

Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft



DISSERTATION

Analysis of the Long Term Dynamics in Thesaurus Developments and its Consequences

Zur Erlangern der Doktorwürde

Philosophische Fakultät I.

Mohammad Tavakolizadeh-Ravari
aus dem Iran

Dekan der Philosophische Fakultät I.: Prof. Dr. Michael Borgolte

Gutachter: 1. Prof. Dr. Walther Umstätter

2. Prof. Dr. Robert Funk

eingereicht: 01.06.2007

Datum der Promotion: 17.07.2007

Zusammenfassung

Die Arbeit beschäftigt sich mit der statistischen Erfassung der intellektuellen Indexierung mit Hilfe von Thesaurusbegriffen. Sie versucht die dynamische Entwicklung und den Gebrauch von Thesaurusbegriffen zu analysieren. Zusätzlich konzentriert sie sich auf die Faktoren, die die Zahl von Indexbegriffen pro Dokument bzw. bei den verschiedenen Zeitschriften beeinflussen. Als interessante Faktoren erwiesen sich: „Länge der Dokumente“, „Vorhandensein von Zusammenfassungen“, „Sprache der Dokumente“, „Datum der Indexierung“, „Journal Impact Factor (JIF)“, und die „Priorität der Journale bei der Indexierung“. Als Untersuchungsobjekt dienten die Medical Subject Headings (MeSH) und die entsprechende Datenbank „MEDLINE“. Bei ihr liegen bekannte frühere Untersuchungen vor, sie existiert seit vielen Jahren und der Gesamtumfang an Dokumenten ist beeindruckend. Die wichtigsten Konsequenzen der Analyse sind, wie folgt:

1. Der MeSH-Thesaurus hat sich durch drei unterschiedliche Phasen jeweils logarithmisch entwickelt. In jeder Phase hat der Bedarf der Optimierung die Wachstumsrate der Thesaurusbegriffe bestimmt, da die exponentielle Zunahme der zu indexierenden Dokumente zu bewältigen war. Das Wachstum eines Thesaurus wie bei den MeSH sollte nach den vorliegenden Untersuchungen der folgenden Gleichung folgen: „ $T = 3.076,6 \ln(d) - 22.695 + 0,0039d$ “ (T = Begriffe, Ln = natürlicher Logarithmus und d = Dokumente). Um solch einen Thesaurus zu konstruieren, muss man demnach etwa 1.600 Dokumente haben, die die unterschiedliche Themen des Bereiches des Thesaurus umfassen, um den Grundstock an Begriffen aufbauen zu können. Die dynamische Entwicklung von Thesauri wie MeSH erfordert die Einführung eines neuen Begriffs pro Indexierung von 256 neuen Dokumenten.
2. Die Verteilung der Thesaurusbegriffe erbrachte drei Kategorien: starke, normale und selten verwendete Headings. Die letzte Gruppe ist in einer Testphase, während in der ersten und zweiten Kategorie die neu hinzukommenden Deskriptoren im Lauf der Zeit zu einem Thesauruswachstum führen.
3. Es gibt ein logarithmisches Verhältnis zwischen der Zahl von Index-Begriffen pro Aufsatz und dessen Seitenzahl. Dieses Verhältnis gilt für den Bereich von Artikeln zwischen einer und einundzwanzig Seiten.
4. Im allgemeinen erhalten Zeitschriftenaufsätze mit Abstracts fast zwei Deskriptoren mehr als die, die in MEDLINE ohne Abstract erscheinen.

5. Die Zahl von Indexbegriffen pro Aufsatz zeigte, dass die Findability der nicht-englischsprachigen Dokumente, wie z.B. Publikationen auf Deutsch in MEDLINE geringer ist als die der englischen Dokumente. Der größte Unterschied ist bei Aufsätzen mit 10 Seiten (33% weniger Deskriptoren) zu verzeichnen.
6. Aufsätze der Zeitschriften mit einem Impact Factor 0 bis fünfzehn erhalten nicht mehr Indexbegriffe als die der anderen von MEDLINE erfassten Zeitschriften.
7. In einem Indexierungssystem haben unterschiedliche Zeitschriften mehr oder weniger Gewicht in ihrem Findability. Die Verteilung der Indexbegriffe pro Seite hat gezeigt, dass es bei MEDLINE drei Kategorien Publikationen gibt. Die mit 2,3, 1,5 und 0,7 von MeSH-Begriffen pro Seite. „Natur“, „Science“ und „Transplant Proc.“ gehören beispielsweise zu den von MEDLINE stark bevorzugten Zeitschriften.

Schlagwörter:

Intellektuelle Indexierung, Sachliche Erschließung, Indexierungsbreite, Indexierungstiefe, Thesaurusaufbau, Thesaurusentwicklung, Verteilung von Thesaurusbegriffen, MEDLINE, MeSH.

Abstract

The current dissertation concerns subject indexing with thesaurus terms. It tries to analyze dynamic development and use of thesauri by statistical methods. In addition, it focuses on the six factors that have affected the number of index terms per document or journal. They are “length of documents”, “presence of abstracts”, “language of documents”, “date of indexing”, “Journal Impact Factor”, and “priority of journals for in-depth indexing”.

Medical Subject Headings (MeSH) and its corresponding well known database “MEDLINE” were established to conduct this research. The main consequences of analyzing the long-term indexing of MEDLINE are as follows:

1. MeSH has developed logarithmically through three different phases. The existence of each phase has been due to the need of optimizing the growth rate of thesaurus terms to cope with the exponential increase of indexed documents. The growth of a thesaurus such as MeSH should consequentially follow the equation “ $T = 3,076.6 \ln(d) - 22,695 + 0.0039d$ ” (T = thesaurus terms, Ln = natural logarithm, and d = documents). To construct such a thesaurus, one needs to have at least 1,600 documents covering different topics of the thesaurus subject area. The dynamic of thesauri such as MeSH is due to the persistent inclusion of one new term per indexing of 256 new documents.
2. The distribution of thesaurus terms yielded three classes: highly, normally, and rarely used terms. The last group is in a test phase, and only growth rates of most frequented terms in the first class and newer terms in the second class were becoming persistent over time.
3. There is a logarithmic relationship between the number of index terms per article and its pages. This relationship will occur if the articles are between one and twenty-one pages.
4. In general, journal articles with abstracts received almost two more terms than those included into MEDLINE without abstracts.
5. The number of index terms per article showed that findability of non-English documents, such as articles written in German and indexed in an American-based database like MEDLINE, is less than that of English documents. The greatest difference is for articles with ten pages (33% more index terms of English articles) and the least is for those with twenty and more pages.
6. Journals with Impact Factors in the range from 0 to fifteen receive roughly the same number of index terms per page.

7. In an indexing system, different journals have more or less weight in their findability. Distribution of index terms per page has shown that there are three regions respectively with 2.3, 1.5, and 0.7 terms per page. In addition to these regions, few journals are the most favored ones and get more index term per page. “Nature”, “Science”, and “Transplant Proc” belong to such journals in MEDLINE.

Keywords:

Manual Indexing, Subject Indexing, Exhausticivity of Indexing, Depth of Indexing, Thesaurus Construction, Thesaurus Development, Use Distribution of Thesaurus Terms, MEDLINE, MeSH.

Table of Contents

Zusammenfassung	2
Abstract	4
Dedication	10
Abbreviations	11
Preface	12
1 Introduction	13
1.1 Overview	13
1.1.1 Aim	13
1.1.2 Research questions	14
1.1.3 Materials and Methods	15
1.1.4 Main results	17
1.2 Thesaurus	18
1.2.1 Linguistic structure of thesaurus	20
1.2.2 Similar problems of Conventional and automatic thesauri	21
1.2.3 MeSH as subject headings and thesaurus	21
1.3 Indexing	22
1.3.1 Depth of indexing	23
1.3.2 Exhaustivity of indexing	24
1.3.3 Specificity of indexing	24
2 Materials and Methods	26
2.1 PubMed	26
2.1.1 PubMed Coverage	27
2.2 MEDLINE	27
2.2.1 MEDLINE format	28
2.3 Medical Subject Headings (MeSH)	28
2.3.1 Growth of MeSH	29
2.4 Use distribution of MeSH headings in MEDLINE	30
2.5 Factors effecting the number of MeSH headings per article in MEDLINE	31
2.5.1 Determining the text tokens and types	31
2.5.2 Determining number of pages per article	32
2.5.3 Determining the presence and form of abstracts	33
2.5.4 Determining journal titles	34
2.5.5 Determining the number of MeSH headings per article	34

2.5.6	Determining the Entrez date.....	35
2.5.7	Determining the indexing priority of Journals	36
2.5.8	Determining Journal Impact Factor (JIF).....	36
2.6	Some notes about programming by Delphi	36
2.6.1	Processes for sorting records.....	37
2.6.2	Processes for determining distinct MeSH headings.....	37
2.6.3	Processes for determining the growth of MeSH	37
2.6.4	Processes for determining the use distribution of MeSH headings.....	38
2.6.5	Processes for determining the average number of MeSH headings per article	38
3	Results	44
3.1	Growth of Medical Subject Headings (MeSH).....	44
3.1.1	Cumulative Growth of Medical Subject Headings (MeSH)	44
3.1.2	Growth of MEDLINE vs. growth of MeSH.....	47
3.1.2.1	Half-Term-Rate (HTR)	48
3.1.3	Absolute growth of Medical Subject Headings (MeSH)	49
3.1.4	Optimization of accurate thesaurus development	52
3.2	Distribution of MeSH headings in MEDLINE	54
3.2.1	Highly frequented headings	57
3.2.2	Normally frequented headings	58
3.2.2.1	Half-Rank-Usage (HRU) of normally frequented headings	59
3.2.3	Rarely frequented headings.....	61
3.3	Factors related to the number of index-terms of articles	62
3.3.1	Article length.....	62
3.3.1.1	Tokens and types.....	63
3.3.1.1.1	Relationship between the text tokens and types.....	63
3.3.1.1.2	Relationship between the number of pages of articles and tokens.....	64
3.3.1.1.3	Tokens of articles and average of MeSH headings assigned to them	65
3.3.1.2	Number of pages and existence of abstracts	65
3.3.1.2.1	Articles with and without abstracts	66
3.3.1.2.1.1	Average of MeSH headings per page.....	67
3.3.1.2.2	Structured and unstructured abstracts	68
3.3.2	Language of articles	69
3.3.3	Date of indexing	72
3.3.3.1	Average lengths of articles over the years	72

3.3.3.2	Average of MeSH headings of articles over the years.....	73
3.3.3.2.1	Role of Abstracts over the years	74
3.3.3.2.2	Role of structured abstracts over the years.....	74
3.3.4	Journal priorities for in-depth indexing.....	75
3.3.4.1	Journal Impact Factor.....	77
4	Discussion.....	80
4.1	Growth of Medical Subject Headings (MeSH).....	80
4.1.1	Interaction between Thesaurus development and in-depth indexing.....	83
4.2	Distribution of MeSH headings in MEDLINE.....	83
4.2.1	Highly frequented headings	84
4.2.2	Normally frequented headings	86
4.2.3	Rarely frequented headings.....	87
4.3	Factors related to the number of index terms of articles.....	89
4.3.1	Length of Articles.....	90
4.3.2	Presence of abstracts	95
4.3.2.1	Structured and unstructured abstracts	99
4.3.3	Language of Articles	101
4.3.4	Date of indexing.....	104
4.3.5	Priority of journals for in-depth indexing	107
4.3.6	Impact Factor.....	107
5	Conclusion.....	110
5.1	Development of thesaurus terms.....	110
5.2	Use distribution of thesaurus terms.....	111
5.3	Factors related to the number of index terms of articles.....	111
5.3.1	Length of articles.....	112
5.3.2	Articles with abstract.....	112
5.3.3	Language of articles	112
5.3.4	Date of indexing.....	113
5.3.5	Journal Impact Factor (JIF).....	114
5.3.6	Priority of journals for in-depth indexing	114
6	Theses.....	115
	References.....	116
	Acknowledgment	123
	List of Figures	124

List of Tables.....	126
List of Equations	127
Curriculum vitae (Lebenslauf)	128
Eidstattliche Erklärung	129

Dedication

*To my parents, who taught me the meaning of love,
and to my wife, Arezoo, for believing in me,
and to my sons Parham and Pedram*

Abbreviations

HRU	Half-Rank Usage
HTR	Half-Term-Rate
IF	Impact Factor
ISI	Institute of Scientific Information
JCR	Journal Citation Report
JIF	Journal Impact Factor
MeSH	Medical Subject Headings
NLM	National Library of Medicine

Preface

Motivation is one of the most important psychological processes. It is constantly interactive, changing, and it enables us to be unique and one-off (Krajnc, A. 1982). Every human activity is motivated. Motivation enables a person to satisfy a need, a goal, which he has set for himself or which has been set for him (Razdevšek-Pučko, C. 1999).

The motives which led me to take the human indexing and development of thesauri into consideration goes back to the time when I was doing my Master. Reading a book guided me to the field of reverse engineering. I learned that engineers develop the technical products by the means of techniques which allow them to reengineer the products. The great number of mechanical products can be reengineered by detecting their three main phenomena: mechanical, dimensional, and operational. My question was how one can detect the phenomena related to an indexing system in order to develop or rebuild the same ones. I presented my idea at a conference that was held in Iran (Tavakolizadeh-Ravari, M., 2002). The literature that I found in this area couldn't satisfy me. The way that could help me to reach the goal was a long term study of a well known indexing system like MEDLINE.

Moving to Germany for completing my Ph.D. made it possible for me to approach this case. Prof. Dr. Walther Umstätter¹ supervised my doctoral project. He offered me the opportunity to include his knowledge and experiences to my work. I learned some main points that helped me to detect the phenomena that had to be focused to get a deep understanding of an indexing system for its reengineering. Three main points were then selected for investigation:

1. Construction and development of a thesaurus in an indexing system.
2. Use of thesaurus terms in its corresponding database.
3. Factors that affect the number of index terms received by articles.

¹ Homepage: <http://www.ib.hu-berlin.de/~wumsta/>

1 Introduction

This dissertation will focus on subject indexing with index-terms. It will consider three main points:

1. **Dynamic developments of thesauri**
2. **Use distribution of thesaurus terms.**
3. **Effecting factors on the average number of index-terms per journal article.**

To conduct the research, MeSH was used, a well-known thesaurus. It corresponds mostly to MEDLINE, which has been recognized as a very dynamic database. These two names have been together for a long time. The great number of indexed documents (over 16,000,000 to date) incorporated dynamically to MEDLINE over four decades made this system into a very attractive source to be used as a sample system to conduct my investigations.

1.1 Overview

This overview will briefly describe the contents and the structure of the dissertation, as well as some essential concepts.

1.1.1 Aim

In the current research, we will try to answer several analytical questions about MEDLINE subject indexing by human indexers. The main goal is to find a number of key consequences related to long term subject indexing with thesaurus terms. To reach this goal, statistical analyses on the following subjects were performed:

1. Development of MeSH through indexing MEDLINE
2. Use distribution of MeSH terms in MEDLINE
3. Factors that affect on the number of MeSH headings per journal article in MEDLINE. They are:
 - i. Length of documents.
 - Number of pages
 - Tokens and types (word frequency and vocabulary size).
 - ii. Abstracts of documents
 - Structured abstracts
 - Unstructured abstracts.
 - iii. Language of documents (comparing the English and German documents).
 - iv. Entrez Date (inclusion date of documents into MEDLINE).

- v. Priority of journals for in-depth indexing.
- vi. Journal Impact Factor (JIF).

1.1.2 Research questions

The following research questions summarize the main points of the dissertation:

1. How does the growth of thesaurus terms correlate with the number of documents in its corresponding database?

The inclusion of new terms into controlled vocabularies indicates the appearance of new concepts in literature. This means there is a relationship between the growth of the publications and the emergence of new terms. Thus, this question addresses the correlation between the growth of the thesaurus terms and the number of new published literature.

2. How is the use distribution of thesaurus terms in a database?

In contrast to the vocabularies of natural texts, the number of thesaurus terms is very limited. Therefore, the use distribution of MeSH terms in MEDLINE may not follow the function that is found by Zipf, G. K. (1949) and we should expect to derive other functions.

3. How is the number of index-terms per article related to the number of pages of articles?

It is clear that longer texts contain more words. However, we are not sure how their profusion affects the amount of concepts inherent within them. The question takes this problem into consideration.

4. Does the inclusion of abstracts in a database reduce the number of index-terms per article?

Abstracts of documents can be counted as an auxiliary tool to present the contents of texts in brief. They make it easier for indexers to detect the concepts within texts. Abstracts bear the key points of documents and allow a pertinent free text search without any need to look up descriptors. The first two points above express that abstracts should increase the number of index-terms per document. But the possibility of free text searching through abstracts opposes the idea of the need for deeper indexing of such documents.

5. Do the journal articles written in German or English have the same chance of findability in a US-Based database?

MEDLINE is a US and English based database, but journals in other languages and from other countries are included within it. A large majority of the journals are US/English and only a small amount is foreign. This fact may affect the number of index terms that indexers give to the non-English articles and reduce their chances of being retrieved.

6. Which events have changed the in-depth indexing of MEDLINE documents over the years?

The average number of MeSH terms that NLM indexers have assigned to documents over the years indicates the different periods of MEDLINE policies for indexing. The events in every period can then show how they (like development of technology) could affect the depth of indexing.

7. How many regions of journals are recognizable through the distribution of the average number of index-terms per journal?

NLM gives priorities to the journals for in-depth indexing. From this point of view, journals are divided into three groups with priorities 1, 2, and 3. These are for in-house use and aren't accessible to others. The average number of MeSH headings per journal will help to rank the journals and find their priorities. We can then determine how deep the given priorities could effect on their indexing.

8. Is there any correlation between the Impact Factor (IF) of Journals and the average number of index-terms they receive?

The question addresses how the IF of Journals could affect the in-depth indexing of their articles.

1.1.3 Materials and Methods

The issue will be discussed in the next chapter (chapter 2) in more detail. The bases of the current dissertation are two products of NLM: MEDLINE and MeSH, which are accessible via PubMed. To find answers to the eight questions above, the needed data was derived sequentially as follows:

- 1. PubMed was searched for the word “up” and retrieved 948,000 records:**
 - i They were then sorted in chronological order of inclusion into MEDLINE.
 - ii. The numbers of records of the sample that returned the first occurrence of every distinct term were recorded.

- 2. The above mentioned sample returned 23,198 distinct terms:**
 - i. Each of them was searched four times. Every time, the searches were limited to a certain interval. The time limitations were “1965 – 1970”, “1965 – 1980”, “1965 – 2000”, and “1965 – 2006”.
 - ii. The number of returned records was recorded following every search. If the number of results was zero, it indicated that the term was not excluded from MeSH in that period.
- 3. PubMed was searched for the words “Humans and Medical”. It retrieved almost 1,000,000 records. The search was limited to the “Entrez Date” between the years 1965 and 2005, “Journal Articles” AND “English”:**
 - i. Check tags were excluded from them
 - ii. Articles longer than thirty pages were also excluded
 - iii. The rest were divided into thirty groups based on their number of pages.
 - iv. The total number of MeSH headings and documents of each group was recorded.
 - v. The total number of MeSH headings of each group was divided by the total number of documents of the corresponding group to get the average number of MeSH headings per article.
 - vi. Nine full-text articles of different lengths were downloaded from the links given by PubMed, and their tokens and types were determined.
- 4. Beside the processes done on the sample in “3.” the concentration was on the “Abstracts Field” of records. This time records that had the “Abstracts’ Field” were taken into consideration and the same processes repeated on them again two times.**
 - i. The first time, the forms of abstracts were not important.
 - ii. The second time, the processes were done on records with structured abstracts. The occurrence of words like AIM, OBJECTIVES and etc. (in uppercase) in abstracts made it possible to distinguish them from others.
- 5. PubMed was searched for journal articles that were written in German between the years 1965 – 2005. About 500,000 records were retrieved. The same processes that were done on the sample in “3.” were repeated on this sample.**
- 6. The sample in “3.” was used again and the same processes were done on it, except for process “iii”. Instead, they were placed into forty groups based on their inclusion dates into MEDLINE.**
- 7. The sample in “3.” was used again. The same processes were applied on it, except for process “iii”. Instead, they were grouped based on their corresponding journals.**

- i. The journals that had more than 500 articles in the sample were taken into consideration, resulting in 454 journals.
- ii. Fulfilling this condition created 454 groups.
- iii. Instead of the process “v.” in “3.”, the total number of pages of each group was divided by the total number of MeSH headings of corresponding group.

8. The titles of the 454 journals mentioned above were searched in JCR to find their IF for the years 2003 and 2004. This resulted in the IF of 245 journals.

- i. The same processes of “7” applied on them, except for “iii”. Instead, the total number of MeSH headings of every journal was divided by the total number of indexed articles of corresponding journal

1.1.4 Main results

The main results of the thesis are numbered between one and eight. Every number corresponds to the questions in section 1.1.2.

1. Medical Subject Headings (MeSH) have grown following three logarithmic functions. The exponents of the functions have increased linearly. Simultaneously, MEDLINE citations have grown following three exponential functions. On the contrary, their exponents have decreased.
2. The use distribution of MeSH headings has shown that they should be divided into three classes. The distribution of the highly frequented class is that of a power law, that of the normally frequented is exponential, and that of the rarely frequented is linear. The majority of MeSH terms belong to the normally frequented class.
3. The correlation between the lengths of journal articles without abstracts and the average number of their MeSH terms is logarithmic in MEDLINE. The function is “ $y = 1.2905 \ln(x) + 5.1966$ ”. It is valid only for articles between one and twenty-one pages.
4. The existence of abstracts could increase the average number of MeSH terms given to journal articles between one and seventeen pages and then reach the level of the articles without abstracts. The exponent of the logarithmic correlation between the lengths of such articles and the average of their MeSH headings increased to “ $y = 2.1816 \ln(x) + 5.2454$ ”. This function is valid only for articles between one and ten pages.
5. The number of indexed MeSH terms has shown, on average, that the articles written in English and consisting of ten pages have 33% greater findability than those written in

German. But for article with twenty pages, the findability is nearly the same for both English and German.

6. The average number of MeSH headings given to journal articles in the years “1965 – 2005” shows that in-depth indexing of MEDLINE has had three periods. A linear increase between the years “1965 – 1974”, a linear decrease between the years “1975 – 1981”, and a linear increase between the years “1982 – 2005”. Mechanisation of Index Medicus and the inclusion of abstracts in 1974 and structured abstracts in 1988 are three main events that could have changed in-depth indexing of journal articles.
7. The articles of three known journals (Nature, Science, and Transplant Proc) were given 3.3 MeSH headings per page. The distribution of MeSH headings per page for other journals revealed the existence of three regions, which agree with the three priority numbers given to the journals by NLM. The average MeSH headings of journals per page for the first to three regions were respectively “2.3”, “1.5”, and “0.7”.
8. The relationship between the average number of MeSH headings per journal page and their Impact Factor is only for journals with IF higher than fifteen verifiable.

1.2 Thesaurus

A thesaurus is etymologically a treasure from the point of view of documentation (Schwartz, I. and Umstätter, W., 1999). In general, a thesaurus is a list of terms. A term can be a word, a composed word or even an expression, often indicating structural relationships between the terms. De Jesus Adriano, H. et al. (2004) stated that: “The term or phrase entries in a thesaurus are commonly listed alphabetically for easy location of entries, with some entries being arranged hierarchically. Entries often indicate which other terms are broader terms (often abbreviated in a printed thesaurus as “BT”) or narrower terms (often abbreviated as “NT”). Broader terms, often representing a superclass, such as mammals, are above narrower or subclass terms, such as primates or ungulates, on the hierarchy”. “Members of a subclass can be said to inherit features of the superclasses to which they belong” (Losee, R. M. 2006).

“The content of a document is represented using the words that appear in it” (Zazo, Angel F. et al. 2005). Indexers use thesaurus terms to control and standardize these content-bearing words. This limits the terms available and increases the possibility that the query will use appropriate terms in retrieval process (Bechhofer, S. and Goble, C. A., 2001). A thesaurus has a two-fold function. One concerns indexing and the other the retrieval of documents.

Thellefsen, M. (2004) discusses the consistency and exhaustivity in indexing, recall, and precision of searching. As he says, “without vocabulary control, indexing becomes fuzzy and messy and subject searching becomes haphazard”.

In fact, controlled vocabularies reveal the subjects of questions that can be answered by a database. From this point of view, a thesaurus contains classes of “questions that presuppose the existence of the documents pertaining to the subject of the question“ (Derr, R. L., 1982, p. 70). Therefore, the numbers of descriptors given to the citations of a database assert how many answers can be returned by the current document. Frequency of each term in the database shows how many answers can be found for the subject of the queried question. In addition, the relatedness between the growth of thesaurus terms and the number of indexed documents indicates the number of publications needed for creation of new questions in the world of knowledge. Consequently, thesauri for information retrieval systems like Biosis, Chemabs, ERIC, MEDLINE, etc. can be understood as compressed collections of questions that can be answered by published literature. Their growth is proportional to the growth of new questions in the different topics.

This phenomenon reveals that a thesaurus is a part of a question-answering system. It copes with this duty when its construction and development follows several system development principles. They are well presented by Chen, H. et al. (1997): “logarithmic vocabulary growth, completeness, term specificity, asymmetric association, relevance feedback, vocabulary overlapping, and spreading activation”. This thesis focuses on some of them.

It is clear that the existence of questions in the area of knowledge will never cease. The development of thesauri is related to new questions that can be answered by literature collected in databases. The persistent creation of new questions makes a thesaurus dynamic. Contrary to the belief that implies an end to the inclusion of new terms to thesauri, there will never be a point of saturation. New questions have been created since the existence of human beings and will continue because it is part of our fundamental nature.

Proliferation of documents in a geometrical way is an indication of answering new questions. Without answering new questions or only dealing with old questions, production of new literature becomes meaningless. Theoretically, almost any publication should solve a problem of knowledge. If every publication provides new questions, the count of thesauri terms should be more than or equal to the number of the documents in their corresponding databases. The number of terms in a thesaurus also corresponds to the level of their specificity. When including new terms in a thesaurus, the count of documents retrieved through a query should

be taken into consideration. Thus, thesaurus terms set very specific questions in their related classes. This act will avoid retrieving only few documents.

1.2.1 Linguistic structure of thesaurus

A thesaurus has also a linguistic structure. Syntax, semantics, and pragmatics aspects of a thesaurus have been discussed for decades. Syntax corresponds to the combining of words to form grammatical phrases or sentences. Semantics address the meaning of words and their combination to form the meaning of sentences. Pragmatics is concerned with the bridging between the sentence meaning and text meaning. In other words, it relates to how the users get the meaning of a sentence or utterance from its context.

Thellefsen, M. (2004) states that the concept of synonymy, hyponymy and meronymy are incorporated and expressed in a thesaurus by BT and NT, which indicate hierarchical relationships, while UF expresses equivalence relationships. In a text, the syntagmatic relations hold between words that collocate in a grammatical string and that have semantic affinities. The cross references (i.e. BT, NT, UF) do the same task in a thesaurus. They could be labelled as structural relations between terms.

Schwartz, I and Umstätter, W. (1999) described this kind of thesaurus as a semantic thesaurus. They describe a semantic thesaurus as being able to make a relationship between objects that derive from their meanings and appear in form of tokens and their relations. They add that many of these relations are clearly represented as hierarchical. Beside the mono- and poly-hierarchical parts, logical or functional relations are also recognized in such semantic thesauri today. The relations in a thesaurus can be illustrated by different syntactic methods. Semantic networks, frame-slot-structures, and graphs or neuronal images are common instances.

Semantics and pragmatics are also regarded in information retrieval. Fidel, R. (1991) depicts the requested topic by a user as the semantic part. A topic presents the subject matter that is of concern to the user. The purpose of a request is concerned with the pragmatics of the search. Different users that request information about the same topic may have different purposes. One user may be interested in just a few highly relevant citations while others' purpose is to get the most recent citations.

Because of semantic relations of thesaurus terms, a user can control his topic by the mean of a thesaurus and formulate the appropriate queries to conduct his search in its corresponding database. The descriptor fields consisted of the same semantic terms that appeared in a

thesaurus. If the topic is available in the thesaurus, information about the requested topic is available. On the other hand, the pragmatics of a user request can be clarified by the index terms assigned to the documents. All assigned terms together express the context of the indexed text. If a document is returned by a query, it means the context of the document and user request are linguistically the same. Thus, the study of the number of terms assigned to documents concern the pragmatics of texts that are reflected by thesaurus terms. To narrow the search by combining more terms yields fewer but more pertinent citations. The more terms are combined the closer the returned results satisfied the user request. In other words, context of user request and documents become closer to each other.

1.2.2 Similar problems of Conventional and automatic thesauri

Conventional and automatic thesauri have similar problems. Most research findings concerning construction and development of one type of thesaurus are applicable to the other one. For example, in the literature of automatic thesauri, we find citations that address the traditional controlled vocabulary. One by Lancaster, F. W. (1986) is about the logarithmic growth of controlled vocabularies. For instance, Chen, H. et al. (1996), Dorbin, Tobun Ng (2000) and Greenberg, J. (2001), who work on the field of automatic thesauri, take his finding into consideration. It reveals that the findings of every field can be used in another field.

1.2.3 MeSH as subject headings and thesaurus

Some literature differs between subject headings and thesauri. Taylor, A. G. (1992, p. 454 -5) reveals several differences between them:

1. Thesauri are composed of terms that represent single concepts, whereas many subject headings represent compound subjects.
2. The relationships between terms in thesauri are defined and displayed according to rules, whereas the relationships between subject headings are at best shown inconsistently.
3. Thesauri are usually limited to coverage of a particular discipline, whereas subject headings attempt to cover the entire realm of recorded knowledge.
4. There are international standard guidelines for the creation of thesauri, while there are none for subject heading lists.

Other differences can be also added into the list above:

5. Subject headings are used generally for post-coordinate indexing and thesaurus terms for pre-coordinate.
6. In general, thesauri have a hierarchical structure while subject headings lack it. This phenomenon has similarities with case “2” mentioned above by Taylor, A. G. (1992).

The question is whether MeSH is a thesaurus or rather a collection of subject headings. Its structure compromises the functions of the two forms of controlled vocabularies. Its usage is extended from the indexing of general types of documents (like books) to very special ones (like patents) and even non-print material.

The most distinct characteristic of a thesaurus is its hierarchical structure. “The core of MeSH is a hierarchical structure that consists of sets of terms” (Ijzereef, L.; Kamps, J. and De Rijke, M. (2005). There are fifteen general categories of headings at the top level. At deeper levels are more specific headings such as Brain infarction (sixth level of Diseases branch) or Dissociative Anesthetics (ninth level of Chemicals and Drugs). The hierarchy is an eleven-level tree structure that contains over 23,000 headings.

MeSH can be used for subject headings. It can be used for indexing of general materials (like books) and is pertinent for pre-coordinate indexing. This type of indexing performed by the use of sub-headings which are known as qualifiers in MeSH.

Understanding MeSH requires an understanding of its structure. It has three major components: the headings themselves, the qualifiers, and the Supplementary Concept Records. Main headings are the meat of the MeSH thesaurus. They are used to describe what a document is "about". MEDLINE uses the term MESH HEADING (MH) to indicate the topics discussed by the work cited. Qualifiers are known as sub-headings. They are used to refine the meaning of MH. Supplementary Concept Records are edited and added to MeSH daily, and preferred names in these records can be assigned to a special data element (Name of Substance) within the MEDLINE record of a citation. As implied by many of the names of data elements, the bulk of these records are related to chemicals and drugs.

1.3 Indexing

Indexing covers a broad area of activities. The general meaning of indexing in the field of documentation is the process of converting a collection of data and documents into a database. This thesis is concerned in particular with “subject indexing”. The indexing manual

of NLM ² expresses „The indexing or subject heading operation is the process of assigning to an article the headings from MEDICAL SUBJECT HEADINGS (MeSH), the MEDLARS authoritative vocabulary, which best describe the content and substance as written by the author“. Mai, J.-E. (2005) states: “The purpose of indexing is to determine the subject matter of documents and express the subject matter in index terms (e.g. descriptors, subject headings, call numbers, classification codes, or index terms) to make subject retrieval possible”.

A document can be indexed either by controlled vocabularies or by natural indexing techniques. A thesaurus (pertaining to controlled vocabularies) helps to control the subject matters of a document. It permits the indexers to assign only selected vocabularies. Subject indexing and a thesaurus are thus two related topics that should be considered together. Indexing through controlled vocabularies can be done either by human expertise or by machines. This thesis focuses on human indexing.

An indexer performs two principal steps: conceptual analysis and translation. Use of a thesaurus is part of the second step. Indexers try to translate the contents of a document into terms of controlled vocabularies (i.e. a thesaurus) and choose those that are permitted to be used. In this process, they bridge between the pragmatics of document contents and thesaurus terms. The selected terms which are assigned to the documents are called descriptors. Searchers can use descriptors to retrieve indexed materials that meet their needs.

1.3.1 Depth of indexing

One of the subjects discussed in this research is the average number of index terms assigned to the indexed documents, in particular, the exhaustivity and specificity of indexing. Anderson, J. D. (1997, p.37); Cleverland, D. B. and Cleverland, A. D. (2001, p.254) indicate their combined effect as the depth of indexing. Wellisch, H. H. (1991, p. 122) claims “[depth of indexing] is not, as often thought, just the equivalent of exhaustivity but is always a combination of exhaustivity and specificity which, when both are at a high level (a large number of terms each of which is also highly specific), results in the greatest possible indexing depth“.

Exhaustivity relates to how depth contents of documents are scanned and specificity addresses the topical broadness of vocabularies used for indexing. As these two terms are close to each other, specificity is explained in the following with regard to exhaustivity.

² See „Indexing Operation“ in the section of “References”

1.3.2 Exhaustivity of indexing

Jones, K. S. (2004) states “the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it”. This definition shows that the number of index terms given to documents indicates the level of exhaustivity of indexing. For example, Raghavan, V. V. et. al., (2004) state “when indexing is exhaustive, it results in a large number of terms assigned to reflect all aspects of the subject matter present in the document“. In fact, the index terms of documents can’t be counted as the only determinant factor of exhaustivity. Soergel, D. (1994) says “the average number of descriptors assigned to an entity in the database being studied is often used — somewhat naively — as a stand-in measure for exhaustivity. This would work if exhaustivity was the only determinant of the number of descriptors per document”. He represents then the determinant of the numbers of descriptors that was expressed by Maron, M. E. (1979). They are the properties of the entity being indexed, the degree of pre-combination, the correctness of indexing, and the indexing policy. He adds viewpoint and importance as two components of exhaustivity, “Viewpoint exhaustivity addresses the question: Are the facets or viewpoints useful for retrieval represented in the index language and thus available for retrieval? The degree to which this question can be answered with "yes" is viewpoint exhaustivity... Importance exhaustivity addresses the question: What is the importance threshold for the assignment of descriptors as prescribed in the indexing rules? For the indexer considering an entity this question takes the form: Which of the concepts associated with this entity are important enough to warrant indexing?”

As we see, other factors relate to the exhaustivity of indexing. Assigning more index terms can also increase the count but not necessarily the exhaustivity. For example, redundancy increases the number of index terms given to documents, but it leads to a poor indexing.

1.3.3 Specificity of indexing

In indexing, a topic should be indexed under the most specific term that entirely covers it (Lancaster, F. W. 1991, p.26). A thesaurus contains both narrower and broader terms that can cover a topic. An indexer tries to assign possibly the narrower of them. For example, the term “animals” is broader than “cats”. Thus “cats” has more specificity than “animals”. If an article discusses cats and the indexer assigns “animals” to it, the indexing will have poor specificity. The best selection in this simple example is the term “cats”. Furthermore, if this article is

assigned to both of the two above mentioned terms, this will create redundancy. Redundancy occurs when unnecessary and ineffective terms are assigned to indexed materials. Thus, a high exhaustivity may occur because of assigning unnecessary and ineffective terms and can't always be an indicator of the quality of indexing.

Poor or good indexing directly influences retrieval. A relevant retrieval relates partly to assigning broader or narrower terms. Documents indexed by broader terms, logically, can satisfy recall preferences and narrower ones the precision. This is also approved by Svenonius, E. (1971) in his study. Giyeong, K. (2006) studied the relationship between specificity and relevance of retrieved materials by users' judgment. He regarded term-document specificity, which is a relationship between an index term and the document indexed with the term. The results show that the relevancy between these two variables is statistically significant from the point of users' judgment.

Jenuwine, E. S. and Floyd, J. A. (2004) used the terms specificity and sensitivity together. As their focus was on retrieval, they defined them from this point of view: "Sensitivity is the ability of a search to retrieve relevant articles. Specificity is the ability of the search to exclude irrelevant articles". On their definition, specificity of indexing is an effort to prevent the retrieval of irrelevant documents. They found that the sensitivity of searching through MeSH terms ranged between 5% to 36% and the specificity varied from 85% to 99%. In their findings, index terms (e.g. MeSH headings) result in higher specificity and prevent the retrieval of irrelevant articles, but they have less ability to retrieve all of relevant ones.

Specificity of indexing relies not only on indexers and indexing policy but on specificity of thesaurus vocabulary as well. If a general thesaurus, for example, is used for indexing of a narrower field, the specificity will reduce.

2 Materials and Methods

This dissertation consists of three related studies:

1. Growth of Medical Subject Headings (MeSH) over the years.
2. Use distribution of MeSH headings in MEDLINE.
3. Factors affecting the number of MeSH headings assigned to the MEDLINE documents:
 - i. Length of documents.
 - ii. Presence of abstracts within documents.
 - iii. Language of documents.
 - iv. Date of entering of documents into MEDLINE.
 - v. Priority of journals for indexing.
 - vi. Journal Impact Factor (JIF).

As the above outlines illustrates, this work concentrates on MEDLINE as a database and MeSH as its corresponding thesaurus. Thus, we decide to use PubMed to get the information needed for conducting the current research.

2.1 PubMed

“PubMed, available via the NCBI, was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), located at the U.S. National Institutes of Health (NIH). Entrez is the text-based search and retrieval system used at NCBI for services including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others. PubMed provides access to citations from biomedical literature. LinkOut provides access to full-text articles at journal Web sites and other related Web resources. PubMed also provides access and links to the other Entrez molecular biology resources. Publishers participating in PubMed electronically submit their citations to NCBI prior to or at the time of publication. If the publisher has a web site that offers full-text of its journals, PubMed provides links to that site as well as biological resources, consumer health information, research tools, and more. There may be a charge to access the text or information. In addition, PubMed provides a Batch Citation Matcher, which allows users to match their citations to PubMed citations using bibliographic information such as journal, volume, issue, page number, and year”³.

³ See the source of quote under „**PubMed & MEDLINE**“in References section.

2.1.1 PubMed Coverage

“PubMed provides access to bibliographic information that includes MEDLINE, OLDMEDLINE, as well as:

- The out-of-scope citations (e.g., articles on plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and chemistry journals, for which the life sciences articles are indexed for MEDLINE.
- Citations that precede the date that a journal was selected for MEDLINE indexing.
- Some additional life science journals that submit full text to PubMedCentral and receive a qualitative review by NLM”.

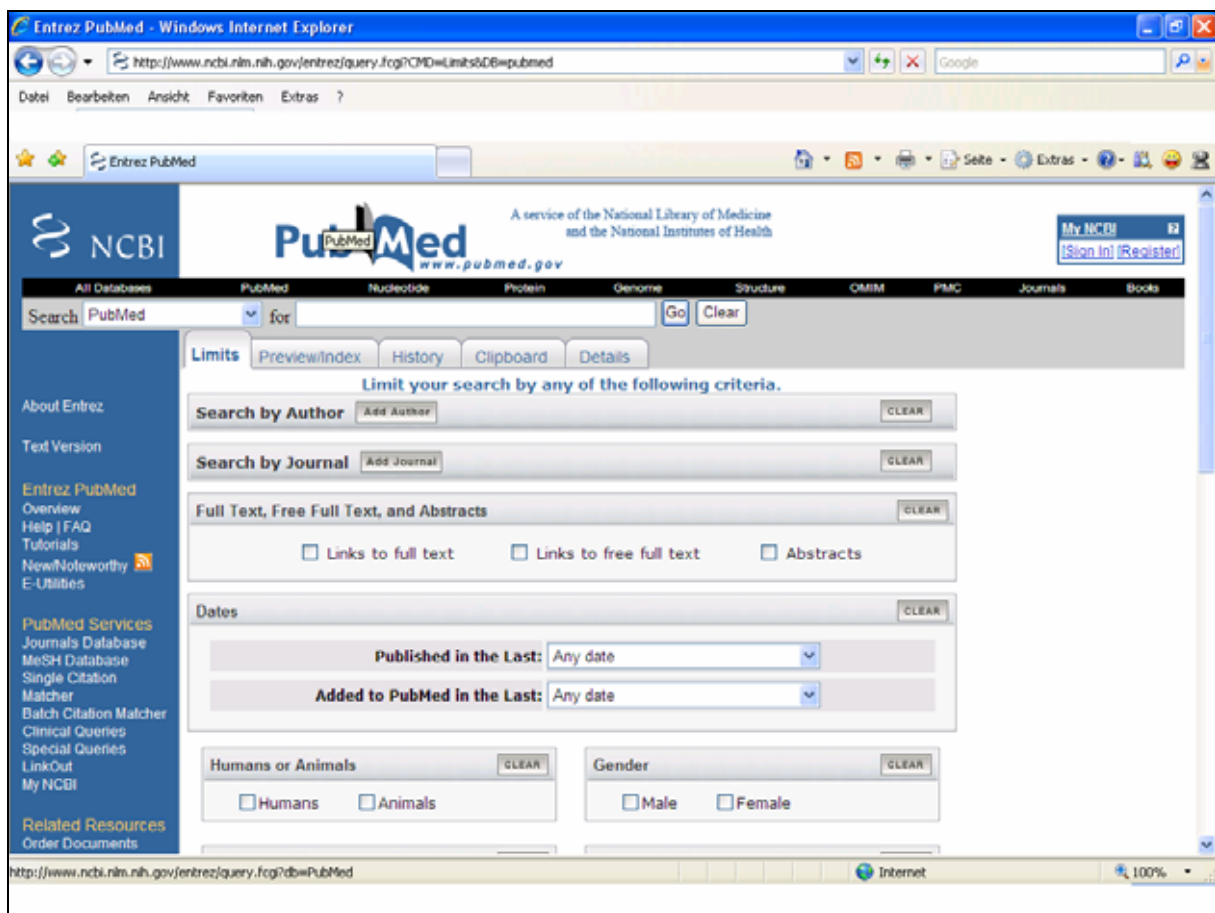


Figure 1: Entrez PubMed Homepage.

2.2 MEDLINE

MEDLINE is the NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences. MEDLINE contains bibliographic citations and author abstracts from more than

5,000 biomedical journals published in the United States and 80 other countries. The database contains over 16 million citations dating back to the mid-1950's. Coverage is world-wide, but most records are from English-language sources or have English abstracts.

2.2.1 MEDLINE format

One of the possibilities within PubMed allows for viewing and saving records in several formats including MEDLINE format. It was suited to this work, because it shows the bibliographic fields in the separated lines and introduces them with abbreviated labels and it eases the text processing through computer programming:

```
PMID- 16610373
OWN - NLM
STAT- MEDLINE
DA - 20060413
DCOM- 20060525
PUBM- Print
IS - 1055-3134 (Print)
VI - 69
IP - 1
DP - 2006 Spring
TI - How to write resolutions.
PG - 116-25
FAU - Smith, Beth
AU - Smith B
LA - eng
PT - Journal Article
PL - United States
TA - Tenn Nurse
JT - Tennessee nurse / Tennessee Nurses Association.
JID - 9102869
SB - N
MH - Humans
MH - *Lobbying
MH - Societies, Nursing/*organization & administration
MH - Tennessee
MH - *Writing
EDAT- 2006/04/14 09:00
MHDA- 2006/05/26 09:00
PST - ppublish
SO - Tenn Nurse. 2006 Spring;69(1):16.
```

Figure 2: An example of MEDLINE Format.

2.3 Medical Subject Headings (MeSH)

“The Medical Subject Headings comprise NLM's controlled vocabulary used for indexing articles, for cataloging books and other holdings, and for searching MeSH-indexed databases, including MEDLINE.

MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. MeSH organises its descriptors in a hierarchical structure so that broad searches will find articles indexed more narrowly. This structure also provides an effective way for searchers to browse MeSH in order to find appropriate descriptors.

The MeSH vocabulary is continually updated by subject specialists in various areas. Each year hundreds of new concepts are added and thousands of modifications are made.”⁴

2.3.1 Growth of MeSH

To find how MeSH grown over time, a sample of about 948,000 MEDLINE records was used. The sample was yielded from querying the term “up” in PubMed. The date of searching was 15.10.2006.

As mentioned before, this amount of records contains almost 23,000 different terms of MeSH. The number of terms yielded from the search was 23,198, but searching them the following week revealed that about 800 of them were deleted from MeSH because of infrequent usage and other policies of NLM.

PubMed results are initially displayed in reverse chronological order of the Entrez date, i.e., last in, first out. They are saved as they are displayed. It was not possible to sort them in normal chronological order, so the records were saved on the reverse (default) order. Their order was changed to the normal order by writing a program in Delphi, so that the record that were entered into MEDLINE earlier were near the beginning of the file and those entered more recently were at the end. The sort order of the file was necessary in this part of study, because the growth of thesaurus could be determined by the initial appearance every of MeSH terms in MEDLINE. In this case, we presumed that those terms that were used in MEDLINE earlier were added earlier. This method let us not only find the order of appearing terms in MESH, but the number of documents needed to produce the amount of thesaurus terms as well. Additionally, the growth of MEDLINE can be compared to the growth of MeSH.

The thesaurus is dynamic and interacts with the corresponding database. It means, new terms will be added to the thesaurus, when none of its terms can describe the content(s) of new documents in database. Thus, when we study the development of thesaurus terms, we learn that emerging new documents with new contents produce new thesaurus terms.

All headings in the field “MH - “ were checked to see if a term was used in the prior records or not. If the answer was negative, one number was added into the amount of the thesaurus terms. If the term was used already, it was ignored for calculation. In addition to that, the number of documents in which the term was used for the first time was noted. For example, 1,000 different MeSH headings are used by the first 900 documents. In record 901 we find out that a term is used for the first time. We can assert that the 1,001st term is produced by the 901st document and so on.

2.4 Use distribution of MeSH headings in MEDLINE

MeSH contains over 23,000 terms. By testing MEDLINE, we considered that a randomly selected sample of about 1,000,000 records contains almost all of the different MeSH terms. This amount of headings was gathered for a list from the MeSH headings field, labeled by “MH - “ (see Figure 2). In the following section (2.6, there are some notes about the Delphi programm), it will be explained how this could be done by computer programming.

The distribution of used MeSH headings in MEDLINE was studied in four different intervals (1965 - 1970, 1965 - 1980, 1965 - 2000, and 1965 - 2006). To avoid the inclusion of documents entered in pre-MEDLINE time, the analysis was focused on the usage of terms from 1965. Searching PubMed by limiting the searches to the above years revealed about 7,700, 13,000, 20,000, and 23,000 distinct terms that were added into MeSH respectively up to 1970, 1980, 2000, and 2006.

For getting the usage of terms, they were searched through PubMed and the number of returned citations was noted after each search. The total number of searches was 63,700. It was equal to the number of distinct terms used between 1965 and the four above mentioned years (i.e. “1965 – 1970”, “1965 – 1980”, “1965 – 2000”, and “1965 – 2006”). In addition to limiting the number of searches to the intended periods, all of them were conducted also by syntax [MH:noexp] to turn off the automatic inclusion of the more specific terms. For example: “Ethics, Medical [MH.noexp]”.

Subheadings are ignored in the current work. They can be determined through a slash (“/”) which indicates that the terms after it are sub-headings.

⁴ See the source of quote under „Medical Subject Headings® - Overview“ in References section.

2.5 Factors effecting the number of MeSH headings per article in MEDLINE

Two samples were taken from PubMed to study the factors that affect the number of MeSH headings assigned to the MEDLINE documents:

1. A sample of 989,281 records by querying two keywords: “Humans AND Medical”. The search conducted at 16.03.2006. It was limited into the following features:
 - i. English language documents.
 - ii. Journal article.
 - iii. Entrez date between 1965 and 2005.
2. A sample of 574,242 records without querying any keywords at 04.09.2006. The search was done only on the following limit features:
 - i. German language documents.
 - ii. Journal articles.
 - iii. Entrez date between 1965 and 2005.

Other limitations were done on the records by Delphi programming that PubMed could not do:

The documents consisting of more than thirty pages were excluded from both of the two samples.

1. The documents that were not indexed by NLM but were entered into MEDLINE.
2. The remaining items were 955,697 in the first sample and 497,313 in the second one.

2.5.1 Determining the text tokens and types

The other effort was focusing on the length of articles regarding the types and tokens. Nine full-text articles with different lengths were processed to determine the impact of article's lengths measured by the amount of words. The bibliographic information of these nine articles is as follows:

Bolding, J. Neurosci. and Biedenkapp (2006), What Can Immediate-Early Gene Expression Tell Us about Spatial Memory Retrieval?, *The Journal of Neuroscience*, 26(6):1659-60.

Feldser, Feldser, Margaret A. Strong, and Carol W. Greider (2006), Ataxia telangiectasia mutated (Atm) is not required for telomerase-mediated elongation of short telomeres, *PNAS*, 103(7): 2249-2251.

Boldrin, F et al. Metallothionein gene from Tetrahymena thermophila with a copper-inducible-repressible promoter. *Eukaryot Cell* 5(2):422-5.

Hittinger, Chris Todd, Antonis Rokas, and Sean B. (2004), Carroll Retention and Loss of Amino Acid Biosynthetic Pathways Based on Analysis of Whole-Genome Sequences *Eukaryot. Cell*, 5(2): 272 - 276.

Araujo, Luiz Felipe Bittencourt de et al. (2006), Effect of conjugated equine estrogens and tamoxifen administration on thyroid gland histomorphology of the rat, *Clinics*;61(4):321-326.

Ghose, J. Neurosci (2006) Steering by Hearing: A Bat's Acoustic Gaze Is Linked to Its Flight Motor Output by a Delayed, Adaptive Linear Law, *the Journal of Neurosciences*, 26(6):1704-1710.

Yin, Zheng Qin(2006), Pre- and post-critical period induced reduction of Cat-301 immunoreactivity in the lateral geniculate nucleus and visual cortex of cats Y-blocked as adults or made strabismic as kittens, *Molecular Vision*, 12: 858-866.

Shi, Yang and Iryna M. Ethell (2006), Integrins Control Dendritic Spine Plasticity in Hippocampal Neurons through NMDA Receptor and Ca²⁺/Calmodulin-Dependent Protein Kinase II-Mediated Actin Reorganization, *The Journal of Neuroscience*, 26(6):1813-1822.

Carole, Torsney and Macdermott, Amy B. (2006), Disinhibition opens the gate to pathological pain signaling in superficial neurokinin 1 receptor-expressing neurons in rat spinal cord, *The Journal of neuroscience*, 26(6): 1833-1843.

To find the amount of tokens and types in each of the nine above articles, the following processes were performed:

1. The full-texts of above mentioned articles saved in the PDF format.
2. They were transferred into Microsoft Word one by one.
3. Using Word's facilities, every space between words replaced with the carriage/return character. This caused every word to be placed on the separate lines.
4. They were copied onto Microsoft Excel.
5. Sorted alphabetically.
6. The numbers and non-alphabetic characters were deleted from the list.
7. The frequency of words determined by Excel's commands.
8. The total number of word frequencies counted as an amount of text tokens.
9. Vocabularies sizes of texts taken as the amount of types.

2.5.2 Determining number of pages per article

To determine the length of articles, the field "PG" was processed. In the Figure 2, following the PG label, we see "116-25". That means that the article covers pages 116 through 125 of the mentioned journal. In this case we may meet some possibilities:

1. The articles consisted of only one page. In this case we see only a number, like “PG - 11”. It tells us that this article appeared only on one page and was published on page eleven of the source journal.
2. The article consisted of two or more pages. In the case of the article above (Figure 2), the first page is separated from the last page by inclusion of a dash (“-“) between them. The first part is always introduced but the last part (the number following the dash) depends on the length of articles and the number of digits used in the first part. To determine the length, we have to consider several possibilities and for each possibility we need a different formula:
 - I. The first and last page consisted of one digit (i.e. “PG - 3-6”). (second part – first part + 1 = length of article).
 - II. The first part consisted of one digit and the last of two and more digits (i.e. “PG - 3-16”). (second part – first part + 1 = length of article).
 - III. The first part consisted of two or more digits and second of one digit (i.e. “PG - 23-9”). (second part – the last digit of the first part + 1 = length of article).
 - IV. The first part of two digits and second of three and more (i.e. “PG - 95-106”). (second part – first part + 1 = length of article).
 - V. The first part of three digits and second of four and more (i.e. “PG - 953-1006”). (Second part – first part + 1 = length of article).

When determining the number of pages of articles, we may encounter some mistakes made by NLM’s typists. They can be distinguished in the following cases:

1. if the last page number was less than the first, the result of article length is a negative number, and
2. if the result of article length is a large number (i.e. 320).

To reduce the errors, our program excluded those records whose lengths were negative or more than 30 pages. Articles which didn’t appear continuous and were introduced in two places of a journal were excluded as well.

2.5.3 Determining the presence and form of abstracts

The presence of the label “AB - ” in a MEDLINE record indicates that the abstract of the corresponding document is included as well. In addition, we had to differentiate between structured and unstructured abstracts. Based on the earlier work conducted by Harbourt, A.

M.; Knecht, L. S. and Humphreys, B. L. (1995), an abstract that bears one of the following terms in upper case is considered as structured:

Table 1: Uppercase words, their existences within an abstract indicate that they are structured.

OBJECTIVE	STUDIES	PATIENT	PURPOSE
SYNTHESIS	IDENTIFICATION	EXTRACTION	SUBJECT
MEASURE	RESULT	BACKGROUND	OUTCOME
STUDY	GOAL	PARTICIPANT	DESIGN
SELECTION	SETTING	CONCLUSION	TYPE
MEASUREMENT	METHOD	AIM	END
DATA	INTERVENTION	MAIN	

2.5.4 Determining journal titles

The label “TA - ” shows the title of journals in which the articles are published. It helps to determine the source of articles in abbreviated form. We didn’t face a problem when programming, because this field was available for all records and the phrases following the label could simply introduce the journal titles.

2.5.5 Determining the number of MeSH headings per article

The focus of this research was on the MeSH headings. They are distinguished by the label “MH - “ as a repeated field, so that every repetition contains the “MH - “ on a separate line (see Figure 2). Major headings which are introduced by the asterisk sign (“*”) prior to MeSH headings were not weighted more than other headings in this work.

The main focus was to determine the number of MeSH headings within documents. This was done by counting the number of lines preceding the “MH - ” lable. The final task in this field was to exclude the check tags when determining the number of index terms (MHs) per article. A check tag is defined as a concept of a 'tag' which must be considered routinely for every article indexed. On the MEDLINE citation, the check tags are usually displayed in the MeSH term field. The following check tags were excluded by the program:

Table 2: List of MeSH check tags.

Humans	Aged, 80 and over	English Abstract	History, 20 th Century
Male	Adolescent	In Virto	History, 19 th Century
Female	Pregnancy	Cricetinae	History, 18 th Century
Infant	Animals	Research Support, N.I.H., Intramural	History, 17 th Century
Child	Mice	Research Support, U.S. Gov' t. P.H.S.	History, 16 th Century
Child, Preschool	Rats	Research Support. N.I.H.. Extramural	History, 15 th Century
Adult	Cats	Research Support, U.S. Gov' t. Non-P.H.S.	
Middle Aged	Dogs	Research Support, Non-U.S. Gov' t	
Aged	Comparative Study	History, 21 th Century	

In Figure 2 (“an example of MEDLINE format”), we see that the label “MH - ” is repeated five times, among them the term “Humans” belongs to the check tags, it is not a real MeSH heading. Thus, the number of index terms assigned to this article in Figure 2 should be counted as four headings instead of five.

2.5.6 Determining the Entrez date

Instead of concentrating on the date of publication, the inclusion date of a citation in MEDLINE was taken as a parameter in this work. NLM marks the inclusion date of documents in MEDLINE as “Entrez Date”. Because of this, we take the term used by NLM.

Some MEDLINE indexing policies may change over the years. That is why the Entrez date was selected.

This field is preceded by the “EDAT- “ label, so that the complete date of inclusion is displayed, including year, month, day and even hour and minute in some cases (i.e. EDAT-2006/04/14 09:00). The year was the only part that was taken into account and other information in this field was eliminated.

2.5.7 Determining the indexing priority of Journals

NLM has given the journals a priority number between one and three for in-depth indexing of their articles. They are only for the indexers’ use and don’t appear in MEDLINE, so this information is not available to others. Despite this, finding the priorities of journals is still possible. We can assume that journal articles with higher priorities get more index terms. It is enough to determine the average of terms assigned to the articles of journals per page. We took only those journals into consideration from which more than 500 articles were indexed in the sample. 454 journals fulfilled this condition.

The total number of index terms assigned to the articles of an instance journal was divided by the total number of pages of its indexed articles. This enables us to determine the average number of index terms per page. If we sort the results decreasingly, their depth of indexing will reduce downwards.

2.5.8 Determining Journal Impact Factor (JIF)

In the case above, the concentration was only on journals which indexed more than 500 times in our sample. 454 journals fulfilled this condition. They were then compared with the list of the JIF presented in the Journal Citation Report (JCR) for the years 2003 and 2004. Only 246 journals from the 454 were covered by the JCR.

2.6 Some notes about programming by Delphi

Delphi is a computer programming language developed by Borland. As the focus of the current work was on the large samples of MEDLINE records, analyzing them manually was not possible. Delphi has some features that facilitate the processing of texts. It makes possible automatic searches in the databases on the WWW as well.

2.6.1 Processes for sorting records

As explained above, sort order of PubMed search results is in reverse chronological order of Entrez date. But for determining the growth of MeSH, we need the chronological order of records. Because of this, they need to be sorted again.

To do this, all of the records were brought into a Delphi memo. The program read the lines from the bottom of the memo upwards and added them to another memo until it reached the line which began with "PMID- ". This showed that a record was completed. The same process was done on the second memo, but the lines were outputted to a text file. The above process was repeated until the cursor reached the first line of the first memo.

One may claim that this process could be done by creating a database of records. But the limits of Delphi professional Edition cause an error when exceeding 40,000 records.

2.6.2 Processes for determining distinct MeSH headings

Following the saving of 948,000 randomly selected records from PubMed, a database of MeSH headings was created. The program looked for the MeSH headings field determined by "MH - ". Every time the program found the MeSH headings field, it searched the headings within its database. If the search result returned "zero", the heading was added into database. This process was repeated until the end of the sample. Finally, 23,198 different headings were included into its database.

2.6.3 Processes for determining the growth of MeSH

After determining distinct MeSH headings, we can find out what number of documents produces what number of new MeSH headings. To do this, two databases were created by Delphi. The first consisted of two integer fields: "Number of Records" and "Number of Terms". The second of one string field is called: "MeSH Heading".

The program began to read the file line by line. By finding the lines containing "PMID - ", it determined that this was a new record and added one to an integer variable to record the record number. By finding the MeSH headings field (MH -), the MeSH term was extracted from it and searched in the second database. If the search returned false, it was added into the database. And then the count of distinct terms was recorded in the field "Number of Terms" of the first database. It showed how many distinct terms were created up to the nth record. If

the search returned “true”, the field “Number of Terms” was filled in by the same value of its prior record and the value of “Record Number” field was equal to the number of record that was on processing.

2.6.4 Processes for determining the use distribution of MeSH headings

The frequency of headings used in MEDLINE was determined by automatic searches in PubMed. As explained above, PubMed was searched 63,700 times. The internal components of Delphi were not suited to be used for this aim. Overbyte (www.overbyte.be) developed a component named “ICS”, which was installed in Delphi. By expanding one of ICS’s examples named “HTTPDMO”, it was possible to search automatically in PubMed. The major part of URL for every search was the same, only the part substituted by **** in the following instance was replaced with the new heading that was taken from the list of 23,198 headings. To limit the searches in the periods of “1965-1970”, “1965-1980”, “1965-2000”, and “1965 – 2006”, the beginning and end years of limited periods were replaced with the places that are underlined in the following URL:

*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=PureSearch&db=pubmed&details_term=%22****%22%5BMeSH%20Terms%3A%20exp%5D%20AND%20%28%221965%22%5BEDAT%5D%20%3A%20%221970%22%5BEDAT%5D%29*

2.6.5 Processes for determining the average number of MeSH headings per article

This dissertation concentrates also on several factors that affect the average number of MeSH headings assigned to the articles by indexers. Those factors are:

1. Length of articles.
2. Presence of abstracts within articles and also the form of abstracts.
3. Language of articles.
4. The inclusion dates of articles were added to MEDLINE.
5. Priority of journals for in-depth indexing.
6. Journals Impact Factor (JIF).

To prepare data for statistical analyses three different databases were created:

1. Page database consisted the fields:

- i. number of pages
- ii. number of articles
- iii. total number of MeSH terms used by articles
- iv. average number of MeSH terms
- v. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the current article, M = the average of MeSH headings and i = number of pages)
- vi. number of articles without abstracts
- vii. total number of MeSH terms used by articles without abstract
- viii. average number of MeSH terms of articles without abstract
- ix. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article without abstract, M = the average of MeSH headings without abstract and i = number of pages)
- x. number of articles with abstract
- xi. total number of MeSH terms used by articles with abstract
- xii. Average number of MeSH terms of articles with abstract
- xiii. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article with abstract, M = the average of MeSH headings with abstract and i = number of pages)
- xiv. average of MeSH headings with abstract and i = number of pages)
- xv. number of articles with normal abstract
- xvi. total number of MeSH terms used by articles with normal abstract
- xvii. average number of MeSH terms of articles with normal abstract
- xviii. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article with normal abstract, M = the average of MeSH headings with normal abstract and i = number of pages)
- xix. number of articles with structured abstract
- xx. total number of MeSH terms used by articles with structured abstract
- xxi. average number of MeSH terms of articles with structured abstract
- xxii. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article with structured abstract, M = the average of MeSH headings with structured abstract and i = number of pages)

2. Date database consisted the fields:

- i. Entrez date
- ii. number of articles

- iii. total number of MeSH terms used by articles
- iv. average number of MeSH terms
- v. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the current article, M = the average of MeSH headings and i = Entrez date)
- vi. number of articles without abstract
- vii. total number of MeSH terms used by articles without abstract
- viii. average number of MeSH terms of articles without abstract
- ix. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article without abstract, M = the average of MeSH headings without abstract and i = Entrez date)
- x. number of articles with abstract
- xi. total number of MeSH terms used by articles with abstract
- xii. Average number of MeSH terms of articles with abstract
- xiii. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article with abstract, M = the average of MeSH headings with abstract and i = Entrez date number of articles with abstract)
- xiii. number of articles with normal abstract
- xiv. total number of MeSH terms used by articles with normal abstract
- xv. average number of MeSH terms of articles with normal abstract
- xvi. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article with normal abstract, M = the average of MeSH headings with normal abstract and i = Entrez date)
- xvi. number of articles with structured abstract
- xvii. total number of MeSH terms used by articles with structured abstract
- xviii. average number of MeSH terms of articles with structured abstract
- xix. $\sum (X_i - M)^2$ (where X = number of MeSH headings of the article with structured abstract, M = the average of MeSH headings with structured abstract and i = Entrez date)
- xx. total number of pages
- xxi. average number of pages
- xxii. $\sum (X_i - M)^2$ (where X = number of pages of the article, M = the average number of pages and i = Entrez date)
- xxiii. total years delayed in indexing of articles of corresponding Entrez year
- xxiv. average years delayed in indexing articles of corresponding Entrez year

3. Source database consisted the fields:

- i. Source title
- ii. number of articles
- iii. total number of MeSH terms used by articles
- iv. average number of MeSH terms
- v. $\sum (X_i - M)^2$ (where X = number of MeSH headings used by the current source, M = the average of MeSH headings and i = source number)
- vi. total number of pages
- vii. average number of pages
- viii. $\sum (X_i - M)^2$ (where X = number of pages of the article, M = the average number of pages and i = source number)

The first fields of databases were considered as the base fields with the values of other fields corresponding to them. For example, in the “Source Database”, the source title was considered as the base field and the values of second field were the number of articles of the corresponding source that were indexed in MEDLINE. The values of the third field were the total number of MeSH headings assigned to the articles of the corresponding source and so on.

For programming, the following fields of MEDLINE records were taken into consideration:

1. “PMID- “, beginning of records,
2. “PG - “, number of pages,
3. “EDAT- “, Entrez date,
4. “PDAT- “, publication date,
5. “TA - “, Title of source,
6. “MH - “, MeSH heading,
7. “SO - “end of records”.

After considering each of above fields, the program chose the corresponding task(s), calculated the variables, and then appended them to the related databases and fields.

The values of databases were transferred into MS-Excel. The two following tables represent the structure of the data yielded by the programs.

Table 3: A sample of the database based on the number of pages. This sample represents only two parts from the five: Not-abstracted and abstracted parts.

Number of pages	Without Abstracts				With Abstracts			
	No. of documents	No. of MeSHs	Average of MeSHs	$\sum (X_i - M)^2$	No. of documents	No. of MeSHs	Average of e	$\sum (X_i - M)^2$
1	15,514	85,907	5.54	106,908.8	1,207	7,123	5.9	8,451.27
2	38,286	228,788	5.98	305,225.5	22,371	140,566	6.28	164,983.23
...

The table above is a small sample from the first database. As is shown, the first column is titled as “Number of pages”, with the other columns arranged on the base field. The first line shows that articles with only one page were in most cases (column 1) 15,514 articles, without abstracts (column 2). These articles received 85,907 MeSH headings in all (column 3) and the articles with one page length received 5.54 MeSH terms on the average (Column 4) and x_i (column 5) represents the number of index terms assigned to the each article. “M” is the average determined in the prior column.

The second database looked like the above table, except it was based on the year of entering articles into MEDLINE.

The last database was based on the source of articles. It contained two parts. The following sample illustrates these two parts.

Table 4: A sample of database that based on the journal titles.

Journal Title	Whole Sample				Pages			
	No. of Documents	No. of MeSH	Average of MeSH	$\sum (X_i - M)^2$	Total pages	No. of pages pro article	Average of pages	$\sum (X_i - M)^2$
J Biol Chem	7,721	122,656	15.89	301,017.70	57,571	7.47	32,943.58	
Cancer	4,504	41,146	9.14	64,865.38	31,677	7.00	30,460.07	
JAMA	4,372	38,671	8.85	66,086.17	19,524	4.47	24,477.85	

The table above is a excerpted from the last database. The first column is the base. We can define the information about the “J Biol Chem” as follows:

7,721 articles of the first journal (J Biol Chem) were indexed in MEDLINE. This number represents only the articles that were in our sample. These articles received 122,656 MeSH headings. On average, every article of this journal was indexed by 15.89 terms.

The second part of table shows that the total number of articles in the journal presented in our sample (file) consisted of 57,571 pages (column 6). Each article consisted of 7.47 pages on average (column 7).

3 Results

In this chapter, we will observe the results in three sections:

3.1 Growth of Medical Subject Headings (MeSH)

In the following, the growth of Medical Subject Headings will be investigated. We will focus on the relationship between the increasing number of MEDLINE documents and the growing number of descriptors in MeSH. In seeking the relationship, we will try to discover, when adding the n^{th} document into database, how many distinct index terms have been produced through its indexing. Thus, the terms “new headings”, “headings used for the first time”, “appearing new headings”, “increasing thesaurus terms” will be used interchangeably.

3.1.1 Cumulative Growth of Medical Subject Headings (MeSH)

The following figure illustrates how the number of distinct MeSH headings has increased when the number of documents added into MEDLINE increased. The records returned through querying PubMed were sorted by the date when they were added to MEDLINE. Every 50 records were taken as a point. This means, it was considered how many new terms were used by the first 50 documents of the sample and how many by the second 50 documents and so on.

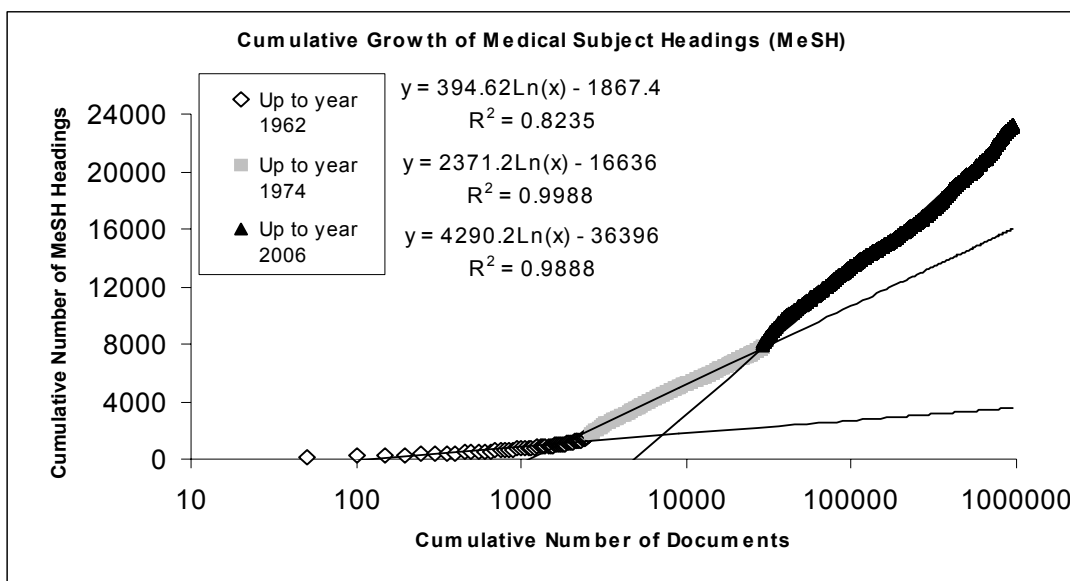


Figure 3: Cumulative growth of the Medical Subject Headings (MeSH). The x-axis is scaled logarithmically.

The figure above shows how many new MeSH headings were used for indexing by indexers when the number of documents added into MEDLINE increased. The X-axis is scaled logarithmically to get straight trend lines. Note that the documents are sorted by the date added to MEDLINE.

As the curves illustrate, the growth of MeSH hasn't followed a single function. We see three different changes in the growth. For the first 2,600 documents of the sample we see that the first 1,656 headings have grown logarithmically “ $y = 394.62 \text{ Ln}(x) - 1,867.4$ ” and from that point to the 28,850th record, the number of unique headings reached 7,853 and the function has changed to “ $y = 2,371.2 \text{ Ln}(x) - 16,636$ ”. Finally, the number of MeSH headings has reached 23,199 headings when the number of records in the sample reached 949,198. In this part, the function is also logarithmic ($y = 4,290.2 \text{ Ln}(x) - 36,369$).

The 2,600th record of the sample was added into MEDLINE in the last month of 1962, the 28,850th in the last month of 1974, and the 949,198th in the last months of the year 2006. We have thus obtained three phases of MeSH growth. The first phase lasts from the beginning up to 1962, the second one between the years 1963 and 1974, and finally the third one covers the years between 1975 and 2006.

The findings show that the exponents of the three functions mentioned above have grown linearly:

$$y = 1947.8x - 1543.6 \text{ and } R^2 = 0.9999 \quad (\text{i.})$$

Due to this fact, if we disregard the cutting points and apply the exponents of every logarithmic function for determining the growth of whole MeSH and then calculate the percentile growth of the yielded results, we will observe that the outcomes of all three exponents produce exactly the same values:

$$y = [394.62 \text{ Ln}(d_i) / \sum 394.62 \text{ Ln}(d_i)] \times 100 \quad (\text{ii.})$$

=

$$y = [2371.2 \text{ Ln}(d_i) / \sum 2371.2 \text{ Ln}(d_i)] \times 100 \quad (\text{iii.})$$

=

$$y = [4290.2 \text{ Ln}(d_i) / \sum 4290.2 \text{ Ln}(d_i)] \times 100 \quad (\text{iv.})$$

where d_i is number of documents. Thus, the above equations could be expressed mathematically:

$$y = 2371.2 \text{ Ln}(d_i) / 394.62 \text{ Ln}(d_i) = 2371.2 / 394.62 = 6.01 \quad (\text{v.})$$

$$y = 4290.2 \text{ Ln}(d_i) / 2371.2 \text{ Ln}(d_i) = 4290.2 / 2371.2 = 1.81 \quad (\text{vi.})$$

The equations “i – vi” make it possible to predict some phenomena: The equation “i” (linear growth of the exponents) shows that the exponent will change from 4290.2 to “4290.2 + 1947.9 = 6238.1” in the future. The equations “iv & v” reveal also that the growth speed of the second function is 6.01 times higher than that of the first function. The growth speed of the third function is 1.81 times higher than that of the second, and it will be 1.45 times greater than that of the third function, when the exponent reaches to 6238.1 in the future (6238.1 / 4290.2 = 1.45).

We can determine some other phenomena of the thesaurus growth through the median of terms. Let us to call each change of the functions a phase and represent their medians on a graph:

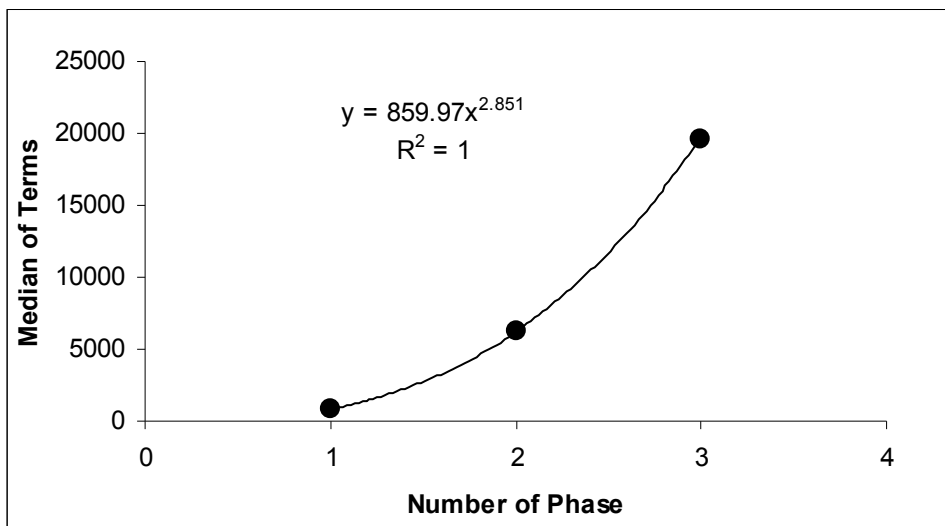


Figure 4: Relationship between the phase number of MeSH growth and the medians of terms.

The medians of terms within each phase are sequentially “857”, “6,263” and “19,596”. The figure above reveals that the correlation between them and the number of phases is very significant:

$$y = 859.97x^{2.851} \text{ and } R^2 = 1 \quad (\text{vii.})$$

Thus, we can predict the median of terms in each phase through the equation “vii” (i.e. $859.97 \times 4^{2.851} = 36,851^{\text{th}}$ term in the fourth phase).

On the other hand, the relationship between the cumulative number of the last terms in each phase and the medians is linear.

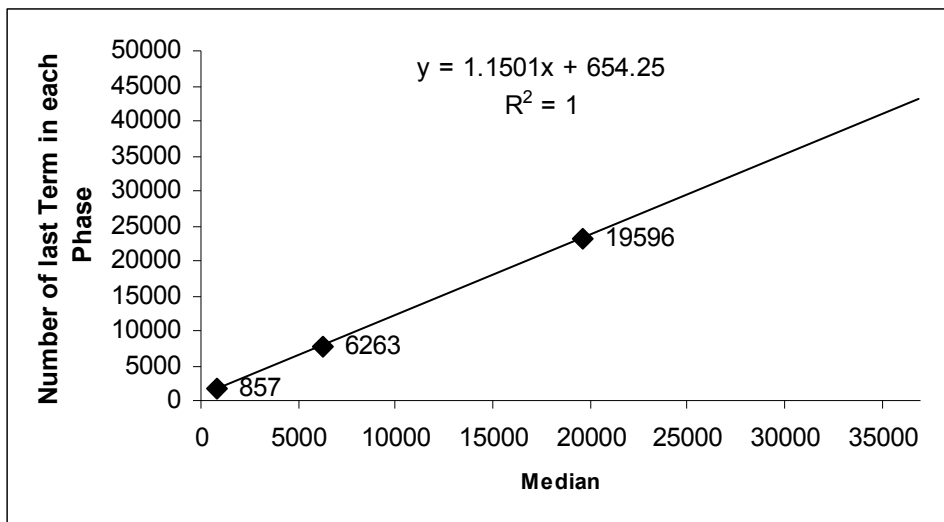


Figure 5: Relationship between the median of each phase and the last term of the phase.

The figure above shows the most probable correlation:

$$y = 1.1501x + 654.25 \text{ and } R^2 = 1 \quad (\text{viii.})$$

Thus, we can predict the last term in each phase and beginning of the next phase by replacing the “y-variable” of the equation “vii” with the “x-variable” of the equation “viii”. For example, we found that the median of terms for the fourth phase should be “36,851”, then the last term in this phase will be the “ $1.1501 \times 36,851 + 654.25 = 43,037^{\text{th}}$ ” term which is following. A new phase will begin.

3.1.2 Growth of MEDLINE vs. growth of MeSH

The question that should be still asked is how the documents of MEDLINE have increased against the growing of MeSH. We can answer this question by exchanging the places of the x- and y-axes of Figure 3:

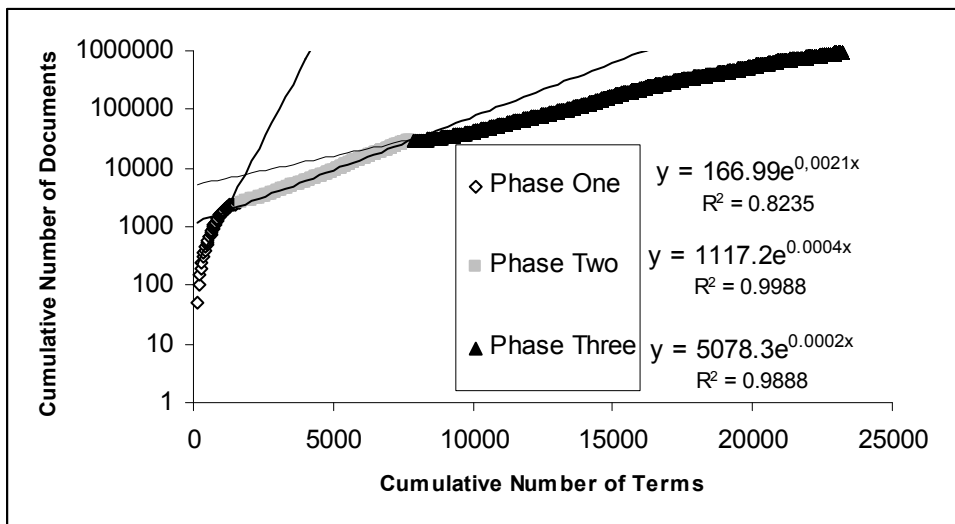


Figure 6: Growth of MEDLINE versus growth of MeSH.

The figure above expresses that the documents of MEDLINE have increased exponentially where the MeSH terms (in Figure 3) have grown logarithmically. The exponents of functions show that the growth speed of documents in each phase has decreased relative to its prior phase.

3.1.2.1 Half-Term-Rate (HTR)

If we apply the equation used for determining the Half-Life and simulate the same case here, we will find Half-Term-Rate (HTR):

$$\text{HTR} = \text{Ln}(2) / \text{exp.} \quad (\text{ix.})$$

Where “Ln” means natural logarithm and “exp.” is the exponent of the functions.

The HTRs of the first to the third phases were determined by the equation “ix”. The results were sequentially “330”, “1,733”, and “3,466”. It means that the number of MEDLINE documents has doubled against the inclusion of every “330”, “1,733” and “3,466” new terms respectively in the first, second, and third phases.

Plotting the values of HTRs against the phase numbers yielded a linear function:

$$\text{HTR} = 1567.8x - 1292.8; R^2 = 0.9963 \quad (\text{x.})$$

Equation “x” allows us to predict the HTR of new phases. Having the HTR, one can determine the exponent of documents distribution in each phase. We need only to change the equation “ix” as follows:

$$\text{Exp} = \text{Ln}(2) / \text{HTR} \quad (\text{xi.})$$

For example, the equation “x” predicts the HTR of the fourth phase will be equal to 4,978 and the equation “xi” shows that the exponent of documents distribution in this phase will be “0.00014”.

Before going to the absolute growth of MeSH, let us see which phenomena were found about the growth of terms:

1. The MeSH has grown through three logarithmic phases.
2. The growth speed of the terms increased 6.01 times in the second phase and 1.81 times in the third.
3. The exponents of logarithmic functions of phases have increased linearly (growth power is equal to 1947.9), so we can predict the exponent of the next phases.
4. The median of terms for every phase increases double logarithmic ($y = 859.97x^{2.851}$, $R^2 = 1$), so we can predict the median of the next phases.
5. There is a linear correlation between the medians of terms in each phase and its last term ($y = 1.1501x + 654.25$), so we can predict the end of a phase and the start of the next phase.
6. The growth of MEDLINE documents versus the growth of MeSH terms was following three exponential functions and their exponents decreased sequentially “0.0021”, “0.0004”, and “0.0002”.
7. Half-Term-Rate (HTR) is increasing linearly ($y = 1567.8x - 1292.8$ and $R^2 = 0.9963$; where x is the number of phase).

3.1.3 Absolute growth of Medical Subject Headings (MeSH)

The result above was derived by observing the growth of MeSH versus the cumulative number of headings. The study of MeSH growth versus the absolute number of unique headings used in the MEDLINE for the first time should yield the three phases mentioned above as well. In Figure 3, every point on the x-axis represent fifty citations, but in the following figure, every point on x-axis covers 2,000 citations. It helps to avoid getting the value zero on its corresponding point on the y-axis since, on occasion; thousands of new documents have not been indexed by even one new distinct heading.

The observation of the cumulative growth of MeSH showed how it has grown, but the study of the absolute growth expresses how the growth rate has fallen over the years.

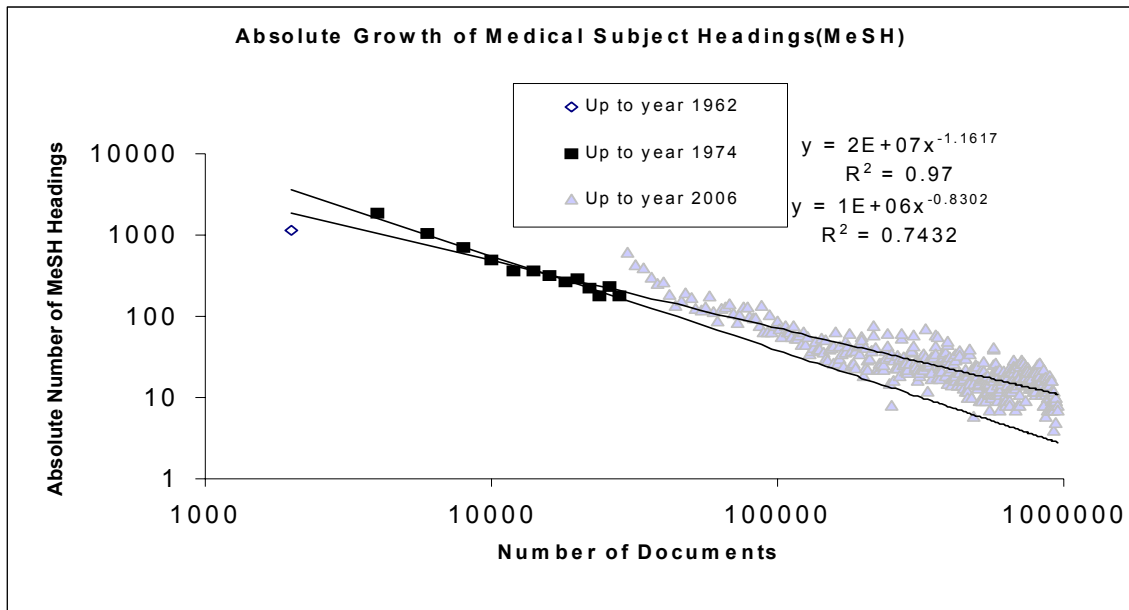


Figure 7: Absolute growth of the Medical Subject Headings (MeSH). The x and y axes are scaled logarithmically.

The Figure above shows, that an increase of the number of indexed documents is followed by a reduced growth of new headings used for indexing. The x and y axes are scaled logarithmically to get straight trend lines.

Let us compare Figure 3 and 7. The figure above expresses almost the same results from other point of view. In the Figure 3, we observed three different phases. We considered that the first 2,600 documents of the sample contained 1,656 new headings. Figure 7 shows that the first 2,000 documents produced first 1,138 new headings, the second phase continued up to 28,000 records and produced 6,197 unique MeSH headings (their total = 1,138 + 6,197 = 7,335). The total is almost close to the results of Figure 3. In this Figure (phase 1 + 2), 28,850 records produced cumulatively 7,853 new headings. The figure above shows that the last phase continued also up to the last record in the sample and produced 15,541 new headings. Thus, the total of three phases in the figure above is the same as the cumulative number of distinct terms in Figure 3.

The first phase is illustrated by one point; because of this it is not possible to derive any function from it. The second and the third phases express that the growth rate of MeSH following a power law function with a very fast decrease from year 1963 to 1974. In the second phase, after a jump, the rate of reduction decreased in the third phase. The function of the second phase (i.e. $y = 2E + 07x^{-1.1617}$) has changed to $y = 1E + 06x^{-0.8302}$ in the third phase.

The R-Squared of the third phase in the figure above is 0.7432. It reveals a greater deviation from the trend line in comparison to the prior phase. The vast deviations at the end of the curve reveal that a huge number of documents should be added to MEDLINE to produce a new distinct term. The following table illustrates this better:

Table 5: A brief summary of the absolute growth of the MeSH. The two first columns represent the growth through the first 36,000 documents, and the third and fourth, the accession of the last headings.

Number of Documents	New headings	MeSH	Number of Documents	New headings	MeSH
1-2,000		1,138	912,001-914,000		7
2,001-4,000		1,864	914,001-916,000		9
4,001-6,000		1,065	916,001-918,000		4
6,001-8,000		705	918,001-920,000		9
8,001-10,000		490	920,001-922,000		8
10,001-12,000		363	922,001-924,000		9
12,001-14,000		358	924,001-926,000		7
14,001-16,000		312	926,001-928,000		11
16,001-18,000		264	928,001-930,000		12
18,001-20,000		286	930,001-932,000		10
20,001-22,000		224	932,001-934,000		9
22,001-24,000		180	934,001-936,000		8
24,001-26,000		230	936,001-938,000		9
26,001-28,000		175	938,001-940,000		5
28,001-30,000		607	940,001-942,000		9
30,001-32,000		422	942,001-944,000		10
32,001-34,000		390	944,001-946,000		8
34,001-36,000		300	946,001-948,000		7

The right part of the table above illustrates the values related to the third phase. It reveals that every 2,000 documents could produce between four and twelve distinct terms. We learned that the HTR of this phase was 3,466. It shows that the number of documents has doubled for producing every 3,466 distinct terms. If we consider the 912,000th document of the sample, the number of documents should reach 1,824,000 for adding other 3,466 distinct terms into

MeSH. Thus, the productivity of every 2,000 MEDLINE documents will show a sloppy curve at the end.

3.1.4 Optimization of accurate thesaurus development

How can the development of MeSH be explained by a single function from the today's point of view? To answer this question, we will try to derive the function from the three logarithmic functions yielded from MeSH development.

First let us determine how many documents produce one more new MeSH term:

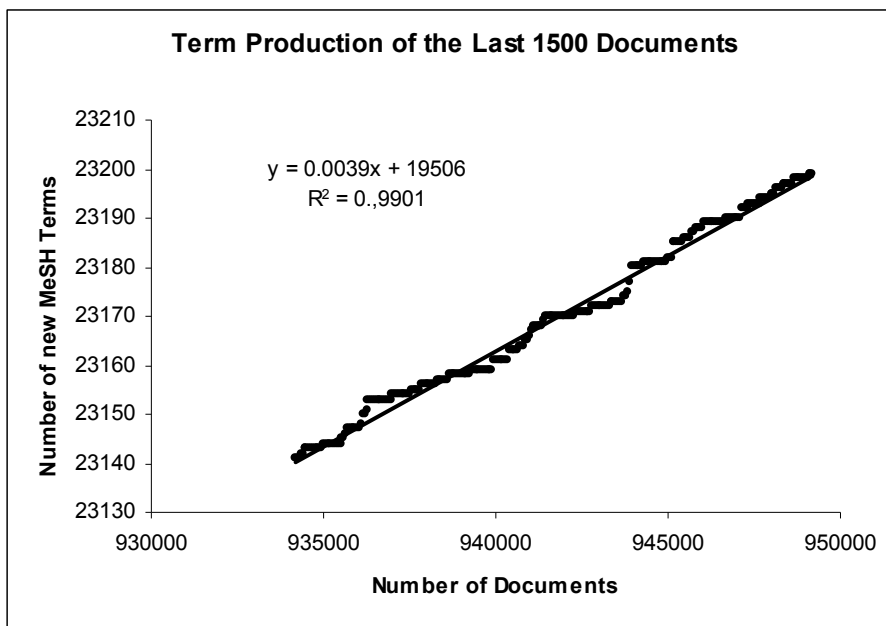


Figure 8: Term production of the last 1,500 documents of current MEDLINE.

The function in figure above illustrates term production of the last 1,500 documents. The equation derived shows a linear correlation.

$$y = 0.0039x + 19,506; R^2 = 0.9901 \quad (\text{xii.})$$

where “y” is equal to the number of thesaurus terms and “x” is the number of indexed documents.

It means, as an average, one new document is producing 0.0039 new index-terms. In other words, inclusion of one more new term to MeSH is the consequence of 256 new documents in MEDLINE. This is the dynamic of nomenclature in medicine with respect to an indexing system like MeSH. If we subtract this linear development, we will get a logarithmic function again:

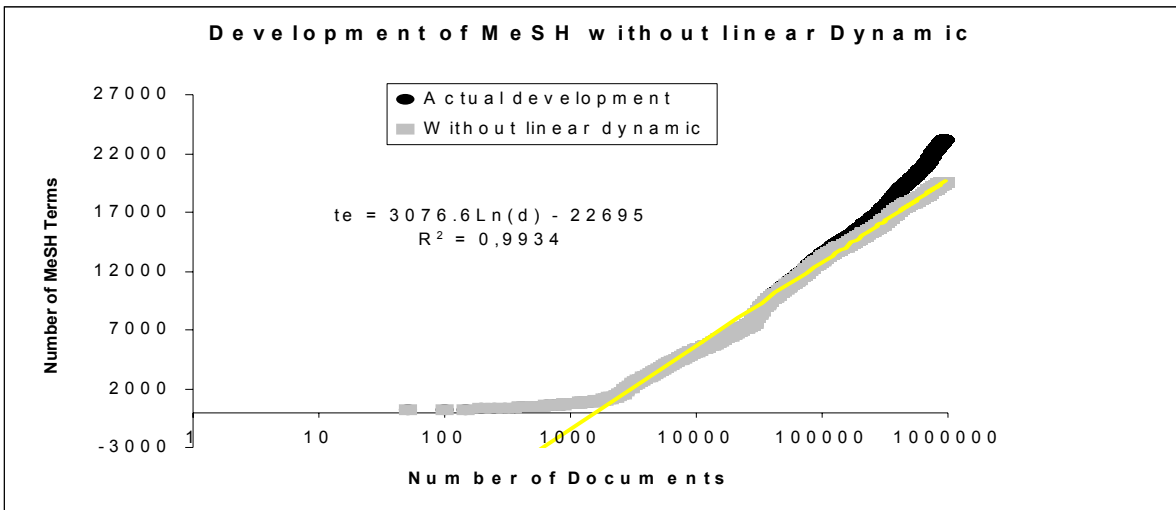


Figure 9: Comparison of MeSH development with and without linear dynamic.

The figure above compares the actual development of MeSH with the optimized one in Figure 8. The function in this figure shows that the values of “ t_e ” become positive if the number of documents is more than 1,600. It indicates the least amount of documents that one needs to create a thesaurus like MeSH.

From today’s point of view, we know that a medical documentation system like MEDLINE needs at least 1,600 different documents to construct a primordial thesaurus like MeSH and it should develop as follows, if we combine the equation of linear dynamic growth of the thesaurus (“Equation xiii”) with its logarithmic growth showed by the Figure 9:

$$t_e = 3,076.6 \ln(d) - 22,695 + 0.0039d \quad (\text{xiii.})$$

where “ t_e ” is the estimated number of thesaurus terms and “ d ” equals the number of documents.

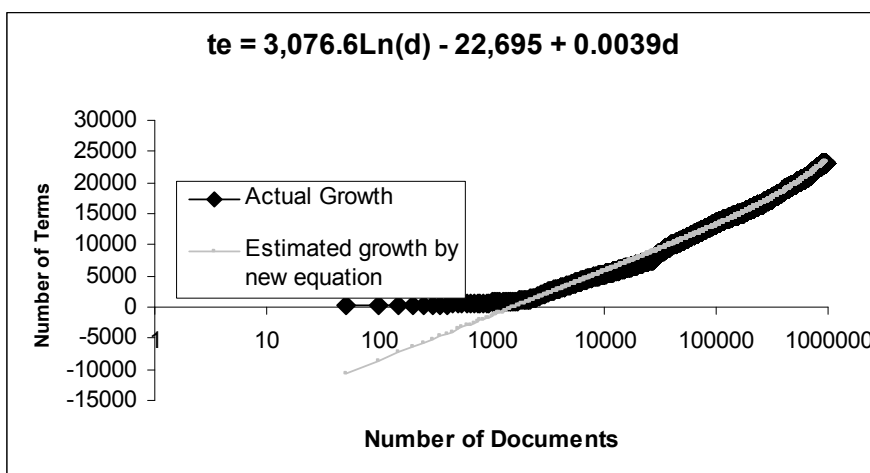


Figure 10: Comparison of development of MeSH based on the equation „ $t_e = 3,076.6 \ln(d) - 22,695 + 0.0039d$ “with the actual ones.

The figure above compares the actual growth of MeSH with the estimated growth by the equation xiii. It reveals that the calculated equation matches the results drawn from the functions of the three phases. From the today's point of view, the development of a thesaurus like MeSH follows the function in Equation xiii.

It is remarkable that a thesaurus is a dynamic system, growing with a first proximity in a linear way with the number of publications. For a medical documentation system like MEDLINE, it needed around forty years to reach the linear growth shown by Equation xii.

3.2 Distribution of MeSH headings in MEDLINE

The use frequency of MeSH headings indexed in MELINE during the intervals “1965 – 1970”, “1965 – 1980”, “1965 – 2000”, and “1965 - 2006” were studied to see how they distributed over the years. The focus was on three facts: The shape of distributions, the types of functions resulting from the distributions, and the parameters of the functions with concentration on their exponents:

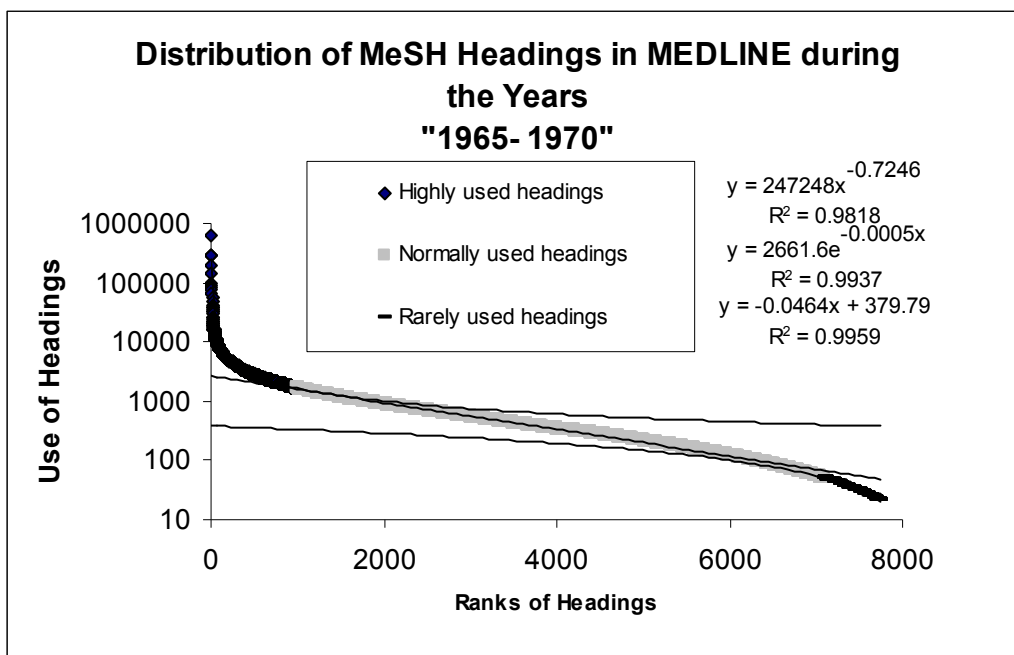


Figure 11: Distribution of MeSH headings in MEDLINE during the years 1965 – 1970. The y-axis is scaled logarithmically.

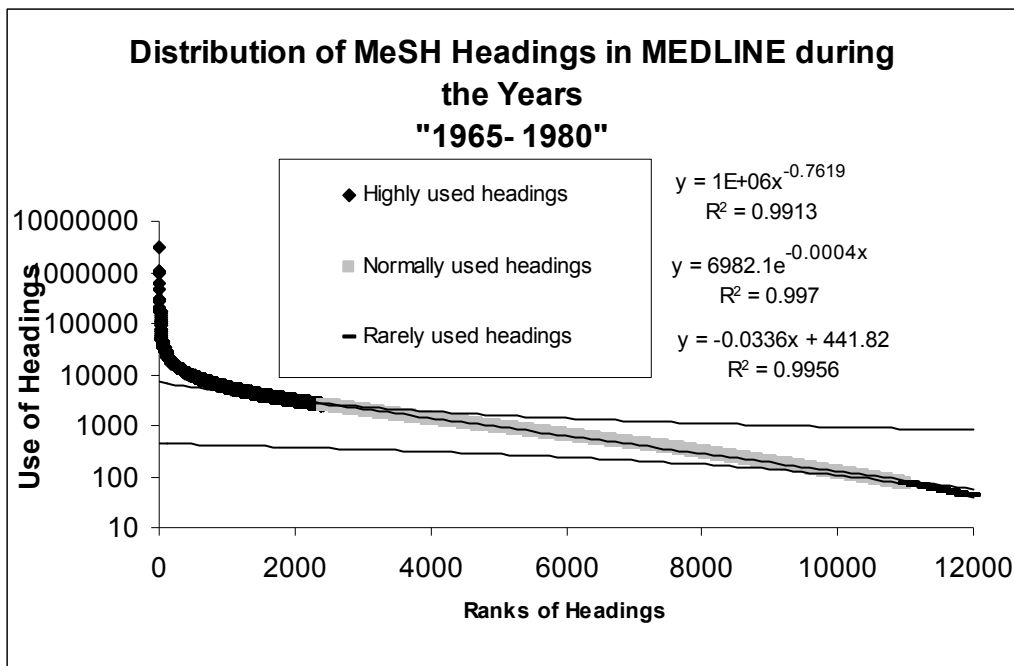


Figure 12: Distribution of MeSH headings in MEDLINE during the years 1965 – 1980. The y-axis is scaled logarithmically.

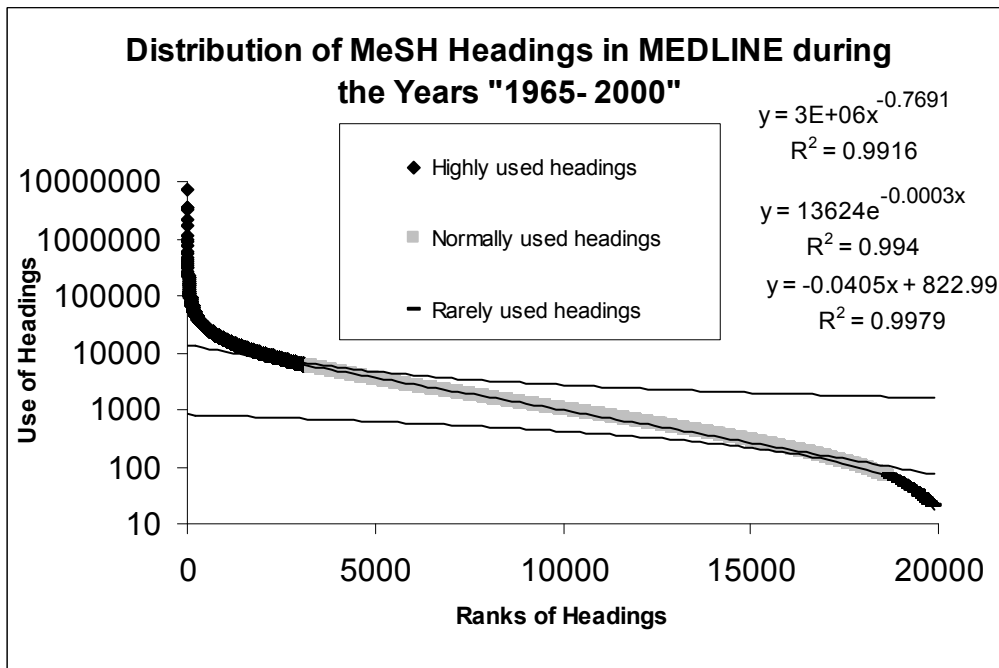


Figure 13: Distribution of MeSH headings in MEDLINE during the years 1965 – 2000. The y-axis is scaled logarithmically.

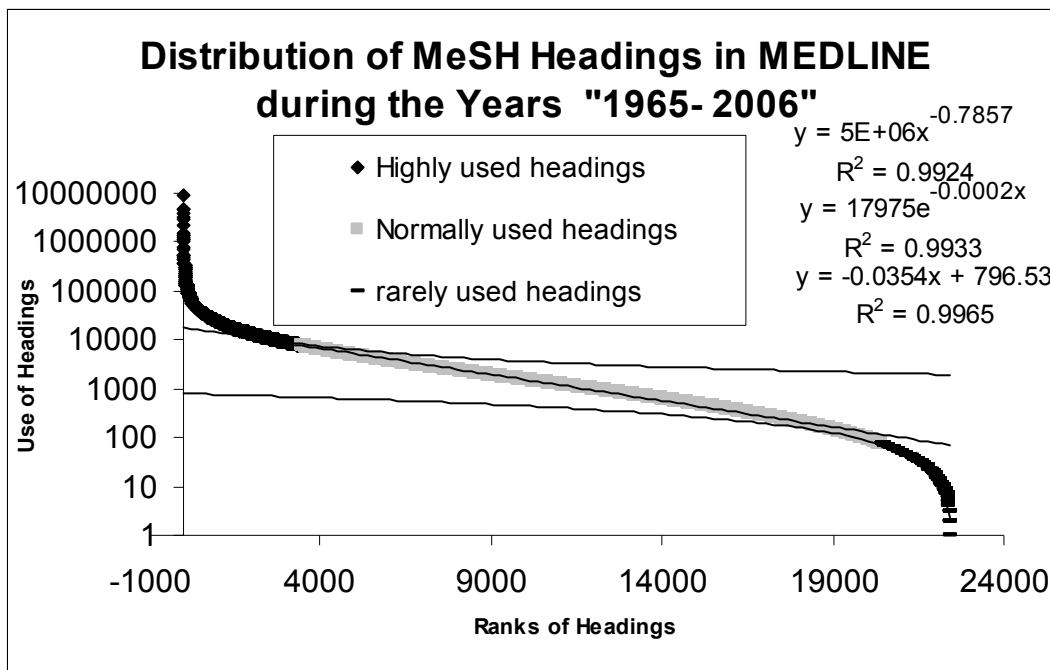


Figure 14: Distribution of MeSH headings in MEDLINE during the years 1965 – 2006. The y-axis is scaled logarithmically.

The four figures above illustrate how the terms in Medical Subject Headings (MeSH) distributed during the intervals “1965 - 1970”, “1965 – 1980”, “1965 – 2000” and “1965 – 2006”. The first fact that can be derived is the existence of three classes of MeSH headings. We shall call them “highly frequented”, “normally frequented” and “rarely frequented”. The distribution of terms in the first class is following a power law function, a function like what Zipf, G. K. (1949) found for the distribution of words in natural texts. The terms in the second class are distributed exponentially and finally, the distribution of rarely frequented terms is following Pearson’s function.

To facilitate understanding, the equations above are represented in Table 6.

Table 6: A brief summary of the distribution of MeSH headings in MEDLINE drawn from the results in Figures 11, 12, 13, and 14.

Highly Used headings	#Headings	1970 997 (~%12.8)	1980 2,403(~%20.0)	2000 3,282(~%16.52)	2006 3,442(~%15.4)
	Function	$y = 247248x^{-0.7246}$	$y = 1E+06x^{-0.7619}$	$y = 3E+06x^{-0.7691}$	$y = 5E+06x^{-0.7857}$
Normally Used headings	#Headings	6,099 (~%78.6)	8,593(~%72.0)	15,370(~%77.4)	16,976(~%75.7)
	Function	$y = 2661.6e^{-0.0005x}$	$y = 6982.1e^{-0.0004x}$	$y = 13624e^{-0.0003x}$	$y = 17975e^{-0.0002x}$
Rarely Used headings	#Headings	668 (~%8.6)	1,001(~%8.3)	1,216(~%6.1)	1,996 (~%8.9)
	Function	$y = -0.0464x + 379.79$	$y = 0.0336x + 441.82$	$y = -0.0405x + 822.99$	$y = -0.0354x + 796.53$

3.2.1 Highly frequented headings

The highly frequented class has covered “%16 ± 4” of headings. As explained above, their distributions follow the power law function. The exponents of the functions changed from – 0.7246 in 1970 to –0.7857 in 2006. These changes were not enough to remodel the shape of the power law distributions heavily, but we can learn some facts from them.

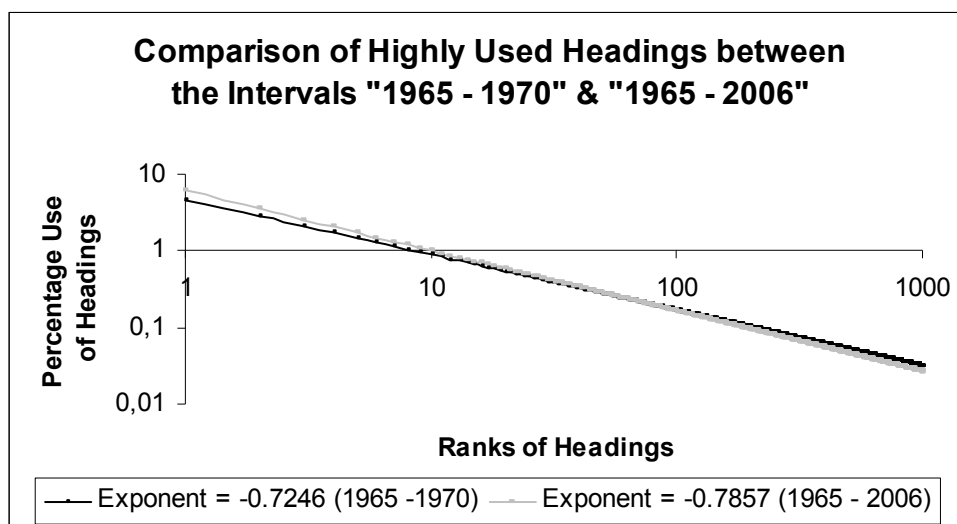


Figure 15: Comparison of highly used headings between the intervals „1965 – 1970“ and „1965 – 2006“.x and y axes are scaled logarithmically.

By changing the exponent from –0.72 to –0.79, the curve will move and the slope becomes sharper. If we consider the x-axis of the plot as the ranks of highly frequented headings and

the y-axis as the percentage use of them, we will find that the difference between the usage of the first and the last terms increases. This means the persistency usage rate of the top terms will never decrease over time. On the other hand, the Gross-Droops express that the difference between the usages of terms is far apart at the outset, but gradually decreases so that the difference between the uses of the terms on the bottom of the class is nearer to each other.

The percentage use of headings shows clearly that the difference between the heading on the top and the one at the bottom of the ranking during the interval “1965 – 2006” is more than that for the interval “1965 – 1970”. In other words, the curve characterised with lighter points, which belongs to the year 1970, is flatter than that of 2006. It starts under the other curve and gradually overtakes it.

The use frequencies of headings in this class varied between 1,654 - 632,032 (in 1970), 2,566 – 3,042,254 (in 1980), 5,768 – 7,136,841 (in 2000) and 7,715 – 9,502,974 (in 2006).

3.2.2 Normally frequented headings

We observed that the headings with normal usage cover “%75 ± 2” of MeSH, meaning that the majority of headings belong to this class of terms. Unlike the previous one, this class has been distributed exponentially. The exponents of the functions have changed from -0.0005 in 1970 to -0.0002 in 2006.

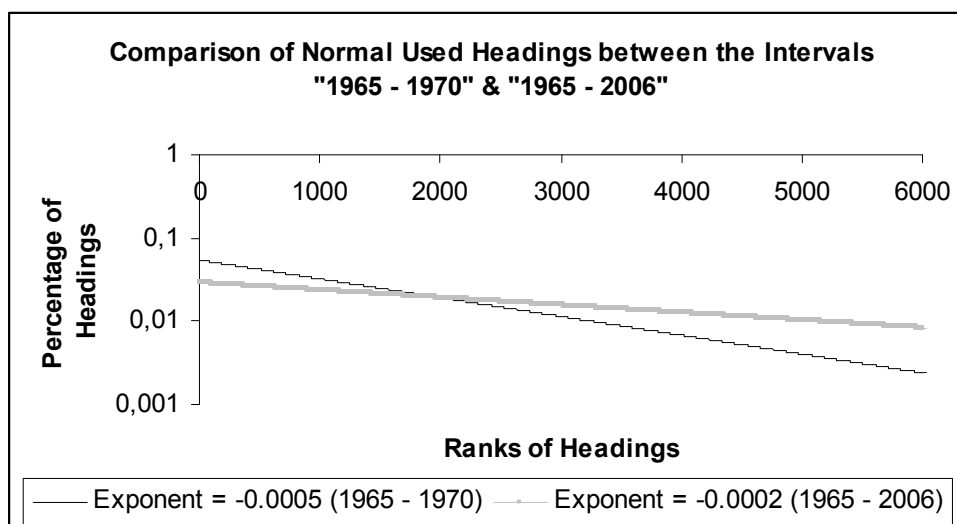


Figure 16: An illustration of how the curve of normally frequented headings is moving to the gentle slope. Y axis is scaled logarithmically.

In the figure above, we observe two curves with different slopes. In other words, increasing the absolute values of exponents moves the curve toward the gentle slope. If we follow the growth speed of values of the y-axis onwards, we will see that the usage rate of the most frequented terms of this class decreases in comparison to those used less. This means that the uses of terms that had a lower frequency in the past will find a higher growth rate in the future. Conversely, the growth rate of those with higher frequency will decrease over time.

The use of headings of this class varied between 51 – 1,653 (in 1970), 74 – 2,566 (in 1980), 68 – 5,766 (in 2000) and 76 – 7,715 (in 2006) (see Table 6).

3.2.2.1 *Half-Rank-Usage (HRU) of normally frequented headings*

The exponential functions have other features that help to study the distribution of use frequency from another point of view. If we consider the use frequency of the heading on the top of the normally frequented class, we can determine the rank of the headings that was used half as many times and also that of the succeeding heading that was used half as much as its predecessor, and so on. Let present this fact with the term Half-Rank-Usage (HRU). It can be calculated as:

$$\mathbf{HRU = Ln (2) / exp} \quad \text{(xiv.)}$$

where the “Ln” is the natural logarithm and “exp” is the value of the exponent derived from the exponential distribution of the headings in MEDLINE. In the current work, the exponents were 0.0005 in 1970 and reached to 0.0002 in 2006.

The value of the HRU for MeSH headings has increased over the years. It was 1,386 in 1970, rose to 1,733 in 1980, kept increasing to 2,310 in 2000 and reached 3,466 in year 2006. To take an example, the explanation of the HRU for the headings in 1970 can clarify its meaning. Based on the HRU, we can assume that the 1,386th heading was used half as many times less than the first heading which is in the top rank of the normally frequented class. And the 2,772nd heading (i.e. 1,386 * 2 = 2,772) was used half as many times as the 1,386th ranked heading and the 4,158th heading (i.e. 1,386 * 3 = 4,158) and also half as many times less than the 2,772nd ranked heading and so on. The values of the following table are based on the above calculations:

Table 7: Ranks of terms compared to the number of their usage based on HRU in four different years. HRU calculated by $HRU = \text{Ln}(2) / \exp$.

HRU of 1970 =1,386		HRU of 1980 =1,733		HRU of 2000 =2,310		HRU of 2006 =3,466	
Rank of Term	Term Usage	Rank of Term	Term Usage	Rank of Term	Term Usage	Rank of Term	Term Usage
1	1,653	1	2,566	1	5,766	1	7,715
1,386	827	1,733	1,283	2,310	2,883	3,466	3,858
2,772	413	3,466	642	4,620	1,442	6,932	1,929
4,158	207	5,199	321	6,930	721	10,398	964
5,544	103	6,932	160	9,240	360	13,864	482
6,930	52	8,665	80	11,550	180	17,330	241

The total number of normally frequented terms in the years 1970, 1980, 2000, and 2006 were sequentially “6,098”, “8,593”, “15,370”, and “16,979”. If we compare these with the last values of the “Rank of Term” columns in the table above, we will see that they are close to each other. The greatest difference belongs to the year 2000 (i.e. 11,550 in the third part of the table above vs. 15,370 which equals the total number of normally frequented terms in 2000). The varieties are mostly because of the sensitivity of HRU against the exactness of the exponents. This means the precise values of exponents should be calculated even up to five or six digits after the decimal point (0.), but the statistical program (Microsoft Excel) could not determine them exactly. If we disregard these deviations and take the values of the above table into consideration, we can accept:

$$\mathbf{HRU = \text{Ln}(2) / \exp \cong (1/5) \times \text{total number of terms} \quad (\text{xv.})}$$

If we halve the used number of the first-ranked term five times, we will yield the used number of the last term in the normally frequented class. Based on HRU, this means the normally used terms, regardless of their total, are always divided into six groups. Due of this fact, the values of the table above are placed on six rows regardless of the total amount of the terms in each of the four cases.

Suppose we have ten terms and the first one has been used one hundred times. The HRU should be two ($HRU = 10/5 = 2$) and the second up to the fifth halves should be sequentially “2 (50)”, “4 (25)”, “6 (12.5)”, “8 (6.25)”, and “10 (3.125)”. In this example, the numbers preceding the parentheses are ranks and those within parentheses are the predicted used numbers of terms.

Let us calculate the HRU through the new equation and then represent the yielded values on a new table:

$$\text{HRU} = \text{total number of terms} / 5$$

(xvi.)

Table 8: Ranks of terms versus the numbers of their usage based on HRU in four different years. HRU calculated by $\text{HRU} = \text{Total number of terms} / 5$.

HRU of 1970 =1,220		HRU of 1980 =1,719		HRU of 2000 =3,074		HRU of 2006 =3,396	
Rank of Term	Term Usage	Rank of Term	Term Usage	Rank of Term	Term Usage	Rank of Term	Term Usage
1	1,653	1	2,566	1	5,766	1	7,715
1,220	827	1,719	1,283	3,074	2,883	3,396	3,858
2,439	413	3,437	642	6,148	1,442	6,792	1,929
3,659	207	5,156	321	9,222	721	10,187	964
4,878	103	6,874	160	12,296	360	13,583	482
6,098	52	8,593	80	15,370	180	16,979	241

We see the new equation has the same function as the other one derived from the exponents. It is an easy way to determine the HRU with more precision.

What the table above reveals is the constant relationship between the richest and poorest terms for all cases. It means: $1,653 / 52 \cong 2,566 / 80 \cong 5,766 / 180 \cong 7,751 / 241 \cong 31$. In addition, such a relationship exists between others groups.

3.2.3 Rarely frequented headings

The last class belongs to the rarely frequented headings. They cover “%7.5 ± 1.5” of MeSH headings. Their use frequency shows that their distributions follow Pearson’s function. The powers of the functions have been almost -0.04 during the different intervals. The rarely frequented terms can be counted as the temporary class, since they shift gradually into the two other categories.

The rarely frequented terms in years 1970 and 1980 were compared with those of 2006. The results were almost unexpected. It was thought that a few terms should share that list. However, not one of them belonged to the rarely frequented terms in 2006. The comparison of the same class of terms between the years 2000 and 2006 revealed that 490 terms from 1,261 rarely frequented class in 2000 still appeared in the list of the rarely frequented for the year 2006.

The use frequencies of headings in this class varied between 21-51 (in 1970), 41 – 73 (in 1980), 20 - 68 (in 2000) and 1 - 76 (in 2006). The shift of headings from two other classes down to this class should not be possible over time.

3.3 Factors related to the number of index-terms of articles

In the current investigation, the factors that affect the number of index terms assigned to the documents by the human indexers are divided into three different groups: 1. content-related; 2. presentation-related; and 3. policy-related factors.

The length of indexed documents and the types of documents' sources (i.e. peer reviewed) are two content-related factors. Some factors like abstract of documents, language in which a document is written, and structure of documents are related to the presentation of the documents. Finally, the year of indexing, the Impact Factor, and the priorities of journals for in-depth indexing can be counted as policy-related factors.

From the factors mentioned above, the following ones will be taken into consideration:

1. Article length.
2. Presence of abstract within article.
3. Language of article.
4. Entrez date (inclusion dates of article into MEDLINE).
5. Priority of journals for in-depth indexing.
6. Journal Impact Factor.

The concentration will be on those articles written in English and indexed in MEDLINE. The exception is journal articles written in German only in the section where the role of language will be determined on the average number of MeSH headings assigned to documents.

3.3.1 Article length

The length of a text can be measured by the means of two scales:

1. Amount of the text tokens and types (number of words).
2. Number of pages.

For determining the relationship between the article's length and the number of MeSH headings, both of the two scales will be taken into consideration. As the tokens and types are a better indicator of the text content semantically, they will be considered first.

3.3.1.1 Tokens and types

In the field of automatic indexing and thesauri development, the tokens and types of texts have an important role. As these activities related mostly to the semantics and the pragmatics of the texts, attention is on which of the types can convey the contents of the texts. The importance of the words or terms in them is determined by using different techniques. One of them is the frequency of the types in the text.

This section will first find the relationship between the types and the tokens. The relationship between the number of pages of articles and their tokens will be determined, followed by a concentration on the relationship between the average number of MeSH headings assigned to the articles and the tokens.

The contents of the current section will be as follows:

1. Relationship between the text tokens and types.
2. Relationship between the number of articles pages and tokens.
3. Relationship between the tokens of the articles and the average of MeSH headings assigned to them.

3.3.1.1.1 Relationship between the text tokens and types

In this section we will observe the relationship between the text tokens and types:

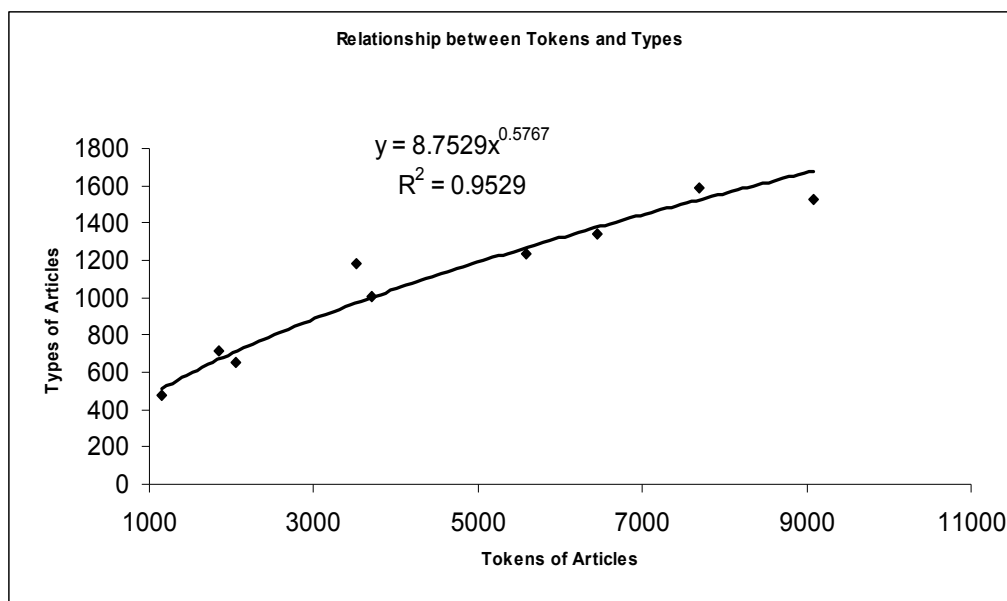


Figure 17: Relationship between the tokens and the types of articles.

The figure above illustrates a power law relationship between the tokens of an article and the types that the article has. This means that the variety of words in shorter articles is larger than in longer ones. Where the variety of the words is more, the occurrence of the content-bearing words is possibly higher. Longer texts contain more types, but when compared to total tokens of texts their variety is less.

3.3.1.1.2 Relationship between the number of pages of articles and tokens

The assumption is that articles consisting of more pages should bear more tokens. This will be tested as follows:

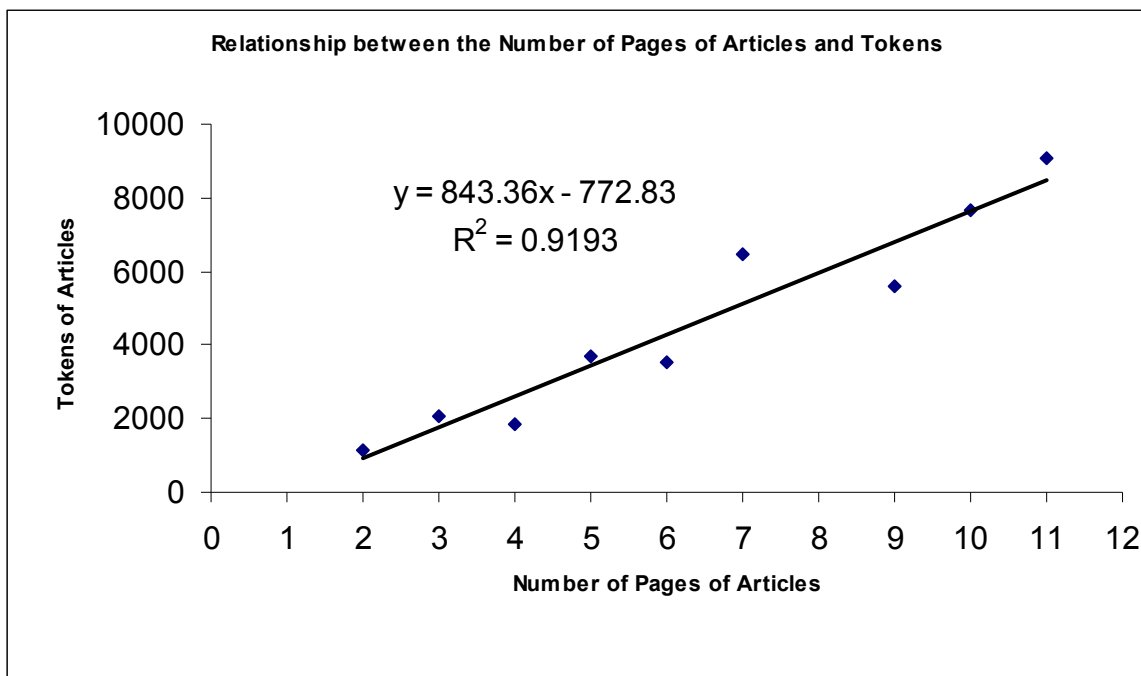


Figure 18 Relationship between the number of pages of articles and their tokens.

The figure above resulted from determining the relationship between the number of article pages and its tokens. It is clear that the higher the number of pages of articles, the bigger the volume of tokens. As shown in this figure, the relationship between them is linear with a growth power of ~843 tokens. By adding one more page to a typical article in the biomedical field, nearly 843 tokens are added to the text.

3.3.1.1.3 Tokens of articles and average of MeSH headings assigned to them

In Figure 18, we observed how the number of pages of articles and their tokens are related to each other. If the tokens influence the number of MeSH headings assigned to the articles, the number of pages of articles should have the same effect on the number of assigned headings.

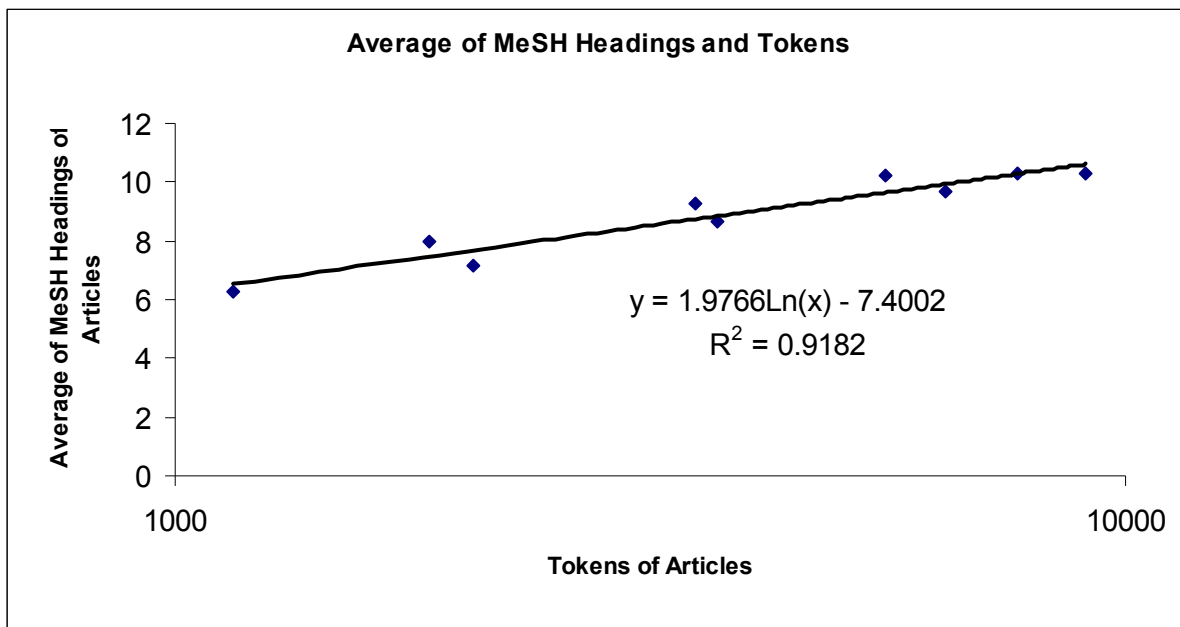


Figure 19: Relationship between the lengths of articles scaled by tokens and their average MeSH headings. The x-axis scaled logarithmically.

As expected, the relationship between the text tokens and the average of MeSH headings assigned to the articles yields a logarithmic function. The number of headings that should be added into an article is nearly two times greater than the natural logarithm of the tokens in the text minus 7.4.

3.3.1.2 *Number of pages and existence of abstracts*

The number of pages and presence of abstracts are two different factors that impact on the average number of MeSH headings assigned to the journal articles. Nevertheless, they will be studied together since the impact of the abstract can be determined only through comparison of articles with and without abstracts of the same length.

Abstracts can be also structured or unstructured. Their form can serve as a determining factor as well. Because of this, in addition to the comparison of articles with and without abstracts, the impact of the abstracts' form will be studied in the sections following the current section.

3.3.1.2.1 Articles with and without abstracts

The following figure will show the impact of the length of journal articles with and without abstracts and allows for the possibility of the comparison between them as well.

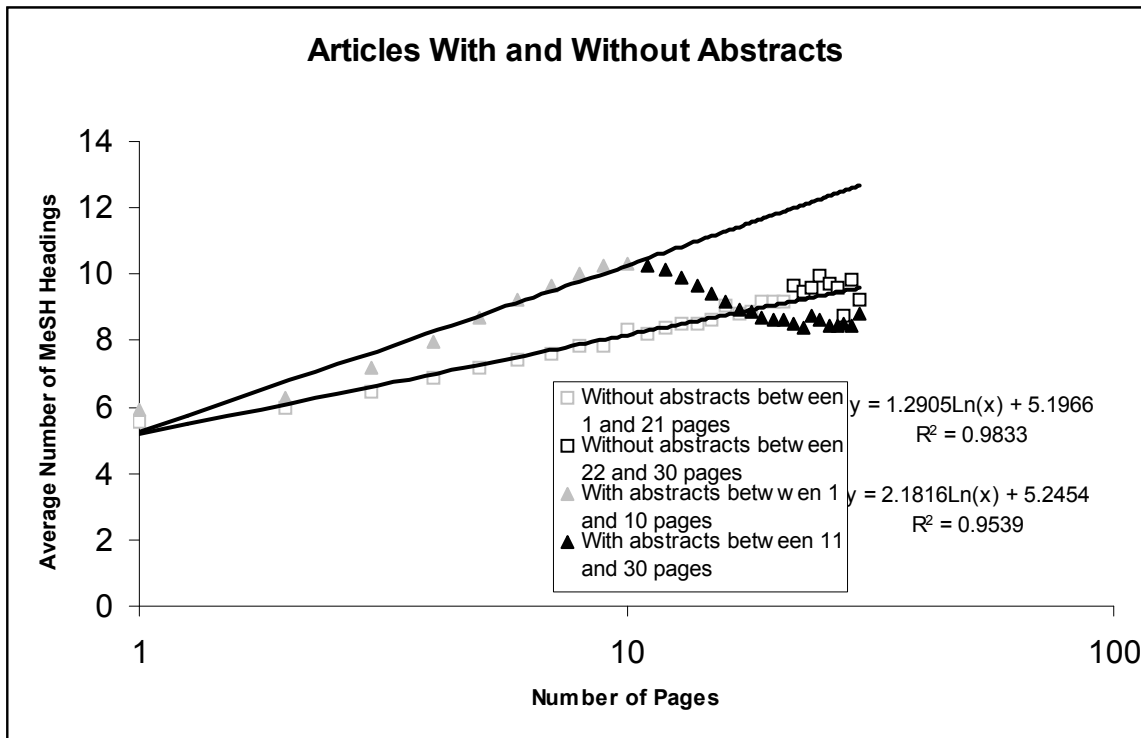


Figure 20: Average number of MeSH headings assigned to articles with and without abstracts. The x-axis is scaled logarithmically.

The figure shows that the articles without abstracts between one and twenty one pages will receive, on average, more MeSH headings logarithmically ($y = 1.2905 \ln(x) + 5.1966$), when the number of pages of articles increases. The last points after the twenty-one make the curve sloppy because of the scarcity of longer articles in the sample.

Articles with abstracts between one and ten pages were also assigned more MeSH headings logarithmically when the number of articles pages increased. The resulting exponent from the function is almost equal to the one yielded from plotting the tokens and average number of MeSH headings (Figure 19). The current exponent is “ $2.1816 \ln(x)$ ” and the former was “ $1.9766 \ln(x)$ ”. In the former, we saw that the average headings is related to the tokens of the text. We see here again that the number of headings of a typical article is almost two times greater than the natural logarithm of the number of pages plus 5.2. The small difference between the two exponents is because of the existence of only nine articles in that sample (Figure 19).

The impact of the abstracts decreases from the eleventh point and reaches the level of the articles without abstracts when the number of pages of articles reaches seventeen. This reduction continues from that point on and makes the curve sloppy. To study why it happened, the average of index terms assigned to both of the two groups of documents are compared point by point by T-Test. At $\alpha = 0.05$, the difference between the average terms of abstracted and non-abstracted articles for those consisting of seventeen and more pages were not statistically significant.

3.3.1.2.1.1 Average of MeSH headings per page

We will now examine the average of MeSH headings assigned to the journal articles per page.

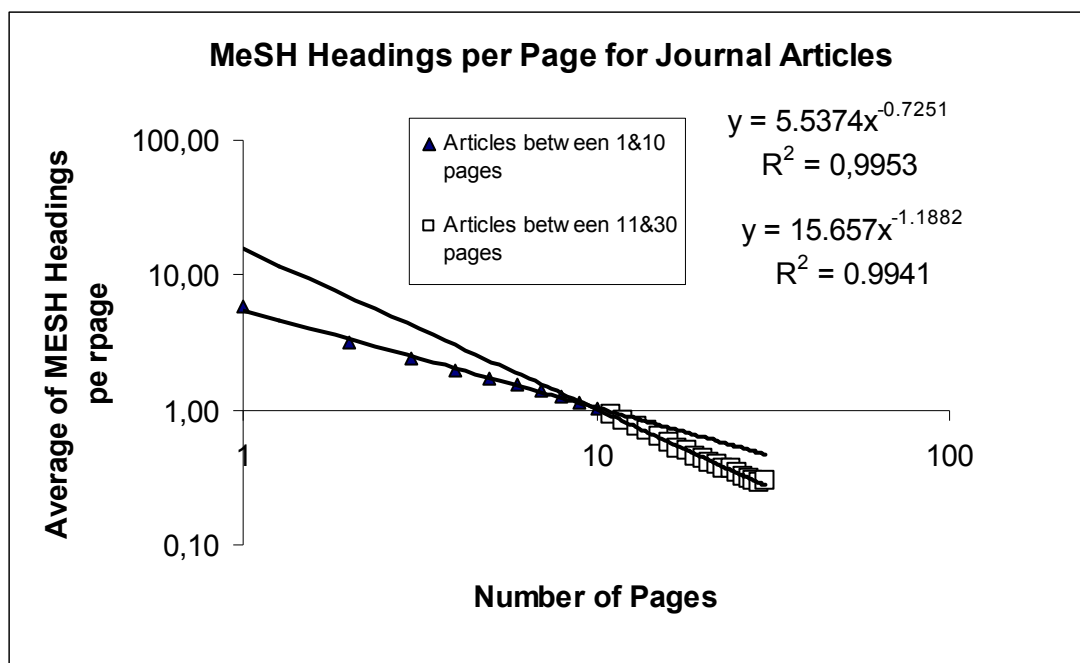


Figure 21: Average of MeSH headings assigned to the journal articles in MEDLINE. The x and y axes are scaled logarithmically.

In the figure above, the average number of headings per page has been shown regardless of whether they are with or without abstracts. It illustrates that the larger articles get less index terms per page. The relationship between the articles' lengths and the headings per page follows two power law functions. The average for the articles between one and ten pages decreases with the function " $y = 5.5374x^{-0.7082}$ " and the reduction continues faster with the function " $y = 15.651x^{-1.124}$ " for the articles between eleven and thirty pages. The results indicate that the articles with only one page yield 5.90 headings on average and when the

lengths of the articles are ten pages, this amount reaches 1.03. Larger articles between eleven and thirty pages have an average of 0.93 to 0.29 headings per page.

The sudden shift of the exponents from -0.7082 to -1.124 indicates that articles larger than ten pages are lengthy due to other reasons, not because they have more content.

3.3.1.2.2 Structured and unstructured abstracts

The form of abstracts in MEDLINE has been discussed widely under two categories: unstructured and structured. The difference between these two forms of abstracts refers to the form of introducing the summarized contents of documents. The structured abstracts do this more precisely than the unstructured, because their format in a way offers the important contents of documents by scanning their main objects so that each scanned part is preceded with an uppercase word, such as AIM, OBJECTIVE and etc.

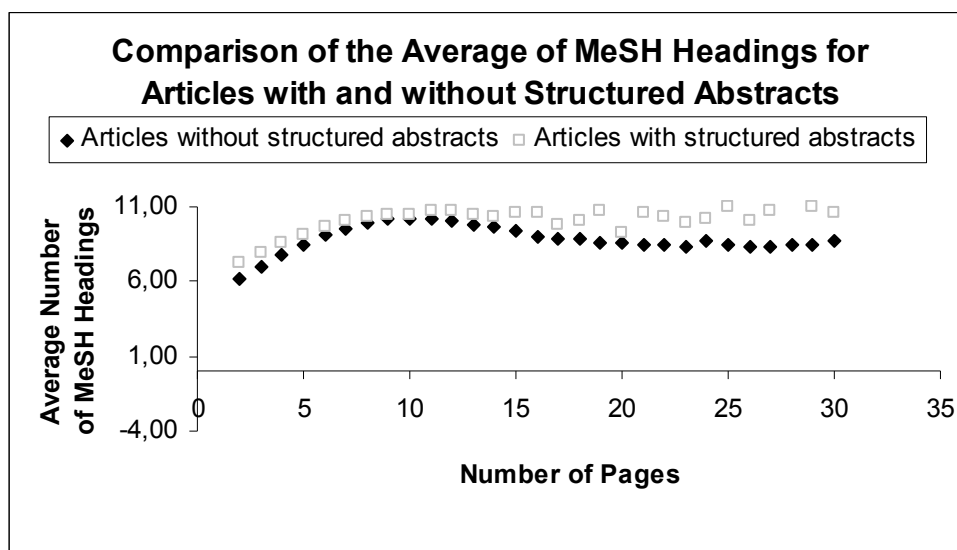


Figure 22: Comparison of the average of MeSH headings assigned to the journal articles with and without structured abstracts.

We saw how abstracts as an additional variable can increase the average number of MeSH headings assigned to articles. Harbourt, A. M.; Knecht, L. S. and Humphreys, B. L. (1995) showed that medical documents with structured abstracts received three more index terms in comparison to others. It showed that the form of abstract can be counted as an additional variable related to the presentation of the text contents.

The figure above shows the role of the abstracts' form on the average number of MeSH headings assigned to the journal articles. The structured abstracts have more headings in 68

comparison to the unstructured ones. The current study revealed that structured abstracts are assigned on average 1.29 more headings. The average of structured abstracts is 10.11 headings, whereas for unstructured ones it is 8.83 (without check tags).

To clarify whether the role of the abstract form is also statistically meaningful, every point of structured abstracts was compared with the corresponding point of unstructured abstracts by T-test. For example, the average number of MeSH headings of articles consisting of four pages and having structured abstracts was compared with unstructured abstracts consisting of four pages. The test results at $\alpha=0.05$ showed that the abstract form has a significant impact on the number of MeSH heading assigned to the article only if the article length is under twenty-one pages. Despite the great differences from point twenty-two and more, the T-tests showed that the impact was not statistically significant.

3.3.2 Language of articles

We learned how the lengths of articles and their abstracts can have an effect on the number of index terms that were assigned to them. The focus was only on journal articles written in English. In the following, the average of MeSH headings assigned to German articles will be compared with the English ones to find the effect of language on the indexing.

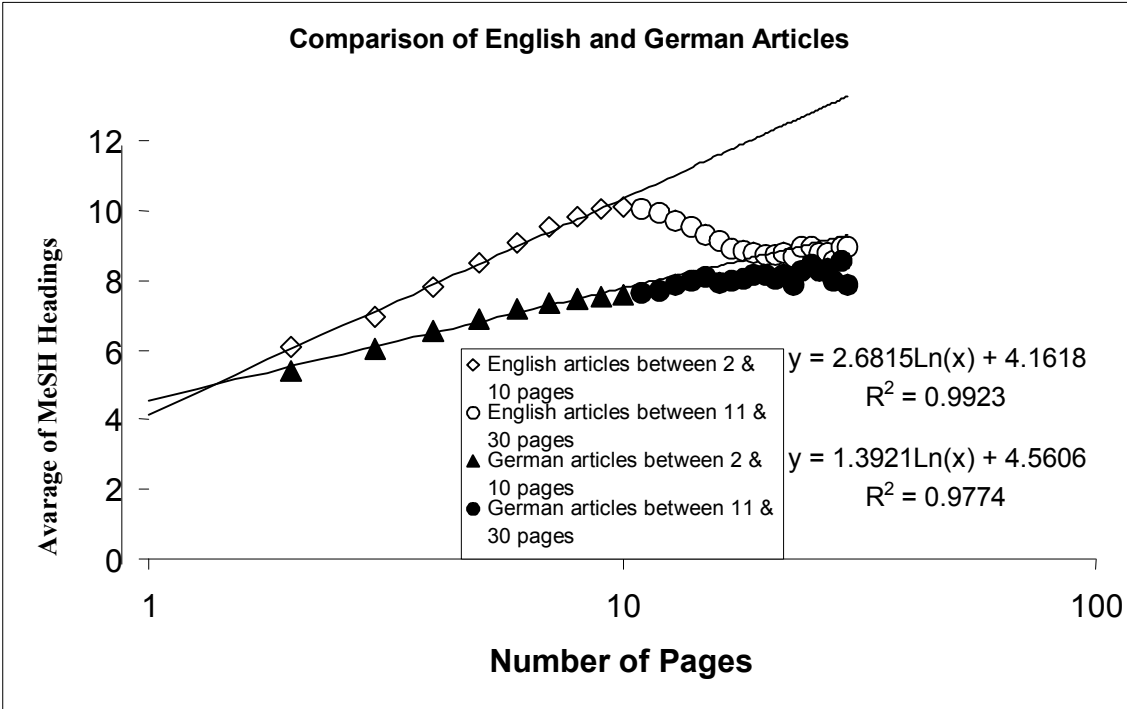


Figure 23: Comparison of the average number of MeSH headings assigned to the English and German journal articles in MEDLINE. The x-axis is scaled logarithmically.

A comparison between the average number of headings assigned to the articles written in English and German shows that language does not change the type of function. The average of MeSH headings of German articles also increases logarithmically for articles one to ten pages in length. The main difference is between their exponents. They show that the average number of headings assigned to English articles is 2.68 times more than the natural logarithm of the number of pages plus 4.2, whereas it is only 1.4 times more for those written in German. The cutting points are almost the same.

The results show that the effect of the language reaches the highest level when the articles are approximately ten pages. The difference in this point is about 2.5 headings (i.e. 33% more findability) and decreases to 0.34 for articles with twenty nine pages.

The effect of language can also be studied from the “headings per page” point of view. We learned that longer texts cause the reduction of the assigned headings per page following two power law functions.

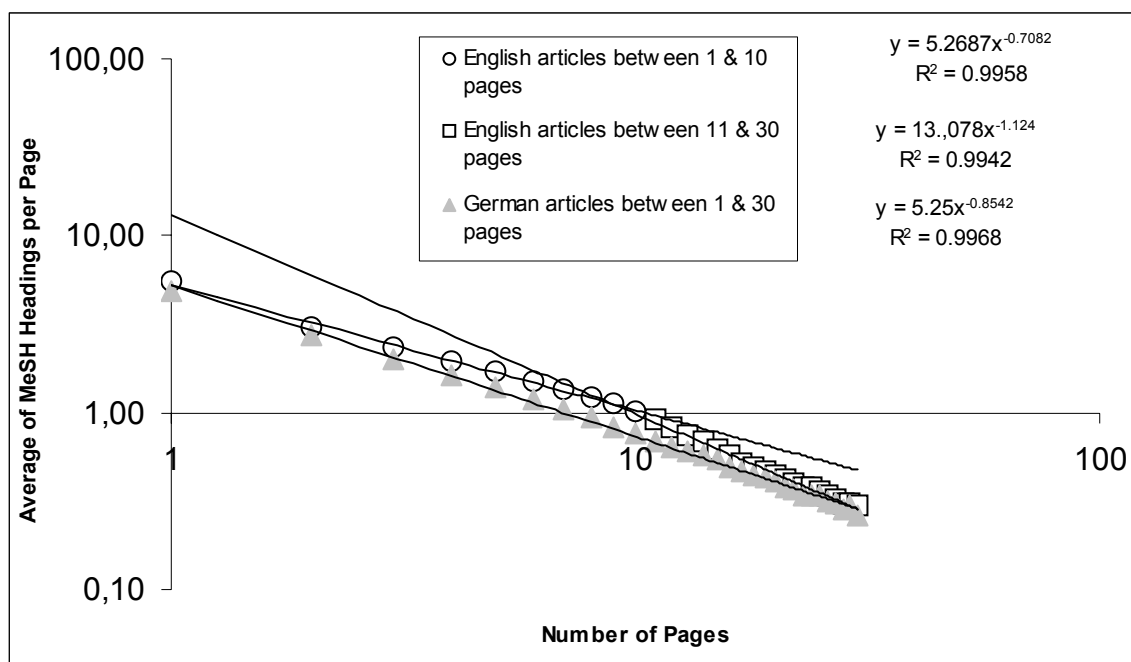


Figure 24: Comparison of the average of MeSH headings assigned to the English and German journal articles per page in MEDLINE. Both of the x and y axes are scaled logarithmically.

The figure above illustrates that assigned MeSH headings per page to German articles also diminishes, following a power law function ($y = 5.25x^{-0.8542}$) on average, when the articles become longer. As the cutting point value of the first function of the English articles is nearly equal to the German ones (respectively 5.27 and 5.25), those with one page get the same amount of headings on average. From point ten, there is a sudden reduction of assigning

headings per page to the English articles, but this does not happen so for the German ones. Thus, we observe that from this point on the curves related to the two languages get closer to each other and the difference of the average number of headings per page decreases to 0.03 in the twentieth point.

The average of headings assigned to German articles is 7.6 and to the English ones is 8.82 (without check tags). It should be noted again that the check tags are not taken into account.

It should be useful to see the effect of the abstracts on the average number of headings assigned to the German articles. The following figure compares the German articles with and without abstracts:

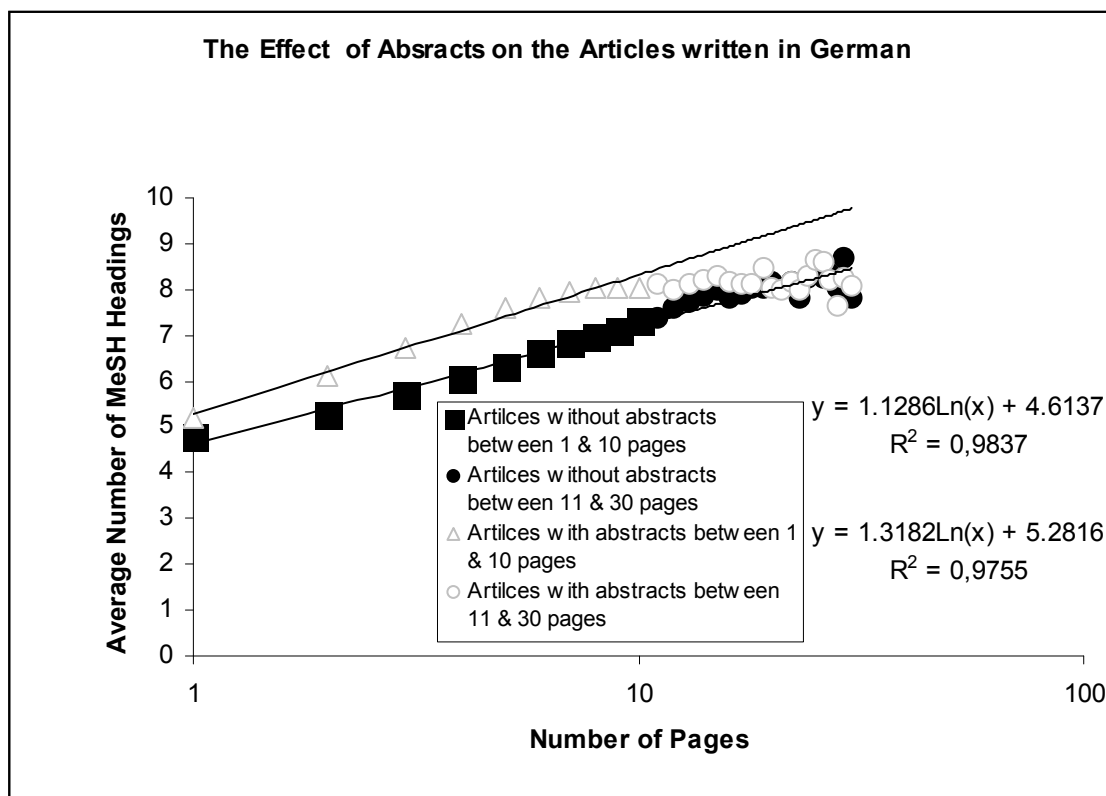


Figure 25: Comparison of the average number of MeSH headings assigned to the German journal articles with and without abstracts in MEDLINE. The x-axis is scaled logarithmically.

The figure above expresses that the difference between the averages of headings assigned to the German articles between one and ten pages is nearly the same on all points. The parallel trend lines on the figure and the almost same exponents illustrate this fact. The difference between the cutting points depict that the articles with abstracts that have equal and less than ten pages received 0.7 more headings on average. From the tenth point, the amount of headings of articles with abstracts decreases and becomes closer to those without abstracts and from the point eighteen they reach the same level.

3.3.3 Date of indexing

The year of indexing documents can be counted as a factor that affects the average number of headings of documents indexed in MEDLINE. This factor is mostly related to the policy of indexing.

3.3.3.1 Average lengths of articles over the years

As the length of articles in this study is considered as a factor in determining the effect of other factors, it is very important to see how the length of articles has changed over time.

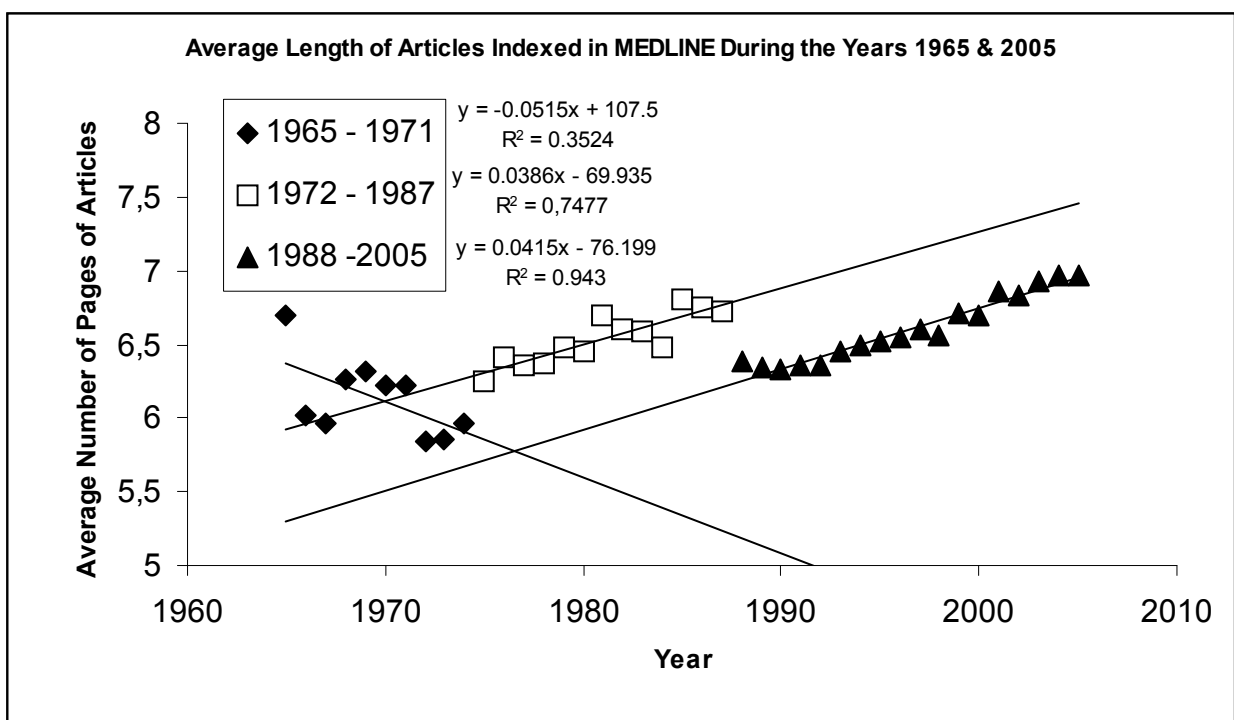


Figure 26: Length of journal articles indexed in MEDLINE during the years 1965 – 2005.

The length of articles decreased from 6.7 pages in the year 1965 down to 6 in 1974. The following year the length increased to 6.2 pages. Growth continued up to the year 1987 and the average of length became 6.7 pages. In the year 1988 it decreased to 6.4 and increased continuously again and reached 7 pages in the year 2005. The functions of the above figure illustrate that the growth rate was -0.052 between the years 1965 and 1971 and changed to a positive rate (0.04) during the years 1972 to 1987. After the reduction in the year 1988, it became 0.042.

As we saw above, the changes were not considerable. To see if small differences in the average of article length are statistically significant, the average lengths of every year were compared with the averages of the years following it. The comparisons were done through T-tests. They showed that in most cases the differences were not significant at $\alpha=0.05$. This is very important for the following findings, because the factor of article length is about the same for all of the years studied.

3.3.3.2 Average of MeSH headings of articles over the years

Depth of document indexing is related to the indexing policy. The average of headings assigned to the articles in MEDLINE during the different years makes it possible to see when and how the policy has changed.

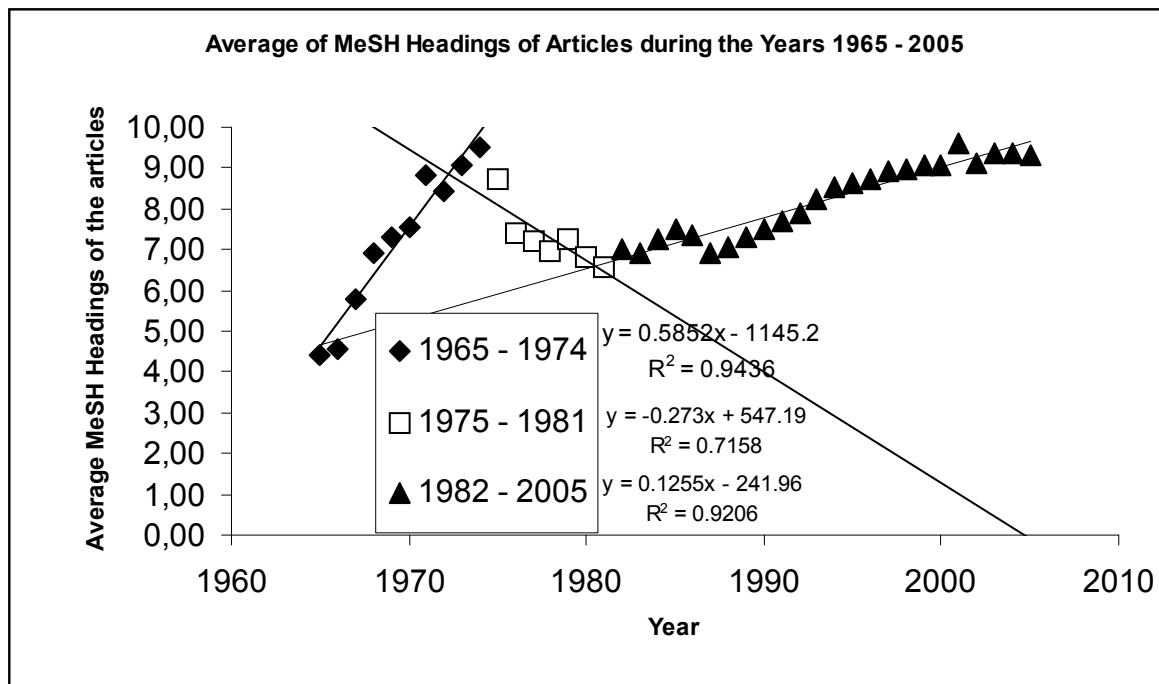


Figure 27: Average of MeSH headings of the journal articles indexed in MEDLINE during the years 1965 – 2005.

The first nine years of the observed average headings show that the assigning of more headings to articles in MEDLINE was considerable. The average grew persistently 0.6 headings every year from 1965 and changed from 4.43 to 9.51 in the year 1974. It decreased during the years “1975 – 1981” with a rate of 0.3 per year and came down to 6.7. During the last twenty-four years, it grew 0.13 headings per year and reached to 9.34 in year 2005.

3.3.3.2.1 Role of Abstracts over the years

We saw that articles with abstracts get more MeSH headings up to a point. In the following we will see how the abstracts of articles could help indexers to achieve greater depth in indexing documents.

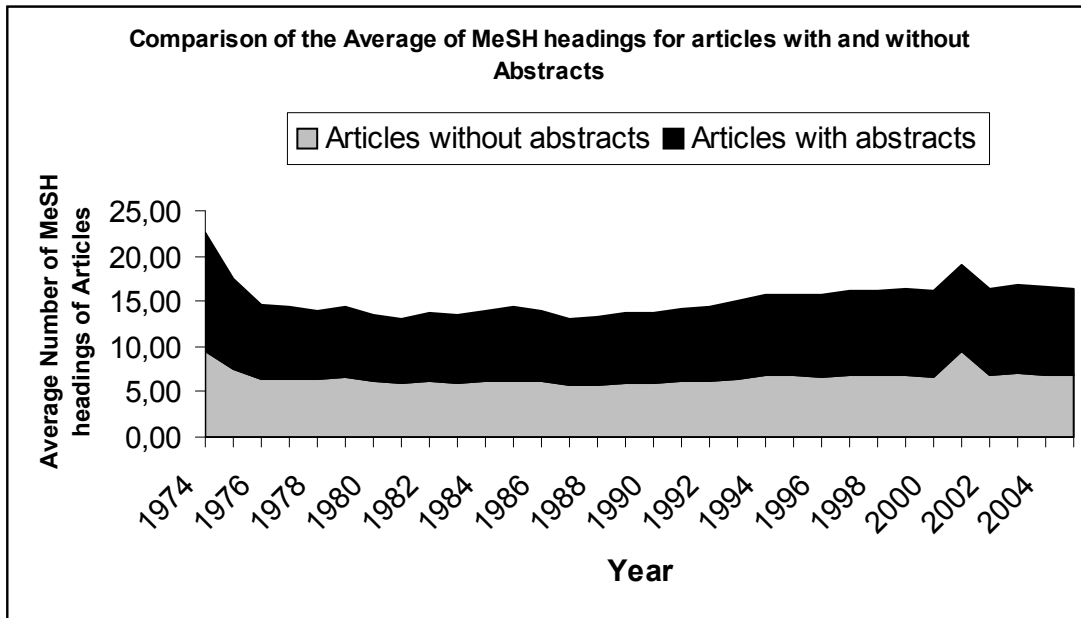


Figure 28: Comparison of the average number of MeSH headings assigned to journal articles with and without abstracts in MEDLINE during the years 1974 – 2005.

Over the course of years, articles with abstracts have gotten more headings in comparison to those without abstracts. The average numbers of their headings are respectively 8.6 and 6.7 (without check tags). The figure above illustrates, when the average number of terms of articles with abstracts has increased or decreased, the average number of terms of articles without abstracts has also followed these increases or decreases in the same time. It means, whenever NLM decided to increase or decrease in-depth indexing of articles, this policy has expanded to all articles without any exception.

3.3.3.2.2 Role of structured abstracts over the years

The introducing of structured abstracts goes back to the year 1988. The following figure will show how they could affect the in-depth indexing of articles over the years.

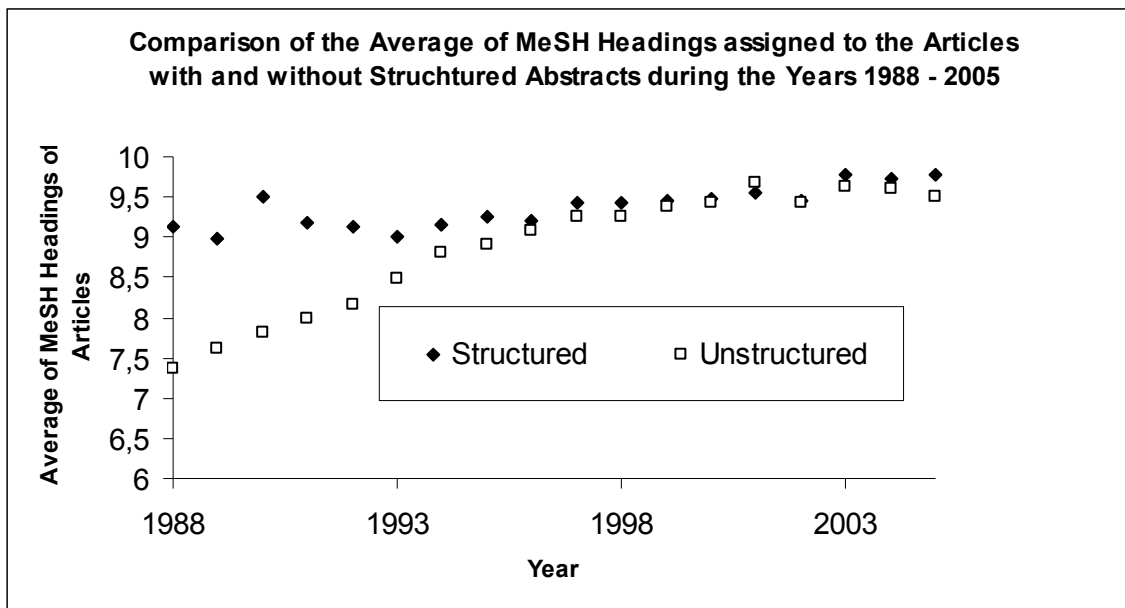


Figure 29: Comparison of the role of structured and unstructured abstracts on the average number of MeSH headings assigned to the journal articles in MEDLINE during the years 1988 – 2005.

In the first eight years after the introduction of structured abstracts they affected the in-depth indexing of articles and from 1996 the average number of terms received by articles with unstructured abstracts reached the same level that received by structured ones.

3.3.4 Journal priorities for in-depth indexing

As explained before, in an indexing system like MEDLINE, the depth of article indexing corresponds with the level of priority given to the journals. NLM gives the journals a number. This is the priority of the journal assigned for depth of indexing. This field is labeled by “PY”(Priority). Valid values are 1, 2, or 3. This is considered to be an in-house data element for management purposes and NLM does not document it for the public online users⁵ Despite this, the priority level of journals for in-depth indexing can be determined through the average number of index-terms assigned to each page of an indexed article.

⁵ See „Indexing priority” in the References section.

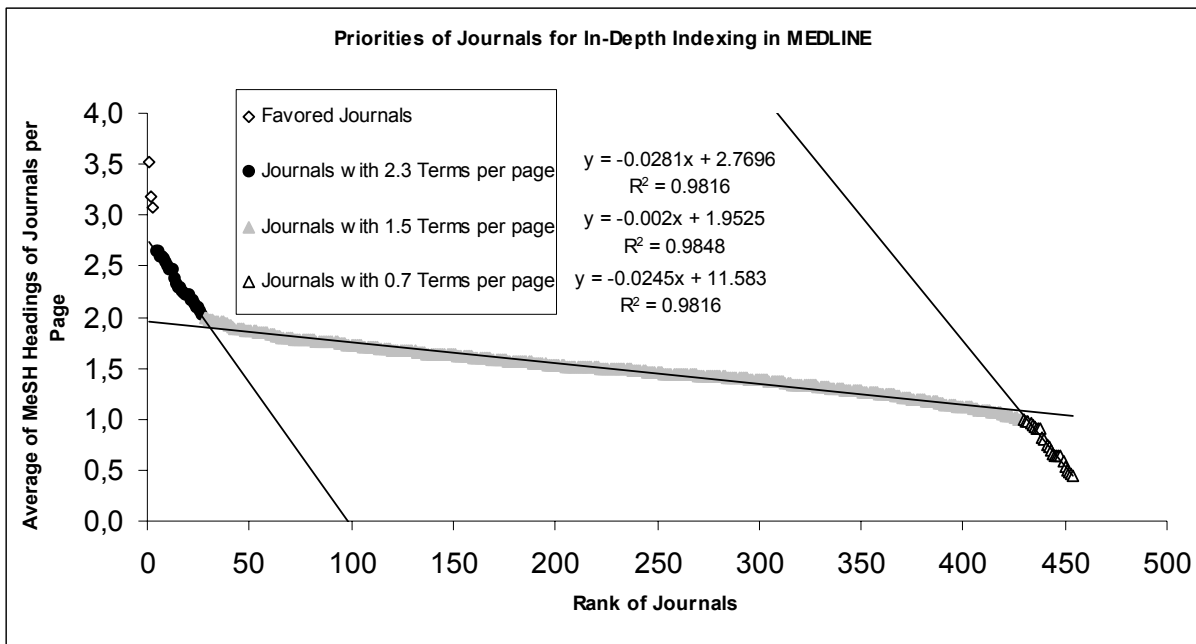


Figure 30: Journal priorities for in-depth indexing. 454 journals were ranked in order of the average number of MeSH headings per page.

The distribution of the MeSH terms per page in the figure above shows several regions of journals for in-depth indexing in MEDLINE. The first three known journals can be counted as the exceptions and as most favored journals in MEDLINE. Except for these, we can observe three different regions that give rise to three priority levels.

The first three journals cover ~0.7% of the collection. Each of them were assigned an average of 3.3 MeSH headings per page (i.e. between 3.1 and 3.5 headings).

Region (1) consisted of ~5.3% of the journals and each journal was assigned an average of 2.3 MeSH headings per page (i.e. between 2 and 2.7 headings).

Region (2) consisted of ~88.5% of the journals and each journal was assigned an average of 1.5 MeSH headings per page (i.e. between 1 and 2 headings).

Region (3) consisted of ~4.8% of the journals and each journal was assigned an average of 0.7 MeSH headings per page (i.e. between 0.4 and 1 heading).

The average of headings per page in every region decreases linearly. In the first region it is “-0.0281” headings on average, in the second very low (-0.002), and in the third region it is “-0.0245” headings on average.

3.3.4.1 Journal Impact Factor

Do journals with higher IF get more headings on average or not? To answer this question, we want to find the role of IF on the priorities given to the journals for in-depth indexing. The distribution of MeSH headings per journal reveals that they can be broken into two groups: IF equal and greater than eight ($IF \geq 8$) and IF smaller than eight ($IF < 8$).

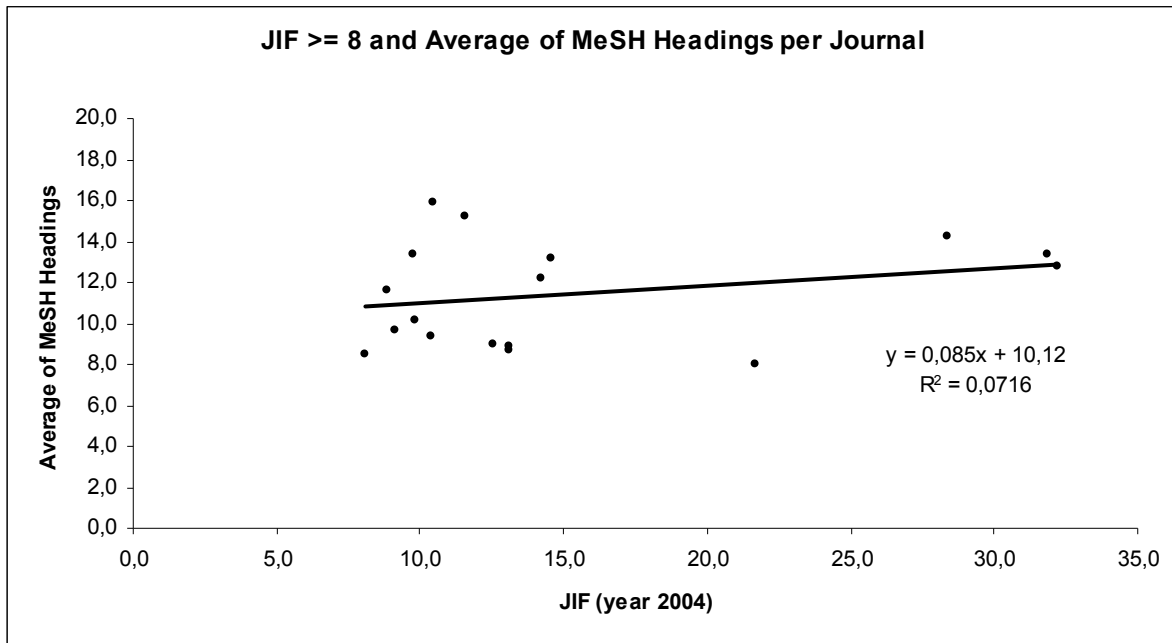


Figure 31: Journal Impact Factor (JIF) ≥ 8 and average number of MeSH Headings per Journal.

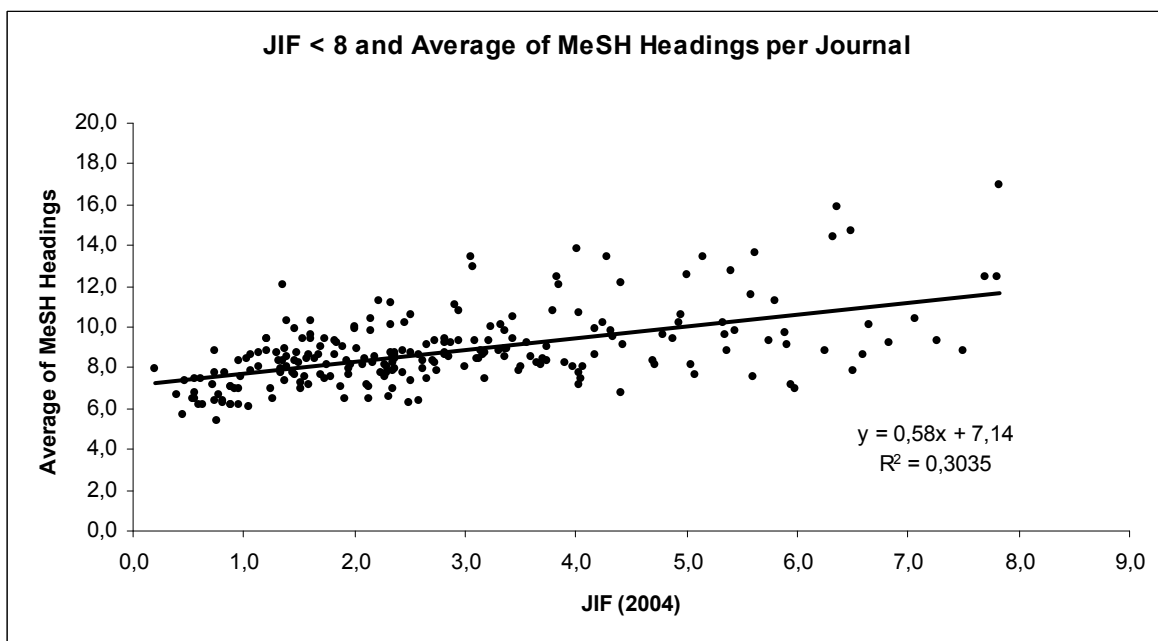


Figure 32: Journal Impact Factor (JIF) < 8 and average number of MeSH Headings per Journal.

As is shown in the figures above (Figures 31 and 32), there are linear relationships between the average number of MeSH headings of journals and their Impact Factors. If the JIF is equal or greater than eight, the increase per IF is nearly 7 times higher than for IF = 0 to 8.

In comparison to this observation, the distribution of MeSH headings per journal page in relation to the JIF, makes clear, that there is no or at least a negligible increase of MeSH-terms per page, in the range of JIF under fifteen (Figure 34). For Impact Factors greater 15, the increase is only 0.06 MeSH-terms per one JIF, with very high scattered values (Figure 33).

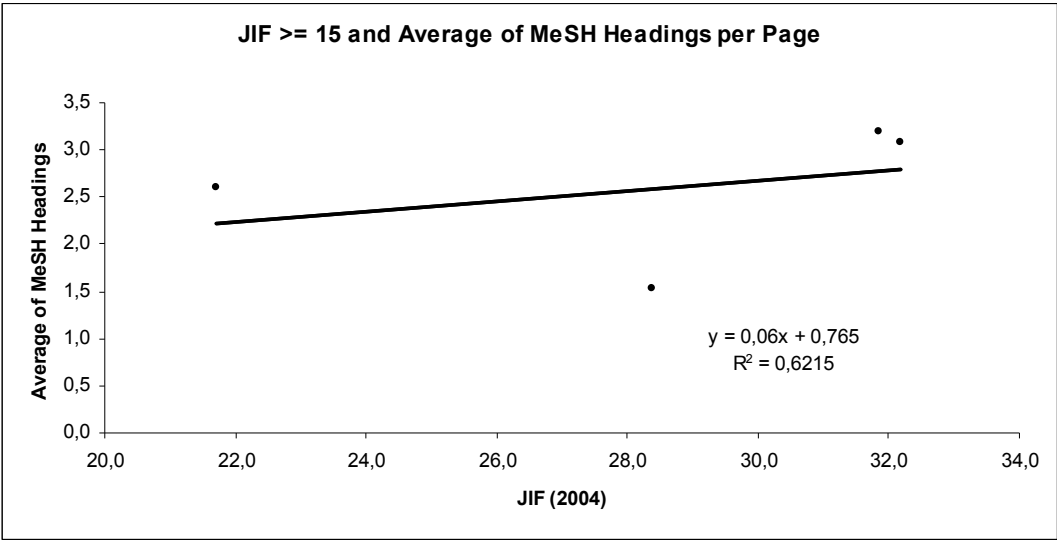


Figure 33: Relationship between JIF >= 15 and average of MeSH Headings assigned to the journal articles per page.

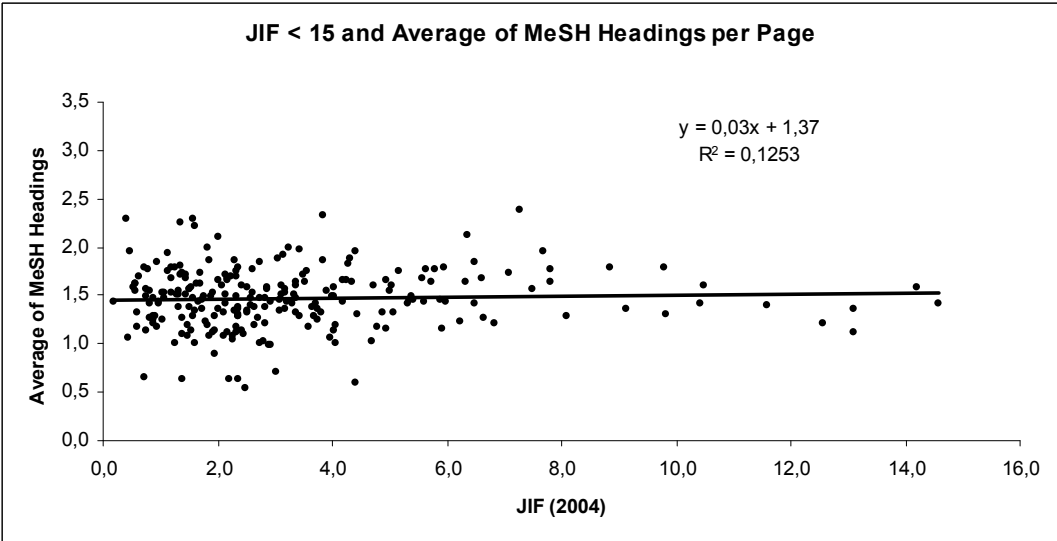


Figure 34: Relationship between JIF < 15 and average of MeSH headings assigned to the journal articles per page.

The difference between Figure 33 and 34 reveals that Impact Factors are in correlation to reviews with more pages than other journal articles. In so far JIF can't be an indicator for priority levels of journals for in-depth indexing. Notwithstanding there are clear priorities for indexers to canalize the findability of the articles in different journals acquired by the NLM. The following table illustrates this point:

Table 9: The first twenty top journals of MEDLINE that are covered by ISI.

Journal Title (abbreviated)	JIF(2004)	Index-Term/ Page
SCIENCE	31.85	3.18
NATURE	32.18	3.07
LANCET	21.71	2.70
NUCLEIC ACIDS RES	7.26	2.38
FEBS LETT	3.84	2.32
CLIN NUCL MED	1.58	2.29
TROP DOCT	0.40	2.28
ONCOL REP	1.36	2.25
J MED ETHICS	1.61	2.21
J BIOL CHEM	6.36	2.13
NEUROSCI LETT	2.02	2.10
AM J EMERG MED	1.82	1.99
AM J PUBLIC HEALTH	3.24	1.99
J CLIN MICROBIOL	3.44	1.98
PEDIATR EMERG CARE	0.47	1.95
CANCER RES	7.69	1.95
INT J CANCER	4.42	1.95
SCAND J INFECT DIS	1.14	1.94
AM J CARDIOL	3.14	1.92
INT J ONCOL	3.06	1.89

The table above is sorted by the average number of terms of journals per page in descending order. It represents the twenty top journals of MEDLINE in the observed sample that are covered by ISI as well. We observe that some journals that are the most important for MEDLINE are not covered by ISI at all. Examples are “TRANSPLANT PROC“with 3.5 or “Nurs Times“with 2.65 MeSH headings per page. By contrast “TROP DOCT” and “PEDIATR EMERG CARE“ are two journals with JIF under 0.5 but they are listed as the first twenty top journals in the table above.

4 Discussion

4.1 Growth of Medical Subject Headings (MeSH)

A thesaurus is dynamic in nature and grows over time. Lancaster, F. W. (1986) says “Of course, a vocabulary developed through an actual indexing operation will grow very fast at the first, but it will reach a plateau after X papers have been indexed ... How large the vocabulary will be depends not only on the subject field but on the specificity of the terms and type of terms used”.

Earlier works which studied the growth of thesaurus terms were restricted to small samples and belonged to the 1960s or 1970s, like Wurm, B. R. (1964) MacClelland, R. M. A. and Mapleson, W. W. (1966), and Blagden, J. F. (1971).

Wurm, B. R. (1964) said “... it is not possible to establish with any accuracy what the relations between file size [number of documents] and term number might be, as the observations made deviate too much from any ideal curve”. He found that the curve that illustrates thesaurus growth should be broken into different curves. He expresses, “... the file growth follows a pattern which can be represented by a mathematical model. This seems to open up the possibility of estimating more exactly the total number of different terms ... This is a probability problem and it seems reasonable to assume that within each category of terms the file size could be presented by combined sum of the sums of a number of geometrical progressions”. It leads him to the following equation:

$$S_n = T_1^{\text{tot}} [1 - (1 - (T_1 / T_1^{\text{tot}}))^n] + T_2^{\text{tot}} [1 - (1 - (T_2 / T_2^{\text{tot}}))^n] + k_n \quad (\text{xvii.})$$

where:

S_n = number of different terms

n = number of documents

T_1 = average number of different low-frequency terms per document

T_1^{tot} = total number of different low frequency terms in file

T_2 = average number of different high-frequency terms per document

T_2^{tot} = total number of different high frequency terms in file

k = average number of different extremely low-frequency terms per document

As we will see in the next section (section 4.2), his mathematical equation showed that the growth of controlled vocabularies is also related to the usage of index terms in the corresponding database.

Another important point of his findings concerns the amount of documents that should be taken from a database as a sample in order to find the growth of its corresponding thesaurus. Wurm claims that a collection of one percent of randomly selected documents makes it possible to measure the total number of descriptor terms from the entire document file. In the current study this is above six percent of the whole MEDLINE (i.e. 948,000 of 16,000,000 documents).

Lancaster, F. W. (1986) has shown that the rate of growth for information (i.e. documents) continues at an exponential pace, while the corresponding rate of growth over the same period of time for number of concepts (keywords and terms) converges logarithmically. The findings of the current thesis support his finding for logarithmic growth of thesaurus. In addition to it, we found that the logarithmic growth rate of the MeSH changed three times:

- From the beginning to the year 1962.
- During the years 1963- 1974.
- During the years 1975 – 2006.

The sample shows that the MeSH has grown following three different functions. The first function is “ $y = 394.62 \ln(x) - 1,867.4$ ” from the beginning to the year 1962, followed by “ $y = 2,371.2 \ln(x) - 16,636$ ” during the years 1963-1974, and continued as “ $y = 2,490.2 \ln(x) - 36,396$ ” up to the year 2006”, where “y” is the number of headings of MeSH and “x” is the number of citations of MEDLINE in the sample. These functions show that the number of headings of MeSH has been related to the number of documents in MEDLINE, revealing that the speed of growth has changed three times.

The functions’ exponents can give some valuable information about the growth of the Medical Subject Headings (MeSH). They present not only three growth phases of MeSH, but indicate the growth rates in each phase as well.

The first phase of growth continued to the year 1962 to create the MeSH thesaurus. This can be called a creation phase. A sudden shift occurred in 1963 when the speed of growth increased 6.8 times and continued to the year 1974. This can be called the first development phase. In 1975 it increased 1.6 times again and began a new period. This can be called a second phase of development. NLM reports that the total of MeSH headings was 6,762 in 1967 (NLM Fiscal Year 1967-68 p. 27) and the amount of headings added to MeSH were

5,000 in 1975(NLM Fiscal Year 1975 p. 38). Thus the beginning of the third phase is motivated by this fact.

Nowadays MeSH has over 23,000 headings. Of them, 7.1% were produced through indexing of 0.3% of documents in the first phase, 26.7% through indexing 2.7% of documents in the second phase and 66.2% through 97% of documents in the third phase.

Other findings show that MEDLINE documents versus MeSH terms have grown exponentially. Lancaster, F. W. (1986) claims that the terms of controlled vocabularies are growing logarithmically, whereas the information (like journal articles) are expanding exponentially. Chen, H. (1994) notes this phenomenon as he sheds light on the information overload problem. If we look at the two following phenomena, we will find how MEDLINE could cope with the problem of information overload of terms:

- Logarithmic growth of a thesaurus has followed several phases not a single function; following each new phase the growth speed has increased, and
- Conversely the rate of inclusion new documents in MEDLINE has decreased simultaneous with each new phase.

These points make clear how the logarithmic growth of MeSH could cope with the exponential growth of its corresponding database to avoid a halt in vocabulary growth and to prevent the problem of information overload. Reducing the exponents has caused a linear increase of HTR. Thus, we should regard the logarithmic growth of thesaurus versus the linear growth of HTR, instead of direct comparison with the exponential growth of information. From this perspective, the problem of information overload should not be of concern. In the next section, we will discuss how the use distribution of headings helps to avoid this problem as well.

The question that should be asked is how to generate the results of MeSH development to the similar thesauri. The current findings show that in order to create a thesaurus like MeSH, a documentation system needs at least ~1,600 different documents which cover the varied topics of interest. Following to the creation of such thesauri, they will develop logarithmically, but their development never reaches the saturation point. The reason is that the inclusion of one more term to them is consequence of 256 documents in their corresponding databases. This fact clarifies the linear dynamic of such thesauri. This rate can be recognised always through analysing the end points of thesaurus development.

To optimise the growth of a thesaurus from today's point of view, two facts of linear dynamics of thesauri and logarithmic growth of them should be taken into consideration. The combination of these two facts yields an equation which matches the actual development of thesauri like MeSH. The yielded equation for the entire MeSH development is $t_e = 3,076.6 \ln(d) - 22,695 + 0.003d$; where "t_e" is the estimated number of terms and "d" the number of documents.

4.1.1 Interaction between Thesaurus development and in-depth indexing

Comparison of the phases existed through MeSH growth (in Figure 7), with the periods of in-depth indexing of MEDLINE (in Figure 27) reveals an interaction between thesaurus development and in-depth indexing. A closer look at the Figures 7 and 27 makes this clear. The problem is that the first phase of MeSH growth was until 1962, whereas our data for studying the changes of MEDLINE in-depth indexing was limited to the years 1965 – 2005. Because of this, the comparison of the in-depth indexing with the first phase of MeSH growth was not possible. The comparison shows that the increase of the average number of MeSH headings per article continued simultaneously with thesaurus growth in the second phase up to the year 1974 and following it, a new period of in-depth indexing began in 1975 as well. From this year on, the average of headings per article has decreased and this reduction continued through 1981. Figure 7 shows at the beginning of the third phase a sudden reduction after a jump, so that the beginning point of this phase remains over the trend line. This occurs simultaneous to a new period of in-depth indexing in MEDLINE. Afterwards, the points follow the trend line.

4.2 Distribution of MeSH headings in MEDLINE

The discussion about the use of index terms and its distribution is not new. Significant works in this area have been published in the last decades. These studies concentrated on small collections that cover a brief period. Lancaster, F. W. (1991) discussed it as an alternative to study the retrievability of items from a database and made a simulation to predict irretrievability.

Various studies tried to find which functions describe the distribution of index terms in databases. Some found it similar to a Zipfian distribution of words in natural texts. For example, Lancaster, F. W. (1986) cited Cleverdon, C. W. et. al (1966) work which reported

that the distribution of index terms is Zipfian. Wall, E. (1964), on other hand, claimed the distribution of term usage is log-normal.

Umstätter, W. (1986) found that the distribution of thesaurus terms yields an exponential function (normal-log) in the GEOLINE database. We saw that the distribution of the majority of MeSH terms agreed with his finding.

The effort in the current work was to find the distribution of the MeSH headings in MEDLINE during four periods with different intervals. Each of them began in the year 1965. The first case concerned the use frequency of the headings up to 1970. The second did the same but for a longer interval up to 1980. The third observed it up to the year 2000 and finally the fourth case had the longest interval and showed the distribution of MeSH headings up to 2006.

The results showed that the MeSH headings can be divided into three classes: highly frequented, normally frequented and rarely frequented. These classes follow the classification of terms done by Salton, G. (1975) and Salton, G. and Yu, C. T. (1973).

The highly frequented headings are always distributed double-logarithmically, normally frequented, exponentially and rarely frequented linearly. This means, in contrary to the distribution of words in a natural text, the distribution of thesaurus terms can't be described by a single function, whereas the distribution of words in the natural texts follows a power law function that is known as Zipfian's distribution. In fact, the distribution of index terms shouldn't follow Zipfian's model because a natural text is static while the databases are dynamic. The vocabulary used for natural texts is very big, but indexing of documents is limited to terms of the corresponding thesaurus.

4.2.1 Highly frequented headings

In the previous section, we observed that the curve of highly frequented terms is tending toward a slope. This indicates that the difference in number between the richer and poorer terms of this class is becoming greater. The reason is rooted in the nature of terms. They are either check tags or broad terms. The check tags are on the top of this class and the number of them is limited. They are topics of potential interest, regardless of the general topic of the documents. Because of this, their usage grows faster than other terms. The broad terms also classify the documents into large divisions of the literature. In fact, terms that have the ability

to cover wider subject areas are always used more because they can be assigned always to the higher number of documents.

We see that the motives of the most frequented words in natural texts are different from the terms of thesauri. The most frequented words are normally function words: articles like “the”, “a” and “an” or proposition like “of”, “in”, “for” or hyper words like “and”. These words do not have a meaning in themselves semantically. In contrast, every term in a thesaurus bears a meaning, regardless of its category. Their similarity is that they are common terms or words. But the objectives of using check tags and broad terms are different from function words.

The following table will give an instance of the top hundred terms of the highly frequented class in 1980.

Table 10: A hundred of highly frequented MeSH headings in 1980.

<i>Rank</i>	<i>MeSH Heading</i>	<i>Frequency</i>	<i>Rank</i>	<i>MeSH Heading</i>	<i>Frequency</i>
1	Humans	3,042,254	51	Chronic Disease	44,545
2	Female	1,022,206	52	Molecular Weight	41,062
3	Male	1,013,376	53	Body Weight	40,702
4	Animals	984,180	54	DNA	40,066
5	Adult	636,412	55	Prognosis	39,299
6	Middle Aged	469,494	56	Lung	38,490
7	English Abstract	305,606	57	Species Specificity	37,678
8	Adolescent	299,746	58	Sodium	37,652
9	Rats	286,989	59	Evaluation Studies	37,368
10	Aged	285,352	60	Carbon Isotopes	37,213
11	Child	278,665	61	Glucose	36,153
12	Time Factors	212,588	62	Anti-Bacterial Agents	35,726
13	Methods	186,091	63	Skin	35,505
14	Research Support, U.S. Gov't, P.H.S.	172,451	64	Potassium	35,372
15	Pregnancy	165,945	65	Acute Disease	34,758
16	Comparative Study	165,119	66	Cell Membrane	34,473
17	Child, Preschool	164,037	67	Spleen	34,257
18	Mice	148,256	68	Lymphocytes	34,032
19	Infant	129,140	69	Myocardium	34,006
20	Rabbits	114,691	70	Swine	33,816
21	In Vitro	114,016	71	Neoplasms	33,710
22	Infant, Newborn	106,693	72	Protein Binding	33,517
23	Liver	105,196	73	Culture Media	33,434
24	Dogs	101,477	74	Clinical Trials	33,099
25	Age Factors	99,206	75	Sheep	32,765
26	United States	95,261	76	Cells, Cultured	32,668
27	Kinetics	92,716	77	Histocytochemistry	32,486
28	Microscopy, Electron	74,891	78	Mathematics	31,881
29	Cattle	73,953	79	Binding Sites	31,870
30	Diagnosis, Differential	72,308	80	Haplorhini	31,867
31	Hydrogen-Ion Concentration	67,595	81	Mutation	31,670
32	Kidney	64,096	82	Hypertension	31,645

33 Brain	58,569	83 Heart Rate	31,425
34 Postoperative Complications	56,677	84 Spectrophotometry	31,310
35 Research Support, U.S. Gov't, Non-P.H.S.	55,467	85 Models, Biological	31,219
		Transplantation,	
36 Blood Pressure	52,517	86 Homologous	31,139
37 Guinea Pigs	51,345	87 Insulin	30,269
38 Muscles	50,389	88 Electroencephalography	29,941
39 Chemistry	50,104	89 Oxygen Consumption	29,849
40 History, 20th Century	49,595	90 Heart	29,795
41 Sex Factors	49,573	91 Oxygen	29,502
42 Cats	49,215	92 Biopsy	29,379
43 Tritium	46,971	93 Antigens	29,307
44 Escherichia coli	46,622	94 Cell Line	29,282
45 Electrocardiography	45,600	95 Electric Stimulation	29,187
		Dose-Response	
46 Calcium	45,320	96 Relationship, Drug	28,837
47 Erythrocytes	45,083	97 Neoplasm Metastasis	28,701
48 Follow-Up Studies	44,952	98 Cell Nucleus	28,092
49 Amino Acids	44,726	99 Myocardial Infarction	27,870
50 Temperature	44,641	100 Antibody Formation	27,688

As the table above shows, most of the first hundred terms are check tags and the others cover a very broad class of literature.

4.2.2 Normally frequented headings

We saw that $75 \pm 2\%$ of headings belong to the normally frequented class and we also observed that their use frequencies decrease exponentially. They are the terms that have more specificity than the most frequented headings and querying them for retrieval should return higher precision.

Use distributions of terms in four different years' intervals showed that the growth rate of those with higher ranks is decreasing and instead, the rate of those with lower ranks is increasing. This indicates that the attention to older terms is gradually decreasing while that for new terms is increasing. Thus, the terms of this category should have an average useful life. The use of an instance term will reach to the end line, when it is overloaded by information (documents). This happens when there is enough literature around it. Besides this phenomenon, new subjects arise and scientists take them into consideration. The inclusion of new terms into a thesaurus is due to this fact.

MEDLINE has to hold the recall and precision ratios at an optimum level. This indicates a positive relationship between the value of HRU and the total number of MeSH terms. The total numbers of terms are always five times more than HRU. In other words, despite

changing the parameters of the use distributions (exponents and cutting points), the use of the richest term has remained thirty-one times more than that of the poorest one (Table 7 and 8).

4.2.3 Rarely frequented headings

A thesaurus is not static and allows new terms to be added to it over time. New headings should be used less logically. But the existence of this category shows that these terms are very narrow and because of this, they have been used less than the others. The findings of the current work reject this idea. In fact, there are only two classes of terms: highly and normally frequented.

Rarely used terms are a temporary class and will shift to the others in the future. As we saw, none of the rarely frequented headings in 1970 and 1980 contributed to the list of rarely used headings in 2006. Thus, none of the least frequented headings are in that category forever.

The following table shows a hundred of the rarely frequented headings in the year 1980. It shows that they are not the narrowest and have not necessarily the most specificity:

Table 11: A hundred of the rarely frequented MeSH headings in 1980.

<i>Rank</i>	<i>MeSH Heading</i>	<i>Frequency</i>	<i>Rank</i>	<i>MeSH Heading</i>	<i>Frequency</i>
11901	Hydropneumothorax	43	11951	Acetoin	42
11902	Myoclonic Cerebellar Dyssynergia	43	11952	Hydroxyestrones	42
11903	Chancre	43	11953	Bongkreic Acid	42
11904	Formate-Tetrahydrofolate Ligase	43	11954	Coumaphos	42
11905	Dithizone	43	11955	Heptachlor Epoxide	42
	Peptococcaceae				
11906		43	11956	Ascaridoidea	42
11907	Alchemy	43	11957	Dexetimide	42
11908	Lactose Factors	43	11958	Indoramin	42
11909	Herpangina	43	11959	Tuber Cinereum	42
11910	Hemerythrin	43	11960	Aminoacetonitrile	42
11911	Arginine-tRNA Ligase	43	11961	Hordeolum	42
11912	Bupranolol	43	11962	Prostaglandins G	42
11913	Flurogestone Acetate	43	11963	Hospitals, Satellite	42
11914	Catalogs, Library	43	11964	Streptomycetaceae	42
11915	Liniments	43	11965	Data Interpretation, Statistical	41
11916	Paraganglia, Chromaffin	43	11966	Evoked Potentials, Visual	41
11917	Sunstroke	43	11967	Vanadates	41
11918	Molecular Sequence Data	42	11968	Infant Welfare	41
11919	Disease Progression	42	11969	Blinking	41
				Synovitis, Pigmented	
11920	Anastomosis, Surgical	42	11970	Villonodular	41
11921	Hospital Information Systems	42	11971	Heavy Ions	41

			Phosphoenolpyruvate	Sugar	
11922	Cranial Nerve Diseases	42	11972	Phosphotransferase Syste	41
11923	Deoxyribonuclease I	42	11973	Wasp Venoms	41
	Phosphotransferases (Alcohol Group				
11924	Acceptor)	42	11974	beta-Alanine	41
11925	Pirenzepine	42	11975	Whole Blood Coagulation Time	41
11926	Abdominal Wall	42	11976	Camelids, New World	41
11927	Mesna	42	11977	Transistors	41
11928	Hearing Loss, High-Frequency	42	11978	Pulmonary Subvalvular Stenosis	41
11929	Fuchs' Endothelial Dystrophy	42	11979	Fused Teeth	41
11930	Reproductive Control Agents	42	11980	Group Practice, Prepaid	41
11931	Muscarine	42	11981	Mucinosis, Follicular	41
11932	Lisuride	42	11982	Aldicarb	41
11933	Oxonic Acid	42	11983	Q-Sort	41
			Physical Therapy Department,		
11934	Oral Submucous Fibrosis	42	11984	Hospital	41
11935	Fluorometholone	42	11985	Guam	41
11936	Cefadroxil	42	11986	Uranyl Nitrate	41
11937	Chronology	42	11987	Acetylthiocholine	41
11938	Deoxycytosine Nucleotides	42	11988	Colposcopes	41
11939	Contracts	42	11989	Acetolactate Synthase	41
11940	Hexanones	42	11990	Coitus Interruptus	41
11941	Maleic Anhydrides	42	11991	Nefopam	41
11942	Exanthema Subitum	42	11992	Ricinoleic Acids	41
11943	Phenylmercuric Acetate	42	11993	Fertility Agents	41
			Oxidoreductases,	O-	
11944	Carnitine Acyltransferases	42	11994	Demethylating	41
11945	Bacteriophage mu	42	11995	Pantetheine	41
11946	Blushing	42	11996	Programming, Linear	41
11947	Herpesvirus 1, Cercopithecine	42	11997	Acidithiobacillus thiooxidans	41
11948	Moxibustion	42	11998	Hexetidine	41
11949	Chlorzoxazone	42	11999	Eliminative Behavior, Animal	41
11950	Linuron	42	12000	Thioglucosides	41

The terms that have very low frequency will be deleted from MeSH or will be replaced with others. These changes will be updated in all documents of MEDLINE automatically. For example, 948,000 randomly selected records contained 23,199 distinct headings, but searching them through PubMed after one week showed that only 22,414 from them were accessible. That means 785 headings that were used very rarely were removed from the MeSH. Thus, the rarely frequented headings in MeSH either will be shifted to the above classes or will be removed.

4.3 Factors related to the number of index terms of articles

The main emphasis of this section is on the *average number of MeSH headings* of journal articles. The aim is to argue the effects of several factors that are assumed to have a relationship with the number of index terms of documents. Since the type of documents could be counted as one of the effecting factors, the study was limited to journal articles in MEDLINE.

The effect of each factor depends on a motive. From this point of view, they can be divided into different groups. In fact, if documents bear more contents, they will get more index terms. The type and length of documents are two determining factors that have a relationship to the contents of documents. How the contents of documents are presented relates to content presentation. Abstracts are tools for mirroring the text contents in smaller dimensions. Well-structured texts, such as research articles, also help to determine the contents of documents. Thus, the abstract and text structure are two presentation-related factors. On the other hand, date of indexing, Journal Impact Factor (JIF), and priorities given to the journals for in-depth indexing have not necessarily any relationship with the contents of documents, but are dependent on the indexing policy. These can be called policy-related factors.

Lufkin, R. C. (1968) presents some parameters that are pertinent to the nature of documents and their effect on the number of index terms:

- „1. Number of pages
2. Document format (arrangement of information within the document)
3. Author's purpose in writing the document
4. Level of approach (academic level of the author's intended audience)
5. Subject area for which the document was selected.”

The rest of the current work will discuss the effects of several factors on the assigning of MeSH headings to the journal articles by NLM indexers: length of articles, abstracts of articles, language of articles, date of inclusion of articles into MEDLINE, priorities of journals for in-depth indexing, and Journal Impact Factor.

Besides discussing the effect of the length of articles as an independent factor, it will be also considered as an indicator for the effects of the other factors.

4.3.1 Length of Articles

We can assume that the size of any indexes, regardless of their types, is influenced by the size of their corresponding texts. The indexes of books are the ones most common example.

Anderson, M. D. (1971, p. 121) is concerned with how much of a book can be run by an index. He claims that the size of an index runs from 1% through 15% of books.

Wellisch, H. H. (1991, p. 208-213) states that different factors influence a book index: space and time allotted for the index, technical data in the text, nature of the text (e.g.. children's books, history book, reference, scientific and technical books), names of persons or specifically named items, subheadings already included in the text and *length of the text*.

Article length influences not only the number of index terms, but the size of other criteria as well. Abt, H. A. and Garfield, E. (2002), for example, studied the effect of the article lengths on the number of cited references. They studied four groups of journals: (1) biochemistry and molecular biology, (2) immunology, (3) general medicine, and (4) the social sciences. Their results show a linear correlation between article length and mean number of references in all of four journal groups. As a result, we can assume that the number of article references should then effect the number of index terms as well, because their number is related to article length.

The effect of length is not only related to the amount of document contents but to the time needed for indexing as well. The indexers require more time to review and scan the contents of lengthy texts. As a result, they pay less attention to large documents. The following table presented by Lufkin, R. C. (1968, p. 36) shows the average time in minutes that indexers consumed to review and index the documents per page:

Table 12: Average review time per page, versus document length, for experienced indexers. This table is derived from Lufkin's work in 1968 (p. 36).

<u>Number of Pages</u>	<u>Number of Documents in Sample</u>	<u>Average Review Time min. per page</u>	<u>Standard Deviation</u>
1	26	10.42	7.13
2	910	4.45	3.71
3	445	3.99	3.14
4	237	3.59	2.59
5	169	2.98	2.12
6	125	2.39	1.63
7	97	1.95	1.16
8	97	1.81	1.30
9	58	2.03	1.49
10	56	1.78	1.33
11	37	1.45	1.02
12	28	1.78	1.08
14.6 (average for sample of 13-19 pages)	32	1.00	1.05
25.2 (average for sample of 20-34 pages)	12	0.91	0.64
over 100	4	0.13	0.02

The table above shows that indexers consume less time for reviewing the larger texts per page. For example, indexers reviewed documents with one page in "10.42" minutes on average, whereas the total time used for documents with ten pages was " $10 \times 1.78 = 17.8$ " minutes (i.e. 1.78 minutes per page).

The subject area of the documents is a determining factor as well, as stated by Lufkin, R. C. (1968). Abt, H. A. (1992) also found that the number of words per page varied in different disciplines. He stated: "the word content of purely textual material varies from 510 words per page in the mathematical journal to 1,190 words per page in the astrophysics journal. The average is 1,000 words per page".

In addition to the fact mentioned above, there is not a certain standard for the number of words per journal page. Every journal has its own policy in this area. This policy can also change over the years. Schulman, E. et. al. (1997) studied some aspects of article lengths in astronomical publications. They found that, due to changes in the article formats in the journals, the number of their words per page increased over the years.

The fact mentioned above leads us to look at the types and tokens of documents to see how they are related to the number of pages. Types are concerned with different unique words and tokens with their frequency within texts. The known work in this area is by Heaps, H. S. (1978). He described the relationship between the size of a text consisting of words and its

distinct vocabulary. In other words, he showed how text tokens and types are related. His finding is called Heap's Law. It can be formulated as $V_R(n) = kn^\beta$; where " V_R " is the subset of the vocabulary " V " represented by the instance text of size n , " n " number of whole words within the text, " k " and " β " are two constants. Thus, Heap's Law is based on a power law function. Its exponent (β) in English texts is between 0.4 – 0.6 and its cutting point (k) between 10 –100. In the current work (Figure 17), the (β) is ~ 0.58 and the k is ~ 8.8 . The exponent follows the law but the cutting point is 1.2 less than what is expected.

On other hand, Kortendick, O. and Fischer, M. (1996) treated the outline of the cultural materials as a usual text. They then plotted the types of outlines with the tokens of them. The result was a linear function with a power of 0.03. Based on their work, every hundred of outline tokens exist with three new distinct words (types), regardless of the tokens amount. Since we observed the increase of whole texts tokens, the variety of the types was decreasing relative to them. It is rooted in the nature of power law functions. By increasing of the values of the x-axis, the growth rate of the y-axis becomes gradually less.

In addition, we saw the relationship between the tokens (text words) and the number of pages follows Pearson's function (Figure 18). Its equation reveals that for medical articles in MEDLINE each page contains an average of 843 words. If we suppose each page of articles has 843 words and multiply this by the number of pages, we can determine the average number of tokens for articles with different numbers of pages. Solving the equation of the power law function yielded from the plotting of the tokens and types in Figure 17, through replacing the x-variable with the values of the tokens will get the number of distinct words (types). Dividing the number of types by the number of pages will yield the types per page. The following table shows the results of the above-mentioned operation.

Table 13: Reduction of the types per page rate in relation to larger articles.

N. Pages	Tokens	Types	Types per Page	N. Pages	Tokens	Types	Types per Page
1	843	426	426	16	13488	2108	132
2	1686	635	318	17	14331	2183	128
3	2529	803	268	18	15174	2256	125
4	3372	948	237	19	16017	2328	123
5	4215	1078	216	20	16860	2398	120
6	5058	1197	200	21	17703	2466	117
7	5901	1309	187	22	18546	2533	115
8	6744	1413	177	23	19389	2599	113
9	7587	1513	168	24	20232	2663	111
10	8430	1608	161	25	21075	2727	109
11	9273	1698	154	26	21918	2789	107
12	10116	1786	149	27	22761	2851	106
13	10959	1870	144	28	23604	2911	104
14	11802	1952	139	29	24447	2971	102
15	12645	2031	135	30	25290	3029	101

The table above illustrates a decrease of the types per page through enlarging the texts. For example, the types per page of articles with ten pages are “2.6” times less than those with one page. The table indicates why the shorter articles get more MeSH terms than the larger ones. But why the logarithmic growth of the average number of index terms per article falls when the article’s length reaches ten pages is another question which is rooted in the time of indexing. Table 12 illustrated this fact. The role of the abstracts in presenting the contents of the documents will be discussed in the next section.

One may argue that the current work relates to controlled index terms, thus the types and tokens of an instance text can’t be counted as an indicator of its contents. We know that most automatic indexing systems follow the patterns of human indexing in some cases to produce the best possible index terms. Methods like vector space and probabilistic models, are based on word frequency of texts and the normalization of their index size. One example was reported by Van Rijsbergen, C. J. (1979, p.11). He relied on the frequency of the words to determine upper and lower cut-offs through excluding the most and least frequented words. In addition, Boroko, H. and Barnier, C. L. (1978) claim “in general 34-86% of the index terms, assigned by human indexers, can be derived from title words only... more in the fields of science and engineering and less in the social sciences and humanities”. This reveals that plenty of text contents are not only embedded in titles but in the text corpuses as well. Most words within the texts convey a special content, except functional words, such as prepositions, conjunctions, or articles, which have no lexical meaning because their function is to express grammatical relationships. Despite their high frequencies within texts, their

variety is limited. These efforts indicate that increasing text types are a sign of the expanding contents of the text.

We can argue better why the increase of papers' lengths resulted in assigning more or less index terms to the articles: increasing the lengths of articles causes an increase in text tokens, and therefore an increase of tokens brings an increase in types. On the other hand, Table 13 illustrated that the types per page decrease gradually through the enlargement of articles. As a result of this gradual reduction, when the lengths of articles grow, the function of the relationship between number of pages of articles and average number of their index terms becomes logarithmic. As it is the case for the logarithmic functions, the values of the y-axis increase very fast at the beginning points and the curve becomes gradually flat.

As Karbasi, S. and Boughanem, M. (2006) found, if the length of a document grows, the degree of importance of the terms within this document decreases. Their study focused on the retrieval and indexing by vector space model, an algebraic model for information filtering, information retrieval, indexing and relevancy of textual documents using natural language processing methods. They also found that the correlation between document length and the estimated degree of importance of the term is higher than the correlation between document length and term frequency. This means that the role of variation of distinct types is higher than that of the size of documents. They also claim that the specificity of the index terms assigned to the shorter documents may be higher, but this point should be studied separately.

The current work is concerned with the quantity of the index terms. Therefore, we can not overlook that the quality of index terms can also be related to the length of documents. Singhal, A. et. al. (1996) showed in the field of automatic indexing "Contrary to the general assumption that the probability of relevance of a document to a query is independent of the document length, in the TREC collection, probability of relevance of a document increases with its length". They add "Long and verbose documents usually use the same terms repeatedly. As a result, the term frequency factors may be large for long documents. Long documents also have numerous different terms. This increases the number of word matches between a query and a long document, increasing its chances of retrieval over shorter documents".

4.3.2 Presence of abstracts

Abstracts are a brief summary of the most important points in a scientific paper. They mirror the significant dimensions of an instance text. This characteristic feature makes it possible for abstracts to be used widely as an auxiliary source for indexing. They help users and indexers to scan the contents of the texts quickly.

The extent of how many concepts of texts can be represented through abstracts was illustrated by Janos, J. (1975). He selected 200 documents which were indexed by human indexers. They were then indexed again only through their abstracts by using an automatic indexing method. Janos found that indexing through automatic indexing of abstracts could find “3.45” descriptors from “4.25” relevant descriptors on average that were assigned to the documents through full-text indexing by humans indexers. This shows that abstracts bear a significant amount of the contents of their corresponding text.

The comparison of articles with and without abstracts in MEDLINE showed on the average that those with abstracts received more index terms. But the abstracts had not any effect when lengths of articles were more than sixteen pages. We should discuss why they are an effective factor on the number of index terms assigned and why they will be ineffectual, when the lengths of articles exceed the sixteen pages. We may immediately envisage that they help the indexers scan the contents of the texts faster than reading them from beginning to end. This factor leads us to consider the indexing process.

Indexing is a process that comprises a number of steps. Mai, J. – E. (2001), Mai, J. - E. (2005), and Lancaster, F. W. (1991) claim that the process of indexing follows two steps: 1) the indexers analyze the documents to determine its subject matter, and 2) they translate the subject matter into index terms. Mai, J.-E. (2001) describes the first step as a response to the presence of the document. He explains: “It consists of the act of examining the document (i.e. the title, the table of contents, the abstract, if there is one, the back of the book index, reviews of the item, and so on) in order to identify its subject“.

Instead of seeing the process as indexing steps, David, C. et. al. (1995) viewed it as a problem space. From this point of view, they divided indexing into two stages: knowledge space and resolution space. “The knowledge space includes the set of declarative and procedural knowledge which are potential components of the problem. The resolution space consists of the major stages as defined by the norms and what we know of the usual indexing procedure”. To them, content analysis and concept selection are two procedures within resolution space.

The presence of abstracts within texts eases the task of indexing. They help indexers to do content analysis and selection in a shorter time. In addition, they give a deeper insight into documents' contents.

Getting deeper insight from abstracts refers to text comprehension. Wang, Y. and Gafurov, D. (2003) define text comprehension as the action or capability of understanding. They believe that the study of the mechanism and process of comprehension is a fundamental issue in cognitive informatics. From their point of view, "in the first step to comprehend a given real entity or concept, the brain searches the corresponding virtual entity and its relations to objects in the abstract layer". The next step depends on the results of the search for relations. "The ideal search result is that adequate relations have been found. In his case, comprehension is almost reached". Some times the person obtains only partial or no comprehension. This would happen where no sufficient relations between the concept and the abstract layer of the brain were found. This fact addresses mostly the knowledge space of the indexers, where the subject is known as their specialty. Though a necessity of indexing, it is not the only one. In addition, finding the relations for understanding the texts-contents is related to the time that an indexer spends reading documents depending on their lengths. As Lufkin, R. C. (1968) found, the spending time per page decreases by increasing the number of pages of documents. Abstracts are alternate tools for solving this problem.

The other argument concerns the indexers' aim for reading documents. Mulvany, N. C. (1994) says that indexers read the texts differently from users who take an interest in them. Indexers read them quickly but accurately to synthesise the text. Her statement leads us to see how the concepts and real entities transferred into memory for understanding and learning of texts.

Indexers don't read texts for learning. They don't try to keep the contents of documents in Long Term Memory (LTM). Farrow, J. F. (1991) presents a cognitive method for indexing. He states that Short Time Memory (STM) is limited to seven plus or minus two items ($STM^{\text{items}} = 7 \pm 2$). These items will be transferred to LTM. Concepts in LTM can last from as little as thirty seconds to as long as decades. As the aim of the indexers is not to read for learning, the analyzed contents remain in most indexers' LTM very briefly. Since abstracts reflect the contents of texts in brief, they can be read in a short time. Thus, the interval between comprehending the concepts and translating them into index terms becomes shorter, causing the indexers to recall more concepts when translating them into index terms in the second step of indexing.

The other question is why abstracts don't affect on the number of index terms, when the length of articles reached to the sixteen pages. The main reason is the restriction of abstracts to a certain number of words. NLM expresses that the maximum length of abstracts in MEDLINE for records created after the year 2000 is 10,000 characters. The original policy on inclusion of abstracts set a limit of 250 words for acceptance by NLM. In 1984, two changes were made in the policy: 1) the limit of words was raised to 400 words for articles of more than ten pages in the core journals identified by National Cancer Institute, and 2) abstracts exceeding the 250- or 400-word limit were to be included in truncated form at the end of the sentence closest to the word limit⁶.

Due to the word limitation of abstracts, their content coverage reaches the point of saturation, the maximum capacity of content representation. In the current work this point is reached with articles of sixteen pages.

We observed that the effect of abstracts on the number of index terms of articles decreased for those longer than ten pages, before reaching the level of those articles without abstracts. The policy of expanding the word limitation of abstracts to 400 for articles with ten and more pages in core journals supports the finding of the current work. We found that the effect of the abstracts decreases from this mentioned point. Expanding word limitation could solve the problem of content-presentation of the large articles to an extent, but would not fix it completely, unless the policy would expand the number of words of abstracts by expanding the length of articles. It raises the following question: When the length of abstracts increases, is the number of their significant words also increasing or not?

Garas, G. J. (1968) illustrated the relationship between the number of significant words and the length of abstracts scaled by the number of words. The following figure is taken from his work:

⁶ See "MEDLINE®/PubMed® Data Element (Field) Descriptions" in References section.

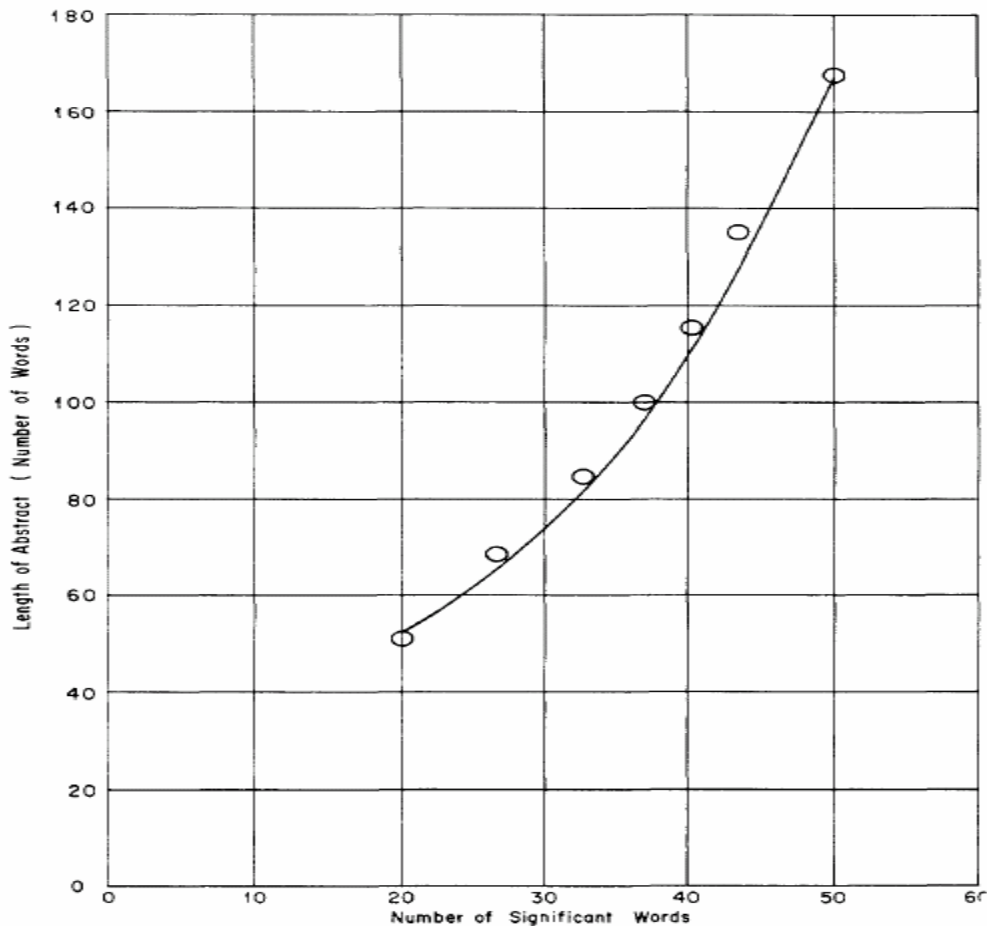


Figure 35: Relationship between the number of words of abstracts and number of significant words. This figure is a cutting from Garas, G. J. (1968).

He explains that the not-significant words were common words such as conjunctions, prepositions, articles, and words possessing little discriminatory power as well as many adjectives and nouns such as method, study, problem, which are not used for indexing. Multiple occurrences of the same significant word in an abstract were ignored; each significant word thus was counted only once in any given abstract.

Besides the arguments above, one other argument is the importance of indexed articles if their abstracts are added to a database. It means that adding abstracts is an indication of the importance of the indexed documents. Because of this, they can receive more index terms than those of which abstracts are not included.

It is remarkable that „if the abstracts are not well made and the titles are not precise, they are not definitive sources for the extraction of concepts” (Munoz Rodriguez, J. V. and Gil Leiva, I., 1997). In addition, titles and abstracts have not enough information for human and Artificial Intelligence (AI) indexers to determine the total contents of an instance text. In

order, indexers need something more to do this. This fact is also important for the field of “data mining”. In this context the experience with MEDLINE has clearly shown that the number of descriptors, in relation to the tokens per page is of high importance. Text mining without the full-text is insufficient. Besides full-texts, there is a need to develop an organic growing thesaurus.

4.3.2.1 Structured and unstructured abstracts

The effect of abstracts on the number of index terms of documents varies depending on their forms.

Garas, G. J. (1968) studied the idea of indexing from abstracts. He wanted to see which forms of abstracts were more suitable: indicative or informative. First, he selected at random one hundred and ninety-nine informative abstracts from International Aerospace Abstracts (A abstracts) and paired them with the same set of indicative abstracts from The Engineering Index (B Abstracts). A pair of abstracts, thus, consisted of one A and one B abstract, both of which referred to the same document. He concluded: “There is a fairly good agreement between the index terms contained in the two types of abstracts as 18 terms were common to both abstracts. Since 71.3% of the index terms of a document were contained in its informative abstract (vs. 52.6% in the indicative abstract), the informative abstract should be preferred as a substitute for the entire document. Over half of the terms found in a document are likely to be contained even in a short, indicative abstract. This type of abstract, therefore, may be an acceptable source of terms for some indexing applications. Obviously, the decision to use abstracts rather than entire documents, and if so which type of abstract, must depend on considerations such as the indexing depth desired, the availability of abstracts and the relative cost of converting to machine readable form”.

He also showed that the average number of index terms assigned to the same document by professional indexers was “22.3” index terms per document. In comparing structured abstracts with unstructured ones, Hartley, J. (2003) saw the following two findings assumed to support the current work: 1. Structured abstracts are 30% longer than unstructured, and 2. “structured abstracts contained significantly more information than did the traditional ones“.

The two works mentioned above support the findings of the current work. We observed that articles with structured abstracts contain 10.11 MeSH headings, whereas those with unstructured contain “8.83” on average (without check tags). This finding varies from what

Harbourt, A. M.; Knecht, L. S. and Humphreys, B. L. (1995) found. They reported the structured abstracts had three more MeSH headings. The comparison of these two forms of abstracts during the years 1988-2005 showed that the effect of unstructured abstracts increased to the level of structured in the year 1996. Since their study of structured abstracts was during the years 1989-1991, their finding lacks this fact.

That structured abstracts can increase the average number of index terms is due to:

their length

their additional content

their form of presentation of the text content.

Guimarães, C. A. (2006) confirms this argument. He says: „The advantage of structured abstracts is that it is easier to understand the text written in shorter paragraphs. ... Structured abstracts contain the most significant data from the paper, and some use them as primary source of information“. His article also strengthens the argument we discussed previously, of how abstracts assigned more index terms relate to easier understanding of the text.

In the following, we will see briefly the advantages and qualifications of structured abstracts gathered from different authors by Hartley, J. (2003):

- *“contain more information (Hartley, 1999a; Hartley and Benjamin, 1998; Haynes, 1993; McIntosh, 1995; McIntosh, Duc and Sedin, 1999; Mulrow, Thacker and Pugh, 1988; Taddio, Pain, Fassos, Boon, Ilersich and Einarson, 1994; Trakas, Addis, Kruk, Buczek, Iskedjian and Einarson, 1997);*
- *are easier to read (Hartley and Benjamin, 1998; Hartley and Sydes, 1997) and to search (Hartley, Sydes and Blurton, 1996) - although some authors have queried this (Booth and O'Rourke, 1997; O'Rourke, 1997);*
- *are possibly easier to recall (Hartley and Sydes, 1995);*
- *facilitate peer-review for conference proceedings (Haynes, Mulrow, Huth, Altman and Gardner, 1990; McIntosh, 1995; McIntosh et al., 1999); and*
- *are generally welcomed by readers and by authors (Hartley and Benjamin, 1998; Haynes et al., 1990; Haynes, 1993; Taddio et al., 1994),*
- *However, there have been some qualifications, Structured abstracts:*
- *take up more space (Harbourt et al., 1995; Hartley, 2002);*

- *sometimes have confusing typographic layouts (Hartley, 2000a); and*
- *may be prone to the same sorts of omission and distortion as are traditional abstracts (Froom and Froom, 1993; Hartley, 2000b; Pitkin and Branagan, 1998; Pitkin, Branagan and Burmeister, 1999; Siebers, 2000, 2001). ..*

4.3.3 Language of Articles

MEDLINE is an English and US-based database. The abstracts prepared on these resources are in English as well, even if the documents are written in other languages. But theoretically, the language of articles shouldn't have any effect on their number of index terms. This is because *Journal Selection for MEDLINE*[®] is based on the following statement: “[MEDLINE] is used internationally to provide access to the world's biomedical journal literature. The decision whether or not to index a journal for this service is an important one and is made by the Director of the National Library of Medicine, based on considerations of both scientific policy and scientific quality”⁷. This means the depth of journal indexing depends only on scientific factors. In fact, the policy of NLM does not assign fewer headings to non-English documents. Thus it can not be counted as a wanted bias. The reason why they are assigned fewer index terms than English articles is based on other issues, not their indexing policy.

Despite of the above statement, we observed that articles written in German (as the second most frequently used language of MEDLINE) have been indexed by fewer index terms on average. This should be seen as the language bias of MEDLINE. It is an issue that has been investigated by different authors, each one having studied it from different points of view. The wide coverage of the documents written in English is one of them.

Tsay, M.-Y. and Yang, Y.-H. (2005) made a “bibliometric analysis of the literature of randomized controlled trials“. Their focus was on articles that their publication type were specified as “Randomized Control Trial” in MEDLINE. In a part of their study, they intended to “find the country and language distributions of the RCT literature from 1990 to 2001“. They found “about 39.9% of the journals and 50.6% of the articles had been published in the United States, England (15.8% of journals and 21.7% of articles) and Germany (6.5% of journals and 6.1% of articles) contribute the 2nd and 3rd most number of articles, followed by Denmark, Switzerland, and the Netherlands, each contributing 2.0% to 4.0% of the total

⁷ See „**Journal Selection for MEDLINE**“ in References Section

journals and articles. Italy, Canada, Ireland, France, and Norway also significantly contribute to the RCT literature“.

Loria, A. and Arroyo, P. (2005) investigated “the language and country preponderance trends in MEDLINE and its causes”. They classified MEDLINE journal articles by country of publication (Anglos/Non-Anglos) and language (English/Non-English) for the years 1966 and from 1970 to 2000 at five-year intervals. Three divisions of the United Kingdom, Australia, Canada, Ireland, New Zealand, and the United States counted as Anglo countries. In a part of their work, they found “non-English papers decreased at a rate of 1,056 fewer papers per year. These trends have led to overwhelming shares of English and Anglo papers in MEDLINE. In 2000, 68% of all papers were published in the 8 Anglo countries and 90% were written in English”. They assume that tendency to publish in English journals from authors in non-English countries were motivated by „ (1) editorial policy changes in MEDLINE and in some journals from Non-Anglo countries, and (2) factors affecting Non-Anglo researchers in the third world (publication constraints, migration, and under [lack of] support).”

Egger, Matthias et. al. (1997) studied the “Language bias in randomised controlled trials published in English and German“. They found key authors of eight leading journals in German, whose works were RCT type. Then they searched them in MEDLINE. If their works were similar in both English and German journals, they were excluded. They found the works which had negative results (in other words, low p values) were published in German journals and those with high “p” value in English. McDonald, S. (2002) concludes from their results: “German-speaking trialists are more likely to report their positive findings in English language journals and their negative findings in local German journals”.

From the above works, we find a tendency to publish the works in English from the authors whose languages are not English. It is clear that authors prefer to publish their works in top and well-known journals, ones that are covered by the top databases or highly scored by known journal evaluating systems such as ISI. Most of the non-English journals are not able to contend with those written in English and, even if covered by the top databases, couldn't belong to the group of journals with high priority of in-depth indexing.

We know that ISI has a considerable role for evaluating journals. Its ranking of journals should be counted as an important parameter that has led the bias toward English journals. Mueller, P. S. et. al. (2006) showed this fact through comparing „the association between impact factors and language of general internal medicine journals“. They compared the impact factors of general internal medicine reported in ISI in 2003. The comparison was done

between English and non-English and also between US and non-US journals. They found that English journals have a higher impact factor than non-English and English journals published in USA more than those written in English but published outside USA. They concluded „Journal impact factor is more associated with journal language (i.e. English versus non-English), rather than journal country of origin“. Narayana, S. M. et. al. (2004) claimed: “most of the non-English journals available on MEDLINE will have an impact factor of zero“.

Debates and researches on JIF challenged its role in assessing the quality and importance of journals in some cases. Dong, P.; Loh, M. and Mondry, A. (2005) express that the IF of non-English journals is lower due to the limited coverage of such journals by the SCI database. Thus, it is not because of their low quality.

We observed that in-depth indexing of MEDLINE has a relationship with JIF, if it is equal or over fifteen. The number of such journals in MEDLINE is very rare. Thus, it can't be counted as a main parameter, even if we suppose the JIF of articles in German was less than those written in English.

The other reason is rooted in the indexers and the nature of indexing. In a most optimistic situation, we can assume that the native language of the indexers who index non-English articles would be the language of the text. In addition to it, they would be specialists in the subject of the articles. Despite this, other problems would still exist.

We learned that indexing has two steps. The first is to read and scan the articles to get their contents, and the second is to translate them into index terms. In the case of controlled indexing, they should be also controlled through corresponding thesaurus or subject headings. This shows that indexers read and get the contents in languages other than English. But in second step they are faced with only English. They should search for appropriate index terms relevant to the contents that have been scanned in other languages through controlled vocabularies in English. This makes the process of indexing more difficult and time consuming. Indexers might choose those contents that they could translate quickly into English index-terms. This can also influence the quality of indexing of non-English documents in an English-based system. But this claim should be investigated.

However, articles written in German have less chance of being retrieved in comparison to English articles, because the number of index terms assigned to the documents reduces the probability of their retrieval. But the effect of language on the number of index terms begins gradually to decrease from the point where articles have more than ten pages.

The question is what are the advantages or disadvantages of this for an indexing system? To answer this question we have to consider the users' intention. Finding relevant documents, in other than English or the own language is in most cases the worst alternative for an information seeker. Users prefer documents they are able to read. Documents in Chinese or Japanese publications can't help most people outside these countries. Because of this, documents other than the main language of the database (in most cases English) should have less weight in their findability. Giving less weight to the journals in other languages is one of the points that the planners should take into consideration when determining the indexing policy.

4.3.4 Date of indexing

Some events over the years, like developments in information technology, can influence indexing policies and the tendency toward assigning more index terms to documents. As the date of inclusion of documents into MEDLINE is a factor that mostly relates to the policy of indexing, reports of NLM in different fiscal years should support this idea.

The report of the fiscal year 1974 by NLM (NLM Fiscal Year 1975 p.21-22) shows that the investigation to find the feasibility of using a computer to publish Index Medicus was begun in April 1959. NLM wanted to see if the computerized version can serve as a basis for an efficient reference and bibliographic service. In 1960 the National Advisory Heart Council approved the transfer of \$500,000 to initiate the project. The National Library of Medicine selected the General Electric Company to begin design and development of the MEDLARS (Medical Literature Analysis and Retrieval System). It took three years and in March 1963 a Honeywell 800-200 computer system was delivered to the Library. After the time of testing, the system was finalized, the first issue of Index Medicus was produced from the MEDLARS system in January 1964.

Hogan, R. (1966) said that one of the objectives of MEDLARS was "to increase the average depth of indexing per article by a factor of five, i.e. 10 headings versus 2". „In pre-MEDLARS days the average number of index terms per article was 1.8". As NLM reported, in 1964 the average number of terms has reached about "6.7" per article (NLM Fiscal Year 1964 p. 23). The findings of the current work show that it was unchanged in 1965.

Besides the aim of increasing in-depth indexing, the indexing of documents per hour was also taken into consideration. The average of documents indexed was about 6.2 per hour in 1964 and decreased to 5.2 in 1965 (NLM Fiscal Year 1965 p.62). In 1966 articles were indexed at

the same rate as the year before (5.7 articles/hour) (NLM Fiscal Year 1966 p. 44). This rate decreased to 4.2 in 1967 (NLM Fiscal Year 1967-68 p. 44). Since 1969, a part of the indexing was performed under contracts and cooperative agreements with external agencies (NLM Fiscal Year 1969 p. 6). Because of this, the rate of indexing per hour can't be determined for the following years. The findings of the current work show that the average number of index terms assigned to documents has increased up to the year 1974.

In 1975 a new period of in-depth indexing of MEDLINE began. NLM decided to include English abstracts written by authors into the searchable MEDLINE file in 1974. It anticipated inputting 100,000 abstracts into MEDLINE each year up to FY 1977. In addition, NLM expected to begin receiving indexing in machine-readable form from some of the non-U.S. MEDLARS centres during FY 1976. This caused in-house indexers and those in other locations to change their methods by keying data directly into the database via online terminals. These decisions could have changed the previous indexing and data entry procedures. Quality control of indexing was also possible with necessary corrections being made by senior indexers (also on-line) prior to release of the citations into the database (NLM Fiscal year 1975, p.38).

Despite the inclusion of abstracts to MEDLINE, we see that the average number of index terms given to the articles has decreased between the years 1975 till 1981. It shows a new policy of NLM to substitute the free-texts searching within abstracts. This policy has changed from 1982 onwards, and the average number of index terms increased gradually. It shows that the intellectual subject indexing by human or Artificial Intelligence (AI) has its own advantages. Thus, the possibility of free-text searching doesn't reduce the need for intellectual indexing. In other words, searching through free-texts (like abstracts or full-texts) and index terms (descriptors) can't be replaced by each other.

As the findings of the current work showed, the inclusion of structured abstracts began in 1987. Harbourt, A. M.; Knecht, L. S. and Humphreys, B. L. (1995) characterized the role of structured abstracts in biomedical journals indexed in MEDLINE® between 1989 – 1991 “as an initial step in exploring their utility in enhancing bibliographic retrieval, “the number of structured abstracts in MEDLINE and the number of MEDLINE journals publishing structured abstracts increased substantially between 1989 and 1991”. They reported also “the average length of the structured abstract was greater than the average length of all abstracts in MEDLINE”.

In 1989, the difference between the average MeSH Headings of articles with structured abstracts became “9.0”, whereas for articles with unstructured abstracts it was “7.6” (without check tags). From 1996, the average number of index terms of articles with and without structured abstracts reached almost the same level. This means structured abstracts had an exceptional effect on the number of assigned MeSH headings in the initial years, and the role of them has reached a saturation point.

Other findings revealed that in year 2001 the average of their assigned headings increased unexceptionally and returned to normal in 2002. The findings indicate the year 2001 was an exception.

In the year 2001, there was a great effort to index past publications. Findings of the current work show in that year, articles were indexed with a delay of “0.82” years on average. This means the NLM concentrated more on past publications in 2001 in comparison with the years prior or following.

Table 14: Yearly delay in indexing between 1990 and 2005.

Indexing Year	#Documents	Total Sum of Delays	Delay in Indexing per Article on Average
1990	35,581	35	0.00
1991	36,330	34	0.00
1992	37,845	47	0.00
1993	40,458	41	0.00
1994	41,909	56	0.00
1995	43,821	56	0.00
1996	45,876	61	0.00
1997	41,521	984	0.02
1998	53,041	10,506	0.20
1999	53,050	12,289	0.23
2000	56,906	12,034	0.21
2001	62,966	51,367	0.82
2002	66,058	15,369	0.23
2003	65,236	10,367	0.16
2004	66,788	7,553	0.11
2005	62,179	7,597	0.12

In every year, besides the current publications, the past publications were indexed as well. This table shows the average interval between the date that articles were published and the date they were indexed. If we add the differences between the year of publication and the year of indexing (Entrez Date) of every document and then divide the sum (column 3) by the number of documents indexed in each year (column2), we will find the delay in indexing as part of a year (e.g. 0.82 * 12 ~10 months).

We see that in year 2001 large funding was given to indexing past publications. The report about “Collection Development and Management” in Fiscal Year 2001 (NLM Fiscal Year 2001, p. 9-10) supports the finding above.

4.3.5 Priority of journals for in-depth indexing

As explained previously, NLM categorizes the journals into three different groups and gives them a priority number one, two or three. On average, the first one is indexed by more and the last one indexed by fewer terms. Since in-depth indexing is related to the number of index terms assigned to articles, we can determine the priority of journals through the average number of index terms per page that have been given to journal articles.

The findings support this idea. Three different regions equal to the three priority numbers could be found for the journals indexed in MEDLINE. Thus, if we exclude the three first known journals in biomedicine, three groups of journals with a different priority will result. The average of their index-terms per page changes from 2.3 down to 1.5 and then down to 0.7. This means the priorities could vary the number of index terms per page for every group of journals. This variation shows that the average effect of the priority factor is 0.7 headings per page. Through reduction of priority, the average of index terms for each group of journals reduces 0.7 headings per page.

Assigning less or more index terms to journals helps to decrease or increase the findability of their articles. The policy of every indexing and retrieval system is to allow the users to receive the most appropriate and important documents. The Search Engines like Google or Yahoo do it by sorting the search results in the order of their appropriateness and importance.

4.3.6 Impact Factor

Impact Factor is known as a measure in evaluating journal quality through citing their articles by others. If it was always so, the correlation between the JIF and depth of indexing would show IF as an indicator for importance of journals. From this we could determine the level of journal priorities for their in-depth indexing. Many articles have discussed the role of IF on measuring the quality and importance of journals, citing some biases and trends as problems that lead to a high JIF. Self-citing (Gami, A. S. et. al., 2004), citing articles within the same journal (Tsay, Ming-Yueh 2006) and many other biases belong to these problems. In addition,

the nature of sciences differs from one to other, influencing the JIF. For instance, Milman, V. (2006) discusses the problem of IF for mathematics journals. He asserts: „This may, perhaps, be a very appropriate approach for, say, medical sciences or biology, where the influence of a publication is decided in the first year or so after publication and, after three or four years many results are already irrelevant. However, what does this mean for mathematics?“.

Dong, P.; Loh, M. and Mondry, A. (2005) introduced other factors that can distort the calculation of the impact factor:

- Coverage and language preference of the SCI database
- Procedures used to collect citations at the ISI
- Algorithm used to calculate the IF
- Citation distribution of journals
- Online availability of publications
- Citations to invalid articles
- Negative citations
- Preference of journal publishers for articles of a certain type
- Publication lag
- Citing behavior across subjects
- Possibility of exertion of influence from journal editors.

There are other works that support the idea of JIF as the journal quality measure from other point of view. For example, Saha, S.; Saint, S. and Christakis, D. (2003) investigated whether the quality of medical journals assessed by individuals are like that determined by IF. Three groups participated in assessing journal quality: Physicians, researchers and practitioners. They stated: “The correlation between impact factor and physicians’ ratings of journal quality was strong ($r^2= 0.82$, $P=0.001$). The correlation was higher for the research group ($r^2= 0.83$, $P= 0.001$) than for the practitioner group ($r^2=0.62$, $P= 0.01$)“. Since the quality of journals is a qualitative feature, assessing it is usually challenging and depends on individual tastes. This is why the IF is a debatable field.

The object of the current work is not to judge the advantages or disadvantages of JIF. We want to investigate the relation between the JIF and in-depth indexing. It will lead us to find whether journal priorities for in-depth indexing in MEDLINE are related to IF of journals or

not. The findings show that there is relationship between them, only if JIF is equal or over fifteen. As we learned, the number of such journals is rare in MEDLINE.

The results reveal that JIF is not a deciding factor for assigning the level of priority to the journals in MEDLINE. This relationship should be obtained by accident. Despite questioning the role of IF in determining the importance of journals, as showed above, it can't be rejected completely. The judgments are only able to show that the JIF alone can't be counted as an indicator for importance of journals and we need to consider other factors too. Some journals are used and cited more, and are known as sources that publish top scientific articles. Of course, there is an interaction between the readers' tendency to journals and JIF. If the readers find them important, they will cite them. This increases the JIF, and those that get a higher IF will be cited more in the future. But the initial motive is users' judgment and attitudes toward journals. Thus, some journals with higher or lower level of priority for indexing had accidentally the higher or lower IF. Otherwise the correlation between JIF and average of index terms assigned to them should be higher.

Since the JIF in 2004 is used for this investigation, one may argue that if the IF of other years were used, we would get very different results and we might see a higher correlation. Because JIF varies from year to year, Garfield, E. (1976) showed that there is a strong relationship between publications and citations. Journals that are cited more will be cited more with a stable constant in the following years. Based on his finding, Nourmohammadi, H. (2007) found that not only was the constant not stable, but it was increasing over the years as well. We can conclude that the calculation of the JIF is related to the number of citations, the positive correlation between them reveals that the ranks of the journals shouldn't change heavily over the course of time.

5 Conclusion

In this Chapter we will summarize what was learned and how it can be applied. The subject area of the current dissertation concerned three related topics. The first topic was to find out how MeSH has developed over time. The second one dealt with the usage of index terms in MEDLINE. The rest of the work covered the factors affecting the number of index terms per document of journals.

5.1 Development of thesaurus terms

The existence of three phases during the development of MeSH indicates an effort to optimize the growth rate of MeSH terms over the course of years. Actually, such changes in an indexing system are due to the inability to predict the future. If the indexing authorities were precisely aware of the future needs, these changes would not happen and the growth would follow a single rate from the beginning.

The results also showed that such systems should consider how dynamic a thesaurus will grow over time. It is now clear that the dynamic of a thesaurus like MeSH is the consequence of the average growth of one new term per ~250 new documents. It is remarkable that such a number of documents need to cover different topics in a thesaurus.

The analysis also revealed that, to construct a new thesaurus, one needs to know at the outset the least number of documents that should be available for indexing. Therefore, the three most important factors for construction and development of an instance thesaurus are:

1. the least number of documents needed for constructing a thesaurus
2. an assessment of the thesaurus development dynamic
3. an assessment of the future growth rate of the literature covered by a thesaurus. Without this assessment, determining growth rate of thesaurus terms is not possible.

In addition to the phenomena mentioned above, the breadth of the subject area covered by a thesaurus should be also taken into consideration. If it deals only with a very specific area such as “Sports Medicine”, the least number of documents needed for constructing the thesaurus will be fewer than estimated in this research (1,600 different documents). And the thesaurus development dynamic won’t be defined as the inclusion of one new term per 256 new documents. Consequently, the growth rate of the thesaurus terms will decrease as well.

5.2 Use distribution of thesaurus terms

Thesaurus terms carry different weights for indexing. Despite this fact, they can be grouped into the different classes. Distribution of MeSH terms in MEDLINE during four different time intervals has shown that they can be classified as highly, normally and rarely-used terms.

The number of highly-used terms is almost limited. They are very broad terms that can be assigned to a great number of publications. Consequently, a thesaurus lacking a classification system can bring the large classes of literature together by the means of such broad terms.

The majority of terms belong to the normal class because thesaurus terms need specificity. On the other hand, the rarely-used terms are actually in a test phase. They will be either shifted into one of the two classes above or will be omitted from the thesaurus. Therefore, the existence of such terms is not because of their high specificity.

Thus, the construction and development of a thesaurus without a classification system concerns other three main points:

1. existence of some broad terms to distinguish the literature into super and broad classes,
2. existence of a possible same level of specificity for the majority of terms,
3. existence of a temporary phase to test the usefulness of terms for indexing.

It is also remarkable that the highly-used terms tend to keep their importance. So the top terms of this class usually stay on the top for ever. This fact applies to normally-used terms in other form. The use of older terms decreased over the course of time while the use rate of newer terms increased gradually. Based on this fact, we can say that such terms have an average life span. The end line of these terms is when they reach the point of saturation. In this case, the term shouldn't be omitted. It should be broken into two or more terms.

5.3 Factors related to the number of index terms of articles

The number of index terms per document depends on different factors. Some of them relate to the nature of the text and some to the indexing system. The current work dealt with documents that were recognized as journal articles by NLM and treated the effects of six factors on the average number of their index terms: "length of articles", "articles with abstracts (and even the form of abstracts)", "language of articles", "date of indexing", "Journal Impact Factor (JIF)", and "priority of journals for in-depth indexing".

5.3.1 Length of articles

We learned that the average number of index terms per article is related to the number of pages of articles. This relationship is logarithmic and applies only when the number of pages of articles is twenty-one or less. The variety of types per token will gradually decrease when the number of pages increases. Since the number of index terms per article depends mostly on the number of their types, the gradual decrease of types variety per token produces a logarithmic relationship between the number of index terms and number of pages of articles. Thus, such a relationship should be expected for an automatic indexing system.

We also found that articles with abstracts receive two more terms than the others. The inclusion of abstracts into a database like MEDLINE can be treated from two different points of views:

1. these types of articles possess a level of importance for the indexing systems,
2. the articles with abstracts help the indexers to index them deeper.

5.3.2 Articles with abstract

The degree of journal importance for an indexing system leads the indexers to decide a level of in-depth indexing that is appropriate to its dedicated scientific value. An indexing system that finds a journal important tries to introduce more details and contents of its articles. Inclusion of abstracts of such journal articles in a database is one way to achieve this. Such articles will receive comparably more index terms than others.

Besides the argument above, the abstracts help the indexers to receive more contents from their corresponding texts in a shorter time. It leads also to assigning more index terms to documents.

5.3.3 Language of articles

One of the aims of databases like MEDLINE is bringing related world literature together. They index articles of important journals written in languages other than English. Comparing the average number of index terms per article between the English and German articles revealed that, in general, the level of findability of German articles (as the second language of MEDLINE) is considerably less than that of English articles. The difference between the

articles with ten pages is the highest (~%33) and disappears almost when the number of pages of articles reaches twenty and more. Several reasons are behind this bias:

1. The first one is the language bias. In an indexing system, tendencies are mostly toward the base language.
2. A number of non-English authors try to publish their significant works in known English journals. Because of this, most scientific journals in foreign languages don't have the acceptance of those in English.
3. Indexing non-English texts in English has its own difficulties. An indexer reads and scans a text in one language and has to index it in the other language. It also leads to assigning fewer index terms to them.

5.3.4 Date of indexing

The date of indexing documents is a factor that is related to the policy of indexing systems. The results yielded from the study of indexing MEDLINE during forty years showed that its in-depth indexing policy has changed three times during the periods “1965 – 1974”, “1975 – 1981”, and “1982 – 2005”. From the second period mentioned above, we can recognize a reduction of average number of MeSH terms per article. Contrary to it, the first and last periods show an increase. If we follow the events of the periods above, we will detect two remarkable facts:

1. The persistent increase of the average number of index terms during the first period is simultaneous with the mechanization of indexing by the NLM. At this time the Index Medicus was migrated to a computerized system. In addition, NLM offered the Index Medicus online. Thus, expanding computer technology has had a significant impact on NLM policy for indexing.
2. The reduction of the average number of index terms per article during the second period is simultaneous to the inclusion of abstracts in MEDLINE. We can be sure that NLM has treated the abstracts as a partial replacement for subject indexing, since they allowed users to conduct free-text searches. This facility apparently reduced the need for in-depth indexing. Every year NLM included more articles with abstracts in MEDLINE. This caused a gradual reduction of average number of index terms per article over several years. It is remarkable that the average number of index terms of articles without abstracts decreased in the same period as well. The other remarkable point of this period

is the reception of more index terms by articles with abstract in comparison with others. The two arguments mentioned above for the role of abstracts supply the reason for in-depth indexing. The existence of the third period reveals a change toward the attitude that inclusion of abstracts can replace partially the in depth indexing. During this period the structured abstracts were also included in MEDLINE.

5.3.5 Journal Impact Factor (JIF)

The study of the role of JIF on the depth of indexing reveals how the known evaluating systems such as ISI could influence the indexing of journals with high Impact Factor. The results showed that only some of the top journals had accidentally higher IFs.

5.3.6 Priority of journals for in-depth indexing

The indexing systems give some priorities to the journals for in-depth indexing. This prioritization can be recognized by the average number of index terms of their articles per page. The distribution of the average number of index terms of journals per page showed three regions of journals in MEDLINE. Three outstanding journals were on the top of these regions: “Nature”, “Science”, and “the Transplant Proc”.

These regions follow the priorities that NLM gives to journals for in-depth indexing. This information is for in-house use, but we can find the priorities of journals by the mean of the method mentioned above. The results express that the three outstanding journals and the three regions of journals receive, on average, respectively 3.3, 2.3, 1.5, and 0.7 terms per page.

As a consequence, journal importance for an indexing system determines the depth of indexing. All journals covered by it do not have the same level of importance. Thus, it should be counted as a bias toward different journals.

6 Theses

1. The development of the number of thesaurus terms is related in a characteristic way to the number of indexed documents. For such a system in medicine (MEDLINE) we can calculate roughly the function “ $T = 3,076.6 \ln(d) - 22,695 + 0.0039d$ ” (T = thesaurus terms, Ln = natural logarithm, and d = documents). That means that we need at first ~1,600 documents out of the planned scope to construct a preliminary thesaurus. Consequently, since 1950 the growth of Medical Subject Headings (MeSH) followed three phases with logarithmic functions that made clear that the NLM had to optimize repeatedly their indexing system. It is remarkable that this function has reached a steady-state of approximately 1 new term from 250 new documents.
2. The Distribution of a well constructed thesaurus without an additional classification (like BIOSIS, Chemical Abstracts, etc.) needs three classes of terms, the highly, the normally, and the rarely used. The last group is in a test phase, and most terms of this group shift to other groups over time. Only the terms in the first and second class were becoming persistent over the years. The first group is growing very fast, despite the attempts to retard this growth.
3. In the range between one and twenty one pages the number of index terms per article is related logarithmically to the number of its pages. Most probable, it is reaching a maximum of 10.3 terms per article (without check tags).
4. The inclusion of abstracts to an indexing system hypothetically reduces the need for in-depth indexing. The inclusion of abstracts to MEDLINE from 1974 to 1981 followed this belief. In general, these most probable important articles with abstracts received as an average two more terms compared to those without abstracts.
5. There is a clear difference between the findability of English and German papers. Articles written in English, with ten pages have in MEDLINE an average of 33% more index terms than those written in German. In articles with twenty or more pages (often Reviews), this difference disappears.
6. Distribution of index terms per journal page in MEDLINE has shown that the relationship between the depth of indexing and Journal Impact Factor in the range $JIF < 15$ is not verifiable.
7. Indexing systems can give different journals more or less weight in their findability. This can be proved roughly by the estimation of index terms per page. The distribution of MeSH terms per page has shown in a sample of ~1 million records that there are three regions with respectively 2.3, 1.5, and 0.7 terms per page. In the first group we found the journals: “Science”, “Nature” and “Transplant Proc”.

References

- Abt, Helmut A. (1992). Publication Practices in Various Sciences. *Scientometrics*. 24(3): 441-447.
- Abt, Helmut A. and Garfield, Eugene (2002). Is the Relationship between Numbers of References and Paper lengths the Same for all Sciences?. *Journal of the American Society for Information Science and Technology*. 53(13): 1106 - 1112.
- Anderson, J. D. (1997). NISO technical report 2: Guidelines for Indexes and Related Information Retrieval Devices. Bethesda, MD: NISO Press.
- Anderson, M. D. (1971). Cambridge Authors' and Printers' Guides: Book indexing. Cambridge, UK: At the University Press.
- Bechhofer, Sean and Goble, Carole A. (2001). Thesaurus Construction through Knowledge Representation. *Data & Knowledge Engineering*. 37(1): 25-45.
- Bibliographic Services Division (BSD), Fact Sheet [WWW Document]
<http://www.nlm.nih.gov/pubs/factsheets/bsd.html> 04-05-2007
- Blagden, John Frederick (1971). Management Information Retrieval: a New Indexing Language. London: Management Publications Ltd. for the British Institute of Management.
- Borko, Harold and Bernier, Charles Llewellyn (1978). Indexing Concepts and Methods. New York: Academic Press.
- Chen, Hsinchun (1994). Collaborative Systems: Solving the Vocabulary Problem. *IEEE Computer*. 27(5):58-66. Special Issue on Computer-Supported Cooper- Supported Cooperative Work.
- Chen, Hsinchun et al. (1996). A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 18(8): 771-782. Special Section on Digital Libraries: Representation and Retrieval.
- Chen, Hsinchun et al. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System. *Journal of the American Society for Information Science*. 48(1): 17-31.
- Cleveland, Donald B. and Cleveland, Ana D. (2001). Introduction to indexing and abstracting (3rd ed.). Englewood, CO: Libraries Unlimited, Inc.
- Cleverdon, Cyril W. et al. (1966). Factors Determining the Performance of Indexing Systems(Vol. 1): Design, Granfield. England: College of Aeronautics. Aslib Granfield Research Project.

- David, Claire et. al. (1995). Indexing as Problem Solving: a Cognitive Approach to Consistency. *Proceedings of the ASIS Annual Meeting*. Vol. 32: 49-55.
- De Jesus Adriano, Holanda et al. (2004). Thesaurus as a complex network. *Physica A*. 344(3-4): 530 – 536.
- Derr, Richard L. (1982). A Classification of Questions in Information Retrieval by Conceptual Presupposition. *Proceedings of the 45th ASIS Annual Meeting*. Vol. 19: 69-71.
- Dong, Peng; Loh, Marie and Mondry, Adrian (2005). The Impact Factor Revisited. *Biomedical Digital Libraries*. 2(7). [www Document]: <http://www.biomediglib.com/content/2/1/7>. 04.05.2007
- Dorbin, Tobun Ng (2000). A Concept Space Approach to Semantic Exchange. Ph.D. Dissertation, University of Arizona, the Graduate College of Business Administration.
- Egger, Matthias et al. (1997). Language Bias in Randomized controlled trials published in English and German. *The Lancet*. 350(9074): 326 – 329.
- Farrow, John F. (1991). A Cognitive Process Model of Document Indexing. *Journal of Documentation*. 47(2): 149-166.
- Fidel, Raya (1991). Searchers' Selection of Search Keys: I. The Selection Routine. *Journal of the American Society for Information Science*. 42(7): 490-500.
- Gami, Apoor S et al. (2004). Author Self-Citation in the Diabetes Literature. *Canadian Medicine Association Journal*. 170(13): 1925 – 1927.
- Garas, Gus J. (1968). Indexing from Abstracts of Documents. *Journal of Chemical Documentation*. 8(1): 20 – 22.
- Garfield, Eugene (1976). Is the Ratio Between Number of Citations and Publications Cited a True Constant?. *Essays of an Information Scientist*. Vol. 2: 419 –425.
- Giyeong, Kim (2006). Relationship between Index Term Specificity and Relevance Judgment. *Information Processing and Management: an International Journal*. 42(5): 1218 – 1229.
- Greenberg, Jane (2001). Automatic Query Expansion via Lexical–Semantic Relationships. *Journal of the American Society for Information Science and Technology*. 52(5): 402–415.
- Guardiola, Elena and Baños, J E (1993). Presence of Abstracts in Non-English Journals Indexed in MEDLINE (1981-1990). *Bulletin of the Medical Library association*. 81(3): 320-322.
- Guimarães, Carlos Alberto (2006). Structured Abstracts: Narrative Review. *Acta Cirúrgica Brasileira*. 21(4): 263 – 268.

- Harbourt, A. M.; Knecht, L. S. and Humphreys, B L (1995). Structured Abstracts in MEDLINE, 1989-1991. *Bulletin of the Medical Library Association*. 83(2): 190 – 195.
- Hartley, James (2003). Improving the Clarity of Journal Abstracts in Psychology: The Case for Structure. *Science Communication*. 24(3): 366 – 379.
- Heaps, H. S. (1978). Information Retrieval - Computational and Theoretical Aspects. Orlando, FL: Academic Press.
- Hogan, Rose (1966). An Evaluation of MEDLARS Output: Demand and Recurring Bibliographies. *Bulletin of the Medical Library Association*. 54(4): 321–324.
- IJzereef, Leonie; Kamps, Jaap and De Rijke, Maarten (2005). Biomedical retrieval: How can a thesaurus help?. *OTM Conferences*. Vol. 2: 1432-1448.
- Indexing Operation. [WWW Document]
http://www.nlm.nih.gov/mesh/indman/chapter_4.html 04.05.07
- Indexing Priority [WWW Document]
http://www.nlm.nih.gov/archive//20070220/databases/license/medlars_elements2.html#py 04.05.2007
- Janos, Jiri (1975). Results of an Experiment with Automatic Indexing Based on the Analysis of the Texts of Abstracts. *Information Processing and Management*. 11(3-4): 115-122.
- Jenuwine, Elizabeth S. and Floyd, Judith A. (2004). Comparison of Medical Subject Headings and Text-Word Searches in MEDLINE to Retrieve Studies on Sleep in Healthy Individuals. *Journal of Medicine Library Association*. 92(3): 349-354.
- Jones, Karen Spärck (2004). A statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*. 60 (5): 493-502.
- Journal Selection for MEDLINE®, National Institutes of Health, National Library of Medicine. [WWW Document] <http://www.nlm.nih.gov/pubs/factsheets/jsel.html> 04.05.2007
- Karbasi, Sohaila and Mohand Boughanem (2006). Effective Level of Term Frequency Impact on Large-Scale Retrieval Performance: By Top-Term Ranking Method. *Sixth Proceedings of the First International Conference on Scalable Information Systems*, May 29-June 1 2006, Hong Kong.
- Kortendick, O and Fischer, M. (1996). Comparative Classifications of Ethnographic Data: Constructing the Outline of Cultural Materials. CSAC Studies in Anthropology. Vol. 11. [WWW Document]
http://lucy.ukc.ac.uk/CSACSIA/Vol14/Papers/Kortendick/OCMPaper/OCMPaper_1.html 04.05.2007

- Krajnc, A. (1982). Motivacija za izobraževanje [in Slavonic]. Slovenia, Ljubljana: Delavska enotnost.
- Lancaster, F. Wilfrid (1986). Vocabulary Control for Information Retrieval(Second Edition). Washington: Information Resources Press.
- Lancaster, F. Wilfrid (1991). Indexing and Abstracting in Theory and Practice. Champaign, IL: University of Illinois. Graduate School of Library and Information Science.
- Loria, Alvar and Arroyo, Pedro (2005). Language and Country Preponderance Trends in MEDLINE and its Causes. *Journal of the Medical Library Association*. 93(3): 381–385.
- Losee, Robert M. (2007). Decisions in Thesaurus Construction and Use. *Information Processing and Management*. 43(4): 958-968.
- Lufkin, Richard C. (1968). Determination and Analysis of Some Parameters Affecting the Subject Indexing Process. Bachelor thesis, Electrical Engineering Department, Massachusetts Institute of Technology.
- MacClelland, R. M. A. And Mapleson, W. W. (1966). Construction and Usage of Classified Schedules and Generic Features in Coordinated Indexing. *ASLIB Proceedings*. Vol. 18: 290-299.
- Mai, Jens-Erik (2001). Semiotics and Indexing: an Analysis of the Subject Indexing Process. *Journal of Documentation*. 57(5): 591-622.
- Mai, Jens-Erik (2005). Analysis in Indexing: Document and Domain Centered Approaches. *Information Processing and Management: An International Journal*. 41(3): 599-611.
- Maron, M. E. (1979). Depth of indexing. *Journal of the American Society of Information Science*. 30(4): 224-228.
- McDonald, Steve (2002). Improving Access to the International Coverage of Reports of Controlled Trials in Electronic Databases: a Search of the Australian Medical Index. *Health Information and Libraries Journal*. 19(1): 14–20.
- Medical Subject Headings® - Overview [WWW Document]
<http://www.nlm.nih.gov/mesh/overview.html> 04.05.2007
- MEDLINE®/PubMed® Data Element (Field) Descriptions, National Institutes of Health, National Library of Medicine. [WWW Document]
<http://www.nlm.nih.gov/bsd/mms/medlineelements.html> . 04.05.2007
- Milman, Vitali (2006). Impact Factor and How It Relates to Quality of Journals. *Notices of the American Mathematical Society*. 53(3): 351-352.

- Mueller, Paul S. et al. (2006). The Association between Impact Factors and Language of General Internal Medicine Journals. *Swiss Medical Weekly*. 136(27-28): 441-443.
- Mulvany, Nancy C. (1994). *Indexing books*. Chicago, IL: The University of Chicago Press.
- Muñoz Rodríguez, J. V. and Gil Leiva, I. (1997). Análisis de los descriptores de diferentes áreas del conocimiento indizadas en bases de datos del CSIC. Aplicación de la indización automática [in Spanish]. *Revista Española de Documentación Científica*. 20(2): 150-160.
- Narayana, S. Murali et. al. (2004). Impact of FUTON and NAA Bias on Visibility of Research. *Mayo Clinic Proceedings*. Vol. 79: 1001–1006.
- NLM Fiscal Year 1964 [WWW Document]
[Http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1964.pdf](http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1964.pdf)
 04.05.2007
- NLM Fiscal Year 1965 [WWW Document]
<http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1965.pdf>
 04.05.2007
- NLM Fiscal Year 1966 [WWW Document]
<http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1966.pdf>
 04.05.2007
- NLM Fiscal Year 1967-68 [WWW Document]
<http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1967-68.pdf>
 04.05.2007
- NLM Fiscal Year 1969 [WWW Document]
<http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1969.pdf>
 04.05.2007
- NLM Fiscal Year 1975 [WWW Document]
<http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1975.pdf>
 04.05.2007
- NLM Fiscal Year 2001 [WWW Document]
<http://www.nlm.nih.gov/ocpl/anreports/fy2001.pdf> 04.05.2007
- Nourmohammadi, Hamzehali (2007). Über die sZientometrische Bedeutung des Impact-Faktors [in German]. Ph.D. Dissertation, Humboldt - Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft.
- Overbyte [WWW Document] http://www.overbyte.be/frame_index.html 04.05.2007

PubMed & MEDLINE [WWW Document]

<http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html#Introduction>
04.05.2007

Raghavan, Vijay V. et. al. (2004). Information Retrieval. *In the Practical Handbook of Internet Computing*, (Munindar P. Singh, ed.), Part-2, Chapter 12, Chapman and Hall/CRC Press.

Razdevšek-Pučko, C. (1999). Motivacija in učenje [in Slovenian]. Slovenia, Ljubljana: Teze predavanj pri predmetu Pedagoška psihologija, Pedagoška fakulteta.

Saha, Somnath; Saint, Sanjay and Christakis, Dimitri A. (2003). Impact Factor: a Valid Measure of Journal Quality?. *Journal of the Medical Library Association*. 91(1): 42-46.

Salton, Gerard (1975). A Theory of Indexing. (volume 18 of Regional Conference): Series in Applied Mathematics. Philadelphia, PA.: Society for Industrial and Applied Mathematics.

Salton, Gerard and Yu, Clement T. (1973). On the Construction of Effective Vocabularies for Information Retrieval. *SIGPLAN/SIGIR Symposium on Programming Languages and Information Retrieval*, Gaithersburg, MD.: 48-60.

Schulman, E. et. al. (1997). Trends in Astronomical Publication Between 1975 and 1996. *Publications of the Astronomical Society of the Pacific*. Vol. 109: 1278-1284.
http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1997PASP..109.1278S&data_type=PDF_HIGH&whole_paper=YES&type=PRINTER&filetype=.pdf 04.05.2007

Schwarz, I. and Umstätter, Walther (1999). Die vernachlässigten Aspekte des Thesaurus: dokumentarische, pragmatische, semantische und syntaktische Einblicke [in German]. *Information - Wissenschaft und Praxis*. 50 (4): 197-203.

Singhal, Amit et. al. (1996). Document Length Normalization. *Information Processing & Management*. 32(5): 619-633.

Soergel, Dagobert (1994). Indexing and Retrieval Performance: The logical evidence. *Journal of the American Society of Information Scientists*. 45(8): 589-599.

Svenonius, Elaine (1971). The Effect of Indexing Specificity on Retrieval Performance. Ph.D. dissertation. University of Chicago.

Tavakolizadeh-Ravari, Mohammad (2002). Production of Reference Resources by the Means of Reverse Engineering Techniques (in Persian). *Proceedings of 2nd Conference of the Role of Information Science on Cultural Development: Book and Information Technology*. Tehran: Khaneh Ketab.

- Taylor, Arlene G. (1992). *Introduction to Cataloging and Classification*. (8th ed.). Englewood, Colorado: Libraries Unlimited.
- Thellefsen, Martin (2004). Concepts and Terminology Reflected from a LIS Perspective: How Do We Reflect Meanings of Concepts?. *Proceedings of the 12th Nordic Conference for Information and Documentation*. 68-75.
- Tsay, Ming-Yueh (2006). Journal Self-Citation Study for Semiconductor Literature: Synchronous and Diachronous Approach. *Information Processing and Management: an International Journal*. 42(6): 1567-1577.
- Tsay, Ming-yueh and Yang, Yen-hsu (2005). Bibliometric Analysis of the Literature of Randomized Controlled Trials. *Journal of the Medical Library Association*. 93(4): 450–458.
- Umstätter, Walther (1986). Informetrische Hilfen durch das 'intelligente Terminal' [in German]. *Deutscher Dokumentartag*. 556-564.
- Van Rijsbergen, C. J (1979). *Information Retrieval* (2nd Edition). London: Butterworths.
- Wall, E. (1964). Further Implications of the Distribution of Index Term Usage. *Proceedings of the American Documentation Institute*. Vol. 1: 457-466.
- Wang, Yingxu and Gafurov, Davrondjon (2003). The Cognitive Process of Comprehension. *Second IEEE International Conference on Cognitive Informatics (ICCI'03)*. 93–97.
- Wellisch, Hans H. (1991). *Indexing from A to Z*. New York: The H. W. Wilson Company.
- Why Citations to Older Articles May Display Before More Recent Ones in PubMed (2002). *NLM Technical Bulletin*. [WWW Document]
http://www.nlm.nih.gov/pubs/techbull/ma02/ma02_display_order.html 04.05.2007
- Wurm, Bengt R. (1964). The Relation between Number of Documents and Number of Terms and their Discriminatory Power in Information Retrieval for U.S. Pharmaceutical Patents. *Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices ICIREPAT*. 349-360.
- Zazo, Angel F. et al. (2005). Reformulation of Queries Using Similarity Thesauri. *Information Processing and Management*. 41(5): 1163–1173.
- Zipf, George K. (1949). *Human Behavior and the Principle of Least-Effort*. Cambridge MA: Addison-Wesley.

Acknowledgment

The completion of this dissertation was made possible through the support and cooperation of many individuals. Special thanks go to my supervisor, Prof. Dr. Walther Umstätter, who provided thoughtful guidance and encouragement through what seemed to be a never-ending process. Thanks also to Prof. Dr. Michael Seadle, Dr. Frank Havemann, Michael Heinz, Mathias Schulz and the other members of the Institute of the Library and Information Sciences of Humboldt University of Berlin, for helping to me to understand the significance of this research.

My thanks go to Dr. Bob Kosovsky, of The New York Public Library and Michael Matrescu, from the University of Michigan, who were my English editors.

Thanks to the “University of Yazd” and the “Iranian Ministry of Science, Research & Technology” for giving me the grant to study in Germany.

Finally, thanks to all my teachers from the beginning to now. Special thanks to the teacher of my first class Mrs. Zahabi (Afsari), who taught me how to learn.

List of Figures

Figure 1: Entrez PubMed Homepage.	27
Figure 2: An example of MEDLINE Format.	28
Figure 3: Cumulative growth of the Medical Subject Headings (MeSH). The x-axis is scaled logarithmically.	44
Figure 4: Relationship between the phase number of MeSH growth and the medians of terms.	46
Figure 5: Relationship between the median of each phase and the last term of the phase.	47
Figure 6: Growth of MEDLINE versus growth of MeSH.	48
Figure 7: Absolute growth of the Medical Subject Headings (MeSH). The x and y axes are scaled logarithmically.	50
Figure 8: Term production of the last 1,500 documents of current MEDLINE.	52
Figure 9: Comparison of MeSH development with and without linear dynamic.	53
Figure 10: Comparison of development of MeSH based on the equation „ $te = 3,076.6 \ln(d) - 22,695 + 0.0039d$ “ with the actual ones.	53
Figure 11: Distribution of MeSH headings in MEDLINE during the years 1965 – 1970. The y-axis is scaled logarithmically.	54
Figure 12: Distribution of MeSH headings in MEDLINE during the years 1965 – 1980. The y-axis is scaled logarithmically.	55
Figure 13: Distribution of MeSH headings in MEDLINE during the years 1965 – 2000. The y-axis is scaled logarithmically.	55
Figure 14: Distribution of MeSH headings in MEDLINE during the years 1965 – 2006. The y-axis is scaled logarithmically.	56
Figure 15: Comparison of highly used headings between the intervals „1965 – 1970“ and „1965 – 2006“. x and y axes are scaled logarithmically.	57
Figure 16: An illustration of how the curve of normally frequented headings is moving to the gentle slope. Y axis is scaled logarithmically.	58
Figure 17: Relationship between the tokens and the types of articles.	63
Figure 18 Relationship between the number of pages of articles and their tokens.	64
Figure 19: Relationship between the lengths of articles scaled by tokens and their average MeSH headings. The x-axis scaled logarithmically.	65
Figure 20: Average number of MeSH headings assigned to articles with and without abstracts. The x-axis is scaled logarithmically.	66

Figure 21: Average of MeSH headings assigned to the journal articles in MEDLINE. The x and y axes are scaled logarithmically.	67
Figure 22: Comparison of the average of MeSH headings assigned to the journal articles with and without structured abstracts.	68
Figure 23: Comparison of the average number of MeSH headings assigned to the English and German journal articles in MEDLINE. The x-axis is scaled logarithmically.	69
Figure 24: Comparison of the average of MeSH headings assigned to the English and German journal articles per page in MEDLINE. Both of the x and y axes are scaled logarithmically.	70
Figure 25: Comparison of the average number of MeSH headings assigned to the German journal articles with and without abstracts in MEDLINE. The x-axis is scaled logarithmically.	71
Figure 26: Length of journal articles indexed in MEDLINE during the years 1965 – 2005. ..	72
Figure 27: Average of MeSH headings of the journal articles indexed in MEDLINE during the years 1965 – 2005.	73
Figure 28: Comparison of the average number of MeSH headings assigned to journal articles with and without abstracts in MEDLINE during the years 1974 – 2005.	74
Figure 29: Comparison of the role of structured and unstructured abstracts on the average number of MeSH headings assigned to the journal articles in MEDLINE during the years 1988 – 2005.	75
Figure 30: Journal priorities for in-depth indexing. 454 journals were ranked in order of the average number of MeSH headings per page.	76
Figure 31: Journal Impact Factor (JIF) ≥ 8 and average number of MeSH Headings per Journal.	77
Figure 32: Journal Impact Factor (JIF) < 8 and average number of MeSH Headings per Journal.	77
Figure 33: Relationship between JIF ≥ 15 and average of MeSH Headings assigned to the journal articles per page.	78
Figure 34: Relationship between JIF < 15 and average of MeSH headings assigned to the journal articles per page.	78
Figure 35: Relationship between the number of words of abstracts and number of significant words. This figure is a cutting from Garas, G. J. (1968).	98

List of Tables

Table 1: Uppercase words, their existences within an abstract indicate that they are structured.	34
Table 2: List of MeSH check tags.....	35
Table 3: A sample of the database based on the number of pages. This sample represents only two parts from the five: Not-abstracted and abstracted parts.	42
Table 4: A sample of database that based on the journal titles.	42
Table 5: A brief summary of the absolute growth of the MeSH. The two first columns represent the growth through the first 36,000 documents, and the third and fourth, the accession of the last headings.	51
Table 6: A brief summary of the distribution of MeSH headings in MEDLINE drawn from the results in Figures 11, 12, 13, and 14.	57
Table 7: Ranks of terms compared to the number of their usage based on HRU in four different years. HRU calculated by $HRU = \ln(2) / \exp$	60
Table 8: Ranks of terms versus the numbers of their usage based on HRU in four different years. HRU calculated by $HRU = \text{Total number of terms} / 5$	61
Table 9: The first twenty top journals of MEDLINE that are covered by ISI.	79
Table 10: A hundred of highly frequented MeSH headings in 1980.	85
Table 11: A hundred of the rarely frequented MeSH headings in 1980.	87
Table 12: Average review time per page, versus document length, for experienced indexers. This table is derived from Lufkin's work in 1968 (p. 36).	91
Table 13: Reduction of the types per page rate in relation to larger articles.....	93
Table 14: Yearly delay in indexing between 1990 and 2005.....	106

List of Equations

$y = 1947.8x - 1543.6$ and $R^2 = 0.9999$ (i.)	45
$y = [394.62 \text{ Ln}(d_i) / \sum 394.62 \text{ Ln}(d_i)] \times 100$ (ii.)	45
$y = [2371.2 \text{ Ln}(d_i) / \sum 2371.2 \text{ Ln}(d_i)] \times 100$ (iii.)	45
$y = [4290.2 \text{ Ln}(d_i) / \sum 4290.2 \text{ Ln}(d_i)] \times 100$ (iv.)	45
$y = 2371.2 \text{ Ln}(d_i) / 394.62 \text{ Ln}(d_i) = 2371.2 / 394.62 = 6.01$ (v.)	45
$y = 4290.2 \text{ Ln}(d_i) / 2371.2 \text{ Ln}(d_i) = 4290.2 / 2371.2 = 1.81$ (vi.)	46
$y = 859.97x^{2.851}$ and $R^2 = 1$ (vii.)	46
$y = 1.1501x + 654.25$ and $R^2 = 1$ (viii.)	47
$\text{HTR} = \text{Ln}(2) / \exp.$ (ix.)	48
$\text{HTR} = 1567.8x - 1292.8$; $R^2 = 0.9963$ (x.)	48
$\text{Exp} = \text{Ln}(2) / \text{HTR}$ (xi.)	48
$y = 0.0039x + 19,506$; $R^2 = 0.9901$ (xii.)	52
$t_e = 3,076.6 \text{ Ln}(d) - 22,695 + 0.0039d$ (xiii.)	53
$\text{HRU} = \text{Ln}(2) / \exp$ (xiv.)	59
$\text{HRU} = \text{Ln}(2) / \exp \cong (1/5) \times \text{total number of terms}$ (xv.)	60
$\text{HRU} = \text{total number of terms} / 5$ (xvi.)	61
$S_n = T_1^{\text{tot}} [1 - (1 - (T_1 / T_1^{\text{tot}}))^n] + T_2^{\text{tot}} [1 - (1 - (T_2 / T_2^{\text{tot}}))^n] + k_n$ (xvii.)	80

Curriculum vitae (Lebenslauf)

Professional Goal: To continually engage in a process of lifetime learning and development so that I can think and act to the best of my abilities, drawing from the resources of my knowledge and experience, in order to ameliorate human knowledge.

I. Education

- 1996 **Master of library and information sciences**; major: Information Science, Tarbiat –Modarres University, Tehran, IRAN.
- 1992 **Bachelor Degree**; major: German Language and Literature, Beheshti University, Tehran, IRAN.
- 1987 **High School Diploma**; major: Experimental Sciences, Shariati High School, Ravar, IRAN.

II. Professional Experiences

- 1996-present **Librarian**, Library and Information Sciences Dept. Yazd university, Yazd, IRAN.
- 1996-2003 **Lecturer**, Library and Information Sciences Dept. Yazd university, Yazd, IRAN.
- 1996-2000 **Library Assistant**, Central Library of Yazd university, Yazd, IRAN.
- 2001-2003 **Director**, Library and Information Sciences Dept. Yazd University, Yazd, IRAN.

Eidstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass die vorliegende Dissertation von mir selbst und ohne unzulässige Hilfe Dritter verfasst wurde, auch in Teilen keine Kopie anderer darstellt und die benutzten Hilfsmittel sowie die Literatur vollständig angegeben sind.

Berlin, den 01.06.2007

Mohammad Tavakolizadeh-Ravari