Use of natural language for information retrieval from Arabic online catalogue

PUBLISHER

Loughborough University of Technology

REPOSITORY RECORD

Al Tayyar, Musaid S.A.. 2021. “Use of Natural Language for Information Retrieval from Arabic Online Catalogue”. Loughborough University. https://doi.org/10.26174/thesis.lboro.14578959.v1.

USE OF NATURAL LANGUAGE FOR INFORMATION RETRIEVAL FROM ARABIC ONLINE CATALOGUE

by

MUSAID SALEH A. AL TAYYAR

A Master`s Dissertation, submitted in partial
fulfilment of the requirements of the award of the
Master of Arts degree of the
Loughborough University of Technology

September 1994

Supervisor: Inese A. Smith, B.A., M. A.
Department of Information and Library Studies

This dissertation is dedicated to my dear parents,
to my wife and my children for without their support is not have been
possible

ABSTRACT

In order to give a context for examining the subject context of book titles for retrieval from Arabic bibliographic databases this study provides a brief description of online catalogues in Saudi Arabia particularly, in Riyadh City. Because they are the most commonly used software packages in Saudi Arabia, A number of difficulties of Arabistion of DOBIS/LIBIS and MINISIS are also discussed. Natural language searching is briefly covered as is the relationship between Arabic language structure and its effect in information retrieval. The study examines and evaluates the efficiency of retrieval using words in Arabic book titles by comparing those words to subject headings which were given to them by King Fahad National Library (KFNL). It is concluded that Arabic book titles appear to provide poor representation of their subject content, with only 42.1 % of words being significant for subject retrieval. It is felt that this proportion is not large enough to enable searchers to depend upon Arabic book titles as sources of subject access points.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervior, Mrs. I. A. Smith, for her assistance and guidance in doing this dissertation.

I am most grateful to the Imam Mohammad Ibun Saud Islamic University for providing me the scholarship to pursue my studies at Loughborough University of Technology.

I would also like to extend my gratitude and appreciation to those people who cooperation and help in this dissertation.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

CHAPTER 2 LITERATURE REVIEW

CHAPTER 3 ARABIC LANGUAGE STRUCTURE

CHAPTER 6 SUMMARY AND CONCLUSION

List of Tables

List of Figures

ABBREVIATION

ALDOC	Arab League Documentation and Information Centre
ALECSO	Arab League Educational, Cultural and Scientific Organistion
AOS/VS	Advanced Operating System / Virtual Storage
ASCII	American Standard Code for Information Interchange
ASMO 449	Arab Standard Metrological Organisation
DOBIS/LIBIS	Dortmunder Bibliothekssystem Leuvens Integrall Bibliotheek System
IDRC	International Development Research Centre
IPA	Institute of Public Administration
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
KFCRIS	King Faisal Centre for Research and Islamic Studies
KFNL	King Fahad National Library
KFUPM	King Fahad University of Petroleum & Minerals
KSU	King Saud University
KWAC	Key Word And Context
KWIC	Key Word In Context
LAPM	Library Automation Project Management
MINISIS	Mini Computer Version of Integrated Set of Information System
OPAC	Online Public Access Catalogue
STAIRS	Storage and Information Retrieval System

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

In recent years a great deal of attention in Arabic literature relating to searching databases and information retrieval has been turned to information linguistics, a term used to highlight the strong relationship between language and information retrieval. As is well known, language is the major vehicle for the transfer of information. As in the foreword to *Linguistics and Information Science* Locke said: "Most information is expressed in language and most language conveys information"(1). Since the computer revolution in the 1960s, language has been the major problem facing those who deal with computers and information retrieval. Such problems with information retrieval include syntax, synonyms, derivation etc. In Arabic language, there are a number of problems facing those who are involved in information retrieval. Therefore, this study attempts to focus on the effectiveness of using natural language for information retrieval in Arabic particuarly as regards subject information in book titles.

1.2 Purpose of the Study

This study aims:

1. to try to identify the language problems facing Arabic information retrieval;
2. to examine Arabic book titles to determine whether they are significant; or have significant subject content;

3. to identify the effectiveness of using Arabic language in information retrieval.

1.3 Methodology

This study is divided into two parts: theoretical and practical. In the theoretical part there will be some discussion of Arabic language structure, while the practical part will examine the effectiveness of using natural language for information retrieval in Arabic. There will also be an examination of Arabic book titles as a subject access point. To achieve this a sample of titles has been taken from King Fahd National Library (KFNL). Whilst choosing the sample a number of factors were kept in mind:

1. All titles should have been written in Arabic language.
2. The sample should cover books from most Arab countries.
3. The sample should cover the widest possible area such as law, religion, economic, history, medicine etc.
4 The sample should cover both new and old books (heritage books).

1.4 Arabic Online Public Access Catalogues

Arabic online public access catalogues have almost the same features as the non-Arabic OPACs. According to Al Dosary and Ekrish(2) there are eight libraries and information centres using OPACs in Saudi Arabia. Table 1.1 shows the locations of the OPAC and the systems which are used.
Considering the inherent limitations of this study it was decided to choose the OPAC system that has been adopted in Riaydh as a model for the study. This is because this systems is fairly typical of the systems that exist throughout Saudi Arabia so it makes a good subject for this study.

Table 1.1: OPACs in Saudi Arabia

INSTITUTION	LOCATION	SYSTEM
1-Arab Security Studies and Training Centre (ASSTC)	Riyadh	Amestral
2-Institute of Public Administration (IPA)	Riyadh	DOBIS/LIBIS
3-King Abdulaziz City for Science and Technology (KACST)	Riyadh	In-House
4-King Fahd University of Petroleum and Minerals Library (KFUPM)	Dammam	DOBIS/LIBIS
5-King Faisal University Library (KFU)	Dammam	MINISIS
6-King Saud University Library (KSU)	Riyadh	DOBIS/LIBIS
7-National Information Centre, Ministry of Finance and National Economy (NIC)	Riyadh	VTLS
8-Saudi Arabian Standards Organisation (SASO)	Riyadh	In-House

From reasous that are not known, Al Dosary and Ekrish made no mention of OPACs in the Imam university Library, King Fahd National Library and King Faisal Centre for Research and Islamic Studies. However, they willbe described briefly in this chapter.

1.4.1 King Saud University Library OPAC

The initial planning for the use of computers in the library started in 1982. According to Ashoor (3), the first initiative of Arbisation of DOBIS/LIBIS was by King Suad University In 1984. The system runs on a "xcom2" computer. By 1985 the Arabic materials were entered into the system(4). The number of Arabic books are 94395 records, while the number of English books are 230000 records(5).

The OPAC at KSU provides users with the following search options:

1. Authors, editors, etc (names)
2. Titles
3. Subjects
4. Publishers
5. Classification
6. ISBN/ISSN

Author search

Searching authors (names) records should be conducted as follows:

- type the family name, then a comma (,)
- press the space bar on the keyboard once
- type the first name or initial
- press the ENTER key

A list of names (up to 14) will be displayed. Users can select the required author by entering the corresponding number from the left column (1 to 14). If the author has more than one book, a short list, showing titles and year of publication, is displayed; and if there is only one book, then full bibliographic data will be given.

Other files (e.g. title, subject) can be searched for in the same way. Guidance is provided to the user at all stages of the retrieval process. At the bottom of the screen a list indicates the keys which need to be pressed for a given purpose:

f	to go forward to the next screen
b	to go back to the previous screen
t	to search for another item in the same catalogue
i	to search another catalogue
e	to end the searching

Title search

The keyword index provides the ability to search for any word or phrase in the title. Users need to know only one word (or more) in the wanted title. Then they can retrieve it and those similar to it.

Truncation and Boolean operators are available in the system but not for public use, because of the complicated way that these searches are structured in the system. Printed instructions are provided to the user about how to search the KSU Library DOBIS/LIBIS online catalogue.

1.4.2 Imam Mohammad Ibn Saud Islamic University OPAC

Planning for library automation at the Imam University Library began in 1984(6). The online public access catalogue is part of a total system which includes circulation, acquisition etc. The system is running on an ECLIPSE MV.4000 DATA GENERAL database. The software used is AOS/VS (Advanced Operating System/Virtual Storage)(7). In-house software is used for library functions. There are 66,000 Arabic records in the catalogue; statistics are not available for foreign books and periodical titles (8). Imam University

OPAC allows users to conduct searches for author, title, subject and classification number.

The system prompts the user for each search by displaying a line number for each command. The following are the commands:

- To search for the author (name), type line (1)
- To search for the title enter line (2)
- To search for the subject enter line (3)
- To search class number, type line (4)

Author Search

The surname (family name) must be entered first, followed by the first name. Author names cannot be retrieved without this inverted order. ال التعريف *al altariaf* [the definite article] which is commonly used in Arabic family names should be omitted.

There is guidance provided to the user at the lower part of the screen:

(م) to go forward to the next screen
(ع) to go back to the previous screen
(ن) to end the searching in the author file
(ج) to go back to the main menu

Title Search

The title should be entered exactly as it appears in the title page. If not, users will not retrieve what they want, because no permuted index for title is provided. Subject headings also should be entered exactly as they appear in the printed list of Arabic subject headings(9). In addition to this there is no provision for Boolean operators. Therefore, it could be said that searching in the Imam University OPAC is poor.

1.4.3 Institute of Public Administration OPAC

The name of the system is Ibn Al Nadeem. It was given this name to commemorate the first person who done the first bibliography for Arabic books. The name of the book (10). According to Ashoor "the third institute that invested tremendous resources in the Arbisation of DOBIS/LIBIS was the Institute of Public Administration (IPA) in 1986"(11). Al Srayi (12) noted that by 1989 the library materials had been entered into the system. The records currently in the system are 188,468 records (13).

The user can access bibliographic records by choosing one of the following commands:

- To search title, use number (1)
- To search names, type number (2)
- To search subject, type number (3)
- Type number (4) for searching classification number.

When one of the above commands has been chosen the system will display the statement "Enter search term" on the screen, which means enter the author's name, subject, title or class number. The system offers keyword searching for the title. In the author search, the surname must be entered first, followed by a common then the first name. Instructions are provided at the bottom of the screen.

1.4.4 King Fahad National Library OPAC

The software which is used in the library is the MINISIS package, which was Arabised by the Arab League Documentation and Information Centre (ALDOC) (14). The library started using this system in 1989. The system runs on a HP3000 computer. KFNL is still not officially open for the public, but users and searchers can have access to library materials.

There are 138,772 records stored in the system (15).

Searching the OPAC

Users are first required to identify themselves to the system by entering a password which is on each terminal at the top right-hand corner of the keyboard. To enter the system one types: Hello KO28203,Q.KFNL(16). When the Enter key is pressed the system will ask the user to enter the name of the database. There are three databases in KFNL:

(library)	Library catalogue
(Shaffull)	Authority file for subjects
(Pafull)	Authority file for names

Shaffull and Pafull are for internal staff use only, whilst the (Library) database is for both staff and other users. To access the bibliographic records the user must enter F8. The commands used in the KFNL OPAC are:

F1 to exist from the system
F2 to display items
F4 to display items in card catalogue format
F5 to search title
F6 to search subject
F7 to search author
ديوي to search class number
ISN to search the book number

Each item in the KFNL database has an identification number which can be used to retrieve a single item quickly. To search by ISN number the user should type the command ISN and then the complete number. The KFNL OPAC is very flexible because it provides keyword searching in most fields, i.e., individual words within names, titles, subject headings and other parts of the bibliographic records may be retrieved by themselves. In addition to this, the system provides Boolean operators which are used for online retrieval (AND, OR, NOT). Truncation is used in the system; the truncation symbol is @. It should only be used at the beginning and the end of the word, not in the middle of a word.

1.4.5 King Faisal Centre for Research and Islamic Studies

In 1985 the KFCRIS started using the MINISIS package for their automation catalogue (17). The centre is the supplier of MINISIS in the Gulf area, with 15 databases stored in the system. There are 125,090 records (18).

Information in the centre is provided by use of the OPAC or by consulting the information specialist, the latter being the most commonly used method. If the users want the information specialist to carry out a search on their behalf then they are requested to fill in a search request form. This form is divided into five parts:

Part one: general information about user (name, address,ect.)
Part two: title of the search (subject of the search)
Part three: current awareness
Part four : type of materials (books, journals etc.)
Part five: for official use

When the form has been completed it is returned to the information specialist who conducts the search on behalf of the user. Sometimes the user needs to be interviewed in order to clearly identify the main points which need to be searched. Then output is given to the user which includes all information which is available in the centre about the requested subject. In the light of this output the user identifies which items are suitable search by marking each item. Then the list is returned to library staff responsible for locating and collecting requested material.

1.5 CONCULSION

This chapter has described OPACs system in Saudi Arabia particular in Riyadh city. Next chapter will be about litrature review related to this study.

REFERENCES

1. Sparck Jones, Karen and Martin Kay. *Linguistics and information science.* 1973, p.xi.

2. Al Dosary, Fahad and Abdurrahaman H. Ekrish. The state of autmation in selected libraries and information centres in Saudi Arabia. *Libri,* 1991, 41 (2), 113.

3. Ashoor, Mohammad Saleh . Arabistion of automated library system in the Arab world: need for compatibility and standardistion. *Libri,* 1989, 39 (4), 296.

4. Interview with computer specialist in KSU, Riyadh, May 1994.

5. Al Burady to M. Al Tayyar, 4 July 1994.

6. Al Zeer, Mohammad, H. *alhasib alaalei fi maktbat jamiat alimam Mohammad Ibn Saud al islamiah* [computers in Imam Mohammad Islamic University Libraries]. Paper presented at the Fourth Symposim on the State and Future of the Libraries in Arab World, 1987, pp.17-18.

7. *Ibid.*

8. Ref.5.

9. King Saud University. *Arabic subject headings list.*[n.d.].

10. Al Srayi, Srayi. Ibn Al Nadeem in the IPA Libraries. In. *Proceedings of Symposium on Using Arabic language in Information Technology*. 1992, p.316.

11. Ashoor,ref.3, p.298.

12. Al Srayi,ref.,10, p.316.

13. Ref.5.

14. Ashoor,ref.3, p.299.

15. Ref.5.

16. Frsony, Fuad. Searching in database. Report, [n.d.], p.1.

17. Al Dosary and Ekrish, ref.2, 113.

18. Ref.5.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will deal with some previous research that has been carried out in this area of the use of natural language in information retrieval. However, it will particularly emphasise those studies that are concerned with Arabisation because of a link between the use of natural language searching and Arabisation. It will also be concerned with the studies which cover the use of Arabic titles of documents as the source of access points.

2.2 Natural language searching

Regarding English literature, there are a number of studies which have discussed natural language searching or processing for information retrieval. Warner(1), who reviewed recent developments in computerised natural language processing covered research published between 1982 and 1986. However, as early as 1979 Lancaster(2) had already discussed natural language information retrieval and believed that there would be increased emphasis on topic in the future. He justified this belief by a number of reasons: "the continued growth in the availability of machine-readable data bases many of which will be in natural language; the continued expansion of online systems, which is likely, eventually, to put terminal in the offices and homes of scientists and other professionals, and a natural language mode of searching seems imperative in this type of application; new developments in computer storage devices will make the storage of very large text files increasingly feasible" (3). In fact, a number of later researchers do agree with Lancaster, for example Doszkocs(4) and Al Atram(5). "Cleverdon has also asserted, on a number of occasions, that performance with natural language can never be lower than that with controlled language"(6).

Use of natural language for information retrieval involves some linguistic problems such as dealing with synonyms, morphology, etc. These issues were discussed by a number of researchers such as Harter(7), Lancaster(8) and Montgomery(9). A good comparison between controlled vocabularies and natural languages was made by Harter(10), who also outlines the advantages and disadvantages of each approach for indexing and online information retrieval.

With regard to Arabic literature, a few studies have also covered these issues. Truly there is no big difference between searching in natural language in Arabic and in other languages. Therefore, a lot of benefit can be gained from English literature in this area. Al Atram(11) investigated the effectiveness of the use of natural language in Arabic information retrieval. He found 57% of the words in title can be used as keywords. He believed that this proportion is not large enough to depend upon the title of documents as sources of access points unless the Arabic titles are improved. Al Sawaydan (12) also agreed with Al Atram on this issue. Al Soynia(13) believed that natural languages and controlled languages each complement the other when used for information retrieval. The relationship between Arabic language and information retrieval has been covered by Al atram(14). Kasem(15) has defined the KWIC indexes and have a brief history of these types of indexing. He discussed the advantages and disadvantages of the KWIC index in the Arabic language.

Because Arabic language has a very complex morphological structure, this has led to differences of opinion between libraries and computer specialists regarding the most suitable way to retrieve information by using roots, stems or words themselves. An experiment was carried out on 355 Arabic bibliographic records by Al Kharashi and Evans(16) to investigate the effect of using words, stems and roots of Arabic words as index terms. They found that using the stem showed better performance over the use of roots and words. Al Kharashi and Evan`s results were encouraging and confirm the results of Al

Atram(17) who did not believe that there is a relationship between reducing the words to their roots and information retrieval, while Ali(18) strongly believed in the importance of reducing the words to their roots. Many morphological analysis algorithms have been suggested and/or implemented. Hegazi and El Sharkawi(19) described a computer aided morphological hierarchy system for a vowelized Arabic text. They based their work on both the morphological rules and phonetic rules of the language. Another paper by Labed, Salhi and Ghazali(20) described the semantic structures for Arabic interrogatives. Ali (21) has criticised Hilal's study about morphological analysis. He mentioned some of the weaknesses in Hilal's study such as the shallow discussion of morphology in Arabic and diacritical marks.

Hegazi and El Sharkawi(22) pointed out there are four phases which are usually considered in the field of natural language processing, these are: the lexical analysis; the syntactic analysis; the semantic analysis, and the factual or pragmatic analysis.

Prefixes and suffixes cause many problems in Arabic information retrieval. To overcome these problems Al Kharashi and Evans(23) suggested that the system should be designed to strip out all suffixes and prefixes from every extracted index term before adding them to the inverted list, while Al Bakhit(24) believes that if the left, right and infix truncations are available in the system these problems will be resolved by using these truncations.

2.3 Arabisation

A great deal of progress has been made to Arabise hardware and software packages in the Arab world in order to introduce automated systems to their libraries and information centres. Ashoor(25) has described the Arabisation of DOBIS/LIBIS, MINISIS and STAIRS. Al Dosary and Ekrish(26) described the state of automation in libraries and information centres in Saudi Arabia.

Aman(27) discussed the technical and linguistic problems which face Arabisation. Booth, Niaz and Al-Swaidan(28) have illustrated these problems. Ashoor(29) considered that the delay of Arabisation in the Arab world was due to the absence of a standard coding system for Arabic characters and the unavailability of completely bilingual terminals. He also added that "no meetings, workshops, seminars or symposia were held to share progress and review problems and prospects of Arabised automated systems"(30). Aman(31) believed that the delay in Arabisation was due the lack of standardisation, while Khurshid(32) felt that there were three obstacles facing the automation development in the Arab world: a shortage of manpower; an absence of standards; and lack of library cooperation. Despite the subsequent success of the Arabisation, Aman(33) believed that the software and hardware still do not answer all the problems posed by Arabic language and culture. Ashoor(34) criticised those institutions which Arabise the same software without consulting with each other. He suggested that "libraries should find ways and means of cooperation with the institutions involved in Arabisation to seek their advice and benefit from their experience"(35).

DOBIS/LIBIS and MINISIS are two widely used Arabised software packages in Saudi Arabia and a comparative study of them was made by Chaudhary and Ashoor(36). Two papers were giving a general description of the DOBIS/LIBIS system were produced by McAllister(37) and McAllister and McAllister(38). The application of DOBIS/LIBIS in Saudi Arabian libraries has been described by Ashoor(39), Khurshid(40), and Booth, Niaz and Al-Swaidan(41). Boyce(42) has also described the MINISIS system. Al Kharashi and Evans(43) suggested two approaches to developing an Arabised computer application for information retrieval. The first approach was to develop the application from scratch and bear in mind the characteristics of the Arabic language. The second approach, however, is based on building an I/O interface for an existing application software built for non-Arabic languages. It seems they preferred the second approach because it has been adapted to Arabise two well known retrieval system software packages, STRIRS and ISIS.

2.4 Evaluation of Arabic titles

If words are used as they appear in documents without modification, this searching is a part of natural language systems. As long ago as 1973, Feinberg justified using natural language for information retrieval because "the growth in the volume of literature in most disciplines, and in the number of users requiring access to that literature, has been frequently remarked upon in recent years. In order to achieve bibliographic control over this growing mass of literature and to satisfy new information needs, the various disciplines have initiated a number of indexing and abstracting services"(44). Using the keywords in the title as an access point is one aspect of automatic computer-based indexing in order to identify the effectiveness of words in titles as subject access points. Al Atram(45) has investigated 279 journal articles. A similar study was done by Al Sawaydan(46) to evaluate the words in titles Arabic journal articles, while Bachir and Buxton(47) evaluated the use of topic sentences to represent Arabic article titles. They found that a mean ranging from 70% to 88% of titles' keywords occurred in the topic sentences. The result of their study cannot be generalised because the sample of this study was taken from Arabic academic journals, the articles of which usually have direct titles. Al Soynia(48) described the permuted indexes and their application in Arabic languages for information retrieval. He discussed in more detail Arabic stoplists. He believes that each library and information centres should establish own their stoplists. On this issue, Al Sawaydan(49) agrees with Al Soynia who recommended 120 stopwords, while Al Atram(50) recommended 202 stopwords which should be stored in the system.

2.5 Special problems related to searching in Arabic

There is a particular problem with the obscure titles of the old Arabic books, the so-celled "heritage books", as the titles have no relevant subject content. Al Soynia(51) noted that the rhyme style is common in these "heritage books". This style reflects the old authors' wishes to attract the attention of the reader and also to make their titles easy to remember.

Several aspects of grammar also pose problems for information retrieval in Arabic. *al- altarif* [the definite article] was discussed by Al Atram(52) and Al Soynia(53). Both presented some solution for this complex issue. Connected particles in Arabic can also cause some difficulty in information retrieval; this issue was discussed by Al Soynia(54) and Al Bakhit(55). Both of these issues are discussed in more detail in Chapter 5.

2.6 Conclusion

This chapter has surveyed several key topic that related to natural language searching, Arabisation etc.These issues will be discussed in the following chapters.

REFERENCES

1. Warner, Amy. Natural language processing. *Annual Review of Information Science and Technology*, 1987, 22, 79-108.

2. Lancaster, F. Wilfrid. *Information retrieval system: characteristics, testing and evaluation,* 1979, pp.279-292.

3. *Ibid.*

4. Doszkocs, Tamas E. Natural language processing in information retrieval. *Journal of the American Society for Information Science*, 1986, 37 (4), 191-196.

4. Pritchard-Schoch. Natural language comes of age. *Online*, 1993, 17 (3), 33-43.

5. Al Atram, Mohammad. *Kafat al lughat al tabiayah fi takshif wa estrjaa alwathiaqe alarabiah* [The effectiveness of natural language for indexing and retrieving the Arabic documents], 1989.

6. Lancaster, ref.2, p.287.

7. Harter, Stephen. *Online information retrieval: concepts, principles and techniques*, 1986, pp.31-53.

8. Lancaster, ref.2,pp.279-292.

9. Montgomery, Christine A. Linguistics and Information Science. *Journal of the American Society for Information Science*, 1972, 23 (3), 195-219.

10. Harter, ref.7, p.51.

11. Al Atram, ref.5, p.3-19.

12. Al Sawaydan, Nasser. *al estrjaa al mawdhey bewasita kalemat al enwan* [subject retrieving by title's words]. In: *Proceedings of Symposium on Using Arabic Language in Information Technology*, 1992, pp.533-568.

13. Al Soynia, Ali, *Kashafat al tabadel wa estrjaa al malonat fi al lughat al arabiah* [Permuted indexes and information in Arabic language], 1988, 13.

14. Al Atram, ref.5, p.4-1.

15. Kasem, Hishmat. *Kashaf al Kalemat al miftahia fi al siyak wa ehtimalatihfi al lughat al arabiah* [KWIC index and its application in Arabic language]. *Alam Al Kutub* [World of Books], 1984, 5 (4), 638-650.

16. Al Kharashi, Ibrahim and Martha W. Evans. Comparing words, stems, and roots as index terms for an Arabic information retrieval system. In: *al moatmar al thani: al lughat alarabiah wa al taquniat al malomatiah al mutqudemah* [Second International Conference: Arabic language and Advanced Information Technology], 1993,pp.297-310.

17. Al Atram, ref.5, p.4-14.

18. Ali, N. *al lughat al arabiah wa al hasub* [Arabic language and computers], 1988, p.322.

19. Hegazi, N.H. and A.A. El Sharkawi. Natural Arabic Language Processing. In: *The Ninth National Computer Conference and Exhibition*, 1986, p.10-5-9.

20. Labed, Lamia, Ridha Salhi and Salem Ghazali. Semantic structures for Arabic interrogatives and coordinated Arabic interrogatives. In: *al moatamar al thani: al lughat alarabiah wa al taquniat al malometiah al mutquademah* [Second International Conference: Arabic Language and Advanced Information Technology], 1993, pp.25-43.

21. Ali, ref.18, p.330.

22. Hegazi and El Sharkawi, ref.19, p.10-5-9.

23. Al Kharashi and Evans, ref.16, p.301.

24. Al Bakhit, Bakhit Suliman. *al bahith fi al enwan fi qawaid al byanat al arabiah* [Title searching in Arabic data bases]. In: *Proceedings of Symposium on using Arabic language in Information Technology*, 1992, pp.569-580.

25. Ashoor, Mohammad Saleh. Arabisation of automated library systems in the Arab World: need for compatibility and standardisation. *Libri*, 1989, 39 (4), 294-302.

26. Al Dosary, Fahad M. and Abdurrahman H. Ekrish. The state of automation in selected libraries and information centres in Saudi

Arabia. *Libri*, 1991, 41 (2), 109-120.

27. Aman, Mohammad M. Use of Arabic in computerised information interchange. *Journal of the American Society for Information Science*, 1984, 35 (4), 204-210.

28. Booth, L.M., Khalid M. Niaz and H M. Al-Swaidan. Arabisation of an automated library system. In: *The Ninth National Computer Conference and Exhibition*, 1986, p.10-32.

29. Ashoor,ref.25, p.294.

30. *Ibid.*, p.301.

31. Aman, ref.27, p.210.

32. Khurshid, Zahiruddin. Application of modern technologies in Arab Libraries. *Libri*, 1983, 33 (2), 107-112.

33. Aman, ref.27, p.210.

34. Ashoor, ref.25, p.310.

35. *Ibid.*

36. Chaudhary, Abussattar and Mohammad Saleh Ashoor. Potential of DOBIS/LIBIS and MINISIS for automating library functions: a comparative study. *Program*, 1990, 24 (2) 109-128.

37. McAllister, Caryl. The online public access catalogue in DOBIS/LIBIS. *Program*, 1987, 21 (1), 25-36.

38. McAllister, Caryl and Stratton McAllister. DOBIS/LIBIS: an integrated online library management system. *Journal of Library Automation*, 1979, 12 (4), 300-313.

39. Ashoor, Mohammad Saleh. Planning for Library automation at the University of Petroleum and Minerals. *Journal of Information Science*, 1982, 5 (5), 193-198.

40. Khurshid, Zahiuddin. Arabic online catalogue. *Information Technology and Libraries*, 1992, 11 (3), 244-251.

41. Booth , Niaz and Al-Swaidan, ref.28.

42. Boyce, Cheryl, MINISIS. *Program*, 1982, 16 (3), 131-141.

43. Al Kharashi and Evans, ref.16, p.301.

44. Feinberg, Hilda. *Title derivative indexing techniques*, 1973, p.ix.

45. Al Atram, ref.5, p.2-2.

46. Al Sawaydan, ref.12.

47. Bachir, Imad and Andrew Buxton. The use of topic sentences for evaluating the representativeness of Arabic article titles. *Journal of Information Science*, 1993, 19 (6), 455-465.

48. Al Soynia, ref.13.

49. Al Sawaydan, ref.12, p.561.

50. Al Atram, ref.5, p.5-3.

51. Al Soynia, ref.13, p.37.

52. Al Atram, ref.5, p.6-12.

53. Al Soynia, ref. 13, p.52.

54. *Ibid.*, p.57.

55. Al Bakhit, ref.24, p.575.

CHAPTER 3
ARABIC LANGUAGE STRUCTURE

3.1 INTRODUCTION

The Arabic language is a member of the semitic family of languages. It is spoken by over 150 million people in 21 Arab countries as the first language (1). An uncertain further number use it as a second language, chiefly in Islamic countries. This chapter will concentrate on Arabic language and its structure, especially those aspects which are important to information retrieval systems, such as affixation, broken plural, morphology, etc. This discussion will be divided into two sections: linguistic considerations and grammatical considerations.

3.2 LINGUISTIC CONSIDERATIONS

3.2.1 The Alphabet

The Arabic alphabet consists of 28 characters which is called حروف الهجاء *hurof alheaja* (Table 3.1). From the twenty-eight characters there are three characters which appear in different shapes as follows (2) :

A- ء *Hamza* [-] sometimes is written with ا *alif* [a] thus أكل *akala* [he ate], sometimes with ئ *ya* [y] thus برئ *barea* [innocent], sometimes with ؤ *waw* [w] thus سؤال *suaal* [question] or without any other characters thus قراءة *geraah* [reading] and similarly ملائم *mulaeem* [suitable].

B- ة *Ta marbuta* [ta] the character ه *haa* [h] with two points above it ة is pronounced like ت *ta* [t]. It is found only at the end of the word (nouns and adjectives) for example سنة *sanatun* [year].

C- ى *alif magsurah* [a] is the character ي *ya* [y] without the point below it. It is represented by the long vowel romanized as in مصطفى *Mustaf* [male name]. The above three characters create a lot of problems in the setting up of an information retrieval system. Therefore some libraries and information centres ignore the *hamaz* and the two points above to unite the input and output for these characters. For example: Title : " التنمية الاجتماعية " *altanmiulat alejtemaih* [social development]. The indexer entered this title into the computer without the two points above ه *ha* [h]. When users want to retrieve this title they must ignore the two points or the title will not be retrieved.

This scenario takes place frequently. Aman (3) pointed out that, while the letters of the Latin alphabet have only one form, the sole exception being the use of capitals, this is not the case with Arabic script, where some characters appear in four, or possibly more different shapes. Two characters, alif and lam ل ا [a, l] have special shapes whenever the alif follows the lam, this being written as لا for example, اب *ab* [father] when this word is prefixed by *lam* [L] it becomes لابي *liabi* [for my father].

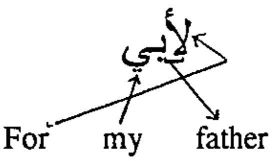


Table 3.1: Arabic Alphabet

Phonemic Symbols for Arabic	Isolated Arabic Letters	Final Arabic Letters	Medial Arbaic Letters	Initail Arbic Letters
?	ء	—	—	—
b	ب	ـب	ـبـ	بب
t	ت	ـت	ـتـ	تت
O	ث	ـث	ـثـ	ثت
j	ج	ـج	ـحب	جب
H	ح	ـح	ـحت	حب
X	خ	ـخ	ـخب	خب
d	د	ـد	ـد	د
6	ذ	ـذ	ـذ	ذ
r	ر	ـر	ـر	ر
z	ز	ـز	ـز	ز
s	س	ـس	ـسب	سب
s	ش	ـش	ـشت	شب
S	ص	ـص	ـصت	صب
D	ض	ـض	ـضت	ضب
T	ط	ـط	ـطـ	طـ
D	ظ	ـظ	ـظـ	ظـ
9	ع	ـع	ـعت	عب
G	غ	ـغ	ـغت	غب
f	ف	ـف	ـفـ	فب
q	ق	ـق	ـقب	قب
k	ك	ـك	ـكب	كب
l	ل	ـل	ـلب	لب
m	م	ـم	ـمب	مب
n	ن	ـن	ـنب	نب
h	ه	ـه	ـهب	هب
w	و	ـو	ـو	و
y	ي	ـي	ـيت	يا

Source : Al Khuli, Ali. *Learn Arabic by yourself,* [1985], p.viii

3.2.2 The Arabic Writing System

The Arabic writing system goes from right to left and most letters in Arabic words are joined together. Twenty-two among the twenty-eight can be joined on both sides and in the process take different shapes depending on their context in a word. "The position can be in the beginning (initial) or in the middle (medial) or at the end (final) of the word"(4). The letter can also be written separately, not connected to another letter in the same word (isolated form) as is shown in Table 3.2.

Table 3.2: Position of letter

Isolated form	ج	زوج	*zawj*	husband
Final	ـج	الحج	*alhaaj*	pilgrimage
Medial	ـجـ	مسجد	*masjid*	mosque
Initial	جـ	جامعة	*jaamiah*	university

3.2.3 Affixation

According to the *Longman dictionary of the English language*, affixation can be defined as "an addition to the beginning or end of, or an insertion in a word, a root, or a whole phrase, serving to produce a derivative word or an inflectional form: an infix, prefix, suffix"(5). Most Arabic words contain some kind of affixation. There is a relationship between affixation and morphology and derivation. (see Section 3.3.3). Most Arabic words have either a prefix, a suffix or an infix. Sometimes all these affixations can be found in one word (6) as in the word مكتتبان *muktatabun* [two subscribers] (see Table 3.3)

Table 3.3: Affixations in Arabic word

prefix	م	*mu*	m
root	كتب	*katab*	write
infix	ت	*ta*	t
suffix	ان	*un*	the mark of dual

The most common prefix is ال , *al* [definite article] as in السيارة *alsyarah* [the car]. There are a number of infixes and suffixes, the nature of which depends on the syntax and derivation of each case.

3.2.4 Homographs

Homographs are "words which have the same spelling but different meanings"(7). In the Arabic language, there are a number of homographs which cannot be understood without vowelization or diacritical marks which are special shapes mainly vowels (*dama* ُ, *fatha* َ, *kasra* ِ, *sokoun* °), placed above and below the Arabic characters to avoid mispronunciation and thus misunderstanding of the word, for example البر has a different meaning according to the diacritical marks as shown in Table 3.4.

Table 3.4: Diacritical marks

Arabic	Diacritical mark	Translation	Transliteration
البُر	dama	wheat	albur
البَر	fatha	land	albar
البِر	kasra	charity	alber

Some words have the same spelling and the same diacritical marks but can only be correctly understood from the context of the whole sentence. For example, المكتبة *almaktaba* which could be (the library) or (the bookshop) depending on the context as in the following examples:

A. أستعرت من المكتبة كتابا *estaarta min almaktabati ketabun* [I borrowed a book from the library]

B. إشتريت من المكتبة كتابا *eshtarut min almaktabati ketabun* [I bought a book from the bookshop]

3.2.5 Synonyms

Synonyms are the opposite of homographs; they are words which have the same meaning but different spelling Words are not synonyms unless "they are close enough in meaning to allow a choice to be made between them in some contexts"(8). Synonyms are widely used in Arabic. Even though there are many Arabic synonym dictionaries, such as *al- alfadh almutradefat almutquarebat alma`ana* [synonyms and near-synonyms], there are few information reteival thesauri incorporating synonym control in Arabic, *Al jaamiah thesauri* is an example. There are aslo a few subject heading lists, for example *Arabic subject headings list* and *Al khazeendar subject headings*. The problem of synonyms will be further elaboratedin Chapter 5 .

3.3 GRAMMATICAL CONSIDERATIONS

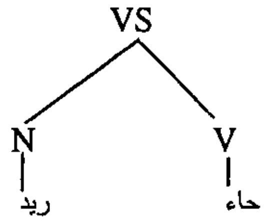
3.3.1 Arabic words

Arabic words are classified into three main categories:

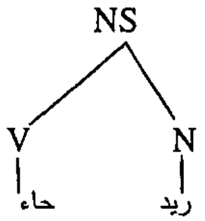
A - إسم *Ism* [noun]
B - فعل *Fil* [verb]
C - حرف *Harf* [particle]

Each category can be subdivided into many types. A single word in Arabic could be a complete sentence, for example قامت *qamat* [she stood], or even the same verb without the pronoun ت [t] gives قام *qama* [he stood](9). The sentence in Arabic might be a verb sentence "VS" or noun sentence "NS" as is shown below:

1)



(2)



1- جاء زيد *Jaa Zaidun* [Zaid came]

2- زيد جاء *Zaidun Jaa* [Zaid came]

3.3.1.1 الإسم *al ism* [the noun]

A noun in Arabic may be classified according to:

1 - number (singular, dual and plural)

2 - case (nominative, genitive and accusative)

3 - and definiteness/indefiniteness (10).

This will be further discussed in Section 3.3.2.

3.3.1.2 الفعل *al fiil* [verb]

Arabic verbs have three main tenses:

1 - ماضي past tense, which is used for all actions which are already completed, e.g. كتب *kataba* [he wrote].

2 - مضارع present tense for all actions not yet complete, e.g. يكتب *yaktubu* [he writes].

3 - أمر command form, e.g. اكتب *uktub* [do write].

Most Arabic verbs can be reduced to a past stem and a present stem, and a standard set of prefixes and suffixes can be added to these stems (11) as is exemplified below in Figs.3. 1-3.

Fig. 3.1: "Past" of standard root KTB (12)

singular	كتب	katab	a	he worte
	كتبت	katab	at	she wrote
	كتبت	katab	ta	you wrote(m)
	كتبت	katab	ti	you wrote(f)
	كتبت	katab	tu	I wrote(c)[1]
dual	كتبا	katab	a	they(two)wrote (m)
	كتبتا	katab	ata	they(two)wrote (f)
	كتبتما	katab	tuma	you(two)wrote (c)
plural	كتبوا	katab	u	they wrote (m)
	كتبن	katab	na	they wrote (f)
	كتبتم	katab	tum	you wrote (m)
	كتبتن	katab	tunna	you wrote (f)
	كتبنا	katab	na	we wrote (c)

---

[1] Common

Fig. 3.2: "Present" of standard root KTB (13)

singular	يكتب	ya	ktub	u	he writes
	تكتب	ta	ktub	u	she writes
	تكتب	ta	ktub	u	you write(m)
	تكتبين	ta	ktub	ina	you write(f)
	أكتب	a	ktub	u	I write(c)
dual	يكتبان	ya	ktub	ani	they(two)write
	تكتبان	ta	ktub	ani	hey(two)write
	تكتبان	ta	ktub	ani	you(two)write
plural	يكتبو ن	ya	ktub	una	they write (m)
	كتبن	ya	ktub	na	they write (f)
	تكتبون	ta	ktub	una	you write (m)
	تكتبن	ta	ktub	na	you write (f)
	نكتب	na	ktub	u	we write (c)

Fig. 3.3: "Command" of standard root KTB

singular	أكتب	u	ktub		do write (m)
	أكتب	u	ktub	i	do write (f)
dual	أكتبا	u	ktub	aa	do (two) write(c)
plural	أكتبوا	u	ktub	u	do write(m)
	أكتبن	u	ktub	na	do write(f)

***harf* [the particle]**

The particle in Arabic is called حرف *harf* meaning letter. It is defined according to "its functional category such as preposition, conjunction"(14) etc. For the purposes of work in information retrieval, most of the particles in the Arabic language would be on a stop list, so there is no need to discuss them in detail. However, some Arabic particles join to other words, such as ب ba [in] بالمدينة *belmadine* [in the city] or ل lam [for] للجامعات *lelljameat* [for universities]. This joining creates difficulty for information retrieval in Arabic, as will be described in Chapter 5.

3.3.2 The number

According to the definition of Crystal, a number is "a grammatical category used for the analysis of word classes, especially nouns which display such contrasts as singular, dual and plural"(15). All the above categories of numbers are used in Arabic. This section will deal with these categories.

3.3.2.1 The singular

Singular in Arabic is divided according to gender:

1) A noun that refers to a male is masculine, such as مدرس *mudarres* [teacher].
2) A noun that refers to a female is feminine, for example مدرسة *mudarresah* [teacher].

3.3.2.2 The dual

Arabic is one of the very few living languages which still has التثنية *altatheriah* [the dual] as a separate form which denotes the number two of things(16). The dual is formed regularly by adding the suffix ان un [mark of the dual] in the nominative case, and ين in [mark of the dual] in oblique and accusative cases. For example: قلم *qalam* [a pen] the dual is expressed by adding the suffix ان [un] to the singular, thus قلمان *qalamun* [two pens] as this is in nominative case. But in the oblique and accusative cases the suffix ين [in] should be added to the singular thus, قلمين *qalamin* [two pens].

As seen above the dual has changed according to its syntax case. There will be more detailed discussion of syntax in Arabic language and its specific effects on sentences in Section 3.3.4.

3.3.2.3 The plural

In Arabic, there are two kinds of plural which are generally known as the sound (or strong) and the broken (or weak).

3.3.2.3.1 The sound plural

The sound plural is subdivided into two categories according to the gender as follows:

1)- The masculine sound plural which is formed by adding the suffix ون un [the mark of the masculine] to the singular in the nominative case. In the oblique and accusative cases the suffix ين [in] is added to the singular.

2)- The feminine sound plural which is formed by adding the suffix ات at [the mark of the feminine] in all syntax cases (nominative, oblique and accusative) (17). For example, as Table 3.5.

Table 3.5: The Sound Plural

	Singular	Masculine		Feminine	
		Nominative	Oblique and Accusative	Nominative	Oblique and Accusative
Arabic	مدرس	مدرسون	مدرسين	مدرسات	مدرسات
Translit.	mudarres	mudrres (un)	mudrerres(in)	mudrresatu	mudrresati
Translation	teacher	teacher(s)	teacher(s)	teacher(s)	teacher(s)

From the above table one can note that in English 's' is added to make the plural, but Arabic is different. It depends upon the type of plural, whether it is masculine or feminine. It also has different shapes, which are called broken plural, as will be shown in the next section. This changing in the suffixes causes some difficulty when the above words are to be retrieved.

3.3.2.3.2 The broken plural

This is the second type of plural in Arabic; it is also known as a weak or irregular plural. It was mentioned above that the sound plural suffixing by ون ، ين , and ات depending on the type of plural, whether it is masculine (nominative or oblique) or feminine. However, in the case of the broken plural, the matter is different, because the broken plural has a number of measures which in Arabic is called أوزان *awzan*. Unfortunately, the singular word does not help in knowing them. Some of the most common types of broken plural are listed in Table [3.6] with their measures(18).

There are more types of broken plurals, but they are not in common use. Also, some nouns have two or more different forms of broken plural such as بحر *bahar* [sea] which can be pluralised as بحور *buhur* or بحار *behar* or أبحر *abhur*. Hence, broken plural in Arabic can cause some difficulty in information retrieval, particularly when natural language is used in indexing.

Table 3.6: The measures of broken pulral

The measure alwazan	Singular			Broken plural			Notes
	Arabic	Transliteration	Translation	Arabic	Transliteration	Translation	

1-Afaal	قلم
	ولد
		qalam
		walad
			a pen
			a boy
				أقلام
				أولاد
					(a)ql(a)am
					(a)wl(a)ad
						pen(s)
						boy(s)
							ا alif [a] is prefixed and alif followed by its becomes لا la [L] added to the second letter.

2-fuool	بيت
	قبر
		bayt
		qabr
			a house
			a grave
				بيوت
				قبور
					buy(oo)t
					qub(oo)r
						house(s)
						grave(s)
							و Waw [w] is added between the second and third letters.

3-fualaa	فقير
	شريك
		faqeer
		shareek
			a poor man
			a partner
				فقراء
				شركاء
					fuqar(aa)
					shurak(aa)
						poor men
						partner(s)
							ا alif [a] and ء hamza [-] are suffixed to the four lettered singular.

4-Afilah	لباس
	لسان
		lebaas
		lesaan
			a dress
			a language
				ألبسة
				ألسنة
					(a)lbesa(h)
					(a)lsena(h)
						dress(es)
						language(s)
							ا alif [a] prefixed and ه haa [h] is suffixed to the four-lettered singular.

5-Mafaail	منزل
	مكتب
		manzel
		maktab
			a house
			an office
				منازل
				مكاتب
					man(aa)zel
					mak(aa)teb
						house(s)
						office(s)
							ا alif [a] is added after the first two letters of the singular.

6-Mafaaeel	منديل
	دستور
		mandeel
		dastoor
			a handkerchief
			regulation
				مناديل
				دساتير
					man(aa)deel
					das(aa)t(ee)r
						handkerc-hief(s)
						regulation(s)
							1 In this category, the singular consists of five letters, ا alif [a] is added after the first two letters
							2 The same as above but و waw [w] is omitted and ي yaa[y] is added after the third letter.

3.3.3 Morphology and derivation

It is helpful before starting a discussion on morphology and derivation in Arabic to define what these terms mean. First, morphology is "a branch of grammar which studies the structure of words"(19). Derivation may be defined as "a major type of word formation where a certain kind of affix is used to form new words. A contrast is intended with process of inflection, which uses another kind of affix in order to form variants of the same word"(20). Arabic language, in common with other semitic languages, has a very complex morphological and derivational structure. As noted by Hegazi ,Ali and Abed (21), the morphological nature of the Arabic language exhibits a high degree of redundancy due to the fact that most Arabic words are morphologically derived from roots, no matter how long the word and how complicated (or how short and apparently simple), for example, كتب *KTB* [write]. This root series can be used to give many different words and meanings as shown in Table 3.7.

Table 37: Some derivations of the root (KTB) (22).

Root	Arabic	Transliteration	Translation	Notes
*K-T-B	كتاب	KiTaaB	book	Simply a choice of internal vowelling with no prefixes or suffixes.
	كتابة	KiTaaBa	writing	Feminine suffix added to the previous word to change its meaning.
	كاتب	KaaTiB	a writer, clerk	Change of internal vowelling
	مكتب	maKTaB	office, desk	The very common prefix o ma [m] here, plus another change in vowelling.
	مكتبة	maKTaBa	library, bookshop	Same as above, but again the feminine ending is used to change the meaning.
	مكاتبة	mukaaTaBa	correspondence	Another common prefix o ma [m] plus another change in vowelling

*The root letters are given in capitals for the sake of clarity

3.3.4 Syntax

Crystal has defined syntax as "a traditional term for the study of the rules governing the way words are combined to form sentences in a language"(23). The meaning of syntax in Arabic is the change which takes place to the word ending according to the part of speech (tools) which precede the word. For example, the suffix of المدرسون *almudarrsun* [teachers] will change according to the previous tools, as shown below:

1. جاء المدرسون *Jaa almudarrsun* [the teachers came].
2. شاهدت المدرسين *shahatu almudarrsin* [I saw the teachers].
3. ذهبت مع المدرسين *dhabatu maa almudarrsin* [I went with the teachers].

It is noted that the suffix of المدرسون changed because the word (teachers) in the first example is فاعل *fa'al* [the subject], while it is the مفعول به *maful* [an object] in the second example. In the third example, the word is إسم مجرور *Ism majurror* [noun in the genitive] the preposition is مع *ma`a* [with]. However, sentences in English do not change in all cases and there is no need to change the termination of the word because of the previous tools. As seen above, the shape of word will change according to the cases of syntax. Hence, the syntax in Arabic also causes some difficulty for Arabic information retrieval and will be discussed in more detail in Chapter 5.

3.4 CONCLUSION

This chapter has discussed the Arabic language structure which has an effective role in information retrieval. Next chapter will deal with Arabistion and natural language searching.

REFERENCES

1. Crystal, David. *An encyclopedic dictionary of language and languages*, 1991, p 25.

2. Kapliwatzky, Jochanan. *Arabic language and grammar,* 1972, pp.19-20, 31- 32.

3. Aman, Mohammed. Use of Arabic in computerised information interchange. *Journal of the Aamerican Society for Information Science,*1984, 35 (4), 205.

4. *Ibid.*

5. *Longman dictionary of the English language*, 1934, p.26.

6. Shaheen, Abdulsabour. *Arabic: the language of science and technology,*[n.d.], p.265.

7. Crystal, ref. 1, p.174.

8. *Ibid.*, p.378.

9. Ibrahim, Farid. *A syntactically based prepocessor for a limited experimental Arabic document retrieval system, 1988,* p.53.

10. Bright, William, ed. *International encyclopedia of linguistics,* 1992, p.51.

11. Alkharashi, Ibrahim and Martha W. Evans. Comparing words, stems, and roots as index terms for an Arabic information retrieval system. In: *Arabic language and information technology,* 1993, p.299.

12. Wickens, G. M. *Arabic grammar*, 1980, p.39.

13. *Ibid.*, p.51.

14. Hegazi, N H. and A. A. Elsharkawi. Natural Arabic language processing. In: *The Ninth National Computer Conference and Exhibition*, 1986, p.10-5-2.

15. Crystal, ref. 1, p.274.

16. Shaikh, Shafi. *A course in spoken Arabic,* 1978, p.18.

17. Al Khuli, Ali. *Learn Arabic by yourself,* [1985], p.113.

18. Shaikh, ref. 16, p.20.

19. Crystal, David. *A dictionary of linguistics and phonetics*, 1991, p.225.

20. Crystal, ref. 1, p.98.

21. Hegazi, Nadia, Nabil Ali and Ehsan Abed. Information content in textual data: revisited for Arabic text. *Journal of the American Society for Information Science*, 1987, 38 (2), 133.

22. Smart, J. R. *Arabic*, 1986, p.69.

23. Crystal, ref. 19, p.341.

CHAPTER 4

ARABISATION AND NATURAL LANGUAGE SEARCHING IN ARABIC

4.1 INTRODUCTION

The previous chapter discussed the structure of Arabic language and its effect on information retrieval. This chapter will discuss the adaptation of hardware and software in order to accommodate Arabic characters. It will also deal with the use of natural language searching in Arabic.

4.2 ARABISATION

In the context of this dissertation, Arabisation refers to "various modifications needed to adapt an automated library system originally designed for Latin scripts to systems that can handle both Latin and Arabic scripts"(1). Over the last few years, libraries, information centres, institutions and universities in the Arab world have introduced automated systems to their libraries(2). Some of these libraries have automated all the library's functions while others have attempted to automate only some functions. Unfortunately, many libraries in Saudi Arabia still do not have automated systems. Booth, Niaz and Al-Swaidan (3) have listed many of the problems faced by Arabisation in its early stages. These problems can be summarised as follows:

1. Arabic script must be entered and displayed with a right-to-left orientation. The available terminals do not compensate for this.
2. Arabic script makes intensive use of letters with diacritical marks, with special pronunciation designated by diacritices over or under the letter characters.
3. Some words, particularly the definite article, are not separated from the following words by a space as the are in Latin language.

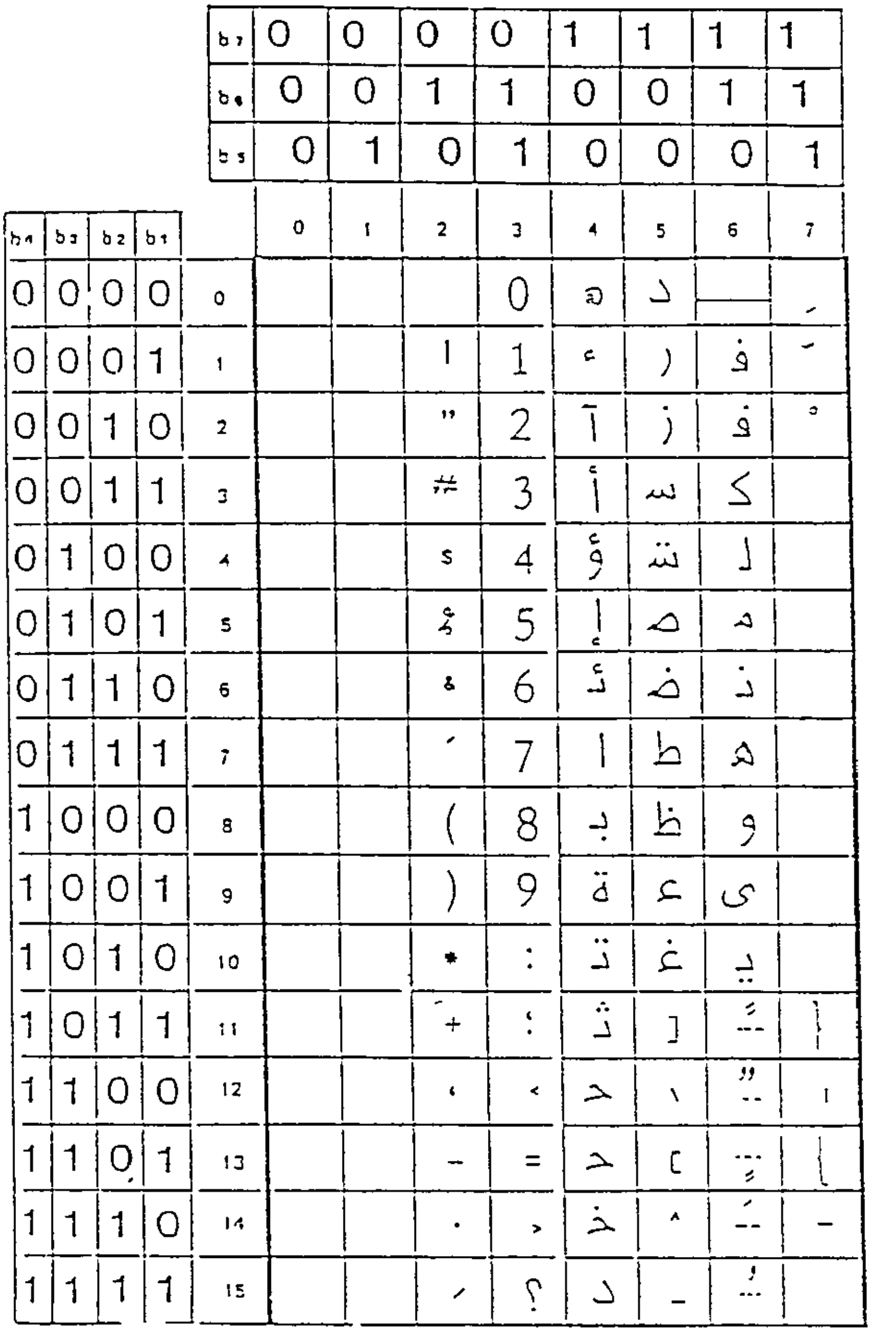
4. Arabic is very context sensitive so that a word can have many meanings depending on use.
5. Some Arabic letters have from two to five forms depending on the positioning of the letter within a word.

A survey has been done by Al Dosary and Ekrish (4), showing that ten different software packages are used in Saudi libraries and information centres which were originally designed for Latin languages, "while four libraries have developed their own in-house system"(5). There are a number of information retrieval systems used in the Arab world, e.g. DOBIS/LIBIS, MINISIS and SMART. The discussion which follows will focus on the two major systems which are used in Saudi Arabian libraries and information centres: DOBIS/LIBIS and MINISIS.

4.2.1 Arab Standard Metrological Organisation (ASMO) system

ASMO449 (Fig.4.1) is a standard coding system for Arabic language. It is a 7-bit coded Arabic set for information interchange. It was developed in October 1982 by the Arab League Educational, Cultural and Scientific Organisation (ALECSO) (6). This code is based on The American Standard Code for Information Interchange (ASCII) (7). In order to develop a unified code for the use of Arabic characters in information, computer experts from the Arab world convened a series of meetings for this purpose. "Some Arab countries have worked together for seven years to develop the appropriate set of Arabic characters that can be used in the computer filed throughout the Arab world" (8).

Fig. 4.1: ASMO 449



				b7	0	0	0	0	1	1	1	1
				b6	0	0	1	1	0	0	1	1
				b5	0	1	0	1	0	0	0	1
b4	b3	b2	b1		0	1	2	3	4	5	6	7
0	0	0	0	0				0	@	ذ	ـ	ـِ
0	0	0	1	1			!	1	ء	ر	ف	ـّ
0	0	1	0	2			"	2	آ	ز	ق	ـْ
0	0	1	1	3			#	3	أ	س	ك	
0	1	0	0	4			$	4	ؤ	ش	ل	
0	1	0	1	5			٪	5	إ	ص	م	
0	1	1	0	6			&	6	ئ	ض	ن	
0	1	1	1	7			'	7	ا	ط	ه	
1	0	0	0	8			(	8	ب	ظ	و	
1	0	0	1	9			)	9	ة	ع	ى	
1	0	1	0	10			*	:	ت	غ	ي	
1	0	1	1	11			+	؛	ث	]	ـً	}
1	1	0	0	12			،	<	ج	\	ـٌ	|
1	1	0	1	13			-	=	ح	[	ـٍ	{
1	1	1	0	14			.	>	خ	^	ـَ	–
1	1	1	1	15			/	؟	د	_	ـُ	

Source: Dchieche, Mohammed. Arabic writing: reform in printing and standardisation in informatics. In: *Proceedings of Symposium on using Arabic language in informationtechnology*, 1992, p.727.

4.2.2 Arabisation of DOBIS/LIBIS

According to McAllister and McAllister (9), DOBIS/LIBIS is the result of the joint work of two universities and IBM. The first university is the University of Dortmund in West Germany, which started work in 1971 on the Dortmunder Bibliothekssystem (DOBIS, Dortmund library system). The second is the Catholic University of Leuven (Louvain) in Belgium which subsequently started development work with Leuven's Integral Bibliotheek System (LIBIS, Leuven library system). DOBIS/LIBIS is an on-line, integrated, interactive system that meets the major library requirements (10), such as cataloguing, acquisition, etc. According to Mc Allister (11), the system is now running in over 150 locations throughout the world.

The first Saudi institution to Arabise DOBIS/LIBIS was King Saud University (KSU), which started the Arabisation in 1984. For its Arabisation of DOBIS/LIBIS KSU began automation with an IBM 3278 model terminal with bilingual (Latin/Arabic) keyboards, which are known as "XCOM 2"(12).
King Fahd University of Petroleum and Minerals (KFUPM) was the second institution in Saudi Arabia which Arabised its system The Library Automation Project Management (LAPM) at KFUPM reviewed and tested various bilingual terminals available in the market place at the time (i e.1985). Finally, the "AL-Arabi" terminal was chosen for its Arabisation for DOBIS/LIBIS. "AL-Arabi" is a bilingual terminal including a complete set of Arabic characters, numbers and diacritic characters that are compatible with ASMO449 and based on ASCII. In 1986 an IBM 3192 computer operating on XBASIC became available, therefore the (LAPM) decided to drop the "AL-Arabi" terminal in favour of the IBM 3192 (XBASIC), because it is more suitable for Arabic characters(13).

In 1986, the Institute of Public Administration (IPA) was the third Saudi institution to Arabise DOBIS/LIBIS. Unfortunately IPA did not benefit from the experience of the Arabisation program of DOBIS/LIBIS at both KSU and KFUPM(14).

DOBIS/LIBIS is now running in five libraries and information centres (Table 4.1) in Saudi Arabia (15).

Table 4.1: DOBIS/LIBIS locations in Saudi Arabia

INSTITUTION	LOCATION
1. Institute of Public Administration Library (IPAL).	Riyadh
2. King Abdulaziz University Library (KAUL).	Jeddah
3. King Fahd University of Petroleum and Minerals Library (KFUPML).	Dhahran
4. King Saud University Library (KSUL).	Riyadh
5. Umm al Qura University Library (UAUL).	Makkah

The DOBIS/LIBIS system provides several indexes which serve as access points to the on-line public catalogue (OPAC). Parton can search for the author, title, subject, etc. in addition to various other entries. Permutation for name and title entries are provided for retrieval of information by keywords. Boolean searching, combinations of dates, material types, etc. and truncation facilities are provided(16).

4.2.3 Arabisation of MINISIS

"MINISIS was developed by the International Development Research Centre (IDRC) of Canada to meet the need for a low-cost hardware/software package for online data entry and interactive retrieval. The system was created primarily for use in a library environment"(17). It can be run in more than one language on the same machine at the same time(18). The system runs on any of the HP3000 series of mini computers. MINISIS was Arabised by the Arab League Documentation Centre (ALDOC) in 1982. The ALDOC assumed full responsibility for the Arabisation of MINISIS. Ashoor (19), mentioned that ALDOC faced various hardware and software problems especially in its early stages (see Section 4.2).

MINISIS is a popular software package specially developed for the Arab world, according to *The ARIS-NET Newsletter*(20).The system is used in 400 locations throughout the world, of which 40 libraries and information centres are in Arab countries, among them four libraries (Table 4.2) in Saudi Arabia(21).

Table 4.2: MINISIS locations in Saudi Arabia

INSTITUTION	LOCATION
1. King Abdulaziz Public Library (KAPL).	Riyadh
2. King Fahd National Library (KFNL).	Riyadh
3. King Faisal Centre for Research and Islamic Studies (KFCRIS).	Riyadh
4. King Faisal University Library (KFU).	Dammam

OPAC facilities are provided in MINISIS through QUERY language. All fields in the database can be searched. The QUERY processor supports the Boolean operations (AND, OR, and NOT) which allow combination of records containing various keys(22). Truncation is provided in the system.

4.3 NATURAL LANGUAGE SEARCHING

In recent years, studying and analysing natural language has become quite important. Despite the attention paid to natural language searching in English, there is no such attention paid to this matter in Arabic literature. Searching in natural language in Arabic is similar to searching in English and other languages. There is no big difference between them except some things which refer to the nature of Arabic language which has been discussed in Chapter Three. The natural language could be defined as "An indexing system in which no index vocabulary controls are imposed"(23). The search terms will be used without modification as they appear in the title on other designated search field. Natural language, as other indexing languages, has advantages and disadvantages. The main ones are as follows.

4.3.1 Advantages

1. the natural language allows a searcher to select the most specific term or phrase or group of words to describe the search(24);
2. it may be the only approach to finding information. The terminology may be too new to have entered into an authority list(25);
3. "natural language permits the conduct of searches of unlimited specificity. Thus, it is possible to look for a document in which individual companies, products, processes or even persons are named"(26).

4.3.2 Disadvantages(27)

1. words in text may have different senses or meanings depending on the domain;
2. the same words may be used in different sentences or phrases concerning totally dissimilar concepts;
3. completely different words may be used to express exactly the same concept;
4. authors may use a variety of expressions to describe their findings.

4.4. STOPWORDS

Stopwords or stoplists are words defined as "those words which carry no useful information, by themselves, such as prepositions, articles, conjunctions, pronouns, introductory prefixes as well as single letters"(28). In addition to this there are a number of words which could be treated as stoplist words, such as verbs, adjectives, etc, which have no value as index terms. Feinberg states that "words which are non-significant in one context may be significant in another"(29). He gave an example of words which may be significant or non-significant depending on the context in which they are used, such as(30):

Table 4.3: Word significant

non-significant	**significant**
sterilization <u>program</u>	computer <u>program</u>
employment <u>division</u>	cell <u>division</u>

Sometimes a single letter or word has an important scientific and technical meaning when combined with another, for example, "Vitamin A" or "on line", etc. With Arabic, some special problems arise. For example, the system can recognise stopwords if they are only a distinct word, that is, separated by spaces or punctuation. Also, there are a number of letters which prefix some words, such as *ba* [b], *fa* [f], *la* [l]. These problems are further discussed in Chapter 5.

Golist words are words that are not found on the stoplist. Words in the golist will be treated as "keywords". The system attempts to match each word in a key with a stoplist. A match means that the word is a stop word.

4.5 BOOLEAN OPERATORS

Boolean operators are a very important technique for information retrieval systems, especially when the natural language is used as an indexing language. "Boolean searching is commonly referred to as post-coordinated searching because the word relationships are not built into the indexes, but must be defined by the searcher in the search request expression"(31) The major Boolean operators used for information retrieval are OR, AND and NOT (32) and their application is demonstrated Figure 4.2.

1. The OR function is used to broaden a search by retrieving all records containing terms A and B; it is called the union of A and B.
2. The AND function is used to narrow a search by specifying that items retrieved must contain terms A and B; it is called the intersection of A and B.
3. The NOT function is used to narrow a search and it retrieves items that contain term A and eliminates items that contain term B; it is called the difference between A and B.

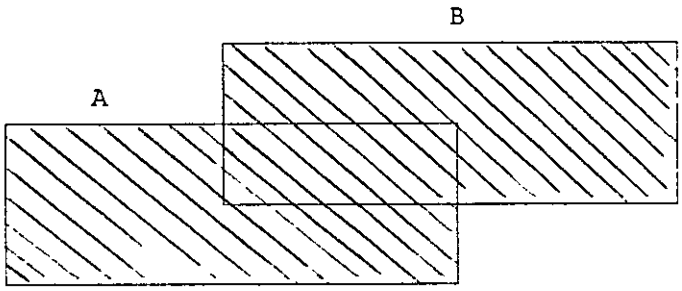
4.6 TRUNCATION

Prefixing, infixing and suffixing are quite prevalent in Arabic and thus truncation for searching in natural language is both very important and necessary. As seen in Chapter Three, Arabic language is a morphological and derivation language. This means that most words in Arabic will have different shapes, especially in the beginning and end of the word, so truncation should be offered in all Arabic information retrieval systems. Unfortunately, there are few systems which offer truncation to search in Arabic, though these systems have facilities for such a procedure.

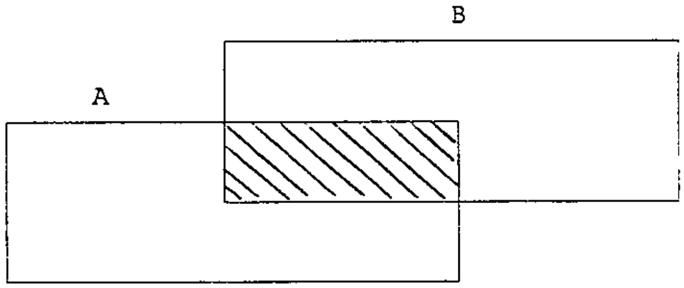
Truncation allows searchers to retrieve a word stem without specifying all possible prefixes and/or suffixes which could occur. It is "a powerful device for searching for large groups of related words. It is also a useful time-saver, for it avoids the need to enter separately a list of terms all having the same stem. It is even useful for handling many singular-plural situations"(33). There are four types of truncation:

1. right truncation, that is, ignoring the ending of a word;
2. left truncation, that is, ignoring the beginning of a word;
3. simultaneous left and right truncation;
4. infix truncation, that is, specifying the beginning and end of a word but leaving the middle unspecified(34).

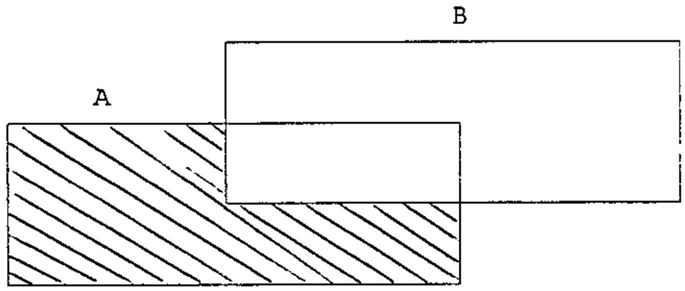
Fig.4.2: Boolean Operators



A OR B



A AND B



A NOT B

4.7 CONCLUSION

This chapter has discussed the main points related to Arabisation in general and of two software packages (DOBIS/LIBIS and MINISIS) in particular. Natural language searching has also been briefly covered, with particular aspects relevant to Arabic being pointed out. Specific problems of subject searching using Arabic book titles will be covered in the next Chapter.

REFERENCES

1 Booth, L M., khalid M. Niaz and H. M. Al-Swaidan. Arabisation of an automated library system. In: *The Ninth National Computer Conference and Exhibition*, 1986, p.10-3-2.

2. Ashoor, Mohammad Saleh. Arabisation of automated library systems in the Arab world: need for compatibility and standardisation. *Libri*, 1989, 39(4), 294.

3. Booth, Niaz and Al-Swaidan, ref.1, p10-3-4.

4. Al Dosary, Fahad, M. and Abdurrahaman H. Ekrish. The state of automation in selected libraries and information centres in Saudi Arabia. *Libri*, 1991,41(2), 115.

5.*Ibid.*

6. Musa, F. A. A system for processing bilingual Arabic/English text. *Journal of the American Society for Information Science*, 1986, 37 (5), 288.

7. Al Dosary and Ekrish, ref. 4, p.112.

8. Aman, Mohammad M. Use of Arabic in computercised information interchange. *Journal of the American Society for Information Science*, 1984, 35(4), 206.

9. McAllister, Caryl and A.Stratton McAllister. DOBIS/LIBIS: an integrated, online library management system. *Journal of Library Information*, 1979, 12 (4), 303.

10. Ashoor, Mohammad Saleh. Planning for library automation at the University of Petroleum and Minerals. *Journal of Information Science*, 1983, 3, 194.

11. McAllister, Caryl. The online public access catalogue in DOBIS/LIBIS. *Program*, 1987, 21(1), 25.

12. Ashoor, ref. 2, p.296.

13. *Ibid.*, p.298.

14.*Ibid.*

15. Al Dosary and Ekrish, ref. 4, p.113.

16. Chaudhry, Abdussattar and Mohammad Saleh Ashoor. Potential of DOBIS/LIBIS and MINISIS for automating library functions: a comparative study. *Program*, 1990, 24(2), 119.

17. Boyce, Cheryl. MINISIS. *Program*, 1982, 16(3), 131.

18. *Ibid.*, p.132.

19. Ashoor, ref. 2, p.295.

20. Arab League Documentation Centre. *ARIS-NET Newsletter*, 1988, 2 (24), 1.

21. Al Dosary and Ekrish, ref. 4, p.113.

22. Chaudhry and Ashoor, ref. 16, p.120.

23. Young, Heartsill, ed. *The ALA Glossary of Library and Information Science,* 1983, p.152.

24. Chamis, Alice Yanosko. *Vocabulary control and search strategies in online searching*, 1991, p.13.

25.*Ibid.*

26. Lancaster, F. Wilfrid. *Information retrieval systems: characteristics, testing and evaluation*, 1979, p.289.

27. Smeaton, Alan F. Prospects for intelligent, language-based information retrieval. *Online Review, 1991*, 15 (6), 373.

28. Bachir, Imad and Andrew Buxton. The use of topic sentences for evaluating the representativeness of Arabic article titles *Journal of Information Science*, 1993, 19, 458

29. Feinberg, Hilda . *Title derivative indexing techniques: a comparative study*, 1973, p.49.

30. *Ibid.*

31. Hildreth, Charles R. To Boolean or not to Boolean? *Information Technology and Libraries*, 1983, 2 (3), 236.

32. Harter, Stephen. *Online information retrieval: concepts, principles and techniques*, 1986, p.76.

33. Lancaster, ref. 26, p 292.

34. *Ibid.*

CHAPTER 5

ARABIC BOOK TITLES AS SUBJECT ACCESS POINTS

5.1 INTRODUCTION

It has been noticed in the previous few years with regard to information retrieval in the Arabic language that the use of words from the title as subject access points for documents. This regard to the large number of articles and books that are published. So a number of researchers (1-3) have evaluated and examined Arabic titles whether they informative or not. There is a debate between librarians and information workers about the effectiveness of using the words in titles as subject access points (4-5). The various points concerning this debate are outlined below:

1. It is impossible to judge the subject of a book or article from its title in many cases.
2. Many titles begin with common phrases such as "an introduction to", "history of", etc.
3. The efficiency of title indexes is highly dependent upon authors' choices of titles for their work.
4. The meaning of the titles may not be clear, or may be misleading or ambiguous.

In addition to the above criticisms, there are a number of comments which were specific to using natural language for information retrieval (see Section 4 3). However, even though there are a number of criticisms for using words in titles for information retrieval there are a number of advantages (6) for this method, which could be summarised as follows:

1. A large number of titles can be processed quickly and cheaply.
2. The KWIC index provides many access points.
3. It is produced with little human intellect.
4. The index uses words given by the author in the title, which are more accurate.

5.2 METHODOLOGY

In order to evaluate the Arabic book titles as subject access points, this study has followed these stages:

Stage 1 . Choice of a sample of Arabic book titles from records in KFNL database. The main sample totalled 351 titles. The analysis of these titles involved consideration of each individual word in each title. After these titles had been studied it was decided that time constraint would make it impractical to carry out a complete detailed analysis of the entire main sample. For this reason, a secondary sample was taken from the original sample. This was achieved by taking the even numbered records of the original sample, thus making a secondary sample of titles from 192 records. Such a number was considered manageable and it was also felt that the coverage would by representative of a broad range of subject areas.

Stage 2. Keywords in titles of all records were compared with the subject headings assigned to these records by KFNL. This compassion was necessary in order to determine or confirm the subject area covered by the book, because it was not possible to have access to the original item during the period when the records were being analysed.

Stage 3. The analysis of the 192 titles from the secondary titles comprised a total of 917 words which were divided into five categories. This division followed the methodology outlined by Kraft (7) with some modifications that were necessary for this particular study (see Fig. 5.1). The five types are as follows:

Type (1): A title which contains a word from of the Arabic subject heading list exactly or in some root form, e.g.

Title: *muquadema fi al reyadheyat* [an introduction to mathematics]

Subject heading: *al reyadheyat* [mathematics]

Type (2): A title which contains a synonym of the Arabic subject heading list, e.g.

Title: *derasat fi al trabiah* [studies in eduction]

Subject heading: *al taalim* [Teaching]

Type (3): A title not of type 1 or 2 but contains significant words not in Arabic subject headings, e.g.

Title: *hurof al shart* [conditional particles]

Subject heading: *Al lughat al arabiah - nahu* [Arabic language - grammar]

Type (4): A title not of type 1, 2 or 3 but containing non-significant words; These are titles containing non-descriptive words, e.g.

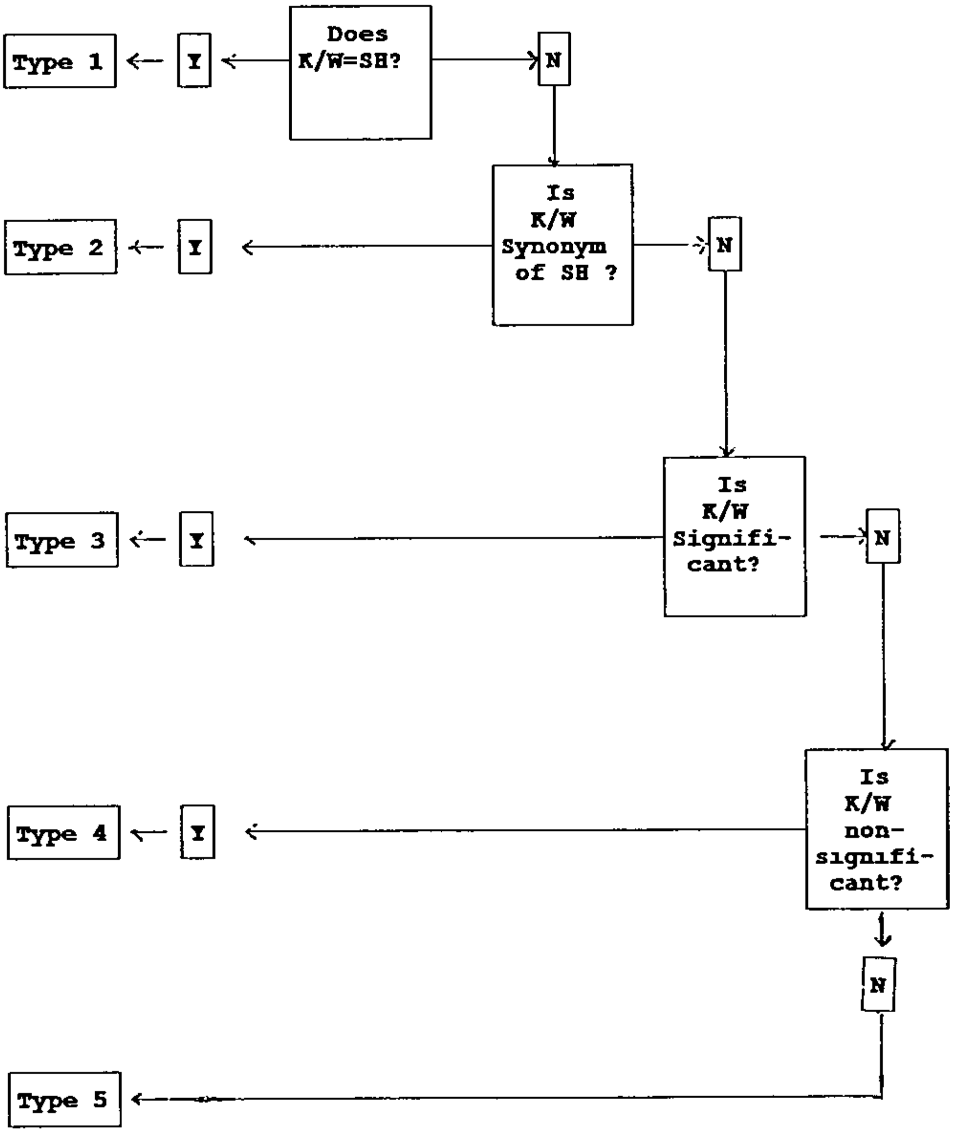
Title: *al sabian* [the seventy]

Subject heading: *al falak* [astronomy]

Type (5): Stop list words.

e.g. to, for, of, etc.

Fig. 5.1: Flowchart of determining word types



Based on: Kraft, Donald. A comparison of keywords in context (KWIC) indexing of titles with a subject heading classification system. *American Documentation*, 1964, 15(1),51.

5.3 RESULTS OF TITLE ANALYSES

The results of the statistical analysis are given in the tables below.

Table 5.1: Types of words in Arabic book titles

Types	Number of Words in titles	% of total
Type 1	266	29.0%
Type 2	31	3.4%
Type 3	89	9.7%
Type 4	338	36.9%
Type 5	193	21.0%
Total	917	100%

Type 1: Key words = subject headings

Type 2: Key words = synonyms of subject headings

Type 3: Key words not covered by subject headings

Type 4: Non-significant words

Type 5: Stop words

The title words in Table 5.1 an also be broken down into these that could be considered significant (i.e. meaningful subject words) or non-significant (i.e. those that do not have subject context) (Table 5.2). Types 1, 2 and 3 were counted as significant words or "keywords" or "indexable words", while Types 4 and 5 were counted as non-significant words

Table 5.2: Significant and non-significant words

	Significant	Non-significant	Total
Number of words in titles	386	531	917
% of total	42.1	57.9	100%

5.4 DISCUSSION

In light of the statistical results which were presented in Section 5.3, this section will discuss the title words in Arabic titles as subject access points.

5.4.1 Significant and non-significant words

Significant words are those words which can be used as an index term or as keywords. Of the 386 significant words in the sample, it was found that 266 (68.9%) of these corresponded to the subject headings. It was also found that 33 (8.5%) of the 386 significant words were synonyms of subject headings which were given by KFNL. Further analysis of these significant words led to the finding that 89 (23%) words were significant words but were not found under the list of subject headings. Of the 531 words that were considered non-significant, 193 (36.3%) were stopwords, while the remainder were found to be words that did not convey subject content. In other words, they were general words, e.g." an introduction to" , "study of " , " history of " ,etc.

It is apparent from the sample in Table 5.1 and Table 5 2 that the words of a title can be utilised to express the meaning of the subject except that the percentage of significant words in the titles of the sample at 42.1% was not seen as sufficaent to be able to utilise this as a method for information retrieval. For this methology to work it is though that at least 70% of titles must contain significant words as has been mentioned by Al Atram (8).

This deficiency was compounded by a further deficiency in the subject headings, where it was found that only 32.6% of the sample was categorised under subject headings suitable for information retrieval. This percentage is obviously too small and makes this methodology not reliable for information retrieval.

It was therefore concluded that for a reliable method for information retrieval to be used one must combine the two methodologies, namely natural language system and vocabulary controlled language for information retrieval. This recommendation has been reached by Al Atram (9), and Al Sawaydan (10).

5.4.2 Obscure/misleading titles

The efficiency of natural language keyword searching is highly dependent on authors' choices of titles for their works. Some authors when they want to title their works attempt to attract the attention of the reader at the expense of the subject content. From the sample it was found that 26 (13.5%) titles were totally obscure, that is to say, these titles did not contain any keywords that could give some idea as to the subject matter addressed under such titles. In some cases, even, some words within the title gave a completely contradictory meaning as to the subject matter addressed and would as a result be totally misleading if used as subject access points. For example القانون المسعودي *al quanun al masuadi* [al masuadi law]. When users are searching for books on law they will retrieve the above title because the title contains the word "law". In fact the subject matter of this book is geography. There is no mention of law in the book at all. This problem is common in the heritage books, and in some cases alsowith modern titles.

A similar study on titles of journal articles was done by Al Atram (11) which showed that among 280 titles, there were 12 (4.2%) titles which were obscure. This differences between the two results with obscure titles refers to the nature of each sample . The sample of this study were books , while the sample of Al atram study were journal articles which is not surprising. This is because journal articles are by their nature more specific than book titles.

To overcome this obscurity in Arabic titles, Al Atram (12) suggested that some

The existence of obscure titles underscores the need for the use of concept analysis for indexing and retrieval. These titles cannot be accessed without the addition of significant words or by the use of subject headings as a method for information retrieval.

In order to identify the clarity of titles the following steps were taken:

1- The fixing of the titles length (1-11 words)

2- The analysis of title clarity by identifying the number of significant words within a given title

3- The following well known evaluation equation was used:

$$\frac{\text{Number of significant words in the title}}{\text{Number of words in the title}} \times 100$$

e.g. a title comprised of four words with only one significant words

$$\frac{1}{4} \times 100 = 25\ \%$$

4- After this the titles were divided up according to their respective levels of clarity, as is presented in Table 5.3.

Table 5.3: Clarity of Arabic book titles

% of title clarity	Occurrences	Percentages
0	21	11%
0-9	0	0
10-19	5	2.6%
20-29	6	3.1%
30-39	20	10.4%
40-49	7	3.6%
50-59	38	19.8%
60-69	32	16.7%
70-79	7	3.6%
80-89	11	5.8%
90-99	0	0
100	45	23.4%
Total	192	100%

It is possible to divide the above Table up into the following three categories: 70-100% having obvious or clear titles, partially clear titles 40-69% and obscure titles 0-39%.

It was observed that if the title was short it invariably led to the subject of the book being more easily retrieved . This indeed is what was observed by Al-Sawyadan (13), who recommended that titles should be as short as possible. Fisher (14) has mentioned that Kenedey has observed, the title should be ideally presented in one short concise sentence.

5.4.3 Stopwords

Stopwords are those words which carry no useful information by themselves. From the sample it was found that 193 (21%) words out of a total 917 words were stopwords. Most of these words were prepositions, conjunction, pronouns, etc. The major problem in Arabic stoplist words is ال التعريف *Al-altarif* [the definite article]. The definite article in Arabic is found in most nouns and adjectives. The difficulties associated with the definite article in Arabic are outlined below.

It is impossible to count this article as a stopword because there are two types of ال *al* :

A)It is used as a definite article such as المدرسة *al madrasah* [the school]. In this type there is no problem because when ال *al* [the] is omitted, the meaning of the word will remain as it was. The word becomes مدرسة *madrasah* [school] without ال *al* [the].

B) Sometimes ال *al* is an integral part of the word. From the sample it was found 7 words that started with ال *al*. Such as in ألمانيا *almainya*

[Germany]. If the ال *al* is omitted, it becomes مانيا *mainya.* In Arabic the word *mainya* does not mean anything. Sometimes ال *al* is placed in the middle of the word, thus الطالب <u>*al*</u> *ta<u>l</u>ib* [a student] . When ال *al* is omitted from this word, the computer will omit the first and the second ال *al*, and the word becomes طب *tib* which means medicine. This is because the computer cannot differentiate between the definite article and ال *al* which is in some cases an integral part of the word. khurshid has mentioned that KFUPM has resolved this problem by designing a list of about 200 words starting with characters similar to those of the definite article is stored in the computer so that the program will not ignore the ال *al* in these words which are counted as a part of them(15).This solution also has presented by Al Soynia(16).

In Arabic it is easy to recognise pronouns, verbs and prepositions as stopwords, while it would be impossible to define a full stop list in Arabic because some words could be keywords in some contexts while they are not in other contexts (see Section 4.4). The same opinion was mentioned by Al Soynia(17). He suggested that each library or information centre should establish their own stoplist according to their needs and the services which they offered. He (18) also listed 120 words which can be counted as stopwords in most libraries and information centres, because these words are common such prepositions, verbs, pronouns etc. Al Atram (19) has listed 202 words as stopwords which indeed have not significant meaning in most content. He (20) suggested to store these words in a computer as a stoplist.

5.4.4 Affixation

Affixation is widely used in Arabic, suffixes being most common in Arabic. On analysis of affixation data in the sample (Table 5.4) confirmed suffix prevalence.

Table 5.4: Some affixation in Arabic word

Prefixes		Infixes		Suffixes	
Type	Occurrence	Type	Occurrence	Type	Occurrence
ب	3	و	3	ية	6
ل	10	أ	1	ون	1
		ا	1	ات	24
				هم	1
				ك	4
				ي	16
				ين	3
				ة	17
				ه	1
				ها	1
				اء	3

There are three of the types of suffix create particular problems in Arabic information retrieval. These are as follows:

1. ية and يه
2. ة and ه
3. ي and ى

The two points which are above or below the characters confuse the searcher and can confuse the cataloguer who inputs the data into the computer. To avoid this problem some libraries ignore the two points above or below the characters when inputting cataloguing and this is not correct in Arabic grammar. The problem sometimes also arises from carelessness when these characters are entered into the computer. Some libraries have a policy to ignore the two points in all cases to overcome this problem such as in KFCRIS. Ali (21) was suggested that the word in question should be returned to the root, in order to list all the definitions of the root. Some researchers such Al Atram (22) did not believe that there is a relationship between referring the word to the root and information retrieval because some words could have the same root, but with the affixation become another word such as the root كتب *kutab* [write]. With the prefix م *maim* [m] gives another meaning, thus مكتب *maktab* [an office] (see Section 3.3.3). This issue is not yet clear for this author, because it needs more searching to decide which method is best by using the root or without using the root. Although there is a study was done by Al Khrashi and Evans (23) reached that the using stem better than the using root or even the word itself.

5.4.5 Connected particles

As outlined in Section 3.2, some particles in Arabic are prefixed with some words. From the sample it was found that there are 12 words which were prefixed with the particle ل *lam* [l], while 6 words were prefixed with the particle ب *baa* [b]. Because the two particles joined the original word, users need to use the right truncation to omit these particles or some items will not be retrieved. For example, the word للكتب *lell kutub* [for books] if the truncation is used the system will retrieve *kutub* [books], *lell kutub*[for books], *bell kutub* [in the books], *wall kutub* [and books], etc. This issue could be solved by offering the truncation in the system which will help users to choose what form of the word they want to search for.

5.4.6 Singular and plural

As would be expected, the sample contained words in both the singular and the plural form. This is problematic with regard to searching as a distinction between these two forms of word expressing the same concept could not be made by the computer. As seen in Section 3.3.2, sound plural marks are suffixed with the end of the word. Therefore, the left truncation can omit the suffixes, then the user can retrieve both the singular and the plural of the word. The most difficult issue is with broken plural, which has different measures (see Section 3.3.2.3.2). Therefore the truncation cannot be effective with this type of plural, because sometimes the measure mark could be in the middle or at the beginning of the word or both. Even sometimes a letter is omitted from the singular to make a plural, e.g. حجرة *hujrah* [a room], the plural of this word is حجر *hujar* [rooms].

5.4.7 Synonyms

It is well known that natural language searching is dependent upon the words which appear in the documents without modification. This leads to many problems (see Section 3.2.5). Synonyms are amongst the problems which lead to a decrease in recall. From the sample of 917 words it was found that 40 (4.4%) synonyms or near-synonyms. Because computers do information retrieval by matching characters, not the concepts, users must try to anticipate all possible words and phrases that might have been used to express the concept of interest, or the retrieval results will suffer from missed information. One solution to this problem is to have a synonym dictionary in the computer in order to control these words, i.e. to have a thesaurus online which links non-descriptors, thus enabling higher recall.

5.5 Conclusion

This chapter has attempted to evaluate the efficiency of using natural language for information retrieval in Arabic. The evaluation was done by examining a sample of Arabic book titles were selected from KFNL database. A comparison was made between the words in titles with subject headings which were given to these titles to define whether these titles are significant or not. The results showed that 42.1% were significant words, while the non-significant words were 57.9% of the total. From the sample it was found that 26 titles were obscure. The results showed that their is a relation between the efficiency of the information retrieval and Arabic language structure.

REFERENCES

1. Al Atram, Mohammad. *Kafat al lughat al tabiayah fi takshif wa estrjaa alwatharqe al arabiah* [the effectiveness of natural language for indexing and retrieving the Arabic documents], 1989.

2. Al Soynia, Ali. *Kashafat al tabadel wa estrjaa al malomat fi al lughat al arabiah* [permuted indexes and information retrieval in Arabic language], 1988.

3. Al Sawaydan, Nasser. *al estrjaa al mawdhey bewasita kalemat al enwan* [subject retriving by title `s words]. In: Proceedings of Symposium on Using Arabic Language in Information Technology, 1992, pp 533-568.

4. Kasem, Hishmat. *kashaf al kalemat al miftahia fi al siyak wa ehtimalatih fi al lughat al arabiah* [KWIC index and its application in Arabic linguage]. Alam Al kutub [World of Books], 1984, 5 (4), 638-650.

5. A1. Sawaydan,ref.3, pp.534-536.

6. Feinberg, Hilda. *Title derivative indexing techniques: a comparative study.* 1973, p.33.

7. Kraft, Donald. A comparison of Keywords in Context (KWIC) indexing of titles with a subject heading classification system. *American Documentation.* 1964, 15 (1), 49.

8. Al Atram, ref.1, p.3-20.

9. *Ibid.*

10. Al Sawaydan, ref.3, p.567.

11. Al Atram,ref.1, p.2-18.

12. *Ibid.*

13. Al Sawaydan, ref.3, p.568.

14. Fischer, Marguerite. The KWIC Index Concept: a retrospective view. *American Documentation,* 1966, 17 (2), 66.

15. Khurshid, Zahiruddin. Arabic online catalogue. *Information Technology and Libraries*, 1992, 11 (3), 251.

16. Al Soynia, ref.2, p.55.

17. *Ibid.*

18. *Ibid.*

19. Al Atram, ref.1, p.5-3.

20. *Ibid*

21. Ali, N. *al lughat al arabiah wa al hasub* [Arabic language and computers], 1988, p.324.

22. Al Atram, ref.1, p.4-13.

23. Al Kharashi, Ibrahim and Martha W. Evans. Comparing words, stems, and roots as index terms for an Arabic information retrieval system. In: *al moatmar al thani: al ughat al arabiah wa al taquniat al malomatiah a*l *mutqudemah* [Second International Conferecne: Arabic Language and Advanced Information Technologhy], 1993, pp 297-310.

CHAPTER 6

SUMMARY AND CONCLUSIONS

6.1 AIMS OF THE STUDY

This study was undertaken to achieve the following aims:

1- To identify the language problem facing information retrieval from Arabic online catalogue systems. It became apparent from the study that there is a link between information retrieval and the structure of the Arabic language. This study found that the morphology, synonyms, singular and plural etc of Arabic. affects the efficiency of information retrieval. A number of suggestions were presented that might resolve some linguistics problems related to information retrieval in Arabic language.

2- The second aim of the study was the evaluation of natural language searching in Arabic book titles by using words in titles as primary access points to the documents. The study showed that the percentage of significant words was 42.1 % of the sample, while the non-significant words represented 57.9 % of total. It was concluded that this ratio is very poor.

3- The third aim was to identify the effectiveness of using Arabic language for information retrieval. Because of the time limitation of this study, it was found that this sort of study requires a far greater sized sample and also a number of an Arabic full text databases. There were not available.

6.2 SUMMARY

This dissertation can be summarized as a follow:

1- A brief description of OPACs in Saudi Arabia was provided particularly in Riyadh City, where a number of OPAC systems such as DOBIS/LIBIS and MINISIS are used. Both systems are briefly described. It was found that little has been written about OPACs in Saudi Arabia in the library literature.

2- The principal problems in Arabisation of online catalogues are the lack of standardisation and the absence of cooperation between institutions and organisations who are involved in this field. In addition to this an important point that should be kept in mind is that most software packages which are used in the Arab world were designed originally for non-Arabic languages mainly English, Naturally, this cause some problems, particularly with Arabic characters.

3- Factors affecting the efficiency of using natural language for information retrieval in Arabic language, e.g. problems of using different words to express the same concept (synonyms), variant forms of words (morphology), singular and plural etc. were described.

4- A random selection of Arabic book titles from records in the KFNL database as examined to identify the efficiency of using the words in titles as subject access points. The investigation clearly indicates that:

a- as in other language, a number of common words which are non-significant are used in Arabic book titles (e.g. "an introduction" , "history of " etc.);

b- a number of titles were obscure (i.e. they did not correspond to the subject content of the book);

c- most obscure titles were found in the "heritage books";

d- shorter titles are more significant than longer titles;

e- suffixes were the most common type of affixation which are used in Arabic words;

f- stopwords are common in Arabic titles;

g- the more technical and scientific the field, the more likely it was that the titles would be useful for retrieval;

h- the stopwords which were recommended by Al Atram and Al Soynia are abase for an Arabic stoplist,

i- there is no way of deciding an ideal Arabic stoplist;

j- truncation is a useful device to resolve affixation, particularly connected particles and singular and plural;

l- little information exists on using natural language for information retrieval in Arabic language;

6.3 SUGGESTIONS

In the light of the results which have been found in this study there are a number of suggestions may be made:

1- Arabisation should be regularised;

2- more attention should be payed to cooperation between libraries and information centres.

3- more attention also should be directed towards to the relationship between linguistics and information retrieval.

4- Arabic information retrieval needs to be improved in order to be compatible with the growth of information.

5- bibliographical standards in the Arab world should be regularised;

6- some authors are not aware of the importance of subject precision when they title their work, so a number of suggestions can be directed to them:

    a- Arab authors and editors must be conscious of information retrieval when they title their publications;
    b- titles should be describe the subject content of the book;
    c- authors should avoid obscurity in their titles;
    d- non-significant words should be avoided if possible.

If the above suggestions could be not achieved or there is no ability to control the titles, the following suggestions may be presented:

    a- authors should be encouraged to write good titles;
    b- librarians and information worker should present their suggestion to publishers and editors about the importance of titles for information retrieval purposes;
    c- significant words should be added to the obscure title to represent the subject content of the documents.

7- the problem of synonymity may be controlled by the compilation of a list synonyms and/or near- synonyms;

8- the solution to variant forms of words or connected particles can be done by offering truncation in the system.

8- the solution to variant forms of words or connected particles can be done by offering truncation in the system.

6.4 FURTHER STUDIES

A number of research studies can be suggested by this author, who believes that they would increase the ability and efficiency of Arabic information retrieval in the future:

1- research needs to be directed towards to the effectiveness of using natural language for information retrieval in Arabic language to determine the efficiency of this method for information retrieval purpose;

2- further studies and data are need to determine the value of retrieving by roots or stems and words themselves;

3- a comparative study between using concept analysis or natural language for indexing purposes is needed;

4- further in depth research to determine the relationship between linguistics and information retrieval in Arabic language would be useful;

5- further studies on Arabic full text searching are needed;

6- Arabic stoplists still need to be studied in more detail.

BIBLIOGRAPHY

Al Atram, Mohammad. *Kafat al lughat al tabiayah fi takshif wa estrjaa alwathiaqe alarabiah* [The effectiveness of natural language for indexing and retrieving the Arabic documents]. Riyadh: King Abulaziz City for Science & Technology, 1989.

Al Dosary, Fahad M. and Abdurrahman H. Ekrish. The state of automation in selected libraries and information centres in Saudi Arabia. *Libri*, 1991, 41 (2), 109-120.1993, pp.569-580.

Al Khuli, Ali. *Learn Arabic by yourself.* Riyadh: [AL-Farazdak Press], 1985.

Al Soynia, Ali. *kashafat al tabadel wa estrijaa al malomat fi al lughat al arabiah* [Permuted indexes and information retrieval in Arabic language]. Riyadh: King Fahad National Library, 1988.

Al Zeer, Mohammad, H. *alhasib alaalei fi maktbat jamiat alimam Mohammad Ibn Saud al islamiah* [computers in Imam Mohammad Islamic University Libraries].Riyadh: Imam University, 1987.

Ali, N. *al lughat al arabiah wa al hasub* [Arabic language and computers]. Kuwait: Ta'reep, 1988.

Aman, Mohammad M. Use of Arabic in computerised information interchange. *Journal of the American Society for Information Science*, 1984, 35 (4), 204-210.

Arab League Documentation Centre. *ARIS-NET Newsletter*, 1988, 2 (24), 1.

Ashoor, Mohammad Saleh. Planning for Library automation at the University of Petroleum and Minerals. *Journal of Information Science*, 1982, 5 (5), 193-198.

Ashoor, Mohammad Saleh. Arabisation of automated library systems in the Arab World: need for compatibility and standardisation. *Libri*, 1989, 39 (4), 294-302.

Bachir, Imad and Andrew Buxton. The use of topic sentences for evaluating the representativeness of Arabic article titles. *Journal of Information Science*, 1993, 19 (6), 455-465.

Boyce, Cheryl, MINISIS. *Program*, 1982, 16 (3), 131-141.

Bright, William, ed. *International encyclopedia of linguistics.* Oxford: Oxford

University Press, 1992.

Chamis, Alice Yanosko. *Vocabulary control and search strategies in online searching.* London: Greenwood, 1991.

Chaudhary, Abussattar and Mohammad Saleh Ashoor. Potential of DOBIS/LIBIS and MINISIS for automating library functions: a comparative study. *Program*, 1990, 24 (2) 109-128.

Crystal, David. *A dictionary of linguistics and phonetics.* Oxford: Blackwell, 1991.

Crystal, David. *An encyclopedic dictionary of language and languages.* Oxford: Blackwell, 1991.

Doszkocs, Tamas E. Natural language processing in information retrieval. *Journal of the American Society for Information Science*, 1986, 37 (4), 191-196.

Feinberg, Hilda. *Title derivative indexing techniques.* Metuchen: The Scarecrow Press, 1973.

Fischer, Marguerite. The KWIC Index Concept: a retrospective view. *American Documentation,* 1966, 17 (2), 57-70.

Fondationdu Roi Abdul-Aziz Al Saoud pour les Etudes Islamiques et les Sciences Humains. *al moatmar al thani: al lughat alarabiah wa al taquniat al malomatiah al mutqudemah* [Second International Conference: Arabic language and Advanced Information Technology]. Casablanca: FASEISH , 1993.

Frsony, Fuad. Searching in database. Report. Riyadh: King Fahad National Library, [n.d.].

Harter, Stephen. *Online information retrieval: concepts, principles and techniques.* Orlando: Academic Press,1986.

Hegazi, Nadia, Nabil Ali and Ehsan Abed. Information content in textual data: revisited for Arabic text. *Journal of the American Society for Information Science,* 1987, 38 (2), 133-137.

Hildreth, Charles R. To Boolean or not to Boolean?. *Information Technology and Libraries*, 1983, 2 (3), 235-237.

Ibrahim, Farid. *A syntactically based preprocessor for a limited experimental Arabic document retrieval system.*ph.D. thesis, Loughborough University of Technology, 1988.

Kapliwatzky, Jochanan. *Arabic language and grammar.* Jerusalem: Rubin Mass, 1972.

Kasem, Hishmat. *Kashaf al Kalemat al miftahia fi al siyak wa ehtimalatihfi al lughat al arabiah* [KWIC index and its application in Arabic language]. *Alam Al Kutub* [World of Books], 1984, 5 (4), 638-650.

Khurshid, Zahiruddin. Application of modern technologies in Arab Libraries. *Libri*, 1983, 33 (2), 107-112.

Khurshid, Zahiuddin. Arabic online catalogue. *Information Technology and Libraries*, 1992, 11 (3), 244-251.

King Abdulaziz Public Library. *Proceedings of Symposium on using Arabic language in Information Technology.* Riyadh: King Abdulaziz Public Library, 1992.

King Saud University. *Arabic subject headings list.* Riyadh: King Saud University, [n.d.].

Kraft, Donald. A comparison of Keywords in Context (KWIC) indexing of titles with a subject heading classification system. *American Documentation.* 1964, 15 (1), 48-52.

Lancaster, F. Wilfrid. *Information retrieval system: characteristics, testing and evaluation.* New York: John Wiley & Sons, 1979.

Longman dictionary of the English language. Harlow: Longman, 1934.

McAllister, Caryl. The online public access catalogue in DOBIS/LIBIS. *Program*, 1987, 21 (1), 25-36.

McAllister, Caryl and Stratton McAllister. DOBIS/LIBIS: an integrated online library management system. *Journal of Library Automation*, 1979, 12 (4), 300-313.

Ministry of Interior. *The Ninth National Computer Conference and Exhibition.* Riyadh: Ministry of Interior, 1986.

Montgomery, Christine A. Linguistics and Information Science. *Journal of the American Society for Information Science.* 1972, 23 (3), 195-219.

Musa, F. A. A system for processing bilingual Arabic/English text. *Journal of the American Society for Information Science*, 1986, 37 (5), 288-293.

Pritchard-Schoch. Natural language comes of age. *Online*, 1993, 17 (3), 33-43.

Shaheen, Abdulsabour. *Arabic: the language of science and technology.* [Riyadh]:[n.p.],[n.d.].

Shaikh, Shafi. *A course in spoken Arabic.*Bombay: Oxford University Press, 1978.

Smart, J. R. *Arabic.* Sevenoaks: Teach yourself Books, 1986.

Smeaton, Alan F. Prospects for intelligent, language-based information retrieval. *Online Review,* 1991, 15 (6), 373-381.

Sparck Jones, Karen and Martin Kay. *Linguistics and information science.* New York: Acamdemic press, 1973.

Warner, Amy. Natural language processing. *Annual Review of Information Science and Technology,* 1987, 22, 79-108.

Wickens, G. M. *Arabic grammar.* Cambridge: Cambridge University Press, 1980.

Young, Heartisill, ed. *The ALA Glossary of Library and Information Science.* Chaicago: American Library Association, 1983.