 3: types of classification according to Cann (1997)

We propose a definition for "taxonomy" in data retrieval, based on ISO/IEC 11179-2 (2005): "a taxonomy provides a hierarchical classification structure that adds the power of inheritance of meaning from generalized taxa to specialized taxa". Classification is a more general term which can be defined as "the grouping of concepts on the basis of shared characteristics". Both structures can be used in medical coding systems.

[9] http://www.springerlink.com

2.2.2. Ontologies

A closer look at the concept of 'ontology' shows that its meaning depends on the domain or the (historical) context in which it is used as well: either philosophy or information science. When used in the context of philosophy, Ontology is often written with an upper-case 'O', whereas ontology with a lower-case 'o' – and with a plural form, ontologies - refers to a representation of reality or to an information retrieval system.

The term 'Ontology' is derived from the Greek words *ὄν* (being) and λογία (science, study, theory) and literally translates into "the science of being". This branch of metaphysics organizes, or attempts to organize the universe and its components into a scheme with explicit formulation of their possible relations. Most dictionaries, such as LONGMAN Dictionary of Contemporary English (Procter, 1978), Oxford English Dictionary (Simpson & Weiner, 1989) and Merriam-Webster (Merriam-Webster Inc., 2008) define Ontology in this context. As a derived meaning used within the context of knowledge management, an ontology can be described as a representation of what exists. Some ontologies, like SNOMED or OpenGalen, are more than just a representation of the concepts within a specific domain with their relationships; they are designed as a coding system or for clinical decision support.

2.3. Bibliographic retrieval

2.3.1. Taxonomies

With the advent of the Internet, taxonomies started covering other purposes than those described in 2.2.1.: they now also function as metadata for information retrieval. The concepts in these taxonomies are used as keywords for tagging documents, or for referencing to these documents. Cann (1997) refers to this type of taxonomy as "documentary classifications" (see 2.2.1.).Their structure offers more transparent and more efficient search options, including explosion of the search term. Term explosion allows the system to search for information about not only the concept itself, but also about its narrower, hyponymic concepts.

Taxonomies can be included in thesauri and ontologies (Beck & Pinto, 2002; Ullrich et al., 2003), and taxonomies and thesauri are often bracketed together as one and the

same concept. So what distinguishes taxonomies from thesauri, and from ontologies? Basically, 'taxonomy' can refer to any hierarchical classification of elements of a group into subgroups according to specific criteria, often visualized as a tree. Its relationships are not specified, i.e. broader and narrower terms can designate the obvious subsumption relationship (parent/child), but also a mereologic relationship (part/whole). Taxonomies do not cover any relationships other than hierarchical. Thesauri and ontologies compensate for this lacuna and give explicit or implicit indications as to the nature of the relationships.

2.3.2. Thesauri

Peter Luhn (IBM) conceived the idea of using a thesaurus, which was previously a purely linguistic tool, for information retrieval. In the 1960s, the first thesauri for information retrieval were published. The *Thesaurus of Engineering and Scientific Terms* (Engineers Joint Council, 1967) sketched the broad outlines of the standard format for thesauri. In this period, thesauri evolved towards their current form, defined by ISO 2788 (International Organization for Standardization, 1986) as "the vocabulary of a controlled indexing language, formally organized so that the *a priori* relationships between concepts (for example as "broader" and "narrower") are made explicit." *Controlled* means that the vocabulary is predetermined and is used as a prescriptive terminology. This implies that the terminology of the subject field is subdivided into preferred terms - also called descriptors- and non-preferred terms or entry terms. A thesaurus is usually organized *hierarchically*, which means that the relationships 'broader term' and 'narrower term' are visible in a tree-like structure or made explicit by the abbreviations BT and NT respectively. ISO 2788 states that there are various ways in which the terms in a thesaurus can be displayed, the most common of which are alphabetical, systematic and graphic display. The standardized relationships in thesauri are the hierarchical, associative and the equivalence relationship. These are *a priori* relationships, which means that they are context-independent, rather than being inferred from the documents they describe.

When used in the context of information and library science, 'thesaurus' refers to a retrieval instrument, used to index and/or search documents. This is often the main or

only purpose of present-day thesauri, and most authors (Aitchison et al., 2000; ANSI/NISO, 2005; Beck & Pinto, 2002; BSI, 2005; Chowdhury, 2003; Hagedorn, 2000; International Organization for Standardization, 1986; Ribeiro-Neto & Baeza-Yates, 1999) define thesaurus in this context. Chowdhury (2003) describes the following main objectives of thesauri for information retrieval:

1. *vocabulary control: a translation of natural language into a more constrained language*
2. *consistency between different indexers*
3. *limitation of the number of terms needed to label the documents*
4. *search aid in information retrieval*

The historical and interdomain shifts – from the linguistic field to the field of information science - described above are reflected in the definitions given by Landau (1984):

1. *A "storehouse" of knowledge such as exhaustive encyclopaedia or dictionaries,*
2. *Exhaustive lists of words from the general language, without definitions, arranged systematically according to the ideas they express.*
3. *A list of subject headings for a particular field of knowledge, arranged in alphabetic or classified order and used for information retrieval and related purposes.*

Due to these shifts, the term 'thesaurus" carries several meanings, and it is thus recommendable to study the context and subject field in which the term occurs before drawing any conclusions as to its meaning.

There are several standards for thesauri. ISO 2788 was created for the design of monolingual thesauri and ISO 5964 (International Organization for Standardization, 1985) documents the design of multilingual thesauri. These standards, however, are outdated (International Organization for Standardization, 2007), as they only refer to printed thesauri. Both standards will be replaced by a new standard, ISO 25964, based on BS 8723[10] (BSI, 2005), the corresponding British standard. ANSI/NISO, the US standardization organization, created its own standard, Z39.19. These guidelines have a somewhat broader scope: they comprise all monolingual controlled vocabularies,

[10] The BS 8723 standard consists of five parts, the first two of which broadly correspond to ISO 2788, whereas the combination of part one and four have approximately the same scope as ISO 5964 (multilingual thesauri). BS 8723-3 covers vocabularies other than thesauri, BS 8723-4 gives recommendations concerning interoperability of vocabularies and BS 8723-5 discusses exchange formats.

including lists, taxonomies, thesauri and synonym rings. There is no single 'worldwide' standard, as the US and other standards (BS, ISO) departed from each other in previous editions. In an interview (Roe & Thomas, 2004), Dr. Amy J. Warner[11] stated that the new ANSI/NISO standard should be more compatible with the existing standards.

In conclusion, the term thesaurus can be used in different contexts, related to different fields of knowledge which came into existence at different points in time. When used in the context of information science, a thesaurus can be defined as *a "controlled vocabulary, which is usually organized hierarchically and which includes standardized, a priori, hierarchical, associative and equivalence relationships between concepts"* (International Organization for Standardization, 1986).

2.3.3. Controlled vocabularies

According to the ANSI/NISO Guidelines (2005), a controlled vocabulary, which is a list of preferred and non-preferred terms, is – or should be - exempt of ambiguities, homonymy and polysemy and all terms should have "an unambiguous, non-redundant definition". Controlled vocabularies can be used for consistent indexing and searching of information. For instance, using a controlled vocabulary in medical information retrieval can help health professionals to describe and classify medical information, optimizing the work of both searchers and indexers.

Compared to natural language, a controlled vocabulary has some weaknesses and some strengths, as stated by Aitchison et al. (2000). Its weaknesses include the relative lack of exhaustivity and specificity, the laboriousness of keeping it accurate and up-to-date and the cost of doing so. Moreover, this language has to be learned by the searcher and efficient exchange is often hampered by the incompatibility of the existing controlled vocabularies. Aitchison et al., however, add that over-exhaustivity may provoke a loss of precision. In addition, a controlled vocabulary can facilitate the search process considerably by expanding the query to its synonyms and excluding ambiguity. A

[11] Project Leader for NISO's Thesaurus Development Team

controlled vocabulary is usually incorporated into a thesaurus, an ontology, a topic map, which, in turn, can be used in an information retrieval system.

2.3.4. Ontologies

In the late twentieth century, the term "ontology" adopted some new properties as it saw its introduction into information architecture and science. Most recent sources (ANSI/NISO, 2005; Beck & Pinto, 2002; Jernst, 2003; Jonker, 2006; Klein & Smith, 2005; Studer et al., 2001; Ullrich et al., 2003; Will, 2007) describe ontology in this field. Its best-known definition is that by Gruber (1995): "an explicit, formal specification of a shared conceptualisation". An analysis of this definition is expedient, as it concentrates some important components. Firstly, 'explicit' means that the concepts included in the ontology are clearly defined, as are the constraints on their use. 'Formal' refers to the language of the ontology. A formal language is computer-readable: the computer 'understands' the relationships –also called 'formal semantics'- within the ontology. This way, they can be used to support computer applications. Examples of formal representation languages for ontologies include RDF (Beckett, 2004) (Resource Description Framework; cf. the Nautilus ontology (Dieng-Kuntz et al., 2006)), F-Logic (Kifer et al., 1990), or Frame Logic (e.g. FLORID (Frohn et al., 1997)), KIF (Knowledge Interchange Format, e.g.), a later version of which – Common Logic - has been submitted to and approved by ISO, OIL (Van Hamelen et al., 2001) (Ontology Inference Layer), DAML+OIL, a combination of DAML (DARPA12 Agent Markup Language) and OIL, and OWL (Bechhofer et al., 2004) (Web Ontology Language; e.g. Basic Clinical Ontology for breast cancer13), which combines OIL and DAML+OIL. Ontologies written in these formal languages can be used for inferencing or to support other software applications.

The last components of the definition, 'shared' and 'conceptualization', imply that this abstract model of phenomena in the world has been agreed upon by a group of users or experts.

[12] DARPA stands for Defense Advanced Research Projects Agency
[13] http://acl.icnet.uk/~mw/MDM0.73.owl

As observed by Garshol (2004), an ontology usually consists of concepts, relations and properties, but "*exactly what is provided around this varies*". The basic elements of an ontology are concepts, grouped into classes. The actual object referred to by the concept, is an individual or instance. Relations between concepts and instances are often called roles. Attributes or properties are assigned to the concepts or instances.

Thesauri and taxonomies, and even glossaries are often considered bedfellows within the category of -simple- ontologies: they organize the concepts or terms of a knowledge domain, and all four can be used for indexing and searching information. An ontology, however, distinguishes itself from the other tools mainly by allowing more types of semantic relationships, which makes the ontology much more versatile, more powerful. In addition, an ontology usually structures its concepts not as a hierarchy, but as a network or a web.

Ontologies were initially conceived as a way to represent knowledge; however now they are "*intended to support the vision of the semantic web through providing structured metadata about resources and a foundation for logical inferencing*" (L.M. Garshol, 2003). They are aimed at giving a truthful reflection of reality, and this has repercussions on their further development for use in information retrieval.

In conclusion, the term 'ontology' is polysemous due to historical and interdomain shifts. Originally, it was the study of being, the outcome of which was a representation of what exists, or 'an ontology'. This later became a schematic representation of fields of knowledge with concepts and their interrelationships. In information science, this structure is formalized and can be used for computer applications, including information indexing and retrieval.

2.3.5. Topic maps

Taxonomies, thesauri and ontologies were originally designed to represent knowledge. Later, and even more so with the advent of the Internet, they started being used as indexing vocabularies, facilitating information and document retrieval. Topic maps, on the other hand; were specifically designed for information indexing and retrieval and consist of a knowledge layer –comparable to an ontology- and a resources layer. The

knowledge layer (called "topic space" in figure 4) is usually a semantic network deduced from the resources layer or pool and not – as an ontology – designed by experts as a representation of reality.

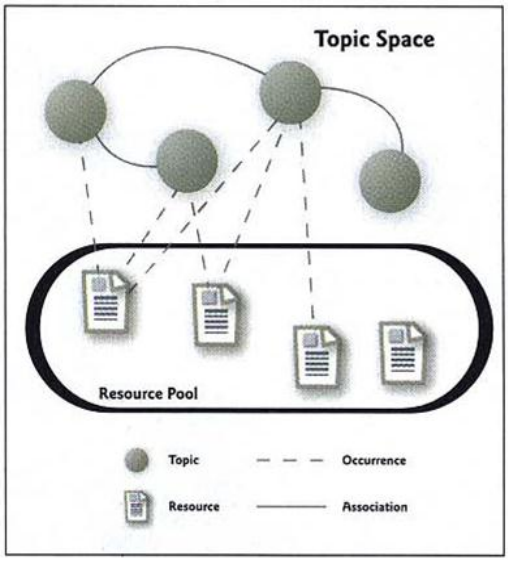


Figure 4: Structure of topic maps (Ahmed, 2002)

The distinction between ontologies and topic maps runs parallel to that between knowledge management and information management: ontologies cover only the knowledge itself, whereas topic maps also involve storing and tracking resources in which this knowledge may be found.

The idea of topic maps emerged in the early nineties when the Davenport Group met to discuss ways to merge indexes, glossaries, thesauri, cross references, etc. This new index was to reflect the structure of the knowledge it represented. Their efforts resulted in 'topic navigation maps', which were adopted as an ISO work item in 1996. In 2000, these topic navigation maps were renamed 'topic maps' and became a new ISO standard[14].

[14] The definition of topic maps proposed in ISO/IEC 13250 is a circular definition, thus not helping to grasp the exact meaning of 'topic maps':

> "a) A set of information resources regarded by a topic map application as a bounded object set whose hub document is a topic map document conforming to the SGML architecture defined by this International Standard.
> b) Any topic map document conforming to the SGML architecture defined by this International Standard, or the document element (topicmap) of such a document.
> c) The document element type (topicmap) of the topic map document architecture."

Ontologies describe concepts -represented by terms- with their attributes and relationships and divide them into classes. These classes consist of concrete or abstract individuals or instances. Topic maps have subjects represented by topics and described by associations and occurrences. Topics are described in more detail by *topic names* and *topic types*, *association types* and *occurrence roles* (see also Pepper (2000)). In addition to this difference in structuring the knowledge layer, topic maps have some other important distinguishing characteristics, mainly concerning their development, initial purpose and standards.

The main differences and similarities are summarized in the following table:

Table 3: Differences between ontologies and topic maps

	Ontologies	Topic maps
Definition	An ontology is a representation of reality.	A topic map is an information retrieval tool which consists of a resources layer linked to a knowledge layer.
Differences	is an organization of **knowledge**	consists of a **knowledge** layer (comparable to an ontology) and a **resources** layer
	can be used as an information retrieval tool when the knowledge is linked to resources	**is designed as** an information retrieval tool
	knowledge structure is designed by **domain expert(s)** and later **linked to** the documents or other resources	knowledge structure is **deduced from** the resources
	the knowledge layer is a **representation of reality** (within a specific domain)	the knowledge layer is a **representation of the knowledge in the resources**
	the knowledge structure consists of **concepts**, **classes, attributes, relations** and **individuals**	the knowledge structure consists of **subjects, topics** (+ names and types), **associations** (+ types) and **occurrences** (+ roles)
	not a standardized format as such	topic maps is an **ISO standard format**

As observed above, the knowledge framework in ontologies is designed from scratch by a domain expert in order to support the vision of the semantic web. In topic maps,

however, this knowledge layer is deduced from the documents or information contained in the resource layer. Pepper (2000) and Hummel (2004) consider the separation into two layers and the standardized format respectively as the topic maps' strengths. These qualities improve the navigational function of topic maps and their interoperability with other topic maps, and even with indexes, thesauri, taxonomies, ontologies and other traditional classification schemes. As confirmed by Garshol (2004), "topic maps do not offer more, but other possibilities with regard to the knowledge represented, i.e. a flexible model with an open vocabulary".

The format of topic maps is captured in an ISO standard, which also improves the efficiency and interoperability with other tools. Ontologies lack this standardization and are thus less suitable for exchange. The format of ontologies is not standardized, but many of their corresponding representation languages (XML, RDF, RDF Schema, and OWL) are.

3. Applications in the (bio)medical domain

The last decades have witnessed an information explosion in the (bio)medical domain, and with it the increasing need for solid vocabularies, terminologies and classification systems. They include – next to the numerous medical glossaries and dictionaries - the UMLS resources, the Gene Ontology, MeSH, SNOMED and OpenGALEN. The present section attempts to characterize these systems in terms of the definitions given above.

3.1. Linguistic tools in the biomedical domain

3.1.1. Medical glossaries, lexicons and dictionaries

Wikipedia's Glossary of medical terms related to communications disorders and the *Ziekenhuis.nl woordenboek* are examples of mono- and bilingual glossaries respectively. They cover terms from the field of medicine or social services, and comply with the definition of 'glossary' given in this article in that they are lists of terms, arranged alphabetically, with definitions.

The *Specialist Lexicon*, which is included the UMLS as one of the Knowledge Sources, meets the criteria for lexicons described in this article. It was designed for use in natural language processing (NLP) and is intended to be a general English lexicon that includes many biomedical terms. The linguistic information includes inflectional variants and derivations, acronyms, spelling variants and, when applicable, verb, noun or adjective complementation. An example of a lexical record can be seen in figure 5:

```
<lexRecord>
        <base>dpn</base>
        <eui>E0023874</eui>
        <cat>noun</cat>
                <inflVars cat="noun" cit="dpn" eui="E0023874" infl="base" type="basic"
                unInfl="dpn">dpn</inflVars>
                <inflVars cat="noun" cit="dpn" eui="E0023874" infl="singular"
                type="basic" unInfl="dpn">dpn</inflVars>
                <inflVars cat="noun" cit="dpn" eui="E0023874" infl="plural" type="inv"
                unInfl="dpn">dpn</inflVars>
                <inflVars cat="noun" cit="dpn" eui="E0023874" infl="plural"
                type="metareg" unInfl="dpn">dpns</inflVars>
                <inflVars cat="noun" cit="dpn" eui="E0023874" infl="plural"
                type="metareg" unInfl="dpn">dpn's</inflVars>
                        <nounEntry>
                                <variants>inv</variants>
                                <variants>metareg</variants>
                        <variants>uncount</variants>
                </nounEntry>
        <acronyms>diphosphopyridine nucleotide|E0023044</acronyms>
        <acronyms>day postnatal</acronyms>
        <abbreviations>transcription factor deadpan</abbreviations>

</lexRecord>
```

Figure 5: Example of a lexical record in the Specialist Lexicon

The *Pinkhof geneeskundig woordenboek* and the *Diccionari d'infermeria* are examples of a monolingual and a multilingual dictionary respectively. They give definitions and information on the origin of the word, which is generally Latin or Greek, and on gender.

3.1.2. The Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages

The *Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages* is a controlled vocabulary in the form of a glossary. Each 'technical' term in this glossary has a popular variant which should be considered as the preferred term in texts intended for patients. The glossary was initiated in the framework of the 92/27/EEC Directive, which made the inclusion of patient information leaflets (PILs) in every medication package mandatory in the Member States of the European Community and

stipulated that the leaflets had to be written in understandable language. As the use of terminology is often an important factor in the readability of these information leaflets, a glossary with popular variants for medical or technical terms was very useful. This controlled vocabulary was thus intended to help writers and translators make their PILs understandable for the general public. The Glossary meets the requirements for glossaries, i.e. it is a list of words with their – English – definitions. However, it is more than just a glossary, as it also contains preferred and non-preferred terms. In summary, this is a controlled vocabulary in the form of a glossary.

3.1.3. The European Multilingual Thesaurus on Health Promotion

The *European Multilingual Thesaurus on Health Promotion* is a merger of 3 thesaurus projects in 12 languages and is used as a linguistic tool: it should stimulate the uniform use of terms related to health promotion and health education in Europe, as a such a shared language supports the efficient exchange of information. This thesaurus is thus used as a controlled vocabulary, with preferred and non-preferred terms. The ISO standards 2788 and 5964 were used as construction guidelines - i.e. the equivalence (UF, USE), associative (RT) and hierarchical relationships (BT, NT) are specified - although the thesaurus is not used for bibliographic retrieval.

3.2. Knowledge management and medical coding

3.2.1. The ATC classification

The ATC (Anatomical Therapeutic Chemical) classification is a system developed by the WHO for the classification of drugs and other medical products. Applying Cann's view to this classification, one could state that this is a specific, documentary, enumerative classification. Specific, because it covers a part of the medical domain, namely medical substances. Documentary, because it functions as an information management and retrieval tool, and enumerative because it lists the classes and subclasses in a specific domain of interest and it is created from the top down.

The classification consists of 14 main classes, each one referring to an anatomical main group, e.g. nervous system (**N**). The next level is indicated by two digits and contains

therapeutic subgroups, e.g. anti-parkinson drugs (N**04**). The third level, which is indicated by one letter, refers to the pharmacological subgroup, e.g. dopaminergic agents (N04**B**). The fourth level, again a letter, is a designation of the chemical subgroup, e.g. dopamine agonists (N04B**C**), and the last two digits indicate the chemical substance, e.g. pramipexole (N04BC**05**; see table 4).

Table 4: Structure of the ATC Classification

	ATC level	ATC code	ATC text
1	Anatomical main group	N	*Nervous system*
2	Therapeutic subgroup	N04	*Anti-parkinson drugs*
3	Pharmacological subgroup	N04B	*Dopaminergic agents*
4	Chemical subgroup	N04BC	*Dopamine agonists*
5	Chemical substance	N04BC05	*Pramipexole*

The ATC classification is mainly used to produce statistics about drug use, but also for the registration process of drugs.

3.2.2. The International Classification of Diseases and Related Health Problems (ICD)

The *International Classification of Diseases and Related Health Problems* is published by the World Health Organization (WHO) and classifies diseases, signs, symptoms, complaints, social circumstances and causes of injury or disease. It is used in statistics, in automated decision support and in reimbursement systems. ICD-10, the tenth revision of ICD, is the most recent version of the classification. The first level of ICD-10 consists of 22 classes, each of which has several subclasses. The first letter in the code refers to the chapter, whereas the following digits specify the disease. For instance, in C18.7, C refers to malignant neoplasms, 18 refers to malignant neoplasms of the colon, and the numeric symbol after the decimal point further specifies the disease, in this case malignant neoplasm of the sigmoid colon.

ICD-10 is a specific, documentary and enumerative classification: it covers a specific domain, it is used to store and retrieve medical data and created from the top down.

3.2.3. *The International Classification of Primary Care (ICPC)*

The International Classification of Primary Care was designed by the WICC (WONCA International Classification Committee) for the classification of reasons for encounter (RFE), problems, diagnoses, interventions and the ordering of these data in an episode of care structure. Chapter ten of the second version of ICPC has been converted into an electronic file, i.e. ICPC-2-E, is specifically designed for use in electronic patient records (EPR) and for research purposes. It is to be used together with the first nine chapters of ICPC-2. As ICD-10 is more fine-grained and allows for documentation at the level of individual patients (Okkes et al., 2000), this classification was the perfect complement to ICPC-2. When ICD-10 was made available, together with its various translations, the WICC decided that all translations of ICPC were to relate to ICD-10, in order to allow for a better structuring of EPRs. For the Netherlands and the Dutch-speaking part of Belgium, this resulted in the ICPC-2/ICD-10 thesaurus (see 3.4.4).

ICPC-2 is a specific, documentary and enumerative classification which has a bi-axial structure. There are 17 main classes with an alpha code referring to the location of the complaint, and 7 components with a two-digit numeric code, which organize each of these classes. ICPC-2 is included in the UMLS (see 3.3.4).

	Chapters																
Components	**A**	**B**	**D**	**F**	**H**	**K**	**L**	**N**	**P**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
1. Symptoms																		
2. Diagnostic, screening, prevention																		
3. Treatment, procedures, medication																		
4. Test results																		
5. Administrative																		
6. Other																		
7. Diagnoses, disease																		

A. General	K. Circulatory	S. Skin	Y.Male genital
B. Blood, blood forming	L. Musculoskeletal	T. Metabolic, endocrine, nutrition	Z. Social
D. Digestive	N. Neurological	U. Urinary	
F. Eye	P. Psychological	W. Pregnancy, family planning	
H. Ear	R. Respiratory	X. Female genital	

Figure 6: Structure of ICPC-2

3.2.4. ICPC-2/ICD-10 thesaurus

The ICPC-2/ICD-10 thesaurus was created at the University of Amsterdam, Department of Family Practice, in collaboration with the Department of General Practice and Primary Health Care of the Ghent University. As stated above, ICD-10 is the perfect complementation for ICPC-2, as it is more fine-grained. The result of this combination is a system with doubly encoded clinical labels: each term has two codes, an ICD-10 and an ICPC-2 code.

This bilingual (English-Dutch) terminology is called a "thesaurus" because it has a hierarchical structure and synonyms for many of the concepts. Moreover, it is a controlled language used to store medical information. However, not all the requirements to designate a vocabulary as a thesaurus are fulfilled: there are no associative relationships. 3BT (Belgian Bilingual Biclassified Thesaurus) is a continuation of the ICPC2/ICD10 Thesaurus, but with French translations added to it. The designation "thesaurus" is a misnomer in this case, as the system does not meet all the criteria described in the ISO standards for thesauri: it has no associative relationships either. However, some terms do have synonyms or entry terms that lead the system to the correct concept. Like ATC, ICD and ICPC, this is a specific, enumerative, documentary classification used for medical coding.

3.2.5. SNOMED CT

The *Systematized Nomenclature of Medicine, Clinical Terms*, or SNOMED CT, provides a comprehensive terminology covering concepts in health care, i.e. diseases, clinical findings and procedures. This terminology, which is also available in German and in Spanish, is designed to support data retrieval and automated inferencing (e.g. for clinical decision support). SNOMED CT is based on the SNOMED Reference Terminology (SNOMED RT) and the British Clinical Terms, version 3. It also cross-maps to a number of existing terminologies and coding systems, such as ICD-9-CM, ICD-10 and LOINC (Logical Observation Identifiers Names and Codes).

The clinical concepts included in SNOMED CT are organized in nineteen hierarchies - alternatively called axes - and linked with definitions in formal logic. Each term in

SNOMED CT has a unique numeric code, a unique name ('fully specified name'), and a 'description' comprising one preferred term and one or more synonyms.

Two main types of relationships are established in this ontology: hierarchical and attribute relationships. Hierarchical 'is-a' relationships are defined within one axis, whereas the attributes link concepts from different hierarchies. Attribute relationships include finding site, causative agent, occurrence, stage, etc.

The prerequisites for an ontology in information science are thus fulfilled: the SNOMED CT terminology represents knowledge from a specific domain (health care), is concept-oriented, and the definitions are formalized. Moreover, almost any semantic relationship can be expressed in this ontology.

3.2.6. OpenGALEN

OpenGALEN is a multilingual terminology and coding system for the classification of surgical procedures, electronic healthcare records (EHCRs), clinical user interfaces, decision support systems, knowledge access systems, and natural language processing.

The OpenGALEN Foundation (Open Galen Foundation s.d.) defines 'ontology' as "the set of primitive, high level categories in a knowledge representation scheme together with any taxonomy which structures those categories". In this view, the OpenGALEN system is an ontology indeed. However, it also fulfils the requirements of an information retrieval ontology in the strict sense: it represents the concepts of a specific domain with formalized relationships, making the ontology re-usable in other applications. Moreover, the ontology allows the expression of extensive semantic relationships, including "kind-of", "part-of", "connects", "branch-of", "serves" and laterality relations.

3.3. Bibliographic retrieval

3.3.1. The NCBI Entrez Taxonomy

The NCBI Entrez Taxonomy[15] is a hierarchical structure which contains all organisms represented in GenBank, with at least one nucleotide or protein sequence. There are seven top classes, i.e. arachea, bacteria, eukaryota, viroids, viruses, other and unclassified. The information provided for each concept is quite elaborate and includes an ID, a rank, a genetic code, synonyms, and information as to the location in the taxonomy ("linkage"; see figure 7).

Taxonomy ID: 242703
Inherited blast name: **crenarchaeotes**
Rank: species
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)
Other names:
synonym: **'Acidilobus saccharovorans'**
Lineage(full)
cellular organisms; Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Acidilobus

Figure 7: Extract from the NCBI Entrez Taxonomy

The Entrez Taxonomy complies with the definition given in 1.2.1: it is a hierarchical classification structure in which meaning is passed from more generalized to more specialized taxa.

3.3.2. MeSH

MeSH is an acronym for **Me**dical **S**ubject **H**eadings, a controlled vocabulary produced by the National Library of Medicine (NLM), geared specifically for information retrieval. MeSH is used for indexing and searching journal articles in MEDLINE and other resources from the NLM Catalog.

[15] http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root

The MeSH vocabulary consists of preferred terms, or descriptors, and entry terms. However, MeSH is more than 'just' a controlled vocabulary, it is a fully fledged thesaurus. The equivalence relationship is established by entry terms, which can be synonyms, near synonyms, abbreviations, or alternate forms of the MeSH term. Besides the equivalence relationship, two other typical thesaurus relations, i.e. hierarchical and associative relations, are represented.

The concepts are structured into a hierarchy, the MeSH tree, with sixteen main branches. Each descriptor can have multiple parents and can consequently appear in several places in the tree. This can be illustrated by looking at a specific example, e.g. the Wolfram syndrome. This descriptor appears under the following subcategories: Nervous System Diseases [C10], Eye Diseases [C11], Male Urogenital Diseases [C12], Female Urogenital Diseases and Pregnancy Complications [C13], Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16], Nutritional and Metabolic Diseases [C18] and Endocrine System Diseases [C19].

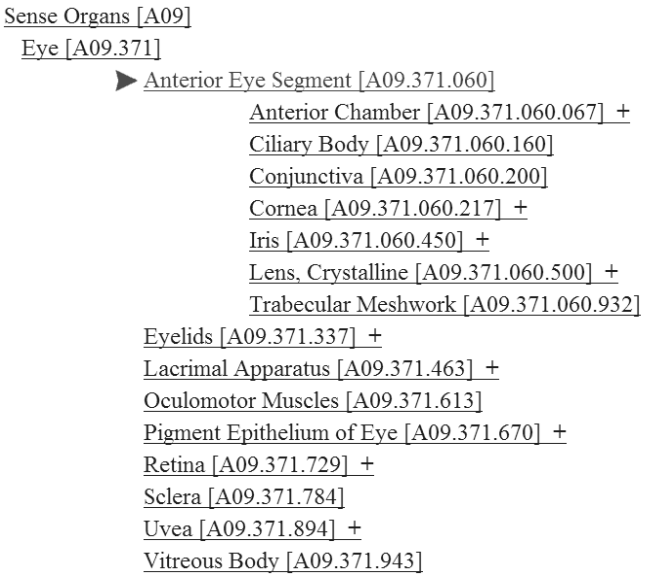


Figure 8: Expressive or hierarchical notation (MeSH)

Each descriptor has a notation – one or several MeSH number(s) - which is an indication of the concept's relationship to its neighboring concepts. This type of notation is referred to by Aitchison et al. (2000) as an "expressive notation" or "hierarchical notation" (as opposed to (semi-)ordinal, synthetic and retroactive notations). The length of the number indicates the specificity of the term: the longer the number, the

more specific the concept. Figure 8 shows that Eye [A09.371] is broader than Anterior Eye Segment [A09.371.060], which, in turn, is broader than Anterior Chamber [A09.371.060.067].

When applied in information retrieval, the MeSH thesaurus can be an extremely valuable tool. It allows explosion of the search terms, and in Entrez PubMed, the terms entered by the searcher are automatically mapped to the appropriate MeSH term (NN/LM, 2006). Term explosion, as described above, is a technique which increases the search yield considerably by searching not only for the term itself, but also for its narrower terms.

When examined for compatibility with the definition of a thesaurus as an information retrieval tool, the MeSH thesaurus proves to fulfill almost all requirements. It is a controlled vocabulary, with its descriptors and its non-preferred entry terms, which lead the searcher to the descriptors. The MeSH tree is organized hierarchically and includes the standardized relations as described in ISO 2788 (International Organization for Standardization, 1986) – the hierarchical, associative and equivalence relationship. These relationships are a priori relationships, i.e. they exist independently of the contents of the articles indexed with MeSH terms. Moreover, each term has a scope note, which contains background information on the usage and scope of the term. Scope notes can contain a definition formulated by the MeSH project partners or copied from other sources, like dictionaries or biomedical publications.

Greenberg (2004), mentions a slight difference between thesauri for information retrieval and subject headings: thesauri generally tend to support post-coordinate searching, whereas subject headings have a pre-coordinated syntax. In pre-coordinated vocabularies, combinations of concepts are made at the indexing stage by the indexers, rather than at the stage of query formulation by the user. This means that the searcher can select very specific, unambiguous and "ready-made" queries instead of combining single-concept terms. Compare, for example, the pre-coordinated MeSH term "Physiological effects of drugs" and the terms "physiological", "effect" and "drugs" in post-coordination. The advantages of pre-coordination described in (Cataloging Policy and Support Office, 2007) include proximity searches, where the searcher uses the relationships between concepts to select the best query. Pre-coordinated terms can be

very useful for browsing, as they enable hierarchical displays. One of the disadvantages stated in (Cataloging Policy and Support Office, 2007) are that pre-coordination requires human manual construction, an expensive and time-consuming task. Another disadvantage of pre-coordination might be that some end-users who are not familiar with this method of searching, might experience some problems. Post-coordination implies that concepts will have to be combined at the searching stage using Boolean operators.

Subject headings have multi-word terms, and often use inverted word order. MeSH can thus be defined as a thesaurus with the syntax of a subject heading list.

3.3.3. Controlled vocabularies

Controlled vocabularies used in bibliographic retrieval are usually incorporated into another structure, like a thesaurus (MeSH) or an ontology (the UMLS knowledge sources combine several controlled vocabularies).

3.3.4. The UMLS Knowledge Sources

The UMLS (Unified Medical Language System) Knowledge Sources combine three of the vocabulary systems described above: a thesaurus (the Metathesaurus), a lexicon (the SPECIALIST Lexicon) and an ontological structure (the Semantic Network).

The Metathesaurus consists of a large number of source vocabularies, including MeSH, SNOMED CT, the Gene Ontology, and other controlled vocabularies. Partly as a consequence of this combination of vocabularies, the Metathesaurus has a polyhierarchical structure. The Metathesaurus can be used in a wide range of applications, including information retrieval, and it becomes more powerful when used in combination with the SPECIALIST Lexicon and the Semantic Network.

The SPECIALIST Lexicon covers both the English general language and concepts from the field of biomedicine. It provides syntactic, morphological and orthographic information about the terms included in the lexicon.

A third component of the UMLS Knowledge Sources is the Semantic Network, which consists of Semantic Types, or broad subject categories, and Semantic Relations between

these Semantic Types. This tool enables a consistent categorization of the concepts in the Metathesaurus.

The combination of the Knowledge Sources could be regarded as an ontology, as it represents knowledge from a specific field, with its concepts and extensive relationships. Furthermore, the Semantic Relations are expressed in a formal language. The combination of Semantic Types and Semantic Relationships makes this knowledge source much more versatile than the average thesaurus.

A medical ontology is being developed by the Lister Hill National Center for Biomedical Communications, a research division of the U.S. National Library of Medicine. This ontology will combine the UMLS with SNOMED-RT, GALEN and MEDLINE citations and will represent a "model for proximity between medical concepts" [16].

3.3.5. The Gene Ontology

The Gene Ontology (GO) is a controlled vocabulary developed by the Gene Ontology Consortium for the annotation of gene products in model organisms. This vocabulary consists of three separate hierarchies, each representing concepts from a different subdomain: cellular components, molecular functions and biological processes. It has a polyhierarchical structure, i.e. a narrower term or hyponym can have more than one broader terms or hypernyms, and it has a simple RDF syntax.

Despite its name, the GO is not an ontology as described in this article. Two types of relationships are present in this controlled vocabulary, namely the hierarchical *is-a* and *part-of* relationships and the equivalence relationship. The term 'ontology' here refers to the fact that knowledge about a specific domain is represented, including the relationships between the concepts.

Smith et al. (2003) give an overview of the requirements for the Gene Ontology to become a cost-effective and semantically consistent system. These changes would convert the Gene Ontology into a system with the relational characteristics of a true

[16] http://lhncbc.nlm.nih.gov/lhc/servlet/Turbine/template/research,langproc,MedicalOntology.vm

ontology. However, making these changes would raise many difficulties. As a result, the Gene Ontology will probably remain in its current form, i.e. a controlled vocabulary.

3.3.6. Topic Maps

Beier and Tesche (2001) developed a medical information retrieval system, using the Medical Subject Headings (in English and German) and their classification as the knowledge layer, and the resources layer includes AHCPR Guidelines, journal articles and selected internet sites. This is a federated search system, i.e. a system which simultaneously searches several databases and/or web resources. The query entered by the user is automatically expanded with the topic name (the preferred term), synonyms, translations and definition.

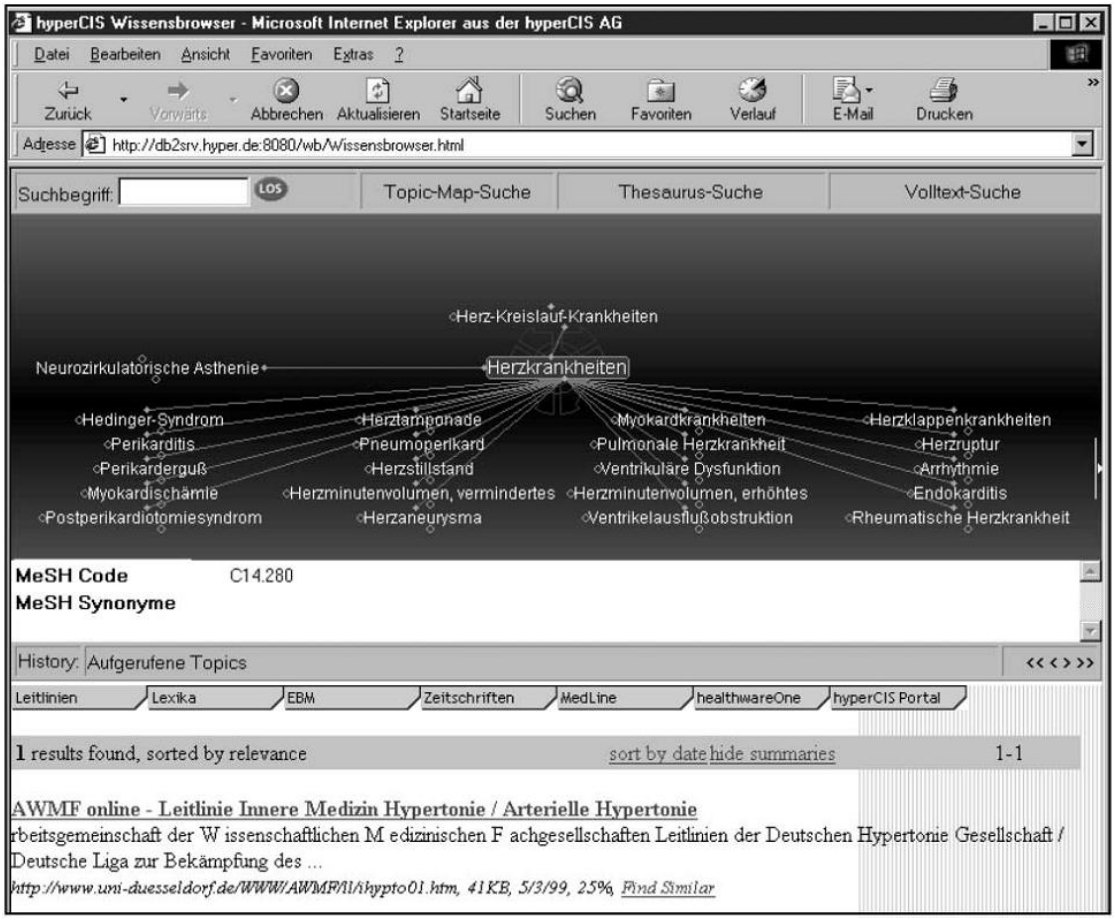


Figure 9: Interface of the MeSH-based topic map created by Beier and Tesche (2001)

The interface (see Figure 9) clearly shows the typical topic map structure of the resources layer and the superimposed knowledge layer. Between both layers, some extra MeSH information (MeSH code, definition and annotations, synonyms and translations) is displayed, in order to help the user find the right topic name for his or her search. The user can select the resources in which he wants the engine to search.

This topic map complies with the ISO standard and with the description of topic maps given in section 2.3.5, except that the knowledge layer was not deduced from the resources.

4. Conclusion

There is a need for consistent terminology in the domains of linguistics, knowledge management and information retrieval, as in most fields of knowledge. Terms such as taxonomy, classification, thesaurus and ontology are often used interchangeably, resulting in definitions which are formulated from different perspectives.

Not only are the terms used in different ways, their scope may also change. When terms are adopted in other fields –a shift which often has a historical aspect- this may cause some confusion.

Unambiguous definitions are proposed for each of the terms in question, depending on the context they are used in, and criteria are presented for a more consistent use of the various competing designations. Some of the best-known vocabularies pertaining to biomedical linguistics, knowledge management and bibliographic retrieval are reviewed and examined for their compatibility with the definitions given in this article. We concluded that the use of the designations ‘ontology’ or ‘thesaurus’ in the biomedical domain - as elsewhere- is not always consistent. More specifically, we found that the ICPC-2/ICD-10 thesaurus and 3BT are not thesauri, but bicoded classifications and that the Gene Ontology is not really an ontology but a controlled vocabulary.

Table 5 below gives an overview of the systems in biomedicine in a two-dimensional structure: according to their domain of application (linguistics, knowledge management – including medical registration- and bibliographical retrieval) and the kind of

vocabulary (taxonomy, classification, thesaurus, controlled vocabulary, ontology or topic maps) they represent.

Table 5: Overview of (bio)medical vocabulary systems

	Linguistics	Knowledge Management	Bibliographic retrieval
Glossary, lexicon and dictionary	Wikipedia's Glossary of medical terms related to communications disorders; Ziekenhuis.nl dictionary; Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages; The Specialist Lexicon; Pinkhof geneeskundig woordenboek; Diccionari d'infermeria		
Taxonomy		Linnaean taxonomy	NCBI Entrez Taxonomy
Classification		ICD, ICPC, 3BT, ICPC2/ICD10 thesaurus	
Thesaurus	European Multilingual Thesaurus on Health Promotion		MeSH
Controlled vocabulary	Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages		MeSH, several vocabularies in the UMLS
Ontology		OpenGalen SNOMED	UMLS
Topic maps			HyperCis Topic Map

References

Agro, Greg. (2004). Classifications and Taxonomies [PowerPoint presentation]. Austin: University of Texas.

Ahmed, Kal. (2002). Introducing Topic Maps: A Powerful, Subject-Oriented Approach to Structuring Sets of Information. (Content Management). *XML Journal, 3*(10), 22-27.

Aitchison, Jean, Gilchrist, Alan, & Bawden, David. (2000). *Thesaurus Construction and Use: A Practical Manual* (4th ed. Vol. 1). London: Aslib IMI.

Ananiadou, Sophia, & McNaught, John. (2006). *Text Mining for Biology and Biomedicine*. Norwood: Artech House.

ANSI/NISO. (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. Bethesda, Maryland: NISO Press.

Bechhofer, Sean, Van Harmelen, Frank, Hendler, Jim, Horrocks, Ian, McGuinness, Deborah L., Patel-Schneider, Peter F., & Stein, Lynn Andrea. (2004). OWL Web Ontology Language Reference. In M. Dean (Ed.), *Schreiber, Guus*: W3C.

Beck, Howard, & Pinto, Helena Sofia. (2002). *Overview of Approach, Methodologies, Standards, and Tools for Ontologies*. draft. University of Florida; Universidade Técnica de Lisboa.

Beckett, Dave. (2004). RDF/XML Syntax Specification (revised). Retrieved 26/11/2008, 2008, from http://www.w3.org/TR/rdf-syntax-grammar/

Beier, Jürgen, & Tesche, Tom. (2001). Navigation and interaction in medical knowledge spaces using topic maps. *International Congress Series, 1230*, 384-388.

BSI. (2005). *Structured vocabularies for information retrieval - guide* (Vol. 1). London: BSI British Standards.

Cann, John. (1997). Principles of classification – suggestions for a procedure to be used by ICIS in developing international classification tables for the construction industry: NBS Services, ICIS.

Cataloging Policy and Support Office. (2007). Pre- vs. Post-Coordination and Related Issues. In A. Management (Ed.): Aquisitions and Bibliographic Access Directorate, Library Services, Library of Congress.

Chowdhury, G.G. (2003). *Introduction to modern information retrieval* (2nd ed.). London: Facet Publishing.

CILF. (1993). *Wörterbuch für Industrie und Technik*. Paris: Conseil International de la Langue Française.

Davis, P.H. , & Heywood, V.H. (1963). *Principles of Angiosperm Taxonomy*. Princeton NJ: Van Nostrand.

Dieng-Kuntz, Rose, Minier, David, Růžička, Marek , Corby, Frédéric , Corby, Olivier , & Alamarguy, Laurent. (2006). Building and using a medical ontology for knowledge management and cooperative work in a health care network. *Computers in Biology and Medicine, 36*(7-8), 871-892.

Engineers Joint Council. (1967). *Thesaurus of Engineering and Scientific Terms: A List of Engineering and Related Terms and their Relationships for Use as a Vocabulary Reference in*

Indexing and Retrieving Technical Information. New York: Engineers Joint Council and the US Department of Defense.

Frohn, Jurgen, Himmeroder, Rainer, Kandzia, Paul-Th., & Lausen, Georg. (1997). *FLORID: A Prototype for F-Logic.* Paper presented at the Proceedings of International Conference on Data Engineering, Birmingham, UK.

Garshol, L.M. (2003). Living with topic maps and RDF. Retrieved 05/01/2007, 2007, from http://www.ontopia.net/topicmaps/materials/tmrdf.html

Garshol, L.M. . (2004). Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all. Retrieved 13/06/2006, from www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html

Greenberg, Jane. (2004). User Comprehension and Searching with Information Retrieval Thesauri. *Cataloging & Classification Quarterly, 37*(3), 103 - 120.

Gruber, Thomas R. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human Computer Studies, 43*(5-6), 907-928.

Hagedorn, Kath. (2000). The Information Architecture Glossary. Retrieved 04/07/2006, 2006, from http://argus-acia.com/white_papers/ia_glossary.pdf

Harter, S.P., & Hert, C.A. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST), vol. 32*, 3-94.

Hummel, Benedikt. (2004). *Einsatz und Nutzenpotentiale von Topic Maps: Ein State-Of-The-Art Bericht.* (Diplom-Bibliothekar), Fachhochschule Potsdam, Potsdam. Retrieved from http://forge.fh-potsdam.de/~buettner/Lehre/Diplomarbeiten/Hummel_F.pdf

Indira Gandhi National Open University. (2006). Part II : Classification Schemes *Indexing languages* (pp. 56-88). New Delhi: Indira Gandhi National Open University.

International Organization for Standardization. (1985). *ISO 5964 - Documentation - Guidelines for the establishment and development of multilingual thesauri*. Geneva: ISO.

International Organization for Standardization. (1986). *ISO 2788 - Guidelines for the Establishment and Development of Monolingual Thesauri*. Geneva: ISO.

International Organization for Standardization. (2007). *Information and documentation. Guidelines for the establishment and development of thesauri [revision of ISO 2788 and 5964]*: ISO/ TC 46/ SC9.

ISO/IEC_11179-2. (2005). *Information technology — Metadata registries (MDR) — Part 2: Classification* (2nd ed. Vol. 2). Geneva: ISO copyright office.

Jernst. (2003, 15/01/2003). What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model? . Retrieved 10/05/2006, from www.metamodel.com/article.php?story=2003011223271

Jonker, Rienk. (2006). Termen en begrippen - Informatiebeheer. Retrieved 15/01/2006, 2007, from http://labyrinth.opweb.nl/files/termenbegrippen.pdf

Kagolovsky, Y., & Moehr, J. R. (2003). Terminological problems in information retrieval. *J Med Syst, 27*(5), 399-408.

Kifer, Michael, Lausen, Georg, & Wu, James. (1990). Logical Foundations of Object-Oriented and Frame-Based Languages: University of Mannheim.

Kilgarriff, Adam, & Yallop, Colin. (2000, May/June). *What's in a thesaurus?* Paper presented at the Proceedings of the Second Conference on Language Resources and Evaluation, Athens, Greece.

Klein, Gunnar O., & Smith, Barry. (2005). *Concept Systems and Ontologies*. Recommendations based on discussions between realist philosophers and ISO/CEN experts concerning the standards addressing "concepts" and related terms. Centre for Medical Terminology, Karolinska Institutet, Stockholm

Department of Philosophy, University at Buffalo, NY. Retrieved from http://ontology.buffalo.edu/concepts/ConceptsandOntologies.pdf

Landau, S. (1984). *Dictionaries: The Art and Craft of Lexicography*. New York: Charles Scribner's Sons.

Mawson, C.O. Sylvester. (1922). *Roget's International Thesaurus of English Words and Phrases* (1st ed.). New York: Thomas Y. Crowell.

Merriam-Webster Inc. (2008). Merriam-Webster Online Dictionary. Retrieved 20/03/2007, 2007, from http://www.merriam-webster.com

NN/LM. (2006). PubMed Expert Searching: Using PubMed to Get Advanced Results. Retrieved 14/03/2007, 2007, from http://nnlm.gov/ner/training/material/NER_PES.doc

Okkes, I. M., Jamoulle, M., Lamberts, H., & Bentzen, N. (2000). ICPC-2-E: the electronic version of ICPC-2. Differences from the printed version and the consequences. *Fam Pract, 17*(2), 101-107. doi: 10.1093/fampra/17.2.101

Pepper, Steve. (2000). *The TAO of Topic Maps.* Paper presented at the XML Europe 2000, Paris, France.

Procter, Paul. (1978). *LONGMAN Dictionary of Contemporary English*. London: Longman Dictionaries.

Ribeiro-Neto, Berthier, & Baeza-Yates, Ricardo A. . (1999). *Modern Information Retrieval*. New York/ Harlow: ACM Press/Addison-Wesley.

Roe, Sandra K., & Thomas, Alan R. (2004). *The Thesaurus: Review, Renaissance and Revision* Binghamton, NY Haworth Information Press.

Roget, P. (1995). Roget's II: The new thesaurus. Third edition. Retrieved 26/02/2009, 2009, from www.bartleby.com/62

Simpson, J.A., & Weiner, S.C. (1989). *Oxford English Dictionary* (2nd ed.). London: Oxford University Press.

Smith, Barry, Williams, Jennifer, & Schulze-Kremer, Steffen. (2003, november 8-12). *The Ontology of the Gene Ontology*. Paper presented at the AMIA Annu Symp Proc, Washington D.C.

Sterkenburg, Piet (Ed.). (2003). *A Practical Guide to Lexicography* (Vol. 6). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Studer, Rudi, Oppermann, Henrik, & Schnurr, Hans-Peter. (2001). *Die Bedeutung von Ontologien für das Wissensmanagement*. Karlsruhe: Ontoprise GmbH.

Ullrich, Mike, Maier, Andreas, & Angele, Jürgen. (2003). *Taxonomie, Thesaurus, Topic Map, Ontologie - ein Vergleich*. White paper. Ontoprise GmbH.

University of London Computer Centre (ULCC). (2003). UNESCO Thesaurus Retrieved 26/02/2009, 2009, from http://www2.ulcc.ac.uk/unesco/#brow

Van Hamelen, Frank, Fensel, Dieter, Horrocks, Ian, McGuinness, Deborah L., & Patel-Schneider, Peter F. (2001). OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems, 16*(2), 38-45.

Van Rees, R. (2003). *Clarity in the usage of the terms ontology, taxonomy and classification*. Paper presented at the CIB73.

Will, Leonard. (2007). Glossary of terms relating to thesauri and other forms of structured vocabulary for information retrieval. Retrieved 01-02-2008, 2008, from http://www.willpowerinfo.co.uk/glossary.htm

Wodtke, Christina. (2002). Mind Your Phraseology! Using Controlled Vocabularies to Improve Findability. Retrieved 24/01/2007, 2007, from http://www.digital-web.com/articles/mind_your_phraseology/

List of figures

List of Tables

2

The role of terminology in medical literature searching

> We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely.
>
> E. O. Wilson, 20th century biologist

Chapter II: PubMed searches by Dutch-speaking nursing students: the impact of language and system experience

Abstract

This study analyzes the search behavior of Dutch speaking nursing students with a nonnative knowledge of English who searched for information in MEDLINE/ PubMed about a specific theme in nursing. We examine whether and to what extent their search efficiency is affected by their language skills. Our task-oriented approach focuses on three stages of the information retrieval process: need articulation, query formulation, and relevance judgment. The test participants completed a pretest questionnaire, which gave us information about their overall experience with the search system and their self-reported computer and language skills. The students were briefly introduced to the use of PubMed and MeSH (**Me**dical **S**ubject **H**eadings) before they conducted their keyword-driven subject search. We assessed the search results in terms of recall and precision, and also analyzed the search process. After the search task, a satisfaction survey and a language test were completed. We conclude that language skills have an impact on the search results. We hypothesize that language support might improve the efficiency of searches conducted by Dutch-speaking users of PubMed.

1. Introduction

The growing amount of information makes it paradoxically difficult to stay abreast of current developments in the biomedical domain and to search for information selectively, even with the help of biomedical bibliographic indexes such as MEDLINE and Embase. Many studies have been devoted to the information retrieval (IR) process (Spink et al., 2001; Sutcliffe et al., 2000), precision and recall, and ways to make this process more efficient (Bin & Lun, 2001; Muin et al., 2005; Wilson, 1999). As English has become the lingua franca of science, the "new Latin" (Eisenberg, 1996), it creates

continuity in the domain, but may also cause problems in the retrieval of information. Scholars whose mother tongue is not English may experience difficulties when conducting a literature search. General language skills are needed for efficient information retrieval (Lankamp, 1989), as well as domain-specific terminology. In addition, searchers have to be familiar with the language of information and documentation science (Mouillet, 1999) to use the interface of the search system effectively. Most studies focusing on query formulation and on the search process in general have been conducted with native English test groups. The present study, however, focuses on difficulties caused by the language barrier for Dutch-speaking users of PubMed[1], a tool designed to search the MEDLINE database and other medical resources through the Internet.

The aim of this study is to describe the efficiency of PubMed searches by Dutch-speaking nursing students (bachelor's and master's level), and to explore the impact of Dutch–English translation problems as well as other characteristics (educational background, computer skills, bibliographic skills) on search efficiency. We focus on performance problems in the need articulation step, on the formulation of efficient queries and on the selection of relevant citations.

2. Method

2.1. Theoretical framework

Sutcliffe and Ennis (1998) distinguish four stages in the information retrieval process: problem identification, need articulation, query formulation, and results evaluation. In the problem identification stage, the user is confronted with an uncertainty or problem about which he or she wants to look up information. Need articulation involves parsing of the problem, which is formulated in natural language, into several knowledge structures (Sutcliffe & Ennis, 1998), i.e., into concepts. Dutch-speaking PubMed users with advanced English-language skills will probably do this parsing in English.

[1] http:// www.ncbi.nlm.nih.gov/pubmed

The query formulation stage is a crucial step in the IR process, as different types of translation actions take place. For native English users, this step includes a transformation of the concepts that resulted from the need articulation stage into search terms, selecting the correct MeSH terms and combining them with Boolean operators, taking into account the specific query syntax of the search system. In our test case, the language barrier also has to be taken into consideration (see Figure 1): the search question is translated into concepts, which are then translated into English search terms. Based on the search terms, PubMed makes one or more suggestions for MeSH terms, from which the user chooses the most appropriate one(s).

Results evaluation or relevance judgment, i.e., comparing the set of retrieved articles to the initial information need and selecting relevant citations, also involves some translation actions, as the searcher needs to read the retrieved information and base relevance judgments on titles and/or abstracts in a foreign language. A first relevance judgment step takes place when the user skims the results and determines whether the set of articles matches his or her information need. If there are some interesting results, the user will start browsing the citations. If not, a new query will be issued. A second, more thorough relevance judgment takes place when the user runs through the individual citations and decides for each of them whether it is relevant or not. If the searcher is not satisfied with the number of citations that result from this search, he or she will formulate a new query.

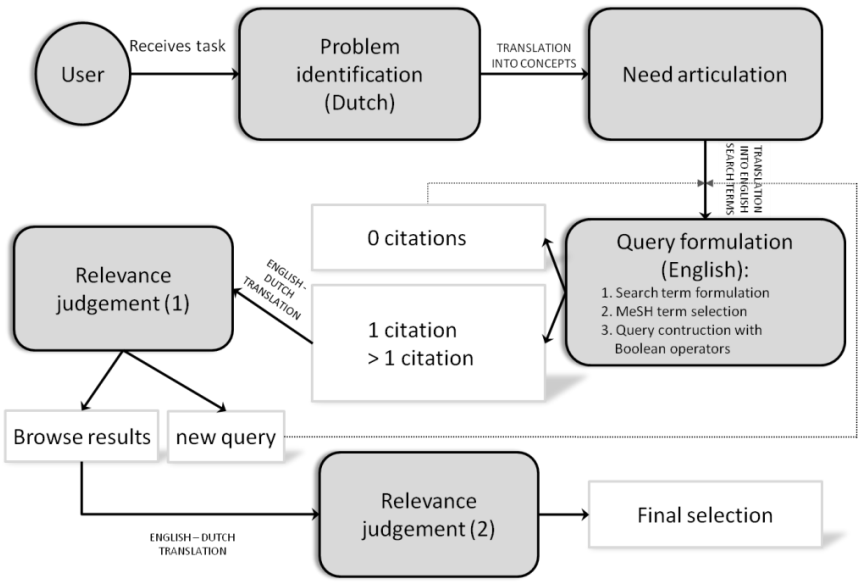


Figure 1: Model for the information retrieval process in a foreign language

2.2. Experimental design

We selected a group consisting of about 60 nursing students pursuing their bachelor's and master's degrees. They had to complete a test which consisted of five parts. First, they completed a pretest questionnaire that focused on computer skills, facility, and experiences with the search system PubMed, and self-assessment of English language skills.

Second, an introduction (10 minutes) was given on the use of MeSH[2] (Medical Subject Headings) in PubMed. MeSH is a controlled vocabulary created by the National Library of Medicine for the purpose of indexing journal articles and books in the biomedical sciences. It helps PubMed users to optimize their literature searches. In this introduction, the advantages and usefulness of MeSH were emphasized, and indexed searching was advocated.

Third, the students conducted a literature search for a specific theme in nursing. This bibliographic task was based on a preformulated question in Dutch (translated: "What is the effect of a multifactorial treatment (i.e., a combination of physiotherapy/ exercise/ medication, etc.) on the risk of falling in elderly living in long-term care facilities, such as nursing homes or homes for the aged?"). We assume that this question was clear to the participants, as it was formulated in their mother tongue. Moreover, we paraphrased the question and explained to the participants what they had to look for orally, and they were free to ask questions at any time during the test. In the posttest questionnaire (see below), we asked whether the search question was formulated in a clear and understandable way.

The participants were advised to use MeSH terms instead of free text and to combine several relevant MeSH terms with Boolean operators to construct a well-formulated query. They had 15 minutes to complete the literature search, which was subsequently assessed in several ways (see Evaluation Methods section).

[2] http://www.ncbi.nlm.nih.gov/mesh

Fourth, a posttest questionnaire was completed to see how the students experienced the test.

Fifth, the participants completed the vocabulary and reading parts of the DIALANG[3] diagnostic language test for English. This test has been internationally validated and was developed by more than 20 major European institutions with the support of the European Commission. It is based on the Common European Framework of Reference for Languages (CEFR)[4] and is available in 14 European languages, including English. The DIALANG language test allowed us to assess the participants' English reading and vocabulary skills on a 6-band scale (see Table 1) and to link the results to their performance on the literature search task.

2.3. Test groups

We recruited 31 undergraduate bachelor's students in the Nursing Department of University College Ghent and 40 master's level students at the Nursing and Midwifery Department of the University of Antwerp. Both institutions are located in Flanders, the Dutch-speaking part of Belgium. The same test was conducted in both institutions in several sessions from November 2008 to December 2009. In the first year of their training, all respondents had taken a compulsory course in which they were briefly initiated into the research domain and learned to search for and understand specialist literature. Additionally, the master's students had attended a program on scientific research in their master's degree training, which includes methodological principles of literature searching, among others in PubMed, and systematic review and analysis of literature. As the master's level students are more experienced searchers, they will be referred to as more experienced compared to the less experienced undergraduate students.

[3] http://www.dialang.org
[4] http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/levels.html

2.4. Development of the gold standard

The gold standard used for the evaluation of the search results was synthesized from the results of three types of searches: the students' searches, an expert search, and a related-citations search. To qualify for the gold standard, citations had to contain the four main elements of the search question, i.e., falls, elderly, long-term care, and multifactorial prevention. If one of the components was not present, the citation was not incorporated into the gold standard. The selection of these citations was done by a linguist in consultation with an expert (a medical doctor with professional expertise in bibliographic retrieval and instruction, and with domain expertise about geriatric pharmacology).

In accordance with the "union of outputs" principle (Miller, 1971), we filtered the relevant citations from the students' selections. This resulted in a set of 51 relevant citations.

In addition, the search task was executed by the expert, who formulated a gold standard query. This query covered all four concepts of the information need (except for the multifactorial aspect), and it consisted of six terms ("Accidental Falls/prevention and control"[Mesh] AND ("homes for the aged"[Mesh] OR "nursing homes"[Mesh]) AND ("aged"[Mesh] OR "Geriatrics"[Mesh])). The extra relevant articles yielded by this query—11 citations—were added to the students' selections.

The total set of relevant articles found by our test subjects and by the expert was expanded with citations retrieved with the "related citations" function in PubMed, as Lin and Smucker (2008) showed that tools based on content similarity can increase recall considerably. In our case, only four extra citations were found with this function.

This three-step procedure resulted in a gold standard of 66 articles in total. However, as the test was conducted in several sessions over a time span of 13 months, we had to take the publication date of the articles in our gold standard into consideration. The gold standard comprised 62, 64, 65, and 66 records for the test groups of November 2008, February 2009, April 2009, and December 2009, respectively. The gold standard query had a recall of 71.2% and was used to calculate concept coverage. The precision of the gold standard query was 17.4% (47 citations out of 270 were relevant).

2.5. Evaluation

2.5.1. Evaluation of the search process

We used the Morae[5] software, a program specifically designed to record and analyze user–computer interaction, for the evaluation of the search process. It registers all onscreen actions performed on the computer. In this way it allows researchers to analyze all operations executed by a user and to log tasks, markers, and marker scores. Tasks take up a period of time, whereas markers are used for events. We defined several tasks, including "Reading the search question," "Searching"—a task that usually consists of several individual PubMed searches—and "Final relevance judgment." One PubMed search includes a querying and a relevance judgment stage. The querying stage is characterized by an alternation of search term formulation and MeSH term selection.

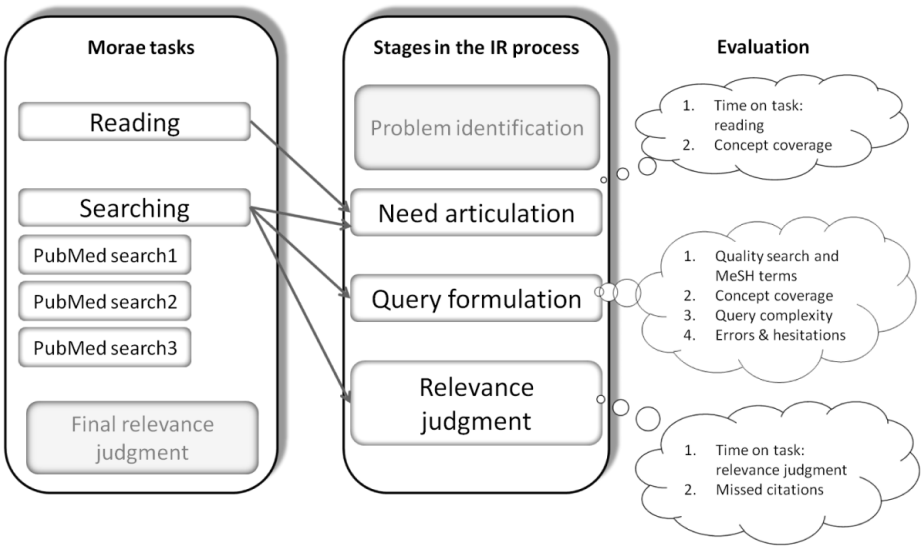


Figure 2: Evaluation of the search process

We also logged "hesitations and errors" as a task. It may be questionable to classify hesitations and errors as a task, but this was the only way to mark events that occurred over a period in time. Only those hesitations that were clearly caused by a lack of experience with the search system were logged, i.e., when it was obvious that the participant did not know what to do next, or when he or she made errors (e.g., going

[5] http://www.techsmith.com/morae.asp

back to the PowerPoint presentation about the use of PubMed, or searching for MeSH terms in PubMed instead of in the MeSH section).

Based on these Morae tasks, we evaluated the need articulation, query formulation, and relevance judgment stages (Figure 2). The problem identification stage was not addressed in this study, as the respondents started from an imposed search question. The need articulation stage as such is an implicit process. However, the result of this need articulation is reflected in the search terms used and in the number of concepts covered by queries. Need articulation was therefore studied in terms of concept coverage, in which we examined how many of the four main concepts (elderly, falls, long-term care, and prevention) were used in the queries. Concept coverage is an indication of how well the participants analyzed the search question and translated it into concepts. In this test, a good query was a query that – did not contain any errors and - contained the four main components of the search question, i.e. falls, elderly, long-term care, and prevention. These concepts or components can be expressed by several MeSH terms. Concept coverage is the proportion of those (four) concepts that were represented in the queries. The query "("Aged"[Mesh] OR "Frail Elderly"[Mesh]) AND "Accidental Falls"[Mesh]" for instance, has a coverage of 50% (two out of four concepts are covered: elderly and falls). The time spent on reading the search question is also considered as an indication of the time spent on need articulation. The query formulation stage was assessed in terms of the quality of search and MeSH terms, concept identification and coverage, query complexity, the use of Boolean operators, hesitations and errors, and zero-result queries.

The final stage of the IR process, relevance judgment, took place each time a PubMed search was executed. Relevance judgment is therefore seen as a part of the search task, following query formulation. The time spent on assessing the citations retrieved is considered as an indication of how thoroughly the relevance judgment process is executed. The effectiveness of this stage can be measured by precision (see Search Results section).

Next, we defined 26 different markers for different events in the search process, the most important of which were "Search term formulation," "MeSH term selection," "Query submission," and "Citation selection" (see Figure 3).

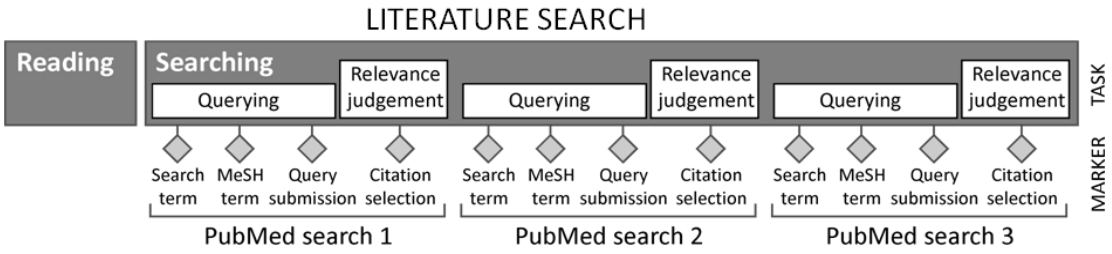


Figure 3: Tasks and markers for search process evaluation

Scores were assigned to the search term formulation and MeSH term selection events: each search term formulated and each MeSH term selected by the participants was assigned 0 (*bad*), 1 (*medium*), or 2 (*good*). These scores were the result of consultation between a linguist and our expert in bibliographic instruction. They were used to assess the quality of the search terms and MeSH terms (see Query formulation stage subsection in the Search Process Characteristics section). Bad search terms included incorrect translations, such as *kine*, *kinesitherapy,* and *kinestics* (instead of *physiotherapy*; translation of the Dutch word *kinesitherapie*), *movingexercises,* or *residention nursinghome*. Also considered as bad search terms were terms that were not relevant for this information search or too general to achieve relevant results (e.g., *resident* or *housesettings*).

Medium search terms included typographical errors (e.g., *physiotherapy progroms* or *resiential care)*. Spelling is a great source of errors too, even in native English users of PubMed (Wilbur et al., 2006). Examples of such orthographical errors from our data are *fysiotherapy* or *multifactoriel intervention*. Spelling and language skills in general are not an issue in the translation into MeSH terms, as the searcher has to select them from a list of suggestions. Bad MeSH terms are terms that are not relevant to the search question; examples include *kinesics* and *residential treatment*. Medium MeSH terms are terms that can be used in the context of the search question, but are not specific enough (e.g., *risk factors*, *hospitals*). A list of acceptable MeSH terms was created by a linguist in consultation with an expert (the same expert who constructed the gold standard query; see Development of the Gold Standard section).

2.5.2. Evaluation of the search results

We calculated the efficiency of the search in terms of recall and precision. Citations that were considered relevant were sent to the clipboard. The result was a list of citations the students deemed relevant to the search question, drawn from the whole search task, which usually consisted of several separate searches. These citations had to contain the four main components of the search question, i.e., elderly, long-term care facility, falls, and (multifactorial) prevention. All four components had to be present for the citation to be classified as relevant. For each participant, the resulting list of citations was compared to the gold standard, and precision and recall were deduced (see Figure 4). It may be noted that we did not intend to measure the performance of the search engine, but the participants' ability to find and select relevant citations in PubMed.

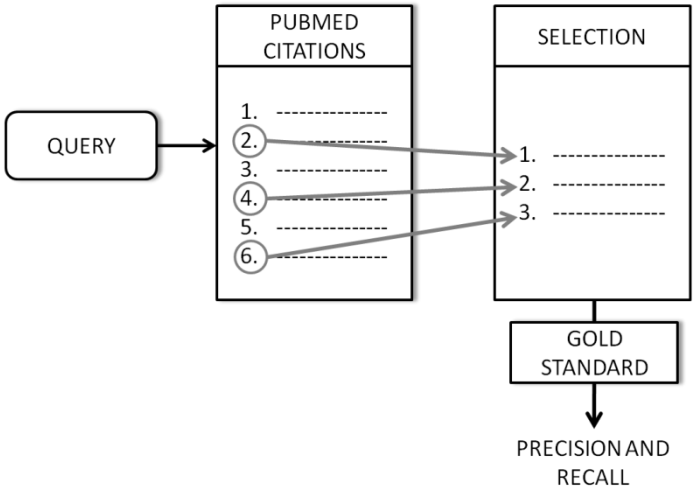


Figure 4: Precision and recall as defined in our analysis

The literature search task came down to a binary classification task, in which the test participants had to select relevant articles and discard the irrelevant ones from the list of citations their query yielded. Precision in our test case therefore referred to the precision of the selection (P_s) of citations made by the test participants. Citations selected by the participants that were also in the gold standard were true-positives (*tp*); false-positives (*fp*) were citations that were wrongly considered to be relevant. P_s can be defined as the proportion of true-positives in the students' selection:

$$P_s = \frac{tp}{tp + fp}$$

Analogously, recall in our test case referred to the recall of the final selection (R_s) of citations. It represented the proportion of citations in the gold standard that was also retrieved and selected (*tp*) by the test participants.

Recall of the students' selection was defined as follows:

$$R_s = \frac{tp}{GoldStandard}$$

We used NLM's E-Utilities[6] to simulate the students' searches to obtain their resulting lists of citations. Taking into account the number of results that were viewed by each participant for each query, we calculated the number of missed citations, i.e., the number of gold standard citations that were returned by a query, but were not selected as being relevant by the participants. This way, we could determine whether false-negatives were the result of a bad query or of bad relevance judgment. The number of false-negatives also allowed us to calculate the potential recall score (R_{pot}), i.e., the recall score the participants would have obtained if they had not overlooked any relevant citations:

$$R_{pot} = \frac{tp + fn}{GoldStandard}$$

The trade-off between recall and precision has been described by many researchers (Alvarez, 2002; Buckland & Gey, 1994; Eysenbach et al., 2001); it forces users to choose which performance measure to optimize. However, as this task did not focus on either one or the other of the two measures explicitly, we assumed that the participants wanted to keep a balance between precision and recall.

2.5.3. Pre- and posttest questionnaire

The students completed a pretest questionnaire that focused on self-perceived English-language and computer skills, and on facility with PubMed. The posttest questionnaire was designed to measure the students' self-perceived test performance. The answers to these questions will be linked to their actual performance on the test to see whether the

[6] http://www.ncbi.nlm.nih.gov/books/NBK25500/

participants had a realistic view of the quality of their search. The self-reported skills, attitudes, and opinions were assessed using 5-point or 7-point Likert scale questions.

2.5.4. Statistical Issues

We analyzed our data with the SPSS PASW 18 package. The Shapiro–Wilk test was used to assess the distribution of the variables. Depending on the types of variables studied, we used the Spearman correlation test, or the Mann–Whitney *U* (distribution-free) test. The minimum significance level used for these tests was 0.05.

For ranked values, we report the median and interquartile ranges as follows: *Mdn* (Q1, Q3; IQR)—median, first and third quartile, and interquartile range, respectively.

Precision of the user's selection is a relative notion: a respondent who selected only two citations, one of which was relevant, achieved a precision of 50%, which may misrepresent the efficiency of the search. We used Spearman's rank correlation to assess relationships between precision and recall, and other variables in the test.

2.5.5. Ethical Issues

We asked the Nursing Departments for formal permission to conduct the test. Students were invited to participate in the test by means of an invitation letter, in which we explained the aim and methods of the test. They were also informed that they could leave the classroom at any time if they no longer wanted to participate.

3. Results

3.1. Respondent characteristics

Seventy-one respondents participated in our test: 31 bachelor's and 40 master's level nursing students. The description of the respondent characteristics below is based on the pre- and posttest questionnaires, and on the results of the DIALANG language test.

3.1.1. Language skills

We assume that at least a B2 level is needed to perform this task successfully, as people with this level of language skills can read and produce more technical texts: 63.4% achieved a B2 level or higher for reading, and 83.1% of the participants reached a B2 level or higher for vocabulary.

Table 1: Results of the DIALANG test

		Participants (n=71)
Reading		
Level	**Corresponding skills**	
A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.	2.8%
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance.	11.3%
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.	22.5%
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.	45.1%
C1	Can understand a wide range of demanding, longer texts, and recognize implicit meaning.	12.7%
C2	Can understand with ease virtually everything heard or read.	5.6%
Vocabulary		
Level	**Corresponding skills**	
A1	Can introduce him/herself and others and can ask and answer questions about personal details.	0.0%
A2	Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.	7.0%
B1	Can produce simple connected text on topics which are familiar or of personal interest.	9.9%
B2	Can produce clear, detailed text on a wide range of subjects.	62.0%
C1	Can use language flexibly and effectively for social, academic and professional purposes.	18.3%
C2	Can express him/herself very fluently and precisely, differentiating finer shades of meaning even in more complex situations.	2.8%

3.1.2. *Self-reported skills*

We asked the students to rate their English-language skills and their computer skills on a scale from 1 (*very poor*) to 7 (*excellent*). Language skills were assigned quite a high score, with a median (*Mdn*) of 5 (4, 5; 1 IQR). With a median (*Mdn*) of 3 (3, 4; 1 IQR), computer skills were assigned lower scores. Although there are very useful biomedical databases, 24% of the students in our test group preferred using Google to look for medical information. More than half of the students indicated that they are used to searching for medical resources in English, as these are also written predominantly in English. However, there is a clear preference for Dutch over English (72%) to read scientific texts. We asked the participants whether the search question was clearly formulated and understandable. Only one student indicated that the search question was not entirely clear.

3.1.3. *Self-reported test performance*

When asked to assess their performance on the search task, 28% answered that they had made a good selection of citations. Sixty-three percent had difficulties finding the right keywords for their searches, and 62% of students were uncertain about the spelling of the search terms they used. After the literature search, most of the students (73%) were enthusiastic about PubMed and indicated that they would like to learn more about the search system.

3.2. Search process characteristics

3.2.1. *Query formulation stage*

- **Quality of search terms and MeSH terms**. On average, half of the search terms entered were good search terms, and 21% of the search terms were scored as bad because they either contained language errors or because they were irrelevant (see Figure 5, chart A). The remaining 29% were medium search terms.

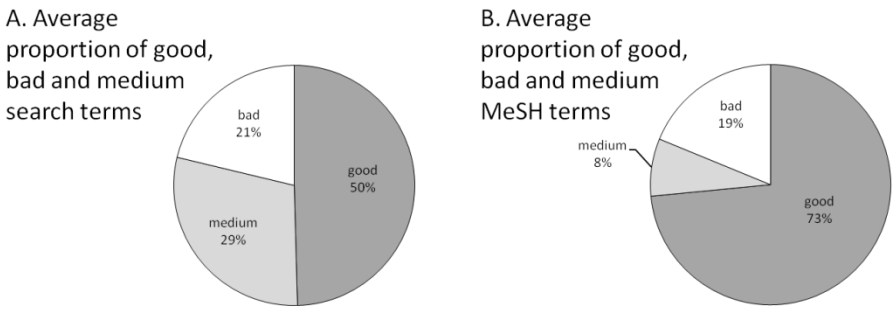


Figure 5: Average proportions of good, medium and bad search and MeSH terms

The translation of a search term into a MeSH term is usually an elimination process: one or more suggestions are provided by the search system, and the user selects the most suitable MeSH term for his information need. Consequently, this translation process is less error-prone than the formulation of free-text search terms (see Figure 5, chart B). This mainly results in a larger proportion of good MeSH terms (73%) and a smaller proportion of medium MeSH terms (8%).

About 50% of the search terms were linguistically incorrect or irrelevant and were therefore assessed as bad or medium, depending on the severity of the error. However, as this is only an intermediate step towards finding MeSH terms, many of those incorrect search terms are filtered by the search system. This corrective effect of subject searching with MeSH resulted in an error rate reduction of 25%. This means that the percentage of medium and bad search terms was reduced by half due to the use of MeSH terms. It should be noted, however, that MeSH terms which were not retrieved were not taken into account here.

- **Concept identification and coverage**. We assume that the participants understood the search question. Only one student—who achieved a relatively high precision and recall score—indicated in the posttest questionnaire that he or she did not completely understand the search question.

  As stated above, a good query has to contain MeSH terms for the four main components of the search, i.e., elderly, long-term care, falls, and (multifactorial) prevention. As there is no MeSH term for the concept "multifactorial," it could not be translated into a MeSH term. Table 2 shows

that the coverage of the concepts "falls" and "elderly" is quite high, and that about half of the participants found a MeSH term for "long-term care." The word "prevention" was not explicitly in the search question, causing many of the participants to overlook this concept.

Table 2: Gold standard concepts and their identification and coverage

Concepts	Concept identification	Concept coverage
elderly	94.37%	73.24%
falls	100%	88.73%
prevention	36.62%	23.94%
long-term care	77.46%	56.34%

To calculate concept coverage, i.e., the number of concepts that were covered by one or more MeSH terms in the participants' queries, we first analyzed the search terms to see which concepts were identified as important (see "Concept Identification" in Table 2). The search term *residention nursinghome*, for instance, which was scored as "bad," shows us that the participant did identify "long-term care" as an important component of the search. As no MeSH term suggestions were made for this search term—and the participant failed to formulate a correct search term—the concept was not covered in the participant's searches. Hence, the absence of a concept does not necessarily indicate that the participant did not identify this concept as important in the search question.

We found three different reasons for non-coverage of concepts. First, sometimes a concept was not identified as important to the search question, and no search terms were formulated for this concept. Consequently, it was not represented in the query. Second, even if a concept was identified as important, the use of an incorrect search term sometimes prevented the participants from finding the correct MeSH term. In other cases, a good search term was formulated, but the participant failed to identify the correct MeSH term.

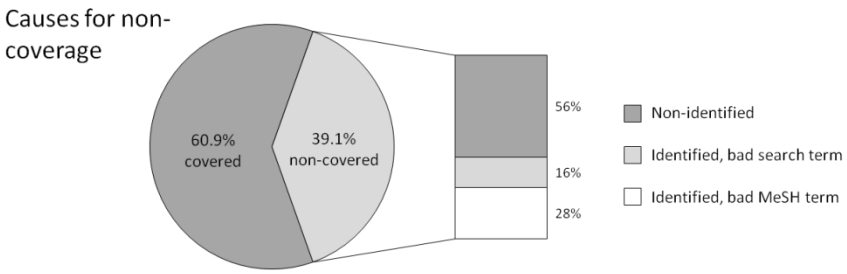


Figure 6: Causes for non-coverage of concepts

Figure 6 shows that 56% of non-covered concepts were absent in the queries because the participants did not identify these concepts in the search question, and therefore did not search for them.

For 16% of the non-covered concepts, the participants did identify the concept, but used a bad search term and consequently did not find an appropriate MeSH term. This category of errors is caused by the lack of active English-language skills. In 28% of the cases, a good search term was formulated, but the participant failed to identify the correct MeSH term.

We can conclude from this data that non-coverage of concepts is caused, in the first place, by the non-identification of concepts in the search question and that the number of bad search terms that lead to non-coverage is limited. Selecting the correct MeSH term seems to be a problem, even when a correct search term was entered. This may be due to the lack of experience with the search system, or to the lack of language skills.

- **Query complexity and the use of Boolean operators**. The average query in our test consisted of 3.36 terms. All test participants constructed queries by combining MeSH—or sometimes free-text search—terms with the Boolean operator AND. About 35% of the students used the OR-operator and none of them used the NOT-operator. The excessive use of the Boolean operator AND (e.g., *"Pharmaceutical Preparations"[Mesh] AND "Aged"[Mesh] AND "Risk Factors"[Mesh] AND "Accidental Falls"[Mesh] AND "Nursing Homes"[Mesh]) AND "Nursing"[Mesh]*) often led to zero results, and it was also found to be one of

the causes of "unproductive searches" by Walker et al. (1991) and Kingsland et al. (1993).

- **Zero-result queries**. A total of 17% of all queries yielded zero results. This is due to either overspecification and the excessive use of AND, or to the incorrect use of MeSH-terms.
- **Hesitations and errors**. We assigned the label "hesitations and errors" when erroneous steps were taken (e.g., searching for a MeSH term in PubMed instead of in the MeSH section), or when the participant clearly hesitated about the next step. Moments of inactivity before formulating a search term were not considered as hesitations. The average total length of hesitations and errors was 2 minutes 4 seconds. The time spent on hesitations and errors can be seen as an indication of search proficiency (see "Associations between respondent and search process characteristics" below).

3.2.2. Relevance judgment stage

- **Time spent on relevance judgment.** During the manual analysis of the screen recordings, we noticed that many participants selected citations too quickly. A combination of the words "elderly" and "falls" in the title was often enough to make them select the citation as relevant. Therefore, we consider the time spent on relevance judgment per search as an indication of how thoroughly this step was executed. The average total time spent on evaluation, i.e., on relevance judgment, during the whole search task was 5 minutes 11 seconds.
- **Selection of citations.** About 1 in 10 participants did not select any citations during the literature search task. On average, the participants selected 6.8 articles with a maximum of 31 and a median of 5 (2, 9; 7 IQR).

3.3. Search results

3.3.1. Number of relevant citations in the set of selected citations

The participants in our test selected 2.2 relevant —max = 13, *Mdn* = 1 (0, 3; 3 IQR)— and 4.6 irrelevant —max = 21, *Mdn* = 3 (1, 7; 6 IQR)— citations. Thirty-seven percent of the

test participants did not select any relevant citations, and consequently had a recall score of 0%. In half of those cases, the potential recall score was also zero. This means that these students' queries did not yield any relevant citations.

In total, 59% of the participants had higher potential than actual recall scores, which indicates that they overlooked relevant citations and hence could have achieved higher recall with the same queries. The average potential recall was 6.8%, almost double the average actual recall score.

3.3.2. Precision

On average, only one in three of the citations selected was relevant: the average precision score was 33.30%. Some students achieved 100% precision; however, as mentioned above (see Statistical Issues section), this may misrepresent the performance of these students.

3.3.3. Recall

The average recall score of the selections made by our test participants was 3.7%, and maximum recall was 20%.

3.4. Exploratory analysis

3.4.1. Associations among respondent characteristics

The students' self-assessment of their English-language skills was quite accurate: students with high scores on the reading and vocabulary tests rated their language skills higher in the pretest questionnaire (Table 3; items 1 and 2). Students with better computer skills used PubMed more often to search for medical information (Table 3; item 3), and those who had a positive perception of their retrieval results indicated that PubMed was a user-friendly search system (Table 3; item 4). Students with lower scores on the language test indicated that they had problems finding the right keywords for their searches, and that they were uncertain about the spelling of the English words (Table 3; items 5–8).

Table 3: Associations among and between respondent characteristics and search process characteristics

		Spearman's	Significance
Associations among respondent characteristics.			
1. Vocabulary test – self-assessment English language skills		r_s= .346	p= .003
2. Reading test - self-assessment English language skills		r_s= .400	p= .001
3. Self-reported computer skills – self-reported exposure to PubMed		r_s= .312	p= .008
4. Self-reported test performance – PubMed = user-friendly		r_s= .463	p= .000
5. Vocabulary test – problems finding right keywords		r_s= -.303	p= .010
6. Vocabulary test – spelling uncertainty		r_s= -.382	p= .001
7. Reading test - problems finding right keywords		r_s= -.394	p= .001
8. Reading test - spelling uncertainty		r_s= -.277	p= .019
	Mann-Whitney	**z**	**Significance**
9. Education level – self-reported language skills	U= 381.00	z= -2.923	p= .000
10. Education level – self-reported computer skills	U= 337.50	z= -3.646	p= .003
11. Education level – self-reported test performance	U= 439.50	z= -2.141	p= .032
Associations among search process characteristics.			
12. Quality of the first search term - Number of bad search terms		r_s= -.286	p= .016
13. Hesitations and errors - number of citations selected		r_s= -.336	p= .004
14. Time on task: reading – bad MeSH terms in "best" query		r_s= -.263	p= .026
Associations between respondent and search process characteristics.			
15. Self-reported exposure to PubMed – query complexity		r_s = .283	p= .017
16. Reading test – hesitations and errors		r_s = -.294	p= .013
17. Vocabulary test – hesitations and errors		r_s = -.252	p= .034
18. Reading test – proportion of good search terms		r_s = .236	p= .048
	Mann-Whitney	**z**	**Significance**
19. Education level – total querying time	U= 406.00	z= -2.481	p= .013
20. Education level – language errors in search terms	U= 432.50	z= -2.218	p= .027
21. Education level – hesitations and errors	U= 444.50	z= -2.049	p= .048

There were also some differences in respondent characteristics between the bachelor's and the master's students. In general, the master's students seemed to be more confident about their skills and performance on the test than the bachelor's students. The bachelor's students rated their language skills lower than the master's students did (see Table 3; item 9). The master's students were also more confident about their computer skills (see Table 3; item 10), and about their performance on the test (see Table 3; item 11).

The master's students used PubMed more often to search for medical information (see Figure 7), whereas most of the bachelor's nursing students rarely or never used this search engine. In summary, the main differences between the bachelor's and master's level students were related to their confidence in their own skills, which is a subjective assessment, and to their experience with PubMed, operationalized as exposure to PubMed and prior training in literature searching. Hence, the division into master's and bachelor's level students can be reduced to the division into more and less experienced PubMed users.

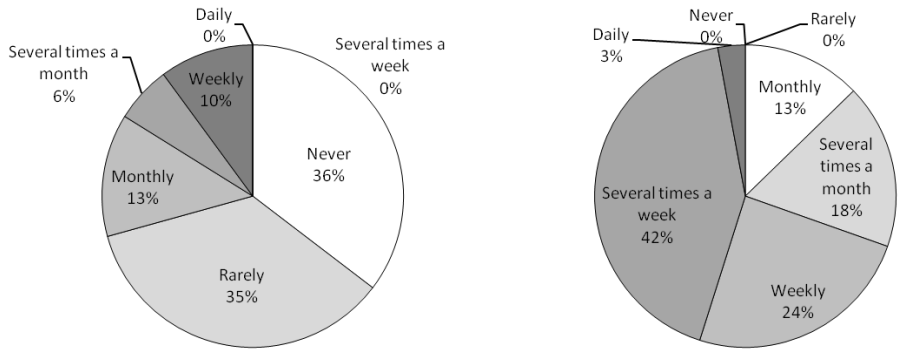


Figure 7: Self-reported exposure to PubMed

3.4.2. Associations among search process characteristics

When the quality of the first search term was low, the rest of the search terms were usually badly formulated as well (Table 3; item 12). This indicates that the effect of human learning (White, Marchionini, & Muresan, 2008) on query formulation was minimal in this test, probably due to the limited time. As can be expected, hesitations have a negative impact on the number of citations that were selected (Table 3; item 13).

The time the students spent on reading the search task was inversely correlated with the number of bad MeSH terms in their best query (Table 3; item 14), i.e., the query that covered the highest number of gold standard concepts. This indicates that a good understanding, interpretation, and articulation of the information need is crucial for the formulation of a good, comprehensive query.

3.4.3. Associations between respondent and search process characteristics

The average number of terms used per query, i.e., query complexity, was affected by PubMed experience (average number of terms: 3.7 vs. 2.9 in the less experienced group): frequent and more experienced users tended to formulate longer queries (Table 3; item 15). Although the construction of a query involves some translation processes, language skills did not seem to play a role in the coverage of gold standard concepts, nor did it influence the proportions of good, bad, and medium search and MeSH terms.

Also remarkable was the relation between language skills and hesitations and errors. Although hesitations in the query formulation stage were not annotated as hesitations and errors, we see that the lower the scores on the language tests are, the more the participants hesitated and made searching errors (Table 3; items 16-17). This might indicate that there were problems with the language of the interface. We also found a significant correlation between the scores on the reading test and the proportion of good search terms (Table 3, item 18).

The more experienced searchers in our test group spent less time on the construction of queries than the less experienced searchers (Table 3; item 19), and also produced less language errors in their search terms (Table 3; item 20). The more experienced searchers constructed queries with a smaller number of bad MeSH terms (16% as opposed to 22% in the less experienced group) in a shorter querying step, which confirms that they are more experienced in searching PubMed and therefore perform smoother searches. Their level of experience was also reflected in a difference in hesitations and errors (Table 3; item 21).

3.4.4. Associations between respondent characteristics and search results

The main aim of this study was to determine the effect of language skills on the efficiency of literature searches in PubMed. We therefore investigated the relationship between scores on the language tests and performance on the literature search. The test showed a significant relation between language skills — both vocabulary (Table 5; item 1) and reading (Table 5; item 2) — and recall. This means that participants with better English-language skills generally performed better on the literature search task. Our data did not show a significant correlation between language skills and relevance judgment, which can be measured by precision. Table 4, however, shows a trend: higher scores on the language test go together with a higher precision and therefore a better judgment of article relevance.

Table 4: Precision and recall per level of English language skills (n=71)

Reading level	Mean precision	Mean recall
A1	.0882	.0227
A2	.2791	.0125
B1	.3386	.0281
B2	.3226	.0410
C1	.4161	.0462
C2	.4357	.0698
Vocabulary level	**Mean precision**	**Mean recall**
A1	.	.
A2	.2818	.0322
B1	.3000	.0111
B2	.3286	.0337
C1	.3385	.0544
C2	.6350	.0873

Participants who indicated that they had difficulties finding the right keywords, and that they were uncertain about the spelling of the English words, achieved lower recall scores (Table 5; item 3 and 4).

Computer skills (Table 5; item 5) and self-reported exposure to PubMed (Table 5; item 6) did not affect efficiency in our test case.

In our posttest questionnaire, we asked the respondents for their opinion about their search process and about the selection of articles they had made. About 28% of the participants indicated that they were quite pleased with their results, although the maximum recall score was 20%. There is, however, a significant correlation between self-reported and actual performance scores (Table 5; item 7).

Table 5: Associations between respondent characteristics and search results (n=71)

Spearman correlations		
	Precision	**Recall**
1. Vocabulary test	r_s= .145 p= .229 (NS)	r_s= .236 p= .048
2. Reading test	r_s= .161 p= .180 (NS)	r_s= .259 p= .029
3. Difficulties finding the right keywords	r_s = -.167 p= .163 (NS)	r_s = -.353 p= .003
4. Spelling uncertainty	r_s = -.134 p= .266 (NS)	r_s = -.380 p= .001
5. Computer skills	r_s= - .154 p= .199 (NS)	r_s= - .092 p= .443 (NS)
6. Self-reported exposure to PubMed	r_s= .060 p= .619 (NS)	r_s= .118 p= .327 (NS)
7. Self-reported performance on search task	r_s= .540 p= .000	r_s= .551 p= .000
Mann-Whitney U Test		
8. Education level (bachelor/master)	U= 604,00 z= -.189 p= .850 (NS)	U= 540,00 z= -.944 p= .345 (NS)

There were some differences between the less and the more experienced searchers with regard to search results: a maximum of six relevant citations were selected in the less experienced group versus 13 in the other group.

The more experienced searchers achieved slightly higher recall (mean *M* = 4.42, *Mdn* = 2.31 (0, 7.61; 7.61 IQR)), than the less experienced students (*M* = 2.69, *Mdn* = 1.59 (0, 4.73; 4.73 IQR). Although this difference in recall is not significant, we do see that the highest recall scores were achieved by the more experienced searchers. The average precision

score in the less experienced group was slightly higher (*M* = 37.58, *Mdn* = 27 (0, 67; 67 IQR)), but not significantly (master's: *M* = 29.96; *Mdn* = 27 (0, 50; 50 IQR)). As this group selected a lower number of citations, it was easier to achieve high precision.

3.4.5. Associations between search process characteristics and search results

Our test participants were advised to use MeSH terms in their searches. The proportion of good (Table 6; item 3) or bad (Table 6; item 1) search terms did not have an influence on precision and recall. However, the selection of bad MeSH terms (Table 6; item 2) did prove to have a negative effect on performance scores and the selection of good MeSH terms resulted in better retrieval (Table 6; item 4).

Other factors that had an impact on retrieval were the number of corrections (Table 6; item 5), querying times (Table 6; item 6), and total evaluation times (Table 6; item 7). Precision and recall decreased with an increasing number of corrections, which might indicate that these participants had problems finding the right keywords. The total time spent on query formulation is inversely correlated with precision and recall. This means that participants who needed more time to formulate their queries selected a smaller number of relevant citations. Long querying times can either indicate that the formulation of the query was done with great consideration, or that the participant hesitated. The second explanation seems more plausible, as precision and recall go down with increasing querying times. This is corroborated by our data, which show positive correlation between hesitations and errors and querying times (R_s = .412; p = .000). The time spent on relevance judgment, on the other hand, was positively correlated with recall. This indicates that a thorough relevance judgment step is crucial for successful retrieval.

Queries covering the four concepts (elderly, falls, long-term care, and prevention) resulted in better recall, but not necessarily in higher precision (Table 6; item 8). This underlines the importance of good relevance judgment: a good query might yield a large number of relevant results, but it is then up to the searcher to make a good selection. The selection of a higher number of citations (Table 6; item 9) resulted in higher recall, which seems logical. However, it also resulted in higher precision, which contradicts the classical trade-off between precision and recall.

Table 6: Associations between search process characteristics and search results (n=71)

	Precision	Recall
Spearman correlation		
1. Proportion of bad search terms	r_s= -.051 p= .675 (NS)	r_s= -.129 p= .284 (NS)
2. Proportion of bad MeSH terms	r_s= -.252 p= .034	r_s= -.302 p= .011
3. Proportion of good search terms	r_s= -.040 p= .738 (NS)	r_s= .036 p= .767 (NS)
4. Proportion of good MeSH terms	r_s= .307 p= .009	r_s= .333 p= .005
5. Corrections	r_s= -.333 p= .005	r_s= -.389 p= .001
6. Querying times	r_s= -.278 p= .019	r_s= -.432 p= .000
7. Total evaluation times	r_s= .127 p= .290 (NS)	r_s= .391 p= .001
8. Concept coverage	r_s= .213 p= .074 (NS)	r_s= .236 p= .048
9. Number of citations selected	r_s= .274 p= .021	r_s= .671 p= .000

Although we did not find a significant correlation between query complexity and search performance, we did see a peak in precision and recall at four to six terms per query (see Figure 8).

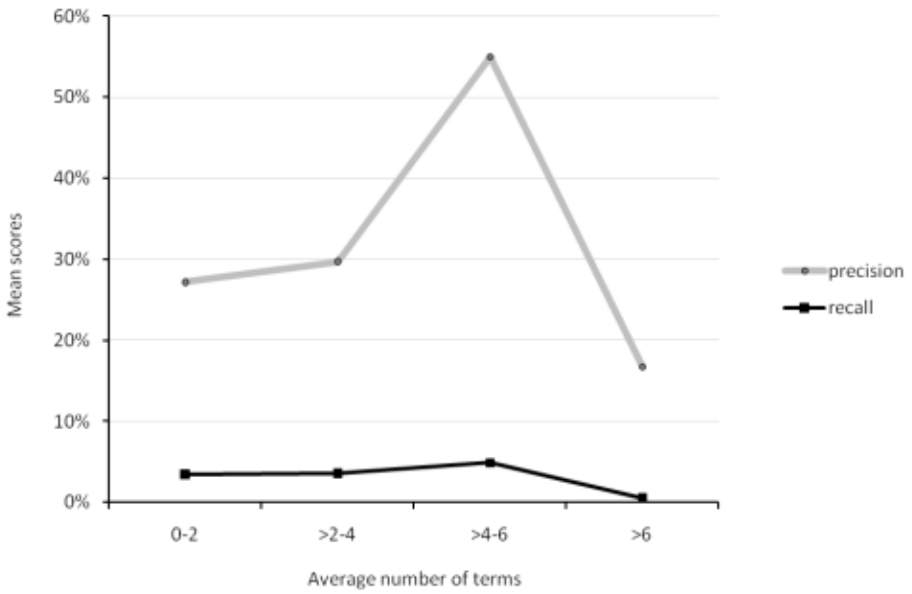


Figure 8: Impact of query complexity on search efficiency

The use of more than six search terms in a query caused a steep drop in these scores, and less than four search terms yielded moderately lower scores as well. It seems logical that the optimal query for this search question contains four terms for the four components of the search to be represented. Some concepts can be translated into a combination of terms, which explains the fact that a query containing more than four search terms can also be successful. Overspecification, i.e., more than six terms, may lead to empty result sets. The ideal query for this task would therefore consist of four to six search terms. In accordance with these findings, our gold standard query consisted of six MeSH terms.

4. Discussion

From a methodological point of view, the main strength of this study is that direct observation using the Morae software allowed us to collect both quantitative and qualitative data without interfering in the IR process or affecting the search results.

This study distinguishes itself from previous work in the field in that it not only analyzes the query formulation process and the resulting citations, but also two very important human interaction steps: need articulation and relevance judgment.

4.1. Main findings

Precision and recall were quite low in the whole test group. The highest recall scores were achieved by master's students, whose searching skills were also reflected in smoother searches with fewer hesitations.

English-language skills were crucial in this cross-language literature searching task: recall correlated positively with reading and vocabulary skills, and there was a positive trend in precision scores with increasing language skills.

The English MeSH terms had a corrective effect when compared to free-text searching and can therefore be considered as a very useful search aid also for nonnative speakers of English.

It is self-evident that high concept coverage, i.e., the number of concepts from the information need that are actually translated into MeSH terms and combined into a query, is a prerequisite for a good query. There are several reasons for non-coverage of concepts: the main cause was the non-identification of concepts in the search question. It is therefore very important that searchers know exactly what they are searching for before they start formulating queries. Other causes were the use of bad search terms, and failure to identify a good MeSH term, even with good search terms.

4.2. Limitations

A limitation of this study was the relatively short period in which the students had to complete the literature search task. However, as the same amount of time was allowed to all participants, we were able to make a valid comparison. Moreover, finding relevant information in a relatively short period can be important in real-life clinical situations.

According to Wendt (1969) and Jacobson and Fusani (1992), the importance of the information need and the motivation of the users in a test case affect the effort made and the results obtained in the search task. In our study, problem identification was admittedly based on a preformulated question rather than on a spontaneous information need, but as this was true to the same extent for all participants, differences in motivation were unlikely to have a major falsifying influence.

We consider high concept coverage as the result of a well-thought-out articulation of the information need combined with the formulation of linguistically correct search terms, but it can probably also be linked to levels of intelligence. This, however, was not studied in this test.

We acknowledge that, in correlating evaluation times with recall, we did not take into account other sources of difficulties, such as poorly written abstracts, problems understanding the texts in English, etc. However, as we noticed that many participants decided too quickly that citations were relevant, and as there was a strong – negative - correlation between evaluation times and recall, we are convinced that a longer and more thorough evaluation step is crucial to a successful search.

4.3. Critical remarks on main findings

4.3.1. The role of search engine experience

Several studies (Aula, 2003; Bernstam et al., 2001; Haynes et al., 1990; Lazonder et al., 2000) conclude that experienced users obtain better results in literature or information search tasks. Fenichel (1981), on the contrary, found that there are only very small differences in the performance of users with different system experience. The more experienced searchers did not perform significantly better on the literature search task. However, we do see that the top 10 recall scores were achieved by these students. Rather than concluding that search engine experience does not have an impact on the efficiency of PubMed searches, we can say that the distinction between the two test groups does not correspond to the distinction between experts and novices made in the aforementioned literature. In other words, the bachelor's students may be designated as novice users, as most of them have no experience with PubMed, but the master's students are not experienced enough to be considered as experts.

4.3.2. Search results

The search results in terms of precision and recall are quite low. This can probably partly be attributed to the limited time in which the participants had to complete the literature search task. An experienced user with a spontaneous, specific information need would try to formulate a query that is as efficient and as comprehensive as

possible. In this artificial situation, users who sometimes had little experience with the search system had to find very specific information in only 15 minutes. This, together with their limited searching skills, resulted in a rather chaotic query formulation stage, mostly based on trial-and-error methods.

Taking into account the time limitation, we considered a search with a yield of five relevant citations or more as a very successful search. This list of citations could then be expanded using the related-citations tool. This cutoff was achieved by 3% of the less experienced and by 28% of the more experienced searchers.

One in five participants had zero potential recall, which means that they did not submit any queries that yielded relevant results. Almost two out of three students had higher potential than actual recall, which means that they overlooked relevant citations and that they could have achieved higher recall with the same queries.

Mouillet (1999) concluded that the MEDLINE/Ovid users in her test group did not have a realistic view of their search results. They seemed to be quite satisfied with their retrieval, despite the fact that "their MEDLINE/Ovid utilization was often irrelevant". As some students in our test reported that they were quite pleased with their results, whereas the maximum recall score was 20%, we could conclude that these students, too, have an unrealistic view of their performance. However, we found a positive correlation between user satisfaction and actual performance, expressed in recall and precision, indicating that the better performing students were more enthusiastic about their results than those who had lower scores.

4.3.3. Search process

Our test participants were asked to use MeSH terms to construct their queries. This implies that they first entered free-text search terms and then selected one or more MeSH terms from the list of suggestions made by PubMed. Our data showed that the quality of the free text search terms does not have an impact on precision and recall. This is not surprising because the actual queries were constructed with MeSH terms and not with free text. Whenever a test participant entered a bad search term (e.g., *kinestherapy* for *physical therapy*), a warning message appeared: "The following term was not found in MeSH: kinestherapy. See Details. No items found." In other cases, the MeSH

terms suggested for the search term were not suitable for the search question (e.g., the search term *multifactorial* yielded the MeSH terms *Multifactorial Inheritance*, *Causality*, *Nephrogenic Fibrosing Dermopathy*, *Typhlitis*, etc.). In many cases, a new—and usually better— search term was then formulated, and there was no impact on the search results. However, these bad search terms are a cause for non-coverage of concepts, which leads to broader and less precise queries. Other reasons for non-coverage were non-identification and failure to select the correct MeSH term.

The use of MeSH terms, although only available in English in the PubMed search interface, reduced the number of medium and bad keywords in the queries by half. This indicates that the MeSH terms are a useful search aid, compensating for badly formulated search terms. However, the use of MeSH terms can also be a stumbling block: in more than two out of five cases, participants failed to select a good MeSH term. We assume that the possibility to search in one's mother tongue might lead to an increase in concept coverage, and consequently also in recall.

4.3.4. Self-reported skills and their effect on search process and results

We investigated the relationship between general computer skills, on the one hand, and query complexity, the quality of search terms, and precision and recall scores on the other. Aula (2003) argues that general computer skills affect the query formulation process. However, we did not find a relation between the self-reported level of computer skills and the quality of the search terms, nor did the subjects' computer skills affect precision scores. Aula also observed that more experienced Web and computer users tend to formulate longer, more specific queries. Students in our test case who estimated their computer skills higher, however, did not formulate longer queries.

As opposed to general Web and computer skills, exposure to the search engine PubMed did prove to have an impact on query complexity. This is in accordance with Sutcliffe et al. (2000), who found that searchers with more MEDLINE experience use more complex queries when compared to novices, who keep their queries rather simple.

Facility with the search engine is also reflected in the participants' pause behavior: participants who were more familiar with the search system paused less during their literature search. This is in accordance with Huang's findings (2003).

According to Herskovic et al. (2007) and Lin and Smucker (2008), between 16 and 20% of all queries submitted to PubMed yield zero results. We found similar results in our data. Our data showed that zero results can be due to many factors, including badly formulated terms or the selection of incorrect MeSH terms, inexperience with the search system, or the formulation of queries that are too narrow or complex.

Several studies (Sewell & Teitelbaum, 1986; Sutcliffe et al., 2000; Vakkari et al., 2003) have shown that more experienced searchers tend to use more advanced Boolean operators, as opposed to novices who mostly use the AND operator. This, however, is not corroborated by our data, probably because the master's students had not reached this level of expertise yet.

4.3.5. Language skills and search results

Higher scores on the DIALANG language test, and therefore better language skills, resulted in higher precision and recall. We assumed that the language barrier would play a crucial role in the stage where active language skills are needed, i.e., the query formulation stage. However, there was no significant correlation between language skills and query formulation in terms of proportions of good, bad and medium and MeSH terms. There was, however, a significant correlation between the scores on the reading test and the proportion of good search terms. Participants with lower scores on the language test indicated that they had problems finding the right keywords and that they hesitated about the spelling of the English words.

So in which stage do these language skills come to play such an important role that they entail higher performance scores? Or, in which stage does the language barrier hamper efficient searching? We already mentioned that nonnative English users of PubMed might have difficulties with the interface. Moreover, participants with better scores on the reading test selected a higher number of relevant citations, which means that language skills play an important role in relevance judgment. The importance of language skills in the relevance judgment stage is also emphasized by Mouillet (1999). She compared the answers of self-trained and librarian-mediated users of MEDLINE/Ovid and Pascal (a French bibliographic database) users to a survey in which she focused on the impact of the language barrier on the understanding of the

MEDLINE/Ovid interface. Although her test did not simulate an information need and a resulting information search, she did conclude that the English language barrier is especially reflected in the erroneous selection of articles.

5. Conclusions

We conducted an experiment to analyze the search behavior of Dutch-speaking nursing students and the efficiency of their literature searches in PubMed, focusing on query formulation and relevance judgment. We found that searching for information about a given topic within a limited time span is a complex and difficult task, the outcome of which is influenced by many factors.

English-language skills proved to have an impact on the efficiency scores: students with higher scores on the language test also performed better on the literature search task. Especially the relevance judgment stage benefits from better language skills: students with better knowledge of the English language were better at detecting highly relevant articles and thus had higher precision and recall scores.

From our test data we cannot conclude that search engine experience has an impact on search efficiency. However, the top recall scores were achieved by the more experienced searchers. Moreover, as they were more familiar with the search system, they hesitated less during the search process and spent less time on querying. Although there was no significant difference in language skills, the more experienced searchers formulated a smaller number of incorrect search terms. In summary, we can state that the students who were more familiar with the search system performed relatively smooth searches, apparently experiencing fewer hitches than less experienced searchers. An analysis of concept coverage showed us that good need articulation, although implicit in this research, is crucial, as higher concept coverage led to higher efficiency scores. The importance of a good interpretation and articulation of the information need, together with good relevance judgment, is underlined by our findings. The translation of an information need into concepts and from concepts into MeSH terms should therefore be an important part in bibliographic instruction, next to the actual use of search engines.

The medical subject headings proved to be a useful language aid, as they compensated for bad search terms. Conversely, the selection of erroneous MeSH terms resulted in an unproductive query. The Medical Subject Headings can therefore be very helpful, but they can easily become a stumbling block when used incorrectly.

In conclusion, the main factors influencing the efficiency of a biomedical literature search in PubMed across language boundaries are language skills, facility with the search engine, a good parsing of the information need into concepts, a careful selection of MeSH terms, and an in-depth evaluation of the relevance of the articles retrieved.

6. Future work

We realize that the current subject matter is quite comprehensive; therefore, not every aspect could be studied. We would like to set up several studies in which we will analyze the query formulation step in more detail. We could, for instance have students construct a query in Dutch, which would allow us to study concept identification. Second, we would like to study the Dutch–English translation step by having students translate a good query from Dutch into English. Another interesting task would be to have the students search for good MeSH terms for a given query, formulated in English. To analyze the relevance judgment step, we would like to give a test group a list of citations from which they have to select the relevant ones.

In addition, a think-aloud protocol study would be interesting to reveal the steps between concept identification and concept coverage.

Acknowledgments

We express our gratitude to Professor Dr. Monique Elseviers and to Liesbeth Van Heck, who helped us organize the test in the Nursing Departments of the University of Antwerp and University College Ghent, respectively. We would also like to thank Marc De Spiegelaere and Annelies Verhaeghe, who offered great help for the statistical analysis of our test data. Thanks are due to Joke Coussement of the Centre for Health Services and Nursing Research, KU Leuven, who gave helpful hints in formulating the search question.

References

Alvarez, S. (2002). An exact analytical relation among recall, precision, and classification accuracy in information retrieval (Vol. Technical Report BCCS-02-01). Boston: Boston College.

Aula, A. (2003). *Query Formulation in Web Information Search.* Paper presented at the IADIS International Conference WWW/Internet Algarve, Portugal.

Bernstam, E, Kamvar, S , Meric, F, Dugan, J, Chizek, S , Stave, C, Troyanskaya, O, & Fagan, L. (2001). *Oncology patient interface to Medline.* Paper presented at the American Society of Clinical Oncology Annual Meeting.

Bin, Li, & Lun, K.C. (2001). The retrieval effectiveness of medical information on the web. *Int J Med Inform, 62*(2-3), 155-163.

Buckland, Michael, & Gey, Fredric. (1994). The relationship between recall and precision. *J Am Soc Inf Sci, 45*(1), 12-19.

Eisenberg, A. (1996). Using English as the international language of science. In D. C. Andrews (Ed.), *International Dimensions of Technical Communication* (pp. 1-4). Arlington: Society for Technical Communication.

Eysenbach, G., Tuische, J., & Diepgen, T. L. (2001). Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. *Med Inform Internet Med, 26*(3), 203-218.

Fenichel, Carol Hansen. (1981). Online Searching: Measures That Discriminate among Users with Different Types of Experiences. *J Am Soc Inf Sci, 32*(1), 23-32.

Haynes, R. B., McKibbon, K. A., Walker, C. J., Ryan, N., Fitzgerald, D., & Ramsden, M. F. (1990). Online access to MEDLINE in clinical settings. A study of use and usefulness. *Ann Intern Med, 112*(1), 78-84.

Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: analysis of a typical day's query log. *J Am Med Inform Assoc, 14*(2), 212-220.

Huang, Mu-hsuan. (2003). Pausal behavior of end-users in online searching. *Information Processing and Management, 39*(3), 425-444. doi: http://dx.doi.org/10.1016/S0306-4573(02)00040-7

Jacobson, T., & Fusani, D. (1992). Computer, system, and subject knowledge in novice searching of a full-text, multifile database. *Library & Information Science Research, 14*(1), 97-106.

Kingsland, L. C., Harbourt, A. M., Syed, E. J., & Schuyler, P. L. (1993). Coach: applying UMLS knowledge sources in an expert searcher environment. *Bull Med Libr Assoc., 81*(2), 178-183.

Lankamp, Robert Eduard (1989). *A Study on the Effect of Terminology on L2 Reading Comprehension. Should Specialist Terms in Medical Texts be avoided?* (PhD doctoral thesis), Eindhoven University of Technology, Eindhoven.

Lazonder, Ard W., Biemans, Harm J. A., & Wopereis, Iwan G. J. H. (2000). Differences between novice and experienced users in searching information on the World Wide Web. *J Am Soc Inf Sci, 51*(6), 576-581. doi: http://dx.doi.org/10.1002/(SICI)1097-4571(2000)51:6<576::AID-ASI9>3.0.CO;2-7

Lin, Jimmy, & Smucker, Mark D. (2008). *How do users find things with PubMed?: towards automatic utility evaluation with user simulations*. Paper presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore.

Miller, William L. (1971). The extension of users' literature awareness as a measure of retrieval performance, and its application to Medlars. *Journal of Documentation, 27*(2), 125-135.

Mouillet, E. (1999). Language barriers and bibliographic retrieval effectiveness: use of MEDLINE by French-speaking end users. *Bull Med Libr Assoc, 87*(4), 451-455.

Muin, M., Fontelo, P., Liu, F., & Ackerman, M. (2005). SLIM: an alternative Web interface for MEDLINE/PubMed searches-a preliminary study. *BMC Med Inform Decis Mak 5*(1), 37.

Sewell, Winifred, & Teitelbaum, Sandra. (1986). Observations of end-user online searching behavior over eleven years. *J Am Soc Inf Sci, 37*(4), 234-245.

Spink, Amanda H., Wolfram, Dietmar, Jansen, Bernard J., & Saracevic, Tefko. (2001). Searching the web : the public and their queries. *J Am Soc Inf Sci, 52*(3), 226-234.

Sutcliffe, A., & Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers, 10*(1998), 321-351.

Sutcliffe, A., Ennis, M., & Watkinson, S. J. (2000). Empirical studies of end-user information searching. *J Am Soc Inf Sci, 51*(13), 1211-1231. doi: http://dx.doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1033>3.0.CO;2-5

Vakkari, Pertti, Pennanen, Mikko, & Serola, Sami. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management, 39*(3), 445-463.

Walker, C. J., McKibbon, K. A., Haynes, R. B., & Ramsden, M. F. (1991). Problems encountered by clinical end users of MEDLINE and GRATEFUL MED. *Bull Med Libr Assoc, 79*(1), 67-69.

Wendt, Dirk. (1969). Value of information for decisions. *Journal of Mathematical Psychology, 6*(3), 430-443.

Wilbur, W. John, Kim, Won, & Xie, Natalie. (2006). Spelling correction in the PubMed search engine. *Inf. Retr., 9*(5), 543-564. doi: http://dx.doi.org/10.1007/s10791-006-9002-8

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation, 55*, 249-270.

List of figures

List of tables

Chapter III: Lost in PubMed. Factors influencing the success of medical information retrieval

Abstract

With the explosion of information available on the Web, finding specific medical information in an efficient way has become a considerable challenge. PubMed/MEDLINE offers an alternative to free-text searching on the web, allowing searchers to do a keyword-based search using Medical Subject Headings. However, finding relevant information within a limited time frame remains a difficult task. The current study is based on an error analysis of data from a retrieval experiment conducted at the nursing departments of two Belgian universities and a British university. We identified the main difficulties in query formulation and relevance judgment and compared the profiles of the best and worst performers in the test.

For the analysis, a query collection was built from the queries submitted by our test participants. The queries in this collection are all aimed at finding the same specific information in PubMed, which allowed us to identify what exactly went wrong in the query formulation step. Another crucial aspect for efficient information retrieval is relevance judgment. Differences between potential and actual recall of each query offered indications of the extent to which participants overlooked relevant citations.

The test participants were divided into "worst", "average" and "best" performers based on the number of relevant citations they selected: zero, one or two and three or more, respectively. We tried to find out what the differences in background and in search behavior were between these three groups.

Highlights ► Categorization of errors in queries submitted during an IR experiment in PubMed. ► Identification of the factors that have a direct

impact on query quality. ► Analysis of the characteristics of the best and worst performers. ► Language skills play an important role in non-native English searchers. ► MeSH terms compensate for limited language skills in non-native speakers of English.

Keywords: Medical information retrieval; Medical Subject Headings; Bibliographic instruction; Nursing education; Information seeking behavior

1. Introduction

Several studies have been devoted to possible causes for search failure in information retrieval (Hofstede et al., 1996; McCray & Tse, 2003; Sutcliffe, 2000), trying to find out why some information searches do not yield satisfactory results. The aim of the present study is to contribute to the understanding of the reasons for failure in bibliographic searches executed by – relatively – untrained PubMed users. This should help us to formulate educational objectives in bibliographic instruction and to draw a profile of the better-performing searchers and compare it to that of the worst-performing searchers. As (Sutcliffe, 2000) claims, training the searchers is sometimes the only remedial action.

The present study focuses on the use of PubMed, an online system to access journal citations and abstracts in MEDLINE. PubMed was developed by the National Center for Biotechnology Information (NCBI) and daily provides hundreds of thousands of users with bibliographic information from the life sciences. It is a global resource of US origin; nevertheless many of its users are non-native speakers of English, which makes efficient retrieval an even more challenging task. Although the recommendation that only MeSH terms should be used is a matter of discussion (Jenuwine & Floyd, 2004), the use of these terms can enhance PubMed searches considerably (Richter & Austin, 2012) – provided that the user understands how search terms map to MeSH terms and how PubMed's search engine works in general. Poor understanding of MeSH is an issue that exceeds the problem of the language barrier: native speakers of English may also experience difficulties in formulating a good query with MeSH terms. Controlled vocabularies can

therefore enhance information retrieval, but they can also be a barrier to finding relevant information in a time- and cost-efficient way.

In this study, we want to do an error analysis of the queries that were submitted by our test participants, focusing mainly on quality in terms of the MeSH terms they contain, and on the differences between their potential and actual recall. Based on an error analysis, we try to formulate advice on how to address retrieval problems. Some searchers succeed in finding relevant results more easily than others. We draw a profile of efficient searchers versus those who have more difficulty in finding relevant citations by comparing their characteristics and search strategies.

We will discuss the methods used in this study in part two. The results section of this paper consists of two main parts: query error analysis, and secondly, a comparison of the best, average and worst performers. In the third part we will discuss some of our main findings, and finally, we will present our conclusions and future work in parts four and five.

2. Methods

2.1. Recruitment and test setup

We conducted a test at the nursing departments of two Flemish universities and one British university. A total of 100 respondents with different educational and linguistic backgrounds participated in the test: 31 Dutch-speaking and 8 native English-speaking bachelor's students, 40 Dutch-speaking and 21 native English master's students.

Prior to the actual retrieval test, the participants completed a pretest questionnaire, which allowed us to capture the participants' search experience and – for the Dutch-speaking respondents – their self-reported English language skills.

After a short introduction into searching PubMed with the use of MeSH terms, they conducted a literature search for a given subject. The participants in our test were stimulated to use MeSH terms, so their query formulation process consisted of several steps: first, they had to find relevant MeSH terms for each of the components of the search question (falls, elderly, long-term care and prevention). In order to find these

MeSH terms, they had to go to the MeSH module in PubMed and enter a free-text search term. Subsequently, PubMed made one or more suggestions for MeSH terms, from which the participants had to select the relevant ones and send them to the search box. This action was repeated until a satisfactory query was obtained. For example, most test participants entered the search term “fall” or “falls” in the MeSH module and then selected the MeSH term “Accidental Falls”. Once they had found the right MeSH terms for the other components of the search question and submitted their queries to PubMed, a list of citations was returned by the search engine. From this list, they had to select only those citations that were relevant to all aspects of the search question. The students were given 15 minutes to complete the search. All individual sessions were recorded with the Morae software, enabling us to time the subtasks and to reconstruct the queries.

After the experimental task, the participants completed a posttest questionnaire which measured their satisfaction with the search results and with the search system. Additionally, all participants completed an English language test, which enabled us to measure their language skills.

2.2. Query collection and error analysis

We collected all the queries submitted during the literature search task. This resulted in a total of 309 queries, issued by 98 participants – two participants did not submit any queries. The number of queries per participant ranges between one and ten, with a median of three.

For each of the queries in our collection, we determined which errors they contained; this allowed us to make a classification of different error types. Queries that contained no errors and covered the information need were labeled as “good queries”.

On the basis of these findings, we will try to make suggestions for the improvement of bibliographic instruction.

2.3. Performance

We developed a gold standard, consisting of 62–66 citations, depending on the moment of the test session (for more information see (Vanopstal et al., 2012). The students' selections were compared against this gold standard in order to calculate recall.

We are especially interested in the students' search strategies and in their relevance judgment, which is reflected in the selection of citations they considered as relevant. We will not report on the typical performance metrics in information retrieval, i.e. proportional recall scores expressed in percentages, but instead we will discuss performance in terms of absolute recall (R_{abs}), i.e. the number of relevant citations selected by the test participants as relevant to the information need.

We consider three relevant citations a good threshold to designate a search as successful, especially in the limited time frame of this test. Three relevant citations is a good starting point for exploratory work using the "related citations" function of PubMed, and it should provide the searcher with a relevant introduction to the research field. Based on this absolute recall, we will subdivide our test group into a "worst" (no citations), "average" (one or two citations) and "best" (three or more citations) performer group (see Section 2.4).

Next to absolute recall, we also will calculate the number of missed citations per query and per participant. Missed citations are relevant citations that were returned by the queries, but were not selected as being relevant. Using NLM's E-Utilities[1], we simulated the students' searches to obtain their resulting lists of citations. Per search, we registered the number of result pages that were viewed. Each page contained 20 citations, so a participant who looked at two result pages, is considered to have viewed 40 citations.

We compared each result list, i.e. only the pages that were actually viewed, to the gold standard. This allowed us to calculate – absolute – potential recall (R_{pot}), the recall the participants would have obtained had they not overlooked any relevant citations.

[1] http://www.ncbi.nlm.nih.gov/books/NBK25500/

Potential recall is the "raw" recall of the query itself, without any intervention or selection by the searcher.

$$R_{pot} = \#\ relevant\ but\ missed\ citations + R_{abs}$$

For instance, if the participant only looked at the first page (with 20 results per page), and there were two relevant citations in that page, the potential recall of that query was two.

2.4. Comparison of the performer types

We will analyze the differences between the worst, average and best performers in our test. This categorization is based on absolute recall. Participants in the worst performer group did submit one or more queries, but did not select any relevant citations. The "average performers" selected one or two relevant citations, and the "best performers" selected three or more.

All comparisons between the performer types were tested using the ANOVA test for variables with normal distribution. The other variables were tested using the nonparametric Kruskal–Wallis test with pairwise comparison and Bonferroni correction. All statistical tests were performed with IBM SPSS Statistics 20.

2.4.1. Search process

We consider the number of queries as an indication of the fluency of the search process. Participants who submitted ten different queries obviously had more problems finding the information they needed than those who submitted only one or two queries.

Other indicators for the fluency of the search process are querying and relevance judgment times. As described in Vanopstal et al. (2012), the querying step is "an alternation of search term formulation and MeSH term selection". It results in the construction of a query and ends when the user submits the query to the search engine. The total querying time is the sum of the querying times that precede each submission of a query.

Total relevance judgment time is the time spent on assessing the lists of citations returned by PubMed after the submission of each query.

2.4.2. Quality-based assessment of queries

In this part of the study we try to find out whether any of the performer groups make a higher or lower number of errors of a specific type. We will analyze three error types: incorrect MeSH term, underspecification, and the incorrect use of Boolean operators.

2.4.3. Outcome-based query analysis

Queries can be labeled as "good" or "bad" based on the number of errors they contain, but another way to classify them is based on their potential recall (see Figure 1: "adequate" versus "inadequate" queries). In this categorization, good or adequate queries yield at least one relevant citation, whereas bad or inadequate queries either lead to an empty result set, or to a list of citations that are not relevant to the information need.

Besides the ability to formulate an adequate query, the participants therefore needed the ability to distinguish relevant from irrelevant citations. We can subdivide the category of adequate queries into queries that led to the selection of relevant citations ("good relevance judgment") and queries that did not ("relevance judgment errors"; see Figure 1).

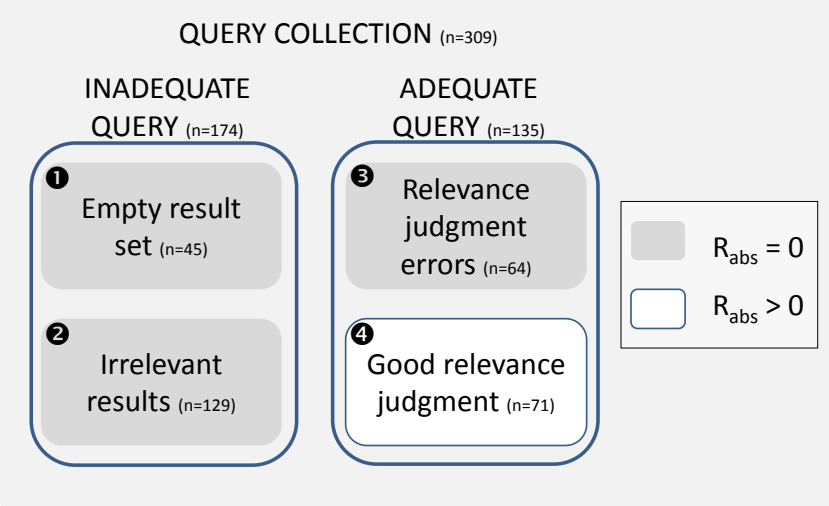


Figure 1: Outcome-based classification of queries

2.4.4. Query reformulation

Another angle from which we can study queries, next to analyzing the errors they contain, is the reformulation strategies used. As mentioned above, the participants had 15 minutes to complete the literature search task. In an ideal situation, they would have entered one comprehensive query, which covered all the components of the information need. However, as many of these students were not familiar with the search system, and as even more of them were not familiar with the subject of the search, most participants had to iterate the process of finding MeSH terms and combining them into a query. We identify different types of strategies and analyze their use by the different performer types.

3. Results

3.1. Sample description

3.1.1. Respondents

A total of 100 respondents participated in the test, two of whom did not formulate any queries and are therefore excluded from the analyses. Although the participants come from different linguistic (English versus Dutch-speaking) and educational (bachelor's versus master's level) backgrounds, a Kruskal–Wallis test indicated that there are no significant differences in recall between these groups, so we can safely concatenate them and use another categorization for the purpose of this study, i.e. best, average and worst performers.

3.1.2. Background

With regard to PubMed experience, our test group was rather heterogeneous: 44% had had an elaborate introduction into the use of the search engine, whereas others had only had a brief introduction (46%). Some (10%) claimed to have had no introduction into PubMed at all, although this was part of their curriculum.

About 97% use a computer several times a week to daily, but only 18% consult PubMed with the same frequency. About 40% of our test participants rarely or never use PubMed to search for medical information.

As for English language skills, 74.4% of the - British and Belgian - students achieved a B2[2] level in reading and 88.8% achieved a B2 level in vocabulary, indicating that they are "independent users" of the English language, and that they should be able to read and understand complex technical texts and "produce detailed text on a wide range of subjects".

3.2. Query analysis

3.2.1. Quality-based query analysis

We analyzed the queries in our collection (n = 309) and distinguish 8 types of errors (see Table 1 for an overview).

These error types are not mutually exclusive, i.e. one query can contain several errors, causing overlap between the error categories. Moreover, some errors induce other errors, e.g. "incorrect operator", and more specifically the excessive use of "AND", automatically leads to overspecification. The fourth column in the table shows the number of times each error occurs in our query collection.

A total of 60 queries did not contain any errors and covered all components of the information need.

[2] For more information about CEFR levels, see http://www.coe.int/t/dg4/Linguistic/Source/Framework_EN.pdf

Table 1: Error types and their frequencies

Error Type	Description	Example	*n*
1. Irrelevant MeSH term	Query contains at least one irrelevant MeSH term.	(("**Multifactorial Inheritance**"[Mesh] AND "Accidental Falls"[Mesh]) AND "Frail Elderly"[Mesh]) AND "Nursing Homes"[Mesh]	89
2. Overspecification	Query is too narrow and therefore yields few or no results.	"Pharmaceutical Preparations"[Mesh] AND "Aged"[Mesh] AND "Risk Factors"[Mesh] AND "Accidental Falls"[Mesh] AND "Nursing Homes"[Mesh]) AND "Nursing"[Mesh]	36
3. Underspecification	Query is too broad; contains only 1 or 2 concepts and yields a long list of citations.	"Accidental Falls" [Mesh]	125
4. Incorrect non-MeSH term	Query contains incorrect free-text search term. The corrective effect of the MeSH terms is lost, and spelling and translation errors corrupt the queries.	multifactorial **programm** and **faling**	42
5. Spelling error	A misspelled and therefore incorrect non-MeSH term	study for **fallprevention**	7
6. Incorrect translation	Query contains an incorrect translation. This can be an incorrect free-text search term, or a MeSH term which is believed to have another meaning than intended.	("Accidental Falls"[Mesh] AND "**Disabled Persons**"[Mesh]) AND "Nursing homes"[Mesh]	7
7. Incorrect operator	The excessive use of AND can lead to overspecification, whereas the exclusive use of OR leads to underspecification.	• ((("Aged"[Mesh] **AND** "Accidental Falls"[Mesh]) **AND** "Residential Facilities"[Mesh]) **AND** "Nursing Homes"[Mesh]) **AND** "Homes for the Aged"[Mesh] • (("Aged"[Mesh]) OR "Residential Facilities"[Mesh]) OR "Accidental Falls"[Mesh]	27
8. Syntax error	query contains unmatched brackets or quotes, or truncated words	• Accidental Falls"[Mesh]) AND**""**Frail Elderly"[Mesh]) AND "Nursing Homes"[Mesh] • "kine* AND ((("Aged"[MeSH] OR "Frail Elderly"[MeSH])) AND "Accidental Falls"[MeSH] AND "Residential Facilities"[MeSH]	17

3.2.2. Impact of query quality on potential recall

We analyzed the impact of the eight different error types on search performance, and noticed that three of those error categories had a significant impact on actual and potential recall: incorrect MeSH terms, underspecification, and the incorrect use of Boolean operators (see Table 2).

Table 2: Impact of query quality on potential recall

	n	$R_{pot} = 0$	Mean R_{pot}
Good queries	60	0	4.05
Queries with incorrect MeSH term	42	73%	0.78
Underspecified queries	125	77%	0.41
Queries with incorrect Boolean operator	27	81%	0.85

- **Irrelevant MeSH terms**. This error was made in almost 1 out of 3 queries (29%). A total of 73% of the queries containing an incorrect MeSH term had zero potential recall, either because of empty result sets (33%), or because the results were irrelevant to the search question (40%). In the remaining 27%, the search did yield some relevant results, despite the use of a MeSH term that was not entirely relevant for this search. Queries containing an incorrect MeSH term yielded less than one (0.78) relevant citation on average.

- **Underspecification**. The error of underspecification, i.e. when queries consist of only one or two terms and are therefore too broad, was made in 125 queries (40%). About 77% of the underspecified queries had zero potential recall. Underspecified queries yielded 0.41 relevant citations on average.

- **Incorrect use of Boolean operators**. In 27 queries (8%), one or more Boolean operators were used incorrectly. This manifests itself mainly in the excessive use of AND (67%) and OR (33%). This error led to zero potential recall in 81%, yielding empty result sets in 37% of the cases, and yielding only citations irrelevant to the search question in 44%.

- **Good queries**. A total of 60 queries (19%) were formulated correctly, with an average potential recall of just above 4 citations. This means that the participants who submitted these queries could have selected an average of four relevant citations, whereas they selected less than two.

3.2.3. Outcome-based query analysis

Next to the quality of the queries in terms of the number and types of errors they contain, we also assembled data on the potential and actual recall for each query. Potential recall data allow us to determine the direct influence of each error (type) on the efficiency of the query (see Section 3.2.2), and differences between potential and actual recall indicate relevance judgment errors.

We can subdivide our query collection into adequate and inadequate queries on the basis of their actual and potential recall. Inadequate queries did not yield any relevant results, either because the result set was empty (Figure 1 box 1), or because it contained only irrelevant citations (Figure 1 box 2). Adequate queries, on the other hand, were well-constructed and covered the information need. However, in some cases relevance judgment errors prevented the searcher from selecting relevant citations (Figure 1 box 3). This means that well-formulated queries do not guarantee high recall in the context of our study.

A total of 71 queries (22.9%, Figure 1 box 4) were well-formulated, and led to the selection of at least one relevant citation.

A total of 45 queries returned empty result sets, and another 129 queries had zero potential recall. This means that 56% of the queries in our collection contained errors and did not cover the information need.

A total of 135 queries (44%) were adequate, i.e. they yielded at least one relevant citation. In almost half of those cases (48%), the query itself was acceptable and – although it may contain one or more (minor) errors – had positive potential recall, but the issuer lacked in relevance judgment skills. The remaining 71 (52%) queries had positive potential recall, and their issuers selected at least one relevant citation from the lists of results.

3.3. Performance

During the search task, our test participants selected six citations on average, two of which were relevant (average $R_{abs} = 2$). The potential recall of their searches was four, which means that their search results contained four relevant citations on average, two of which were overlooked by our test participants.

3.4. Comparison of the performer types

3.4.1. Division into performer types

As mentioned in Section 2.3, we divided our test group into three performer groups, based on the number of relevant citations they selected. A total of 38 participants are labeled as "worst performers", 28 as "average performers" and 32 as "best performers".

A chi-square test did not reveal any significant differences in the distribution of the student types over the types of performers (see Table 3). However, there are more Dutch-speaking master's students in the best performer group than we would statistically expect (observed: 17, expected: 12.8; 53% of the best performers are Belgian master's students).

Table 3: Distribution of participants over 3 performer types

		worst performers (n=38)		average performers (n=28)		best performers (n=32)	
		%	n	%	n	%	n
Dutch	bachelor	27%	10	39%	11	28%	9
	master	39%	15	29%	8	53%	17
English	bachelor	13%	5	3%	1	6%	2
	master	21%	8	29%	8	13%	4

3.4.2. Background of the performer types

There are no significant differences in language skills between the performer types: the average level in all three groups (including the native speakers of English) is B2 for both reading and vocabulary.

A Kruskal–Wallis test showed no significant differences between the performer types in prior experience with PubMed, general computer skills, or in general usage of the

Internet to search for information. Although the difference is not significant, we do see that more than half of the participants in the best performer group (53%) are students who had received an elaborate introduction into the use of PubMed.

In the posttest questionnaire, we asked the students whether they were satisfied with their search results and their search process. A one-way ANOVA test ($F(2, 97) = 28.917$; $p < .001$) showed that the worst performers were significantly less satisfied with their search results than the average and best performers (Bonferroni correction; $p < .001$ for both groups). The worst performers also experienced their search process as less fluent than the other two groups ($F(2, 97) = 22.796$; $p < .001$; Bonferroni correction: $p < .001$) and one in three of the worst performers find PubMed not so user-friendly, as opposed to less than one in five in the average and best performer groups.

3.4.3. Search process

On average, all three performer types submitted three queries during the search. However, we do see that the number of participants who needed only one query to conduct their search task is higher in the best performer group than in the other groups. This means that their searches are more focused from the beginning, whereas the other participants needed more queries to find what they were looking for.

We measured the time spent on querying, i.e. the time spent on searching for MeSH terms and combining them into a query. As we explained in our previous study (Vanopstal et al., 2012), longer querying times can indicate hesitation. A one-way ANOVA test showed that there were significant differences in querying times between the performer types ($F(2, 95) = 11.896$, $p < .001$). Bonferroni post-hoc comparisons of the three groups indicated that the worst performers needed significantly more time to formulate their queries than the average ($p = .001$) and best ($p < .001$) performers.

Total evaluation time is the time spent on skimming the result list(s) for relevant results. As the total evaluation times were not distributed normally, we used a Kruskal–Wallis test to find any differences between the three performer types ($H = 18.18$, $p < .001$). Post-hoc tests for pairwise comparison showed us that the average and best performers spent significantly more time on the evaluation of the search results ($p = .003$ and $p < .001$, respectively).

3.4.4. Quality-based query analysis per performer type

Figure 2 shows a summary of the errors that will be discussed in this section. Although we also see some clear differences in the number of bad MeSH terms used by the performer types, and we have already shown the impact of incorrect MeSH terms on recall (see Section 3.2.2), we only found significant differences in the number of underspecification errors and in the incorrect use of Boolean operators. We refer to the error analysis for an analysis of the direct impact of different types of errors on recall.

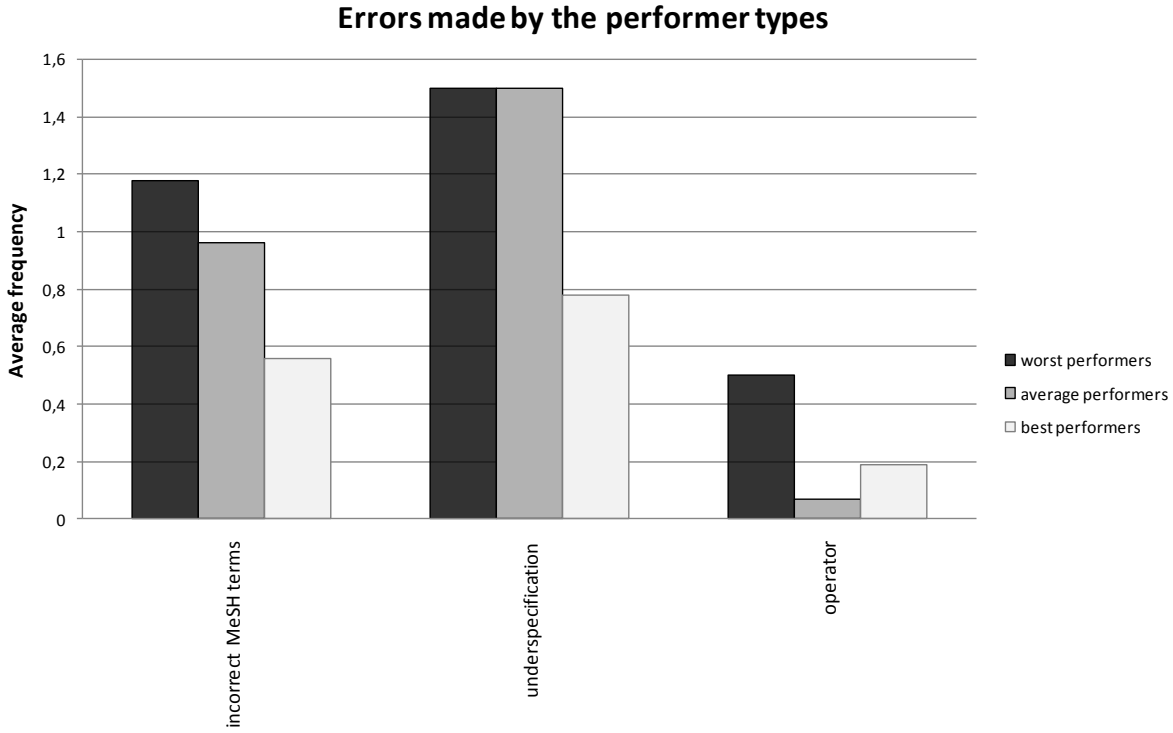


Figure 2: Summary of errors per performer type

- **MeSH terms**

  As described above (see Section 2.1), our test participants were instructed to use MeSH terms. In previous research (Vanopstal et al., 2012), we have shown that MeSH terms have a corrective effect; they compensate for possible errors in the free-text search terms that were entered in the MeSH module. Although these free-text search terms have no direct effect on recall, they may have an impact on the fluency of the search process. The worst performers formulated

significantly more search terms than the other two groups ($H = 9.95$, $p = .007$), indicating that they struggled to find the right MeSH terms for their search.

The best performers selected a smaller number of incorrect MeSH terms, which enabled them to construct better queries. Although there is a clear trend in the number of badly chosen MeSH terms, the differences between the performer types is not significant.

- Underspecification

Both worst and average performers made a high number of underspecification errors: 1.5 times on average during the search. A Kruskal–Wallis test showed a significant difference in occurrence of this error between the performer types ($H = 8.030$; $p = .018$), more specifically between worst and best performers (Bonferroni correction; $p = .028$).

- Boolean operators

A Kruskal–Wallis test revealed a significant difference between the performer types in the incorrect use of Boolean operators ($H = 8.037$, $p = .018$), usually the excessive use of AND or OR. This is only true for the worst and average performers (Bonferroni correction; $p = .014$). There are no significant differences in the incorrect use of Boolean operators between the best and worst performers.

3.4.5. Differences between actual and potential recall as an indication of relevance judgment quality

Figure 3 below gives an overview of the number of citations viewed by each performer group and the proportions of relevant and irrelevant citations. For each PubMed search, we registered how many – titles of – citations in the result list were viewed. When a participant performed more than one search, we added up this number from the several searches. On average, 67 citations were viewed. The worst performers viewed 57 citations on average, 55 (96%) of which were irrelevant. Although the remaining two (4%) were relevant, this group failed to distinguish them from the relevant ones. The average performers viewed 72 citations on average, 69 (96%) of which were irrelevant. They missed some citations, but succeeded in identifying some too. On the other hand,

this group also selected more irrelevant citations than the worst performers. Finally, the best performers viewed 92 citations on average, 10% of which were relevant, indicating that their queries were better constructed than those in the other two groups. They were also better at identifying the relevant citations, as they only missed 38% of the relevant ones in the results lists. However, they also selected a relatively high number of irrelevant citations.

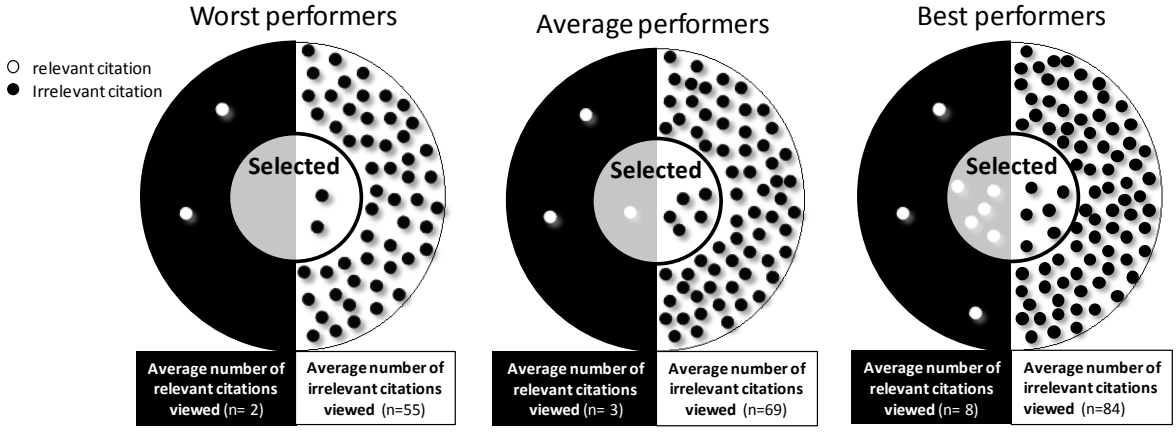


Figure 3: Relevant versus irrelevant citations selected by the performer types

3.4.6. Outcome-based comparison

We already stated above (see Section 3.2.3) that low recall can be caused by either ill-formulated queries, or bad relevance judgment. In Figure 4, this information is linked to the performer types.

Ill-formulated queries can lead to empty result sets, or to zero potential recall. About 74% of the queries issued by the worst performers were ill-formulated, which is almost double of the erroneous queries in the group of average (44%) and best (41%) performers.

About 60% of the queries submitted by the best performers were adequate, i.e. they yielded at least one relevant citation (potential recall > 0). In the group of average performers, this was 56%, whereas no more than 26% of the queries in the worst performer group yielded relevant results.

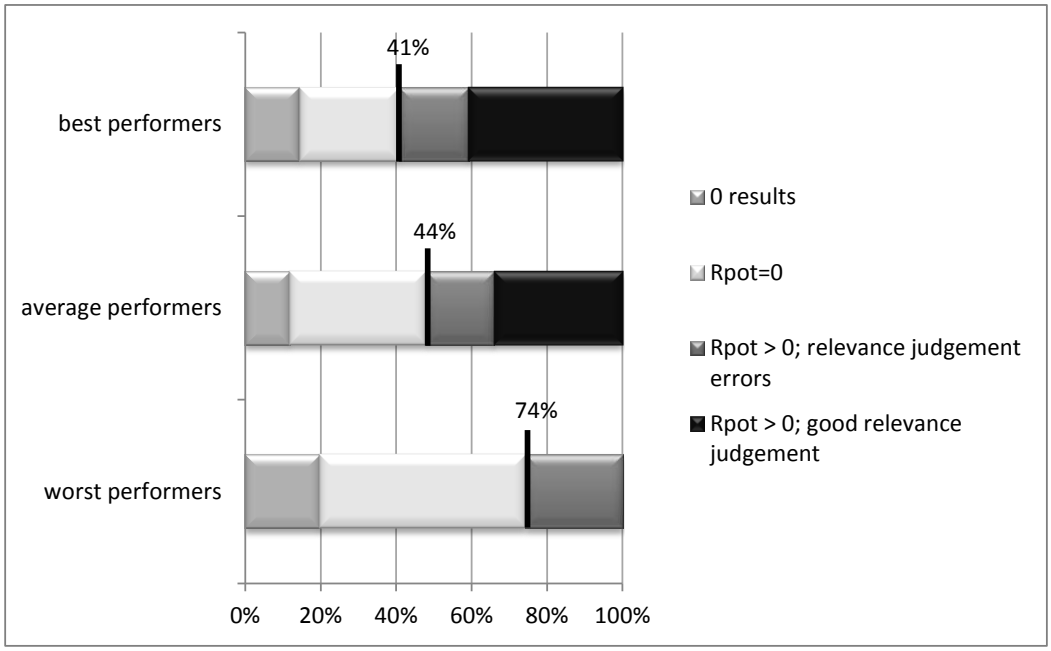


Figure 4: Percentage of zero and positive potential recall queries per performer type

Due to bad relevance judgment, the worst performers failed to identify any of the relevant citations yielded by those 26% of good queries. The best and average performers failed to identify any of the relevant citations yielded by their adequate queries in 18% of the cases.

3.4.7. Query reformulation

The formulation of a good query requires a conceptual analysis of the information need, and a thorough understanding of the syntax used by the search engine. When a query does not yield satisfactory results, a searcher may have problems finding alternative ways to formulate it. It takes some insight to see what exactly went wrong in a query for a searcher to be able to correct that error.

We distinguish six different types of reformulation: narrowing, broadening, substitution, repetition, trial and error, and a last category which we call “one”.

- **Narrowing:** a more general query is made more specific by adding one or more MeSH terms

  e.g. Query 1 = "Housing for the Elderly [MeSH] AND Accidental Falls [MeSH]"; Query 2 = "(Housing for the Elderly [MeSH] AND Accidental Falls [MeSH]) AND Accident Prevention [MeSH])"

- **Broadening:** a query that is too specific - and therefore often yields an empty results set – is made more general by omitting one or more terms from the query

  e.g. Query 1 = "(Housing for the elderly [MeSH] AND Accident Prevention [MeSH]) AND Nursing Homes [MeSH]"
  Query 2 = "Accident Prevention [MeSH]) AND Nursing Homes [MeSH]"

- **Substitution:** one MeSH term is substituted for another

  e.g. Query 1 = "Accidental Falls [MeSH] AND Frail Elderly [MeSH]"
  Query 2 = "Accidental Falls [MeSH] AND Elderly [MeSH]"
  Query 3 = "Accidental Falls [MeSH] AND Residential Treatment [MeSH]"
  Query 4 = "Accidental Falls [MeSH] AND Combined Modality therapy [MeSH]"

- **Repetition:** re-use of a previous query

- **Trial and error:** formulation of a completely different query, as the previous one did not appear to yield any satisfying results.

  e.g. Query 1 = "Critical pathways [MeSH]"
  Query 2 = "Accident Prevention [MeSH]"
  Query 3 = "(Aged [MeSH] OR Frail Elderly [MeSH] OR Housing for the elderly [MeSH])"

- **One:** only one query was submitted.

In general, there are no significant differences in the use of one specific reformulation strategy between the three performer types, except for the trial and error strategy (Kruskal–Wallis $H = 9.010$; $p = .011$). The worst and average performers use this strategy significantly more often than the best performers (Bonferroni correction, $p = .046$ and $p = .018$, respectively). This may be another indication that their searches are less fluent.

As pointed out above (see Section 3.4.4), the best performers used a lower number of incorrect MeSH terms in their queries than the worst performers did. There are three ways in which this error can be corrected: by removing the incorrect MeSH term, which is a way of broadening the query, by replacing the incorrect MeSH term (substitution), or by formulating a completely new query (trial and error). The errors that were made in the best performer group were corrected in 60% of the cases, as opposed to 48% in the worst performer group.

We already showed that there were no significant differences in the incorrect use of Boolean operators between the worst and best performers. The difference between the two groups lies more in their reaction to the - usually poor - results of these searches. In only 26% of the cases did the worst performers succeed in correcting the erroneous query. The other queries either repeated the error, or they were replaced by another erroneous query. This indicates that the searchers did not know exactly what went wrong. The best performers, on the contrary, corrected 83% of the queries containing an error of this type. Correction is done by either replacing the operator (substitution), removing a component of an overspecified query (broadening), or by formulating a completely new query (trial and error).

The best way to correct an underspecified query is to narrow it down to a more specific one. About 60% of the underspecified queries were corrected this way in the best performer group, as opposed to 34% and 31% in the worst and average performer groups, respectively.

An overspecified query should be corrected by broadening. This reformulation strategy was used in 15%, 25% and 33% of the queries in the worst, average and best performer groups, respectively.

Incorrect free-text terms seem to be very difficult to correct, as the searcher mostly does not realize that there is an error in the query. These free-text terms were replaced in nine (out of 43) queries, but only in two of those queries did the searcher (best performer type) replace the incorrect free-text term with a correct one.

4. Discussion

4.1. Main findings

When we look at the separate queries, there are three error types which have a direct impact on potential recall, i.e. which cause the query to yield few or no relevant citations: incorrect MeSH terms, underspecification and incorrect Boolean operators. Between 73% and 81% of the queries containing these error types had zero potential recall.

Good queries do not guarantee high recall: in almost half of the queries with positive potential recall, students failed to identify the relevant citations. This indicates that the participants experienced some problems during the relevance assessment step.

None of the four student types (Dutch-speaking bachelor's and master's students, native English bachelor's and master's students) outperformed the others, whereas we had expected the English (master's) students to be the better-performing ones. The Dutch-speaking master's students are better represented in the best-performing group. This group had had the most elaborate introduction into PubMed during their training. This may indicate that language skills – although obviously important - do not compensate for the lack of facility with the search engine.

There are no significant differences between the performer types in the scores on the language tests, educational background or computer skills. The worst performers did not select any relevant citations, and they are well aware of their poor performance. One in three of these participants assessed the PubMed search system as "not so user-friendly".

The worst performers struggled to find the correct MeSH terms for their searches and generally needed more time to formulate their queries. On the other hand, they spent less time on the evaluation of the search results, a crucial step in information retrieval.

Making errors may be one indication of poor research skills. However, the correction of an error in the next query demonstrates a certain level of understanding of the system. This study showed that the ability to correct one's own errors distinguishes better performing searchers from the less successful ones.

4.2. Strengths and limitations

One of the limitations of this analysis is the small number of queries available for research. It is difficult to find significant results for such a small dataset. However, we do believe that the fact that these queries were all meant to fulfill the same information need – as opposed to queries from logs, where the information need is unknown – adds to the validity of our conclusions.

4.3. Critical remarks on main findings

4.3.1. Impact of query quality

As argued by Dogan et al. (2009) the quality of a query depends on 3 factors: the searcher's understanding of the information need, his searching skills, and system design on the search engine's side. The retrieval experiment described in this paper was set up to enable us to formulate advice for the improvement of bibliographic instruction. In an earlier paper, we concluded that the non-identification of concepts in the information need was the main cause for non-coverage. The first factor, i.e. understanding of the information need, is therefore a problem that should be tackled in bibliographic instruction. The second factor, searching skills, should be addressed in bibliographic instruction as well, focusing on three error types: incorrect use of MeSH terms and of Boolean operators, and the formulation of underspecified queries.

Most of the queries that contained an *incorrect MeSH term* did not lead to the selection of any relevant citations, either because of empty result sets, or because the query only yielded irrelevant results, or because relevant citations were overlooked.

Underspecification, also referred to as "the million hits syndrome" (Mulligen et al., 2004), leads to very long lists of results, which discourage the searcher from skimming the results. In almost two out of three of the underspecified queries, test participants considered cost-effectiveness too low and constructed a new query. Only 12% made the

effort of going through the results, and succeeded in identifying at least one relevant citation. Underspecification in itself therefore does not render a query completely useless; however, it makes the relevance judgment step much more labor-intensive and causes people to give up.

The danger of using *incorrect operators* lies especially in overspecification, which usually results in queries with zero potential recall.

Medical students should learn how to construct comprehensive queries that cover the information need, without overspecifying. They need to gain more insight into the use and structure of MeSH, practice combining the terms to a good query, and learn to interpret the MeSH terms assigned to the citations that were retrieved. In this respect, the incorporation of MeSH translations into the search engine may be useful for non-native speakers of English. An understanding of the indexing and relevance sorting algorithms may also help to formulate better queries (Aula, 2003).

The absence of errors in queries, however, does not guarantee positive recall: bad relevance judgment may cause searchers to overlook relevant citations, as it did in about 25% of the queries. More experience in reading scientific articles, and more familiarity with the display settings in PubMed may facilitate relevance assessment of citations based on their abstract.

4.3.2. Performer profiles

There are no significant differences in the distribution of the two student levels in the groups of performers (see Table 3), although the Belgian master students are better represented in the best performer group. We assumed that native speakers of English would do better on a literature search task in PubMed, and therefore that a larger proportion of the native English participants would be in the best performer group. However, their language skills do not seem to compensate for the lack of searching skills.

Although there are no significant differences between the performer types with regard to PubMed familiarity or frequency of use, we do see that more than half of the best

performers were Belgian master students – the most experienced PubMed users in our test group. Searching skills therefore definitely play a role in search efficiency.

We did not find any significant differences in language skills between the performer types. However, when we only look at the non-native speakers of English, a Kruskal–Wallis test shows that the best performers scored better on the reading test than the average and worst performers ($H = 3.968$; $p = .047$): 81 percent of the best performers achieved a B2 level or higher, as opposed to 60 and 44 percent in the worst and average performer groups, respectively. The differences in scores on the vocabulary test are less obvious, as the scores are relatively high in all three groups. This means that English – reading – skills do play a role in information retrieval, more specifically in non-native speakers of English.

4.3.3. Errors made by the different performer types

Long citation lists resulting from underspecified queries discourage most searchers from scrolling through them. Participants of the worst performer type who made this error failed to select any relevant citations, whereas some of the average and best performers did. This means that the latter are either more perseverant, or their relevance judgment skills compensate for a low-quality query. Underspecification therefore especially has an impact on recall in those searchers who lack in relevance judgment skills.

The incorrect use of Boolean operators was especially found in queries submitted by the worst and best performers, whereas only three average performers committed this error. Differences in system experience may partly explain this difference between worst and average performers, whereas the differences between average and best performers may be caused by the length of the queries. Query length in the average performer group was 4.1, in the best performer group 5.8. Longer queries automatically contain more operators, which makes them more error-prone.

We consider citations that do not contain the crucial components *falls* and *fall prevention* as completely irrelevant to the search question: citations in which these two components are not represented contain too little information to answer the information need. Surprisingly, we see that the best performers selected a significantly

larger number of citations without the components *falls* and *prevention* than the worst performers. They selected more relevant, but also more completely irrelevant citations. This illustrates the classical trade-off between precision and recall: the students' selections contain an increasing number of irrelevant citations with increasing performance (r_s = .344, p = .000, n = 98). In other words, the higher the recall, the more "noise" we see in the students' selections.

The main difference between bad and average or good performers lies in the query formulation step. The worst performers failed to construct a comprehensive query with relevant MeSH terms and no syntax errors. This issue should clearly be addressed in bibliographic instruction. The difference between average and good performers is subtler, and also mainly originates in the query formulation step. This is illustrated by the average potential recall scores in each of the performer types: average recall in the worst performer group was 0.5, and 1 and 3 in the average and best performer groups, respectively. Although their queries were still rather unsuccessful, the average performers did succeed in identifying some of the relevant citations their queries yielded. The best performers' queries were better-constructed and yielded more relevant results, which, in turn, made it easier for the participants to identify them. The best performers spent more time on relevance judgment, probably because they made strategic decisions in allocating enough time to this crucial last phase.

4.3.4. Query reformulation

Incorrect free-text terms are rarely (twice in our query set) corrected by our test participants, rather they are repeated, or replaced by another incorrect free-text term. This corroborates our previous finding that the extra step of selecting MeSH terms can be very useful to prevent errors from percolating to the final query (Vanopstal et al., 2012).

Another error that seems very difficult to correct, is the error of overspecification. About one in three of these errors were corrected. This error therefore also deserves some extra attention in bibliographic instruction.

The incorrect use of MeSH terms, and underspecification and overspecification errors are problems that need extra attention, especially in the instruction of novice searchers.

They seem to have more difficulty in correcting these errors than the better-performing searchers.

5. Conclusions

We conducted a retrieval experiment in a group of nursing students with mixed linguistic and education level backgrounds: Dutch-speaking master's and bachelor's nursing students, and native English-speaking master's and bachelor's nursing students. The aim of this study was twofold: to formulate advice for the improvement of bibliographic information retrieval instruction, and to draw a profile of the best, average and worst performers in the test.

An analysis of the queries submitted by our test participants allowed us to identify the errors with a direct impact on recall, and to determine a focus for bibliographic information retrieval instruction. Although broad queries can be good for a searcher's orientation within a specific domain, exercises on the translation of an information need into a good query should prevent the students from formulating broad or underspecified queries (only). The skills required for this include a thorough analysis of the components of the information need, the translation of these components into free-text search terms and subsequently into MeSH terms. Students may benefit from some practice in the use of these MeSH terms, which can enhance a search considerably, provided the terms are used correctly. We agree with Aula's (2003) assertion that an understanding of the indexing and relevance sorting algorithms may also help to formulate better queries. Combining MeSH terms using Boolean operators to obtain a comprehensive query is a difficult task which should also be addressed in bibliographic retrieval instruction.

Another problem in information retrieval using PubMed is the relevance judgment step. Relevant citations are often overlooked, even by native English speaking searchers. Skimming exercises may help the students to detect the structure and contents of abstracts more easily. General familiarity with scientific texts may also facilitate the relevance judgment step.

We tried to draw a profile of the "efficient searchers" in our test group and analyzed what they did differently from the less efficient searchers. In non-native speakers of

English, the level of English language skills plays an important role in retrieval, as the best performers are those with the highest scores on the English language tests.

More than half of the best performers proved to be Belgian master's students, the group who had received an elaborate introduction into the use of PubMed in their master's training.

The best performers generally formulated better queries, were better at detecting and correcting the errors in their queries and had less difficulty in identifying the relevant citations in the result sets. The correction of one's own errors in queries requires insight into the search system and a critical analysis of the queries. The best performers are better at correcting errors pertaining to incorrect MeSH terms, Boolean operators and underspecification. They do, however, also have problems detecting and correcting the apparently more complex errors of overspecification and incorrect free-text terms.

6. Future work

We would like to experiment with some techniques that facilitate both query formulation and relevance judgment for non-native English searchers. A translated version of the Medical Subject Headings can help them to formulate a good query. This translation can also be integrated for relevance judgment: listing the translated MeSH terms that are assigned to each citation can be helpful do decide whether an article is relevant to the information need or not. We would also like to experiment with simplified abstracts using automatic paraphrasing techniques, and with wikification (He et al., 2011), which may also make the selection of relevant abstracts easier. Applying comprehensibility assessment techniques like OCSLA (Liu & Lu, 2009) to the abstracts in PubMed may provide some insight into the reasons why some texts are more easily understood – and selected – than others.

References

Aula, A. (2003). *Query Formulation in Web Information Search.* Paper presented at the IADIS International Conference WWW/Internet Algarve, Portugal.

Dogan, R. I., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database, 2009*, bap018. doi: 10.1093/database/bap018

He, J., Rijke, de M., & Sevenster, M. (2011). *Generating links to background knowledge for medical content*. Paper presented at the Second International Workshop on Web Science and Information Exchange in the Medical Web (MedEX 2011). http://dare.uva.nl/record/420713

Hofstede, A.H.M, Proper, H.A., & van der Weide, T. P. (1996). *Query formulation as an information retrieval problem* (Vol. 39). Oxford: Oxford University Press.

Jenuwine, E. S., & Floyd, J. A. (2004). Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. *J Med Libr Assoc, 92*(3), 349-353.

Liu, Rey-Long, & Lu, Yun-Ling. (2009). Online assessment of content skill levels for medical texts. *Expert Systems with Applications, 36*(10), 12272-12280.

McCray, A. T., & Tse, T. (2003). Understanding search failures in consumer health information systems. *AMIA Annu Symp Proc*, 430-434. doi: D030003312 [pii]

Mulligen, E. van, Diwersy, M., Schijvenaars, B., Weeber, M., van der Eijk, C., Jelier, R., Schuemie, M., Kors, J., & Mons, B. (2004). Contextual annotation of web pages for interactive browsing. *Medinfo, 11*(Pt 1), 94-98.

Richter, Randy R., & Austin, Tricia M. (2012). Using MeSH (Medical Subject Headings) to Enhance PubMed Search Strategies for Evidence-Based Practice in Physical Therapy. *Physical Therapy, 92*(1), 124-132. doi: 10.2522/ptj.20100178

Sutcliffe, AlistairRyan MicheleDoubleday AnnSpringett Mark (Writer). (2000). Model mismatch analysis: towards a deeper explanation of users' usability problems [Article], *Behaviour & Information Technology*: Taylor & Francis Ltd.

Vanopstal, Klaar, Vander Stichele, Robert, Laureys, Godelieve, & Buysschaert, Joost. (2012). PubMed Searches by Dutch-Speaking Nursing Students: The Impact of Language and System Experience. *JASIST, 63*(8), 1538-1552.

List of figures

List of tables

Chapter IV: Query formulation and relevance judgment in native and non-native English-speaking PubMed users

Abstract

Objective To investigate the impact of the language handicap of non-native English-speaking users of PubMed, together with the impact of system experience.

Materials and Methods We set up a 15-minute retrieval experiment with a specific information retrieval task in PubMed in which participants were instructed to use MeSH terms. The search process and output were recorded and analyzed, together with keystroke logging. This allowed us to study both the query formulation and the relevance judgment step. Moreover, an in-depth analysis of recall was performed.

Results Forty Dutch-speaking and 21 native English-speaking master students in nursing participated. The English-speaking students had better language skills, whereas the Dutch-speaking students had more system experience with PubMed. During the test, the Dutch-speaking students experienced more difficulties in covering concepts and finding the correct terms, but they used MeSH more efficiently, i.e. in combination with free-text terms. Their queries yielded more relevant articles (5 versus 2 on average), and their selections had a higher informative value (weighted recall 44 versus 21 on average).

Conclusion Dutch-speaking users of PubMed have a linguistic disadvantage which leads to poorer performance in the initial stages of query formulation (concept coverage and search term formulation). Training which focuses on searching skills, on a more advanced use of MeSH terms, and on better relevance judgment can compensate for this handicap. The Dutch-speaking students' system experience resulted in higher recall than in the native

English-speaking group, who had had no prior formal searching skills training.

Keywords: Information Storage and Retrieval, Medical Subject Headings, Education, Language, Nursing

1. Introduction

With the evolution of medical sciences and the explosion of the internet, efficient literature searching has become crucial to professionals working in the medical field, and especially in evidence-based medicine. One of the major tools used for biomedical information retrieval is PubMed, a search interface that provides free online access to MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed).

Several studies (Dogan et al., 2009; Herskovic et al., 2007; Lu et al., 2009; Silverstein et al., 1999) have been devoted to the analysis of PubMed queries through the analysis of large query logs in order to respond to the needs of the users and to improve the search system. On the basis of such a large query log analysis, Dogan et al. (Dogan et al., 2009) concluded that large result sets seem to have a discouraging effect on the selection of citations. They also found that queries are often reformulated and that searchers would benefit from author disambiguation, and from optimized ranking techniques. Lu et al. (2009) report on their query log analysis which resulted in the implementation of the Related Queries component in PubMed. Analysis of smaller query logs (e.g. Hoogendam et al. (2008)), on the other hand, allow researchers to focus on a specific group of users, and are therefore more likely to result in actions on the end-users' side, such as suggestions for the improvement of bibliographic instruction and methods to facilitate query formulation.

The Medical Subject Headings, a thesaurus and controlled vocabulary designed by the NLM to enable more focused searching, have been translated into several languages in order to support non-native speakers of English in their search for (bio)medical information (Anne et al., 2010; Fontelo et al., 2007; Liu et al., 2006; Thirion et al., 2007). Although – or because – it seems logical that non-native speakers of English have

difficulties searching for very specific information in a foreign language, the impact of the language handicap in medical information retrieval has not been studied in detail yet. In this paper, we focus on the interaction between system experience and English language skills. The aim of the study was to investigate the impact of the language handicap of non-native English-speaking users of PubMed, also taking into account the impact of system experience.

2. Method

We used individual query logs from a sample of a limited number of participants, complemented with a recording of the entire search process of each individual, and an in-depth analysis of recall. This enabled us to detect obstacles in the retrieval process, and to identify and compare these obstacles in the search process and in the resulting output of non-native versus native speakers of English.

2.1. Experimental setup

We set up a literature searching task in two convenience samples of master's nursing students: a group of Belgian, Dutch-speaking students, and a group of British, native English-speaking students. They completed the same literature searching task, from which we extracted information about characteristics of the search process on the one hand, and about the outcome of the search on the other hand. These data will be compared for both test groups.

The participants had to search for citations that were relevant to a pre-formulated search question: "What is the effect of a multifactorial treatment on the incidence of falls in elderly who live in long-term care facilities?". The participants were instructed to use MeSH terms and combine them into PubMed queries. To ensure that all participants had a basic understanding of the query formulation process using MeSH terms, they were given a short tutorial which explained the three steps in the formulation process: entering free-text search terms, selecting MeSH terms, and combining them with Boolean operators into a more complex query.

2.2. Recruitment

We recruited Dutch-speaking master's students at the Nursing and Midwifery Department of Antwerp University in Belgium, and native English-speaking master's students at the School of Nursing of the University of Nottingham.

2.3. Measurements

2.3.1. Respondent characteristics

A pretest questionnaire provided us with information about the participants' sex, age, self-reported language skills, educational background, and about their experience with PubMed. This information allowed us to take into account any biases in our samples. In order to assess the test participants' language skills in an objective way, they completed the DIALANG[1] language tests which focus on reading and vocabulary skills.

2.3.2. Query formulation process

In this qualitative analysis, we analyze both the process and the outcome of the query formulation step.

- **Process indicators**

We used Morae[2], a software package designed to test system usability, to register screen views and keystrokes during the search process.

- *Concept coverage*

  A first difference between the search processes of Belgian versus British students was that the British students started from a search question in their own language, whereas the Belgian students had to translate the question that was formulated in Dutch, into English concepts. Concept coverage is therefore an interesting aspect in the comparison of the two groups. We consider a concept as "covered" when its corresponding MeSH term occurs in at least one of the queries submitted by the participants.

[1] http://www.lancs.ac.uk/researchenterprise/dialang/about
[2] http://www.techsmith.com/morae.html

- *Quality of search terms and MeSH terms*

We assigned a quality label to each of the search terms and MeSH terms entered or selected by the test participants: 0 for a bad, 1 for a medium and 2 for a good search or MeSH term. For more details about scoring search and MeSH terms, see Vanopstal et al. (2012).

- *Mixed queries*

Although they were instructed to search for MeSH terms for each component of the search and to combine these MeSH terms with Boolean operators, some participants also used free-text terms in their queries. We will compare the use of "mixed" queries in the two groups. As the Boolean operator *OR* is typically used to express parallel relationships, and is often used to combine a MeSH term with a free-text term, we also analyze the use of *OR* in this section.

- *Error types*

In an earlier study (Vanopstal et al., 2013), we distinguished eight types of errors in the queries submitted by the test participants: incorrect MeSH terms, underspecification, overspecification, spelling errors, incorrect translations, incorrect non-MeSH terms, incorrect use of Boolean operators, and syntax. We compare the English-speaking and Dutch-speaking groups to see whether they both make the same types of errors.

- Outcome indicators

The average potential recall score (see also Vanopstal et al.(2013)) was calculated for each participant. We used NLM's E-Utilities[3] to reconstruct the output of the participants' searches. The potential recall score is an indication of the quality of the queries, and is calculated on the basis of a gold standard list of citations. The gold standard list was developed using three principles: union of outputs, a gold standard query, and an evaluation of the related citations.

[3] http://www.ncbi.nlm.nih.gov/books/NBK25500/

Potential recall indicates how many relevant citations the queries logs contained, irrespective of whether they were selected by the participants or not. We only took into consideration the citations (20 per page) that were actually viewed by the participants. Potential recall is based on a binary criterion: citations that are in the gold standard are relevant, those which are not are considered irrelevant.

2.3.3. Relevance judgment

- **Process indicators**

The time spent on relevance judgment can be considered as an indication of how fluently and thoroughly relevance judgment was executed.

- **Outcome indicators**

In this paper, relevance is studied from a user-oriented (Park, 1994) or subjective (Swanson, 1986) perspective. We assess the selection made by the test participants from the system's output, using three different measurements: absolute recall, the correlation between potential and absolute recall, and weighted recall.

- *Absolute recall*

  Absolute recall expresses the number of relevant citations selected by our test participants, i.e. citations that were also in the gold standard.

- *Correlation between potential and absolute recall scores*

  We consider the correlation between potential and absolute recall as an indication of relevance judgment quality. A high potential recall score means that the results yielded by a query contained a high number of relevant citations. A high absolute recall score means that the participant was able select these relevant citations from the system's output. Stronger correlations between potential and absolute recall therefore suggest better relevance judgment.

- *Weighted recall*

  In our previous studies, recall was calculated on the basis of binary relevance criteria only: citations were either relevant or irrelevant. However, these measures

do not take into account the degree of relevance. Citations that did not cover all aspects of the search question may nevertheless also contain relevant information and can therefore also help the user to satisfy his or her information need. Hence we decided to use a more fine-grained scoring system for the students' selections, next to absolute recall. Each citation in the students' selections was assigned a score which indicates how many of the components were present in that citation. We assigned a heavier weight to the more important components of the search question: the crucial components of *falls* and *prevention* received a score of two, the other components were assigned a score of one. A citation containing the components *falls*, *elderly*, *long-term care*, and *prevention*, for instance, received a weighted recall score of six. The scores are added up for the total number of selected citations, which results in a total weighted recall score per participant. This total score is an indication of the information gain achieved after a 15-minute PubMed search. This may provide better insight into the relevance judgment skills of our test participants.

2.4. Statistical analysis

The results of this analysis will be presented as a comparison between the Dutch-speaking and the native English-speaking groups. As the test groups are relatively small, and most variables were not normally distributed, we used the non-parametric Mann Whitney U statistic to test the significance of differences between the two groups. We tested the correlation between potential and absolute recall with the Spearman Rank Correlation test in both groups. We used the Chi-square (x^2) test for nominal variables.

3. Results

3.1. Respondent characteristics

3.1.1. Demographics

A total of 61 master's students participated in the test: 40 Dutch-speaking and 21 native English-speaking nursing students. Forty-seven of them were female, 14 male, all between 21 and 24 years old. The Belgian students were in the fifth year of their Nursing and Midwifery master's training at Antwerp University, Belgium; the British

participants were fifth-year nursing students at the University of Nottingham, UK. Their curricula were more or less parallel, so we assumed that their educational backgrounds were comparable.

3.1.2. PubMed experience

Although both test groups had the same age and training level (master's nursing students), there was a clear difference in PubMed experience. All 40 Belgian students had had an elaborate introduction into the use of PubMed during their master's training, 70% of them used PubMed once or several times a week at the time of the test. They had all heard of the Medical Subject Headings, and 78% used MeSH terms from time to time to construct PubMed queries.

The British students, on the other hand, had only received a short introduction into the use of PubMed during their training as nurses, and 67% of them indicated that they rarely or never used PubMed. Only one of the British students had heard of MeSH terms, but he/she had never used them.

3.1.3. Language skills

Another major difference was the mother tongue of the two groups: Dutch versus English. A Mann-Whitney U-test shows a significant difference between the two groups in the results on both the reading and the vocabulary test (U= 222.5, p= .002 and U=151, p =.000, respectively. See table 1).

Table 1: PubMed experience (self-reported) and language skills

	Dutch (n= 40)	**English (n= 21)**	**statistical test Mann-Whitney U/ x^2**
% notion of MeSH	100%	4.8%	x^2 (1, n= 61)= 56.678; p= .000
% using MeSH occasionally	78%	0%	x^2 (2, n= 61)= 46.116; p= .000
% using PUBMED > 1 x a week	70%	0%	U= 35.00; z= -5.955; p= .000
% Fluent English (C1 or C2) – vocabulary	25%	81%	x^2 (1, n= 61)= 17.474; p= .000
% Fluent English (C1 or C2) – reading	23%	62%	x^2 (1, n= 61)= 9.273; p= .002

3.2. Analysis of the query formulation process

3.2.1. Process indicators

- **Concept coverage**

In comparison with the Dutch-speaking group, the English-speaking group achieved higher coverage for the concepts of *elderly*, *falls* and *long-term care*, and the same – low – coverage for *prevention*. Most of the participants did not identify the concept of *prevention* from the search question and therefore did not look for a corresponding MeSH term to add to their queries.

Table 2: Query formulation process

	Dutch (n= 41)	English (n= 21)	statistical test
PROCESS EVALUATION			
A. Concept coverage			**X^2**
1. Concept: elderly	75%	95.24%	$x^2(1, n= 61)=.476$; NS
2. Concept: falls	87.5%	100%	$x^2(1, n= 61)= .039$; NS
3. Concept: prevention	10%	9.52%	$x^2(1, n= 61)= .168$; NS
4. Concept: long-term care	60%	80.95%	$x^2(1, n= 61)= 1.137$ NS
B. Search terms and MeSH terms			**Mann-Whitney**
1. % well-formulated search terms	51%	90%	$U= 56.000$; $z= -5.564$; $p= .000$
2. % correct MeSH terms	74%	83%	$U= 319.000$; $z= -1.591$; NS
C. Mixed queries			**Mann-Whitney/ x^2**
1. Average number of free-text terms	1	0	$U= 294.000$; $z= -2.374$; $p= .018$
2. % participants who used "OR"	75%	33%	$x^2(1, n= 61)= 10.018$; $p= .002$
D. Error Types			
Average number of incorrect free-text terms in queries	1	.09	$U= 322.500$; $z= -2.133$; $p= .033$
OUTCOME EVALUATION			
Mean potential recall	5	2	$U= 331.00$; $z= -1.371$; NS

- Quality of search terms and MeSH terms

The participants were instructed to build queries by finding appropriate MeSH terms for each of the components of the search question. In order to find these MeSH terms, they first had to enter a search term into the MeSH module (e.g. *falls*), then select the MeSH term that best represented the concept they were looking for (*Accidental Falls*), and add it to the search builder. In order to construct more complex queries, i.e. queries which consist of multiple concepts, MeSH terms were to be combined using Boolean operators.

The Belgian participants entered an average of 12 free-text search terms to select a total of five MeSH terms. Of those 12 search terms, 51% were well-formulated and relevant to the information need.

The British test group needed ten free-text search terms to select a total of six MeSH terms. About 90% of the search terms were well-formulated and relevant to the information need.

The English-speaking group selected a slightly larger proportion (NS) of good MeSH terms (see table 2, B.2.).

In summary, the search terms entered by the English-speaking participants were of better quality than those formulated by the Dutch-speaking participants. Nevertheless, there is only a minor difference in quality of the selected MeSH terms.

- Mixed queries

Although they were instructed to use MeSH terms, some participants also used free-text terms in their queries. The Dutch-speaking students used more free-text search terms in their queries than the English speaking students (see table 2, C.1.), but combined them more often with MeSH terms using the Boolean operator *OR* (see table 2, C.2.).

- Error types

There are no significant differences between the English-speaking and Dutch-speaking participants in the types of errors they make, except for the use of incorrect free-text terms in their queries. Incorrect free-text (non-MeSH) terms include spelling and translation errors, as well as irrelevant terms. The queries submitted by the Dutch-

speaking participants contained a higher number of such errors (see table 2, D), which can be explained by the simple fact that they used more free-text terms.

3.2.2. Outcome indicators

The result lists yielded by the queries our participants submitted contained four relevant citations on average, i.e. their potential recall was four. The queries submitted by the Belgian participants had a mean potential recall of five (*Mdn*= 3 (IQR 0-8), *Max*= 21), those submitted by the British participants two (*Mdn*= 2 (IQR 0-4), *Max*= 8). Although the difference is not significant (see table 2, Outcome evaluation), there is a trend indicating that the Dutch-speaking students' queries were generally of better quality than the queries submitted by the British participants.

3.3. Analysis of relevance judgment

3.3.1. Process indicators

There was no significant difference in total evaluation times (see table 3, Process evaluation).

3.3.2. Outcome indicators

- **Absolute recall**

The participants in our test selected two relevant citations, i.e. citations that were also in the gold standard, on average. The Dutch-speaking test group selected three relevant citations (*Mdn*=2 (IQR 0-5)), whereas the English-speaking students selected only one relevant citation (*Mdn*= 1 (IQR 0-2)). Although it is clear that the Dutch-speaking participants performed better, a Mann-Whitney test showed that this difference in absolute recall is not significant (see table 2).

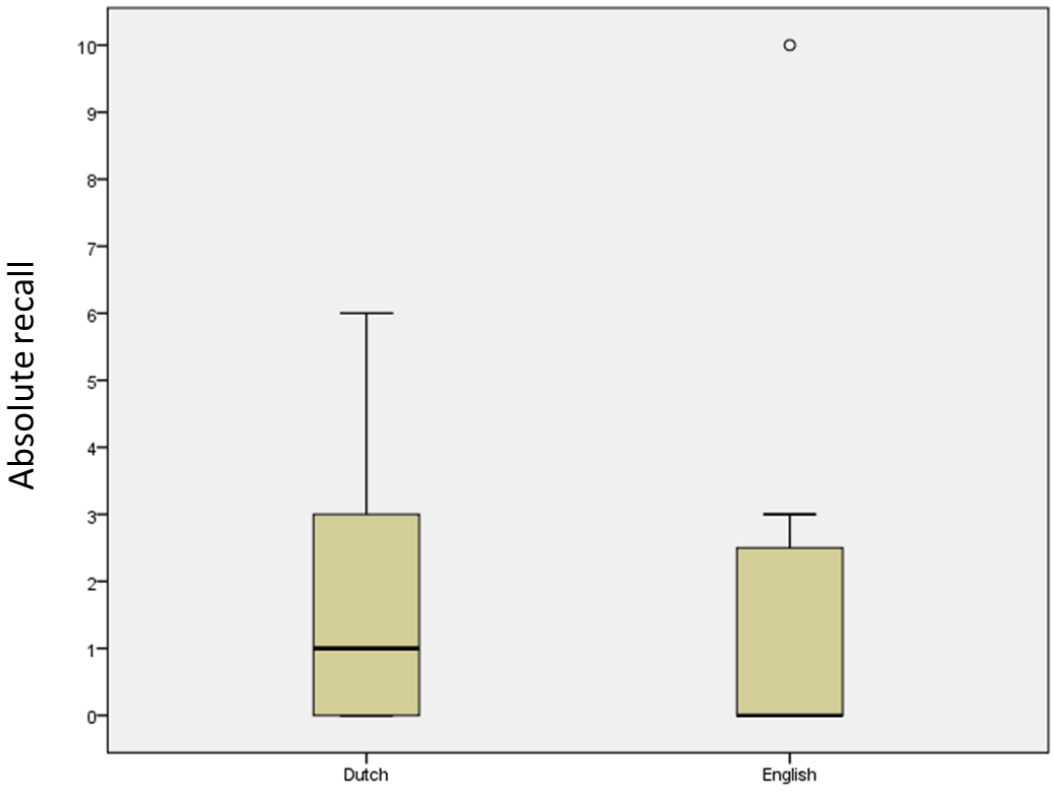


Figure 1: Box plot showing absolute recall in the two test groups.

A more detailed analysis of the results shows that the Belgian data for this variable are more dispersed, with more outliers (scores range between zero and 13, see figure 1) whereas the data in the British test group are more concentrated around the mean, ranging between zero and 4.

Table 3: Relevance judgment

	Dutch (n=40)	English (n=21)	Statistical test
PROCESS EVALUATION			
Average time spent on relevance judgment (minutes)	5.22	4.07	*U*= 332.000; *z*= -1.336;NS
OUTCOME EVALUATION			
1. Absolute recall (all components)	3	1	*U* = 325.50; *z* = -1.486; *p*=.137
2. Correlation coefficient between potential and actual recall (Spearman correlation)	.917	.651	*U*= 242.50; *z*= -2.727; *p*= .006
3. Weighted recall	44	21	*U* = 277.50; *z* = -2.170; *p*=.030

- **Correlation between potential and actual recall score**

A Spearman correlation test showed stronger correlations in the Dutch-speaking group when compared to the English-speaking group (see table 3, Outcome evaluation, 2). The difference in correlation between the two groups was significant, indicating that the Dutch-speaking participants' relevance judgment was better.

- **Weighted recall**

The Dutch-speaking test group achieved a mean weighted recall score of 44 (*Mdn*= 35 (IQR 12-62)), whereas the British students achieved a weighted recall score of 21 (*Mdn*= 14 (IQR 1-28)). The difference in weighted recall between the two groups is significant (see table 3, Outcome evaluation, 3), which means that the citations the Dutch-speaking participants selected contain more information that can help them solve the search question. Hence, their relevance judgment is better.

4. Discussion

To our knowledge, this is the first study to explore the impact of language on search quality empirically, based on log files of queries and output, and a qualitative analysis of the search process.

4.1. Main findings

The British participants were at an advantage during this test as they conducted the PubMed search in their own native language. The Dutch-speaking participants, however, were relatively well-trained in the use of PubMed and MeSH when compared to the English-speaking group. Although the search process of the latter was more fluent, with higher concept coverage and higher search term quality, the information gain in the Dutch-speaking participants' selection was significantly higher. We had expected a significant difference, but one that was the adverse of the result that we obtained. This means that the disadvantage the Dutch-speaking students had of searching in a non-native language was compensated for by their experience with the search system.

4.2. Strengths of the study

The main strength of this work resides in the fact that we studied a combination of the impact of language and PubMed experience on retrieval, focusing on query formulation as well as on relevance judgment. Each of these stages is analyzed according to both process and outcome indicators.

We refined measurements used to analyze outcome: next to absolute recall (a rather rigid measurement given its binary nature), we used weighted recall to calculate the informative value of the students' selection. Weighted recall is a more balanced and fine-grained measure to assess relevance and information gain. This work sheds some light on the performance of novice end-users with either no system experience or some formal bibliographic training and system experience.

The Dutch-speaking group achieved higher potential recall, despite their struggle to find correct English terms. However, as Jenuwine and Floyd (2004) argue, subject and text-word searches complement each other "and should be used together for maximal retrieval". The Belgian students submitted a significantly higher number of "mixed" queries. This combined strategy enhances their recall, despite a higher number of incorrect free-text terms in their queries.

4.3. Limitations

A limitation to this study is the limited sample size, which resulted in a lack of power and failure to show statistical significance for relevant trends (Bèta-error).

As we already mentioned in a previous paper (Vanopstal et al., 2012), some decisions or actions in the search process may be linked to different levels of intelligence. This, however, is not taken into account in the present study.

The success of a PubMed search is determined by several components , such as the searching skills of the participants, their ability to distinguish between relevant and irrelevant documents, intelligence and the accuracy of the system when it matches the query against the indexing terms assigned to the documents. As the focus of the present study is on the end-user perspective and not on system design, we assume that the

system's accuracy is perfect. We did not take into consideration different levels of intelligence, which may be considered as a limitation to this study.

5. Conclusions

We conducted a bibliographic retrieval experiment with two test groups of master's nursing students: a native English-speaking and a Dutch-speaking group. They were given a specific information task, and were instructed to search for relevant citations using MeSH terms in PubMed.

Despite their linguistic disadvantage, the Dutch-speaking students in our test achieved higher overall information gain, which we measured by calculating weighted recall. Moreover, the correlation between potential and absolute recall was stronger in the Dutch-speaking group, indicating that they were better at distinguishing between relevant and irrelevant citations. This may be attributed to their experience with the search engine and with literature searching in general.

We can conclude that non-native English-speaking searchers have a disadvantage, which, however, can be compensated for by thorough training of searching skills in general, and of the use of MeSH terms, where necessary in combination with free-text terms. Nevertheless, language support in the form of translated MeSH terms is likely to make the query formulation process more fluent.

6. Future work

Our study showed that the Dutch-speaking participants experienced some difficulties during the query formulation process, especially when they had to translate the search question into free-text search terms. It would therefore be interesting to set up a test in which the impact of language support in the form of translated MeSH terms (Buysschaert, 2006) is tested.

The methodology developed in this work can also be applied to research into the quality of medical registration and the impact of the use of multilingual end-user terminology on the performance and semantic interoperability of E-health systems.

7. Acknowledgements

We would like to thank the Nursing (and Midwifery) departments at Antwerp University and Nottingham University for allowing us to conduct the experiment. We also express our gratitude to Prof. Dr. Monique Elseviers, Dr. Philip Clissett and Dr. Stacey Johnson, who helped us to find candidates for our test. Thanks are due to Joke Coussement of the Centre for Health Services and Nursing Research, KU Leuven, who gave helpful hints in formulating the search question.

References

Anne, A., Bagayoko, C. O., & Fontelo, P. (2010). Évaluation de BabelMeSH en français. *IRBM, 31*(3), 170-174. doi: http://dx.doi.org/10.1016/j.irbm.2009.06.006

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

Dogan, R. I., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database (Oxford), 2009*, bap018. doi: 10.1093/database/bap018

Fontelo, P., Liu, F., Leon, S., Anne, A., & Ackerman, M. (2007). PICO Linguist and BabelMeSH: development and partial evaluation of evidence-based multilanguage search tools for MEDLINE/PubMed. *Stud Health Technol Inform, 129*(Pt 1), 817-821.

Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: analysis of a typical day's query log. *J Am Med Inform Assoc, 14*(2), 212-220.

Hoogendam, A., Stalenhoef, A. F., Robbe, P. F., & Overbeke, A. J. (2008). Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. *BMC Med Inform Decis Mak., 8*, 42. doi: 10.1186/1472-6947-8-42

Jenuwine, E. S., & Floyd, J. A. (2004). Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. *J Med Libr Assoc, 92*(3), 349-353.

Liu, F., Fontelo, P., & Ackerman, M. (2006). *BabelMeSH: Development of a Cross-Language Tool for MEDLINE/PubMed.* Paper presented at the AMIA Annu Symp Proc, Washington.

Lu, Z., Wilbur, W. J., McEntyre, J. R., Iskhakov, A., & Szilagyi, L. (2009). Finding query suggestions for PubMed. *AMIA Annu Symp Proc, 2009*, 396-400.

Park, T.K. (1994). Toward a theory of user-based relevance: a call for a new paradigm of inquiry. *JASIST, 45*(3), 135-141.

Silverstein, Craig, Marais, Hannes, Henzinger, Monika, & Moricz, Michael. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6-12. doi: 10.1145/331403.331405

Swanson, D.R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Libr Q, 56*, 389-398.

Thirion, Benoit, Pereira, Susanne, Névéol, Aurélie, Dahamna, Badisse, & Darmoni, Stéphan J. (2007). French MeSH Browser: a cross-language tool to access MEDLINE/PubMed. *AMIA Annu Symp Proc*, 1132.

Vanopstal, Klaar, Buysschaert, Joost, Laureys, Godelieve, & Vander Stichele, Robert. (2013). Lost in Pubmed. Factors influencing the success of medical information retrieval. *Expert Syst Appl, 40*(10), 4106-4114.

Vanopstal, Klaar, Vander Stichele, Robert, Laureys, Godelieve, & Buysschaert, Joost. (2012). PubMed Searches by Dutch-Speaking Nursing Students: The Impact of Language and System Experience. *JASIST, 63*(8), 1538-1552.

List of figures

List of tables

3

Discussion and conclusions

1. Part 1: The terminology of information retrieval

1.1. Research questions

The research questions in this part were the following:

1. Which definitions of glossary, taxonomy, controlled vocabulary, thesaurus, ontology and topic maps can be found in the literature? Are they consistent?
2. What causes inconsistencies in the use of these terms?
3. Is it possible to formulate a domain-independent definition for the concepts "thesaurus" and "controlled vocabulary"? How do the Medical Subject Headings relate to this definition?

For each of the terms used to designate vocabularies for information retrieval, the literature gives multiple diverging definitions which are sometimes incompatible. We assembled a corpus of definitions from the literature, which allowed us to study the use of the terms in different contexts. An analysis of these definitions showed that the polysemous and sometimes even incorrect use of the terms taxonomy, thesaurus and ontology was caused by historical and interdomain shifts. Hence, it was not possible to formulate consensus definitions for each of the terms in this study.

The terms glossary, thesaurus and controlled vocabulary were first used within the field of linguistics. When they were later adopted in the fields of knowledge management and/or bibliographic retrieval, their meaning shifted, causing confusion and incorrect use of the terms. In the first part of this dissertation, we tried to provide a solution for this confusion by listing a definition for each of the terms and for each of the fields in which they are used: linguistics, knowledge management and/or bibliographic retrieval. Figure 1 below shows how adding the field of knowledge as an extra dimension helped to provide clear and unambiguous definitions.

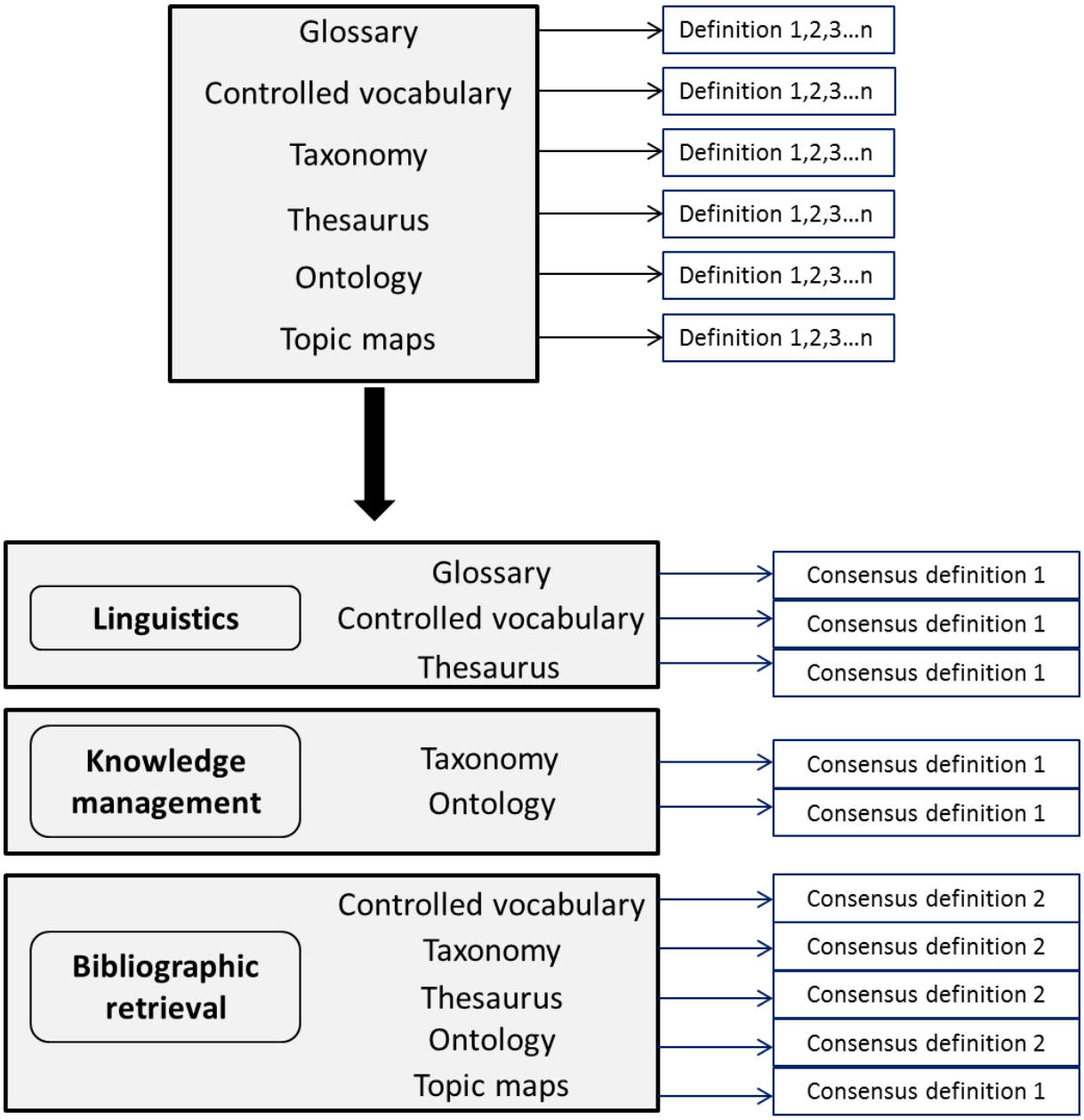


Figure 1: A layered schematic representation of the definitions in Chapter I

In view of the second part of this dissertation, controlled vocabularies and thesauri for information retrieval were of particular interest. Both thesauri and controlled vocabularies can be used as a purely linguistic tool; they then have a prescriptive character and are aimed at creating consistency in language use by making a distinction between preferred and non-preferred terms. We define controlled vocabulary in the field of linguistics as "a set of terms which provides a standard language for a very specific domain". In the same context, we define a thesaurus as "a rich set of terms which provides a standard language for a field of knowledge". The difference between a

controlled vocabulary and a thesaurus in the field of linguistics can be found in the size: a controlled vocabulary is usually limited to a (sub)domain of knowledge whereas a thesaurus is a "treasure of words", with a broader scope.

Next to being a linguistic tool, a controlled vocabulary can also serve as a basis for information retrieval thesauri and other information retrieval vocabularies. Thesauri for information retrieval are controlled vocabularies with the additional specification of hierarchical, associative and equivalence relationships. We define controlled vocabularies for information retrieval as follows: "a list of preferred terms and their non-preferred variants". For thesauri, we adopt the ISO definition for thesauri: "a controlled vocabulary, which is usually organized hierarchically and which includes standardized, a priori, hierarchical, associative and equivalence relationships between concepts."

The Medical Subject Headings are compliant with the definition of a thesaurus given by ISO 2788; however, subject headings are pre-coordinated, which is atypical for thesauri. The MeSH browser visualizes the hierarchical structure of the vocabulary, and provides its users with related terms and a scope note. When used in information retrieval, the hierarchical relationship in this vocabulary enables term explosion, whereas synonyms (non-preferred terms) are mapped to their preferred terms, thus enabling more focused searching.

For the definitions of the other vocabularies, and for a detailed discussion of the designations of other medical vocabularies, we refer to the article.

1.2. Update of the research data

Since the publication of the article "*Vocabularies and retrieval tools in biomedicine: disentangling the terminological knot*", a new standard has been published: the joint British-American standard ISO 25964 (International Standards Organization, 2011, 2013). This new standard replaces ISO 2788 and ISO 5964, the standards for monolingual and multilingual thesauri, respectively. It consists of two parts: *Thesauri for information retrieval (International Standards Organization, 2011)* and *Interoperability with other vocabularies (International Standards Organization, 2013)*. Part 1 describes the aspects of developing and maintaining both monolingual and multilingual thesauri. It also

provides a data model and an XML schema for the exchange of data. Part 2 discusses interoperability issues and gives recommendations for mapping between thesauri and with other vocabularies for information retrieval.

The main novelties in this standard are:

- a shift in focus from paper thesauri to computer and information retrieval applications;
- a clearly-defined concept-oriënted approach ;
- a model for interoperability with subject headings and other vocabularies for information retrieval.

Although ISO 2788 professed to be concept-based, it described relationships between terms rather than concepts. The new standard offers a more convincing concept-oriented data model, which should also enhance the interoperability of thesauri.

One important component in this new standard in the light of this dissertation is that it provides a structure for interoperability between thesauri and subject headings (such as the MeSH vocabulary used in our PubMed experiment). The terms in thesauri are usually used in post-coordination, i.e. they contain individual, single concepts which can be combined into compound concepts by searchers and indexers. MeSH concepts are pre-coordinated, which is a distinguishing feature of subject headings. ISO 25964-2 provides guidelines for handling pre-coordination, enabling mappings between thesauri and subject headings.

As explained above, the focus of this study was on thesauri, and more specifically on MesH. In the second part of this dissertation, we studied the impact of experience with and use of MeSH and PubMed on the search process and results in Dutch-speaking and English-speaking nursing students.

2. Part 2: The role of terminology in medical literature searching

2.1. Research questions

The quest to design an ideal information retrieval system has been ongoing for the past 50 years (Sanderson & Croft, 2012). In light of this quest, most studies have focused on the architecture of the systems, and on ranking algorithms. End-users, if they are involved in the evaluation at all, are usually considered as a medium to evaluate the system rather than being the focus of research. In this dissertation medical information retrieval was studied from the end-user perspective.

The research questions to be answered in part 2 were:

1. Do English language skills in Dutch-speaking users of PubMed affect the efficiency of their literature searches?
2. How can we distinguish between best and worst performers? Can their characteristics be linked to the errors they make when they search PubMed?
3. To what extent do language skills and searching skills in native and non-native speakers of English contribute to the outcome of literature searches in PubMed?

We conducted a retrieval experiment, the resulting data of which were used in three separate analyses:

1. A contrastive analysis of need articulation, query formulation and relevance assessment in Dutch-speaking bachelor's and master's nursing students, with a focus on the impact of English language skills assessed through a language test. (Chapter II)
2. A contrastive error analysis of the queries constructed by the best and worst performers. (Chapter III)
3. A contrastive analysis of the search process and outcome of Dutch-speaking versus native English-speaking master's nursing students, with a focus on the interaction between English language skills and system experience. (Chapter IV)

2.2. Description of the search process

We developed an information retrieval model for non-native searchers on the basis of Sutcliffe and Ennis' (1998) findings. This model describes four main stages: problem identification, need articulation, query formulation and relevance judgment. Different kinds of translation take place on different levels in this model.

As the participants in our test started from a pre-formulated question, they did not have to go through the first stage, viz. problem identification. Consequently, this stage was left out of our analyses.

Need articulation, the second stage, involves parsing of the problem, which is formulated in natural language, into several concepts. Although the search terms that were formulated in the next step give us an idea about which concepts were identified, need articulation itself is implicit in this test. It is a mental process which involves intralingual (Jakobson, 1981) or intrasystemic (Torop, 2002) translation. Jakobson defines intralingual translation as "the interpretation of verbal signs by means of other signs of the same language". In our test case, this is the translation of the Dutch search question into (Dutch) concepts.

The third stage, query formulation, consists of two steps: search term formulation and MeSH term selection. In the search term formulation step, the concepts identified during need articulation are translated into English search terms. Although concepts are supposed to be language-independent, we hypothesize that there is some kind of translation of "Dutch" concepts into English search terms. We assessed the quality of the search terms formulated by our test participants, and found that this quality did not have a direct impact on recall. However, badly formulated search terms were a cause for non-coverage of concepts with MeSH terms. The other two causes were non-identification (error resulting from stage two) and the failure to select the correct MeSH term (see below). Once the search terms have been entered into the MeSH module of PubMed, the searcher has to select the appropriate MeSH terms. This can also be designated as intralingual translation: the translation of English search terms into English MeSH terms. We assessed the quality of the MeSH terms selected by our test participants, and we found that this quality had a direct impact on the number of

relevant results the query returned (potential recall). MeSH terms can be a very useful aid when searching PubMed, even if they are only available in English: whereas about 50% of the search terms were incorrect, the intermediate step of MeSH term selection resulted in an error rate reduction of 25%. We refer to this phenomenon as "the corrective effect of (subject searching with) MeSH".

The fourth stage, relevance judgment, involves skimming the list of results and selecting relevant citations. Searchers now have to map the results of the search to their information need and select the citations that are relevant to the search question. As the titles and abstract of the citations are written in English, we assume that an English to Dutch translation process is also involved in this stage. The quality of relevance judgment was studied in terms of relevance judgment times, the number of missed citations, absolute recall and its correlation with potential recall.

This dissertation focuses mainly on query formulation and on relevance judgment.

2.3. Query formulation

2.3.1. Process indicators

- *Error analysis*

An error analysis of the queries submitted by all participants during the literature search task resulted in the identification of eight error types: (in order of descending frequency) underspecification, irrelevant MeSH terms, incorrect free-text terms, overspecification, incorrect use of Boolean operators, syntax errors, spelling errors, and incorrect translations.

Three errors had a direct impact on the number of relevant results returned by a query (potential recall): irrelevant MeSH terms, underspecification and incorrect use of Boolean operators. About 80% of the queries containing one of those errors led to zero potential recall.

Most queries (81%) contained one or more errors. However, some of these queries nevertheless yielded relevant results, indicating that minor errors (mainly incorrect free-text terms and overspecification) do not always render queries useless.

- *Comparison of the query formulation process in different performer types*

We divided our test groups in three groups on the basis of their performance: worst performers did not select any relevant citations, average performers selected one or two, and the best performers selected three or more relevant citations. We performed an error analysis across these performer types, which allowed us to describe the search behaviour in these groups.

This error analysis led to the conclusion that the best performers did not necessarily make fewer errors (except for underspecification errors and the incorrect use of Boolean operators), rather they were better at correcting errors. This means that the worst performers made errors in one query, and subsequently submitted a query that either contained the same error, or another one. The best performers, on the other hand, succeeded in correcting incorrect MeSH terms, incorrect Boolean operators and underspecified queries (in 60%, 83% and 60% of the cases, respectively). The correction of overspecified queries and incorrect free-text terms seemed to be more difficult than the correction of the other error types, even in the best performer group.

The best performers formulated better queries with a potential recall of 8 relevant citations (versus 2 in the worst performer group), which gave them a head start in the relevance judgment stage.

2.3.2. Outcome indicators

We introduced potential recall as a new measure to assess the quality or effectiveness of a query. It indicates how many relevant citations the query yielded. We only took into consideration the citations the participants actually viewed in order to calculate this score. If a participant for instance only looked at the first 40 citations, we counted how many relevant citations this list of 40 citations contained. The total potential recall score (sum of the potential recall of all queries submitted by one participant) ranged between 0 and 21 relevant citations.

On the basis of potential recall, we can divide the queries issued during our test into adequate ($R_{pot} > 0$) and inadequate ($R_{pot} = 0$) queries. A total of 44% of the queries were adequate; the rest of the queries either returned no results, or only irrelevant ones.

High potential recall did not necessarily lead to high absolute recall: 47% of the queries with positive potential recall did not lead to the selection of any relevant citations.

2.4. Relevance judgment

2.4.1. Process indicators

The second analysis showed that the best performers spent less time on querying and more time on relevance judgment. In other words, they reached a more productive balance between the two most important stages of information retrieval.

2.4.2. Outcome indicators

We assessed the outcome of the relevance judgment stage in terms of different types of recall and – to a lesser extent – precision.

- *Proportional and absolute recall*

Recall was initially calculated on the basis of a gold standard: we calculated the number of relevant citations in the participants' selection as a proportion of the number of gold standard citations. However, using absolute numbers (e.g. "4 relevant citations") proved to be much more illustrative than the use of percentages (e.g. "recall of 6.25%"). Consequently, we decided to only mention absolute recall,which ranged between 0 and 13).

- *Weighted recall*

We introduced weighted recall as an alternative to proportional and absolute recall. Weighted recall is more fine-grained and less rigid than proportional and absolute recall in that it measures the information gain in the participants' selection. The search question contained five main components: falls, elderly, long-term care, multifactorial, and prevention. In the calculation of proportional and absolute recall, citations which lacked one of these components were considered to be irrelevant, whereas the underlying idea of weighted recall is that these citations may also contribute to the fulfillment of the information need.

We analyzed the citations selected by our test participants and counted how many of these components were present. A heavier weight was assigned to the more important components of the search question: the crucial components of *falls* and *prevention* received a weight of two, the other components were assigned a weight of one. A weighted recall score was calculated for each participant by adding up the scores for each individual citation in their selection. The total weighted recall score ranged between 0 (indicating that the participant did not select any citations) and 186.

- *Precision*

Precision was calculated as the proportion of relevant citations in the participants' selection. It ranged between 0 and 1.

- *Correlation between potential and absolute recall as an indication of relevance judgment*

We consider the correlation between potential and absolute recall as an indication of relevance judgment quality. Creating a good query with high potential recall is an accomplishment in itself; however, it is then a matter of distinguishing the relevant citations from the irrelevant ones. Strong correlations between potential and absolute recall indicate that the searcher succeeded in doing exactly that.

An analysis of the results of Dutch-speaking bachelor's and master's students showed that relevance judgment (measured by the correlation between potential and absolute recall[1]) was significantly better in participants who achieved the highest levels (C1 or C2) on the language test[2]. However, the third analysis provides evidence for better relevance judgment in the more experienced, Dutch-speaking participants than in the native English-speaking participants. This indicates that there are other factors than language skills which play a role in efficient relevance judgment, such as general research skills, or experience with reading scientific literature.

[1] vocabulary ≥ C1: r_s= .897, p= .000
vocabulary < C1: r_s= .719, p= .000
reading ≥ C1: r_s= .953, p= .000
reading < C1: r_s= .753, p= .000

[2] vocabulary: Mann-Whitney U= 204.000, z= -3.055, p= .002
reading: Mann-Whitney U= 227.000, z= -2.240, p= .025

2.5. Impact of English language skills

2.5.1. Comparison based on the results of the DIALANG language test

The first analysis in this dissertation (among Dutch-speaking nursing students) showed that there is a positive correlation between English language skills in Dutch-speaking PubMed users and their recall.

The same analysis showed that language skills have an impact on several factors in the query formulation process: participants with better language skills formulated a higher proportion of good search terms, hesitated less during the search, and had fewer doubts about the spelling of their search terms.

This analysis did not reveal a significant correlation between language skills and the quality of MeSH terms.

2.5.2. Comparison of best and worst performers

The best-worst performer analysis, which included all participants (master's and bachelor's students, Dutch-speaking and native English-speaking) did not show a relation between language skills and performer type. However, in section 2.5.1, we did conclude that there was a positive correlation between recall and language skills. If we only consider the Dutch-speaking participants we see that the best performers did score significantly higher on the reading test than the worst performers.

2.5.3. Comparison based on mother tongue

In the first analysis, we hypothesized that searching in one's own mother tongue would have an influence on concept coverage (see p.81). The third analysis showed that the native English-speaking students did achieve slightly higher concept coverage. This means that they succeeded in identifying MeSH terms for most of the components in the information need.

The English-speaking participants had less difficulty in translating the concepts of the search question into search terms, as they were not hampered by the Dutch to English translation step. However, as we have shown in an earlier study (Chapter II), the quality of these search terms has little impact on the outcome of the search process, as they

were translated into MeSH terms. The Dutch-speaking group, who were more familiar with the use of MeSH terms, benefited more from the corrective effect of MeSH terms with an error reduction of 25%. This effect was not found in the query formulation process of the native English-speaking participants.

There were no significant differences in the types of errors the English-speaking and Dutch-speaking participants made, except for the use of incorrect free-text terms. The Dutch-speaking participants used a higher number of free-text terms in their queries. It is not clear whether this is due to the fact that they simply could not find the right MeSH term, or to their experience with PubMed, which made them more "adventurous". Although there were no significant differences in the error types made by Dutch-speaking or English-speaking participants, we did see that participants who achieved a C1 level or higher on the vocabulary test, formulated a significantly higher number of error-free queries during the literature search task (new analysis; U= 950.00, z= -1.983, p= .047).

In summary, we can state that English language skills have an impact on the fluency of the query formulation stage. Our data did not provide evidence that language skills also resulted in queries that returned a higher number of relevant citations

2.6. Impact of searching skills

We define searching skills as the participants' prior experience with PubMed, facility with the interface, and the ability to use MeSH in an appropriate way.

There are several factors in our data indicating that good English language skills do not guarantee a successful PubMed search: the fact that there were only 6 native English-speaking participants in the best performer group shows that they also had difficulties in conducting an effective search. Moreover, the contrastive analysis between Dutch-speaking master's students and native English-speaking master's students (Chapter IV) showed that the Dutch-speaking participants outperformed the English-speaking participants by compensating for their relatively weaker language skills with better searching skills.

The first analysis, which includes only the Dutch-speaking participants, showed that the selection of MeSH terms is influenced by prior experience with PubMed and MeSH. The third analysis (Chapter IV), however, showed that the native English-speaking participants were slightly better at MeSH term selection than the more experienced Dutch-speaking searchers. This indicates that the selection of MeSH terms is influenced by both language skills and system experience, and it implicates that especially inexperienced searchers with weak English language skills would benefit from the incorporation of translated MeSH terms into PubMed.

Our final analysis (Chapter IV) indicated that the correlation between potential and actual recall was stronger in the more experienced searchers, even though they were non-native speakers of English. This means that their relevance judgment was of higher quality. We tested this finding in the group of the first analysis, and came to the same conclusion: relevance judgment was better in the more experienced searchers than it was in the group of novices (Mann-Whitney U= 328.00, z= -3.400, p= .001).

This higher-quality relevance judgment in the Dutch-speaking master's students resulted in significantly higher information gain (measured by weighted recall) than in the native English-speaking group.

In summary, the adequate use of MeSH and relevance judgment is especially influenced by searching skills.

2.7. Balance between language skills and system experience

The results of this research suggest that non-native speakers of English who search PubMed can compensate for their language handicap with more advanced searching skills. English language skills in non-native speakers of English do have an impact on the outcome of a PubMed search, but the Dutch-speaking master's students' performance shows that more factors are involved than language alone.

2.8. Suggestions for further research

We hypothesize that language support in the form of MeSH terms may facilitate the search process of non-native speakers of English. The more experienced searchers in our test compensated for the fact that they had to search in a non-native language with their more advanced searching skills. This suggests that a translation of the MeSH terms (Buysschaert, 2006) may benefit Dutch-speaking novice searchers the most.

In order to investigate this hypothesis, an experiment would be needed with MeSH translations in groups with different levels of PubMed experience to see what the impact of language support is and how much it contributes to better searching at different levels of PubMed experience. The translation can also be integrated for relevance judgment: listing the translated MeSH terms that are assigned to each citation can be helpful do decide whether an article is relevant to the information need or not.

Defective concept identification was one of the main causes for non-coverage of concepts in the queries of our test participants. It would, for instance, be interesting to isolate the query formulating step from the rest of the search process, and have students construct a query in English, with a control group who construct a query in Dutch. A think-aloud protocol would allow us to study problems related to concept identification and the translation of these concepts into search terms. A think-aloud protocol would also allow us to study the use of free-text terms in more detail.

The relevance judgment step can be studied by giving a group of students the same list of citations from which they have to select the relevant ones. This would eliminate the effect of bad queries, so that relevance judgment can be studied on its own.

Research on how to support the selection of relevance judgment (e.g. simplified abstracts or wikification (He et al., 2011)) would also provide insight in methods to facilitate the retrieval process for non-native English-speaking users of PubMed.

References

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

He, J., Rijke, de M., & Sevenster, M. (2011). *Generating links to background knowledge for medical content*. Paper presented at the Second International Workshop on Web Science and Information Exchange in the Medical Web (MedEX 2011). http://dare.uva.nl/record/420713

International Standards Organization. (2011). ISO 25964-1:2011, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva.

International Standards Organization. (2013). ISO 25964-2:Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies. Geneva.

Jakobson, Roman. (1981). *On linguistic aspects of translation*. The Hague: Mouton.

Sanderson, M., & Croft, W.B. (2012). The History of Information Retrieval Research. *Proceedings of the IEEE, 100*(Special Centennial Issue), 1444-1451.

Sutcliffe, A., & Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers, 10*(1998), 321-351.

Torop, Peeter. (2002). Translation as translating as culture. *Sign Systems Studies, 30*(2), 593-605. doi: citeulike-article-id:345897

4

Appendix

A. Summary

This dissertation consists of two parts: a literature study (Part 1) and three experimental studies among different populations of nursing students (Dutch-speaking bachelor's students; Dutch-speaking master's students; English-speaking bachelor's students; English-speaking master's students) (Part 2).

The first part presents a theoretical study of vocabularies for medical information retrieval, and the way they are defined in the literature. The starting point of this study was MeSH (Medical Subject Headings), a vocabulary used to index and retrieve information. This vocabulary was used in the retrieval experiment in Part 2.

We assembled a corpus of definitions from the literature for the terms *thesaurus*, *controlled vocabulary*, *glossary*, *ontology*, *taxonomy*, and *topic maps*. This corpus allowed us to study the use of these terms in different contexts. An analysis showed that the polysemous and sometimes even incorrect use of the terms *taxonomy*, *thesaurus* and *ontology* was caused by historical and interdomain shifts. We tried to provide a solution for this confusion by listing a definition for each of the terms and for each of the fields in which they are used: linguistics, knowledge management and/or bibliographic retrieval. We concluded that MeSH is a thesaurus with the syntax of subject headings.

The second part elaborates on medical information retrieval and the difficulties nursing students experience when they search for medical information in PubMed/MEDLINE. It consists of three separate analyses of data assembled during a retrieval experiment with Dutch-speaking and native English-speaking bachelor's and master's nursing students:

1. A contrastive analysis of need articulation, query formulation and relevance judgment in Dutch-speaking bachelor's and master's nursing students, with a focus on the impact of English language skills.

2. A contrastive error analysis of the queries constructed by the best and worst performers. For this study, we analyzed the queries of all four test groups.

3. A contrastive analysis of the search process and outcome in Dutch-speaking versus native English-speaking master's nursing students, focusing on the interaction between English language skills and system experience.

In the first analysis, we studied several factors in the query formulation process (e.g. quality of search terms and MeSH terms, concept coverage, hesitations, etc.) and found that the English language skills in Dutch-speaking searchers especially had an impact on the fluency of the query formulation step. The more experienced searchers were better at selecting the appropriate MeSH terms for their search, and at distinguishing relevant citations from irrelevant ones. This is probably due to their generally more advanced research skills.

The main difference in search behavior between best and worst performers lies in the correction of errors: the best performers were better at correcting their errors, except when they concerned overspecification and the use of incorrect free text terms. Our data showed a relation between the English language skills in the Dutch-speaking participants and their distribution over the performer types.

A contrastive analysis between Dutch-speaking and native English-speaking nursing students showed that the query formulation process was more fluent in the native speakers. Nevertheless, they did not achieve better results or higher information gain. On the contrary: the Dutch-speaking, more experienced students achieved higher weighted recall, and our analysis showed that they were better at relevance judgment than the English-speaking students, who were novice searchers.

In conclusion, language skills have an impact on the fluency of the search process, but the overall success of the search depends on other factors as well, such as searching skills and general research skills.

B. Samenvatting

Dit proefschrift gaat over medische information retrieval en bestaat uit twee delen: een literatuurstudie (Deel 1) en drie experimentele studies uitgevoerd bij verschillende groepen verpleegkundestudenten (Nederlandstalige bachelor- en masterstudenten, en Engelstalige bachelor- en masterstudenten) (Deel 2).

Het eerste deel behandelt verschillende soorten vocabularia die gebruikt worden bij medische informatieopzoekingen. Concreet worden de volgende termen bestudeerd: *thesaurus*, *gecontroleerd vocabularium*, *glossarium*, *taxonomie*, *ontologie* en *topic maps*. Het uitgangspunt voor deze studie was MeSH (Medical Subject Headings), een vocabularium dat bij de ontsluiting van biomedische informatie wordt gebruikt om teksten te indexeren en later ook terug te vinden. De Medical Subject Headings werden ook gebruikt in de experimenten voor Deel 2.

Voor elk van de termen in kwestie werden definities verzameld uit de literatuur. Dit liet ons toe het gebruik van de termen in verschillende contexten te bestuderen. Hieruit bleek dat polysemie en het soms incorrecte gebruik van de termen *taxonomie*, *thesaurus* en *ontologie* veroorzaakt worden door historische verschuivingen en het overnemen van de termen door andere wetenschappelijke disciplines. In deze studie worden daarom eenduidige definities voorgesteld voor elk van de termen op basis van het vakgebied waarin ze worden gebruikt: linguïstiek, kennismanagement en/of bibliografische retrieval. Enkele voorbeelden van vocabularia uit het medische domein werden vergeleken met deze definities. Hieruit kunnen we besluiten dat MeSH een thesaurus voor bibliografische retrieval is, met de syntax van subject headings.

In het tweede deel van dit proefschrift ligt de focus op retrieval van medische informatie en de moeilijkheden die studenten verpleegkunde ondervinden wanneer zij PubMed/MEDLINE gebruiken. Dit deel is gebaseerd op een experiment waarbij Nederlandstalige en Engelstalige studenten verpleegkunde op zoek gingen naar specifieke medische informatie. Pre- en posttestvragenlijsten gaven ons meer

informatie over de achtergrond van de studenten. Dit experiment resulteerde in drie analyses:

1. Een contrastieve analyse van verschillende stadia in het zoekproces (need articulation, query formulation en relevance judgment) bij Nederlandstalige bachelor- en masterstudenten verpleegkunde. Bij deze analyse ligt de nadruk op de invloed van taal op het zoekproces en de resultaten daarvan.

2. Een contrastieve foutenanalyse waarbij het zoekgedrag van de beste en de slechtste zoekers werd onderzocht en vergeleken. Voor deze studie werden het zoekproces en de resultaten van alle testgroepen geanalyseerd.

3. Een contrastieve analyse van het zoekproces van Nederlandstalige en Engelstalige masterstudenten verpleegkunde. Hierbij werd vooral gekeken naar de interactie tussen taalvaardigheid en ervaring met het zoeksysteem.

In de eerste analyse werden verschillende factoren van het zoekproces bestudeerd. Het formuleren van een goede query is een complex proces waarin verschillende variabelen een belangrijke rol spelen. Zo werden naast de kwaliteit van de zoektermen bijvoorbeeld ook aarzelingen, de vertaling van de zoekvraag in concepten en het gebruik van MeSH-termen bestudeerd. Uit deze analyse kunnen we besluiten dat Engelse taalvaardigheid wel degelijk een invloed heeft bij het opzoeken van medische informatie, meer bepaald op de vlotheid waarmee query's worden geformuleerd. De meer ervaren gebruikers van de zoekmachine waren bedrevener in het gebruik van MeSH-termen en bovendien konden ze beter het onderscheid maken tussen relevante en irrelevante artikels voor deze zoekopdracht. Dit is waarschijnlijk te wijten aan hun vertrouwdheid met onderzoek in het algemeen.

De "beste zoekers" onderscheiden zich vooral van de "slechtste zoekers" door de manier waarop ze op hun eigen fouten reageren: hoewel zij ook fouten maakten, waren ze telkens in staat om deze te corrigeren. Enkel wanneer het om overspecificatie ging, of over het gebruik van incorrecte "vrije zoektermen" (i.p.v. MeSH-termen) bleken ook zij moeilijkheden te hebben om hun eigen fouten te verbeteren. Uit onze data bleek verder dat de meest taalvaardige studenten eerder in de groep van "beste zoekers" zaten, en

dat studenten die lager scoorden op de taaltest eerder tot de “slechtste zoekers” behoorden.

Een contrastieve analyse van het zoekproces van de Nederlandstalige en Engelstalige masterstudenten verpleegkunde toont aan dat de Engelstaligen minder moeilijkheden ondervonden bij het formuleren van query’s. Dit uit zich echter niet in betere zoekresultaten: ze behaalden geen hogere recallscore, noch had hun selectie een hogere informatieve waarde, of “information gain”. Integendeel, de Nederlandstaligen behaalden de hoogste scores en hun selectie van artikels had ook de hoogste informatieve waarde. Daarenboven blijkt uit onze analyse dat de Nederlandstaligen beter het onderscheid konden maken tussen relevante en irrelevante artikels dan de Engelstaligen. Dit is waarschijnlijk toe te schrijven aan een meer uitgebreide algemene ervaring met het opzoeken van informatie in vergelijking met de Engelstaligen, die slechts een beperkte ervaring hadden met het zoeksysteem en met information retrieval.

We kunnen uit dit onderzoek besluiten dat taalvaardigheid zeker een invloed heeft op het zoekproces en de vlotheid daarvan, maar dat het uiteindelijke welslagen van een zoekopdracht ook door andere factoren wordt beïnvloed, zoals ervaring met het zoeksysteem en algemene onderzoeksvaardigheden.

C. Pre- and posttest questionnaires (Dutch)

1. <u>Algemeen</u>:

Nummer computer	.....
Geslacht:	☐ **man** ☐ **vrouw**
Instelling:	.....
Opleiding en jaar:	.....
Datum afname:	.....

2. <u>Info en instructies</u>:

Beste student,

Bedankt om deel te nemen aan dit onderzoek! Voor je aan de test begint, neem je best rustig de volgende informatie door. Als je vragen hebt, aarzel vooral niet om ze te stellen.

<u>Instructies</u>

Deze test maakt deel uit van een doctoraatsonderzoek dat loopt aan de Hogeschool Gent rond de efficiëntie van Engelstalige (bio)medische zoeksystemen. De test zal zowel in een Nederlandstalige als in een Engelstalige groep worden afgenomen.

Op de volgende bladzijden vind je:

- een **vragenlijst** "Enkele vragen vooraf"
- een **zoekopdracht** in PubMed
- een **vragenlijst** “Enkele vragen achteraf”
- **instructies** voor de DIALANG taaltest

De test zal ongeveer 1 uur 15 minuten duren.

1. De vragenlijsten

De vragenlijsten bevatten 2 soorten vragen:

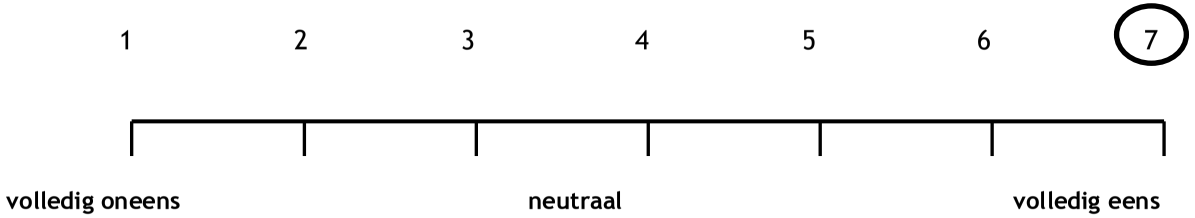
- Vragen waarbij je het juiste antwoord **aankruist**.
  vb. “Ik studeer aan de Universiteit Antwerpen”

  ☒ ja

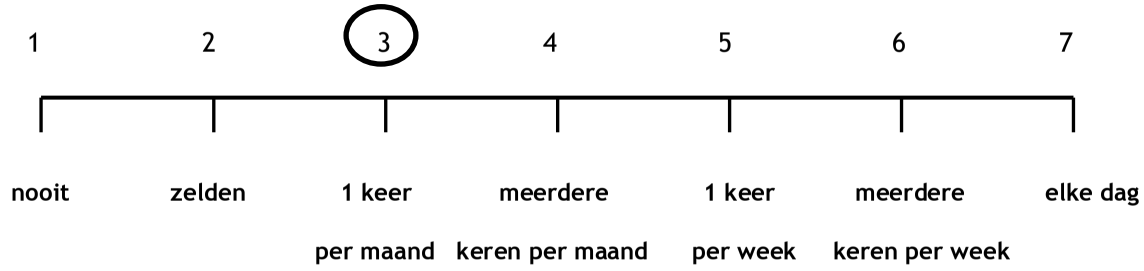
  ☐ neen

- Vragen waarbij je je mening of een frequentie op een **schaal** van 1 tot 7 weergeeft. Hierbij **omcirkel** je het antwoord dat het beste bij je mening aansluit.

  vb.1 "Ik vind deze test interessant."



  vb.2 "Hoe vaak ga je naar de cinema?"



Probeer zo **volledig** mogelijk te werken, zodat we nadien een betrouwbare beoordeling bij dit onderzoek kunnen opmaken.

Aarzel niet om de verantwoordelijke op de hoogte te brengen van eventuele **problemen**, hoe onbelangrijk ze ook mogen lijken. Op het einde van de vragenlijst heb je ook plaats om je **opmerkingen/commentaren/problemen** te rapporteren.

2. De zoekopdracht

Na het invullen van de vragenlijst krijg je een zoekopdracht die je zal uitvoeren met behulp van **PubMed** (geen andere bronnen)**.** Lees deze opdracht aandachtig en ga vervolgens op zoek naar de meest relevante artikels. Met behulp van het programma Morae zullen je handelingen op de computer worden geregistreerd, zodat wij achteraf kunnen bepalen waar zich eventuele moeilijkheden voordeden.

Bedankt voor je tijd!

3. Enkele vragen vooraf

Als je bepaalde vragen onduidelijk vindt, dan kan je dit melden of onderaan noteren bij "Opmerkingen".

	Vraag	Antwoord
Computervaardigheden	1. Hoe vaak werk je met een computer (PC/ laptop)?	1 — 2 — 3 — 4 — 5 — 6 — 7 **nooit** — **zelden** — **1 keer per maand** — **meerdere keren per maand** — **1 keer per week** — **meerdere keren per week** — **elke dag**
	2. Waarvoor gebruik je je computer het meest? Kruis de **3 meest toepasselijke** antwoorden aan.	☐ **tekstverwerking** ☐ **communicatie (e-mail, chat,...)** ☐ **analyse van gegevens** ☐ **zoeken van informatie op het internet** ☐ **programmeren van hardware of software** ☐ **downloaden van muziek en/of films** ☐ **andere: _____**

	3. Hoe vaak zoek je (algemene) informatie op het internet?	1 2 3 4 5 6 7 nooit zelden 1 keer per maand meerdere keren per maand 1 keer per week meerdere keren per week elke dag
	4. Wanneer je naar (bio)medische informatie zoekt op het internet, doe je dat dan in het Engels of eerder in het Nederlands?	☐ Engels ☐ Nederlands ☐ Even vaak in het Engels als in het Nederlands
	5. Wanneer je naar (bio)medische informatie zoekt, welke kanalen gebruik je dan? (meerdere antwoorden mogelijk)	☐ Google ☐ Biomedische databases, nl.: _____ ☐ Andere, nl.: _____
	6. Hoe schat je je computervaardigheden zelf in?	1 2 3 4 5 heel slecht slecht middelmatig goed uitstekend

Vertrouwdheid met PubMed	7. Heb je al kennis gemaakt met PubMed? (vb. via lessen)	☐ **ja, heel kort** ☐ **ja, uitgebreid** ☐ **neen**
	8. Hoe vaak gebruik je PubMed om biomedische informatie op te zoeken?	1 2 3 4 5 6 7 **nooit** **zelden** **1 keer per maand** **meerdere keren per maand** **1 keer per week** **meerdere keren per week** **elke dag**
	9. Ken je ook de geavanceerde zoekopties van PubMed?	☐ **ja, en ik gebruik ze ook** ☐ **ja, maar ik gebruik ze niet.** ☐ **neen**
	10. Ik vind de geavanceerde zoekopties van PubMed nuttig.	1 2 3 4 5 6 7 **volledig oneens** **neutraal** **volledig eens** ☐ **niet van toepassing**

Zoekstrategieën	11. Ken je MeSH (**Me**dical **S**ubject **H**eadings)?	☐ **ja** ☐ **neen**
	12. Maak je gebruik van de MeSH wanneer je iets in PubMed opzoekt?	☐ **ja, soms** ☐ **ja, altijd** ☐ **neen**
Engelse taal	13. Hoe schat je je kennis van het Engels in?	**1 2 3 4 5** **heel slecht slecht middelmatig goed uitstekend**
	14. Een zoekopdracht in het Engels uitvoeren is moeilijker dan een zoekopdracht in het Nederlands.	1 2 3 4 5 6 7 **volledig oneens neutraal volledig eens**
	15. Ik ben het gewoon om informatie in het Engels op te zoeken.	1 2 3 4 5 6 7 **volledig oneens neutraal volledig eens**

16. Ik lees een **per maand** gemiddeld ... medische artikels in het Engels	☐ **geen** ☐ **hoogstens 1** ☐ **2 tot 4** ☐ **5 tot 7** ☐ **8 tot 10** ☐ **meer**
17. Ik heb zelf al medische teksten in het Engels geschreven.	☐ **neen, nog nooit** ☐ **ja, al 1 keer** ☐ **ja, al meerdere keren** ☐ **ja, ik schrijf regelmatig medische teksten in het Engels**
18. Ongeveer … % van het cursusmateriaal is in het Engels geschreven.	…… %

19. Ik lees liever Nederlandse dan Engelse medische teksten.	1 2 3 4 5 6 7 **volledig oneens** **neutraal** **volledig eens**
20. Ik doe er langer over om een medische tekst in het Engels te lezen dan in het Nederlands.	1 2 3 4 5 6 7 **volledig oneens** **neutraal** **volledig eens**
Opmerkingen	

5. Enkele vragen achteraf

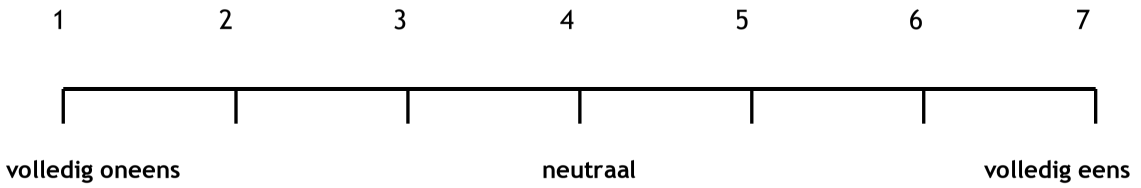
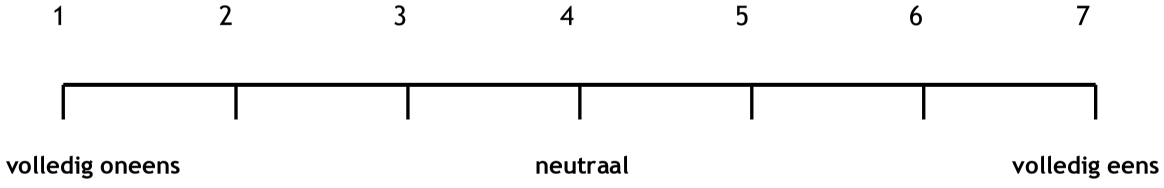
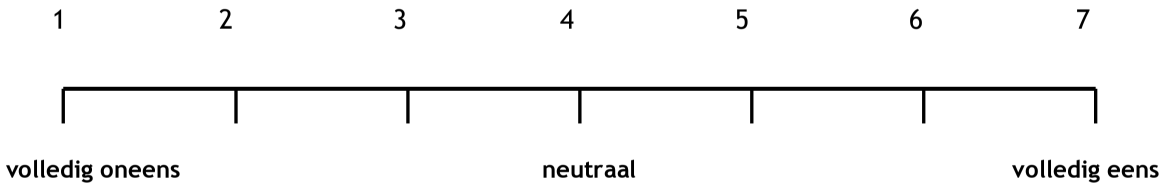
Nummer computer: .....

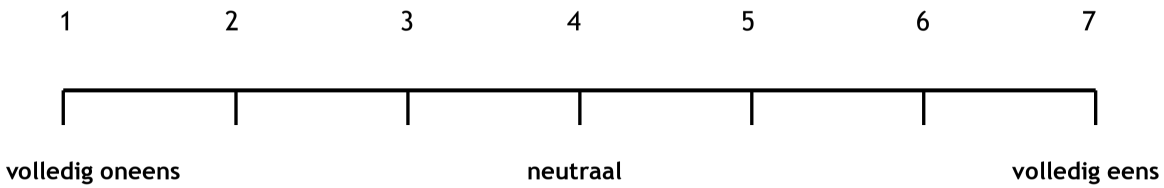
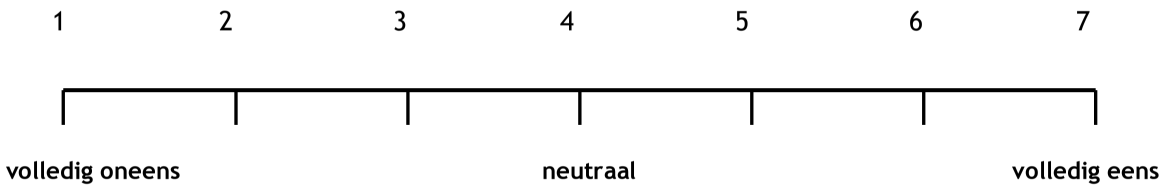
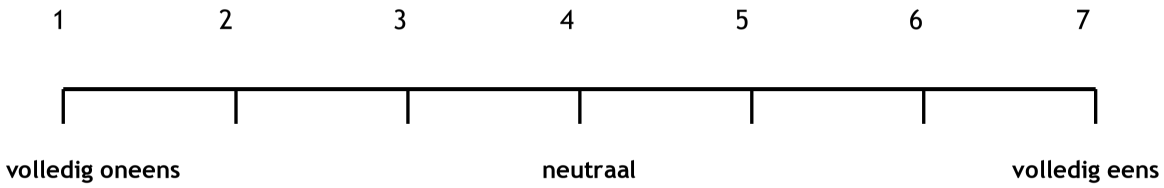
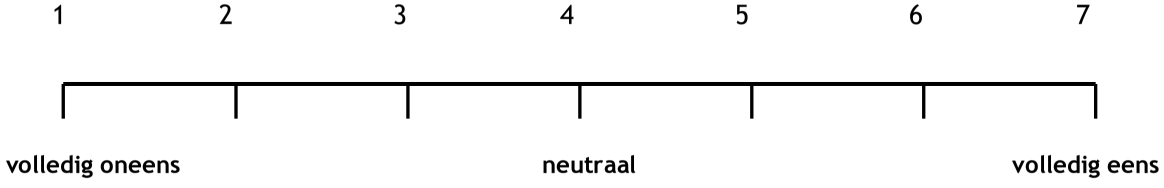
Geslacht: ☐ man ☐ vrouw

Instelling: Universiteit Antwerpen

Opleiding en jaar: .....

	Vraag	Antwoord
Algemeen	1. Heb je de test volledig afgelegd?	☐ **ja** ☐ **neen, omdat:** ....
Zoekopdracht	2. De zoekopdracht was duidelijk geformuleerd.	1 2 3 4 5 6 7 **volledig oneens** **neutraal** **volledig eens**

	3. Ik heb ooit al informatie opgezocht over dit onderwerp.	☐ ja ☐ neen
Tevredenheid	4. Ik heb vlot de gewenste informatie gevonden.	
	5. Ik heb een goede selectie gemaakt van relevante artikels.	
Zoeksysteem	6. PubMed heeft een gebruiksvriendelijk zoeksysteem.	

	7. Ik vind dat het zoeksysteem van PubMed logisch in elkaar zit.	
	8. Ik zou graag meer leren over de zoekmogelijkheden in PubMed.	
Taal van het systeem	9. Ik moest vaak zoeken naar het juiste woord in het Engels.	
	10. Ik twijfelde vaak aan de schrijfwijze van de Engelse woorden.	

Opmerkingen

6. Taalbeheersingstest

Nummer computer: .....

Instructies

1. Start de DIALANG software door op het icoontje (bureaublad) te klikken.

2. Kies “Instructies in het Nederlands”

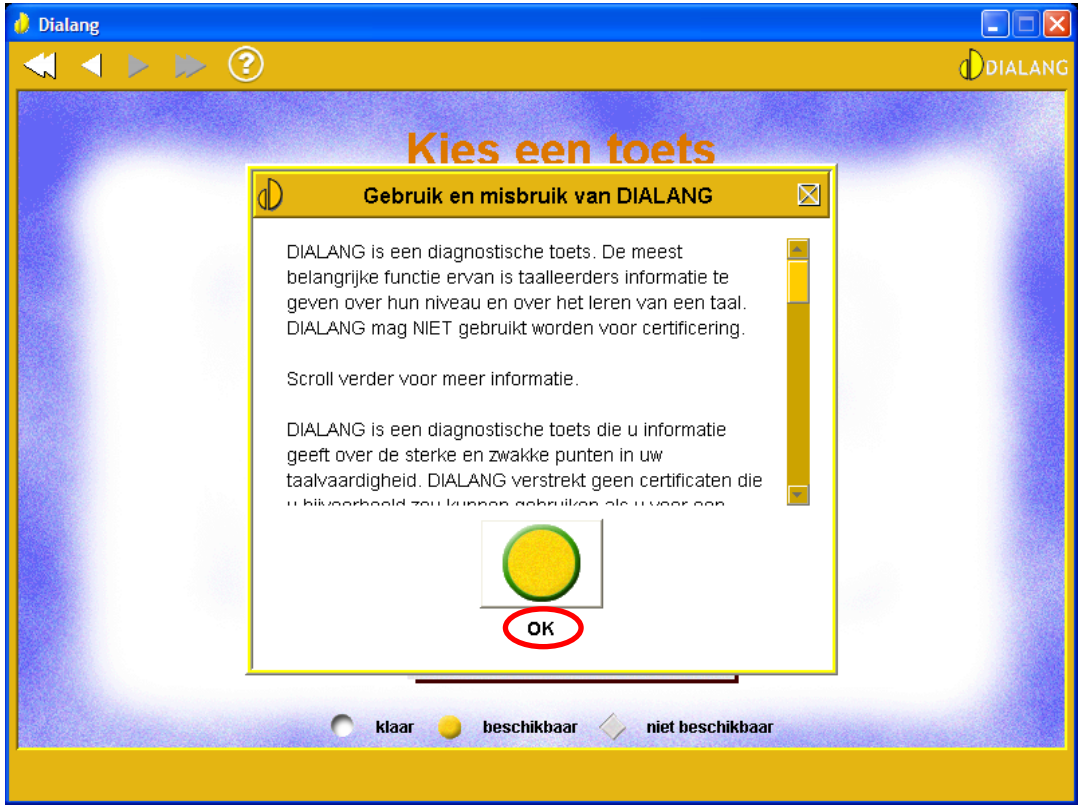
3. Je krijgt een scherm met een aantal knoppen die je tijdens de test nodig hebt:

4. Klik op de “volgende”-toets bovenaan:

5. Je krijgt een scherm met de procedure:

Klik opnieuw op de “volgende”-toets bovenaan:

6. Op het volgende scherm krijg je wat uitleg over de aard van de test; Klik op OK:



7. Kies vervolgens de taal en het soort test: Engels - Lezen

8. <u>Sla de plaatsingstoets en daarna de zelfevaluatie over</u> door op de forward-knop te klikken:

Klik telkens op "ja" om je keuze te bevestigen.

9. Klik op de "play"-toets om de test te starten:

10. <u>Herhaal</u> de test voor het onderdeel Engels, Woordenschat.
<u>RESULTAAT:</u> Ik behaalde …… op de <u>lees</u>test. Ik behaalde …… op de <u>woordenschat</u>test.

D. Pre- and posttest questionnaires (English)

1. General:

Computer	.....
Sex:	☐ male ☐ female
Institution:	.....
Training and year:	.....
Date of the test:	.....

2. Info and instructions:

Dear student,

First of all, thank you for participating in this test! Please read the instructions carefully before you start. Should you have any questions, please do not hesitate to ask them.

Instructions

This test is part of a PhD research project at the University College of Ghent (Belgium) about the efficiency of English (bio)medical retrieval systems. Both Dutch-speaking and English-speaking students will do this test.

The test will take about one hour and consists of 5 parts:

- A questionnaire about your experience with PubMed, your computer skills, etc. (5 min)
- An introduction into searching PubMed (10 min)
- A literature search in PubMed (15 min)
- A language test (20 min)
- A satisfaction survey (5 min)

1. Questionnaires

The questionnaires contain two types of questions :

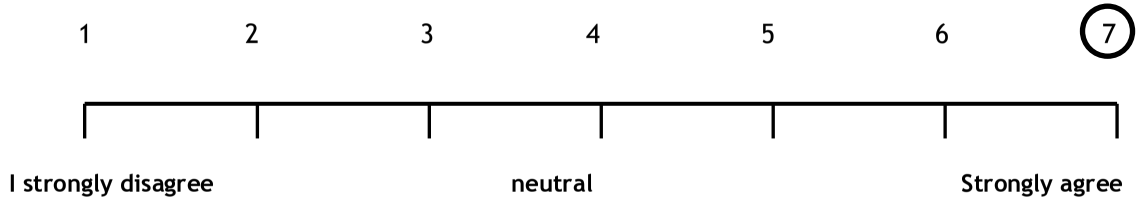
- Questions where you have to tick the appropriate answer.
  vb. *"I am a student at the University of Nottingham"*

  ☒ *yes*

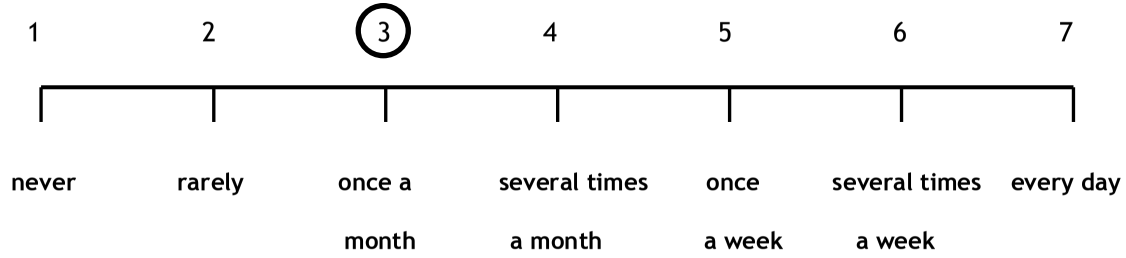
  ☐ *no*

- Opinion or frequency questions, in which you circle on a scale from 1 to 7, the extent to which you agree, or the appropriate to frequency.

e.g. *"I think this is an interesting test."*



e.g. *"How often do you go to the cinema?"*



Please try to answer the questions as truthfully as possible, so we can make a balanced assessment. Do not hesitate to address the person in charge of this test if any problems should occur. In the last section of this test, the satisfaction survey, some space is provided for your comments and/or remarks.

2. PubMed tutorial

Once you have completed the questionnaire, you will be asked to have a look at a **PubMed tutorial** (PowerPoint presentation). This tutorial will give you an introduction into searching for medical information with **MeSH** (Medical Subject Headings).

3. PubMed search

After the tutorial you will be asked to do a **literature search** in PubMed. Please read the assignment carefully and then look for **relevant articles about this subject**. Every step in your search process will be registered and recorded, so we can determine where potential problems occur. At the end of this part, you should have a list of the articles you found most relevant. Send every article you think is relevant to the subject to the PubMed clipboard.

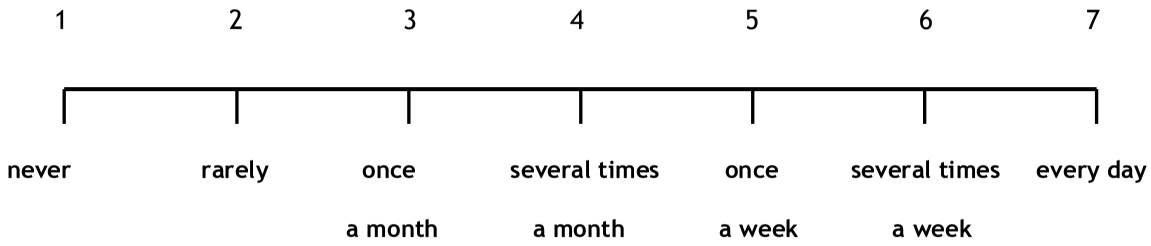
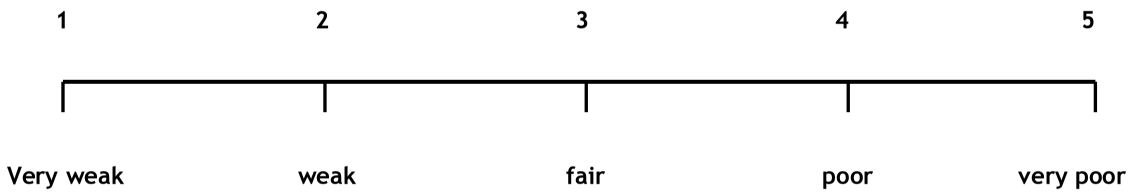
4. DIALANG language test

For an objective analysis of your language skills and their influence on search performance, both test groups -Belgian and British- do a language test, which consists of a reading test and a vocabulary test. This will take about 20 minutes.

Thank you very much for your time!

3. Pretest questionnaire

	Question	Answer
Computer skills	1. How often do you use a computer (desktop/laptop)?	1 never 2 rarely 3 once a month 4 several times a month 5 once a week 6 several times a week 7 every day
	2. What are the main purposes you use your computer for? Tick the **3 most appropriate answers**.	☐ word processing ☐ communication (e-mail, chat,...) ☐ data analysis ☐ looking for information on the Internet ☐ programming ☐ downloading music and/or movies ☐ other: _____

	3. How often do you search for (general) information on the internet?	
	4. Where do you usually look for (bio)medical information? (several answers possible)	☐ Google ☐ Biomedical databases, namely: _____ ☐ Other, namely.: _____
	5. How would you assess your own computer skills?	

Familiarity with PubMed	6. Have you been introduced to PubMed? (e.g. in courses)	☐ **yes, very briefly** ☐ **yes, extensively** ☐ **no**
	7. How often do you use PubMed to search for biomedical information?	1 2 3 4 5 6 7 **never** **rarely** **once a month** **several times a month** **once a week** **several times a week** **every day**
	8. Do you know PubMed's advanced search options (filters, limits, clipboard,…)?	☐ **yes, and I use them myself** ☐ **yes, but I don't use them** ☐ **no**
	9. I find PubMed's advanced search options useful.	1 2 3 4 5 6 7 **Strongly disagree** **neutral** **Strongly agree** ☐ **not applicable** (if you answered "no" or "yes, but I don't use them" in the previous question)

Search strategies	11. Do you know the MeSH (**Me**dical **S**ubject **H**eadings)?	☐ **yes** ☐ **no**
	12.Do you use the Medical Subject Headings in PubMed?	☐ **yes, sometimes** ☐ **yes, always** ☐ **no**

4. <u>PubMed search</u>

5. Posttest questionnaire

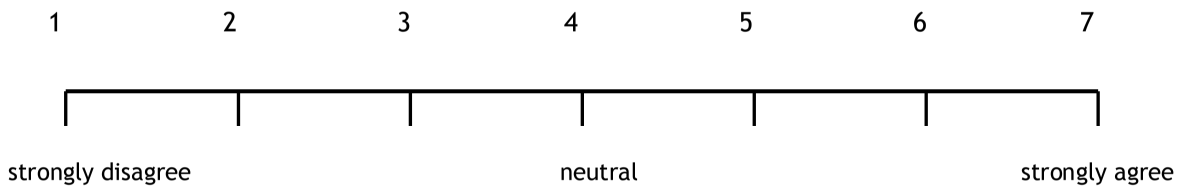
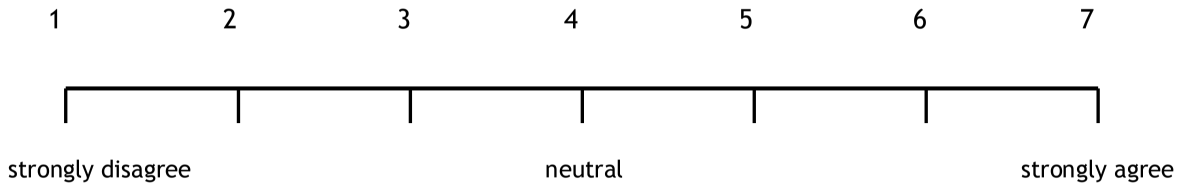
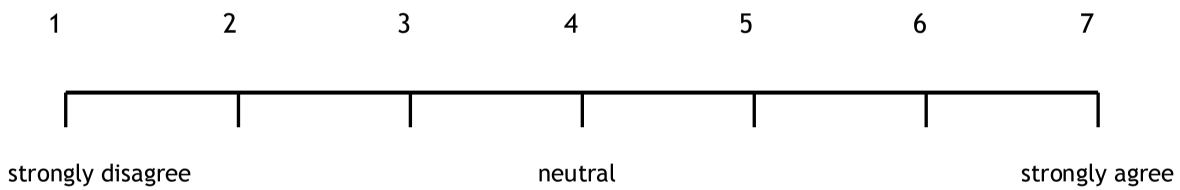
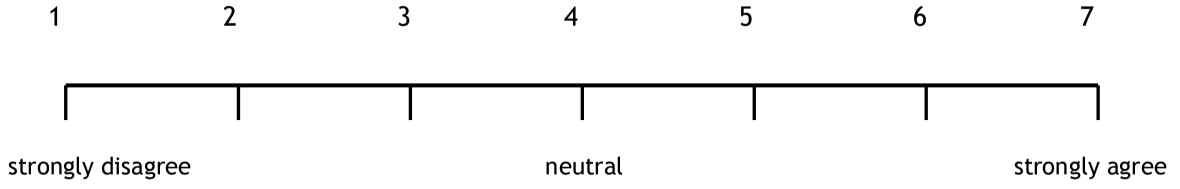
Computer number: .....

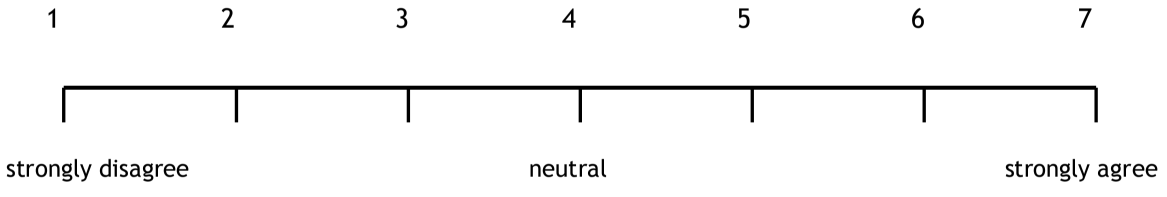
Sex: ☐ male ☐ female

Institution: .....

Training and year: .....

	Question	Answer
Algemeen	1. Did you complete the whole test?	☐ **yes** ☐ **no, because:** ....
Zoekopdracht	2. The subject of the PubMed search was formulated clearly.	1 2 3 4 5 6 7 Strongly disagree neutral strongly agree

	3. I had looked up information about this subject in the past.	☐ **yes** ☐ **no**
Tevredenheid	4. I found the information I needed quite easily in PubMed.	
	5. I made a good selection of relevant articles.	
PubMed	6. PubMed has a user-friendly search system.	
	7. PubMed's search system is logical.	

	8. I would like to learn more about PubMed's search options.	
Remarks/ comments		

6. Language test

Computer number: .....

Instructions

1. Start de DIALANG software by clicking on the Dialang icon

2. Choose “Instructions in English”

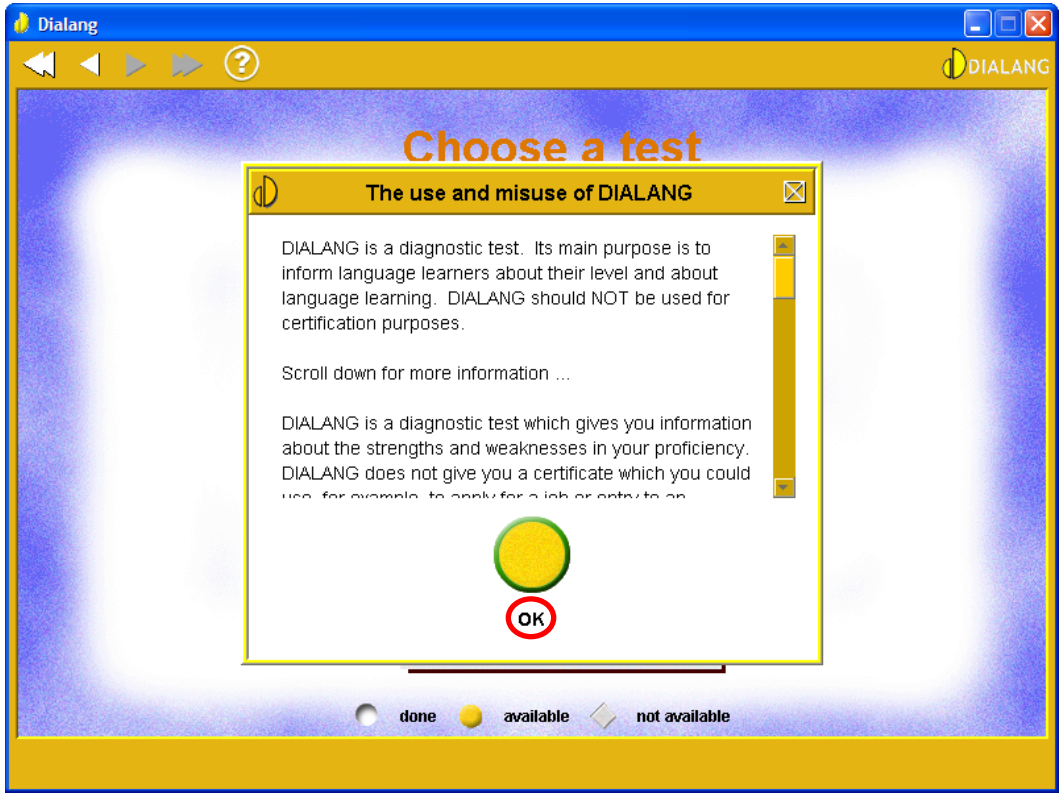
3. A screen appears, explaining the symbols used during the test:

4. Click on “next”:

5. On the next screen, the procedure of the test is explained:

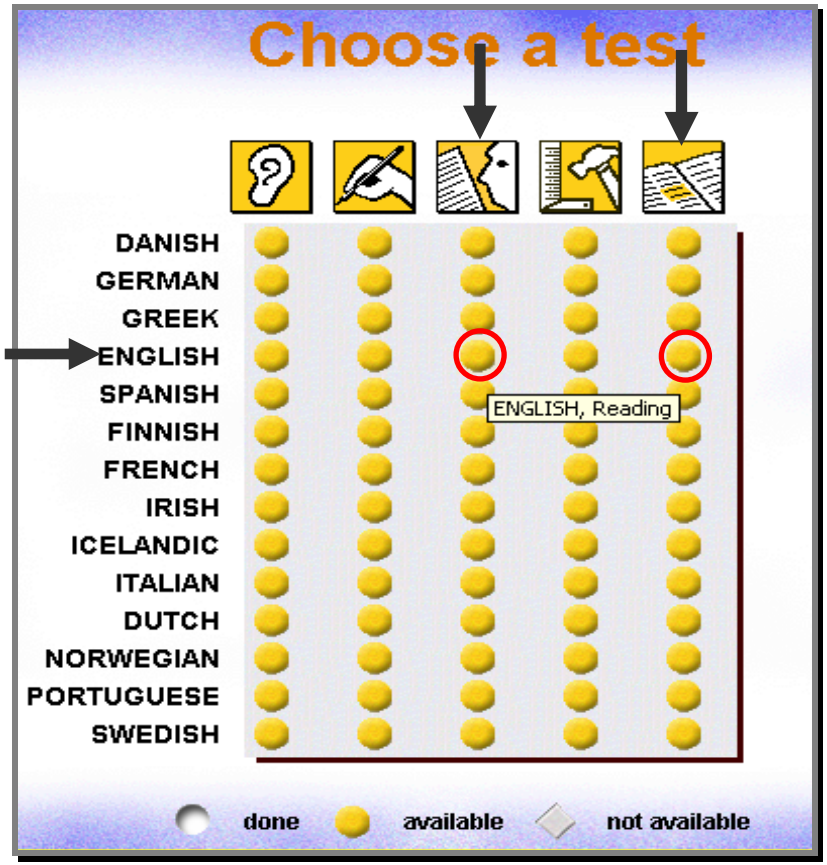
Again, click the “next” tab: 

6. On the next screen, the use of DIALANG is explained:



Click on “OK”.

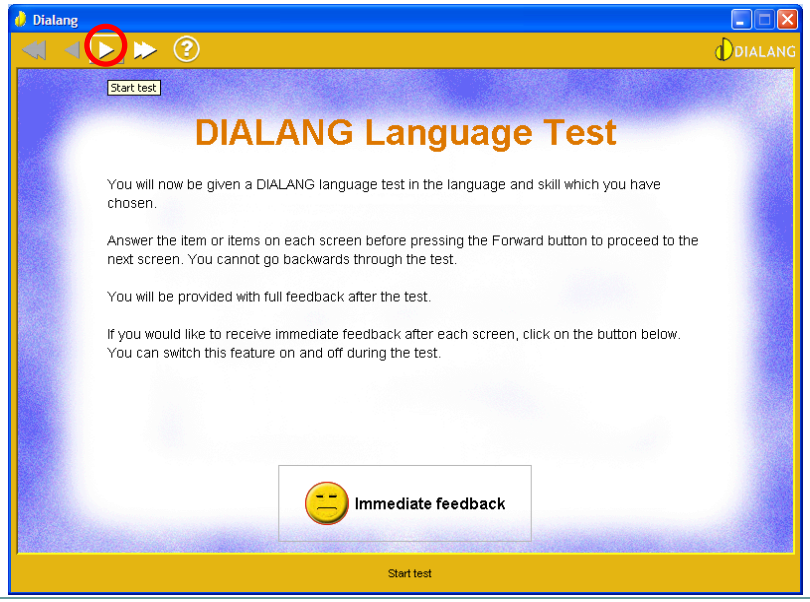
7. Choose your language (English) and the test type (Reading)



8. Skip the Placement Test by clicking on the forward button, and click "Yes" to go confirm your choice:

9. Skip the Self-Assessment Test, and click "Yes" to confirm:

10. Click on "play" to start the test:



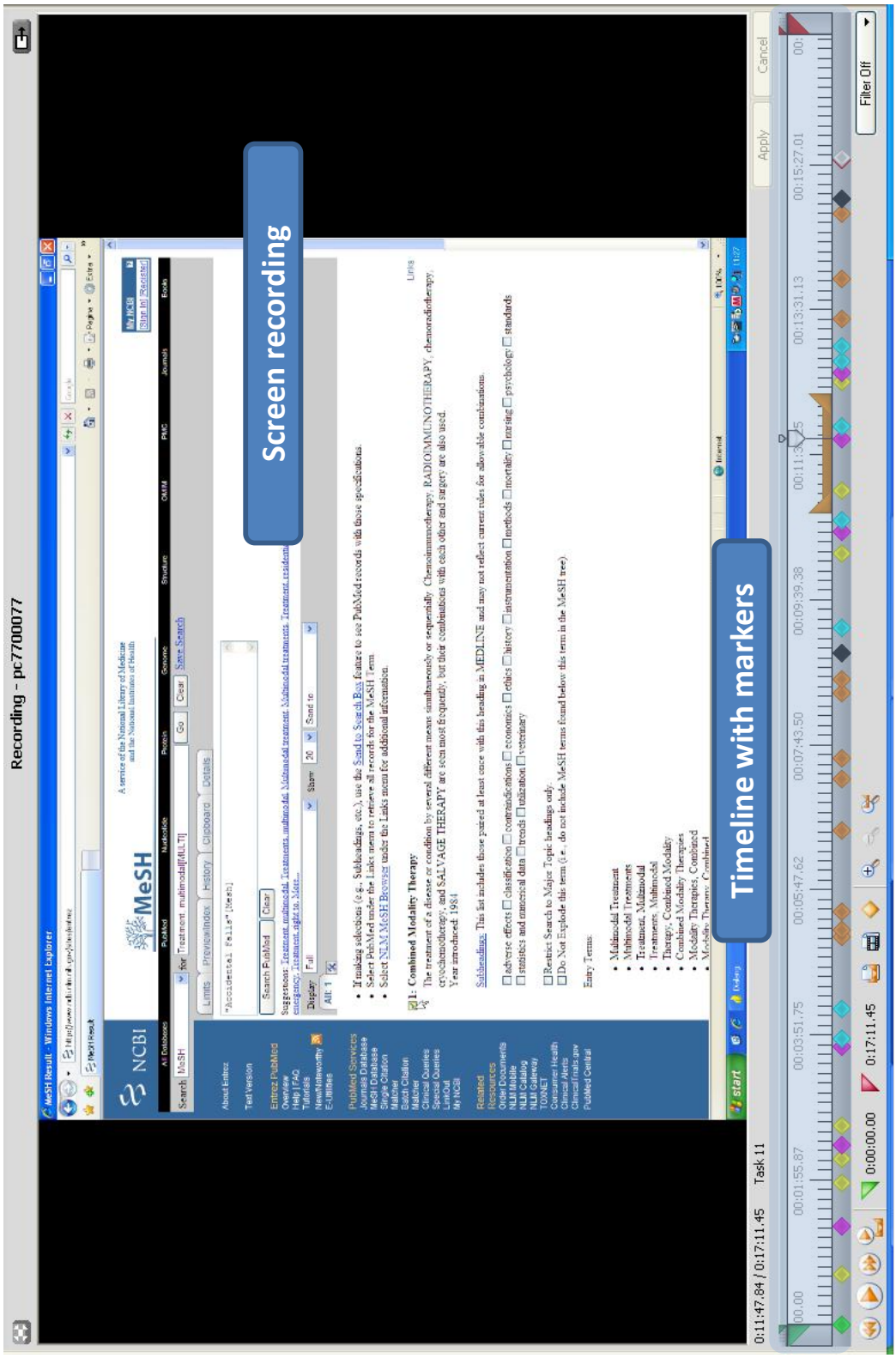
11. Click “next” to go to the next question

RESULT:

I scored …… on the reading skills test.

I scored …… on the vocabulary test.

E. Morae screenshot



F. List of publications

A1

Vanopstal, K., Buysschaert, J., Laureys, G. and Vander Stichele, R. (2013), Query formulation and relevance judgment in native and non-native English-speaking PubMed users, *Journal of the American Medical Informatics Association* (submitted).

Vanopstal, K., Buysschaert, J., Laureys, G. and Vander Stichele, R. (2013), Lost in PubMed: factors influencing the success of medical information retrieval*, Expert Systems with Application, 40 (10): 4106-4114.*

Vanopstal, K., Vander Stichele, R., Laureys, G. and Buysschaert, J. (2012), PubMed searches by Dutch-speaking nursing students : the impact of language and system experience, *Journal of the American Society for Information Science and Technology*, 63 (8) : 1538-1552.

Vanopstal, K., Vander Stichele, R., Laureys, G. and Buysschaert, J. (2011), Vocabularies and retrieval tools in biomedicine : disentangling the terminological knot, *Journal of Medical Systems*, 35 (4) : 527-543.

Hoste, V., Vanopstal, K., Lefever, E., Delaere, I. (2010), Classification-based scientific term detection in patient information, *Terminology*, 2010. 16 (1) : 1-29, John Benjamins Publishing Company, Amsterdam, Netherlands.

B1

Buysschaert, J., Vanopstal, K., Kovács, L. (2008), ELeCT 3.0. Electronic lexicon of communication terminology, *Communication and Cognition.*

C1

Vanopstal, K., Vander Stichele, R., Laureys, G., & Buysschaert, J. (2010). *Assessing the impact of English language skills and education level on PubMed searches by Dutch speaking users.* In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (eds.), Proceedings of the seventh International Conference on Language

Resources and Evaluation (LREC'10). European Language Resources Association, Valletta, Malta.

Vanopstal, K., Desmet, B., & Hoste, V. (2010). *Towards a Learning Approach for Abbreviation Detection and Resolution*. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (eds.), Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association, Valletta, Malta.

Hoste, V., Lefever, E., Vanopstal, K., & Delaere, I. (2008*). Learning-based Detection of Scientific Terms in Patient Information.* In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias (eds.), Proceedings of the Sixth Conference on International Language Resources and Evaluation (LREC'08). European Language Resources Association, Marrakech, Morocco.

Hoste, V., Vanopstal, K., & Lefever, E. (2007). *The Automatic Detection of Scientific Terms in Patient Information*. Proceedings of the Workshop on Acquisition and Management of Multilingual Lexicons. Borovetz, Bulgaria.

Vanopstal, K., & Van Wiele, K. (2007). *Incorporation of two Terminology Projects into a System for Information Retrieval Using NLP for Term Expansion.* In M. Campos Pardillos and A.G. González-Jover (ed.), 1st International conference on Language and Health Care. Alicante, Spain.

C3

Vanopstal, K., Vander Stichele, R., Laureys, G., & Buysschaert, J. (2011), *Native versus non-native English-speaking PubMed users : an interactive study*, poster presented at the AMIA Annual Symposium, American Medical Informatics Association.

Vanopstal, K., Vander Stichele, R., Laureys, G., & Buysschaert, J. (2011), *Causes for poor information retrieval in PubMed*, poster presented at the AMIA Annual Symposium, American Medical Informatics Association.

Hoste, V., Vanopstal, K., Lefever, E., Delaere, I. (2008) *Classification-based scientific term detection in patient information*, paper presented at Atila 2008.