The Linguistic dimension in Information Sciences: An applied study to use indexing languages in the work of search engines

A Research Submitted for the fulfilment of A PhD. Degree in Library & Information Sciences (LIS)

By:
Sabri Elhaj Elmubarak Elamin

Supervised By:
Ustaz. Abu baker El siddiq Osman

2007
Acknowledgements

I am particularly grateful to Ustaz Abu baker El siddiq Osman for his supervision that he has spent a lot of time reading and guiding this study at its different stages. Special thanks are directed to my friends Mr. Malik Osman El faki; Housam Eldin Mohamed Awaed alla and Miss Sohir Abdin for their continuous assistance. Thanks also are directed to Dr. Radia Adam Mohammed, Dr. Afaf Karoom, Afaf Mohamed Elhasan and Mai Abbas Fadul, and many people who have offered advice and suggestions. Last but not least, warm thanks for my family members my mother, brother Adil, sisters special thanks for my sister Wisal for her warm kind, and for her help in typing this research, Sawsan, Tahani, Amani, Daughter of my sisters Nisreen, Shiren and Ahmed. My Wife Maisa for her continuous support, help and pray for completing this research and my daughter Nour.
Abstract

Thesis Title: the Linguistic dimension in Information Sciences: An applied study to use indexing languages in the work of search engines.

Candidates Name: Sabri Elhaj Elmubarak Elamin

Degree: PhD in information & Library Sciences
Graduates college – University of Khartoum

Supervised By: Ustaz. Abu baker Elsiddig Osman
Year: 2007

Along with increasing use of international databases, cross-cultural indexing is becoming more common and tools like multilingual thesaurus are urgently needed. This study concentrates on cultural and linguistic problems in multilingual thesaurus. Very little attention has been paid to this topic in Arabic research and the need for this kind of study has grown rapidly. This study has proved that all the engines tested had ranking schemes that were not well documented, being based on how many times search words were mentioned, whether or not they appeared early in the document, whether or not they appeared close together, and how many search terms were matched. I did not find the
ranking schemes very useful, as relevant and irrelevant pages frequently had the same scores. It is found that aggregating meaning is possible on the Internet because there are many easily accessible semantic objects to be harvested. Analysis of the aggregations can suggest patterns of high likelihood that permit applications to recommend, adapt, profile, forecast and so on. It has been proved that the terminology and thesaurus construction standards and guidebooks provide very little details and consideration about equivalence. It has been indicated that Google may index billions of Web pages, but it will never exhaust the store of meaning of the Web. The reason is that Google's aggregation strategy is only one of many different strategies that could be applied to the semantic objects in public Web space. It has been found that web search engines do not conspire to suppress controversy, but their strategies do lead to organizationally dominated search results depriving searchers of a richer experience and, sometimes, of essential decision–making information. These experiments suggest that bias exists, in one form or another, on the Web and should, in turn, force thinking about content on the Web in a more controversial light
المستخلص

Joshua 1:13

I am just as confident as you are, and I am not afraid. I am not afraid, because the Lord is with me.

ى سأتبعك، لأنه معني.

V
تمكن يعتبر الكلي المعنى أنه البحث ي وجود على الإنترنت مع الكوادترات تيسيل الدالالة مشاركة ووجود بسبب وذلك إلى الكلي ي، إلئ الى اضافة أنماطية يقدمنا الكلي التحليل أنه في تساهيم تطبيقا تتيح التشبّه كبيرة وموضوعات الدارسة أثبّت ذلك وبجانب وتبدو التوافق حول تذكّر تفصّلات تقدّم لا الأدلة بجانب الكائنات بناء.

وقد ذلك البحث محرك أنه نحن عليه أنتان (بلاين) يكشف انترنت الشبكة على ذلك لأن الدايرتين تنتهي البحث المحرك الكلي الاستراتيجية بناء ذلك وثبت ذلك قولة (يعد من فقط واحده الشبكة مساحة في الدالية الأهداف على تطبيقه يمكن كثيرا الاستراتيجيّات العامة).

ولكن الشبكة في البحث محركات أن إلى الدارسة توصلت ذلك وبجانب حداه يتحمل الرقم بجانب بحث تيجة إلى استراتيجيةها وأن القرار صنع معلومات به ينادي وبالتالي الشبكة في حيدة أكبر.
Content

Acknowledgments ………………………………………………………( II)
Abstract……………………………………………………………………(III)
Table of Contents…………………………………………………………..(V)
List of Tables………………………………………………………………..(XI)
List of Figures………………………………………………………………..(XIII)

Chapter one

Introduction & methodology:

Introduction: .................................................................(2)

1. Statement of Research problem & Objectives…………………..(3)

1.1 Equivalence………………………………………………….(3)

1.2 Aims, Methods and Materials………………………………….(4)

1.2.1 Aims………………………………………………….(4)

1.2.2 Methods………………………………………………….(4)

1.2.3 Main tools for analysis at the indexing term level……….(6)

1.2.4 Main tools for analysis at conceptual level………………..(7)

1.3 Material……………………………………………………….(8)

1.3.1 Research questions and hypotheses………………………..(9)

1.3.2 Significance of the results………………………………….(11)

1.3.3 Research Definitions………………………………………..(12)
1.4 Structure of the study………………………………………………(12)

1.5 Perception Studies: A review of Trends…………………………(14)

Chapter Two
The Development of Classical Arabic :

2.1 Introduction …………………………………………………………(17)

2.2 The development of Orthography ………………………………(18)

2.3 The standardization of the language ……………………………(24)

2.4 The development of An Arabic Literary style …………………….……(36)

2.5 The official status of Arabic ………………………………………(43)

chapter Three
Literary and Modern Standard Arabic :

3.1 Introduction…………………………………………………………(47)

3.2 Dialectical Phonologies…………………………………………..(49)

3.3 Grammar………………………………………………………………(50)

3.4 Arabic alphabet:

3.5.1 Computers and the Arabic alphabet…………………………(67)

3.5.2 Arabic Chat Alphabet: Online…………………………………(69)

Chapter Four
Arabic Language Features :

4.1 Introduction……………………………………………………………(74)
4.2 Contextual Analysis ................................................. (78)
4.3 Design of BiDi Software ........................................... (84)
4.4 Technical Issues in Developing BiDi Software .............. (93)

Chapter Five:

Arabic on the Internet:

5.1 History of Arabic on Computers................................. (101)
5.2 Microsoft and Arabization....................................... (106)
5.3 The Issue of Platform and Browser Independence......... (109)
5.4 Transliteration: Using Latin Characters for Arabic....... (111)
5.5 Arabic on the Web .................................................. (112)

Chapter Six:

Information Retrieval & The using of file structures in document retrieval:

6.1 Introduction........................................................... (116)
6.2 Effectiveness and efficiency ................................. (122)
6.3 Automatic Text analysis ........................................... (123)
6.4 Indexing ............................................................... (124)
6.5 Automatic keyword classification........................... (131)
6.6 A language for describing file structures................... (133)
6.7 Sequential files ...................................................... (134)
6.8 Index-sequential files.................................................(136)
6.9 Scatter storage .........................................................(137)
6.10 Clustered files ..........................................................(142)

Chapter Seven:
Evaluation of Information Retrieval System:
7.1 Introduction .............................................................(144)
7.2 Relevance ...............................................................(149)
7.3 Composite measures ................................................(151)
7.4 The SMART measures ..............................................(152)
7.5 A normalized symmetric difference.........................(156)

Chapter Eight:
Web search engines & suppressing controversy:
8.1 Introduction .............................................................(172)
8.2 How Software Agents and Search Engines Work .........(173)
8.3 Search Engine Features..............................................(179)
8.4 The nature of meaning in the age of Google ..............(194)
8.5 Aggregating Meaning ...............................................(197)
8.6 Information and spam..............................................(200)
8.7 Traditional methods of constructing meaning…………………..(205)

8.8 Understanding Web behavior ………………………………..(214)

8.9 Why does visibility of controversy matter?…………………..(218)

8.10 The Organizational Web dominates the Analytic Web………(230)

8.11 A simulated objective Web……………………………………..(233)

8.12 Appendixes ……………………………………………………..(237)

Chapter Nine:

Summaries, Conclusion and Recommendations…………………..(254)

Bibliography…………………………………………………………(258)
List of Tables

3.1 Personal pronouns ............................................ (57)
3.2 Enclitic pronouns ............................................ (57)
3.3 Prefix and Suffix ............................................. (62)
6.1 Data Retrieval (DR) Information Retrieval (IR) ............ (117)
7.1 Search Engine Test Results .................................... (190)
8.1 Google visits to the Brooks Website .......................... (204)
8.2 Broad topics and controversial subtopics ....................... (224)
8.3 Broad topics and other subtopics (Sample URLs) ........... (225)
8.4 Comparison of Revealing and Deep (Controversy, Broad Topic) ......................................................... (246)
8.5 Appendix B.4: Search engine characteristics for all URLs (Simple, Controversy, Factual) ........................................ (251)
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Logical –to- physical Transformation</td>
<td>86</td>
</tr>
<tr>
<td>4.2</td>
<td>Semantic Inversion of Insertion</td>
<td>76</td>
</tr>
<tr>
<td>4.3</td>
<td>Contextual Analysis</td>
<td>79</td>
</tr>
<tr>
<td>4.4</td>
<td>Effect of Diacritics on the meaning of a word</td>
<td>80</td>
</tr>
<tr>
<td>4.5</td>
<td>Vertical Kerning of Diacritics</td>
<td>81</td>
</tr>
<tr>
<td>4.6</td>
<td>Arabic Ligatures</td>
<td>82</td>
</tr>
<tr>
<td>4.7</td>
<td>Traditional Arabic font</td>
<td>83</td>
</tr>
<tr>
<td>4.8</td>
<td>Same Number with different Representations</td>
<td>83</td>
</tr>
<tr>
<td>4.9</td>
<td>Arabic strung display with numbers</td>
<td>84</td>
</tr>
<tr>
<td>4.10</td>
<td>Latin Insertion in Arabic line</td>
<td>85</td>
</tr>
<tr>
<td>4.11</td>
<td>Logical &amp; Visual Block selections</td>
<td>87</td>
</tr>
<tr>
<td>4.12</td>
<td>Kashida Justification</td>
<td>88</td>
</tr>
<tr>
<td>6.1</td>
<td>A diagram illustrating Thomen condition</td>
<td>161</td>
</tr>
<tr>
<td>6.2</td>
<td>The diagram show the relative position of points</td>
<td>166</td>
</tr>
<tr>
<td>7.1</td>
<td>Search Engine Test Results</td>
<td>190</td>
</tr>
<tr>
<td>8.1</td>
<td>Combine repeated visits of googlebot</td>
<td>204</td>
</tr>
<tr>
<td>8.2</td>
<td>Broad Topics &amp;controversial subtopics</td>
<td>224</td>
</tr>
</tbody>
</table>
8.3 Broad Topics & controversial subtopics ...........(225)
8.4 The Results of Queries from all engines ..........(243)
8.5 Search Engine characteristics for all URLs .......(251)
Introduction

The rapid development of modern societies changes language and terminologies in the social sciences. Together with a widening globalization, this increases the need for high quality documentation tools for locating up to date information from multilingual resources.

On-line databases and Internet have become common and information retrieval and seeking over national borders are increasing all the time. The success of creating and using international information resources depends on common tools and understanding of the concepts used.

Along with increasing use of international databases, cross-cultural indexing is becoming more common and tools like multilingual thesauri are needed. This study concentrates on cultural and linguistic problems in multilingual thesauri. Very little attention has been paid to this topic in Arabic research and the need for this kind of study has grown rapidly.

Traditionally, a thesaurus can be defined as a documentary language and it is developed to be a tool in information
retrieval and documentation. Terms in a thesaurus have many equivalents in natural language - especially in the spoken language. Also the manner of representation is artificial - hierarchical and associative relationships being the context of the thesaurus terms. This kind of representation aims at helping an information seeker to broaden or specify the search.

Guidebooks and standards for thesauri construction are very strongly focused on linguistic matters when defining a good indexing term. - A good term should be transparent, consistent, practical, brief, productive, easy to pronounce and spell, and unambiguous. It should also otherwise be linguistically suitable and preferably based on one's own language. Very little attention has been paid to cultural and semantic aspects. Nevertheless, constructing a multilingual thesaurus demands operating distinctly in a multicultural environment.

1. Statement of Research problem and Objective

1.1 Equivalence
In multilingual and multicultural terminology work one of the key problems is equivalence. In defining equivalence level a search engine constructor tries to anticipate what kind of term most likely will be used in information seeking and retrieval.

what kind of term most likely will be used in information seeking and retrieval.

Usually equivalence means similarity, (perfect) equality. Terminology and thesaurus construction standards and guidebooks provide very little details and consideration about equivalence. In International Organization for Standardization's standards *Documentation - Guidelines for the establishment and development of multilingual thesauri* equivalence is divided into *exact equivalence, inexact equivalence, partial equivalence, single-to-multiple equivalence and non-equivalence*.

In translation studies equivalence is understood in many different ways, but a general consensus today seems to be that there is no sense in demanding "mirror-translation". Still even today translating is not seen as an easy act. A translator is commonly seen as a "prisoner" of his/her own culture.
Eugene Nida and William Reyburn (1981) have found, that a translator usually understands the message in accordance with his/her own cultural-linguistic context. Usually a translator is aware of that and tries to solve the problem by using foreign terms. Regardless of that a concept in the source language is not always semantically equivalent with the same concept in the target language. Generally descriptive phrases are therefore better (more equivalent) than foreign terms in translating a message into another culture. This can also be seen in thesaurus construction standards and guidelines where loan words are not recommended.

2. Aim, Methods and Material

2.1 Aim
The focus of this study is the translatability of English social science indexing terms into Arabic language and culture at a conceptual, term and indexing term level. The perspectives are both linguistic and sociological - a combination through which a broader understanding of the phenomena is being aimed at in the general frame of information science.
2.2 Methods

This study is quantitative-qualitative. Qualitatively is emphasized in developing the analysis method for this specific research material. Traditional terminology sources will be supplemented by the theories and concepts of translation studies, linguistics and communication theories in order to achieve a broader and more pragmatic perspective. Thus, a new research method can be developed. Linguistic and cultural matters cannot be clearly separated, because they go hand in hand. Therefore, in this study concentrating on the cultural or linguistic matters is more like an emphasis or a perspective.

The study uses multiple cases aiming at theoretical replication. It is thus an empirical case study and in its nature descriptive, generating hypotheses and illustrating a new theory of "pragmatic indexing equivalence". The samples studied are theoretical (not random) and thus cases assumed to give answers to the questions of this study.

Several data collection and analysis methods will be used in order to construct a theory by triangulation of evidence. The research process is thus linear starting from hypotheses
(statements) and ending through observations and generalizations at new hypotheses and a construction of a theory.

This study examines Arabic equivalents for British-English indexing terms. The key terms of the research corpus will be studied:

- 1) at a conceptual level by interviewing people representing different cultures (Arabic versus British) and subcultures (indexers, specialists, multilingual thesaurus constructors);
- 2) by a modification of co-word analysis (the use and the indexing practice of the key terms studied in Arabic, British and international databases) and;
- 3) by comparing several thesauri.

2.2.1 Main tools for analysis at the indexing term level

Reasons and motivations for the chosen equivalents will be studied by interviewing the specialists, thesaurus constructors and indexers.
Comparing equivalents explanations and datasets are explored by classification and dividing terms into different groups. Comparisons inside and between these groups are made due to e.g. their type of equivalence, nativeness in the Arabic language and culture-boundness.

Problems will be analyzed in order to make a classification for these problematic terms. Difficulties can be caused by, for example:

- **(1) Society:**
  The same kind of phenomenon or a concept as in British society does not exist in the Arabic society. *(homemaker)*

- **(2) Language:**
  The concept exists in both cultures, but to find a term to describe it conforming Arabic common language usage is problematic. *(illegitimate births)*

- **(3) Euphemisms:**
  Socially difficult concepts are described with euphemisms. *(family disorganization)*
• **(4) Institutional differences:**
  To express the concept conforming Arabic indexing language thesaurus constructors have to use factoring. *(married women workers)*

• **(5) Citation loans, foreign words:**
  There is not a Arabic native word to describe the concept and/or at term level a citation loan or foreign word is commonly established. *(nationalismi)*

Arabic, British and multicultural indexing practice will be studied and compared with the key terms of the corpus. Indexing frequency of the key terms of this study and their equivalents will be studied by examining the use of terms and their synonyms in several databases using a modification of a co-word analysis.

**2.2.2 Main tools for analysis at conceptual level**

To achieve a broader perspective and a better understanding of the content of the key terms, people representing different cultures and subcultures will be interviewed. The aim is to find out the tacit assumptions and practices, which are linked with the key terms. Terminologies and thesauri
usually operate at the denotation level of the words, but human communication - both formal and informal - uses also connotative meanings.

The recognition of the equivalence type is in this research done by component analysis. It is commonly used by translators and known also as denotation analysis. In component analysis the meaning of the word (denotation and connotation) is divided into smaller components, semantic characteristics. A component analysis is helpful especially in polysemy cases and in co-ordinated and related cases.

In component analysis one can use, for instance, a semantic characteristics matrix, where words are defined by semantic characteristics. With a matrix we can recognize the common characteristics of different words and the distinctive components. In this study component analysis will be used especially in defining the meaning of synonymous or near synonymous terms.

2.3 Material

The terms for this study are selected according to three criteria:
1. There exist for one English term in Arabic several (quasi-)synonyms
2. There is no exact Arabic equivalent for an English term
3. A theme - terms related to gender.

The selected databases in order to get pools of documents about the topics in the above-mentioned groups are:
1) Reference Database for Arabic Articles ARTO and Arabic Virtual Library and;

The thesauri used to examine and compare the indexing terms used in the received documents in the database searches are: The UNESCO Thesaurus, Eurovoc, The OECD Macrothesaurus, ELSST, ERIC, Thesaurus of Sociological Indexing Terms, HASSET and YSA.

Arabic and British indexers will be asked to index the same documents, that is, articles retrieved from Arabic and British databases having the studied terms as one of the central ones (e.g. in title and/or as a descriptor and thus likely to be used in indexing).
3. Research questions and hypotheses:

Translatability is examined by trying to answer the following research questions:

- **Equivalence:**

  What kind of equivalence type is aimed and is that kind of equivalence possible to achieve in a documentary language translation?

  How do indexing term equivalence and concept equivalence differ from each other?

  What kinds of terms are considered problematic?

  Do equivalence problems mainly stem from the nature of language and/or culture or from the nature of the documentary language or domain?

  How are terms understood and defined by people representing different cultures and subcultures?

  In what way do equivalents given by specialists and indexers differ from each other?

- **Indexing practices:**
How do indexing practices differ from each other in Arabic, British and international databases?

How do Arabic and British indexers index the same document?

How are the terms studied represented in well-known and widely used thesauri?

The central hypotheses (statements) of this research are:

(1) Intercultural misunderstandings are often caused by - Subconscious blinders. People are not aware of their own assumptions and their cultural basis.

- The lack of cultural self-awareness. It's a common belief that the main challenge in cross-cultural communication is to know the foreign culture, when in fact it is to know our own culture and how it affects our behavior and thinking.

- Projected similarity. In real life other people differ more from oneself than assumed. Another person's situation differs also more than is usually assumed. Differences can then be expected, imagined and discovered as similarities.

- Cross-cultural misevaluation. We use our own culture as a standard of measurement.
(2) Languages and ideas expressed by languages are created mainly in accordance with expression needs of the surrounding culture and thus, a thesaurus can be seen as "a cultural product". The content of a thesaurus depends on the documentary needs of the datasources in the surrounding culture. Arabic indexers are guided by the Arabic culture and its traditions.

(3) Translation problems of Arabic and non-Arabic concepts occur for instance on the basis of the thesaurus' (hierarchical) structure, the differences between languages and cultures, the time factors which influence the content of concepts.

(4) Indexing terms are part of a documentary language and because of this they can sometimes greatly differ from natural language and common usage. Specialists' and indexers' equivalents differ from each other. When selecting an indexing term indexers are more likely to take into consideration the standards' and guidebooks' requirements for a good indexing term.

(5) Indexing term equivalence and concept equivalence are not necessarily the same. We can have concept equivalence
between a Arabic and a British concept, but at the indexing term level equivalence levels differ from each other and vice versa.

4. Significance of the results

Research on multilingual thesauri on the internet is useful for all the aspects of production, management, and use of digital information resources. This research aims at defining the nature of multicultural and multilingual terminology work and clarifying traditional standards and guidebooks for multilingual thesaurus construction and multicultural content management. A cross-domain approach will bring new perspectives to the research field and helps to develop the research analysis method by borrowing expressive and useful concepts and tools especially from translation and communication studies.

Research results will be useful in a large variety of different fields - from private and public sectors to the developers and actors in global Internet communication. The project has also novelty value describing a small language area's culture and language bound problems in development and studies of multilingual indexing tools.
5. Research definitions:

For this research different terms and terminologies have been used. To avoid any ambiguity or / and misunderstanding, some definitions were provided where the term is mentioned for the first time for the first time in the research text.

6. Research structure:

This research has been structured into nine chapters.

Chapter one is devoted to introduction and Methodology framework. It includes introduction statement of the problem, justification and hypothesis of the research objective of the study, research methodology and structure of the study.

Chapter two and chapter three provides theoretical framework: these chapters’ deals with three topics connected with the process of standardization. The most important prerequisite for the written codification of the language was the invention of an orthography, or rather the adaptation of existing scribal practices to the new situation.
Chapter four: is devoted to Structure of the Arabic alphabet, Arabic Chat Alphabet: Online and Arabic language features: contextual analysis, design of BiDi Software, and technical Issues in developing BiDi Software.

Chapter six discusses the nature, history and current characteristics of Language Engineering, which is contrasted with Natural Language Processing and Computational Linguistics, and shown to have attained its own distinct identity in recent years.

Chapter Seven: devoted to Structure of Arabic on the Internet: History of Arabic on Computers, Microsoft and Arabization, the issue of platform and browser Independence, transliteration: Using Latin Characters for Arabic and Arabic on the Web.


Chapter eleven and Chapter twelve provides analytical framework: The using of file structures in document...
retrieval: Logical or physical organization and data independence, a language for describing file structures, Sequential files, Index-sequential, scatter storage and Clustered files.

**Chapter Thirteen** provides analytical framework: evaluation of Information Retrieval System: relevance, Precision and recall, averaging techniques, composite measures the SMART measures, a normalized symmetric difference.


7. **Perception Studies: A review of Trends:**
Chapter one
Introduction & methodology
Chapter one

Introduction & methodology

1.1 Introduction

The rapid development of modern societies changes language and terminologies in the social sciences. Together with a widening globalization, this increases the need for high quality documentation tools for locating up to date information from multilingual resources.

On-line databases and Internet have become common and information retrieval and seeking over national borders are increasing all the time. The success of creating and using international information resources depends on common tools and understanding of the concepts used.

Traditionally, a thesaurus can be defined as a documentary language and it is developed to be a tool in information retrieval and documentation. Terms in a thesaurus have many equivalents in natural language - especially in the spoken language. Also the manner of representation is artificial - hierarchical and associative relationships being the context of the thesaurus terms. This kind of representation aims at helping an information seeker to broaden or specify the search.

Guidebooks and standards for thesauri construction are very strongly focused on linguistic matters when defining a good indexing term. - A good term should be transparent, consistent, practical, brief, productive,
easy to pronounce and spell, and unambiguous. It should also otherwise be linguistically suitable and preferably based on one's own language. Very little attention has been paid to cultural and semantic aspects. Nevertheless, constructing a multilingual thesaurus demands operating distinctly in a multicultural environment.

1. Statement of Research problem and Objective

1.1 Equivalence

In multilingual and multicultural terminology work one of the key problems is equivalence. In defining equivalence level a search engine constructor tries to anticipate what kind of term most likely will be used in information seeking and retrieval. Usually equivalence means similarity, (perfect) equality. Terminology and thesaurus construction standards and guidebooks provide very little details and consideration about equivalence. In International Organization for Standardization's standards Documentation - Guidelines for the establishment and development of multilingual thesauri equivalence is divided into exact equivalence, inexact equivalence, partial equivalence, single-to-multiple equivalence and non-equivalence.

In translation studies equivalence is understood in many different ways, but a general consensus today seems to be that there is no sense in demanding "mirror-translation". Still even today translating is not seen as an easy act. A translator is commonly seen as a "prisoner" of his/her own
culture. Eugene Nida and William Reyburn (1981) have found, that a translator usually understands the message in accordance with his/her own cultural-linguistic context. Usually a translator is aware of that and tries to solve the problem by using foreign terms. Regardless of that a concept in the source language is not always semantically equivalent with the same concept in the target language. Generally descriptive phrases are therefore better (more equivalent) than foreign terms in translating a message into another culture. This can also be seen in thesaurus construction standards and guidelines where loan words are not recommended.

1.2 Aim, Methods and Material

1.2.1 Aim

The focus of this study is the translatability of English social science indexing terms into Arabic language and culture at a conceptual, term and indexing term level. The perspectives are both linguistic and sociological - a combination through which a broader understanding of the phenomena is being aimed at in the general frame of information science.

1.2.2 Methods

This study is quantitative-qualitative. Qualitativness is emphasized in developing the analysis method for this specific research material. Traditional terminology sources will be supplemented by the theories and concepts of translation studies, linguistics and communication theories in order to achieve a broader and more pragmatic perspective. Thus, a new
research method can be developed. Linguistic and cultural matters cannot be clearly separated, because they go hand in hand. Therefore, in this study concentrating on the cultural or linguistic matters is more like an emphasis or a perspective.

The study uses multiple cases aiming at theoretical replication. It is thus an empirical case study and in its nature descriptive, generating hypotheses and illustrating a new theory of "pragmatic indexing equivalence". The samples studied are theoretical (not random) and thus cases assumed to give answers to the questions of this study.

Several data collection and analysis methods will be used in order to construct a theory by triangulation of evidence. The research process is thus linear starting from hypotheses (statements) and ending through observations and generalizations at new hypotheses and a construction of a theory.

This study examines Arabic equivalents for British-English indexing terms. The key terms of the research corpus will be studied:

- **1)** at a conceptual level by interviewing people representing different cultures (Arabic versus British) and subcultures (indexers, specialists, multilingual thesaurus constructors);
- **2)** by a modification of co-word analysis (the use and the indexing practice of the key terms studied in Arabic, British and international databases) and;
- **3)** by comparing several thesauri.
1.2.3 Main tools for analysis at the indexing term level

Reasons and motivations for the chosen equivalents will be studied by interviewing the specialists, thesaurus constructors and indexers.

Comparing equivalents explanations and datasets are explored by classification and dividing terms into different groups. Comparisons inside and between these groups are made due to e.g. their type of equivalence, nativeness in the Arabic language and culture-boundness.

Problems will be analyzed in order to make a classification for these problematic terms. Difficulties can be caused by, for example:

- **(1) Society:**
  The same kind of phenomenon or a concept as in British society does not exist in the Arabic society. (*homemaker*)
- **(2) Language:**
  The concept exists in both cultures, but to find a term to describe it conforming Arabic common language usage is problematic. (*illegitimate births*)
- **(3) Euphemisms:**
  Socially difficult concepts are described with euphemisms. (*family disorganization*)
- **(4) Institutional differences:**
  To express the concept conforming Arabic indexing language thesaurus constructors have to use factoring. (*married women workers*)
Citation loans, foreign words:
There is not an Arabic native word to describe the concept and/or at term level a citation loan or foreign word is commonly established. (*nationalism*)

Arabic, British and multicultural indexing practice will be studied and compared with the key terms of the corpus. Indexing frequency of the key terms of this study and their equivalents will be studied by examining the use of terms and their synonyms in several databases using a modification of a co-word analysis.

**1.3.4 Main tools for analysis at conceptual level**

To achieve a broader perspective and a better understanding of the content of the key terms, people representing different cultures and subcultures will be interviewed. The aim is to find out the tacit assumptions and practices, which are linked with the key terms. Terminologies and thesauri usually operate at the denotation level of the words, but human communication - both formal and informal - uses also connotative meanings.

The recognition of the equivalence type is in this research done by component analysis. It is commonly used by translators and known also as denotation analysis. In component analysis the meaning of the word (denotation and connotation) is divided into smaller components, semantic characteristics. A component analysis is helpful especially in polysemy cases and in co-ordinated and related cases.
In component analysis one can use, for instance, a semantic characteristics matrix, where words are defined by semantic characteristics. With a matrix we can recognize the common characteristics of different words and the distinctive components. In this study component analysis will be used especially in defining the meaning of synonymous or near synonymous terms.

1.3 Material

The terms for this study are selected according to three criteria:

1. There exist for one English term in Arabic several (quasi-)synonyms
2. There is no exact Arabic equivalent for an English term
3. A theme - terms related to gender.

The selected databases in order to get pools of documents about the topics in the above-mentioned groups are:

1) Reference Database for Arabic Articles ARTO and Arabic Virtual Library and;

The thesauri used to examine and compare the indexing terms used in the received documents in the database searches are: The UNESCO Thesaurus, Eurovoc, The OECD Macrothesaurus, ELSST, ERIC, Thesaurus of Sociological Indexing Terms, HASSET and YSA.
Arabic and British indexers will be asked to index the same documents, that is, articles retrieved from Arabic and British databases having the studied terms as one of the central ones (e.g. in title and/or as a descriptor and thus likely to be used in indexing).

1.3.1 Research questions and hypotheses:

Translatability is examined by trying to answer the following research questions:

- **Equivalence:**

  **What** kind of equivalence type is aimed and is that kind of equivalence possible to achieve in a documentary language translation?

  **How** do indexing term equivalence and concept equivalence differ from each other?

  **What** kinds of terms are considered problematic?

  **Do** equivalence problems mainly stem from the nature of language and/or culture or from the nature of the documentary language or domain?

  **How** are terms understood and defined by people representing different cultures and subcultures?

  **In what** way do equivalents given by specialists and indexers differ from each other?
- Indexing practices:

**How do** indexing practices differ from each other in Arabic, British and international databases?

**How do** Arabic and British indexers index the same document?

**How** are the terms studied represented in well-known and widely used thesauri?

The central **hypotheses** (statements) of this research are:

(1) Intercultural misunderstandings are often caused by - Subconscious blinders. People are not aware of their own assumptions and their cultural basis.

- The lack of cultural self-awareness. It's a common belief that the main challenge in cross-cultural communication is to know the foreign culture, when in fact it is to know our own culture and how it affects our behavior and thinking.

- Projected similarity. In real life other people differ more from oneself than assumed. Another person's situation differs also more than is usually assumed. Differences can then be expected, imagined and discovered as similarities.

- Cross-cultural misevaluation. We use our own culture as a standard of measurement.

(2) Languages and ideas expressed by languages are created mainly in accordance with expression needs of the surrounding culture and thus, a
A thesaurus can be seen as "a cultural product". The content of a thesaurus depends on the documentary needs of the datasources in the surrounding culture. Arabic indexers are guided by the Arabic culture and its traditions.

(3) Translation problems of Arabic and non-Arabic concepts occur for instance on the basis of the thesaurus' (hierarchical) structure, the differences between languages and cultures, the time factors which influence the content of concepts.

(4) Indexing terms are part of a documentary language and because of this they can sometimes greatly differ from natural language and common usage. Specialists' and indexers' equivalents differ from each other. When selecting an indexing term indexers are more likely to take into consideration the standards' and guidebooks' requirements for a good indexing term.

(5) Indexing term equivalence and concept equivalence are not necessarily the same. We can have concept equivalence between Arabic and a British concept, but at the indexing term level equivalence levels differ from each other and vice versa.

1.3.2 Significance of the results

Research on multilingual thesauri on the internet is useful for all the aspects of production, management, and use of digital information resources. This research aims at defining the nature of multicultural and multilingual terminology work and clarifying traditional standards and
guidebooks for multilingual thesaurus construction and multicultural content management. A cross-domain approach will bring new perspectives to the research field and helps to develop the research analysis method by borrowing expressive and useful concepts and tools especially from translation and communication studies.

Research results will be useful in a large variety of different fields - from private and public sectors to the developers and actors in global Internet communication. The project has also novelty value describing a small language area's culture and language bound problems in development and studies of multilingual indexing tools.

1.3.3 Research definitions:

For this research different terms and terminologies have been used. To avoid any ambiguity or / and misunderstanding, some definitions were provided where the term is mentioned for the first time in the research text.

1.4 Research structure:

This research has been structured into nine chapters.

Chapter one is devoted to introduction and Methodology framework. It includes introduction statement of the problem, justification and hypothesis of the research objective of the study, research methodology and structure of the study.
Chapter two and chapter three provides theoretical framework: these chapters’ deals with three topics connected with the process of standardization. The most important prerequisite for the written codification of the language was the invention of orthography, or rather the adaptation of existing scribal practices to the new situation.

Chapter four: is devoted to Structure of the Arabic alphabet, Arabic Chat Alphabet: Online and Arabic language features: contextual analysis, design of BiDi Software, and technical Issues in developing BiDi Software.

Chapter five: devoted to Structure of Arabic on the Internet: History of Arabic on Computers, Microsoft and Arabization, the issue of platform and browser Independence, transliteration: Using Latin Characters for Arabic and Arabic on the Web.


Chapter Seven provides analytical framework: evaluation of Information Retrieval System: relevance, Precision and recall, averaging techniques, composite measures the SMART measures, a normalized symmetric difference.

Chapter Nine has been devoted to provide some summaries, conclusion, and the study recommendations. The bibliography and some relevant annexes have been attached at end of the thesis.

1.5 Perception Studies: A review of Trends:

(1) The Linguistic Dimension of Terminology: Principles And Methods of Term Formation. By:

Kostas Valeontis:
Physicist-Electronic Engineer, President of ELETO

Elena Mantzari:
Linguist-Researcher, Deputy Secretary General of ELETO

In this Study, authors focus on the presentation of general linguistic principles concerning term formation, during primary naming of an original concept in a source language and secondary term formation in a target language. Special reference will be made to the application of these principles in the Greek language. Linguistic aspects of term formation are of major interest to terminologists, terminographers and subject field specialists, but also to translators, interpreters and technical writers; specially when translators happen to work with less widely used
languages such as Greek, where the lack of adequately developed reference tools such as specialized dictionaries and glossaries very often compels them to become neologists.

(2) A Study of the Linguistic Dimension of Information Processing as a Function of Cognitive Complexity

Gary F. Soldow (Baruch College, The City University of New York)

The study is concerned with maximizing communicative effectiveness by tailoring messages in light of the receiver's information processing ability, that ability being a function of the number of cognitive categories comprising the "mind” and referred to as cognitive complexity. It was hypothesized that cognitive complexity would cause people to process sentences differently depending upon whether they were high or low in cognitive complexity. Sentences varied in terms of the number of clauses and the location of those clauses relative to the main subject-verb combination. Cognitive complexity did, in fact, appear to predict ability to recall these sentences, particularly those with clauses occurring after the main subject-verb unit. People high in cognitive complexity had significantly better recall scores than people low in cognitive complexity.
Chapter Two

The Development of Classical Arabic
Chapter Two

The Development of Classical Arabic

2.1 Introduction

At the beginning of the Islamic period, only two sources of literary Arabic were available, the Qur’aan and the pre-Islamic poems. It is not surprising, then, that these two sources were to play a crucial role in the standardization and development of the Arabic language. It is not surprising, either, that the first scholarly activities in Islam concentrated on the text of the Qur’aan, which had to be transmitted and explained, both on the level of the text and on that of the contents. At the same time, when the direct ties with the desert were broken, the living practice of poetry was very soon replaced by scholarly interest in the pre-Islamic poems. The transmission of both 'texts' had taken place orally and informally, but in the rapidly expanding empire such a form of transmission could no longer be trusted.

The language itself, too, underwent a process of standardization. While in pre-Islamic times the Bedouin regarded themselves as members of one speech community, they had no single linguistic norm, and even in the language of poetry, which was supposed to be supra-tribal, a great deal of variation was accepted. After the conquests, when Arabic became the language of an empire, there was an urgent need to standardize the language for three reasons. First, the divergence between the language of the Bedouin and the various
colloquial varieties that emerged became a real threat to communication in the empire. **Second**, the policy of the central government, first in Damascus and later in Baghdad, aimed at the control of the subjects, not only in economical and religious but also in linguistic matters. Obviously, if Arabic was to be used as the language of the central administration, it had to be standardized. **Third**, the changed situation called forth a rapid expansion of the lexicon, which had to be regulated in order to achieve some measure of uniformity. This chapter deals with three topics connected with the process of standardization. The most important prerequisite for the written codification of the language was the invention of orthography, or rather the adaptation of existing scribal practices to the new situation. Then a standardized norm for the language was elaborated, and the lexicon was inventoried and expanded. Subsequently, when these requirements had been met, a stylistic standard was developed. The existing Bedouin model was instrumental in the development of a stylistic standard for poetry, but the emergence of an Arabic prose style marked the real beginning of Classical Arabic as we know it. In the final section of this chapter, we shall deal with the official status of the Arabic language.

### 2.2 The development of orthography

The first concern of Islamic scholars was to codify the texts with which they worked. Even though oral transmission continued to remain an essential component of Islamic culture, the risk of major discrepancies in the transmission became too large to ignore. The
need for an authoritative text was imperative above all in the case of the Revealed Book. Clearly, the central government had a major stake in the acceptance of a uniform Book throughout the empire as the basis for all religious and political activities.

The codification of the Qur’aan was a crucial moment in the development of a written standard for the Arabic language. On a practical level, the writing-down of the text involved all kinds of decisions concerning the orthography of the Arabic script and the elaboration of a number of conventions to make writing less ambiguous and more manageable than it had been in the Jaahiliyya. We have know that writing was not unknown in the peninsula in the pre-Islamic period. But, for religious reasons, early Islamic sources emphasized, perhaps even exaggerated, the illiteracy of the Prophet and, by extension, of the entire Jahilee society. The Prophet had been an 'ummee, someone who could not read nor write, and this was what made the revelation of the Qur’aan and his recitation of the text a miracle. In the biography of the Prophet, there are many references to his using scribes for his correspondence with the Arab tribes and for the writing of treaties, for instance the famous treaty with the settlements in North Arabia. This treaty which was signed in the course of the expedition to Tabook in year 9 of the Hijra, laid down for the first time the relations between Muslims and people of other religions. In the account preserved by the historians, the scribe and the witnesses are mentioned, as well as the fact that the Prophet signed it with his fingernail (cf. al-Waaqidee, Maghaazee III, 1,025ff.). This
last detail is probably added to underscore the fact that the Prophet himself could not write. The Prophet may well have been illiterate himself, but there were scribes on whom he could rely, just as his fellow Meccans used scribes in the management of their affairs. In the beginning, the revelation consisted of short messages which the Prophet brought to the believers and which could easily be committed to memory. But very soon, the messages grew longer and longer, and it became essential to have a written aid to memory, while the recitation of the text continued to take place orally. Tradition has preserved the names of several scribes to whom Muhammad dictated the messages, chief among them being Zayd ibn Thaabit (d. 45/665). The text of the Qur’an itself documents this shift from recitation to collected text. The current term Qur’an in the early suras (possibly borrowed from Syriac qeryaanaa 'recitation') is replaced increasingly often in the later suras with the term kitaab 'book'.

Both Islamic tradition and Western scholars agree that there was no complete collection of the revelation during the Prophet's lifetime, but there were fragments of all kinds of material, on which parts of the messages were recorded. The actual collection of all these fragments took place after the death of the Prophet. According to the tradition, the third caliph, 'Uthmaan (r. 25/644-35/656), ordered the establishment of an authoritative codex of the Qur’an. He entrusted this edition to Muhammad's scribe Zayd, who had already been involved in the recording of the text during the Prophet's lifetime. When the work was finished, the codex was sent to the important
centers of the Islamic empire, where it was to replace all existing alternative readings. Acceptance of this text, usually called al-mu S Haf, was slow, and non-canonical variants continued to be transmitted; but eventually, by the end of the second century of the Hijra, the 'Uthmaanic text had become the basis for religious teaching and recitation almost everywhere. In the first grammatical treatise of Arabic, Seebawayhi's (d. 177/793) Kitaab, all deviations from the consonantal text of the codex are rejected and only some divergence in the vocalisation of the text is allowed. Around the variant readings ( qiraa’aat ), a massive literature arose which at the same time contributed to the linguistic study of the text and the language of the Qur’aan. Apart from the problems of unification encountered during the codification of the text, the main problem confronting Zayd ibn Thaabit and his committee of text-editors was the ambiguity of the Arabic script. The type of script which the Meccan traders had at their disposal was still a primitive one. Basically, there were two problems connected with this primitive form of the Arabic alphabet. In the first place, there were as yet no diacritic dots to distinguish between certain phonemes, and many of the letters of the alphabet indicated two or even more phonemes, in the case of seen/sheen, Saad/ Daad, baa/taa/tha/noon/yaa', faa/qaaf, daal/ th aal, raal/zaay, Taa/ Da a'. This was the heritage of the Nabataean script that had been the model for the earliest form of Arabic script and that did not contain all of the Arabic phonemes. The second problem was connected with a general trait of all Semitic scripts, namely the fact that these scripts do not indicate the short vowels. In the case of the Nabataean model, even
many of the long vowels were written defectively (cf. above, p. 30). The former problem may already have been solved in pre-Islamic times. There are some indications that, very early on, scribes had used diacritic dots to distinguish between homographs. They may have borrowed this device from a Syriac model, since in the Syriac script dots are used to distinguish between allophonic variants of phonemes. According to some scholars, there are even examples of the use of dots in the Nabataean Script. The notation of the short vowels was an altogether more complicated problem. During the first century of Islam, when people started to collect and record the fragments of the Qur’aanic revelation, the need for a uniform and unambiguous system for the short vowels made itself felt. Various grammarians, among them the legendary 'inventor' of grammar, 'Aboo, l-'Aswad ad-Du'alee (d. 69/688?), are credited with the introduction of a system of (coloured) dots below and above the letters to indicate the three short vowels. Two other innovations attributed to 'Aboo l'Aswad concern the notation of the hamza (glottal stop) and the shadda (gemination). Both signs are absent in the Nabataean script. We have seen in Chapter 4 (p. 42) that in the Hijaaaz the hamza had probably disappeared, but in the variety of the language in which the Qur’aan was revealed and the pre-Islamic poems were composed, the hamza was pronounced. Because of the prestige of the language of poetry and the Qur’aan, the Hijaaazee scribes had to devise a way of recording the glottal stop. Since in their own speech the hamza had been replaced in many cases by a long vowel, they spelled words containing a hamza with a long vowel, indicated by a semiconsonant
w, y or 'alif. According to the tradition, 'Aboo l-‘Aswad improved this system by using a small letter 'ayn above the semi-consonant; this 'ayn indicated the presence of a guttural sound, namely the glottal stop. The gemination of a consonant was noted by a diacritic dot.

A substantial improvement in the system of short-vowel notation is usually attributed to the first lexicographer of the Arabic language, al-Khaleel ibn 'A Ḥmad (d. 175/791). He replaced the system of dots with specific shapes for the three short vowels, a small waaw for the vowel \( u \), a small 'alif for the vowel \( a \), and a (part of a) small yaa' for the vowel \( i \). He also changed the sign for the shadda, using a small seen (short for shadeed 'geminated') instead. When a single consonant was intended, a small khaa’ (short for khaafeef 'light') could be used. Originally, this system had been devised for writing down poetry, which also went through a period of codification, but gradually it spread to Qur’aanic manuscripts written in cursive script as well. It was considerably less ambiguous than the old system, in which the dots had to perform various functions. In the first period of the establishment of the Islamic empire, the government, therefore, opted to use Greek-speaking clerks in Syria and Egypt and Persian-speaking clerks in the East for purposes of administration and taxation. In the sources, the shift from Greek to Arabic in the tax register (deewaan) is traditionally connected with the name of the caliph 'Abd al-Malik. According to this story, the caliph ordered the clerks to shift to Arabic in the year 81/700, allegedly because one of the Greek clerks was caught urinating in an inkwell (al-Balaath uree, Futoo H 196-7).
Whatever the truth of that story, the shift is a sign of the growing self-confidence of the Arabs and their increased familiarity with a practical writing system.

2.3 The Standardisation of the Language

Even before the language shift of the deewaan, Arabic was used as a written language: the earliest papyri date from year 22 of the Hijra, and at the end of the first century of the Hijra quite a number of papyrus texts must have been circulating. The language of these papyri is highly irregular from the point of view of the codified grammar of Classical Arabic, but the fact that they contain a large number of hypercorrections demonstrates that the scribes tried to emulate a linguistic ideal. In Chapter 8, on the so-called Middle Arabic texts, we shall deal with the linguistic features of the corpus of papyri. In this chapter, our main purpose is to sketch the process of standardisation that was soon under way.

The Qur’aanic language, though virtually identical with the language of pre-Islamic poetry, has a typically religious flavour, manifesting itself in peculiarities of style and language that must have been absent in other registers. Likewise, the language of the poems was marked by poetic licences that did not occur in ordinary language. Although both sources constituted a model for correct Arabic, they could hardly serve as a model for ordinary prose. The arbiters of linguistic correctness, the Bedouin, were frequently called in for help in linguistic matters, but they were in no position to enforce a standard
language, if only because of their own linguistic differences. We have seen above (Chapter 4) that in the period of the Jaahiliyya the language of the various tribes varied to a certain extent; and, even though it is reasonable to assume that there were no real problems of communication, there was no general standard either. On the other hand, the growing sedentary population with a more or less complete command of the language was very much in need of such a standard, but could hardly be expected to devote themselves to decisions about linguistic correctness. As a matter of fact, their slipshod use of the language for practical purposes, as in the texts which we find in the papyri, was one of the reasons for a growing concern on the part of those who regarded themselves as the true heirs of Bedouin civilisation, the pure Arabs. Even if we do not believe the account of Muslim historians such as Ibn Khaldoon about the corruption of speech as the main motive behind the 'invention' of grammar (cf. p. 102), it can hardly be denied that in the early decades of Islam there was an increasing call for specialists who could provide adequate teaching in Arabic.

According to most of our sources, the fourth caliph 'Alee (r. 35/656-40/661) was the first to insist that something be done about the growing number of mistakes in speech (other sources mention the governor of the two Iraqs, Ziyaad ibn 'Abeehi). The person whose name has become connected with the first efforts to standardise and codify the language was the same 'Aboo l'Aswad whom we met above as the reformer of the writing system. Several stories are told about
his reluctance to accept this job; according to some historians, he was finally persuaded when his own daughter made a terrible mistake in the use of the declensional endings, by confusing the expressions maa 'a Hsana s-samaa'a 'how beautiful is the sky!' and maa 'a Hsanu s-samaa'i 'what is the most beautiful thing in the sky?' (as-Seeraafee, 'Akhbaar, ed. F. Krenkow, Beirut, 1936, p. 19). Another version of this story, in which the mistakes occur in the recitation of the Qur’aan, has been mentioned above (Chapter 4, p. 50).

The historicity of these anecdotes is, of course, doubtful, and Talmon (1985) has shown that the figure of 'Aboo l-'Aswad was used by later grammarians as some kind of eponym for their own grammatical school. But the point remains that grammarians must have played an important role in the standardisation of the language. The earliest scholarly efforts concerned the exegesis of the Revealed Book, but since study of the language of the Qur’aan could hardly ignore that other source of pre-Islamic Arabic, the poems, very soon the two main components of the corpus of texts that was to become canonical for the linguistic study of Arabic were combined in the writings of the grammarians.

The first grammarian to give an account of the entire language in what was probably the first publication in book form in Arabic prose, Seebawayhi, was not of Arab stock himself, but a Persian from Hamadhan. His example set the trend for all subsequent generations of grammarians. The grammarians believed that their main task was to provide an explanation for every single phenomenon in Arabic,
rather than a mere description, let alone a set of precepts on how to talk Arabic correctly. Consequently, they distinguished between what was transmitted and what was theoretically possible in language. In principle, they accepted everything that was transmitted from a reliable source: in the first place the language of the Qur’aan, which was sacrosanct anyway, in the second place everything that had been preserved from pre-Islamic poetry, and in the third place testimonies from trustworthy Bedouin informants. In this framework, even singularities or deviant forms were incorporated without, however, being accepted as productive forms that could constitute the basis for a theoretical linguistic reasoning. Such a distinction is characteristic of Islamic science as a whole, where 'aql 'logical reasoning' is always carefully distinguished from naql 'transmitted knowledge'. In this way, a separation was realised between the study of attested forms and the theories of the grammarians, and without being prescriptive the grammarians could still impose a canonical norm of the language.

The codification of grammatical structure went hand in hand with the exploration of the lexicon and its necessary expansion. These two aspects of the process of standardisation are connected. Just as the grammarians were needed because of the perceived 'corruption' of the language, the first aim of the lexicographers seems to have been the preservation of the old Bedouin lexicon, which was at risk. There are several reasons for the lexicographers' worries. In the first place, the sedentary civilisation of early Islam was markedly different from that of the desert tribes, who had been the guardians of the special
vocabulary of the pre-Islamic poems. No city-dweller could be expected to know all the subtle nuances of a vocabulary connected with camels and animal wildlife and tents. There are several anecdotes about grammarians that stress this component of a grammarian's activities. Thus, the grammarian 'Aboo 'Amr ibn al-'Alaa' (d. 154/770), when he started lecturing about language and poetry, was confronted by a real Bedouin, who interrogated him about the explanation of obscure words. When the grammarian passed the test, the Bedouin said khus th oo 'anhu fa-'innahu daabba munkara 'transmit from him, because he is an extraordinary beast of burden [i.e. a depository of knowledge]!' (az-Zajjaajee, Majaalis, ed. Haaroon, Kuwait, 1962, p. 262). This anecdote shows how grammarians had to prove their worth by their knowledge of the Bedouin lexicon. For the ordinary speaker, who had grown up in an Islamic city and knew nothing about the Bedouin milieu, even ordinary Arabic words had become unfamiliar. From one of the earliest commentaries on the Qur’aan, we can get an idea about which words had fallen into disuse. Muqaatil ibn Sulaymaan's (d. 150/767) Tafseer contains a large number of paraphrases of Qur’aanic words that he felt to be in need of explanation, e.g. 'aleem 'painful' (replaced by wajee’ ), mubeen 'clear' (replaced by bayyin ), naba'un 'news' (replaced by hadeethun ), naseeb 'share' (replaced by Ha D D ), the verb 'aataa ' to give' (replaced by 'a’ Taa ) and the interrogative adverb 'ayyaan 'when?' (replaced by mataa ).
The second threat to the lexicon had to do with the contact with other languages. When the Arabs became acquainted with the sedentary culture of the conquered territories, they encountered new things and notions for which there did not yet exist Arabic words. The most obvious sources for terms to indicate the new notions were, of course, the languages spoken in the new Islamic empire. And this was precisely what some of the Arab scholars feared. They were convinced that the influx of words from other cultures would corrupt the Arabic language, which had been chosen by God for His last revelation to mankind. In the first century of the Hijra, this attitude did not yet make itself felt, as the comments by the earliest exegetes on the vocabulary of the Qur’aan demonstrate. In preIslamic times, the Arabs had taken over a considerable number of words from the surrounding cultures. Most of them were borrowed either through the Jewish/Aramaic language of Syria, or through the Christian/Syriac language in Mesopotamia, where al-Heera was the most important centre for cultural and linguistic contacts. Examples of early borrowings that occur both in pre-Islamic poetry and in the Qur’aan are the following:

from Middle Persian (Pahlavi) through Syriac/Aramaic:

zanjabeel 'well in paradise' from Syriac zangabeel

from Pahlavi singa b ēr 'ginger'

warda 'rose' from Aramaic warđa from Avestan varṣa.
Some words must have been borrowed directly from Middle Persian, such as:

- istabraq 'brocade' from Pahlavi *sta b r* 'thick (of clothing)'
  + suffix -ak

- jund 'army' from Pahlavi *gund* 'army, troop'

- kanz 'treasure' from Pahlavi *gan* 'treasure'

- dirham 'silver coin' from Pahlavi *draxm* from Greek *drachmè*

or from Greek/Latin through Syriac/Aramaic:

- burj 'tower' from Syriac *boorgaa* from Greek *púrgos*

- zawj 'pair, married couple' from Syriac *zoogaa* 'yoke', *bar zoogaa* 'husband, wife' from Greek *zeūgos* 'yoke'

- 'deenaar 'gold coin' from Syriac *deenaraa* from Greek *dénárian* from Latin *denarius*

- qa Sr 'castle' from Aramaic *qa Sraa* from Greek *kástron* from Latin *castrum, castra*

- Siraa T 'path' from Aramaic *is Traatiyaa* from Greek *stráta* from Latin *strata*

- yaqut 'sapphire' from Syriac *yagoon* *Taa* from Greek *huákinthos* 'hyacinth'
qir Tas 'scroll of paper' from Syriac qar Teesaa, kar

Teesaa from Greek chartès.

And, of course, there was a large number of words that came in straight from Syriac/Aramaic, such as:

Salaat 'prayer' from Aramaic Slōthaa

teen 'fig' from Aramaic teenaa

sifr 'large book' from Aramaic sifraa

masjid 'place of worship' from Aramaic/Nabataean msgd'.

A special category of loanwords is constituted by those words that came in by a southern route, from languages such as South Arabian or Ethiopic, e.g.:

Sanam 'idol' from South Arabian Snm, Safaa'itic Snmt.

The oldest commentaries on the Qur’aan, such as the one by Mujaahid (d. 104/722), had no qualms in assigning words in the Qur’aan to a foreign origin. Mujaahid stated, for instance, that the word Toor 'mountain' came from Syriac, the word sijjeel 'baked clay' from Persian or Nabataean, and the word qis Taas 'balance' from Greek. In the cases mentioned here, he was not that far off, since Toor comes indeed from Syriac Toor, sijjeel from Pahlavi sang 'stone' + geel 'clay', and qis Taas perhaps ultimately derives from Greek dikastèς 'judge', through Syriac deeqas Toos. Some of the etymologies quoted by the commentators may be fanciful, but the important thing is that they looked upon the enrichment of the vocabulary as an advantage.
and as a sign of the superiority of the creative genius evidenced in the Qur’aan. By the end of the second century of the Hijra, however, some philologists had started to attack the notion that the Qur’aan could contain foreign loanwords, and attempted to connect the vocabulary of the Qur’aan with a Bedouin etymology. Thus, for instance, 'Aboo 'Ubayda (d. 210/825) says that 'the Qur’aan was revealed in clear Arabic language, and whosoever claims that the word taahaa is Nabataean makes a big mistake' (nazala l- Qur’aanu bi-lisaanin 'arabiyyin mubeenin fa-man za'ama 'anna taahaa bi-n-Nabatiyyati fa-qad 'akbara, Majaaz I, ed. F. Sezgin, Cairo, 1954, p. 17). Although most Arab lexicographers, such as as-Suyootee (d. 911/1505), continued to assign a foreign origin to many Arabic words, the idea of the purity of the Arabic language remained the prevalent attitude among some Islamic scholars, and attempts by Western scholars to find traces of other languages in the Qur’aan were and still are vehemently rejected.

The real problem arises in the case of Qur’aanic words that have developed a new technical meaning not supported by the semantics of the Arabic root. In such cases, the exegetes go out of their way to find a connection. Thus, for instance, for the expression yawm al-qiyaama 'the day of resurrection', the standard explanation in the commentaries is that it is connected with the root q-w-m 'to stand up', but most likely the Christian Syriac term qiyaametaa as a translation of the Greek anástasis 'resurrection' prompted the semantic extension of the Arabic word. Similar examples are those of zakaat 'alms', masjid 'mosque',
suhuf 'scriptures', sabt 'Saturday', soora 'portion of the Qur’an', and such central notions in the Qur’anic message as kitaab 'book', saa’a 'hour' etc. The term Su Huf 'scriptures', plural of Sa Heefa, is connected by the Arab commentators with a root S Hf, which occurs only as a denominative in the second measure with the meaning of 'making a mistake in reading'. In pre-Islamic poetry, Sa Heefa (plural Sa Haa’if) is used in the sense of 'page of writing'. The Qur’anic use of the word in the sense of 'scriptures' (e.g. Q 20/133 as- Su Huf al-oolaa 'the first scriptures') is difficult to explain from this, which is why Western commentaries often connect it with an Old South Arabian word S Hfi or with the common Ethiopic root s’- H-f ‘to write’. In line with the idea of the purity of the language, the semantic extension of an existing word was regarded as the most appropriate device for the expansion of the lexicon. The model for this procedure was believed to have been given by the language of the Qur’an itself. Since the grammarians analysed many religious terms such as Salaat 'prayer', zakaat 'alms', and the term 'islaam itself, as old Bedouin words which had received a specialised meaning in the religious context, semantic extension became an accepted method of creating new terminology. They were doubtless right in the sense that part of the religious vocabulary of the Qur’an is the result of an internal development without external influence. A case in point is the word 'islaam, which meant in general 'to surrender oneself', but came to mean 'to surrender oneself to God, to convert to the new religion brought by the Prophet'. Besides, even when the new meanings of
existing words were calqued on cognate words in other languages, their occurrence in the Qur’aan canonised the new meaning.

The large-scale influx of new notions and ideas in the early Islamic period could not be handled by giving new meanings to existing words alone. In spite of the purists' opposition, many words from other languages were simply taken over, either in their original form or with some slight adaptation to Arabic phonology or morphology. Loanwords from Persian abound in the domains of pharmacology, mineralogy and botany, for instance in the name of plants: banafsaj 'violet'; sankhaar 'gladiolus'; baaboonij 'camomile'; banj 'henbane'; fustuq 'pistachio'; khashkhash 'poppy'; narjis 'narcissus'. In the earliest translations of Greek logical, medical and philosophical writings, some of the technical terms are simply transliterations of a Greek word for which the translators were unable to find an Arabic equivalent. Thus we have, for instance, hayoolaa 'substance' (from Greek ὑλή), bulghum 'phlegm' (from Greek φλέγμα) and 'u S Tuquss 'element' (from Greek stoichĩon). The next best solution was to create a new word on the basis of an existing root by the application of one of the numerous morphological patterns of Arabic. In the beginning, each translator created in this way his own set of terms. The ensuing confusion was more or less ended with the establishment of the Bayt al- Hikma 'House of Wisdom', the translators' academy founded by the Caliph al-Ma'moon in 215/830. The Greek term katègoroúmenon 'predicate', for instance, had been variably translated as ma Ḥmool, maqool, Sifa or na't, until it was
standardised as ma Hmool. The Greek term *apóphansis* 'proposition' had been translated by as many as five different terms (Hukm, khabar, qawl jaazim, qawl qaa Ti', qa Diiyya), until qa Diiyya became the usual term. The use of patterns to create neologisms from existing roots was particularly useful in the translation of Greek medical terminology. A few examples may suffice to illustrate this method of inventing new vocabulary items. In his terminology of the skins of the eye, Hunayn ibn 'Ishaq translated Greek words in -eidès with abstract adjectives, e.g. qarniyya (Greek *keratoeidès*) 'cornea', zujaajiiyya (Greek *hualoeidès*) 'corpus vitreum', 'inabiyya (Greek *rhagoeidès*) 'uvea', shabakiyya (Greek *amphiblèstroeidès*) 'retina'. The pattern fu'aal was used to systematise the names of illnesses, e.g. zukaam 'catarrh', Sudaa 'headache', Sufaar 'jaundice', duwaar 'dizziness', Tu Haal 'infection of the spleen', and even khumaar 'hangover'.

A prerequisite for the creative use of the existing lexicon was its codification. The first complete dictionary of the Arabic language was composed by Seebawayhi's teacher, al-Khaleel ibn 'A Hmad, who had also been involved in the reform of the Arabic script (cf. above, p. 56) and who is generally acclaimed as the inventor of Arabic metrical theory. The professed aim of the Kitaab al-'ayn, which goes under his name, was the inclusion of all Arabic roots. In the introduction, a sketch is given of the phonetic structure of Arabic, and the dictionary fully uses the available corpus of Arabic by including quotations from the Qur’aan and from the numerous pre-Islamic poems, which had both undergone a process of codification and written transmission by
the hands of the grammarians. The arrangement of al-Khaleel's dictionary, which seems to have been completed by his pupils, set the trend for many subsequent lexicographical writings. The dictionary is divided into books, one for each letter, starting with that of the letter 'ayn, hence the name of the dictionary. Each book is divided into chapters, each dedicated to one set of radicals and containing all the permutations of these radicals. Thus, for instance, the chapter on the radical 'q-z, contains the roots 'z-q, q-z-', z'-q, and z-q-', which are the ones actually used in the language (musta'malaat). Perhaps this reflects some idea of a higher semantic connection between the permutations of radicals, although al-Khaleel does not mention such a connection. The system of the Kitaab al-'ayn remained in use for a long time, even after a new system had been introduced by the grammarian al-Jawharee, (d. 393/1003) in his Si Haa H. He arranged all roots in a kind of rhyming order, that is, alphabetically according to the last radical, then the first, then the second. This system became the current dictionary arrangement with the Lisan al-'Arab by Ibn Man D oor (d. 711/1311), the most popular dictionary ever written in the Arab world.

2.4 The Development of an Arabic Literary Style

The history of literary style in Arabic went hand in hand with the standardisation of the language. The development of such a style did not have to start from scratch. The same two sources that had been available for the standardisation of the language, the Qur’aan and the pre-Islamic poems, became the initial models for a literary style. As in
other cultures, the structured composition of poetry in Arabic preceded the emergence of a literary prose style. But here, too, the desert type of poetry did not satisfy all the needs of a new, elegant sedentary civilisation. New forms of poetry developed under the dynasty of the 'Umayyads, at whose court love poems became a new fashion (e.g. the poems of 'Umar ibn 'Abee Rabee'a, d. 43/712). Inevitably, this led to a looser use of language and to the development of new, often strophic types of poetry, that were not as heavily dependent on the Bedouin model. In such forms of poetry, there was easier access for popular expressions reflecting the new environment of Arabic culture. Some deviations in morphology, syntax and lexicon became gradually accepted, e.g. the use of contracted forms such as naseehi (from nasiyahu), baqee (from baqiya), or the confusion of the fourth and the first verbal form (cf. Fück 1950: 73ff.). In rajaz, poets could experiment with the creation of new words and word forms to a much higher degree than was permitted in official poetry. In general, the muwalladoon, the new Arabs, who had never seen the desert, could not be expected to be as excellent connoisseurs of Arabic as the pre-Islamic poets. Although for a long time the Bedouin model continued to serve as a strict canon, in Seebawayhi's Kitaab the poems of the muwalladoon are not excluded as evidence: the 1,000-plus quotations from poetry in the Kitaab include both Jaahilee poets and those from the urban milieu of the 'Umayyad period, such as 'Umar ibn 'Abee Rabee'a; he even quotes from rajaz poetry.
Gradually, a distinction came into being between the official brand of poetry that clung to the old models and took pleasure in using obsolete vocabulary and avoiding any adaptation to the new modes of speaking, on the one hand, and a new, 'faster', kind of poetry, often improvised, often in strophic form, and very often containing vulgarisms, on the other. In the course of time, these two kinds of poetry grew further apart. Official poetry became more and more erudite, until it could no longer be understood without explanation. The poet al-Mutanabbee (d. 355/965), for instance, published his poems together with a learned commentary. The more popular form of poetry, on the other hand, went through a different development. In its most developed form, the strophic muwashsha and the zajal, it included the use of colloquial forms in a refrain. This kind of poetry became especially popular in the Islamic West (cf. below, p.227).

Because of its idiosyncrasies, poetry is of lesser importance in the standardisation of language than prose. We have seen above that for commercial and administrative purposes Arabic was used from the beginning of the Islamic empire. Such written documents had no literary pretensions whatsoever, although their scribes did try to maintain a Classical norm, which means that already at this time there was a standard (on the language of the papyri see below, Chapter 8). But there were other forms of speech, some of them with roots in the Jaahiliyya. In the first place, Arabic culture had a reputation of long standing for its ability to put speech to rhetorical use. The Bedouin admired verbal prowess, and the tradition of delivering public
speeches was continued in early Islam. A second genre of texts with roots in the pre-Islamic period is the art of story-telling. From early times onwards, storytellers (qussaas) had played an important role in the life of the tribe by transmitting the stories about the exploits of the tribes ('ayyaam al-'Arab), and this tradition was continued in a modified form in early Islam when storytellers went around to tell about the events in the life of the Prophet, the early Islamic expeditions and the conquests of foreign countries. These stories were meant for the general public and they were no doubt told in a lively style, full of fictitious conversations and without any literary embellishments. The topics dealt with by the professional storytellers were also studied by scholars. They had in common with the storytellers a certain aversion to writing down their reports: only the Qur’aan could be a written Book. They did use written notes for recording their own memories and those of their informants, but these were intended for private use only. The earliest efforts to put down in writing systematically the traditions about Muhammad and the early period of the conquests did not start until the end of the first century of the Hijra, at a time when the last people who had actually met the Prophet were old men and women who were bound to die soon. This period witnessed a feverish activity on the part of scholars to collect all they could from the last witnesses still alive. Scholars such as az-Zuhree (d. 124/742) compiled collections of Hadeeths, that were eagerly sought by the caliphal court and were probably deposited in the palace. The best-documented genre in early Islam is the epistolary one. The earliest examples of epistolary texts are found in the
accounts of the correspondence between the Prophet and the tribal chieftains. During the period of the conquests, there must have been a constant stream of letters between the central authorities in Medina and the commanders in the field. The contents of these letters were mostly commercial, but no doubt some epistolary conventions existed even then. It is impossible to determine to what degree the texts of those letters that have been preserved by later historians are authentic. Some historians refer to actual documents, for instance the treaty between the Prophet and the community of Doomat al-Jandal, which al-Waaqidee (Maghaazee III, 1,030) claims to have seen personally. But in general we have no guarantee about the authenticity of the exact wording, although the historians may well have preserved the gist of the contents. The same conclusion applies to such texts as the letters of the early raashidoon or the arbitration pact of Siffeen.

Since most of the scribes (kuttaab) in the early period were Syrians or Persians, or perhaps even Christian Arabs from the tribes outside the peninsula, some foreign examples and conventions may have found their way into Arabic literary products at this period. The reform of the caliph 'Abd al-Malik (r. 65/685-691/705), who as we have seen was responsible for the shift of language in the deewaan, must have been the starting point for a new fashion in writing Arabic for official purposes. Since the secretaries were responsible for the composition of official documents and letters, their role in the development of a chancellery style was essential. Under 'Abd al-Malik's successor Hishaam (r. 105/724-125/743), the foundation was laid for the administrative system that was later taken over and
perfected by the 'Abbaasid caliphs. From the beginning of the 'Umayyad dynasty, the sponsorship of the caliphs was an important factor in the production of texts, both literary and administrative. According to some sources as early as Mu'aawiya's (r. 41/661-60/680) reign, the caliph had some kind of library in which he deposited written versions of Hadeeths, some of which had been collected at his request. His grandson Khaalid ibn Yazeed ibn Mu'aawiya had a keen interest in alchemy and may have commissioned the first translations from Greek into Arabic. Certainly there are enough reports about the later 'Umayyads requesting translations of Greek or Syriac books, mostly on medicine, to warrant the conclusion that a depository (khizaana) of books belonged to the normal appurtenances of the caliphal court. Although the 'Abbaasids, did their best to suppress any favourable report about the 'Umayyads, it is fairly certain that the 'Umayyad caliphs actively supported the activities of scholars such as az-Zuhree in the field of Hadeeth-collecting. The development of a written Arabic style went hand in hand with the development of a literary prose corpus consisting of translations from Persian, including the Kitaab fee s-siyaasa al-'aammiyya mufa S Salan 'Treatise on general administration, with full particulars' that is sometimes attributed to Hishaam's secretary 'Aboo l-'Alaa' Saalim. The epistolary style was perfected by his successor 'Abd al-Hameed ibn Ya Hya (d. after 132/750), secretary of Marwaan II (r. 127/744-132/750), who used this style in treatises, some of which have been preserved, such as his Risaala 'ilaa l-kuttaab 'Letter to the scribes'. He used an ornate style, with an extensive eulogy at the beginning of the treatise, ample
use of parallelism, in a quantitative rhythm, sometimes in rhymed prose (saj’), sometimes in a loose parallel structure of patterns. On the other hand, his style does not include the use of intricate rhetorical figures or rare vocabulary. The first sermons and epistles such as those by al-Hasan al-Basree (d. 110/728) adopted the form of the epistolary genre by addressing them to the caliph, but adapted the epistolary style to the topic at hand. The tradition of caliphal sponsorship of bookwriting that was initiated by the ’Umayyad caliphs was continued under the ‘Abbaasid dynasty. At the request of some of the caliphs, books were composed, mostly by foreigners that were to acquaint the intellectual elite with the achievements of other cultures. Scholars such as the Persian Ibn al-Muqaffa’ (d. ±142/759), a near-contemporary of ’Abd al-Hameed, produced literary translations from Pahlavi. His most famous translation was that of the Indian fables of Kaleela wa-Dimna, but he also composed new original treatises, such as the Kitaab al-’adab al-kabeer and the Risaala fee S-Sa Haaiba. These treatises were mostly concerned with court etiquette and the behavioural code in the relations between rulers and ruled.

The coexistence of and the conflict between a high and a low variety of the language in Islamic culture made its presence felt from the time of the earliest papyri. Through the Middle Arabic texts, this diglossia was introduced in the domain of literary and semi-literary products. We shall see below that this conflict has never disappeared since. In Modern Arabic literature, just like in that of the Classical age, authors have to choose the level of speech in which they wish to write. But
the main constraint for all written production in Arabic is the position of Classical Arabic as the language of prestige. Whether in an 'elevated' or in a 'lower' style, the ultimate model remains the standard language, and even when an author deliberately sets out to write in the vernacular, in the end he can never escape the framework of the written language.

2.5 The official Status of Arabic

Throughout the classical period of Islam, Arabic remained the language of prestige that was used for all religious, cultural, administrative and scholarly purposes. In none of these functions was it ever seriously threatened in the first centuries of Islam. In their attitude towards other languages, the speakers of Arabic took it for granted that there could be no alternative to the Arabic language. This explains the disappearance of all other cultural languages in the Islamic empire, such as Coptic, Greek, Syriac and even Persian. With very few exceptions, the Arab grammarians showed no inclination to study other languages, and speakers of these languages only very seldom found anything to boast of in their own language, preferring to speak and write in Arabic instead. During the first centuries of the Hijra, speakers of Persian tended to regard their own language as inferior to Arabic. We have already seen that the author of the first linguistic description of Arabic, Seebawayhi, was himself a speaker of Persian, but there are absolutely no traces in his Kitaab of any interest in the Persian language. Another famous grammarian, al-Faarisee (d. 377/987), on being asked by his pupil Ibn Jinnee about his mother
tongue, Persian, stated unequivocally that there could be no comparison between the two languages, since Arabic was far superior to Persian (Kha Saa'i S I, 243). Eventually, a countermovement of Persian ethnic feeling (shu'ooobiyya) arose which opposed the monopoly of the Arabs but did not challenge the position of Arabic. From the ninth century onwards, however, Persian became increasingly used as a literary language, first of all in Eastern Iran, where Arabic culture had never gained a foothold. At the court of the more or less independent dynasties in the East, New Persian or Farsee was used in poetry. Under the dynasty of the Samaanids (tenth century), it replaced Arabic as the language of culture. After the fall of Baghdad (657/1258) during the Mongol invasion, Arabic lost its position as the prestigious language in the entire Islamic East to Persian, except in matters of religion. In Iran itself, the Safavid dynasty under Shah 'Ismaa'eel (906/1501) adopted Farsee and the Shi'ite form of Islam as the national language and religion. In all other regions, Arabic kept its position for a long time. A case in point is Mamluk Egypt. The Arabs had always looked down on the Turks, whom they regarded as good soldiers and therefore useful as protectors of Islam, but without any gift for culture. Their Arabic, if they spoke it at all, was deficient. Yet, Mamluk trainees received intensive instruction in Arabic, and most Mamluks must at least have understood the language. In the biographical sources about the Mamluks (e.g. a S-Safadee's al-Waafee bi-l-wafayaat), mention is made of many Mamluk scholars who occupied themselves with the religious and grammatical literature in Arabic, and even when in the
fourteenth century they started to produce scholarly writings in Qipčaq and Oghuz Turkic, Arabic remained in use in Egypt as the main literary language. When the Seljuks conquered Anatolia, Turkish became the official language of their empire, with Persian as the literary language; but even then, Arabic remained important, in the first place as a source of loanwords in Turkic (cf. below, Chapter 13, p. 234), and in the second place as the language of religion. It lost, however, its place as administrative language of the empire to Turkish. At the end of the nineteenth century, during the Renaissance (Nah Da) of Arabic (cf. below, Chapter II), attempts were made to reintroduce Arabic as the language of administration, but with the advent of the colonial period these attempts turned out to be short-lived, and it was not until the independence of the Arab countries as political entities in the twentieth century that it became once again the language in which matters of state and administration could be expressed.
chapter Three

Literary and Modern Standard Arabic
Chapter Three

Literary and Modern Standard Arabic

3.1 Introduction:

The term "Arabic" may refer either to literary Arabic, which no Arab speaks as a mother tongue, or Modern Standard Arabic or to the many spoken varieties of Arabic commonly called "colloquial Arabic." Arabs consider literary Arabic as the standard language and tend to view everything else as mere dialects. Literary Arabic, *al-luğatu-l-arabīyyatu-l-fusḥā* (Literally: "the most eloquent Arabic language" — الفصحى) refers both to the language of present-day media across North Africa and the Middle East and to the more archaic language of the Qur'an. (The expression *media* here includes most television and radio, and all written matter, including all books, newspapers, magazines, documents of every kind, and reading primers for small children.) "Colloquial" or "dialectal" Arabic refers to the many national or regional dialects/languages derived from Classical Arabic, spoken daily across North Africa and the Middle East, which constitute the everyday spoken language. These sometimes differ enough to be mutually incomprehensible. These dialects are not typically written, although a certain amount of literature (particularly plays and poetry) exists in many of them. They are often used to varying degrees in informal spoken media, such as soap operas and talk shows.
The sociolinguistic situation of Arabic in modern times provides a prime example of the linguistic phenomenon of Diglossia—the normal use of two separate varieties of the same language, usually in different social situations. In the case of Arabic, educated Arabs of whatever nationality can be assumed to speak both their local dialect and their school-taught literary Arabic (to an equal or lesser degree). This diglossic situation facilitates code switching in which a speaker switches back and forth unaware between the two varieties of the language, sometimes even within the same sentence. In instances in which Arabs of different nationalities engage in conversation only to find their dialects mutually unintelligible (e.g. a Moroccan speaking with a Lebanese), both should be able to code switch into Literary Arabic for the sake of communication. Since the written Arabic of today differs substantially from the written Arabic of the Qur'anic era, it has become customary in western scholarship and among non-Arab scholars of Arabic to refer to the language of the Qur'an as Classical Arabic and the modern language of the media and of formal speeches as Modern Standard Arabic. Arabs, on the other hand, often use the term Fushā to refer to both forms, thus placing greater emphasis on the similarities between the two. Quite a few English words are ultimately derived from Arabic, often through other European languages, especially Spanish, among them every-day vocabulary like sugar (sukkar), cotton (qutn) or magazine (maḏāzin).

More recognizable are words like algorithm, algebra, alchemy, alcohol, azimuth, nadir, and zenith (see List of English words of Arabic origin). The Maltese language spoken on the Mediterranean island of Malta is the
only surviving European language to derive primarily from Arabic (a North African dialect), though it contains a large number of Italian and English borrowings.

3.2 Dialectical Phonologies

In some dialects, there may be more or fewer phonemes than those listed in the chart above. For example, non-Arabic [v] is used in the Maghreb dialects as well in the written language mostly for foreign names. Semitic [p] became [f] extremely early on in Arabic before it was written down; a few modern Arabic dialects, such as Iraqi (influenced by Persian) distinguish between [p] and [b]. Interdental fricatives ([θ] and [ð]) are rendered as stops [t] and [d] in some dialects (principally Levantine and Egyptian) and as [s] and [z] in "learned" words from the Standard language. Early in the expansion of Arabic, the separate emphatic phonemes [d] and [ð] coalesced into a single phoneme, becoming one or the other. Predictably, dialects without interdental fricatives use [d] exclusively, while those with such fricatives use [ð]. Again, in "learned" words from the Standard language, [ð] is rendered as [z] in dialects without interdental fricatives. Another key distinguishing mark of Arabic dialects is how they render Standard [q] (a voiceless uvular stop): it retains its original pronunciation in widely scattered regions such as Yemen and Morocco (and among the Druze), while it is rendered [g] in Gulf Arabic, Iraqi Arabic, Upper Egypt and less urban parts of the Levant (e.g. Jordan) and as a glottal stop [ʔ] in many prestige dialects, such as those spoken in Cairo, Beirut and Damascus. Thus, Arabs instantly give
away their geographical (and class) origin by their pronunciation of a word such as *qamar* "moon": [qamar], [gamar] or [ʔamar].

### 3.4 Grammar

**History**

Due to the rapid expansion of Islam in the 8th century, many people learned Arabic as a lingua franca. For this reason, the earliest grammatical treatises on Arabic are often written by non-native speakers. The earliest grammarian who is known to us is Abd Allāh ibn Abī Isḥāq (died 735 (117 H)). The efforts of three generations of grammarians culminated in the book of the Persian scholar *Sibawayhi* (ca. 760–793).

Traditionally, the grammatical sciences are divided into four branches:

- *al-luḡah* (lexicon) concerned with collecting and explaining vocabulary
- *at-taṣrif* (morphology) determining the form of the individual words
- *an-naḥw* (syntax) primarily concerned with inflection (*i Ḳab*) which had already been lost in dialects.
- *al-ištiqāq* (derivation) examining the origin of the words
Phonology

Classical Arabic has 28 consonantal phonemes (including two semi-vowels), originally corresponding to the 28 letters of the Arabic alphabet. (By Koranic times, however, the letter 'alif' no longer represented a glottal stop, but a long ā. As a result, a diacritic symbol, 'hamza', was introduced to represent this sound. In addition, some of these phonemes have coalesced in the various modern dialects, while new phonemes have been introduced through borrowing or phonemic splits.) Arabic has six vowel phonemes (three short vowels and three long vowels); they appear as various allophones, depending on the preceding consonant. Short vowels are not usually represented in written language, although they may be indicated with diacritics. List of phonemes as transliterated in this article:

- 26 consonants:  b t ā h ā d r z s š s ā t ż ĝ f q k l m n h
- 2 semi-vowels: w y
- 6 vowels: a ā i ī u ū

Note that Arabic is particularly rich in uvular, pharyngeal, and pharyngealized ("emphatic") sounds. The emphatic sounds are generally considered to be š, ḍ, ṭ, and ž. (Sometimes q is wrongly included -- wrongly, because only the four emphatics, and not q, cause assimilation of emphasis to an adjacent non-emphatic consonant.)
It is generally believed that Classical Arabic phonology is extremely conservative, and is close to that of Proto-Semitic; only the South Arabian languages are more conservative in their phonology. The six vowels are inherited without change from Proto-Semitic, and of the 29 Proto-Semitic consonants, only one has been lost (/ʃ/, with merged with /s/). In addition, various sounds have been changed. An original lateral fricative /ɬ/ became /ʃ/, restoring a previously lost sound. Another complex lateral sound, /dɬʒ/ (voiced pharyngealized lateral affricate), became /d/ with loss of the lateral sound, although the original sound appears to have still existed at the time of the Koran. (Hence the Classical appellation 'luğatu l-ḍād' or "language of the ḍād" for Arabic, where 'ḍād' is the letter corresponding to this sound, which was considered by Arabs to be the most unusual sound in Arabic.) An original /p/ became /f/, and /ɡ/ became palatalized /ɡʲ/ at the time of the Koran, and /dʒ/ in the standard modern pronunciation of Classical Arabic. (The dialects variously have /dʒ/ (Arabian Peninsula), /ɡ/ (Cairo), /ʒ/ (North Africa), /j/ (Persian Gulf area), and original /ɡʲ/ (a few isolated pockets here and there).) Other changes may have occurred as well, especially in the emphatic consonants, depending on how Proto-Semitic is reconstructed. The syllable structure of Arabic is such that there may be clusters of two, but not of three consecutive consonants. A cluster of two
consonants at the beginning of an utterance will be preceded by an auxiliary vowel (*alif al-wasl*).

**Noun**

**State**

The Arabic noun can take one of three states of definiteness: definite, indefinite or construct state. The definite state is marked by the article *al-*.

The indefinite state is marked by an ending -*n* (nunciation). The construct state is unmarked and occurs in the first member of a genitive construction.

**Article**

The article (*adātu-t-tarīf*) *al-* is indeclinable and expresses definite state of a noun of any gender and number. The initial vowel (*hamzatu-l-wasl*), is volatile in the sense that it disappears in sandhi, the article becoming mere -*l-* (although the alif is retained in orthography in any case for clarity).

Also, the *l* is assimilated to a number of consonants (dentals and sibilants), so that in these cases, the article in pronunciation is expressed only by geminating the initial consonant of the noun (while in orthography, the writing *alif lam* is retained, and the gemination may be expressed by putting šadda on the following letter).
The consonants causing assimilation (trivially including \( l \)) are: \( t, \, ṯ, \, d, \, ḍ, \, r, \)
\( z, \, s, \, š, \, š, \, ḍ, \, ṯ, \, z, \, l, \, n \). These 14 letters are called 'solar letters' (\( ḥuruf šamsiyyat \)), while the remaining 14 are called 'lunar letters' (\( ḥuruf qamariyyat \)). The solar letters all have in common that they are dental, alveolar and postalveolar consonants in the classical language, and the lunar consonants are not. (\( ẓ \) is pronounced postalveolar in most varieties of Arabic today, but was actually a palatalized voiced velar plosive in the classical language, and is thus considered a lunar letter.)

**Inflection**

An Arabic noun can take three cases: nominative, genitive and accusative, and three numbers: singular, dual and plural. Normally, nouns take the ending \(-u(n)\) in the nominative, \(-i(n)\) in the genitive and \(-a(n)\) in the accusative. The case endings are only present in formal or literary language. Technically, every noun has such an ending, although at the end of a sentence, no inflection is pronounced, even in formal speech, because of the rules of 'pause'. The plural of a noun is formed by a suffix in some cases (sound plurals), but frequently, the vowel structure of a word is changed to form the plural (broken plurals). There are a number of patterns of how this is done. Some singular nouns take several plurals. There could be traces of broken plurals in other Semitic languages, but nowhere are they as widespread as in Arabic. The plurals of nouns representing humans usually use sound plurals. Masculine sound plurals take the forms "-ūn" in the nominative and "-īn" in the genitive and
accusative. In the feminine, the ending is "-āt" and is limited in its declension to the nominative and genitive endings. For example, "-ātun" and "-ātin" are possible, but not "-ātan". This pattern can also be used with for plurals of non-human nouns.

**Gender:**

Arabic has two genders, expressed by pronominal, verbal and adjectival agreement. Agreement with numerals shows a peculiar 'polarity', c.f. the section on numerals. The genders are usually referred to as masculine and feminine, but the situation is more complicated than that. The 'feminine' singular forms are also used to express 'singulatives', which are plurals of inanimate objects of both grammatical genders. The marker for the feminine gender is a -t- suffix, but some nouns without this marker also take feminine agreement (e.g. ʿumm 'mother', ʿarḍ 'earth'). Already in Classical Arabic, the -t marker was not pronounced in pausa. It is written with a special letter (ta marbuta) indicating that a t sound is to be pronounced in sandhi, but not in pausa.

**Genitive construction (Iḏāfa)**

A noun may be defined more closely by a subsequent noun in the genitive. The relation is hierarchical; the first term (al-muḏāf) governs the second term (al-muḏāf ilayhi). E.g. ʿaytu raḏūlin 'house of a man'. The construction as a whole represents a nominal phrase, the state of which is
inherited from the state of the second term. The first term must be in construct state, and thus cannot be marked definite or indefinite. Genitive constructions of multiple terms are possible. In this case, all but the final term take construct state, and all but the first member take genitive case. This construction is typical for a Semitic language. In many cases the two members become a fixed coined phrase, the idāfa being used as the equivalent of nominal composition in Indo-European languages (which does not exist in Semitic). *baytu-ťalabati* thus may mean either 'house of the (certain, known) students' or 'the student hostel'.

**Nisba**

The Nisba (*an-nisbatu*) is a common suffix to form adjectives of relation or pertinence. The suffix is -*iyy-* for masculine and -*iyyat-* for feminine gender (in other words, it is -*iyy-* and is inserted before the gender marker). E. g. *lubnānu* 'Lebanon', *lubnāniyyun* 'Lebanese'.

A construction noun + nisba-adjective is often equivalent to nominal composition in Indo-European languages.

**Pronoun**

A pronominal paradigm consists of 12 forms: In singular and plural, the 2nd and 3rd persons differentiate gender, while the 1st person does not. In the dual, there is no 1st person, and only a single form for each 2nd and 3rd person. Traditionally, the pronouns are listed in order 3rd, 2nd, 1st.
Personal pronouns: (Table 3.1 Personal pronouns)

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
<th>Dual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd (m)</td>
<td>Huwa</td>
<td>hum</td>
<td>humā</td>
</tr>
<tr>
<td>3rd (f)</td>
<td>Hiya</td>
<td>hunna</td>
<td></td>
</tr>
<tr>
<td>2nd (m)</td>
<td>Anta</td>
<td>antum</td>
<td>antumā</td>
</tr>
<tr>
<td>2nd (f)</td>
<td>Anti</td>
<td>antunna</td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>Ana</td>
<td>naḥnu</td>
<td>(n/a)</td>
</tr>
</tbody>
</table>

Enclitic pronouns

Enclitic forms of the pronoun (אך-ἡμāʾiru al-muttaṣilatu) may be affixed to nouns (representing genitive case, i.e. possession) and to verbs (representing accusative, i.e. a direct object). Most of them are clearly related to the full personal pronouns. They are identical in form in both cases, except for the 1st person singular, which is -ī after nouns (genitive) and -nī after verbs (accusative). (Table 3.2 Enclitic pronouns)

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
<th>Dual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd (m)</td>
<td>-hu</td>
<td>-hum</td>
<td>-humā</td>
</tr>
<tr>
<td>3rd (f)</td>
<td>-hā</td>
<td>-hunna</td>
<td></td>
</tr>
<tr>
<td>2nd (m)</td>
<td>-ka</td>
<td>-kum</td>
<td>-kumā</td>
</tr>
<tr>
<td>2nd (f)</td>
<td>-ki</td>
<td>-kunna</td>
<td></td>
</tr>
</tbody>
</table>
Demonstratives

There are two demonstratives (asmāʾu al-išāratu), near-deictic ('this') and far-deictic ('that'):

- **hāda**, f. hādihi, pl. hāulāi 'this, these'
- **dālika**, f. tilka, pl. ultā ika 'that, those'

Numerals

Cardinal numerals

Cardinal numerals (al-adād al-ašliyyat) from 1-10 (zero is ṣifr, from which the English words "cipher" and "zero" are ultimately derived)

- 1 *waḥidun*
- 2 *iḥānu*
- 3 *ṭalāṭatu*
- 4 *arbaatu*
- 5 *ḥamsatu*
- 6 *sittatu*
- 7 *sabatu*
- 8 *ṭamāniyyatu*
• 9 tisatu 
• 10 ašaratu

The numerals 1 and 2 are adjectives; 3-10 are diptotes (the ending -(t)u is dropped in oral usage). Numerals 3-10 have a peculiar rule of agreement known as polarity: A feminine referrer agrees with a numeral in masculine gender and vice versa, e.g. falāḥu fatayātīn 'three girls'.

Numerals 11-19 are indeclinable, and they show gender agreement (not polarity). The noun counted takes accusative singular.

• 11 aḥada ašara 
• 12 iḥā ašara 
• 13 ṭalāḵata ašara

The numerals 20-99 are followed by a noun in the accusative singular as well. There is agreement in gender with the numerals 1 and 2, and polarity for numerals 3-9.

• 20 išrūna (dual of '10')
• 21 aḥadun wa išrūna 
• 22 iḥāni wa išrūna 
• 23 falāḵatu wa išrūna
• 30 falāḵīna
Whole hundreds, thousands etc. appear as first terms of genitive constructions, e.g. *alfu laylati wa laylatu* '1001 nights'.

- 40 *arbaïna*

Fractions of a whole smaller than "half" are expressed by the structure *sg. ful, pl. afâl.*

- *niṣfūn* "half"
- *ṯulḥūn* "one third"
- *ṯulṯâni* "two thirds"
- *rubūn* "one quarter"
- *ṯâlatâtu arbâin* "three quarters"

Ordinal numerals

Ordinal numerals (*al-adâd at-tartiyabiyyat*) higher than "first" are formed using the structure *fâilun, fâilatun*:

- m. *awwalu*, f. īlā "first"
- m. *fânin*, f. *fâniyât* "second"
- m. *fâliḥun*, f. *fâliḥatun* "third"
- m. *râbiun*, f. *râbiatun* "fourth"
They are adjectives, hence, there is agreement in gender with the noun, not polarity as with the cardinal numbers.

**Verb** : As in many other Semitic languages, the Arabic word formation is based on a (usually) triconsonantal root, which is not a word in itself but contains the semantic core. The consonants **k-t-b**, for example, indicate 'write', **q-r-** indicate 'read', **-k-l** indicate 'eat' etc.; Words are formed by supplying the root with a vowel structure and with affixes. Traditionally, Arabic grammarians have used the root **f--l** 'do' as a template to discuss word formation. The personal forms a verb can take correspond to the forms of the pronouns, except that in the 3rd person dual, gender is differentiated, yielding paradigms of 13 forms.

**Perfect** : In the perfect conjugation, the perfect stem **faal** is affixed with a personal ending, e.g. **kataba** 'he wrote', **qaraa** 'he read'. The perfect expresses a completed action, i.e. mostly past tense.

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
<th>Dual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd (m)</td>
<td>faal-ā</td>
<td>faal-ū</td>
<td>faal-ā</td>
</tr>
<tr>
<td>3rd (f)</td>
<td>faal-at</td>
<td>faal-na</td>
<td>faal-tā</td>
</tr>
<tr>
<td>2nd (m)</td>
<td>faal-ta</td>
<td>faal-tum</td>
<td>faal-tumā</td>
</tr>
<tr>
<td>2nd (f)</td>
<td>faal-tā</td>
<td>faal-tunna</td>
<td>--</td>
</tr>
<tr>
<td>1st</td>
<td>faal-tu</td>
<td>faal-nā</td>
<td>(n/a)</td>
</tr>
</tbody>
</table>
Imperfect

The imperfect expresses an action in progress, i.e. mostly present tense. There are several vowel patterns (a-a, a-u,a-i) the root can take. The root takes a prefix as well as a suffix to build the verb form. E. g. yaktubu 'he is writing'. Note the co-incidence of 3rd f. sg. and 2nd m. sg.

Table : 3.3 Prefix & Suffix

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
<th>Dual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd (m)</td>
<td>ya-fal-u</td>
<td>ya-fal-īna</td>
<td>ya-fal-āni</td>
</tr>
<tr>
<td>3rd (f)</td>
<td>ta-fal-u</td>
<td>ya-fal-na</td>
<td>ta-fal-āni</td>
</tr>
<tr>
<td>2nd (m)</td>
<td>ta-fal-u</td>
<td>ta-fal-īna</td>
<td>ta-fal-āni</td>
</tr>
<tr>
<td>2nd (f)</td>
<td>ta-fal-īna</td>
<td>ta-fal-na</td>
<td>--</td>
</tr>
<tr>
<td>1st</td>
<td>a-fal-u</td>
<td>na-fal-u</td>
<td>(n/a)</td>
</tr>
</tbody>
</table>

Mood

From the imperfect stem, modal forms can be derived: the subjunctive by (roughly speaking) replacing the final vowel by a, the jussive by dropping this a of the subjunctive, and the imperative (only 2nd person) also by dropping the verbal prefix. The subjunctive is used in subordinate clauses after certain conjunctions. The jussive is used in negation, in negative imperatives, and in the hortative li+jussive. For example: 2. sg. m.:

- imperfect indicative tafalu 'you are doing'
- subjunctive an tafala 'that you do'
• jussive la tafal 'do not!'
• imperative ifal 'do!'.

Voice

Arabic has two verbal voices, active and passive. The passive voice is expressed by a change in vocalization and is normally not expressed in unvocalized writing. For example:

• active faala 'he did', yafalu 'he is doing'
• passive fuila 'it was done', yufalu 'it is being done'

Weak verbs

Roots containing one or two of the radicals ـ (wāw), ـ (yā) or ـ (hamza) are subject to special phonological rules because these radicals can be influenced by their surroundings. Such verbs are called 'weak' (verba infirmae, 'verbs of weak [radical]) and their paradigms must be given special attention. In the case of hamza, these peculiarities are mainly orthographical, since hamza is not subject to elision (the orthography of hamza and alif is unsystematic due to confusion in early Islamic times). According to the position of the weak radical in the root, these verbs are called primae infirmae, mediae infirmae or tertiae infirmae. Another special class of roots are such that their second and third radicals are identical. These roots are called mediae geminatae.
Stem formation

Derived verbs are variations on the shape of the primary *kataba* stem, such as *kattaba, kātaba, inkataba, takattaba*. Semantically, these formations correspond to changes in meaning such as intensive, reflexive, and causative, though the exact meaning varies from verb to verb and needs to be recorded in the lexicon. Classical Arabic has a great number of derived stems, not all of which are still in use. For the modern language, it is mostly sufficient to consider stems I-VIII and X.

- I. *faal-* (the basic stem)
- II. *faal-* (gemination of the middle radical)
- III. *fāal-* (lengthening of the vowel following the first radical)
- IV. *afal-* (clustering of first and second radical)
- V. *tafaal-* (prefix *ta-* and gemination of middle radical)
- VI. *tafāal-* (prefix *ta-* and lengthening of the vowel following the first radical)
- VII "*infaal*" (prefix "*in-*")
- VIII. *iftaal-* (infix -*ta-* after first radical)
- X. *istafal-* (prefix *(i)st-*)

The exact vocalisation will be dependent on the word form.

Common uses of those stems include:

- *faala* is often used to make an intransitive verb transitive. Eg: *karuma* is "be noble" but *karrama* is "make (someone) to be noble", or, more idiomatically, to "honor".
• *infaala* to give a passive meaning. Eg: *kasara* "break" and *inkasara" be broken".

**Participle**

The Arabic participle is a verbal noun formed from one of the derived verbal stems. E.g. *muallimun* 'teacher' is the active participle to stem II. of the root *-l-m* ('know').

- The passive participle to Stem I is *mafūlun*
- Stems II-X take prefix *mu-* and nominal endings (e.g. II. *mu-fail-un*.)

**Infinitive**

There is a second type of verbal noun besides the participle that is referred to as 'infinitive' because it often translates to infinitive constructions in Indo-European languages. It is strictly speaking not an infinitive, it would be more correct to speak of "verbal noun I" and "verbal noun II", but the name *infinitive* is too widespread to abandon it.

- infinitive formation to stem I is irregular.
- the infinitive to stem II is *tafilun.*
- stems III-X simply take nominal endings (for stem III, the passive participle is often substituted).

For example: *tarīḥun* 'date, history' is the infinitive to stem II. of *-r-h* ('date').
Syntax

In Arabic, a word is classified as either a noun (ism), a verb (fil), a pronoun or a preposition (ḥarf). Adverbials are expressed with nominal forms. Repetitive use of the same root in verb and noun in a sentence is considered good style, especially with derived forms of the same verb. Such as the root "`alm" which in Form I is "`alama" meaning "to know" but in form II "`allama" with the middle radical(letter) doubled, changing the meaning to "to teach". Also considered good form is constructing a long sentence joined together with connectors (Adawaat al RabT) which are like conjunctions which allow for many clauses to run on and on in the same sentence.

- For example: qara’a al-kitāba qirā’atan baṭīatan, literally: "he read the book a slow reading", i.e., "He read the book slowly". This type of construction is known as the "absolute accusative."

- The Masdar, verbal nouns which are irregular for Form I and regular for all other forms. It functions sometimes like an infinitive and sometimes like the noun which encompasses the concept of the verb.

- Active and Passive partiples, called Ism Fa’l or Ism Maf’ūl after the pattern into which the roots are put, function sometimes like adjectives, sometimes present partiles, and sometimes like nouns
such as "Doer" and "Doneto". So: kātib is "writer" and maktūb is "written".

There are many types of sentences:

- the nominal sentence, consisting of a subject then a predicate (al-bayt kabir - "the house big" viz., "the house is big")
- the verbal sentence, which usually follows the VSO pattern (yafhamu aiman al-muh.ad.arat => Ayman understands the lecture);
- the amma... fa-sentence

3.5 Arabic alphabet

3.5.1 Computers and the Arabic alphabet

The Arabic alphabet can be encoded using several character sets, including ISO-8859-6 and Unicode, in the latter thanks to the "Arabic segment", entries U+0600 to U+06FF. However, neither of these sets indicate the form each character should take in context. It is left to the rendering engine to select the proper glyph to display for each character.

When one wants to encode a particular written form of a character, there are extra code points provided in Unicode which can be used to express the exact written form desired. The Arabic presentation forms A (U+FB50 to U+FDFF) and Arabic presentation forms B (U+FE70 to U+FEFF) contain most of the characters with contextual variation as well as the extended characters appropriate for other languages. These effects
are better achieved in Unicode by using the zero width joiner and non-joiner, as these presentation forms are deprecated in Unicode, and should generally only be used within the internals of text-rendering software, when using Unicode as an intermediate form for conversion between character encodings, or for backwards compatibility with implementations that rely on the hard-coding of glyph forms. Finally, the Unicode encoding of Arabic is in logical order, that is, the characters are entered, and stored in computer memory, in the order that they are written and pronounced without worrying about the direction in which they will be displayed on paper or on the screen. Again, it is left to the rendering engine to present the characters in the correct direction, using Unicode's bi-directional text features. In this regard, if the Arabic words on this page are written left to right, it is an indication that the Unicode rendering engine used to display them is out-of-date. For more information about encoding Arabic, consult the Unicode manual available at http://www.unicode.org/

- Multilingual Computing in Arabic with Windows, major word processors, web browsers, Arabic keyboards, and Arabic transliteration fonts

**Arabic keyboard layout**

The Arabic alphabet derives from the Aramaic script (which variety - Nabataean or Syriac - is a matter of scholarly dispute), to which it bears a loose resemblance like that of Coptic or Cyrillic script to Greek script. Traditionally, there were several differences between the Western
(Maghrebi) and Eastern version of the alphabet—in particular, the fa and qaf had a dot underneath and a single dot above respectively in the Maghreb, and the order of the letters was slightly different (at least when they were used as numerals). However, the old Maghrebi variant has been abandoned except for calligraphic purposes in the Maghreb itself, and remains in use mainly in the Quranic schools (zaouias) of West Africa. Arabic, like other Semitic languages, is written from right to left.

3.5.2 Arabic Chat Alphabet : Online

Online communication is often restricted to an ASCII environment in which not only the Arabic letters themselves but also Roman characters with diacritics are unavailable. This problem is faced by most speakers of languages that use non-Roman alphabets, or heavily modified ones. An ad hoc solution consists of using Arabic numerals which mirror or resemble the relevant Arabic.

The Arabic Chat Alphabet is used to communicate in the Arabic language over the Internet or for sending messages via cellular phones when the actual Arabic alphabet is unavailable for technical reasons. It is mainly a character encoding of Arabic to the Latin alphabet. Users of this alphabet have developed some special notations to transliterate some of the letters that do not exist in the Latin alphabet.

Use:

Online communication, such as IRC, bulletin board systems, and blogs, are often run on systems or over protocols which don't support codepages
or alternate character sets. This system has gained common use and can be seen even in domain names such as Qal3ah. It is most commonly used by youths in the Arab world in very informal settings, for example communicating with friends or other youths. The Arabic Chat Alphabet is never used in formal settings and is rarely, if ever, used for long communications. The length of any single communication in ACA rarely ever exceeds more than a few sentences at a time.

**Comparison table**

Because of the informal nature of this system, there is no single "correct" way, so some character usage overlaps (eg 6, which is used for both ظ and ح). Most of the characters in the system make use of the roman character (as used in English) that best approximates phonetically the Arabic letter that one wants to express (for example, ظ corresponds to k). This may sometimes vary due to regional variations in the pronunciation of the Arabic letter (eg. ج might be transliterated as j in the Gulf dialect, or as g in the Egyptian dialect). Those letters that do not have a close phonetic approximate in roman are often expressed using numerals or other characters. These have been selected so that the numeral graphically approximate the Arabic letter that one wants to express (eg. "ع" is represented using the numeral 3 because the latter looks like a horizontal reflection of the former). Since many letters are distinguished from others solely by a dot above or below the main character, the conversions frequently used the same letter or number with a comma or apostrophe added before or after (eg. 3' is used to represent د).
Examples

thahab ma3a alree7

Standard Arabic: 甘肃省

English: Gone With The Wind

alsalam 3alikom wa ra7mato allah wa barakatuh

Standard Arabic: 甘肃省

Transliteration: Al salam alikom wa rahmato Allah wa barakatuh

ba9al aw 6ama6em

Standard Arabic: 甘肃省

English: Onion or tomato

bri6ania al3othma

Standard Arabic: 甘肃省

English: Great Britain

There are a number of different standards of Arabic transliteration: methods of accurately and efficiently representing Arabic with the Latin alphabet. The more scientific standards allow the reader to recreate the exact word using the Arabic alphabet. However, these systems are heavily reliant on diacritical marks, which may be difficult to pronounce at first sight. Other, less scientific, systems often use digraphs (like sh and kh), which are usually more simple to read, but sacrifice the definiteness
of the scientific systems. During the last few decades and especially since the 1990s, Western-invented text communication technologies have become prevalent in the Arab world, such as personal computers, the World Wide Web, email, Bulletin board systems, IRC, instant messaging and mobile phone text messaging. Most of these technologies originally had the ability to communicate using the Latin alphabet only, and some of them still do not have the Arabic alphabet as an optional feature. As a result, Arabic speaking users communicated in these technologies by transliterating the Arabic text using the Latin script. To handle those Arabic letters that do not have an approximate equivalent in the Latin script, numerals and other characters were appropriated. E.g., the Latin numeral "3" is used to represent the Arabic letter "خ" ("ayn"). There is no universal name for this type of transliteration, but some have named it Arabic Chat Alphabet.
Chapter Four

Arabic Language Features
Chapter Four

Arabic Language Features

4.1 Introduction:

There are a number of obvious differences between Arabic text and English text.

1. The characters are laid out in RTL order, with the exception of numbers, which are laid out LTR as in English. Text is right aligned on the page, and written from top to bottom (like English).

2. Letters change shape depending on context. Each letter has up to four forms: the initial form, where it is the first letter in a word, the final form, where it is the last letter in a word, the isolated form, and the medial form. This means that as a word is typed, the shape of previously-entered letters changes as well as a new letter being added.

3. Arabic and Hebrew both include diacritics. These are marks placed above or below letters which typically represent vowel sounds or other modifiers. These are primarily used for children's and religious text, but we choose to support them.

4. Arabic includes ligatures. This is the process whereby two letters printed together are replaced by a single new character. English uses the same process in typesetting (for
example, ff, fi, fl, ffl, ffi); it should be noted that Apple's upcoming Quickdraw product will include this facility for English.

5. Numbers are handled differently in different Arabic-speaking countries. Some use Hindi digits (see below), others use Arabic digits. However, numbers are displayed from left to right in all BiDi countries.

Each of these five issues will be discussed in the sections which follow.

Layout

Arabic-enabled applications and/or systems always include a logical-to-physical transformation algorithm. All text input is in *logical* order—all output in *physical* order. The physical appearance of a bilingual string can differ enormously from its internal representation. Non-contiguous cursor movement and selection and semantic re-ordering of sections of text must be taken into account by BiDi-enabled applications. The major language, or *reading order*, of a body of text also affects the logical-to-physical transformation. Figure 1 demonstrates how a single string is displayed differently depending on whether the document language is Arabic or English.

![Logical (storage) Layout](image)
Figure 4. 1. Logical-to-Physical Transformation.

Figure 2: demonstrates how changing a single character can change the physical layout of an entire line of text. In this example, the string "War and Peace" is a Latin insertion in an Arabic right-to-left paragraph. Replacing the word "and" with its Arabic equivalent effectively breaks the original insertion into two and causes the two insertions to exchange positions in the line.

War & Peace: منال

War Peace: منال

Figure 4. 2. Semantic Inversion of Language Insertions.

One can argue for many different rules as to how mixed language text should be displayed (an issue referred to as layout). Our resolution to this problem is based on the Unicode standard (see Appendix A of *The Unicode Standard, Volume One*). BiDi Windows implements this with only minor changes. In general, there have been two solutions used in the
Middle East to deal with layout, depending on whether the backing store for text is stored in logical order or in visual order. Following Unicode, we store text in logical order. You might think of this as being the same order in which you would press the keys to produce the given text. The Unicode concept is that it should be possible to display mixed-language text without carrying data about the text. Usually the Unicode layout algorithm works well, but there are circumstances in which the results may not meet the user's wishes. Partly this is because no single algorithm can satisfy all needs. As an example, consider the hyphen. In Arabic text, how should the logical string 555-1212 be displayed? If this is a telephone number, the entire number should be displayed as shown. If it is a subtraction, it should be displayed 1212-555. The computer can't tell, absent information from the user, which is required. To resolve some of these ambiguities, Unicode provides support for right-to-left marks and left-to-right marks. These are non-printing characters which tell the system how to interpret the direction of the character(s) that follow. BiDi Windows has enhanced these marks by making them optionally visible, based on a control panel selection. This is helpful to users, as otherwise it may be very difficult to understand or edit layout based on the impact of invisible marks. Unicode also calls for marks for left-to-right or right-to-left embedding, as well as a pop directional format mark to end embedding. These marks were not implemented in BiDi Windows in order to avoid the added complexity they entail. Instead, Microsoft believes that users will want finer control over character layout primarily in applications like Word Processors which carry data about text. Because of this additional data, the developer can know more about the user's
intention than when dealing with raw text. For example, in Windows Write, all neutrals are assumed to have the directionality of the keyboard language in which they were entered. This provides more intuitive results to the user than Unicode, and gives the user more control. Applications may also change the layout algorithm when dealing with formulas, as in Microsoft Excel or Visual Basic. The Unicode algorithm generates undesirable effects when displaying formulas with mixed language text. Fortunately, the BiDiLayout API permits the developer to specify the desired direction of characters, overriding its default behavior.

4.2 Contextual Analysis (Shaping)

The Arabic alphabet has thirty-six alphabets (no upper case), ten numeric symbols, and a few special alphanumeric characters. An Arabic code page usually consists of all these symbols combined with the English alphabet. Although documents and files are stored in their code page form—that is, one character code per symbol—all Arabic alphabetic characters can have up to four display representations depending on their relative position in a word: initial, final, medial, or isolated. The particular shape which an Arabic character will assume is based on its left and right neighbors. The algorithm determining which of the four shapes to use is called contextual analysis. The total number of possible shapes, or glyphs, far exceeds the number of character codes. The average number of shapes required to render good-quality Arabic text is about 250. A better system, like a high-end DTP or typesetting system, could include as many as 900 different glyphs.
Important note: When using a proportional font, different shapes of the same character have differing widths. An application which assumes each character code has only one particular width, without taking contextual information into account, cannot accurately measure the width of an Arabic string. As a result, even elementary operations like cursor positioning will fail.

Figure 4.3 below shows the four forms of the Arabic characters 'ain', 'ba', 'qaf' and 'ha' respectively.

<table>
<thead>
<tr>
<th>Initial</th>
<th>Isolated</th>
<th>Medical</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>مع</td>
<td>ع ع</td>
<td>ع ع</td>
<td>ع ع</td>
</tr>
<tr>
<td>حليب</td>
<td>ب ب</td>
<td>بر</td>
<td>بر</td>
</tr>
<tr>
<td>دمشق</td>
<td>ق ق</td>
<td>قلم</td>
<td>قلم</td>
</tr>
<tr>
<td>نبيه</td>
<td>هاءه</td>
<td>هاءه</td>
<td>هاءه</td>
</tr>
</tbody>
</table>

Figure 4.3. Contextual Analysis.

As indicated, shaping is the process of selecting the appropriate glyphs to represent a set of codes. The Unicode term for this process is rendering, and it is also sometimes referred to as glyphing. This must take into account letter shapes and also ligatures and diacritics (see below). The choice of a zero-width glyph to display RTL or RTL marks is also a shaping issue. Another shaping issue is symmetric swapping. This is the process of changing the direction of a neutral character depending on language. The characters concerned are < and >, [ and ], ( and ), and { and }. In Latin, we write the logical string 3 < 4 as 3 < 4, while in RTL, it
is displayed as $4 > 3$. Because it is read right to left, the shape of the less than symbol has changed. It should be noted that when selecting a BiDi keyboard, the positions of the open and close parentheses are swapped, so the upper-case zero becomes the open parenthesis. This causes it to display with the same visual direction as the keycap. This was done to be consistent with user expectations. BiDi Windows handles the large number of glyphs required by always having two fonts available, one Latin and one BiDi. Numbers and several neutrals are present in both fonts. For example the English % (percent) is displayed as % in the Arabic font. In cases where two versions of a glyph are available, the one from the Latin font is used if the character is being laid out LTR, while the one from the BiDi font is used if the character is being laid out RTL.

**Diacritics**

_Diacritics_ are diacritical marks for Arabic alphabetic characters which primarily represent vowel sounds. There are eight different diacritics. Diacritics are displayed above (single or double diacritic) or below (single diacritic) any of the thirty-six Arabic characters. (The diacritics are not the same as the dots appearing above and below some Arabic letters, which are part of the letter). Our Arabic keyboard requires a separate keystroke to type diacritics. Characters are typed first, followed by the diacritic(s). Diacritics are superimposed on the character. Complex grammatical rules come into play to select the correct diacritic. Daily documents like newspapers do not use diacritics except in a few cases to disambiguate identical words. See Figure 4.4
Figure 4. 4 Effect of Diacritics on the Meaning of a Word.

<table>
<thead>
<tr>
<th></th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single diacritics:</td>
<td>مَهَّهَهُ</td>
</tr>
<tr>
<td>Double diacritics:</td>
<td>ءُلُّعُلُّ</td>
</tr>
<tr>
<td>Shadda:</td>
<td>سُ</td>
</tr>
<tr>
<td>Combined diacritics:</td>
<td>مَتَأ، تَأَءَصَأَءُ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>دَرَسَ</td>
<td>he studied</td>
</tr>
<tr>
<td>دَرَسْ</td>
<td>a lesson</td>
</tr>
<tr>
<td>دَرَسَ</td>
<td>he taught</td>
</tr>
<tr>
<td>دَرَسْ</td>
<td>it was studied</td>
</tr>
</tbody>
</table>

Vertical kerning of diacritics is a desirable feature as it is considered primitive to display diacritics all at the same height. See Figure 5.

Figure 4. 5 Vertical Kerning of Diacritics.
BiDi Windows handles placement of diacritics based on tables built into GDI. With certain decorative fonts, this causes diacritic placement to be sub-optimal. A more sophisticated technique would be to carry diacritic placement information within the TrueType font itself. BiDi Windows did not take this approach, partly because it still supports bitmap fonts.

**Ligatures**

Ligatures are combinations of two--and sometimes three--characters into one shape. See Figure 6. Ligature selection is dependent not only on the characters themselves but also on the selected Arabic font. Some fonts do not use ligatures at all and others may have as many as 200 different ligatures defined. Note also that ligatures affect the positioning of diacritics.

![Figure 4.6. Arabic Ligatures.](image)
Ligatures found in the Traditional Arabic font:

Because of font page limitations, these are the only ligatures supported by Arabic Windows.

**Number Formats**

Some Arabic countries require a different shape for digits, called *Hindi* digits. Therefore, digit shapes are a property of countries and not language or font.

![Number Representation](image)

**Figure 4.8 Same Number with Different Representations (including punctuation).**

Numbers are always entered and read from left to right, just like Latin text. Because numbers are entered and stored most-significant digit first, *push mode* support is required by Arabic-enabled applications (see next
section). Likewise, to display numbers properly, a logical-to-physical transformation is required.

Figure 4.9. Arabic String Display with Numbers.

4.3 Design of BiDi Software

BiDi Text Display

Bi-directionality and language insertions

If Arabic documents contained only Arabic text, life would be easier. In practice, one basic requirement of BiDi software is its ability to handle input, display, editing, and printing of Latin text (such as English or French character strings). Arabic documents can, and often do, contain Latin text. Although Arabic is read from right to left, Latin strings are still read from left to right, as are Arabic numbers. It is common to have documents where the page is divided into two columns, the left column being Latin (left-to-right orientation) and the right column Arabic (right-to-left orientation). Such a feature is available in competitive word processors and is referred to as dual column. Lines, paragraphs and documents always have a language attribute which denotes whether this
is a right-to-left text element or a left-to-right text element. Strings in the 'other' language are called **language insertions**. To avoid typing Latin text in reverse, Arabic-enabled software uses a concept called *push mode* whereby the cursor remains stationary and the Latin string is 'pushed' to the left to make room for the next character, like a calculator display. When the user finishes typing the Latin text (signaled by changing keyboard language), the cursor moves to the extreme left of the Latin string, waiting for an Arabic character. See Figure 9, which demonstrates a progressive language insertion of the word "Latin" in an Arabic string. The numbers near the bottom of the figure indicate the input sequence of the individual characters.

![Figure 4.10 Latin Insertion in an Arabic Line.](image)

11 12 13 14 10 9 8 7 6 5 4 3 2 1

**Figure 4.10 Latin Insertion in an Arabic Line.**
Text editing

Routines which edit text need to be aware of special BiDi requirements. First is the caret, which should have a direction pointer showing the selected keyboard language. Second, highlighting and caret movement are based on logical text, not visual (some applications, like Windows Write, offer visual caret movement as an option). This means that as the right arrow key, say, is repeatedly pressed, the caret moves through the logical text. In a LTR paragraph, this will cause it to move to the right in Latin text and to the left in BiDi text, jumping when it hits boundaries. Because of the discontinuous layout, highlighting of a logical sub-string can cause multiple visual blocks to be highlighted (up to five). Note that justification is handled differently in Arabic than English; text is justified by stretching the connecting lines between certain letters rather than by stretching the spaces between words (this is called \textit{kashida justification}, based on the name of the stretcher character).

Text selection

Text selection, typically implemented by marking a body of text using a highlighted block, is complicated when the internal and physical representations of text differ. Currently, all Arabic-enabled software implements \textit{logical selection and marking} where the selected block may be non-contiguous (see Figure 10). A sophisticated application might also wish to implement \textit{physical selection and marking} where the operations are intuitively more apparent to the users. Such a solution requires some substantial changes to the underlying software. In the upper example of
Figure 10, representing text in LTR reading order, the user started a block mark at the word 'text' and dragged the mouse to the right, to a position just between the first two Arabic words. Because the highlight crossed a language insertion, the resulting selection is non-contiguous physically, but contiguous logically.

Figure 4.11 Logical and Visual Block Selections.

Text justification

A Latin paragraph is justified by inserting spaces between words. Unlike Latin lines, which are justified by adding extra white space between words, Arabic line segments are justified by stretching the connectors between certain letters inside words. This stretching is done on computers by adding an extra character called a \textit{kashida} which looks like a horizontal connecting line. See Figure 11. Extending characters follows Arabic calligraphic rules which favor some characters to be extended over others. Like ligatures, kashida use is a property of the font, as it is incorrect to use kashida justification for some Arabic fonts. This English paragraph is justified by extending the spaces between words so that words touch the left and right margins.
BiDi applications should in general be modified so they have a "right-to-left feel." Examples would be that text will be right aligned, scroll bars on the left, and so forth. At the same time, an application may be written so that it still supports Latin text, in which case it may require two modes of operation. Microsoft Excel is an example of this, offering both RTL and LTR spreadsheets. Dialog boxes for localized applications should be laid out with design elements going RTL, and RTL controls should be used. Edit controls support both RTL and LTR text entry and display, user selectable by using the CTRL+left SHIFT and CTRL+right SHIFT key combinations. Pressing these keys does three things:

- Makes the text align to the appropriate side of the edit control
• Changes the reading order to match the text direction
• Changes the keyboard language to match the selected direction

The requirements for enabling the main window of an application must be designed carefully, and no general guidelines can be given. Study of the applets included in BiDi Windows is recommended. An example would be Windows Paintbrush: the tools were switched to the right and their column order reversed. The vertical scroll bar was placed on the left and the horizontal scroll bar defaults to right aligned. Text is displayed to the left of the selected point, and mixed text is supported, along with use of the dual font dialog. The icon for text entry was changed to display Arabic letters instead of English ones. Internal code changes were also required in order to display text properly.

Other Design Issues

Right-to-left editing

Document origin for Arabic text is at the top right-hand side of the page. Other modifications required for Arabic text editors:

• Rulers should originate on the right-hand side and read from right to left
• The first column of tables or columnar text is on the right, last on the left. Spreadsheet columns originate at the right-hand side of the page
• Arabic paragraphs can be right justified or left justified, but still read from right to left
• The first item of a menu bar should be on the right-hand-side, and the vertical scrollbar should be on the left-hand-side

This does not mean one needs to switch everything; we don't swap the system menu or minimize/maximize icon locations. After all, the Arabic model is that of a right-to-left document, not a right-to-left desktop.

**Dates**

Arabic countries use both Gregorian (western) and *Hijri* (Islamic) dates. Hijri date is the official date in Saudi Arabia, our largest Arabic market. A conversion and selection utility is required for most applications with an option to display the date in either Gregorian or Hijri formats. Four different sets of month names are used and applications need to provide for a user selectable set of names. Month names can be either Gregorian or Hijri months, and either the Latin or Arabic alphabet might be used to write them. At the same time, an application should respect the Date preferences given by the user in Control Panel.

**Keyboard input**

Arabic Windows provides a bilingual keyboard with access to both Latin and Arabic keys. The user can switch keyboards either by pressing a keyboard shortcut (ALT+left SHIFT for Latin and ALT+right SHIFT for Arabic) or by clicking on the L or Ayn icons which have been added
to the menu bar of the active window. As mentioned above in "Contextual Analysis," directional neutrals change shape when the language changes. Also, as mentioned in "Text Editing," arrow keys work visually "backward" when moving the caret through text of the opposite language from the reading order. However, in an application like Microsoft Excel, direction keys such as arrow keys should move in the expected direction irrespective of the sheet direction, even though movement keys such as HOME or END will now go to the (right-hand) origin or (left-hand) "end" of a RTL sheet. BiDi Windows supports use of ALT+accelerator keys for menus and dialog boxes irrespective of the current keyboard language. Designers should avoid using an accelerator key that is part of a Lam-Alif ligature.

**Spelling and grammar checker (automatic diacritizer)**

Spelling mistakes are rare in Arabic due to the nature of the language. Words are spelled as they are pronounced, thus the necessity of a spelling checker is limited. However, grammar is much more complex especially when using diacritics; semantic parsing of the sentence is required to define the correct diacritics. A useful product is an *automatic diacritizer* that can put diacritics on a sentence after grammatically checking the sentence's semantics. An Arabic grammatical checker is a complex program which needs a lot of research and linguistic resources.

**Searching and sorting**
Sorting Arabic text is based on the code page and not the font page. However, there are a few Arabic characters which have the same weight and should be treated identically in sorting unless they are the only difference between two words. Kashidas should be ignored in sorts, and applications should give users the option to ignore diacritics. Ideally, users should be given the option whether to sort Latin characters before or after Arabic.

Chart types and direction

Chart types are the same as those used for Latin. In general, charts are laid out right to left in Arabic and left to right in Hebrew, but it is best to provide users the option to change direction; for example, scientific charts in Arabic still have the X-axis increase LTR.

Macros

It is an open question whether macro languages should use terms in Latin or the BiDi language. The benefit to Latin terms is the availability of documentation and the fact that there are no standard terms in BiDi languages for certain issues. The benefit of local language is that it will be understood by a wider range of users. Providing a bilingual solution is optimal.

Mathematical and financial formulas

Formulas in Arabic are written from right to left, while in Hebrew they are written left to right. The issue of function names is the same as that
for Macros (see above). However, widely used function names (such are sin, cos, and so forth) are generally widely familiar in the BiDi world. Programmers in the BiDi world are used to programming languages with verbs in English.

4.4 Technical Issues in Developing BiDi Software

Transparency

Applications are divided into aware applications, for example, those written specifically to run under BiDi Windows, and unaware applications, which were written to run under Windows with no knowledge of BiDi. Developers can provide a substantial amount of BiDi support in many unaware applications through use of transparency. This refers to BiDi Windows' modification of standard Windows APIs so they provide BiDi support. Examples include ExtTextOut and DrawText, which under BiDi Windows will successfully display strings of mixed text. The other aspect of transparency is enhanced controls (especially single-line edit and multi-line edit) under BiDi Windows. Together, these aspects of transparency provide enough BiDi support to allow many English applications to be used with BiDi data under BiDi Windows with no modification. When using an unaware application, the user can typically select a font to apply to some text. However, BiDi Windows allows the user to switch languages at any point in time. It handles this situation by making available a default font in the non-chosen language. Clearly this must be a scaleable font so it can be matched in size to the selected font. The
default BiDi font is selected in the WIN.INI file; the system default font is used for a default Latin font. Applications which do not do complex text handling, and which use standard Windows controls and **ExtTextOut**, can be used in the BiDi market simply by localizing their strings and re-drawing their dialogs to use BiDi controls (see next section). On the other hand, word processors, desktop publishers, and so forth, will require careful rewriting to handle the painting and processing of text. Transparency does not handle kashida justification.

**String Objects**

In order to provide support for certain applications, BiDi Windows contains a set of APIs supporting string objects. These provide a mechanism for encapsulating the details of text handling and making it easy for the programmer to handle editing tasks such as caret placement, text highlighting, and so forth. These string object APIs are discussed in the Windows SDK Hebrew Supplement. A string object contains text with a single Latin font and style and a single BiDi font and style. They work only for screen display, and do not support the display of italics, OEM or Symbol fonts, and so forth. Although simple to use, they are very slow compared to using the complex APIs (see below).

**Code Page and Font Issues**

Because Arabic speakers in North Africa typically speak French, the Arabic code page does not overlap with French characters. The problem this creates is that although there are enough code points to represent both
Latin and Arabic characters in 256 codes, this in an inadequate number to hold both the differing shapes and ligatures. The solution implemented in BiDi Windows is to have two font pages always available, one ANSI, and the other Arabic (or Hebrew). This also allows neutrals (characters not clearly associated with a language, such as % or &) to be represented in both fonts. Even 256 font points is not enough to display all the Arabic letter shapes, ligatures, neutrals, and so forth, so certain compromises had to be struck. This was done primarily by having two separate font pages for Arabic, referred to as Simplified Arabic and Traditional Arabic. Simplified Arabic has only one ligature, "Lam-alif," which is required. This leaves room for the font to hold both neutrals and Arabic (or Latin) numerals as well as Hindi Digits (or Arabic style numbers. This naming is unfortunate but is traditional). We identify each type of font by a unique character set. Aware applications have access to a dual font picker which allows the user to select both a Latin and a BiDi font, each with the size and style preferred. It is the application's responsibility to measure and paint text using these fonts. Arabic Windows includes a Control Panel setting which allows the user to select preferred display formats for numbers: Arabic, Hindi, or context. The transparent APIs handle the appropriate display automatically; complex applications (those not using transparency or String Objects to display text) will in general also get the appropriate character type by calling BiDiFontComplement. Another font issues has to do with the inconsistent appearance of Arabic and Latin fonts of the same height. Fonts are matched based on height (not point size), but Arabic fonts need space above and below the characters to hold diacritics. As a result, Arabic text of a given height appears much smaller
than Latin text of the same height. As a partial solution to this problem, Arabic Windows includes the Arabic Transparent font, which has no diacritics, and hence can be a better match in terms of size. Even this font, however, appears small compared to Latin text of the same height. This is especially a problem in running transparent applications, which may use 10-point (Latin) text for display, causing the Arabic display to be in 8-or 9-point text, which is barely legible on low-resolution displays. To summarize, then, there are two fundamental ways in which Arabic text is different from Latin text which makes Arabic text handling more than just a localization issue. Each of these is based on an assumption so deeply ingrained in Latin programmers that they do not realize they are making it, and which it is our task to break to support BiDi text:

1. The assumption that the character (backing store) sequence is the display sequence, and
2. The assumption that the code points have a one-to-one relationship with the font points.

An unfortunate corollary is that making each of these assumptions saves time, so that BiDi text processing is inevitably slower than Latin text processing.

Localization

The simplest part of RTL support is within dialog boxes. BiDi Windows has added additional flags to most controls to support BiDi features. For example, radio buttons or check boxes can have the text on
the left of the button, right aligned. List boxes and combo boxes can have
text right aligned with scroll bars on the left. These flags can be modified
directly in the resource file or the BiDi Dialog Editor from our SDK can
be used to paint the controls. Similarly, menus can be made RTL, either
statically or by calling the API \texttt{MakeMenuRtoL} (there is also an API
named \texttt{MakeScrollBarsRtoL} for windows with scroll bars). In order to
support use of the bilingual user interface, a new entry has been added to
the resource file, LANGUAGE ARABIC (or whatever). The resource IDs
for different language versions of a given resource should be identical for
each of the supported languages. The system will take care of changing
resources. However, the application is responsible for intercepting
\texttt{WM\_LANGUAGE} messages and taking appropriate action if required,
such as changing dynamically modified menus or dynamically displayed
strings. More details on multilingual development are available in the
SDK manual.

\textbf{Complex APIs}

Code which cannot use transparency to handle text, such as word
processors, use the \textit{complex APIs}. These include \texttt{BiDiLayout},
\texttt{BiDiFontComplement}, \texttt{BiDiShape}, and \texttt{BiDiSetKashida}. Despite their
name, these APIs are not especially complex once you have become
familiar with them. They provide detailed support for text handling of all
types within a Window. Typically the actual display of the text is done
using \texttt{RawExtTextOut} using the glyph array produced by \texttt{BiDiShape}.
Each \textit{run} (string of text sharing common language and font attributes)
must be printed separately, so it is the responsibility of the programmer to
loop through the text rather than treating it as a single homogeneous string. **BiDiShape** also returns an array of glyph widths which can be used to calculate caret placement, highlighting, location in text of a mouse click, and so forth.

**MS-DOS Files**

Unfortunately, there is no standard for MS-DOS files in the Arabic world. Instead, a variety of code pages are used by different layers (TSRs which provide BiDi support to MS-DOS), and different font pages are used for Arabic as well. Some of these are logical, and some are visual. Even Microsoft's Arabic DOS has a different code page from Arabic Windows. As a result, it is necessary to convert files when passing them between MS-DOS and Windows (the Windows MS-DOS box, when windowed, also does this conversion on the fly). Applications (such as word processors) which might read MS-DOS files, should provide conversion support as part of their File Open dialog (and perhaps File Save dialog). Windows provides APIs which support this conversion, namely **BiDiAnsiToOemCode** and **BiDiOemCodeToAnsi**.

**MS-DOS Applications**

In general, the way MS-DOS is supported in the Middle East is by loading a layer, or TSR which handles the layout and shaping issues. Microsoft sells such layers. The most common Hebrew layers save text in visual order and have the Hebrew alphabet beginning at location 80 (hex) in the code page. In Arabic countries, there are a variety of competing
layers. Some applications are designed to run on U.S. MS-DOS and essentially have their layer bundled into the application (MS Hebrew Works for MS-DOS is an example). BiDi Windows provides support for these layers by having special MS-DOS box handling. When running a windowed MS-DOS box, the user can set the code page and font page, and text will be displayed appropriately. Copy and Paste functions perform code page conversion on the fly. Multiple MS-DOS boxes can run concurrently with different code pages. Thus Windows serves rather as a lingua franca for the conflicting MS-DOS applications in the region.
Chapter Five

Arabic on the Internet
Chapter Five

Arabic on the Internet

5.1 History of Arabic on Computers

This chapter provides a historical overview on how computers handled Arabic in the past, and how it is handled at present. Arabization in the computer world was initially a hardware vendor activity. Vendors either had in house resources (Systems Engineers), or subcontracted others to do the job for them. Some vendors established centers of expertise to develop Arabic, such as the NCR-ALIF (Arabic Language Integration Facility) in Nicosia, Cyprus then in Cairo, Egypt. IBM and ICL had similar facilities. Each major vendor had several standards as improvements were gradually made.

bit In Lower Case

Initially, Arabization was very rudimentary. The lower case English characters were replaced by the Arabic characters, and were stored as 7-Bits. Lower case English was not used much in computers in the 1960s and 1970s, until UNIX came to being. Some printers were not even capable of printing lower case English letters. This made the replacement of lower case somewhat acceptable in those days. Texas Instruments minicomputers used this scheme for the Arabization in Egypt in the early 1980s.
Limited Character Set

8-Bit character sets were then used. These used the upper part of the ASCII table (128 decimal/80 Hexadecimal and above). Initially, this was somewhat primitive, with just one representation for the several different shapes an Arabic characters could have. For example, The second letter in the Arabic alphabet, the ب BA character and its like (TA ت, THA ث) were only represented as the beginning of word shape, which also functioned for the end of word shape as well. An example of this era is **NCR-64**. It provided limited shaping, the above three letters had only one variant, instead of having separate shapes for beginning, middle and end of word. The 'Ein غ and Ghein غ letters only had two variants, not the usual four variants for them. Moreover, this character set catered for Farsi (Persian) as well, and had the 4 extra letters for Gaf, Pa, Ja, Tch sounds. It also had a sorting problem: for some unknown reason, it had the Waw ظ before the Ha’ُ too!

Enhanced Character Set

As improvements were done, richer character sets came into being, which had more shapes, and thus was more visually acceptable. An example is **NCR-96**. Storage was still in 7-bit most of the time, and terminals used SI/SO (control characters Shift In, Shift Out) to switch languages. Auto shaping was done in the ROM of terminal, which was a step forward. For printing, applications had to call special system routines to do the automatic shaping (called **Context Analysis** at NCR). Still, every vendor
had their own character sets, and interchanging data was quite a chore, just like the pre-ASCII days in the 1950s and 1960s in the West.

**IBM EBCDIC and its idiosyncrasies**

IBM, since they use EBCDIC had a peculiar character set, where Arabic and English were interspersed. The detection of whether the 8th bit was on or off was not a guarantee that this is an Arabic letter. Sorting was problematic too.

**Standardized Character Sets**

Finally, the era of standards arrived. These were cross-vendor standards accepted by the industry, and set by standard bodies.

**ASMO 449**

One of the first vendor independent standards was ASMO 449 (Arabic Standards and Measurements Organization). However, it was still 7-bit, and required escape characters (normally the plain text braces { and }, which was kind of odd to use as escape codes on the terminal).

**ASMO 708**

By the time ASMO-708 was introduced, things were getting better. This was a true transparent 8-bit vendor-independent standard. Terminals and Terminal Emulators were now sophisticated enough to do the shaping and work in 8-bits, distinguishing between Latin and Arabic automatically by whether the 8th bit is set or not, as well as obeying certain escape
sequences to shift the keyboard language/direction as well as the screen language and direction. Printing was done in the firmware of the printers (e.g. from Alis). Where there were no Alis printers, the UNIX spooler did the work through a custom filter written in C that does the context analysis. ASMO 708 was adopted by the International Standards Organization as the ISO 8859-6 standard. Still in some cases, vendors did silly things: NCR and ICL each had their own ASMO-708 derivative. For example, in Egypt customers insisted that the Lam-Alef ﻃ is one character, and not stored as a separate Lam ﻟ and Alef ﺟ! Also the Ya’ ﻳ at the end of a word, with two dots under it was virtually unknown in Egypt, and the version without the dots (actually and Alef sound) was used instead. This led to NCR-ASMO and such vendor variants.

Other approaches

ICL DRS-80 system

Even in the mid 1980s, some companies did get Arabization right, such as ICL (later bought by Fujitsu). They had this DRS-80 system which had totally seamless Arabization. One could enter Arabic right in the program code using the editor, be it COBOL or BASIC, a feat which was uncommon in other systems in those days, where development environments, editors and other programmers’ tools could not handle 8-bit input. This system also had other features, such as calling programs from one language to another language, such as COBOL and BASIC, and the ability to recover from power failures via a dump facility.
The advent of the PC in the early 1980s also influenced how Arabization (and other internationalization) is to be done. Several code pages were developed by Microsoft, including CP 720 for Arabic DOS. Later, when Windows arrived, another code page was developed, Windows CP 1256.

Unicode

Unicode finally arrived. It was an international standard, and neither vendor specific nor developed by Arabic standards organizations. The scheme required 2 bytes to represent every character on earth, from Arabic to Chinese.

UTF-8

Due to the limitations of Unicode requiring 2-bytes, some scientists developed a scheme called UTF-8, which is basically Unicode that allows for existing ASCII to be encoded in just one byte (8-bits). You can read How UTF-8 was born, by Rob Pike. Also, Joel Spolsky has an article on the use of Unicode in software from a developer's perspective. Remember that this is just a standard for encoding, and not presentation. Therefore, it is not Arabic specific, nor does not address the problems with direction, but it is a standard nonetheless, and hence very useful in general. Still other developers have proposed other encoding systems.
5.2 Microsoft and Arabization

Microsoft, being the main producer of operating systems and applications for personal computers for about a half century, had a lot of impact on Arabic and Computing. This article details some of that impact, both from the technology point of view, as well as from the business angle.

The Beginning

Initially, Microsoft did not supply Arabic for their MS-DOS operating system. They seemed disinterested in the market at first, and users often complained about this lack of attention. Microsoft attributed that to the size of the market being too small to warrant attention. This piece of the market left to independent software vendors (ISVs). Some like 01 Systems had good commercial products, such as al-Nafitha, which allowed the user to select from several different code pages. Various small vendors in the Arab region produced their own Arabization, as well as drivers for printers, ...etc.

Microsoft Competes

When Windows 3.x arrived in the early 1990s, it was not Arabized at first. Making it Arabic-capable was left again to ISVs. Sakhr, which was then a subsidiary of the Kuwait based al-Alamiah, was first with an Arabization for it. Again, this was customizable, allowing various character sets and code pages to be selected. It is at this point that Microsoft started to behave in a predatory manner, with the monopolistic tendencies it has shown in other segments of the IT market. Microsoft
lured the lead technical architect of Arabic at Sakhr to become one of its own staff. I do not recall his exact name, but he seems to be of Syrian or Lebanese origin, and with a Western first name, but Arab family name. Sakhr sued Microsoft for this, but the lawsuit was later settled.

**Microsoft Becomes a Monopoly**

From that point on, Arabization of Microsoft Operating Systems became solely a Microsoft game. ISVs continued to provide applications and peripherals, such as terminal emulators, printers, and such, but the platform itself was not an area open for competition. Arabized versions of operating systems (e.g. Windows 95 and 98) and applications (e.g. Word) were released at a later date than the original launch of the product, often a year late. In the 1980s, up to the mid-1990s, Microsoft did turn a blind eye for rampant piracy in the region. This strategy helped popularize its products with end users and developers alike, since they were virtually "free". This is similar to what happened elsewhere. **Bill Gates was quoted** as late as 1998 that if companies were to pirate software, he would rather that they did so with *his software* than with the competition.

**Calling in the Inquisition**

By the mid 1990s Microsoft began to assert its "intellectual property" rights, and pushed the law enforcement authorities in the region to make sure that piracy is fought. These actions ensured Microsoft total dominance and monopoly status. Arabs are currently around 300 million, and Arabic is the 5th most widely spoken language in the world. For a
company to gain a monopoly on an entire culture like this is simply wrong and unacceptable, but it did happen.

**On the Internet**

On the internet front, the Arab world joined a bit later than the rest of the world. The Microsoft monopoly on the desktop caused this type of monopoly to migrate to the internet as well. By that time, Netscape Navigator was killed off by Microsoft bundling Internet Explorer for free in its systems. Moreover, Netscape Navigator never had good Arabic support. Again, some ISVs did release Arabic add ones for it, but they were not very good. Also, many developers knew only Microsoft products, and thought that the world is only Microsoft. They not only use Microsoft Windows 2000 or Windows XP as an operating system, but *must* also use IIS, ASP, and MS SQL as well. If this was on the server side only, it would not bother me one bit. The problem comes when these developers assume that *all users have Microsoft operating systems and Internet Explorer* and nothing else exists. I once had a debate with a developer of Arabic technology at a major vendor, when he said that his application worked only with Microsoft Internet Explorer 5.x or better. I asked "What about people who use Mac?". There was no good answer for it except that this is the way it is!

**Some Hope for Interoperability**

First, Linux distributions, such as Mandrake 10.x, Fedora, and others do support Arabic to varying degrees. Konquerer, the browser of the KDE
desktop product is also Arabic capable. Fire Fox 1.0 does support Arabic in an acceptable fashion, both on Windows and on Linux. The reasons for the move are neither technical, philosophical nor ideological. They are pragmatic, and mainly due to the poor security model of Microsoft ActiveX that is in Internet Explorer, causing the current epidemic of spy ware, malware, Trojans, viruses and what have you. Therefore, people are moving towards Fire Fox, Opera and alternative browsers, as well as to Mac OS X and Safari out of disgust with Microsoft insecurity. Some sites are still Microsoft centric and would not display properly except in Microsoft Internet Explorer. Yet some of the worst offenders realize this problem, and are slowly moving to standards. Al Jazeera used not to display properly, but as of the fall of 2004 has converted to a new Content Management System (made by Microsoft) that does a good job on non MS browsers and platforms. When I browse a site that I think is important to others, but does not function in Fire Fox, I take the time to write to the site's webmaster about the problem. Sadly, at many sites, I never get a reply, and often I get a bounce on the email because the mailbox is full or something silly like that. I have discussed a Arabic Browser Hall of Shame with the Egyptian Linux User Group to try to gain attention to this problem. Perhaps this will work where other methods did not.

5.4 The Issue of Platform and Browser Independence

A sizable population of web developers write their sites pretending that only one browser exists (Microsoft Internet Explorer) on one Operating System (Windows). This is a terrible mistake that even large web sites
fall into. Even Al Jazeera did so until they overhauled their site in late 2004. Such web developers are really locked into a "Microsoft-only world", and really ignore recognized standards set by W3C and other organizations. Until recently, this was a serious limitation for people who do not use Windows, such as those who like Linux, the Apple Macintosh, or other web sites. It was also a limitation for people who do use Windows but do not have Arabic Windows installed, such as researches and scholar in the West, or Arabs living in non-Arab countries. However, when Windows 2000 and XP came out, they solved this problem, by not requiring a dedicated Arabized version of the operating system.

**A picture with Arabic in Graphics**

The only truly portable form of Arabic on the internet, is a picture with Arabic in Graphics. This means that the page to be published can be scanned to a .GIF or .JPG file and published as an image. This can be read by any graphical browser on any platform. Even if the user is an Arabic speaker who happens to be temporarily abroad, using UNIX or Apple Macintosh, he will be able to read the site. There is no need for an arabized version of the Operating System, or a specifically arabized browser. There are several sites that use this technique successfully, such as some Egyptian Newspapers, and a very funny joke sites, containing lots of Upper Egyptian Jokes at Nokat Nos Com. The **advantages** of this solution are:

- Works from any graphics-capable browser worldwide.
- Independent from platform and/or Operating System.
• No need for any plug-in or font.
• Does not suffer from belonging to one Arabic code set standard or the other.

Of course, this is not a perfect solution, and has its disadvantages:

• Arabic is not represented by a code set.
• Cannot be indexed on keywords, and therefore cannot be searched by normal search engines.
• Can be slow loading if the page contains color, ...etc.

There is a brief discussion of why you should not do it the way I suggest at this IHorizons Page. I am presenting it here, since they are all valid points. The key issue here is that different code sets are used for Arabic, and this becomes very painful for software companies, web publishers, and in turn end users. Here is an article by a Lotus Arabization expert - Hussam Eid describing the issue and urging Arab standard bodies to develop a unified standard for Arabic character sets. Until this happens, images are the most portable way - sorry!

5.4 Transliteration: Using Latin Characters for Arabic

Can you understand this?

"9aba7 el kheir...Ya alf nhar abya'9"

This says in Arabic : يا صباح الخير ... يا ألف نهار أبيض :
Which is actually a newly emerging "folk-standard" that has been evolving in the later half of the 1990s, whereby the Arabic language can be represented using the Latin alphabet. This seems to have started using Usenet and E-Mail, but really caught on the chat services: mainly IRC and ICQ. A language is mainly phonetic (sound-based) and most languages of the world have not been represented in writing until recent times, and therefore, representing languages in different character sets is possible. It has been used in the past due to various reasons (political, ease of use, ...etc.) Recent examples of this include the representation of the Turkish language in a Latin-based script, instead of the Arabic script used for hundreds of years, and the Malay language used to be represented using both Arabic and Latin scripts. There are a few variations and not a single standard, depending on the virtual community that evolved this standard. Therefore different networks in IRC may use different standards, and even different rooms may have their own dialects. However, there is a lot of "cultural borrowing" going on as people span networks and rooms. This is a perfect example where the Internet is giving rise to new cultures.

To predict the future: Will we ever see the day where Al Ahram or Ash-Sharq Al Awsat Newspaper published in Latin characters? Maybe not in the near future, but yet again, maybe sooner than you expect...

5.5 Arabic on the Web

- Arabic 2000: How to read Arabic on W3? - This is a comprehensive page covering all platforms including Windows, Macintosh, and UNIX/Linux. Readable in English.
• Internet Horizons: How to view Arabic - Another comprehensive page covering all platforms. Readable in English
• Assr.org: How to read Arabic on the Internet - A summary page in English.
• Al Saha Arabization Web Page - This page addresses Microsoft Windows only. It is in Arabic, and is readable using any Graphics-capable browser on any platform.
• Linux4Arab How to view Arabic Text - A page with a collection of links for viewing Arabic on different platforms.
• Nicholas Heer's Page on HTML and Arabic - A very comprehensive set of links on Arabic and Persian HTML for various platforms, including DOS, Windows, Linux, Mac, ...etc.
• Yamada Arabic WWW Guide - Arabic Links to Fonts, Newsgroups, some web sites, ...etc.
• Non-English and the Net - Written by Knut S. Vikom Norway. A survey on problems and solutions of Non-English languages (including Arabic) on the net. Focuses on Mac and Eudora.
• Some links about Arabic on the net by Ivo Spira. Mainly for non Arabic speakers who are studying the language, not computer specific.
• Arabic, The Big Problem is an article by Mohamed Sameer on issues concerning Arabic in software.

Specific Platform Issues

• Arabic Mosaic for Linux and UNIX - Langbox, a company specializing in UNIX arabization, has an Arabic/Latin Web
browser for various UNIX and Linux platforms. They also have a brief discussion on various Arabic character sets, and other Arabization issues on the Web. It is a bit out of date but still very useful.

- Ahmed AbdelHamid ACON - Arabic Console for Linux.
- Mozilla Bi-Di Support Project - This is a project aiming at providing Mozilla (which is an Open Source version of Netscape) with bidirectional (Bi-Di) support for languages that use right to left direction (Arabic and Hebrew).
- Mac Arabic on the Net - A discussion on using Arabic Macintosh on the Internet. Very useful for Mac lovers.
Chapter Six

Information Retrieval
6.1 Introduction

Information retrieval is a wide, often loosely-defined term but in this chapter I shall be concerned only with automatic information retrieval systems. Automatic as opposed to manual and information as opposed to data or fact. Unfortunately the word information can be very misleading. In the context of information retrieval (IR), information, in the technical meaning given in Shannon's theory of communication, is not readily measured (Shannon and Weaver). In fact, in many cases one can adequately describe the kind of retrieval by simply substituting 'document' for 'information'. Nevertheless, 'information retrieval' has become accepted as a description of the kind of work published by Cleverdon, Salton, Sparck Jones, Lancaster and others. A perfectly straightforward definition along these lines is given by Lancaster: 'Information retrieval is the term conventionally, though somewhat inaccurately, applied to the type of activity discussed in this volume. An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.' This specifically excludes Question-Answering systems as typified by Winograd and those described by Minsky. It also excludes data retrieval systems such as used by, say, the stock exchange for on-line quotations. To make clear the difference between data retrieval (DR) and
information retrieval (IR), I have listed in Table 1.1 some of the distinguishing properties of data and information retrieval. One

**Table 6.1 Data Retrieval or Information Retrieval?**

<table>
<thead>
<tr>
<th>Matching</th>
<th>Exact match</th>
<th>Partial match, best match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inference</td>
<td>Deduction</td>
<td>Induction</td>
</tr>
<tr>
<td>Model</td>
<td>Deterministic</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>Classification</td>
<td>Monothetic</td>
<td>Polythetic</td>
</tr>
<tr>
<td>Query language</td>
<td>Artificial</td>
<td>Natural</td>
</tr>
<tr>
<td>Query specification</td>
<td>Complete</td>
<td>Incomplete</td>
</tr>
<tr>
<td>Items wanted</td>
<td>Matching</td>
<td>Relevant</td>
</tr>
<tr>
<td>Error response</td>
<td>Sensitive</td>
<td>Insensitive</td>
</tr>
</tbody>
</table>

may want to criticize this dichotomy on the grounds that the boundary between the two is a vague one. And so it is, but it is a useful one in that it illustrates the range of complexity associated with each mode of retrieval. Let us now take each item in the table in turn and look at it more closely. In data retrieval we are normally looking for an exact match, that is, we are checking to see whether an item is or is not present in the file. In information retrieval this may sometimes be of interest but more generally we want to find those items which partially match the request and then select from those a few of the best matching ones. The inference used in data retrieval is of the simple deductive kind, that is, \( a R b \) and \( b R c \) then \( a R c \). In information retrieval it is far more common to use inductive inference; relations are only specified with a degree of
certainty or uncertainty and hence our confidence in the inference is variable. This distinction leads one to describe data retrieval as deterministic but information retrieval as probabilistic. Frequently Bayes' Theorem is invoked to carry out inferences in IR, but in DR probabilities do not enter into the processing. Another distinction can be made in terms of classifications that are likely to be useful. In DR we are most likely to be interested in a monothetic classification, that is, one with classes defined by objects possessing attributes both necessary and sufficient to belong to a class. In IR such a classification is one the whole not very useful, in fact more often a polythetic classification is what is wanted. In such a classification each individual in a class will possess only a proportion of all the attributes possessed by all the members of that class. Hence no attribute is necessary nor sufficient for membership to a class. The query language for DR will generally be of the artificial kind, one with restricted syntax and vocabulary; in IR we prefer to use natural language although there are some notable exceptions. In DR the query is generally a complete specification of what is wanted; in IR it is invariably incomplete. This last difference arises partly from the fact that in IR we are searching for relevant documents as opposed to exactly matching items. The extent of the match in IR is assumed to indicate the likelihood of the relevance of that item. One simple consequence of this difference is that DR is more sensitive to error in the sense that, an error in matching will not retrieve the wanted item which implies a total failure of the system. In IR small errors in matching generally do not affect performance of the system significantly.
Many automatic information retrieval systems are experimental. I only make occasional reference to operational systems. Experimental IR is mainly carried on in a 'laboratory' situation whereas operational systems are commercial systems which charge for the service they provide. Naturally the two systems are evaluated differently. The 'real world' IR systems are evaluated in terms of 'user satisfaction' and the price the user is willing to pay for its service. Experimental IR systems are evaluated by comparing the retrieval experiments with standards specially constructed for the purpose. I believe that a book on experimental information retrieval, covering the design and evaluation of retrieval systems from a point of view which is independent of any particular system, will be a great help to other workers in the field and indeed is long overdue. Many of the techniques I shall discuss will not have proved themselves incontrovertibly superior to all other techniques, but they have promise and their promise will only be realised when they are understood. Information about new techniques has been so scattered through the literature that to find out about them you need to be an expert before you begin to look. I hope that I will be able to take the reader to the point where he will have little trouble in implementing some of the new techniques. Also, that some people will then go on to experiment with them, and generate new, convincing evidence of their efficiency and effectiveness. My aim throughout has been to give a complete coverage of the more important ideas current in various special areas of information retrieval. Inevitably some ideas have been elaborated at the expense of others. In particular, emphasis is placed on the use of automatic classification techniques and rigorous methods of measurement.
of effectiveness. On the other hand, automatic content analysis is given only a superficial coverage. The reasons are straightforward, firstly the material reflects my own bias, and secondly, no adequate coverage of the first two topics has been given before whereas automatic content analysis has been documented very well elsewhere. A subsidiary reason for emphasising automatic classification is that little appears to be known or understood about it in the context of IR so that research workers are loath to experiment with it.

**An information retrieval system**

Let me illustrate by means of a black box what a typical IR system would look like. The diagram shows three components: input, processor and output. Such a trichotomy may seem a little trite, but the components constitute a convenient set of pegs upon which to hang a discussion. Starting with the input side of things. The main problem here is to obtain a representation of each document and query suitable for a computer to use. Let me emphasise that most computer-based retrieval systems store only a representation of the document (or query) which means that the text of a document is lost once it has been processed for the purpose of generating its representation. A *document representative* could, for example, be a list of extracted words considered to be significant. Rather than have the computer process the natural language, an alternative approach is to have an artificial language within which all queries and documents can be formulated. There is some evidence to show that this can be effective. Of course it presupposes that a user is willing to be taught to express his information need in the language. When the retrieval
system is on-line, it is possible for the user to change his request during one search session in the light of sample retrieval, thereby; it is hoped, improving the subsequent retrieval run. Such a procedure is commonly referred to as feedback. An example of a sophisticated on-line retrieval system is the MEDLINE system and, I think it is fair to say that it will be only a short time before all retrieval systems will be on-line. Secondly, the processor, that part of the retrieval system concerned with the retrieval process. The process may involve structuring the information in some appropriate way, such as classifying it. It will also involve performing the actual retrieval function that is, executing the search strategy in response to a query. In the diagram, the documents have been placed in a separate box to emphasise the fact that they are not just input but can be used during the retrieval process in such a way that their structure is more correctly seen as part of the retrieval process. Finally, we come to the output, which is usually a set of citations or document numbers. In an operational system the story ends here. However, in an experimental system it leaves the evaluation to be done.

6.2 Effectiveness and efficiency

Much of the research and development in information retrieval is aimed at improving the effectiveness and efficiency of retrieval. Efficiency is usually measured in terms of the computer resources used such as core, backing store, and C.P.U. time. It is difficult to measure efficiency in a machine independent way. In any case, it should be measured in conjunction with effectiveness to obtain some idea of the benefit in terms of unit cost. In the previous section I mentioned that effectiveness
is commonly measured in terms of precision and recall. I repeat here that 
\textit{precision} is the ratio of the number of relevant documents retrieved to the 
total number of documents retrieved, and \textit{recall} is the ratio of the number 
of relevant documents retrieved to the total number of relevant documents 
(both retrieved and not retrieved). The reason for emphasising these two 
measures is that frequent reference is made to retrieval effectiveness but 
its detailed discussion is delayed until next Chapter. It will suffice until 
we reach that chapter to think of retrieval effectiveness in terms of 
precision and recall. It would have been possible to give the chapter on 
evaluation before any of the other material but this, in my view, would 
have been like putting the cart before the horse. Before we can appreciate 
the evaluation of observations we need to understand what gave rise to 
the observations. Hence I have delayed discussing evaluation until some 
understanding of what makes an information retrieval system tick has 
been gained.

\section*{6.3 Automatic Text Analysis}

Before a computerized information retrieval system can actually operate 
to retrieve some information, that information must have already been 
stored inside the computer. Originally it will usually have been in the 
form old documents. The computer, however, is not likely to have stored 
the complete text of each document in the natural language in which it 
was written. It will have, instead, a document representative which may 
have been produced from the documents either manually or 
automatically. The starting point of the text analysis process may be the 
complete document text, an abstract, the title only, or perhaps a list of
words only. From it the process must produce a document representative in a form which the computer can handle. The developments and advances in the process of representation have been reviewed every year by the appropriate chapters of Cuadra's *Annual Review of Information Science and Technology*. The emphasis in this Chapter is on the statistical (a word used loosely here: it usually simply implies counting) rather than linguistic approaches to automatic text analysis. The reasons for this emphasis are varied. **Firstly**, there is the limit on space. Were I to attempt a discussion of semantic and syntactic methods applicable to automatic text analysis, it would probably fill another book. Luckily such a book has recently been written by Sparck Jones and Kay. Also Montgomery has written a paper surveying linguistics in information science. **Secondly**, linguistic analysis has proved to be expensive to implement and it is not clear how to use it to enhance information retrieval. Part of the problem has been that very little progress has been made in formal semantic theory. However, there is some reason for optimism on this front, see, for example, Keenan. Undoubtedly a theory of language will be of extreme importance to the development of intelligent IR systems. But, to date, no such theory has been sufficiently developed for it to be applied successfully to IR. In any case satisfactory, possibly even very good, document retrieval systems can be built without such a theory. **Thirdly**, the statistical approach has been examined and tried ever since the days of Luhn and has been found to be moderately successful. The chapter therefore starts with the original ideas of Luhn on which much of automatic text analysis has been built, and then goes on to describe a concrete way of generating document representatives. Furthermore, ways
of exploiting and improving document representatives through weighting or classifying keywords are discussed. In passing, some of the evidence for automatic indexing is presented.

6.4 Indexing

An index language is the language used to describe documents and requests. The elements of the index language are index terms, which may be derived from the text of the document to be described, or may be arrived at independently. Index languages may be described as pre-coordinate or post-coordinate, the first indicates that terms are coordinated at the time of indexing and the latter at the time of searching. More specifically, in pre-coordinate indexing a logical combination of any index terms may be used as a label to identify a class of documents, whereas in post-coordinate indexing the same class would be identified at search time by combining the classes of documents labelled with the individual index terms. One last distinction, the vocabulary of an index language may be controlled or uncontrolled. The former refers to a list of approved index terms that an indexer may use, such as for example used by MEDLARS. The controls on the language may also include hierarchic relationships between the index terms. Or, one may insist that certain terms can only be used as adjectives (or qualifiers). There is really no limit to the kind of syntactic controls one may put on a language. The index language which comes out of the conflation algorithm in the previous section may be described as uncontrolled, post-coordinate and derived. The vocabulary of index terms at any stage in the evolution of the document collection is just the set of all conflation class names.
There is much controversy about the kind of index language which is best for document retrieval. The recommendations range from the complicated relational languages of Farradane et al and the Syntol group (see Coates for a description) to the simple index terms extracted by text processing systems just described. The main debate is really about whether automatic indexing is as good as or better than manual indexing. Each can be done to various levels of complexity. However, there seems to be mounting evidence that in both cases, manual and automatic indexing, adding complexity in the form of controls more elaborate than index term weighting do not pay dividends. This has been demonstrated by the results obtained by Cleverdon et al., Aitchison et al., Comparative Systems Laboratory and more recently Keen and Digger. The message is that uncontrolled vocabularies based on natural language achieve retrieval effectiveness comparable to vocabularies with elaborate controls. This is extremely encouraging, since the simple index language is the easiest to automate. Probably the most substantial evidence for automatic indexing has come out of the SMART Project (1966). Salton recently summarized its conclusions: '... on the average the simplest indexing procedures which identify a given document or query by a set of terms, weighted or unweighted, obtained from document or query text are also the most effective'. Its recommendations are clear, automatic text analysis should use weighted terms derived from document excerpts whose length is at least that of a document abstract. The document representatives used by the SMART project are more sophisticated than just the lists of stems extracted by conflation. There is no doubt that stems rather than ordinary word forms are more effective (Carroll and Debruyn\[1\]). On top of this the
SMART project adds index term weighting, where an index term may be a stem or some concept class arrived at through the use of various dictionaries. For details of the way in which SMART elaborates its document representatives see Salton. In the next sections I shall give a simple discussion of the kind of frequency information that may be used to weight document descriptors and explain the use of automatically constructed term classes to aid retrieval.

**Index term weighting**

Traditionally the two most important factors governing the effectiveness of an index language have been thought to be the *exhaustivity* of indexing and the *specificity* of the index language. There has been much debate about the exact meaning of these two terms. Not wishing to enter into this controversy I shall follow Keen and Digger in giving a working definition of each. For any document, *indexing exhaustivity* is defined as the number of different topics indexed, and the *index language specificity* is the ability of the index language to describe topics precisely. Keen and Digger further define *indexing specificity* as the level of precision with which a document is *actually* indexed. It is very difficult to quantify these factors. Human indexers are able to rank their indexing approximately in order of increasing exhaustivity or specificity. However, the same is not easily done for *automatic* indexing. It is of some importance to be able to quantify the notions of indexing exhaustivity and specificity because of the predictable effect they have on retrieval effectiveness. It has been recognized (Lancaster) that a high level of exhaustivity of indexing leads to high recall and low precision. Conversely, a low level of exhaustivity
leads to low recall and high precision. The converse is true for levels of indexing specificity, high specificity leads to high precision and low recall, etc. It would seem, therefore, that there is an optimum level of indexing exhaustivity and specificity for a given user population. Quite a few people (Sparck Jones, Salton and Yang), have attempted to relate these two factors to document collection statistics. For example, exhaustivity can be assumed to be related to the number of index terms assigned to a given document, and specificity related to the number of documents to which a given term is assigned in a given collection. The importance of this rather vague relationship is that the two factors are related to the distribution of index terms in the collection. The relationships postulated are consistent with the observed trade-off between precision and recall just mentioned. Changes in the number of index terms per document lead to corresponding changes in the number of documents per term and vice versa. I am arguing that in using distributional information about index terms to provide, say, index term weighting we are really attacking the old problem of controlling exhaustivity and specificity. If we go back to Luhn's original ideas, we remember that he postulated a varying discrimination power for index terms as a function of the rank order of their frequency of occurrence, the highest discrimination power being associated with the middle frequencies. His model was proposed for the selection of significant terms from a document. However, the same frequency counts can be used to provide a weighting scheme for the individual terms in a document. In fact, there is a common weighting scheme in use which gives each index term a weight directly proportional to its frequency of occurrence in the
document. At first this scheme would appear to be inconsistent with Luhn's hypothesis that the discrimination power drops off at higher frequencies. However, referring back to Figure 2.1, the scheme would be consistent if the upper cut-off is moved to the point where the peak occurs. It is likely that this is in fact what has happened in experiments using this particular form of weighting. Attempts have been made to apply weighting based on the way the index terms are distributed in the entire collection. The index term vocabulary of a document collection often has a Zipfian distribution, that is, if we count the number of documents in which each index term occurs and plot them according to rank order, then we obtain the usual hyperbolic shape. Sparck Jones showed experimentally that if there are $N$ documents and an index term occurs in $n$ of them then a weight of $\log(N/n) + 1$ leads to more effective retrieval than if the term were used unweighted. If indexing specificity is assumed to be inversely proportional to the number of documents in which an index term occurs then the weighting can be seen to be attaching more importance to the more specific terms. The difference between the last mode of weighting and the previous one may be summarized by saying that document frequency weighting places emphasis on content description whereas weighting by specificity attempts to emphasise the ability of terms to discriminate one document from another. Salton and Yang have recently attempted to combine both methods of weighting by looking at both inter document frequencies and intra document frequencies. Their conclusions are really an extension of those reached by Luhn. By considering both the total frequency of occurrence of a term and its distribution over the documents, that is, how
many times it occurs in each document, they were able to draw several conclusions. A term with high total frequency of occurrence is not very useful in retrieval irrespective of its distribution. Middle frequency terms are most useful particularly if the distribution is skewed. Rare terms with a skewed distribution are likely to be useful but less so than the middle frequency ones. Very rare terms are also quite useful but come bottom of the list except for the ones with a high total frequency. The experimental evidence for these conclusions is insufficient to make a more precise statement of their merits. Salton and his co-workers have developed an interesting tool for describing whether an index is 'good' or 'bad'. They assume that a good index term is one which, when assigned as an index term to a collection of documents, renders the documents as dissimilar as possible, whereas a bad term is one which renders the documents more similar. This is quantified through a term discrimination value which for a particular term measures the increase or decrease in the average dissimilarity between documents on the removal of that term. Therefore, a good term is one which on removal from the collection of documents, leads to a decrease in the average dissimilarity (adding it would hence lead to an increase), whereas a bad term is one which leads on removal to an increase. The idea is that a greater separation between documents will enhance retrieval effectiveness but that less separation will depress retrieval effectiveness. Although superficially this appears reasonable, what really is required is that the relevant documents become less separated in relation to the non-relevant ones. Experiments using the term discrimination model have been reported. A connection between term discrimination and inter document frequency has also been made
supporting the earlier results reported by Salton, Wong and Yang. The main results have been conveniently summarised by Yu and Salton, where also some formal proofs of retrieval effectiveness improvement are given for strategies based on frequency data. For example, the inverse document frequency weighting scheme described above, that is assigning a weight proportional to \( \log (N/n) + 1 \), is shown to be formally more effective than not using these weights. Of course, to achieve a proof of this kind some specific assumptions about how to measure effectiveness and how to match documents with queries have to be made. They also establish the effectiveness of a technique used to conflate low frequency terms, which increases recall, and of a technique used to combine high frequency terms into phrases, which increases precision.

6.5 Automatic keyword classification

Many automatic retrieval systems rely on thesauri to modify queries and document representatives to improve the chance of retrieving relevant documents. Salton has experimented with many different kinds of thesauri and concluded that many of the simple ones justify themselves in terms of improved retrieval effectiveness. In practice many of thesauri are constructed manually. They have mainly been constructed in two ways:

(1) Words which are deemed to be about the same topic are linked;

(2) Words which are deemed to be about related things are linked.

The first kind of thesaurus connects words which are intersubstitutable, that is, it puts them into equivalence classes. Then one word could be
chosen to represent each class and a list of these words could be used to form a controlled vocabulary. From this an indexer could be instructed to select the words to index a document, or the user could be instructed to select the words to express his query. The same thesaurus could be used in an automatic way to identify the words of a query for the purpose of retrieval.

The second kind of thesaurus uses semantic links between words to, for example, relate them hierarchically. The manually constructed thesaurus used by the MEDLARS system is of this type. However, methods have been proposed to construct thesauri automatically. Whereas, the manual thesauri are semantically based (e.g. they recognise synonyms, more general or more specific relationships) the automatic thesauri tend to be syntactically and statistically based. Again the use of syntax has proved to be of little value, so I shall concentrate on the statistical methods. These are based mainly on the patterns of co-occurrence of words in documents. These 'words' are often the descriptive items which were introduced earlier as terms of keywords. The basic relationship underlying the automatic construction of keyword classes is as follows: If keyword $a$ and $b$ are substitutable for one another in the sense that we are prepared to accept a document containing one in response to a request containing the other, this will be because they have the same meaning or refer to a common subject or topic. One way of finding out whether two keywords are related is by looking at the documents in which they occur. If they tend to co-occur in the same documents, the chances are that they have to do with the same subject and so can be substituted for one another. It is
not difficult to see that, based on this principle, a classification of keywords can be automatically constructed, of which the classes are used analogously to those of the manual thesaurus mentioned before. More specifically we can identify two main approaches to the use of keyword classifications:

(1) Replace each keyword in a document (and query) representative by the name of the class in which it occurs;

(2) Replace each keyword by all the keywords occurring in the class to which it belongs.

If we think of a simple retrieval strategy as operating by matching on the descriptors, whether they be keyword names or class names, then 'expanding' representatives in either of these ways will have the effect of increasing the number of matches between document and query, and hence tends to improve recall. The second way will improve precision as well. Sparck Jones [2] has reported a large number of experiments using automatic keyword classifications and found that in general one obtained a better retrieval performance with the aid of automatic keyword classification than with the unclassified keywords alone. Unfortunately, even here the evidence has not been conclusive. The work by Minker et al. has not confirmed the findings of Sparck Jones, and in fact they have shown that in some cases keyword classification can be detrimental to retrieval effectiveness. Salton, in a review of the work of Minker et al., has questioned their experimental design which leaves the question of the
effectiveness of keyword classification still to be resolved by further research.

6.6 A language for describing file structures

Like all subjects in computer science the terminology of file structures has evolved higgledy-piggledy without much concern for consistency, ambiguity, or whether it was possible to make the kind of distinctions that were important. It was only much later that the need for a well-defined, unambiguous language to describe file structures became apparent. In particular, there arose a need to communicate ideas about file structures without getting bogged down by hardware considerations. This section will present a formal description of file structures. The framework described is important for the understanding of any file structure. The terminology is based on that introduced by Hsiao and Harary (but also see Hsiao and Manola and Hsiao). Their terminology has been modified and extended by Severance, a summary of this can be found in van Rijsbergen. Jonkers has formalised a different framework which provides an interesting contrast to the one described here.

6.7 Sequential files

A sequential file is the most primitive of all file structures. It has no directory and no linking pointers. The records are generally organized in lexicographic order on the value of some key. In other words, a particular attribute is chosen whose value will determine the order of the records. Sometimes when the attribute value is constant for a large number of
records a second key is chosen to give an order when the first key fails to discriminate. The implementation of this file structure requires the use of a sorting routine. Its main advantages are:

(1) It is easy to implement;

(2) It provides fast access to the next record using lexicographic order.

Its disadvantages:

(1) It is difficult to update - inserting a new record may require moving a large proportion of the file;

(2) Random access is extremely slow.

Sometimes a file is considered to be sequentially organised despite the fact that it is not ordered according to any key. Perhaps the date of acquisition is considered to be the key value; the newest entries are added to the end of the file and therefore pose no difficulty to updating.

**Inverted files**

The importance of this file structure will become more apparent when Boolean Searches are discussed in the next chapter. For the moment we limit ourselves to describing its structure. An *inverted file* is a file structure in which every list contains only one record. Remember that a list is defined with respect to a keyword $K$, so every $K$-list contains only one record. This implies that the directory will be such that $ni = hi$ for all $i$, that is, the number of records containing $Ki$ will equal the number of
$Ki$-lists. So the directory will have an address for each record containing $Ki$. For document retrieval this means that given a keyword we can immediately locate the addresses of all the documents containing that keyword. For the previous example let us assume that a non-black entry in the field corresponding to an attribute indicates the presence of a keyword and a black entry its absence. Then the directory will point to the file in the way shown in Figure 4.3. The definition of an inverted file does not require that the addresses in the directory are in any order. However, to facilitate operations such as conjunction ('and') and disjunction ('or') on any two inverted lists, the addresses are normally kept in record number order. This means that 'and' and 'or' operations can be performed with one pass through both lists. The penalty we pay is of course that the inverted file becomes slower to update.

### 6.8 Index-sequential files

An index-sequential file is an inverted file in which for every keyword $K_i$, we have $n_i = h_i = 1$ and $a_{11} < a_{21} \ldots < a_{m1}$. This situation can only arise if each record has just one unique keyword, or one unique attribute-value. In practice therefore, this set of records may be order sequentially by a key. Each key value appears in the directory with the associated address of its record. An obvious interpretation of a key of this kind would be the record number. In our example none of the attributes would do the job except the record number. Diagrammatically the index-sequential file would therefore appear as shown in Figure 4.4. I have deliberately written $R_i$ instead of $K_i$ to emphasise the nature of the key. In the literature an index-sequential file is usually thought of as a sequential
file with a hierarchy of indices. This does not contradict the previous
definition; it merely describes the way in which the directory is
implemented. It is not surprising therefore that the indexes ('index' =
'directory' here) are often oriented to the characteristics of the storage
medium. For example (see Figure 4.5) there might be three levels of
indexing: track, cylinder and master. Each entry in the track index will
contain enough information to locate the start of the track, and the key of
the last record in the track which is also normally the highest value on
that track. There is a track index for each cylinder. Each entry in the
cylinder index gives the last record on each cylinder and the address of
the track index for that cylinder. If the cylinder index itself is stored on
tracks, then the master index will give the highest key referenced for each
track of the cylinder index and the starting address of that track. No
mention has been made of the possibility of overflow during an updating
process. Normally provision is made in the directory to administer an
overflow area. This of course increases the number of book-keeping
entries in each entry of the index. Were we to group documents according
to the keywords they shared, then for each keyword we would have a
group of documents, namely, those which had that keyword in common.
Ci would then be the field containing the keyword uniting that particular
group. The rings would of course overlap (Figure 4.9), as in this
example:

\[ D1 = \{K1, K2\} \]

\[ D2 = \{K2, K3\} \]
\[ D3 = \{K1, K4\} \]

The usefulness of this kind of structure will become more apparent when we discuss searching of classifications. If each ring has associated with it a record which contains identifying information for its members, then, a search strategy searching a structure such as this will first look at \( C_i \) (or \( K_i \) in the second example) to determine whether to proceed or abandon the search.

### 6.9 Scatter storage or hash addressing

One file structure which does not relate very well to the ones mentioned before is known as Scatter Storage. The technique by which the file structure is implemented is often called Hash Addressing. Its underlying principle is appealingly simple. Given that we may access the data through a number of keys \( K_i \), then the address of the data in store is located through a key transformation function \( f \) which when applied to \( K_i \) evaluates to give the address of the associated data. We are assuming here that with each key is associated only one data item. Also for convenience we will assume that each record (data and key) fits into one location, whose address is in the image space of \( f \). The addresses given by the application of \( f \) to the keys \( K_i \) are called the hash addresses and \( f \) is called a hashing function. Ideally \( f \) should be such that it spreads the hash addresses uniformly over the available storage. Of course this would be achieved if the function were one-to-one. Unfortunately this cannot be so because the range of possible key values is usually considerably larger than the range of the available storage addresses. Therefore, given any
hashing function we have to contend with the fact that two distinct keys 
\( K_i \) and \( K_j \) are likely to map to the same address \( f(K_i) = f(K_j) \). Before I 
explain some of the ways of dealing with this I shall give a few examples 
of hashing functions. Let us assume that the available storage is of size 
\( 2^m \) then three simple transformations are as follows:

1) if \( K_i \) is the key, then take the square of its binary representation and 
select \( m \) bits from the middle of the result;

2) cut the binary representation of \( K_i \) into pieces each of \( m \) bits and add 
these together. Now select the \( m \) least significant bits of the sum as the 
hash address;

3) divide the integer corresponding to \( K_i \) by the length of the available 
store \( 2^m \) and use the remainder as the hash address.

Each of these methods has disadvantages. For example, the last one may 
given the same address rather frequently if there are patterns in the keys. 
Before using a particular method, the reader is advised to consult the now 
extensive literature on the subject, e.g. Morris, or Lum et al.. As 
mentioned before there is the problem of collisions, that is, when two 
distinct keys hash to the same address. The first point to be made about 
this problem is that it destroys some of the simplicity of hashing. Initially 
it may have been thought that the key need not be stored with the data at 
the hash address. Unfortunately this is not so. No matter what method we 
use to resolve collisions we still need to store the key with the data so that 
at search time when a key is hashed we can distinguish its data from the
data associated with keys which have hashed to the same address. There are a number of strategies for dealing with collisions. Essentially they fall into two classes, those which use pointers to link together collided keys and those which do not. Let us first look at the ones which do not use pointers. These have a mechanism for searching the store, starting at the address where the collision occurred, for an empty storage location if a record needs to be inserted, or, for a matching key value at retrieval time. The simplest of these advances from the hash address each time moving along a fixed number of locations, say $s$, until an empty location or the matching key value is found. The collision strategy thus traces out a well defined sequence of locations. This method of dealing with collisions is called the **linear** method. The tendency with this method is to store collided records as closely to the initial hash address as possible. This leads to an undesirable effect called **primary clustering**. In this context all this means is that the records tend to concentrate in groups or bunch-up. It destroys the uniform nature of the hashing function. To be more precise, it is desirable that hash addresses are equally likely, however, the first empty location at the end of a collision sequence increases in likelihood in proportion to the number of records in the collision sequence. To see this one needs only to realise that a key hashed to any location in the sequence will have its record stored at the end of the sequence. Therefore big groups of records tend to grow even bigger. This phenomenon is aggravated by a small step size $s$ when seeking an empty location. Sometimes $s = 1$ is used in which case the collision strategy is known as the **open addressing technique**. Primary clustering is also worse when the hash table (available storage) is relatively full. Variations in the linear
method which avoid primary clustering involve making the step size a variable. One way is to set $s$ equal to $ai + bi[2]$ on the $i$th step. Another is to invoke a random number generator which calculates the step size afresh each time. These last two collision handling methods are called the quadratic and random method respectively. Although they avoid primary clustering they are nevertheless subject to secondary clustering, which is caused by keys hashing to the same address and following the same sequence in search of an empty location. Even this can be avoided, see for example Bell and Kaman. The second class of collision handling methods involves extra storage space which is used to chain together collided records. When a collision occurs at a hash address it may be because it is the head of a chain of records which have all hashed to that address, or it may be that a record is stored there which belongs to a chain starting at some other address. In both cases a free location is needed which in the first case is simply linked in and stores the new record, in the second case the intermediate chain element is moved to the free location and the new record is stored at its own hash address thus starting a new chain (a one-element chain so far). A variation on this method is to use a two-level store. At the first level we have a hash table, at the second level we have a bump table which contains all the collided records. At a hash address in the hash table we will find either, a record if no collisions have taken place at that address, or, a pointer to a chain of records which collided at that address. This latter chaining method has the advantage that records need never be moved once they have been entered in the bump table. The storage overhead is larger since records are put in the bump table before the hash table is full. For both classes of collision
strategies one needs to be careful about deletions. For the linear, quadratic etc. collision handling strategies we must ensure that when we delete a record at an address we do not make records which collided at that address unreachable. Similarly with the chaining method we must ensure that a deleted record does not leave a gap in the chain, that is, after deletion the chain must be reconnected. The advantages of hashing are several. Firstly it is simple. Secondly its insertion and search strategies are identical. Insertion is merely a failed search. If $K_i$ is the hashed key, then if a search of the collision sequence fails to turn up a match in $K_i$, its record is simply inserted at the end of the sequence at the next free location. Thirdly, the search time is independent of the number of keys to be inserted. The application of hashing in IR has tended to be in the area of table construction and look-up procedures. An obvious application is when constructing the set of conflation classes during text processing. In Chapter 2, I gave an example of a document representative as simply a list of class names, each name standing for a set of equivalent words. During a retrieval operation, a query will first be converted into a list of class names. To do this each significant word needs to be looked up in a dictionary which gives the name of the class to which it belongs. Clearly there is a case for hashing. We simply apply the hashing function to the word and find the name of the conflation class to which it belongs at the hash address. A similar example is given in great detail by Murray. Finally, let me recommend two very readable discussions on hashing, one is in Page and Wilson, the other is in Knuth's third volume.
6.10 Clustered files

It is now common practice to refer to a file processed by a clustering algorithm as a clustered file, and to refer to the resulting structure as a file structure. For example Salton[34] (p. 288) lists a clustered file as an alternative organisation to inverted, serial, chained files, etc. Although it may be convenient terminologically, it does disguise the real status of cluster methods. Cluster methods (or automatic classification methods) are more profitably discussed at the level of abstraction at which relations are discussed in connection with data bases, that is, in a thoroughly data independent way. In other words, selecting an appropriate cluster method and implementing it are two separate problems. Unfortunately not all users of clustering techniques see it this way, and so the current scene is rather confused. One factor contributing to the confusion is that clustering techniques have been used at a very low level of implementation of system software, for example, to reduce the number of page exceptions in a virtual memory. Therefore, those who use clustering merely to increase retrieval efficiency (in terms of storage and speed) will tend to see a classification structure as a file structure, whereas those who see clustering as a means of discovering (or summarizing) some inherent structure in the data will look upon the same structure as a description of the data. Of course, this description may be used to achieve more efficient retrieval (and in IR more effective retrieval in terms of say precision and recall). Furthermore, if one looks carefully at some of the implementations of cluster methods one discovers that the classificatory
system is represented inside the computer by one of the more conventional file structures.
Chapter Seven

Evaluation Of Information

Retrieval Systems
Chapter Seven

Evaluation Of Information Retrieval Systems

7.1 Introduction

Much effort and research has gone into solving the problem of evaluation of information retrieval systems. However, it is probably fair to say that most people active in the field of information storage and retrieval still feel that the problem is far from solved. One may get an idea of the extent of the effort by looking at the numerous survey articles that have been published on the topic. In a research of this nature it will be impossible to cover all work to date about evaluation. Instead I shall attempt to explicate the conventional, most commonly used method of evaluation, followed by a survey of the more promising attempts to improve on the older methods of evaluation. To put the problem of evaluation in perspective let me pose three questions: (1) Why evaluate? (2) What to evaluate? (3) How to evaluate? The answers to these questions pretty well cover the whole field of evaluation. There is much controversy about each and although I do not wish to add to the controversy I shall attempt an answer to each one in turn. The answer to the first question is mainly a social and economic one. The social part is fairly intangible, but mainly relates to the desire to put a measure on the benefits (or disadvantages) to be got from information retrieval systems. I use 'benefit' here in a much wider sense than just the benefit accruing due to acquisition of relevant
documents. For example, what benefit will users obtain (or what harm will be done) by replacing the traditional sources of information by a fully automatic and interactive retrieval system? Studies to gauge this are going on but results are hard to interpret. For some kinds of retrieval systems the benefit may be more easily measured than for others (compare statute or case law retrieval with document retrieval). The economic answer amounts to a statement of how much it is going to cost you to use one of these systems, and coupled with this is the question 'is it worth it?'. Even a simple statement of cost is difficult to make. The computer costs may be easy to estimate, but the costs in terms of personal effort are much harder to ascertain. Then whether it is worth it or not depends on the individual user. It should be apparent now that in evaluating an information retrieval system we are mainly concerned with providing data so that users can make a decision as to (1) whether they want such a system (social question) and (2) whether it will be worth it. Furthermore, these methods of evaluation are used in a comparative way to measure whether certain changes will lead to an improvement in performance. In other words, when a claim is made for say a particular search strategy, the yardstick of evaluation can be applied to determine whether the claim is a valid one. The second question (what to evaluate?) boils down to what can we measure that will reflect the ability of the system to satisfy the user. Since this book is mainly concerned with automatic document retrieval systems I shall answer it in this context. In fact, as early as 1966, Cleverdon gave an answer to this. He listed six main measurable quantities:
(1) The coverage of the collection, that is, the extent to which the system includes relevant matter;

(2) the time lag, that is, the average interval between the time the search request is made and the time an answer is given;

(3) the form of presentation of the output;

(4) the effort involved on the part of the user in obtaining answers to his search requests;

(5) the recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request;

(6) the precision of the system, that is, the proportion of retrieved material that is actually relevant.

It is claimed that (1)-(4) are readily assessed. It is recall and precision which attempt to measure what is now known as the effectiveness of the retrieval system. In other words it is a measure of the ability of the system to retrieve relevant documents while at the same time holding back non-relevant one. It is assumed that the more effective the system the more it will satisfy the user. It is also assumed that precision and recall are sufficient for the measurement of effectiveness. There has been much debate in the past as to whether precision and recall are in fact the appropriate quantities to use as measures of effectiveness. A popular alternative has been recall and fall-out (the proportion of non-relevant documents retrieved). However, all the alternatives still require the
determination of relevance in some way. The relationship between the various measures and their dependence on relevance will be made more explicit later. Later in the chapter a theory of evaluation is presented based on precision and recall. The advantages of basing it on precision and recall are that they are:

(1) the most commonly used pair;

(2) fairly well understood quantities.

The final question (How to evaluate?) has a large technical answer. In fact, most of the remainder of this chapter may be said to be concerned with this. It is interesting to note that the technique of measuring retrieval effectiveness has been largely influenced by the particular retrieval strategy adopted and the form of its output. For example, when the output is a ranking of documents an obvious parameter such as rank position is immediately available for control. Using the rank position as cut-off, a series of precision recall values could then be calculated, one part for each cut-off value. The results could then be summarized in the form of a set of points joined by a smooth curve. The path along the curve would then have the immediate interpretation of varying effectiveness with the cut-off value. Unfortunately, the kind of question this form of evaluation does not answer is, for example, how many queries did better than average and how many did worse? Nevertheless, we shall need to spend more time explaining this approach to the measurement of effectiveness since it is the most common approach and needs to be understood.
Before proceeding to the technical details relating to the measurement of effectiveness it is as well to examine more closely the concept of relevance which underlies it.

7.2 Relevance

Relevance is a *subjective* notion. Different users may differ about the relevance or non-relevance of particular documents to given questions. However, the difference is not large enough to invalidate experiments which have been made with document collections for which test questions with corresponding relevance assessments are available. These questions are usually elicited from bona fide users, that is, users in a particular discipline who have an information need. The relevance assessments are made by a panel of experts in that discipline. So we now have the situation where a number of questions exist for which the 'correct' responses are known. It is a general assumption in the field of IR that should a retrieval strategy fare well under a large number of experimental conditions then it is likely to perform well in an operational situation where relevance is *not* known in advance. There is a concept of relevance which can be said to be *objective* and which deserves mention as an interesting source of speculation. This notion of relevance has been explicated by Cooper. It is properly termed 'logical relevance'. Its usefulness in present day retrieval systems is limited. However, it can be shown to be of some importance when it is related to the development of question-answering systems, such as the one recently designed by T. Winograd at Massachusetts Institute of Technology. Logical relevance is most easily explicated if the questions are restricted to the yes-no type.
This restriction may be lifted - for details see Cooper's original paper. Relevance is defined in terms of logical consequence. To make this possible a question is represented by a set of sentences. In the case of a yes-no question it is represented by two formal statements of the form 'p' and 'not-p'. For example, if the query were 'Is hydrogen a halogen element?', the part of statements would be the formal language equivalent of 'Hydrogen is a halogen element' and 'Hydrogen is not a halogen element'. More complicated questions of the 'which' and 'whether' type can be transformed in this manner, for details the reader is referred to Belnap. If the two statements representing the question are termed component statements then the subset of the set of stored sentences is a premiss set for a component statement if an only if the component statement is a logical consequence of that subset. (Note we are now temporarily talking about stored sentences rather than stored documents.)

A minimal premiss set for a component statement is one that is as small as possible in the sense that if any of its members were deleted, the component statement would no longer be a logical consequence of the resulting set. Logical relevance is now defined as a two-place relation between stored sentences and information need representations (that is, the question represented as component statements). The final definition is as follows:

A stored sentence is logically relevant to (a representation of) information need if and only if it is a member of some minimal premiss set of stored sentences for some component statement of that need. Although logical relevance is initially only defined between sentences it can easily be
extended to apply to stored documents. A document is relevant to an
information need if and only if it contains at least one sentence which is
relevant to that need. Earlier on I stated that this notion of relevance was
only of limited use at the moment. The main reason for this is that the
kind of system which would be required to implement a retrieval strategy
which would retrieve only the logically relevant documents has not been
built yet. However, the components of such a system do exist to a certain
extent. Firstly, theorem provers, which can prove theorems within formal
languages such as the first-order predicate calculus, have reached quite a
level of sophistication now (see, for example, Chang and Lee). Secondly,
Winograd's system is capable of answering questions about its simple
universe blocks in natural language. In principle this system could be
extended to construct a universe of documents, that is, the content of a
document is analyzed and incorporated into the universe of currently
'understood' documents. It may be that the scale of a system of this kind
will be too large for present day computers; only the future will tell.
Saracevic has given a thorough review of the notion of relevance in
information science. Robertson has summarized some of the more recent
work on probabilistic interpretations of relevance.

7.3 Composite measures

Dissatisfaction in the past with methods of measuring effectiveness by a
pair of numbers (e.g. precision and recall) which may co-vary in a loosely
specified way has led to attempts to invest composite measures. These are
still based on the 'contingency' table but combine parts of it into a single
number measure. Unfortunately many of these measures are rather ad hoc
and cannot be justified in any rational way. The simplest example of this kind of measure is the sum of precision and recall

\[ S = P + R \]

This is simply related to a measure suggested by Borko

\[ BK = P + R - 1 \]

More complicated ones are

\[ Q = \frac{P^* - F}{P^* + F - 2RF} \quad (F = \text{Fallo}ut) \]

\[ F = 1 - \frac{1}{2\left(\frac{1}{P}ight) + 2\left(\frac{1}{R}\right) - 3} \]

Vickery's measure \( V \) can be shown to be a special case of a general measure which will be derived below. Some single-number measures have derivations which can be justified in a rational manner. Some of them will be given individual attention later on. Suffice it here to point out that it is the model underlying the derivation of these measures that is important.

7.4 The SMART measures

In 1966, Rocchio gave a derivation of two overall indices of merit based on recall and precision. They were proposed for the evaluation of retrieval systems which ranked documents, and were designed to be independent of cut-off. The first of these indices is normalised recall. It
roughly measures the effectiveness of the ranking in relation to the best possible and worst possible ranking. The situation is illustrated in Figure 7.9 for 25 documents where we plot on the y-axis and the ranks on the x-axis. Normalized recall (Rnorm) is the area between the actual case and the worst as a proportion of the area between the best and the worst. If \( n \) is the number of relevant documents, and \( ri \) the rank at which the \( i \)th document is retrieved, then the area between the best and actual case can be shown to be (after a bit of algebra):

\[
A_b - A_a = \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} r_i}{N}
\]

(see Salton, page 285).

A convenient explicit form of normalised recall is:

\[
\hat{R}_{\text{norm}} = 1 - \frac{\sum_{i=1}^{N-n} r_i - \sum_{i=1}^{N-n} r_i}{N(N-n)}
\]

where \( N \) is the number of documents in the system and \( N - n \) the area between the best and the worst case (to see this substitute \( ri = N - i + 1 \) in the formula for \( Ab - Aa \)). The form ensures that \( Rnorm \) lies between 0 (for the worst case) and 1 (for the best case). In an analogous manner normalised precision is worked out. In Figure 7.10 we once more have three curves showing (1) the best case, (2) the actual case, and (3) the worst case in terms of the precision values at different rank positions.
The calculation of the areas is a bit more messy but simple to do (see Salton, page 298). The area between the actual and best case is now given by:

\[ A_a - A_b = \sum_{i=1}^{n} \log f_i - \sum_{i=1}^{n} \log f_i \]

The log function appears as a result of approximating \( \frac{1}{r} \) by its continuous analogue \( \int \frac{1}{r} \, dr \), which is \( \log r + \text{constant} \). The area between the worst and best case is obtained in the same way as before using the same substitution, and is:

\[ \log \left( \frac{M!}{(N-M)! \, M!} \right) \]

The explicit form, with appropriate normalization, for normalized precision is therefore:

\[ F_{\text{norm}} = 1 - \frac{\sum \log f_i - \sum \log f_i}{\log \left( \frac{M!}{(N-M)! \, M!} \right)} \]

Once again it varies between 0 (worst) and 1 (best).

A few comments about these measures are now in order. **Firstly** their behavior is consistent in the sense that if one of them is 0 (or 1) then the other is 0 (or 1). In other words they both agree on the best and worst performance. **Secondly**, they differ in the weights assigned to arbitrary positions of the precision-recall curve, and these weights may differ
considerably from those which the user feels are pertinent (Senko). Or, as Salton (page 289) puts it: 'the normalised precision measure assigns a much larger weight to the initial (low) document ranks than to the later ones, whereas the normalised recall measure assigns a uniform weight to all relevant documents'. Unfortunately, the weighting is arbitrary and given. Thirdly, it can be shown that normalised recall and precision have interpretations as approximations to the average recall and precision values for all possible cut-off levels. That is, if \( R(i) \) is the recall at rank position \( i \), and \( P(i) \) the corresponding precision value, then:

\[
R_{norm} \sim \frac{1}{N} \sum_{i=1}^{N} R(i)
\]

\[
P_{norm} \sim \frac{1}{N} \sum_{i=1}^{N} P(i)
\]

Fourthly, whereas Cooper has gone to some trouble to take account of the random element introduced by ties in the matching function, it is largely ignored in the derivation of \( P_{norm} \) and \( R_{norm} \). One further comment of interest is that Robertson15 has shown that normalised recall has an interpretation as the area under the Recall-Fallout curve used by Swets. Finally mention should be made of two similar but simpler measures used by the SMART system. They are:

\[
\text{Rank Recall} = \frac{\sum_{j=1}^{z} i_j}{\sum_{j=1}^{z} r_j}
\]

\[
\text{Log Precision} = \frac{\sum_{j=1}^{z} \ln i_j}{\sum_{j=1}^{z} \ln r_j}
\]
And do not take into account the collection size $N$, $n$ is here the number of relevant documents for the particular test query.

### 7.5 A normalized symmetric difference

Let us now return to basics and consider how it is that users could simply measure retrieval effectiveness. We are considering the common situation where a set of documents is retrieved in response to a query, the possible ordering of this set is ignored. Ideally the set should consist only of documents relevant to the request, that is giving 100 per cent precision and 100 per cent recall (and by implication 0 per cent fallout). In practice, however, this is rarely the case, and the retrieved set consists of both relevant and non-relevant documents. The situation may therefore be pictured as shown in Figure 7.11, where $A$ is the set of relevant documents, $B$ the set of retrieved documents, and $A \cap B$ the set of retrieved documents which are relevant. Now, an intuitive way of measuring the adequacy of the retrieved set is to measure the size of the shaded area. Or to put it differently, to measure to what extent the two sets do not match. The area is in fact the symmetric difference: $A \Delta B$ (or $A \cup B - A \cap B$). Since we are more interested in the proportion (rather than absolute number) of relevant and non-relevant documents retrieved, we need to normalize this measure. A simple normalization gives:
\[ E = \frac{|A \triangle E|}{|A| + |E|} \]

In terms of \( A \) and \( E \) we have:

\[ E = 1 - \frac{1}{\frac{1}{2} \left( \frac{1}{\bar{A}} \right) + \frac{1}{2} \left( \frac{1}{E} \right)} \]

which is a simple composite measure.

The preceding argument in itself is not sufficient to justify the use of this particular composite measure. However, I shall now introduce a framework within which a general measure may be derived which among others has \( E \) as one of its special cases.

**Foundation**

Problems of measurement have arisen in physics, psychology, and more recently, the social sciences. Clarification of these problems has been sought with the help of the theory of measurement. I shall attempt to do the same for information retrieval. My purpose is to construct a framework, based on the mathematical theory of measurement within which measures of effectiveness for retrieval systems can be derived. The basic mathematical notions underlying the measurement ideas will be introduced, but for their deeper understanding the reader is referred to the excellent book by Krantz et al. It would be fair to say that the theory developed there is applied here. Also of interest are the books by Ellis and Lieberman. The problems of measurement in information retrieval differ from those encountered in the physical sciences in one important
aspect. In the physical sciences there is usually an empirical ordering of the quantities we wish to measure. For example, we can establish empirically by means of a scale which masses are equal, and which are greater or less than others. Such a situation does not hold in information retrieval. In the case of the measurement of effectiveness by precision and recall, there is no absolute sense in which one can say that one particular pair of precision-recall values is better or worse than some other pair, or, for that matter, that they are comparable at all. However, to leave it at that is to admit defeat. There is no reason why we cannot postulate a particular ordering, or, to put it more mildly, why we can not show that a certain model for the measurement of effectiveness has acceptable properties. The immediate consequence of proceeding in this fashion is that each property ascribed to the model may be challenged. The only defence one has against this is that:

(1) all properties ascribed are consistent;

(2) they bring out into the open all the assumptions made in measuring effectiveness;

(3) each property has an acceptable interpretation;

(4) the model leads to a plausible measure of effectiveness.

It is as well to point out here that it does not lead to a unique measure, but it does show that certain classes of measures can be regarded as being equivalent.
The model

We start by examining the structure which it is reasonable to assume for the measurement of effectiveness. Put in other words, we examine the conditions that the factors determining effectiveness can be expected to satisfy. We limit the discussion here to two factors, namely precision and recall, although this is no restriction, different factors could be analyzed, and, as will be indicated later, more than two factors can simplify the analysis. If $R$ is the set of possible recall values and $P$ is the set of possible precision values then we are interested in the set $R \times P$ with a relation on it. We shall refer to this as a relational structure and denote it $<R \times P, \geq >$ where $\geq$ is the binary relation on $R \times P$. (We shall use the same symbol for less than or equal to, the context will make clear what the domain is.) All we are saying here is that for any given point $(R, P)$ we wish to be able to say whether it indicates more, less or equal effectiveness than that indicated by some other point. The kind of order relation is a weak order. To be more precise:

Definition 1. The relational structure $<R \times P, \geq >$ is a weak order if and only if for $e_1, e_2, e_3 [[\text{proper subset}]] R \times P$ the following axioms are satisfied:

(1) Connectedness: either $e_1 \geq e_2$ or $e_2 \geq e_1$

(2) Transitivity: if $e_1 \geq e_2$ and $e_2 \geq e_3$ then $e_1 \geq e_3$

We insist that if two pairs can be ordered both ways then $(R_1, P_1) \sim (R_2, P_2)$, i.e. equivalent not necessarily equal. The transitivity condition is
obviously desirable. We now turn to a second condition which is commonly called *independence*. This notion captures the idea that the two components contribute their effects independently to the effectiveness. *Definition 2.* A relation $\geq$ on $R \times P$ is independent if and only if, for $R_1, R_2$ $[[proper subset]] R$, $(R_1, P) \geq (R_2, P')$ for some $P$ $[[proper subset]] P$ implies $(R_1, P') \geq (R_2, P')$ for every $P'$ $[[proper subset]] P$; and for $P_1, P_2$ $[[proper subset]] P$, $(R, P_1) \geq (R, P_2)$ for some $R$ $[[proper subset]] R$ implies $(R', P_1) \geq (R', P_2)$ for every $R'$ $[[proper subset]] R$. All we are saying here is, given that at a constant recall (precision) we find a difference in effectiveness for two values of precision (recall) then this difference cannot be removed or reversed by changing the constant value. We now come to a condition which is not quite as obvious as the preceding ones. To make it more meaningful I shall need to use a diagram, *Figure 7.12*, which represents the ordering we have got so far with definitions 1 and 2. The lines $l_1$ and $l_2$ are lines of equal effectiveness that is any two points $(R, P), (R', P')$ $[[proper subset]]l_1$ are such that $(R, P) \sim (R', P')$ (where $\sim$ indicates *equal* effectiveness). Now let us assume that we have the points on $l_1$ and $l_2$ but wish to deduce the relative ordering in between these two lines. One may think of this as an interpolation procedure.
Figure: 6.1 A diagram illustrating the Thomen condition.

**Definition 3** (Thomsen condition). For every $R_1, R_2, R_3$ [[proper subset]] $R$ and $P_1, P_2, P_3$ [[proper subset]] $P$, $(R_1, P_3) \sim (R_3, P_2)$ and $(R_3, P_1) \sim (R_2, P_3)$ imply that $(R_1, P_1) \sim (R_2, P_2)$.

Intuitively this can be reasoned as follows. The intervals $R_1, R_3$ and $P_2, P_3$ are equivalent since an increase in the $R$-factor by $R_1, R_3$ and an increase in the $P$-factor by $P_2, P_3$ starting from $(R_1, P_3)$ lead to the same effectiveness (points on $l_2$). It therefore follows that a decrease in each factor starting from equal effectiveness, in this case the two points $(R_3, R_1)$ and $(R_2, P_3)$ on $l_1$, should lead to equal effectiveness. The fourth condition is one concerned with the continuity of each component. It makes precise what intuitively we would expect when considering the existence of intermediate values. **Definition 4** (Restricted Solvability). A relation $\geq$ on $R \times P$ satisfies restricted solvability provided that:
(1) whenever \( R, \ R, \ R \) \( [\text{proper subset}] \) \( R \) and \( P, \ P' \) \( [\text{proper subset}] \) \( P \) for which \( (\ R, \ P') \geq (R, P) \geq (R, P') \) then there exists \( R \) \( [\text{proper subset}] \) \( R \) s.t. \( (R, P') \sim (R, P) \);

(2) a similar condition holds on the second component.

In other words we are ensuring that the equation \( (R', P') \sim (R, P) \) is soluble for \( R' \) provided that there exist \( R, R \) such that \( (R, P') \geq (R, P') \) \( \geq (R, P') \). An assumption of continuity of the precision and recall factors would ensure this. The fifth condition is not limiting in any way but needs to be stated. It requires, in a precise way, that each component is essential. **Definition 5.** Component \( R \) is *essential* if and only if there exist \( R_1, \ R_2 \) \( [\text{proper subset}] \) \( R \) and \( P_1 \) \( [\text{proper subset}] \) \( P \) such that it is *not* the case that \( (R, P) \sim (R_1, P_1) \). A similar definition holds for \( P \).

Thus we require that variation in one while leaving the other constant gives a variation in effectiveness. Finally we need a technical condition which will not be explained here, that is the *Archimedean property* for each component. It merely ensures that the intervals on a component are comparable. For details the reader is referred to Krantz *et al.* We now have six conditions on the relational structure \( \langle R \times P, \geq \rangle \) which in the theory of measurement are necessary and sufficient conditions* for it to be an *additive conjoint structure*. This is enough for us to state the main *representation theorem*. It is a theorem asserting that if a given relational structure satisfies certain conditions (axioms), then a homomorphism into the real numbers is often referred to as a scale. Measurement may therefore be regarded as the construction of homomorphisms for
empirical relational structures of interest into numerical relational structures that are useful. In our case we can therefore expect to find real-valued functions $[[\Phi]]_1$ on $R$ and $[[\Phi]]_2$ on $P$ and a function $F$ from $Re \times Re$ into $Re$, 1:1 in each variable, such that, for all $R, R'$ [[proper subset]] $R$ and $P, P'$ [[proper subset]] $P$ we have:

$$(R, P) \geq (R', P') \iff F [[[\Phi]]_1 (R), [[\Phi]]_2 (P)] \geq F [[[\Phi]]_1 (R'), [[\Phi]]_2 (P')]$$

(Note that although the same symbol $\geq$ is used, the first is a binary relation on $R \times P$, the second is the usual one on $Re$, the set of reals.) In other words there are numerical scales $[[\Phi]]_i$ on the two components and a rule $F$ for combining them such that the resultant measure preserves the qualitative ordering of effectiveness. When such a representation exists we say that the structure is decomposable. In this representation the components ($R$ and $P$) contribute to the effectiveness measure independently. It is not true that all relational structures are decomposable. What is true, however, is that non-decomposable structures are extremely difficult to analyze. A further simplification of the measurement function may be achieved by requiring a special kind of non-interaction of the components which has become known as additive independence. This requires that the equation for decomposable structures is reduced to:

$$(R, P) \geq (R', P') \iff F [[[\Phi]]_1 (R) + [[\Phi]]_2 (P)] \geq F [[[\Phi]]_1 (R') + [[\Phi]]_2 (P')]$$

where $F$ is simply the addition function. An example of a non-decomposable structure is given by:
\[(R, P) \geq (R', P') \iff [(\text{Phi}1)(R) + [(\text{Phi}2)(P) + [(\text{Phi}1)(R' + [(\text{Phi}2)(P')] \geq [(\text{Phi}1)(R') + [(\text{Phi}2)(P')] + [(\text{Phi}1)(R' + [(\text{Phi}2)(P'])

* It can be shown that (starting at the other end) given an additively independent representation the properties defined in 1 and 3, and the Archimedean property are necessary. The structural conditions 4 and 5 are sufficient. Here the term \([(\text{Phi}1) [\text{Phi}2] is referred to as the interaction term, its absence accounts for the non-interaction in the previous condition. We are now in a position to state the main representation theorem.

**Theorem**

Suppose \(<R \times P, \geq > is an additive conjoint structure, then there exist functions, \([(\text{Phi}1) from R, and \([(\text{Phi}2) from P into the real numbers such that, for all \(R, R' \subset R and \(P, P' \subset P:

\[(R, P) \geq (R', P') \iff [(\text{Phi}1)(R) + [(\text{Phi}2)(P) \geq [(\text{Phi}1)(R') + [(\text{Phi}2)(P')

If \([(\text{Phi}1']) are two other functions with the same property, then there exist constants \([(\text{Theta}) > 0, [(\text{gamma})], and \([(\text{gamma})] such that

\[
[(\text{Phi}1')] = [(\text{Theta})]([\text{Phi}1] + [(\text{gamma})]1 ([(\text{Phi}2')] = [(\text{Theta})]([\text{Phi}2] + [(\text{gamma})]2

The proof of this theorem may be found in Krantz et al.[15].
Let us stop and take stock of this situation. So far we have discussed the properties of an additive conjoint structure and justified its use for the measurement of effectiveness based on precision and recall. We have also shown that an additively independent representation (unique up to a linear transformation) exists for this kind of relational structure. The explicit form of \([\Phi_i]\) has been left unspecified. To determine the form of \([\Phi_i]\) we need to introduce some extrinsic considerations. Although the representation \(F = [\Phi_i]1 + [\Phi_i]2\), this is not the most convenient form for expressing the further conditions we require of \(F\), nor for its interpretation. So, in spite of the fact that we are seeking an additively independent representation we consider conditions on a general \(F\). It will turn out that the \(F\) which is appropriate can be simply transformed into an additive representation. The transformation is \(f(F) = -(F - 1)[-1]\) which is strictly monotonically increasing in the range \(0 \leq F \leq 1\), which is the range of interest. In any case, when measuring retrieval effectiveness any strictly monotone transformation of the measure will do just as well.

**Explicit measures of effectiveness**

I shall now argue for a specific form of \([\Phi_i]\) and \(F\), based on a model for the user. In other words, the form \([\Phi_i]\) and \(F\) are partly determined by the user. We start by showing how the ordering on \(R \times P\) in fact induces an ordering of intervals on each factor. From Figure 7.13 we have that \((R3, P1) \geq (R1, P2), (R3, P1) \geq (R1, P1)\) and \((R1, P2) \geq (R1, P1)\). Therefore the increment (interval) \(R1R3\) is preferred to the increment \(P1P2\). But \((R2, P2) \geq (R4, P1)\), which gives \(P1P2\) is preferred to \(R2R4\). Hence \(R1R3 \geq 1 R2, R4\) where \(\geq 1\) is the induced order relation on
R. We now have a method of comparing each interval on R with a fixed interval on P.

Since we have assumed that effectiveness is determined by precision and recall we have committed ourselves to the importance of proportions of documents rather than absolute numbers. Consistent with this is the assumption of decreasing marginal effectiveness. Let me illustrate this with an example. Suppose the user is willing to sacrifice one unit of precision for an increase of one unit of recall, but will not sacrifice another unit of precision for a further unit increase in recall, i.e.

\[(R + 1, P - 1) > (R, P)\]

but
\((R + 1, P) > (R + 2, P - 1)\)

We conclude that the interval between \(R + 1\) and \(R\) exceeds the interval between \(P\) and \(P - 1\) whereas the interval between \(R + 1\) and \(R + 2\) is smaller. Hence the marginal effectiveness of recall is decreasing. (A similar argument can be given for precision.) The implication of this for the shape of the curves of equal effectiveness is that they are convex towards the origin.

Finally, we incorporate into our measurement procedure the fact that users may attach different relative importance to precision and recall. What we want is therefore a parameter \(\beta\) to characterise the measurement function in such a way that we can say: it measures the effectiveness of retrieval with respect to a user who attaches \(\beta\) times as much importance to recall as precision. The simplest way I know of quantifying this is to specify the \(P/R\) ratio at which the user is willing to trade an increment in precision for an equal loss in recall.

**Definition 6.** The relative importance a user attaches to precision and recall is the \(P/R\) ratio at which \(\frac{\partial}{\partial R} E = \frac{\partial}{\partial P} E\), where \(E = E(P, R)\) is the measure of effectiveness based on precision and recall.

Can we find a function satisfying all these conditions? If so, can we also interpret it in an intuitively simple way? The answer to both these questions is yes. It involves:
The scale functions are therefore, \( \Phi_1(P) = \alpha \left( \frac{1}{P} \right), \) and \( \Phi_2(R) = (1 - \alpha) \left( \frac{1}{R} \right) \). The 'combination' function \( F \) is now chosen to satisfy definition 6 without violating the additive independence. We get:

\[
F(\Phi_1, \Phi_2) = 1 - \frac{1}{\Phi_1 \cdot \Phi_2}
\]

We now have the effectiveness measure. In terms of \( P \) and \( R \) it will be:

\[
E = 1 - \frac{1}{\alpha \left( \frac{1}{P} \right) + (1 - \alpha) \left( \frac{1}{R} \right)}
\]

To facilitate interpretation of the function, we transform according to \( \alpha = 1/(\beta^2 + 1) \), and find that \( F \) is now:

\[
E = \frac{1}{\alpha \left( \frac{1}{P} \right) + (1 - \alpha) \left( \frac{1}{R} \right)}
\]

If \( A \) is the set of relevant documents and \( B \) the set of retrieval documents, then:

\[
P = \frac{|A \cap B|}{|B|} \quad \text{and} \quad R = \frac{|A \cap E|}{|A|}
\]

\( E \) now gives rise to the following special cases:

(1) When \( \alpha = 1/2 \) (\( \beta = 1 \)) \( E = \frac{|A \Delta B|}{|A| + |B|} \), a normalised symmetric difference between sets \( A \) and \( B \) (\( A \Delta B = A \cup B - A \cap B \)). It corresponds to a user who attaches equal importance to precision and recall.
(2) $E \rightarrow 1 - R$ when $[[\alpha]] \rightarrow 0 (\beta \rightarrow *)$, which corresponds to a user who attaches no important to precision.

(3) $E \rightarrow 1 - P$ when $[[\alpha]] \rightarrow 1 (\beta \rightarrow 0)$, which corresponds to a user who attaches no importance to recall.

It is now a simple matter to show that certain other measures given in the literature are special cases of the general form $E$. By the representation theorem, the $[[\Phi]]i$'s are uniquely determined up to a linear transformation, that is, $[[\Phi]]i'' = [[\Theta]][[\Phi]]i + [[\Gamma]]i$ would serve equally well as scale functions. If we now set $[[\Phi]]1'' = 2[[\Phi]]1 - 1/2$, $[[\Phi]]2'' = 2[[\Phi]]2 - 1/2$, and $\beta = 1$ then we have:

$$E = 1 - \frac{\frac{1}{2} + \frac{1}{2}}{\frac{1}{2} + \frac{1}{2} - 1}$$

which is the measure recommended by Heine[3].

One final example is the measure suggested by Vickery in 1965 which was documented by Cleverdon et al.[27]. Here we set:

$$\Phi_1 = 4\Phi_1 - \frac{3}{2}, \Phi_2 = 4\Phi_2 - \frac{3}{2}, \text{ and } \beta = 1$$

and obtain

$$E = 1 - \frac{1}{2 \left( \frac{1}{2} \right)^2 + 2 \left( \frac{1}{2} \right)^{-3}}$$

which is Vickery's measure (apart from a scale factor of 100).
To summaries, we have shown that it is reasonable to assume that effectiveness in terms of precision and recall determines an additive conjoint structure. This guarantees the existence of an additively independent representation. We then found the representation satisfying some user requirements and also having special cases which are simple to interpret. The analysis is not limited to the two factors precision and recall; it could equally well be carried out for say the pair fallout and recall. Furthermore, it is not necessary to restrict the model to two factors. If appropriate variables need to be incorporated the model readily extends to $n$ factors. In fact, for more than two dimensions the Thomsen condition is not required for the representation theorem.
Chapter Eight
Web search engines & suppressing controversy :
Chapter Eight

Web search engines & suppressing controversy:

8.1 Introduction

The Web evolved beyond FTP archives not just by becoming a graphically rich multi-media world, but by evolving tools which made it possible to find and access this richness. Oldsters like this author remember that before browsers there was WAIS (released 1991), and the XWAIS version provided a user-friendly GUI way to find information. However, this system required servers to organize information according to a specific format. GOPHER, another information serving system with some user-friendliness, was released the same year. One of the earliest search engines like those today, Lycos, began in the spring of 1994 when John Leavitt's spider was linked to an indexing program by Michael Mauldin. Yahoo!, a catalog, became available the same year. Compare this to the appearance of NCSA Mosaic in 1993 and Netscape in 1994. Today there are a score or more of "Web location services." A search engine proper is a database and the tools to generate that database and search it; a catalog is an organizational method and related database plus the tools for generating it. There are sites out there, however, that try to be a complete front end for the Internet. They provide news, libraries, dictionaries, and other resources that are not just a search engine or a catalog, and some of these can be really useful. Yahoo!, for example,
emphasizes cataloging, while others such as Alta Vista or Excite emphasize providing the largest search database. Some Web location services do not own any of their search engine technology - other services are their main thrust. Companies such as Inktomi (after a Native American word for spider) provide the search technology. These Web location services have put amazing power into every user's hands, making life much better for all of us and it's all free.

8.2 How Software Agents and Search Engines Work

There are at least three elements to search engines that I think are important:

(1) Information discovery & the database

(2) the user search (3) the presentation and ranking of results.

1. Discovery and Database

A search engine finds information for its database by accepting listings sent in by authors wanting exposure, or by getting the information from their "Web crawlers," "spiders," or "robots," programs that roam the Internet storing links to and information about each page they visit. Web crawler programs are a subset of "software agents," programs with an unusual degree of autonomy which perform tasks for the user. How do these really work? Do they go across the net by IP number one by one? Do they store all or most of everything on the Web? According to The WWW Robot Page, these agents normally start with a historical list of
links, such as server lists, and lists of the most popular or best sites, and follow the links on these pages to find more links to add to the database. This makes most engines, without a doubt, biased toward more popular sites. A Web crawler could send back just the title and URL of each page it visits, or just parse some HTML tags, or it could send back the entire text of each page. Alta Vista is clearly hell-bent on indexing anything and everything, with over 30 million pages indexed (7/96). Excite actually claims more pages. OpenText, on the other hand, indexes the full text of less than a million pages (5/96), but stores many more URLs. Inktomi has implemented HotBot as a distributed computing solution, which they claim can grow with the Web and index it in entirety no matter how many users or how many pages are on the Web. By the way, in case you are worrying about software agents taking over the world, or your Web site, look over the Robot Attack Page. Normally, "good" robots can be excluded by a bit of Exclusion Standard code on your site. It seems unfair, but developers aren't rewarded much by location services for sending in the URLs of their pages for indexing. The typical time from sending your URL in to getting it into the database seems to be 6-8 weeks. Not only that, but a submission for one of my sites expired very rapidly, no longer appearing in searches after a month or two, apparently because I didn't update it often enough. Most search engines check their databases to see if URLs still exist and to see if they are recently updated.

2. User Search

What can the user do besides typing a few relevant words into the search form? Can they specify that words must be in the title of a page? What
about specifying that words must be in an URL, or perhaps in a special HTML tag? Can they use all logical operators between words like **AND**, **OR**, and **NOT**? Most engines allow you to type in a few words, and then search for occurrences of these words in their database. Each one has their own way of deciding what to do about approximate spellings, plural variations, and truncation. If you just type words into the "basic search" interface you get from the search engine's main page, you also can get different logical expressions binding the different words together. Excite! actually uses a kind of "fuzzy" logic, searching for the **AND** of multiple words as well as the **OR** of the words. Most engines have separate advanced search forms where you can be more specific, and form complex Boolean searches. Some search tools parse HTML tags, allowing you to look for things specifically as links, or as a title or URL without consideration of the text on the page. By searching only in titles, one can eliminate pages with only brief mentions of a concept, and only retrieve pages that really focus on your concept. By searching links, one can determine how many and which pages point at your site. Understanding what each page does with the non-standard pluralization, truncation, etc. can be quite important in how successful your searches will be. For example, if you search for "bikes" you won't get "bicycle," "bicycles," or "bike." In this case, I would use a search engine that allowed "truncation," that is, one that allowed the search word "bike" to match "bikes" as well, and I would search for "bicycle OR bike OR cycle" ("bicycle* OR bike* OR cycle*" in Alta Vista).
Query Syntax Checklist

How does your engine handle:

**Truncation, Pluralization & Capitalization:**

Macintosh, Mac, Macintoshes, Macs, macintosh, macintoshes, mac, macs, could all yield different results. Most engines interpret lower case as unspecified, but upper case will match only upper case, but there are exceptions. There is no standard at all for truncation, and worse yet, it is probably different in general and advanced search mode for every engine.

**Presentation & Ranking**

With databases that can keep the entire Web at the fingertips of the search engines, there will always be relevant pages, but how do you get rid of the less relevant and emphasize the more relevant? Most engines find more sites from a typical search query than you could ever wade through. Search engines give each document they find some measure of the quality of the match to your search query, a relevance score. Relevance scores reflect the number of times a search term appears, if it appears in the title, if it appears at the beginning of the document, and if all the search terms are near each other; some details are given in engine help pages. Some engines allow the user to control the relevance score by giving different weights to each search word. One thing that all engines do, however, is to use alphabetical order at some point in their display algorithm. If relevance scores are not very different for various matches, then you end up with this sorry default. Zeb's [Whatever] page will never fare very
well in this case, regardless of the quality of its content. For most uses, a
good summary is more useful than a ranking. The summary is usually
composed of the title of a document and some text from the beginning of
the document, but can include an author-specified summary given in a
meta-tag. Scanning summaries really saves you time if your search
returns more than a few items.

**Get More Hits By Understanding Search Engines**

Knowing just the little bit above can give you ideas of how to give your
page more exposure.

**Hustle for Links**

Most software agents find your site by links from other pages. Even if
you have sent in your URL, your site can be indexed longer and ranked
higher in search results if many links lead to your site. One of my sites
that couldn't show up in the most casual search got most of its hits from
links on other sites. Links can be crucial in achieving good exposure.

**Use Titles Early In the Alphabet**

All engines that I used displayed results with equal scores in alphabetical
order.

**Submit Your URL to Multi-Database Pages**

It is best to use a multiple-database submission service such as SubmitIt!
to save you the time of contacting each search service separately.
Remember, it takes 6-8 weeks to become indexed.
Control Your Page's Summary
You can use the meta tag name="description" to stand out in search results. Appear in search summaries as "Experienced Web service, competitive prices" not "Hello and welcome. This page is about."

Search Reverse Engineering
Simulate your audience's search for your page (have all your friends list all the searches they might try), then see what you need to do to come up first on their search engine's results list.

1. Use the meta-tag name="keywords" to put an invisible keyword list at the beginning of your document that would match keywords your audience would use. Most search engines rate your page higher if keywords appear near the beginning.

2. How many times do the keywords appear in the text? It usually demonstrates good writing if you don't repeat the same words over and over. However, search engines penalize you for this, usually rating your page higher for repetitions of keywords, inane or not. Some authors combat this by putting yet more keywords at the bottom of their pages in invisible text. Look at the source code for this article, and you'll see what I mean; the words are just in the same color as the background.

SPAMMERS BEWARE
"Spamming" is net-lingo for spreading a lot of junk everywhere; keyword spamming is putting hidden keywords a huge number of times in your document just so yours will be rated higher by search engines.
1. Search engines typically limit you to 25 keywords or less, and one I know of truncates your list when they see an unreasonable number of repetitions.

2. Invisible text at the end of your pages puts blank space there, which looks bad and slows loading. Services which rate pages will enjoy marking you down for this.

**Responsible Keyword Use**: If an important keyword doesn't appear at least four times in your document, I hereby give you the right to add invisible text until it appears a maximum of five times.

### 8.3 Search Engine Features

Web location services typically specialize in one of the following: their search tools (how you specify a search and how the results are presented), the size of their database, or their catalog service. Most engines deliver too many matches in a casual search, so the overriding factor in their usefulness is the quality of their search tools. Every search engine I used had a nice GUI interface that allowed one to type words into their form, such as "(burger not cheeseburger) or (pizza AND pepperoni)." They also allowed one to form Boolean searches (except Hotbot as of 7/1/96, which promises to install this feature later), i. e. they allowed the user to specify combinations of words. In Alta Vista and Lycos, one does this by adding a "+" or a "-" sign before each word, or in Alta Vista you can choose to use the very strict syntax Boolean "advanced search." This advanced search was by far the hardest to use, but also the one most completely in the user's control (except for
OpenText). In most other engines, you just use the words AND, NOT, and OR to get Boolean logic. By far the best service for carefully specifying a search was Open Text. This form has great menus, making a complex Boolean search fast and easy. Best of all, this service permits you to specify that you want to search only titles or URLs. But then there's Alta Vista's little known "keyword" search syntax, now as powerful as OpenText, but not as easy to use. You can constrain a search to phrases in anchors, pages from a specific host, image titles, links, text, document titles, or URLs using this feature with the syntax keyword: search-word. There is an additional set of keywords just for searching Usenet. (To my knowledge, Alta Vista's keywords were undocumented before 7/19/96, so tell your friends you heard it here first!) . What could really make engines with large data bases shine, however, would be an improvement in the way they rank and present results. All engines I tested had ranking schemes that were not well documented, based on how many times your search words were mentioned, whether or not they appeared early in the document, whether or not they appeared close together, and how many search terms were matched. I did not find the ranking schemes very useful, as relevant and irrelevant pages frequently had the same scores.
Catalogs

I have only been disappointed by catalog services. In practice, they seem to aim for the lowest common denominator, and reflect very little thought to how and when they might be useful instead of search engines. All the ones I tested were directed toward novices and favored popular commercial sites. I would have thought they would be very good for finding software at least, but this was not the case. See the example below trying to find Web server related software.

Advanced or Boolean Queries

Making queries very carefully in Boolean terms to narrow a search rarely produces useful results for me (but see below). In practice, other

Useful Non-Search Goodies

E-mail address books:
Most engines allow you to search for someone's name if you quote it "John Q. Webhead", but you have to be careful about exact spelling, use of initials, etc.

News Services:
Yahoo! has the best news, in my humble opinion, as they have Reuters international news headlines. Most other news are ultra-brief summaries which read like "MacPaper."
ways of specifying a search besides detailed logic are much more useful. Specification of exact vs. approximate spelling, specification that search terms must appear as section headings or URLs, using more keywords, and just specifying the language of the document would have been more valuable in all of my search examples.

Example: Eliminating Unwanted Matches

The exception to this is the **AND NOT** operator - it is essential to exclude unwanted but close matches when they outnumber the desired matches. An example of when to use this operator is given by the problem of finding information on growing apples, because you will be deluged by information on Apple computers. With enough work, you can start to see apples with stems, not cords, but it isn't easy. Using Alta Vista, "+apple -mac* -comp* -soft* -hard* -vendor" got me information on the Payson-Santaquin apple farming region and a federal apple agriculture database on the first page of results.

Figure 7.1 Useful Search Features

Useful Search Features

**Find Images to Steal (Alta Vista)**

I bet you will all use this at one time or another, so I insist you credit this article and [webreference.com](http://webreference.com) for this goodie: With Alta Vista, you can limit your search to image titles by using the format: `image:title_string`
This was the only way I could find a useful picture of a nose for a physician's page - I had searched through jillions of clip art pages, and even contacted graphic artists, and they couldn't come up with anything as good as I found for free! USE THIS.

Try it now (replace ansel with your choice of image search string):

**Alta Vista Search:**

```
image:ansel
```

Search the Web

**Search for Strings in Titles (Alta Vista, OpenText)** for faster results. If applicable, this kind of search eliminates chaff by sticking to the pages that center on your subject, not ones that just mention a lexically related word. Use the syntax:

```
title:search_string
```

in Alta Vista, or just use the simple pull-down menus in OpenText's "advanced search mode."

**Find the Links to Your Own Site (Alta Vista)**

Alta Vista claims that you can get all the links to your own site by searching with the keyword construction: +link:http://mysite.com/ -host:mysite in the Simple query

...I found that the most important link to one of my sites was missing from this search, so I was not impressed; however, my editor swears by this. Try it now (replace webreference below with your site name):
Alta Vista Search:

+link:http://www.webreference.com-host:webreference

Search the Web

Find the Number of Links to Your Own Site (Alta Vista)

For a more accurate estimate of the actual number of links to your site (or backlinks), use Alta Vista's advanced search, and display the results as a "count only." The above method will give you links, but approximates their number, this method more accurately estimates the number of backlinks. Try it now (replace webreference below with your site name) ABK-12-29-96:

Search the Web and Display the Results as a Count only

Selection Criteria: Please use Advanced Syntax (AND, OR, NOT, NEAR).

Which is the Best Search Engine?

To decide which search engine I would choose as the best, I decided that nothing but useful results would count. Previous pages have emphasized quantified measures for speed and database sizes, but I found these had little relevance for the best performance in actual searches. By now, all engines have great hardware and fast net links, and none show
any significant delay time to work on your search or return the results. Instead, I just came up with a few topics that represented, I felt, tough but typical problems encountered by people who work on the net: First, I tried a search with "background noise", a topic where a lot of closely related but unwanted information exists. Next, I tried a search for something very obscure. Finally, I tried a search for keywords which overlapped with a very, very popular search keyword. I defined a search as successful only if the desired or relevant sites were returned on the first page of results.

Example - Search Terms Which Yield Too Many Matches

For the first type of search, I wanted to find a copy of Wusage to download, free software that lets you keep track of how often your server or a specific page is accessed, a common tool for HTML developers. This site is hard to find because output files are produced by the program on every machine running it that have the string "wusage" in their title and text. When I simply typed "wusage" into search page forms, Infoseek and Lycos were the only engines to find the free version of the software I wanted. (Note I gave no credit for finding the version for sale. A careful search of the sale version's page, did not produce any links to the free version's download site.) Infoseek's summaries were very poor, however, and all matches had to be checked.
Always Search As Specifically As Possible

Most engines failed to find their quarry because the search was too broad. After all, how is the engine supposed to know I want the free version? After spending a long time to find out the exact name of what I wanted, "wusage 3.2", Infoseek, Excite, Magellan, and Lycos all found the site I was interested in. Alta Vista, Hotbot, and OpenText yielded nothing of interest on their first page. Magellan came out the clear winner on this search, as the site summary was by far the best. (Asking Alta Vista to display a detailed version of the results didn't change things at all!) Infoseek and Excite performed well, but Lycos listed a much older version of wusage (2.4) first.

Think About Search Terms

It eventually occurred to me to search for "wusage AND free" to find the free copy of wusage. In some sense, Lycos was the winner this time because the free version was the first match listed; however, its summary was not very useful. While it did a better job than Infoseek, it didn't tell me whether each site was relevant or not. Magellan's response was very good, as it included a link leading to the software on the first page of matches, again with an excellent summary. Yahoo and Alta Vista also found it, but all these engines rated the fee version higher than the free version. OpenText did very well here, but only in advanced search mode where it was possible to specify that wusage must be in the title, and "free" could be anywhere in the text. Wusage3.2 was listed as the second
of only two entries - no digging here! Excite failed to find the site at all, and HotBot found only 10 matches for statistics of a server in Omaha.

Curiously, a search for "download wusage" did not improve the results over the single-word searches for any of the search engines! (It may be time for rudimentary standardized categories to be used on the Web: e.g. this is a download archive, this is an information only site, this is an authoritative site, etc.) The lesson here may just be "if at first you don't succeed..."

Catalogs

Catalogs were not helpful. Yahoo!, under computers/software had nothing whatever to try for wusage: no http, no HTML, no wusage, not even servers. In Excite!, under computing/www/web ware, three more clicks got me to wusage, but -surprise!- I could not get to the free version. See why you don't want anyone else filtering your information?

Example:- Finding The Really Obscure

For this example, let's try to find out how to care for a "tegu", a South American lizard that is only moderately popular even among lizard enthusiasts. (If that's not an adequate example of obscure information, I don't know what is.) I know that a page exists called "TEGU INTRO" at http://www.concentric.net/~tegu/tegu.html, but we will simulate a blind search here. This search was full of surprises. First I began by just searching for the string "tegu." Infoseek's first match was a tegu page I did NOT know about! Still, the one I wanted was not listed on the first
page. Excite yielded nothing about tegus, only information on a vaguely related reptile, the "dwarf tegu." A search on the string "tegu care" yielded nothing relevant. (A search on their handy Usenet database did find the old tegu article I was looking for, three weeks old, which was no longer on my local news server. Other engines found this as well.) Lycos came up with the URL Infoseek found, plus two more, however, the additional listings were only pictures, not information. Searching for the string "tegu care" got nothing. Alta Vista found nothing useful either way, just ads for lizard food. OpenText found nothing, even when I searched for "tegu lizard." Hotbot found a picture of a tegu with "tegu care," but it did not return any relevant information with any search.

None of the searches I tried came up the URL I knew about. The lesson here is that you can really find new things on the Web with search engines, but if you need to find a specific page, it will always be a crap shoot. Advanced searches yielded nothing more with any engine ("tegu in title AND (care or lizard)", etc.) Some way to require that the searches were only among English language documents would have been much more helpful. Some northern-European sounding language apparently has the word tegu in it, not referring to a lizard, and many foreign language pages fouled my results on some engines. Another feature that would really have made a difference would be a filter for sales pages -- most of the mentions of tegu on the net are ads for "Monitor and Tegu Food", containing no care information. As expected, Yahoo! and Excite! Catalogs were useless here as well.

Example :- Selectivity: Apple Trees NOT Apple Computers
There are gobs of stuff on the net about Apple Computers, but what about growing apple trees? Surprisingly, this search was very easy! apple* alone always yielded lots of stuff about the computers, and one often had to add as many as five excluded terms (apple* -vendor* -hard* -soft* -comp* -mac*) before receiving any matches for apples you can eat. Surprisingly, however, just apple* tree* usually yielded detailed information on growing apple trees on the first page of results. The poorer results required one to increase the search command to apple* tree* grow*.

And The Winner Is. . .

I don't really want to pick a winner. . . All right, if you insist: The "Search Test Results . . ." table, below, lists the engines in order of their ranking. Lycos is therefore the official heavy weight search engine champion of the universe, based on the tests above. However, I think this is missing the point. As shown in the table, "Which Search Page . . .?", above, you should choose different engines for different tasks. None of the engines tested were able to limit their searches to images except for Alta Vista. This engine must therefore surely be the best one for graphics designers if they are allowed to use only one, but for most other purposes, the user will have to wade through the mountains of chaff and drek to find what they want. It is more beneficial to use different engines for different tasks; at most only a few are required.
**Table: 7.1 Search Engine Test Results**

<table>
<thead>
<tr>
<th>Engine</th>
<th>&quot;One Item Among Many Related Pages&quot; Test</th>
<th>&quot;Obscure Item&quot; Test</th>
<th>&quot;Selectivity: Apple Trees Not Computers&quot; Test</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lycos</td>
<td>Found item with broad search word and exact name. Found item first on results list with two search terms.</td>
<td>Found unknown item, but not known item.</td>
<td>Just apple$ tree$ yielded good results.</td>
<td>Returned the most relevant matches in the tests, but requires more time to check bad matches than Magellan.</td>
</tr>
<tr>
<td>Infoseek</td>
<td>Found item with broad search word</td>
<td>Found unknown</td>
<td>Just apple$</td>
<td>Poor Summarie</td>
</tr>
<tr>
<td><strong>OpenText</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Found ususage in title search</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good results with 2 or 3 terms, most useful with 3 terms due to superior summaries.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to specify title searches very useful and user-friendly. Summaries very good.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Alta Vista</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed with approximate and exact words. Found item low on first page with two search terms.</td>
</tr>
<tr>
<td>Good results with apple* tree* grow*.</td>
</tr>
<tr>
<td>Keyword searches for images, titles, etc. are very</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
</tbody>
</table>
| **Magellan** | Found with exact name.  
Found item low on first page with two search terms. | Found nothing | Required three search terms: apple*  
tree*  
grow*  
Superior summarie s always save you surf time. |
| **Excite** | Found with exact name, failed with two word search. | Found nothing | Required third search term: apple*  
tree*  
grow*, even then irrelevant results were first. |
<table>
<thead>
<tr>
<th><strong>HotBot</strong></th>
<th>Failed all searches</th>
<th>Failed all searches</th>
<th>Found only images, and did worse when grow* was added!!!</th>
<th>Poorest Performer (excluding catalogs).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Excite! Catalog (not engine)</strong></td>
<td>Failed all searches</td>
<td>Failed all searches</td>
<td>Failed all searches</td>
<td>Catalogs not at all useful.</td>
</tr>
<tr>
<td><strong>Yahoo! Catalog (not engine)</strong></td>
<td>Failed all searches</td>
<td>Failed all searches</td>
<td>Failed all searches</td>
<td>Catalogs not at all useful.</td>
</tr>
</tbody>
</table>

**IV. Conclusion**

Different engines have different strong points; use the engine and feature that best fits the job you need to do. One thing is obvious; the engine with the most pages in the database IS NOT the best. Not surprisingly, you can get the most out of your engine by using your head
to select search words, knowing your search engine to avoid mistakes with spelling and truncation, and using the special tools available such as specifiers for titles, images, links, etc. The hardware power for rapid searches and databases covering a large fraction of the net is yesterday's accomplishment. We, as users, are living in a special time when search engines are undergoing a more profound evolution, the refinement of their special tools. I believe that very soon the Web will evolve standards, such as standard categories, ways of automatically classifying information into these categories, and the search tools to take advantage of them, that will really improve searching. I think it's exciting to be on the Web in this era, to be able to watch all the changes, and to evolve along with the Web as we use it.

8.4. The nature of meaning in the age of Google

The culture of lay indexing has been created by the aggregation strategy employed by Web search engines such as Google. Meaning is constructed in this culture by harvesting semantic content from Web pages and using hyperlinks as a plebiscite for the most important Web pages. The characteristic tension of the culture of lay indexing is between genuine information and spam. Google's success requires maintaining the secrecy of its parsing algorithm despite the efforts of Web authors to gain advantage over the Googlebot. Legacy methods of asserting meaning such as the META keywords tag and Dublin Core are inappropriate in the lawless meaning space of the open Web. A writing guide is urged as a necessary aid for Web authors who must balance enhancing expression versus the use of technologies that limit the aggregation of their work.
The age of Google

Financial markets anticipate Google's initial public stock offering to be valued at $15 billion to $25 billion (Martinuzzi, 2003). The magnitude of these figures reflects Google's pre-eminence as a Web search engine:

- The readers of Search Engine Watch voted Google 'Outstanding Search Service' three years in a row (Sullivan, 2003, January 28).
- Google performs the greatest number of searches per day: 250 million in February 2003 (Sullivan, 2003, February 25).
- Google dominates statistical comparisons of search engines on factors such as 'relative size', 'freshness', 'change over time' and so on (Notess, 2003, May 28).
- 'To google' has become a verb. Google was the unanimous choice for 'Most Useful Word' for 2002 by the American Dialect Society.

"I recently went to Silicon Valley to visit the offices of Google, the world's most popular search engine. It is a mind-bending experience. You can actually sit in front of a monitor and watch a sample of everything that everyone in the world is searching for. (Hint: sex, God, jobs and, oh my word, professional wrestling usually top the lists.)... In the past three years, Google has gone from processing 100 million searches per day to over 200 million searches per day. And get this: only one-third come from inside the U.S. The rest are in 88 other languages." (Friedman, 2003, June 29)
Google Harvests *Lay Indexing*

Google harvests the content placed in public Web space by millions of anonymous, independent Web authors. Google parses the text found in Web pages and uses hyperlinks among Web pages to calculate a PageRank score. The PageRank calculation includes the number of incoming to and outgoing links from a Web page, and favorably weights in-coming links from Web pages that have large PageRank scores.

*The citation (link) graph of the Web is an important resource that has largely gone unused in existing Web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a Web page's 'PageRank', an objective measure of its citation importance that corresponds well with people's subjective idea of importance. (Brin & Page, 1998.)*

Probability dictates that PageRank will successfully capture the subjective sense of Web-page importance. If a large number of Web users in the role of authors create content that points at certain Web pages, then it is highly probable that those same Web pages presented as query results will satisfy a large number of Web users in the role of searchers. In other words, Google satisfies the average Web searcher so well because it has aggregated the valuations of the average Web author. In this way, Google transforms Web authors into *lay indexers* of Web content where the linkages they set is a plebiscite for the most 'important' Web pages.
For example, a recent search for 'dogs' returned a retrieval set of more than 14.5 million Web pages with these three first:

1. I-love-dogs.com (PageRank = 6/10 on January 29, 2004),
2. Guide dogs for the blind (PageRank = 6/10), and
3. American Kennel Club (PageRank = 6/10)

The combination of the PageRank of these Web pages, their use of the word 'dogs', and the hyperlink text pointing at these Web pages permits Google to bet that these are the most likely Web pages to satisfy the average Web searcher looking for 'dogs'. Google's pre-eminence as a Web search engine is clear evidence that this is a winning bet most of the time.

8.5 Aggregating Meaning

Google's innovation, which is worth billions, is to crawl rapidly over public Web space each month or so, and then reflect back to the Web community the words and valuations of Web content that the Web community itself has placed there. In this way Google aggregates the meaning expressed by lay indexers in their textual Web content, their hyperlinks and hyperlink text. Utilizing hyperlink text has a distinguished pedigree: Henry Small (1978) suggested that citations in text act as concept symbols more than thirty years ago. Aggregating meaning is possible on the Internet because there are many easily accessible semantic objects to be harvested. Analysis of the aggregations can suggest patterns of high likelihood that permit applications to recommend, adapt, profile,
forecast and so on. An aggregation strategy permits Google to suggest the most likely Website to satisfy your query, Amazon.com to suggest a likely book for purchase, and governments to collect clues about terrorists. These are all examples of aggregating the meaning, taste, judgment, knowledge, etc., of a large universe of anonymous, independent agents to determine a common value. In a similar fashion a stock market pools multiple buys and sells to find a price for an equity.

Some examples of Internet aggregator applications include:

- Blogdex tracks the diffusion of information through the blogosphere. **Blogdex tracks the diffusion of information through the blogosphere.** *Logdex uses the links made by Webloggers as a proxy to the things they are talking about. Webloggers typically contextualize their writing with hypertext links which act as markers for the subjects they are discussing.... Blogdex crawls all of the Weblogs in its database every time they are updated and collects the links that have been made since the last time it was updated. The system then looks across all Weblogs and generates a list of fastest spreading ideas.* (About Blogdex.)

- Iowa Electronic Markets is a real-money futures market where participants can buy contracts on future economic and political events such as elections. The efficient market hypothesis suggests that market prices reflect the cumulative knowledge of large numbers of individual investors. This technique was briefly considered as a clever method of anticipating future terrorist activity, until it was found to be politically unacceptable:
'Research indicates that markets are extremely efficient, effective and timely aggregators of dispersed and even hidden information,' the Defense Department said in a statement. 'Futures markets have proven themselves to be good at predicting such things as elections results; they are often better than expert opinions.' (Hulse, 2003, July 29.)

- Retailers like Amazon.com use recommender systems to personalize shopping.

  At Amazon.com, we use recommendation algorithms to personalize the online store for each customer. The store radically changes based on customer interests, showing programming titles to a software engineer and baby toys to a new mother. The click-through and conversion rates—two important measures of Web-based and email advertising effectiveness—vastly exceed those of untargeted content such as banner advertisements and top-seller lists. (Linden, et al., January 2003.)

- Specialized aggregators focus on price comparisons (Madnick, et al., 2000, October 22), news feeds (Singel, 2003, August 18) and so on. On the horizon, unbeknownst to you, a new entity, whose plans are to overturn the familiar business landscape, is fast emerging. A shopbot-like aggregator can selectively extract information from your Website, couple it with additional data from other sources including those of your competitors, and make the
necessary fine tuning for intelligent comparisons. (Madnick, et al., 2000, October 22.)

While semantic objects are readily available for collection on the Internet, the possibility always exists that someone has anticipated your collection and is fooling you. In short, the convenience of surreptitiously collecting information from other people is matched by the fear that they may be manipulating your Web-bot aggregator to their advantage. This introduces the characteristic tension between information and spam in the culture of lay indexing.

8.6 Information and spam

Google's most important corporate asset is its ability to collect genuine Web authorship, i.e., the Web community going about their daily lives creating content and linking to Web pages that they find useful. Bad faith occurs when a Web author attempts to gain an advantage over Google, and assert his singular meaning in place of the meaning aggregated from the Web community. A common bad faith technique is loading a Web page with words that the Googlebot will find, but are invisible to Web readers. It also includes link farming, a cooperative sharing arrangement of links, and Google bombing, which coordinates a large number of linkages to a single page. 'Cloaking' occurs when a Web server recognizes a request from the Googlebot and responds with special content: The term 'cloaking' is used to describe a Website that returns altered Webpages to search engines crawling the site. In other words, the Web server is programmed to return different content to Google than it
returns to regular users, usually in an attempt to distort search engine rankings. This can mislead users about what they'll find when they click on a search result. To preserve the accuracy and quality of our search results, Google may permanently ban from our index any sites or site authors that engage in cloaking to distort their search rankings. (Google Information for Webmasters).

Unfortunately for Google and Internet aggregators in general, bad faith is attractive because it can have a big pay-off. Goldhaber's (1997) 'attention economy' compares the deluge of available digital information to the limited supply of human time and attention. In the attention economy, information is plentiful and human attention is scarce. Huberman's (2001) survey indicates that 0.1% of Websites capture 32.3% of activity, indicating that the vast majority of Web content languishes in obscurity. Therefore, a hyperlink from a stranger who has made an unforced choice to highlight your Web content has great value. Imagine the by-passed author's chagrin at the neglect of his Web pages, and the temptation to finagle just a little bit to propel his Web pages out of the obscurity of the retrieval set of 14.5 millions to appear beside the top three Web pages for the query 'dogs'. Search engines are constantly adding and removing pages, as well as altering the algorithms they use to rank pages. However, there's a great obsession with Google because of the large amounts of traffic it can deliver. Of the four most popular search engines—Google, Yahoo, AOL and MSN Search—Google's results are used at the first three. (Sullivan, 2003, December 1).
The controversy between Google and Daniel Brandt, author of NameBase, illustrates the obsession with Google's ability to shine the spotlight of attention and the dangers of bad faith. If you misperceive Google to be a large Web database under the control of a system administrator, and you found your Web content indexed but ignored, you would probably conclude that you need only lobby the administrator to get the spotlight of attention to shine on your content.

'My problem has been to get Google to go deep enough into my site,' he says. In other words, Brandt wants Google to index the 100,000 names he has in his database, so that a Google search for 'Donald Rumsfeld' will bring up NameBase's page for the secretary of defense. (Manjoo, 2002).

But Google's rankings are not the result of a systems administrator's arbitrary judgment. If Google accedes to Brandt and adjusts the valuation of the content on the NameBase Website, then it wounds itself by permitting Brandt, and not the community of lay indexers, to assert the meaning and value of the NameBase Web content. Google's concession to Brandt would lower the quality of Google's retrieval because search results would no longer reflect the average Web user, but a single individual's judgment of the value of the NameBase Website. Google's continued success depends on its ability to collect unaffected Web content, which means that it must avoid the single individual's assertion of meaning. This strategy implies that any metadata scheme for the Web that promotes the meaning assertion of a single Web author (i.e., My Web page means this) will be avoided by aggregators. The strategy of aggregation, the enlistment of Web authors as lay indexers, and the
temptation of bad faith points to the importance of maintaining the ignorance of lay indexers.

**The importance of ignorance**

Consider for a moment the various strategies Google could pursue to maximize the collection of genuine Web authorship and minimize bad faith. Google could, for example, publicize its algorithms and then admonish everyone to behave. The Internet is, however, a network of anonymous, independent agents characterized by viruses, worms, spyware, music piracy, identity theft, etc., that transcends national borders, invades personal privacy, abuses enterprise intranets, etc. The Internet often appears to be *beyond any law*; therefore, it would be foolish to believe that anyone would behave. Google's only possible survival strategy is to keep its parsing and ranking algorithms absolute secrets. In short, the culture of lay indexing is one of mistrust and ignorance: The lay indexer's ignorance of when, if, and how her work will be used, and Google's mistrust of lay indexers, whom it must assume are constantly scheming to gain an advantage over the Googlebot. For example, current interest focuses on a 'filter test' ([Sullivan, 2003, December 1](#)) of systematically adding and subtracting query terms in hopes of revealing Google's underlying algorithm. *Google's order of results is automatically determined by more than 100 factors, including our PageRank algorithm.... Due to the nature of our business and our interest in protecting the integrity of our search results, this is the only information we make available to the public about our ranking system.* ([PageRank Information](#)). Compounding the lay indexer's ignorance of Google's
algorithm is the unpredictable traversal of Web space. The following table gives the 2002-2003 Googlebot monthly page requests of my own Website. During this two-year period, the number of my Web pages did not change dramatically, nor were there any substantial changes in Website architecture, password use, hosting server address, etc.

Table 8.1 combine repeated visits of the Googlebot in the same month, if any repeated visits were made.

<table>
<thead>
<tr>
<th></th>
<th>2002 Page Requests</th>
<th>2003 Page Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>307</td>
<td>January</td>
</tr>
<tr>
<td>February</td>
<td>132</td>
<td>February</td>
</tr>
<tr>
<td>March</td>
<td>309</td>
<td>March</td>
</tr>
<tr>
<td>April</td>
<td>325</td>
<td>April</td>
</tr>
<tr>
<td>May</td>
<td>766</td>
<td>May</td>
</tr>
<tr>
<td>June</td>
<td>345</td>
<td>June</td>
</tr>
<tr>
<td>September</td>
<td>179</td>
<td>September</td>
</tr>
<tr>
<td>October</td>
<td>695</td>
<td>October</td>
</tr>
<tr>
<td>November</td>
<td>504</td>
<td>November</td>
</tr>
<tr>
<td>December</td>
<td>528</td>
<td>December</td>
</tr>
</tbody>
</table>
If Google's most important corporate asset is its ability to collect ineffective Web authorship, then maintaining a lay indexing culture of absolute ignorance is the best guarantor of future success. Web authors outraged at their helplessness might seek help from SEOs (Search Engine Optimizers) who promise to promote or manage the visibility of Websites, but Google warns of the consequences of unscrupulous activity:

If an SEO creates deceptive or misleading content on your behalf, such as doorway pages or 'throwaway' domains, your site could be removed entirely from Google's index. (Search Engine Optimizers).

Probably the best strategy for the average Web author is simply to construct Web pages that are as welcoming to the Googlebot as possible, and then wait patiently for the Googlebot to come by and visit them. Setting out feed for wild birds is an analogous activity. Struggling to maintain the ignorance of lay indexers in the culture of lay indexing contrasts sharply with the historical treatment of indexers. During the last several hundred years in the craft of book arts and scholarly journals, indexers have been honoured and respected. In this legacy culture of indexing, indexer ignorance was an anathema to be avoided, not enhanced.

8.7 Traditional methods of constructing meaning

We inherit a tradition of constructing meaning by trusting the expertise of a few. For example, the claim has been made that indexers possess a special skill for denoting the meaning of text:
Above all, what may be called the 'index sense' is required—that is, the ability to feel instinctively, at the first glance, what and how subjects should be indexed in all their ramifications; the sense that is in touch with searchers, and appreciates just how subjects will be looked for and how to arrange so that they can most readily be found. Experience is the only school in which these qualifications can be gained. (Nichols, 1892: 406). Meaning and trust are also implicit in database management. When the U.S. Department of Education builds a database of education resources (e.g., the ERIC database), a submission is evaluated by subject experts who select topical terms to express its meaning. *A document sent to ERIC is evaluated by subject experts (Submitting Documents to ERIC)* ....The indexer, or abstractor/indexer, examines the document, chooses the concepts to be indexed, and translates these concepts into the indexing terminology of the system. (ERIC Processing Manual). One reason that traditional information systems could rely on the meaning assertion of a few individuals was that these systems were devised, built and managed by information professionals. Professionals were known, publicly accessible and held to high standards of ethics. Information professionals, such as librarians, were considered to be operating a public trust with a view to the best interests of society. Rare was the librarian who abused collection policy to overload a public library with books she penned herself. Rare was the database administrator who filled a public database with his own database records. Professionals who abused the trust given to them by society could be brought to account. Another reason that traditional information systems could rely on the meaning assertion of a few individuals was that access to these systems was tightly
controlled. It was not the case that an anonymous individual could defy responsible information professionals and arbitrarily add an item to a library or database, and furthermore, independently declare its meaning:

*Another big difference between the Web and traditional well controlled collections is that there is virtually no control over what people can put on the Web. Couple this flexibility to publish anything with the enormous influence of search engines to route traffic and companies... deliberately manipulating search engines for profit become[s] a serious problem. This problem that has not been addressed in traditional closed information retrieval systems.* (Brin & Page, 1998).

Traditional closed information systems honored the assertion of meaning by a single individual, but to succeed Google must distrust it. This is the social consequence of a network technology that permits anyone to conflate the roles of author, indexer and publisher. That is, the Internet is an 'open' system where anyone can author anything and declare its meaning, i.e., *a lawless meaning space*. A lawless meaning space is a novelty that most traditional meaning technologies have not anticipated. Being able to operate successfully in a lawless meaning space is, however, the key success criterion for legacy meaning technologies that are applied to Web space.

**Technologies for asserting meaning on the Web**

The notion that the Web community would cooperate to construct information objects and then share them freely is very compelling. It echoes historical ambitions of amassing all world knowledge, e.g., the *World Brain* suggestion of H.G. Wells (1937), and using associative links
to create trails among pieces of information, e.g., the memex device of Vannevar Bush (1945 July). Recently the notion of a cooperating Web community has been expressed as the 'Semantic Web':

*The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.*

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. For the semantic Web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. (Berners-Lee, 2001, May 17).

At this time the Semantic Web remains more aspiration than reality, but clearly the vision would include 'software agents roaming from page to page' making determinations of meaning by using 'structured collections of information and sets of inference rules.' If structured collections of Web content mean metadata created by the author of the Web page, then this would be another example of privileging the assertion of meaning by a single individual, just what Google must avoid. Structured metadata created by Web page authors are another form of the Daniel Brandt controversy where a single individual attempts to promote his single meaning ahead of the meaning and value given to his Web content by the Web community. Example technologies that privilege the single assertion of meaning:
- The HTML recommendation of the World Wide Web Consortium suggests that the META element be used to specify keywords to help a search engine determine the meaning of a Web page. The recommendation offers the following example to indicate that a Web page is about vacationing in sunny Greece.

  <META name="keywords" content="vacation, Greece, sunshine">

Since the META keywords tag is designed for a Web author to claim My Web page means this, it's hardly surprising that Sullivan (2003, March 20) reports that the keywords tag is avoided by all major search engines.

- The Dublin Core metadata set is designed to facilitate interoperability, i.e., the sharing of metadata across applications. It is, however, avoided by aggregators as "spam" (FAQ).

- The Resource Description Framework (RDF)

  ...will make retrieval far faster and more accurate than it is now. Because the Web has no librarians and every Webmaster wants, above all else, to be found, we expect that RDF will achieve a typically astonishing Internet growth rate once its power becomes apparent. (Bosak & Bray, 1999, May).

Three years later Eberhart (2002, August 15) reports that "RDF has not caught on with a large user community."
Formal metadata schemes that require cooperation and good faith to work have been applied to the Web, but remain marginal:

A discouraging aspect of metadata usage trends on the public Web over the last five years is the seeming reluctance of content creators to adopt formal metadata schemes with which to describe their documents. For example, Dublin Core metadata appeared on only 0.5 percent of public Website home pages in 1998; that figure increased almost imperceptibly to 0.7 percent in 2002. The vast majority of metadata provided on the public Web is ad hoc in its creation, unstructured by any formal metadata scheme. (O'Neill, 2003). Of course Google has always disdained structured metadata in the open Web as bad faith:

Also, it is interesting to note that metadata efforts have largely failed with Web search engines, because any text on the page which is not directly represented to the user is abused to manipulate search engines. There are even numerous companies which specialize in manipulating search engines for profit. (Brin & Page, 1998).

Since the Web is a lawless meaning space, you may garnish your Web pages with any sort of metadata scheme you like. But formal metadata schemes that require cooperation and good faith of a community of Web authors will probably have a greater chance of working in 'closed' Web applications that honor the meaning assertions of single individuals, establish trust among strangers and enforce norms of application. Examples may be corporate intranets and digital libraries. Pity the poor Web author! Condemned to a culture of ignorance and denied any direct assertion of meaning of her content! She is encouraged to act naturally,
constructing her Web content and linking to Web pages of interest. Acting naturally, however, is not without hazard in a rapidly changing, technologically complex environment where it is easy to do something 'neat' that inadvertently makes your content unpalatable to the visiting Googlebot. There is a fine line between using technology to jazz up your Web page and using technology that unintentionally limits the aggregation of your content.

**Cool tricks and feeding the Googlebot**

The irony of constructing content for the open Web does not know how aggregators will use it. Any trick you employ to reduce your ignorance (i.e., you successfully spam the Googlebot) will be ultimately neutralized, throwing you back to the position of total ignorance:

> *Google prefers developing scalable and automated solutions to problems, so we attempt to minimize hand-to-hand spam fighting. The spam reports we receive are used to create scalable algorithms that recognize and block future spam attempts.* (Google Information for Webmasters).

The SEO industry awaits for incredulous authors who do not believe that Google will protect its most precious corporate asset: our ignorance of its parsing algorithm. It is helpful to remember that the motivation of the SEO industry is to make money. Pandia SEO, for example, offers a book for sale titled *The unfair advantage book on winning the search engine wars*, which warned in January 2004: *Beware of Google's new Over-Optimization Penalty!!! ...what was a strategy for top positioning is now*
a formula for disaster. Pages that were showing in the top ten have slipped all the way down under 1000 in the rankings. Even worse, the penalty appears to be permanent so this is a mistake to avoid at all costs. (Planet Ocean Communications, 2004). As an example, a SEO might suggest that you use more than four, but fewer than seven keywords in a META field. If such a stratagem were actually to work, then it would be rapidly employed by everyone else, thus diluting its effect and throwing you back again to the position of having no special advantage. Furthermore, Google is constantly tweaking its parsing formula so you're aiming at a moving target: In its latest makeover, Google also tweaked the closely guarded formula that determines which Websites are most relevant to a search request. Google has made five significant changes to its algorithmic formulas in the past two weeks, Brin said. (Liedtke, 2004, February 18). I argue the need for a survival guide for Web authors (without attempting to provide one here). A survival guide helps someone survive, largely by avoiding hazards, as opposed to being a bag of tricks for besting someone else. The need for a survival guide becomes compelling when you witness someone writing for the Web, but doing it in a manner that offends the Googlebot. Google has a list of technological hazards to avoid such as JavaScript, cookies, session IDs, frames, DHTML and Flash. The penalty of inhibiting the Googlebot is limiting the exposure of your work. One would think that poets would be anxious to place their work before a large public, but consider the submission guidelines of Poems That Go: Poems that Go publishes Web-specific new media, hypermedia, and electronic poetry, prose, and short narrative. We are open to all forms of multimedia, computer-generated,
and interactive work that include (but are not limited to) HTML, Shockwave, Quicktime, streaming media, Flash, Java, and DHTML content. Because Poems that Go focuses on how sound, image, motion, and interactivity intersect with literary uses of the Web, we regretfully do not accept text-based poetry or written work in the traditional sense. (Submission guidelines). Such is the gulf that exists between creating cool stuff for the Web and preparing something appetizing for the Googlebot. This problem is also illustrated by the PAD project (Preservation, Archiving and Dissemination) of the Electronic Literature Organization. PAD struggles to maintain access to classic e-texts in formats such as HyperCard, Story space, and BZT ('Better than Zork'), a proprietary system that sold commercially for less than a year. Other classic e-texts require a melange of DHTML, Flash, RealAudio, VRML, animated gifs and so on, none of which are tasty to the Googlebot. It may be that some digital artists are willing to sacrifice exposure and wide dissemination of their work to achieve eye-popping technical effects, but I argue that the average Web author needs a survival guide to help her avoid self wounding in the pursuit of the cool.

The meaning Google misses

Google may index billions of Web pages, but it will never exhaust the store of meaning of the Web. The reason is that Google's aggregation strategy is only one of many different strategies that could be applied to the semantic objects in public Web space. Hidden in the 'dogs' retrieval set of 14.5 millions are special, singular, obscure, unpopular, etc., Web pages that await a different aggregation strategy that would expose their
special meanings. To charge that Google has a bias against obscure Websites (Gerhart, 2004), or that we suffer under a 'Googlearchy' (Hindman, et al., 2003) of a few heavily linked Websites is to expect Google to be something other than Google. Google finds the common meanings. Many other meanings exist on the Web and await their aggregators.

8.8 Understanding Web behavior: Politics, technology and users

Web behavior depends upon three interlocking communities:

(1) authors whose Web pages link to other pages;

(2) search engines indexing and ranking those pages; and

(3) information seekers whose queries and surfing reward authors and support search engines.

Systematic suppression of controversial topics would indicate a flaw in the Web’s ideology of openness and informative ness. This chapter explores search engines’ bias by asking: Is a specific well–known controversy revealed in a simple search? Experimental topics include: distance learning, Albert Einstein, St. John’s Wort, female astronauts, and Belize. The experiments suggest simple queries tend to overly present the "sunny side" of these topics, with minimal controversy. A more "Objective Web" is analyzed where: (a) Web page authors adopt research citation practices; (b) search engines balance organizational and analytic content; and, (c) searchers practice more wary multi–searching. Recent
potential policy changes open the possibility for greater concentration of media ownership, possibly leading to fewer and more overlapping sources of information. How does the Web weigh into this picture? Harvard Digital Government researchers argue the Internet is not really a counterbalance to traditional media bias because the Web itself is dominated by a few gatekeeper and winner–take–all link–accumulating sites. Is the Web’s apparent openness, diversity, and cost effective information dissemination an illusion? One empirical study investigated link characteristics of highly controversial topics such as gun control and abortion. Crawling three million pages confirmed that a few sites accrued most of the inlinks, suggesting a normal surfer would be "pulled" toward those sites, rather than traveling to smaller and more diverse sites. Because search engines heavily use page links to rank results, searchers often start surfing at these same popular sites. This study also found Web site size and linking often mirrors well established traditional organizations, with only occasional Web–based newcomer groups. While a topic may be presented from many viewpoints and published cheaply, less popular sites are not necessarily easily accessible via search engines — nor by surfing. *Retrievability and visibility are quite different, but often confused in technical/political discussions.* Recognizing such bias is not an accusation of unfairness but rather is symptomatic of our growing understanding of complex Web technology operating at a scale of billions of pages and hundreds of millions of users. Although not precisely characterized, search engines collectively cover much less than the whole Web and individual search engines index different parts of the Web, further exposing alternative orderings by their ranking strategies. Indeed,
those biases attract searchers to favor one engine over another. Search engine bias has been mathematically characterized as deviation from norm or ideal, e.g. for benchmarking search engines on consumer decision–making queries such as "brand names of refrigerators." Such bias is termed indexical, versus concerns of propaganda and misinformation as content bias. Search engines are thus similar to media organizations, warning "too much consolidation [of either] limits the options for both information producers and information seekers." In practical terms, this study counters the widely held misconception of search engines behaving somewhat like objective and well–informed librarians. Further political concerns are voiced about the role of search engines in supporting or thwarting the inclusiveness ideology of the Web. A cause–effect description of search engine indexing practices illustrates certain political consequences. Indeed, the Web might be thought of as an economy of links, valued as both monetary and intellectual currency. The commercial search engine optimization industry raises other bias issues tracking ever changing and unclear search engine strategies. Consumer surveys about paid placement in search results reinforce a common theme: The general public that is growing increasingly dependent upon search engine technology has relatively low understanding of how the technology works or their responsibilities for its proper use. Beyond economics and politics, the Web also shares phenomena of complex, dynamic physical systems. Recent studies have explored system theories based on Web–wide regularities of structure, growth dynamics, and patterns of authorship and reading. These studies show the Web having an overall "power law" feature, where size of sites, number of incoming,
and number of outgoing links all follow a non-modal distribution \((i.e., \) no average behavior). These laws describe a world where a few \(X\) have high \(Y\) following a rapidly descending curve with most \(X\) having low \(Y\) (\(X\) is a number of sites or pages, \(Y\) is a number of links or measure of traffic). A few sites receive a disproportionate amount of traffic and links, while most are effectively islands. These phenomena are further associated with "winner take all" market forces where the more links a site has, the larger its share as more pages join the Web. However, "winner take all" applies to a lesser extent when a topic is more uniformly interesting or a community is more competitive. Usability expertise contends no adverse effect on controversial topics from this power law structure because:

\(\textbf{a})\) Topics rarely have overlapping top sites;

\(\textbf{b})\) even if the top sites accrue the bulk of the links, normal searchers will ask more specialized questions that surface smaller sites; and,

\(\textbf{c})\) cheap advertising greatly diversifies the Web by exposing new or smaller sites.

The "search advice" literature bases key rules on this knowledge about Web behavior: Use multiple engines (because no engine indexes even half the Web, each has biases), use specialized search engines (much Web content is invisible, hard to index), authenticate and evaluate the quality of Web sites (many sites are only for self-promotion), don’t depend solely on Web searching (fee-based services and traditional library materials may be better catalogued with higher quality). A
disturbing analysis of critical thinking deficiencies of college students shows specific examples of potentially harmful Web content bias tricking unwary searchers. Furthermore, it’s not reliable to issue retractions or corrections if the modifications are not hyperlinked or reported with the original erroneous articles. Thus, an understanding of controversial topics on the Web must address three interlocking communities:

(1) Web page authors who bestow links on other pages;

(2) search engines that partially rely on links for crawling and indexing pages and their ranking of search results; and,

3) Searchers (humans) who query for topics and then make browsing selections based on results that serve both engines and authors. Certainly, a search engine could have corporate political biases of its own, e.g. by not crawling designated Web sites or by eliminating pages with particular keywords. However, a search engine basically reflects the biases of Web page authors, and their institutions and objectives. Surfers and searchers reinforce both authors and engines.

**8.9 Why does visibility of controversy matter?**

Highly polarized controversial topics are usually clearly observable in Web contexts. They often bear the identity of well–known organizations and evoke immediate emotional responses in informed readers. Of course, there are proprietary databases for legal research, e.g. Westlaw, that contain exclusively controversial matters and disputes. Another type of controversy exists as a murky subtopic of a broader topic. It may be a
different set of facts about a historical episode, or a strongly opposing point of view on a generally accepted practice, or an emerging change in social standards. In Web terms, the controversy may appear as a subweb loosely connected to a topic’s main theme or by pages widely distributed across sites, not inter–linked, but sharing controversy–related keywords. Controversies are easy to miss unless searchers read closely or are actively seeking opposing views. Why might we care whether controversy is omitted from search results or is missed in an initial topic search?

- Controversies often express the richness and depth of a topic. Consider one of our case studies, Albert Einstein. If all top search results are overly bland or redundant biographies, then a person curious about Einstein’s life and work could well miss richer Web content regarding his personal and political life.

- Controversies dramatize change. Einstein’s public persona and scientific work might seem well settled, but the past two decades have opened up new facets of his career and personal life. For feminists, Einstein’s first wife is a great target of study regarding marriages and scientific collaborations in the early twentieth century. Several Web sites also acclaim her significance as a Serbian woman scientist, reminding us of recent international political changes.
• Controversies may make a critical difference in life–altering decisions. Medical treatments and educational choices require extensive personal research, more than the first superficial search. If controversies are buried in search results or require several links of surfing, then the searcher faces increased risk of missing critical information. Furthermore, an initial search may deter further search by portraying an overly bland and boring picture of the topic (e.g. 27 biographies of Einstein) or a misperception of accepted practice (see Distance Learning and St. John’s Wort case studies).

• Scientists, journalists, and intelligence analysts are professionally required to address multiple perspectives, facts, authorities, and opinions on topics. Search engines may significantly decrease their productivity or conceal incompetence if controversies are overly difficult to investigate.

The dilemma with controversies is that the searcher often doesn’t know what to query for on a broad topic. Asking "(Topic) AND controversy" works some times, but "controversy" has many synonyms. A searcher often doesn’t know the keywords, proper nouns, real life organizations, or link paths to expose a controversial side of the Web. And, there’s a difficult psychological step to "look for trouble" or "peek into the dark side," especially if search results provide an initial positive, or even boring, impression.
Case studies of controversial topics

In order to analyze controversy, we must formulate a more precise question, develop an experimental approach to collect and analyze data, then interpret the results. For example, one study [5] presents a mathematical model of bias, collects search engine query data, and offers an interactive Web site for exploring search engine bias. Our starting assumption is: *A controversial subtopic is revealed or suppressed to the degree its URLs are recognizable in the query for the broad topic.* This paper uses the hypothesis a given, well–known specific controversy will *not* be revealed in the top search results. *A priori,* we would expect that among the 1000s of URLs on a topic, few subtopics will be exposed in top search results. However, using this hypothesis, we might find some surprising refutations where, indeed, controversial subtopics are well represented in top search results. Or we might find absolutely no evidence of controversy in top search results, suggesting systematic bias. Or we might (and did) find a mix for different kinds of topics and subtopics. Each outcome leads us to ask about factors suppressing or revealing the particular controversy, which provides the main contribution of this chapter. The outline of each topic experiment is:

1. Select a broad topic and define the simple, perhaps naïve, query a searcher might ask to start learning about the subject matter.
2. Collect the top 50 URLs from each of three popular modern search engines and the top 100 from two meta–searchers (send the query to multiple engines and collate the results).
3. Identify a controversial subtopic, then define a query using specific keywords for the controversy, and query the same engines in the same amounts.

4. Browse the search results to identify pages revealing the controversy, especially those deeply into the controversy itself.

5. Select a second, more factual, somewhat related subtopic with specific keywords, but (to reduce experimenter exhaustion) simply assume all URLs are relevant to the topic.

6. Compare the results of the subtopics to the simple query and to each other to identify overlapping URLs revealing the underlying controversy.

As we browsed Web pages, we made three levels of judgment:

- Deep: This page digs into the details of the topic. If you (an information seeker) ran across the page within search results or while surfing, you would definitely recognize the existence of controversy, which this page explains in some detail.

- Revealing: This page has links to pages of the controversy or short discussions in passing. However, you might well miss the existence of controversy, in contrast to deep pages definitely revealing the controversy.

- Other: Usually very informative pages, but not mentioning or linking into the controversy.
Figure 1 pictorially shows the various subtopics, with images denoting the three categories of Web pages relative to a controversy. The areas we’re particularly interested in are those overlapping the controversy subtopic with the main theme (results of the simple query) and the factual query. Each area represents the results of queries in the combined results from all search engines. More details of the experimental methodology are discussed in Appendix A. The reader of this paper should think of these experiments as prolonged browsing sessions aided by a URL calculator, note-taking, and an iterative process. Appendix B provides a brief list of revealing URLs as examples, tables and graphs for the counts and computed percentages of URLs in simple, controversial, and factual queries (merged from all engines). The interested reader is invited to perform a search in their favorite engine for the simple query to get a feeling of the type of results expected (see Appendix A.2). However, search engine results are always changing and query settings may give different results from ours. **Figure:8.1 Topics and subtopics**
Table 8.2 provides a listing of the broad topics and identifies the selected controversial subtopic. Table 8.3 shows the factual subtopics and representative other subtopics, along with sample URLs from the 200+ URLs in each query group. Following are brief descriptions of each topic and subtopic for context with brief summaries of results and preliminary explanations. **Table 8.2 Broad topics and controversial subtopics (queries are vitalized).**

<table>
<thead>
<tr>
<th>Broad Topic</th>
<th>Controversial Subtopic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance Learning</td>
<td>&quot;Digital Diploma Mills&quot;, the trend toward commercialization of education, as characterized by technology historian <strong>David Noble.</strong></td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>Did Einstein’s first wife, <strong>Mileva Maric</strong>, receive appropriate credit for scientific contributions to Einstein’s early work?</td>
</tr>
<tr>
<td>Female Astronauts</td>
<td>Did the U.S. space program discriminate against the &quot;<strong>Mercury 13</strong>&quot; women pilots who passed preliminary astronaut screening tests?</td>
</tr>
<tr>
<td>St. John’s Wort</td>
<td>Does this popular herbal remedy work effectively for depression and mood improvement? Recent medical trials differ in their results.</td>
</tr>
<tr>
<td>Belize</td>
<td>This small Central American country has a long, and ongoing, <strong>border dispute with Guatemala</strong> with deep historical roots in Spanish and British colonialism.</td>
</tr>
</tbody>
</table>
Table 8.3 : Broad topics and other subtopics (Sample URLs).

<table>
<thead>
<tr>
<th>Broad Topic</th>
<th>Controversial</th>
<th>Related Factual</th>
<th>Other subtopics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance Learning</td>
<td>&quot;Digital Diploma Mills&quot;, David Noble (Sample).</td>
<td>Formative Evaluation (Sample).</td>
<td>accreditation, faculty careers, intellectual property, Web–based training, home schooling, interactive media</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>Einstein’s first wife, Mileva Maric (Sample).</td>
<td>Einstein’s pacifism and political activism (Sample).</td>
<td>religion, humanism, quotations, correspondence, relativity theory, cosmology, philosophy</td>
</tr>
<tr>
<td>Female Astronauts</td>
<td>&quot;Mercury 13&quot; women pilots (Sample).</td>
<td>Astronaut selection criteria and procedures (Sample).</td>
<td>space shuttle, space station, Columbia, Challenger, Mir, mission control, astronaut training, astronaut qualification</td>
</tr>
<tr>
<td>St. John’s Wort</td>
<td>Effectiveness (Sample).</td>
<td>Dosage recommendations (Sample).</td>
<td>side effects, botany, depression, psychotherapy,</td>
</tr>
</tbody>
</table>
"Distance learning" (DL) refers to modes of instruction which are not same–time/same–place. The topic blends technology with instructional practice with the educational enterprise. The Web reflects these dimensions describing "how Web–based technology works and is used in DL," "how instructional practice benefits from technology," and "what (our) university/company’s DL program offers." In 1997, a technology historian, David Noble, challenged the distance learning movement through a series of widely disseminated articles, inciting numerous debates. Faculty were concerned by loss of control over their intellectual products, as well as contact with students. Active commercialization of DL was causing visible strife at certain universities among faculty, administration, and external interests. Noble coined the term "digital diploma mills" to refer to the new breed of university programs. Our experiment found only one revealing page in the simple "distance learning" query: An annotated webliography by a library. The top search results are primarily organizational: DL trade associations, universities
offering DL programs, as well as several explanatory pages on the nature of DL. Our second subtopic is "formative evaluation," the standard terminology for educational methodology evaluation of technologies, materials, and experiences with them. We wondered if effectiveness and ongoing evaluation influenced David Noble or vice versa. However, the simple, controversy, and factual queries are nearly disjoint. This topic might be described as suppressing controversy by organizational clout. The David Noble/Digital Diploma Mill subweb is linked around Noble’s writings but perhaps its participating organizations are of less Web status (links, size) than the organizational ones. Readers interested in this particular controversy are also more likely to be inside the academic enterprise, rather than consumers of DL products. The DL "search engine personality" is to "help you find the DL program right for you," rather than "provide pro and con arguments for going the DL route." An information seeker might well assume distance learning is a fully accepted and well–paved alternative cheaper education, when issues of quality, accreditation, and durability of degrees are still unresolved. A few pages address other perceived negative aspects of DL, such as pure degree mills, faculty workload, rich vs. poor, and intellectual property.

Suppressed Topic: Albert Einstein/Mileva Maric

Einstein’s biographies discuss not only his role within normal science but also his life as a political activist during the World War II era. Because Einstein received Nobel recognition in 1921 and then settled at Princeton’s Institute for Advanced Study, his opinions were widely sought and his pithy quotations broadly disseminated. In the past two
decades, other facets of Einstein’s personal life have emerged, notably the extent to which his first wife, Mileva Maric, might have contributed rather more than recognized in co–authorship of his technical work. Numerous family related facts appear frequently in biographies, including an illegitimate daughter lost to history, the failure of Maric’s scientific career, their messy divorce (her settlement included his Nobel prize money), and Einstein’s estrangement from his children. Compounding a benchmark of genius, *Time’s* "Person of the Century" and widely quotable writings, there’s no shortage of pages and Web sites about Albert Einstein. We focus on the controversy: Did Mileva Maric provide significant scientific support for Einstein’s earlier work? Web authors reference books and articles about "our work," co–authorship listing on submitted papers, indications of Maric’s superior math skills as well as her physics training. However, there is no definitive evidence one way or the other. This controversy was mentioned once in passing in a page from the simple "Albert Einstein" query. Maric is mentioned as student, sometimes colleague, but mostly unhappy wife and mother abandoned for a second wife and family. Our factual topic is Einstein’s pacifism, as expressed in this public writing and speaking. Pacifism, shows a greater number of its revealing URLs in the simple query. Some Maric URLs are revealed in the pacifism query – primarily in pages with more substantive biographies or personality analyses. One explanation for the absence of controversial content is simple: Albert Einstein is a prime target for term papers, an obligatory biography for science sites, and a fertile source of quotations. These top Web pages are mostly bland, featuring the photogenic Einstein. Such pages essentially squeeze out not only
controversy, but any deeper discussions of Einstein’s life and work. The Web material about Einstein is somewhat organizational, with a few museums and affinity groups, plus the Einstein Page and a few other hubs. Many Web pages simply represent organizations named after Einstein. Many pages contain little more Einstein content than a quotation, relevant or not. The Maric Web has the unfortunate characteristic of lacking a central site for the various pages to link to. Using our criteria that a subtopic appear in the top results for the simple query, the Mileva Maric Einstein controversy is suppressed by sheer volume of quotations, biographies, and passing mentions.

**Suppressed Topic: Belize/Guatemala Border Dispute**

Belize is a recently independent Central American country known for its barrier reef, jungles, and Mayan ruins. A former British colony, Belize has been caught in a long-standing (since 1821) territorial dispute with Guatemala. Recent international mediation has reduced the level of dispute to "confidence building" negotiations and free trade agreements. The settlement of this dispute remains a major current political issue and a significant one in Belizean history. We chose this controversy because it relates to certain aspects of tourism (visiting the border area) and provides a major contrast with the North American experience. Our factual subject is "narcotics," a recent problem for Belize as its location provides a transfer point from South to North America. Both topics delve into how a small country with few military resources deals with international issues. The experiment found about seven percent of the simple "Belize" queries controversy–revealing, typically pages like the
CIA or an encyclopedia fact sheet. One search engine surfaced a Belizean government Web site on Guatemalan relations and the dispute. Three percent of the Belize narcotics controversy were revealing, mentioned on pages about modern Belizean politics and country facts. Other controversies/disputes include dam placement, logging, elections, and human rights. Our experiments suggest this topic as another controversy suppressed by organizational clout. A simple "Belize" query returns mostly tourism information and some general country information, with relatively little depth in historical, political, or current events. Major Belizean newspapers are online but much of the political and cultural material comes from either U.S. or British sources. As a one-time tourist (mainly for the reefs and ruins, in the 1990 pre-Web era), this author is now disappointed not to have had this cultural and historical background. Trading off a few tourism pages for more history would have presented a better rounded portrayal of "Belize" as a starting query.

8.10 The Organizational Web dominates the Analytic Web:

Two broad communities of interest on the Web might be grouped as "organizational" and "analytic." Organizational communities are actual or virtual: Companies, universities, trade associations, consortia, alliances, and government agencies. Distance learning purveyors, science reference sites, NASA, medical advisors, and tourist associations have off-Web organizations, resources, and motivation to be big Web players. The "analytic" Web consists of online full-text (or tables of contents) journals, technical reports and preprints, opinion pages, bibliographies, and pages of links to these. Analytic literature is widely available on the
Web, but is not as extensively interlinked nor represented by large dominant Web sites. The Organizational Web tends to link to key Web sites to an extent search engine ranking strategies naturally assume these organizations are what searchers are looking for first and foremost. And they’re usually right. Controversial subtopics are widely distributed, poorly defined, and weakly organized. Mercury 13 illustrates a cohesive controversy, with book Web pages, online stories by protagonists, and broad reach into news Web sites. In contrast, Mileva Maric shows how a few pages without a core Web site or organizing group does not change the primary perspective of its topic. The Distance Learning controversy is represented by many long articles distributed in some limited-interest publications without continuing interaction with DL purveyors.

**Toward a more objective Web**

Knowing that pages with controversial content may be found sometimes, but not consistently, suggests alternative strategies toward more prominence of controversial and analytic pages. Web authors might adopt more rigorous linking practices to

(a) reach a higher standard of objectivity and/or

(b) exert greater influence on search engine ranking.

Academic researchers learn early in their careers to pass stringent peer reviewing by addressing not only the data favoring their approach but also opposing views, contradictions, and unknowns. If this practice was adopted by Web content creators, how would Web linking change?
Search engine optimization experts now practice link exchange in recognition that links have economic value. I simulate a more objective scenario later in this chapter. Many factors work against this change in linking practice. Most organizations invest heavily to promote themselves and certainly won’t link to competitors. This is characteristic of the Organizational Web, but now consider the Analytic Web. Many safe sites (e.g., Yahoo or Google directories) are common hubs for links, leaving the information seeker to sort out the link’s actual target. Well-established authorities, often traditional real-life organizations, are other safe and useful targets. It takes some courage to link to a lesser known or "lone genius" site. Linking to a less stable site with dead links or unpredictably changing pages is risky, even if its content is well-written and provocative. As in writing a thorough research paper, considerable extra thought must go into creating accurate and balanced references (links). We’re still looking for the opposing links for this paper’s novel topic. However, linking only to popular, but possibly bland or less relevant, sites reinforces their status in search rankings over more specialized sites. Search engines might alter their ranking strategies to provide more openness to the "Analytic Web" (controversy + data + evaluation + detail). Search engine ranking strategies remain proprietary, often mysterious, in flux, and a challenge to searchers. It would be unreasonable, and probably undesirable, to urge search engine companies to adopt a more objective standard, e.g., 25 percent of rankings include controversies, 25 percent traditional organization Web sites, and 50 percent objective pages, even if technically feasible. However, ranking might include more predictable and recognizable distribution of the
organizational (institutions vs. individuals) or analytic (glossy vs. technical) subwebs, e.g. Teoma’s "Link Collections from Experts and Enthusiasts." Search trainers and professional searchers might alter their search practices. This paper shows how five different topics only weakly present controversies at the naïve query level. Advice to query specifically for controversy belongs with advice to be wary of deceptive Web pages. Librarians have long advised the Web is often not the best source of analytic information. Search pros recognize Web content must be authenticated and checked from different sources before accepted. Mistrusting search engine results falls into this category. of advice, as do countermeasures for balancing organizational and analytic Web content. Engines, authors, and searchers might advance toward alternative paradigms for working the Web. Today’s metaphor is "search." Another metaphor is "Collect", emphasizing broader search, filtering, and ranking to produce well-rounded, multiple-use collections of URLs rather than a single-answer URL. Another paradigm is explicit markup on concepts driven by ontologies (concept specifications) in the Semantic Web.

8.11 A simulated objective Web

Here’s how the current Web appears. Figure 2 shows the number of in links for sites of the "distance learning" data graphed against the number of sites with this number of in links. Note the area near the Y axis with few sites having large numbers of in links, leading to a long tail of the curve with most sites having 0 or 1 in links. Figure 3 shows a similar curve for pages of "Albert Einstein." These curves apply for links among
pages in the respective collections not the data about the Web as a whole, but sharing the same distributions as fractals.

Figure 8.2: Power curve of Inlinks to Sites of "Distance Learning"

Figure 8.3: Power curve of Inlinks to URLs for "Albert Einstein"

Let’s idealize a situation where many pages are objective; examples are starred in Appendix A.2. A purely objective page might hold a carefully selected list of links or provide multiple sides of a topic. Of course, most pages would be pro or con some issue, usually favoring their organization’s position. Now, suppose we "require" all Web authors identify and link to opposing and contrasting views, as well as relevant
objective pages. Examples of this linking strategy include: Distance Learning pages linking to pages refuting its effectiveness, experimental results, as well as the negative business side of degrees; Einstein biographies linking to Maric biographies and providing more depth on the controversy; Female Astronaut lists also showing the Mercury 13 pilots (many do), plus pages analyzing NASA’s acceptance of women; alternative medicine stores linking to NIH advisories, and revisions to knowledge as clinical trials advance; and, Belize histories linking more to the Guatemala dispute coverage from both sides of the border. Below is the resulting graph when 10 percent of 500 URLs are objective and Web pages distribute their links 50 percent to supporting, 25 percent to opposing views, and 25 percent to objective pages. Numbers of links and links among pages are randomly generated, constraining pages to either both agree or both disagree. This new graph, Figure 4, shows a redistributed power curve with many links to the objective pages (near the Y axis) and a more even distribution of the number of in links with more pages getting in links. Figure 5 shows more evenness in distribution of inlinks with a higher percentage of objective pages. Of course, this simulated Objective Web differs from Figures 1 and 2 not only in link distribution policy but also in number of links.
Figure 8.4: Simulation of an inlinks in an Objective Web with 500 pages, 10 percent objective, and 50:25:25 Agreeing:Opposing:Objective Distribution

Figure 8.5: Simulation of an Objective Web with 500 pages, 30 percent objective, and 30:30:40 Agreeing:Opposing:Objective Distribution
The results are not surprising but visually emphatic, contrasting today’s Web with a hypothetical Web having a different standard of linking behavior. This model confirms that a more objective policy would exhibit an extreme change in some Web characteristics. Further effects would then appear in search engines ranking highly by inlinks, shifting preference to objective pages. The requirement to link more extensively and non–exclusively to agreeing or popular pages would distribute links more evenly across all pages.

8.12 Appendix A: Experimental methodology details

This appendix describes the details of search engine queries (engines used, specific search terms), an analysis of the limits of the experimental methodology, and a summary of the twURL URL analyzer used in the experiments.

Appendix A.1: Search engine details

Search engines used

- Google
- Teoma
- AllTheWeb (FAST)

and two multi–searchers

- Web–based Profusion (Altavista, About, AOL, Lycos, Raging Search, Wisenut, Metacrawler, MSN, Adobe PDF, Looksmart, Netscape, Teoma, AllTheWeb)

Queries used:

Simple: distance learning, Albert Einstein, female astronauts, St. John’s Wort, Belize Controversy: distance learning and David Noble, distance learning and digital diploma mills; Mileva Maric, Albert Einstein and Mileva; Mercury 13, Jerrie Cobb; St. John’s Wort and (Glaxo, Lichter Pharma, Pfizer, NCCAM, ASP-ASIM, Nutrition Alert); Belize and Guatemala dispute Factual: distance learning and formative evaluation; Albert Einstein and pacifism; astronaut selection; St. John’s Wort and dosage; Belize and narcotics Queries posed as Advanced Search, using phrase (vs. all words) where possible. One hundred URLs drawn from the results, then reduced to the top 50. For multi–searchers, top 30 drawn from each engine for Copernic, all for Profusion, as far as the multi–searchers would go, then the top 100 used.

Appendix A.2: Limits of the experiments

• Results are based on the naïve, simple query and a single controversial topic. No established experimental methodology for testing bias in search engines yet exists. Our approach — comparing a simple query with a controversial one — provides a baseline experimental
hypothesis: A specific controversy, though well–known, would not be significantly represented in top search results. This hypothesis proved false for two topics and true for three. The experiments also generated questions and explanations for some behavior of search engines, Web authors, and information seekers.

- The experiments were complicated by messy duplicates, dead links, ads, paid placement, and irrelevant pages. The same Web content often appears under many different variations of URLs or mirrored pages. We considered the pages different until the end analysis where duplicates effectively merged into single URLs. Dead links, paid placement, and browsed pages with no apparent relevance were eliminated. Analysis focused on identifying Deep Web pages first, and those Revealing next, with a goal of missing a few Revealing, but no Deep.

- Identifying Deep and Revealing URLs required subjectivity. We viewed most of the nearly 3,500 URLs in Internet Explorer, using a stream of URLs fed by twURL. Growing familiarity with a topic clarified URLs deep into the controversy (you "couldn’t miss the controversy in the title and page scan"). The Revealing category took more work to find pages with links to or passing references in larger articles. An independent review would likely show some differences of opinion on specific URLs.
The browsing process was incomplete. We used a controlled vocabulary for each topic to identify possibly Deep or Revealing pages. Our keywords applied only to snippets from pages, thus sometimes missing relevant pages. Although these pages were already filtered by the indexing algorithms of search engines, we sometimes found relevant pages which search engines had missed or ranked too low to appear in our results. We also followed links to or from URLs found to be Deep, prioritizing browsing to the deeper controversy, then the revealing, and finally the remaining pages.

The search results are biased by the specific queries used. Are the broad topics using the correct keywords, e.g. "women astronauts" instead of "female astronauts" or "distance education" instead of "distance learning"? Variant queries do give different results, and might lead to somewhat different conclusions. For example, there may be a greater correlation in the terminology of Web contributors between "David Noble" and "distance education" (versus "distance learning"). Since some search engines are also sensitive to plurals, we queried where we thought necessary for both singular and plural of "female astronaut" and "digital diploma mill". Furthermore, advanced searches sometimes twist combinations of "must include" into obscure Boolean expressions. The unpredictability of ranking algorithms together with the imprecise definition of "advanced search"
results is an underlying dilemma: How would one validate a query to assure it’s posed correctly and the search engine delivered its promised results?

- The experiments are time–sensitive. Any query results will vary from one instance to the next and probably more over longer periods of time. Books on "Mercury 13" possibly accelerated the "female astronauts" controversy; these books were published in mid–2003 and this experiment was performed in August 2003, although related articles from other sources go back several years. Forty– and twenty–year panniers of "first female astronauts" also may have influenced the results. Other topic search results will likely show declining or increasing amount of controversy as Web authors write more or less and information seekers interest grows or declines. Search engines will index new pages and eliminate other pages from their indexes. Search engine ranking algorithms will change and the relative rankings of sites will also vary. The same experiments repeated six months from now will likely show some differences.

- The experiments are sensitive to search engine choices. Indeed, more engines could have been chosen, but our experience indicated three major engines plus two multi–searchers (covering other engines) would suffice. Indeed, it is surprising how much the chosen engines are in agreement and how consistent their overlap with other engines.
In summary, we have an outline of an experimental methodology no more valid nor reliable than its underlying technologies. A valid experiment is not biased away from producing the data needed. These experiments addressed bias by targeting recognizably deeper, then revealing, then other Web pages. A reliable experiment would produce the same results for all samples taken and with different experimenters. This should be addressed simply by replicated and more exhaustive experiments. Thus, the results of these experiments should not be overly interpreted until the experimental methodology is refined and variations performed to address the above problems.

**Appendix A.3: Summary of twURL**

twURL is a Windows desktop tool for analyzing URL collections. URLs are extracted from saved search engine pages (Google, Teoma, AlltheWeb, and Profusion) or exported files (Copernic) and merged into a URL base. Views of the URLs then include: Internet domains, number of links, number of sources, ranking by query and engine, keywords from a topic–specific controlled vocabulary. Once organized into Views, URLs are browsed in order within sub views, *e.g.* all URLs with keyword "Milevav or URLs with most links in. Browsing consists of automatically loading the Web page into Microsoft Internet Explorer in the order given in the current View. As URLs are browsed decisions may be made and recorded to identify Deep, Relevant, and Other pages. The data was imported into Microsoft Excel for further analysis, *e.g.* decisions per engine, overlaps among engines, decisions per query, etc. HTML reports
of the entire collection and the controversy pages are available at http://www.twurl.com/Controversy/Data/.

Appendix B: Results of experiments

This appendix provides the numerical counts and percentages of deep and revealing URLs for each topic, a graphical comparison of these results, samples of URLs showing characteristics of controversial and factual pages, and a comparison of search engine overlaps and effectiveness.

Appendix B.1: Query results

This table shows the results of queries from all engines. Revealing Web pages suggest the existence of controversy, among other topics. Deep Web pages provide details in full force regarding the controversy. A searcher might well miss a Revealing URL description but would definitely be presented the controversy in a Deep Web page.

Table 8.4 The results of queries from all engines

<table>
<thead>
<tr>
<th></th>
<th>number of URLs</th>
<th>Revealing</th>
<th>Deep</th>
<th>Percentage Deep</th>
<th>Percentage Deep or Revealing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance Learning (totals)</td>
<td>661</td>
<td>152</td>
<td>100</td>
<td>15</td>
<td>38</td>
</tr>
<tr>
<td>controversy</td>
<td>281</td>
<td>147</td>
<td>100</td>
<td>36</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
<td>Count</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>david noble</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>factual</strong></td>
<td>179</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>&lt;1</td>
</tr>
<tr>
<td><strong>formative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>evaluation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>simple</strong></td>
<td>212</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>&lt;1</td>
</tr>
<tr>
<td><strong>distance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>759</td>
<td>110</td>
<td>4</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>(totals)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>controversy</strong></td>
<td>504</td>
<td>91</td>
<td>58</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>Mileva Maric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>factual</strong></td>
<td>180</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>Albert Einstein Pacifist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>simple</strong></td>
<td>171</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belize (totals)</td>
<td>443</td>
<td>89</td>
<td>22</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td><strong>controversy</strong></td>
<td>144</td>
<td>84</td>
<td>23</td>
<td>16</td>
<td>74</td>
</tr>
<tr>
<td>Belize Guatemala dispute</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>factual</strong></td>
<td>167</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Belize Narcotics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>simple</strong></td>
<td>169</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------</td>
<td>-------</td>
<td>-----</td>
<td>---</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Female Astronauts (totals)</td>
<td>663</td>
<td>85</td>
<td>90</td>
<td>14</td>
<td>26</td>
</tr>
<tr>
<td>controversy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mercury 13, Jerrie Cobb</td>
<td>210</td>
<td>66</td>
<td>88</td>
<td>41</td>
<td>76</td>
</tr>
<tr>
<td>factual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>astronaut selection</td>
<td>167</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>simple female astronauts</td>
<td>349</td>
<td>25</td>
<td>32</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>St. John’s Wort (totals)</td>
<td>948</td>
<td>106</td>
<td>145</td>
<td>15</td>
<td>27</td>
</tr>
<tr>
<td>controversy St. John’s Wort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>effectiveness</td>
<td>648</td>
<td>87</td>
<td>135</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>factual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. John’s Wort dosage</td>
<td>221</td>
<td>16</td>
<td>19</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>simple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St. John’s Wort</td>
<td>267</td>
<td>48</td>
<td>29</td>
<td>11</td>
<td>29</td>
</tr>
</tbody>
</table>
Figure 8.6  Comparison of Revealing and Deep (Controversy. Broad Topic)

Appendix B.3: Samples of most Revealing sites/URLs
Distance Learning

Controversy "David Noble", "digital diploma mills"

- "Digital Diploma Mills: The Automation of Higher Education"
  http://www.firstmonday.dk/issues/issue3_1/noble; "Digital Diploma Mills: A Dissenting Voice"
  http://www.firstmonday.dk/issues/issue4_7/white; "Brave New Universities" by Michael Margolis
  http://www.firstmonday.dk/issues/issue3_5/margolis
- Distance Education http://communication.ucsd.edu/dl;
  Untitled
  http://www.rohan.sdsu.edu/faculty/feenberg/TELE3.HTM
  (deep URL below ranking cut-off)
- Digital Diplomas
  http://www.motherjones.com/mother_jones/JF01/diplomas.html
- Review: Digital Diploma Mills
  http://mtprof.msun.edu/Win2003/lockrev.html
- The Dark Side of ALN
  http://eies.njit.edu/~turoff/Papers/darkaln.html (pre-dates David Noble, similar theme)
• ***Library Support for Distance Learning
  http://www.lis.uiuc.edu/~b-sloan/libdist.htm (Also in factual and simple queries)

  **Factual "formative evaluation"**

  • Archived — The Integrated Studies of Educational Technology — ED Office of Educational Technology
    http://www.ed.gov/Technology/iset.html
  • theNode.org: Ellen Herbeson — a formative evaluation of distance education: experiences of learners
    http://node.on.ca/tfl/notes/herbeson.html

**Albert Einstein**

**Controversy "Mileva Maric"**

  • ***Mileva Maric — Wikipedia
    http://www.wikipedia.org/wiki/Mileva_Maric
  • Student Searches for Albert Einstein’s Wife and Daughter
    http://www.mtholyoke.edu/offices/comm/csj/961122/ein.html
  • Albert Einstein/Mileva Maric — Love Letters
    http://www.semcoop.com/detail/0691088861
  • Mileva Maric http://www.teslasociety.com/Mileva.htm
  • Mileva Maric Einstein — ‘Out From The Shadows Of Great Men’ http://www.compuserb.com/mileva02.htm
Factual "Pacifism"

- Albert Einstein — Wikipedia
  http://www.wikipedia.org/wiki/Albert_Einstein

Female Astronauts

Controversy "Mercury 13"

- Mercury 13 — the Women of the Mercury Era
- CNN — First female astronaut still hoping to go up —
  October 28, 1998
  naut
- Military Women Astronauts
  http://userpages.aug.com/captbarb/astroaurots.html

Factual "Astronaut Selection"

- Women in Space — Female Astronauts and Cosmonauts
  http://womenshistory.about.com/cs/aviationspace

St. John’s Wort
Controversy "effectiveness"

- St. John’s Wort (*Hypericum perforatum*) and the Treatment of Depression http://nccam.nih.gov/health/stjohnswort
- ***ACP–ASIM Pressroom http://www.acponline.org/college/pressroom/antidep.htm

Factual "Dosage"


Belize

Controversy "Belize Guatemala Dispute"
- ***Anglo–Belize/Guatemala Territorial Issue Title Page
  http://www.belizenet.com/bzguat.html
- Belize — Consular Information Sheet
  http://travel.state.gov/belize.html
- ***CIA — World Fact Book 2002 — Belize

_Factual "Belize narcotics"

- Guardian Unlimited | The Guardian | Why the secret world of Belize business alarmed US drug agents
  http://www.guardian.co.uk/uk_news/story/0,3604,500941,00.html

**Appendix B.4: Search engine characteristics for all URLs (Simple, Controversy, Factual)**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Percentage Deep (Range)</th>
<th>Percentage Revealing or Deep (Range)</th>
<th>Revealing or Deep Low/High Engine</th>
<th>Percentage Overlapping search/multi–search (Range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance Learning</td>
<td>17–21</td>
<td>33–49</td>
<td>google/teoma</td>
<td>35–46/62–74</td>
</tr>
<tr>
<td>Location</td>
<td>Cited</td>
<td>Related</td>
<td>Search Engine(s)</td>
<td>HTML</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>---------</td>
<td>------------------</td>
<td>------</td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>6–11</td>
<td>14–21</td>
<td>alltheweb/google</td>
<td>26–33/49–82</td>
</tr>
<tr>
<td>Female Astronauts</td>
<td>16–24</td>
<td>29–37</td>
<td>copernic/google</td>
<td>27–34/56–84</td>
</tr>
<tr>
<td>St. John’s Wort</td>
<td>15–22</td>
<td>25–35</td>
<td>alltheweb/teoma</td>
<td>27–34/66–95</td>
</tr>
<tr>
<td>Belize</td>
<td>4–8</td>
<td>20–32</td>
<td>alltheweb/google, teoma</td>
<td>232–35/66–855</td>
</tr>
</tbody>
</table>

kept=on-topic and not 404, percentage deep = #deep/#kept, percentage revealing or deep = (#deep + # revealing)/#kept, overlapping = #overlapping second engine/#kept
Chapter Nine:

Summaries, Conclusion and Recommendations
Chapter Nine :

Summaries, Conclusion and Recommendations

9.1 Summaries:

The focus of this study is the translatability of English social science indexing terms into Arabic language and culture at a conceptual, term and indexing term level. The perspectives are both linguistic and sociological - a combination through which a broader understanding of the phenomena is being aimed at in the general frame of information science. This study is quantitative-qualitative. Qualitatively is emphasized in developing the analysis method for this specific research material. Traditional terminology sources will be supplemented by the theories and concepts of translation studies, linguistics and communication theories in order to achieve a broader and more pragmatic perspective. Thus, a new research method can be developed. Linguistic and cultural matters cannot be clearly separated, because they go hand in hand. Therefore, in this study concentrating on the cultural or linguistic matters is more like an emphasis or a perspective. The study uses multiple cases aiming at theoretical replication. It is thus an empirical case study and in its nature descriptive, generating hypotheses and illustrating a new theory of "pragmatic indexing equivalence". The samples studied are theoretical (not random) and thus cases assumed to give answers to the questions of this study. Several data collection and analysis methods will be used in order to construct a theory by triangulation of evidence. The research process is thus linear starting from hypotheses (statements) and ending
through observations and generalizations at new hypotheses and a construction of a theory. This study examines Arabic equivalents for British-English indexing terms. The key terms of the research corpus will be studied:

- at a conceptual level by interviewing people representing different cultures (Arabic versus British) and subcultures (indexers, specialists, multilingual thesaurus constructors);
- by a modification of co-word analysis (the use and the indexing practice of the key terms studied in Arabic, British and international databases) and;
- By comparing several thesauri.

9.2 Conclusion and Findings:

Regarding the Linguistic dimension in Information Sciences, the following conclusions are drawn:

* It has been proved that the All engines are tested had ranking schemes that were not well documented, based on how many times your search words were mentioned, whether or not they appeared early in the document, whether or not they appeared close together, and how many search terms were matched. I did not find the ranking schemes very useful, as relevant and irrelevant pages frequently had the same scores.

* It is found that aggregating meaning is possible on the Internet because there are many easily accessible semantic objects to be
harvested. Analysis of the aggregations can suggest patterns of high likelihood that permit applications to recommend, adapt, profile, forecast and so on.

* It has been proved that the terminology and thesaurus construction standards and guidebooks provide very little details and consideration about equivalence.

* It has been indicated that Google may index billions of Web pages, but it will never exhaust the store of meaning of the Web. The reason is that Google's aggregation strategy is only one of many different strategies that could be applied to the semantic objects in public Web space.

* It has been found that web search engines do not conspire to suppress controversy, but their strategies do lead to organizationally dominated search results depriving searchers of a richer experience and, sometimes, of essential decision–making information. These experiments suggest that bias exists, in one form or another, on the Web and should, in turn, force thinking about content on the Web in a more controversial light.

**9.3 Recommendations:**
This part of the study is devoted to provide some recommendations based on the findings of the fieldwork study. The following recommendations are drawn as far as this research is concerned.
• It is found that many languages are in serious danger of being lost and if nothing is done to prevent it, half of the world's approximately 6,500 languages will disappear in the next 100 years. Language data are central to the research of a large social science community, including linguists, anthropologists, archeologists, historians, sociologists, and political scientists interested in the culture of indigenous people. The death of a language entails the loss of a community's traditional culture, for the language is a unique vehicle for its traditions and culture. So in this regard, it is recommended that to preserve endangered languages it must use the state-of-the-art information technologies.

• Besides, it is recommended that it must build a showroom of best practice for digital archives of endangered language data, where the data from ten endangered languages is archived in a way such as to demonstrate the best practice and the best way to design and store material for such an archive.

• Also, it is recommended that it must build a linguistic ontology which would serve as an Interlingua for the various linguistic markups used, so as to allow searching of diverse material.

• Also, it is recommended that it must build (FIELD) which facilitates the work of linguists in storing endangered languages material to conform to best practice.

• It is recommended that it must develop and implement a federated multimedia Web-based digital library that integrates various endangered languages data sources including E-MELD to design
and develop an annotation management system to support domain experts to annotate various data related to endangered languages.

- It is recommended that it must provide support for intelligent querying and retrieval of language data;
- It is recommended that it must provide support for ontology-based multilingual, cross-theoretic querying and retrieval;
- It is recommended that it must provide tools for analysis and cross-linguistic comparison; provide methods for accurate display of linguistic documentations of various media formats for various display devices; and to enhance FIELD to allow for storage of audio and visual materials.
Bibliography


(2) Abbott, Nabia (1941), ’The development of early Islamic scripts’. *Ars Islamica* 8, 65-104.


(33) **Cisco, Susan. 1993.** Indexing Documents for Imaging Systems: A Roadmap to Success. Austin, TX: Marketfinders.

(34) **Cleveland, Donald B., and Cleveland, Ana D. 1983.** Introduction to Indexing and Abstracting. Littleton, CO: Libraries Unlimited.


(36) **Collison, Robert. 1959.** Indexes and Indexing. London: Earnest Bean.


(38) **Craven, Timothy C. 1986.** String Indexing. Orlando, FL: Academic Press.


of the Seventh Message Understanding Conference (MUC-7). Morgan Kaufmann, California.


265


(58) Gaizauskas, R. 1995. Investigations into the grammar underlying the Penn Treebank II. Research Memorandum CS-95-25, Department of Computer Science, University of Sheffield.


(62) Gamma, E., R. Helm, R. Johnson, and J. Vlissides. 1995. Design Patterns - Elements of Reusable Object-Oriented Software. Addison-Wesley.


(75) **Huang, R. (2003).** *ImageSpace: a DAML+OIL based image ontology and annotation tool.* Unpublished master's dissertation, Wayne State University, Detroit, Michigan, USA.


(78) **Jacobs, P.S., editor. 1992.** *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval.* Lawrence Erlbaum, Hillsdale, NJ.


(80) **Kay, M. 1997a.** It's Still the Proper Place. *Machine Translation,* 12:3-23.


(95) applications (CATA'2004), Seattle, WA, USA, March 18-20, 2004.


freiburg.de/~dbis/Publications/2K/TR136-InfoExtr.ps


(130) **Spencer, Harvey. February 1, 2002.** The Recent Evolution Of Data Entry: The Internet,OCR/ICR and Off-Shore Data Entry Can
Dramatically Reduce the Cost of Capturing Data From Existing Forms. IW: 37-39.


