

Constructing An Automatic Lexicon for Arabic Language

Riyad Al-Shalabi Ghassan Kanaan
Yarmouk University - Irbid-Jordan
shalabi@yu.edu.joshalabi ghassank@yu.edu.jo

Abstract: *In this paper, we have designed and implemented a system for building an Automatic Lexicon for the Arabic language. Our Arabic Lexicon contains word specific information. These pieces of information include; morphological information such as the root (stem) of the word, its pattern and its affixes, the part-of-speech tag of the word, which classifies it as a noun, verb or particle; lexical attributes such as gender, number, person, case, definiteness, aspect, and mood are also extracted and stored with the word in the lexicon. A lexicon its a collection of representations for words used by a natural language processor as a source of words specific information; this representation may contain information about the morphology, phonology, syntactic argument structure and semantics of the word. A good lexicon is badly needed for many Natural Language applications such as: parsing, text generation, noun phrase and verb phrase construction and so on. Many rules based on the grammar of the Arabic language were used in our system to identify the part-of-speech tag and the related lexical attributes of the word [13]. We have tested our system using a vowelized and non-vowelized Arabic text documents taken from the holly Qur'an and 242 Arabic abstracts chosen randomly from the proceedings of the Saudi Arabian National Computer Conference, and we achieved an accuracy of about 96%. We discuss the factors behind these errors and how this accuracy rate can be enhanced.*

Keywords: *Lexicon, Stem, Tagger, Vowelized, Pattern, Root, Stopwords, Affixes.*

Received: March 18, 2004 | **Revised:** January 31, 2005 | **Accepted:** February 17, 2005

1. Introduction

Lexicography is the branch of applied linguistics concerned with the design and construction of lexica for practical use. Lexica can range from a paper dictionary or encyclopedia designed for human use and shelf storage to the electronic lexicon used in a variety of human language technology systems, such as word databases, word processors, software for read back by speech synthesis in Text-to-Speech systems and dictation by automatic speech recognition systems. At a more generic level, a lexicon may be a lexicographic knowledge base from which lexica of all these different kinds can be derived automatically. [10] Lexicology [10] is the branch of descriptive linguistics concerned with the linguistic theory and the methodology for describing lexical information, often focusing specifically on issues of meaning. Traditionally, lexicology has been mainly concerned with:

- Lexical collocations and idioms,
- Lexical semantics,
- The structure of word fields and meaning components and relations.

Linguistic theory in the 1990s has gradually been integrating these dimensions of lexical information. Thus lexical information includes; lexical semantics, and the study of the syntactic and morphological and phonological properties of words [10].

Lexicon theory is the study of the universal, in particular, the formal properties of lexica, from the points of view of theoretical linguistics, general knowledge representation languages in artificial intelligence, lexicon construction, access algorithms in computational linguistics, or the cognitive conditions on human lexical abilities in empirical psycholinguistics [10].

Lexical knowledge is the knowledge about individual words in the language. It is essential for all types of natural language processing [12].

A lexicon its a collection of representations for words used by a linguistic processor as a source of words specific information; this representation may contain information about the morphology, phonology, syntactic argument structure and semantics of the word [13].

An important question is how to store lexical information. The format should be standardized, many

programs need a lexicon to accomplish their tasks and many people build this lexicon manually [13].

Lexicon theorists have increasingly made use of extensive lexicological and lexicographic descriptions as models for testing their theories, and lexicographers are increasingly making use of theoretically interesting formalisms such as regular expression calculus in order to drive parsing, tagging and learning algorithms for extracting lexical information from text corpora. Furthermore, the computer has accelerated work in practical lexicography, and has also gradually led to a convergence within these lexical sciences [7, 10, 23].

Lexica are necessary for natural language processing systems such as system for information extraction / retrieval or dialog systems. For some applications, at least, a phrasal lexicon is vitally important [21].

Developers of machine translation systems, which from the beginning have involved large vocabularies, have long recognized the lexicon as a critical system resource [18].

An important critical step towards avoiding duplication of efforts, and consequently towards a more productive course of action for the realization of resources, is to build and make publicly available to the community large-scale lexical resources, with broad coverage and basic types of information, generic enough to be reusable in different application frameworks [18].

One application area where lexica are used is speech technology, particularly for dictation (speech recognition) and readback (text-to-speech) software. The size of the Lexicon needed for such applications has leapt from a few hundred words in the early nineties to tens of thousands today. Software technologies are being developed for generating all word form variants from the stem forms, and for automatically inducing large lexica from text and transcription corpora with statistical and symbolic classification algorithms. The development of lexica for these purposes is a small but growing industry [10].

2. Types of information

The lexicon may contain a wide range of word-specific information, depending on the structure and task of the natural language processing system [18].

A basic lexicon will typically include information about morphology. On the syntactic level, the lexicon will include in particular the complement structure of each word or word sense. A more complex lexicon may also include semantic information, such as a classification hierarchy and selectional patterns or case frames stated in terms of this hierarchy. For machine translation the lexicon will also have a record correspondences between lexical items in the source language and the target language; for speech understanding and generation it will have to include information about the pronunciation of individual words [11, 22].

3. The Model of the Arabic word

A word is defined as an alphanumeric string between any two non-alphanumeric characters. An Arabic word is a word in which all the characters are bare or diacriticized Arabic alphabets characters. It may be either an original Arabic word, or an Arabized word. The original Arabic words are divided in turn into two sub categories:

- **Derived Arabic words:** These are the verbs and nouns that are built according to the Arabic derivation rules. The sweeping majority of Arabic words belong to this category.
- **Fixed Arabic words:** These are a set of words molded by Arabs in ancient times that do not obey the Arabic derivation rules. Most of these fixed words are neither verbs nor nouns, most of them are functional words like pronouns, prepositions, conjunctions, question words, and the like. They may be best regarded as the *glue* that ties the words of the Arabic sentence together [16].

Arabized words are words borrowed from foreign languages (perhaps with some phonetic adjustments to suit the Arabic pronunciation) that have become common among the native Arabic speakers. To preserve the purity of the Arabic language, we prefer to avoid a word in this category unless its meaning has no counterpart among the original Arabic words. Figure 1 [14] summarizes this classification of the Arabic words.

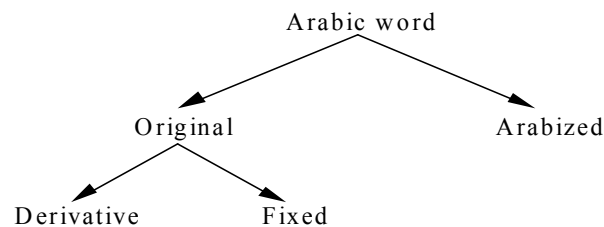


Figure 1. The Classification of Arabic Words

Any treatment of Arabic must treat all of these categories with the same degree of care [15].

4. Arabic is a Diacritized Language

The pronunciation of a word in a *non-diacritized* script is almost always fully determined by its constituent characters, so that the sequence of consonants and vowels determines the correct phonetics. Such a language is Spanish or Finnish [15].

On the other hand, in *diacritized* scripts the pronunciation of the words cannot be fully determined by their constituent characters only, special marks are put above or below the characters to determine the correct pronunciation. These marks are called *diacritics*. In such languages, two different words may have identical spelling whereas their pronunciations and meanings are totally different. Arabic script

involves an elaborate diacritization system. Table 1 shows the Arabic diacritics and the significance of each one [2, 16].

During the process of assigning diacritics we need to determine each two kinds of diacritic information about the character: things:

1. The shadda state of the character. (with /without Shadda.)
2. The diacritic of the character.

Unfortunately, in most Arabic writing today, people do not explicitly include diacritics. They expect their readers to depend on their knowledge of the language and the context to supply the missing diacritics while reading a non-diacritized text. They only mention diacritics in writing when a severe ambiguity is feared or in texts designed for educational purposes. [16].

Table 1: The Arabic Diacritics and the Shadda

Diacritic	Name	Sounds like	Examples	Comments	
(1)	عَ	Fateha فتحَة	a	مَكْفَأَةٌ، مَصْنَعٌ، بِرَاءَةٌ، عِلْمٌ	—
(2)	عُ	Damma ضَمَّة	o	كُتِبَ، هُمُومٌ، صُرَاخٌ، عُودٌ	—
(3)	عِ	Kasra كَسْرَة	e	كِتَابٌ، مِهْنَةٌ، عِيَالٌ، هِمَمٌ	—
(4)	عْ	Sokoon سُكُونٌ	A non vowelized consonant	عَوْنٌ، إِسْنَانٌ، رَأَى، اسْتَعْمَالَ	—
(5)	عَا	Tanween fateha تنوين فتحَة	Fateha + نَ	كِتَابًا، هَيَاةً، طَعَامًا، ثَرَاءً	Only the last character may be assigned this diacritic
(6)	عُ	Tanween damma تنوين ضمة	Damma + نَ	حَصْرٌ، قَصُورٌ، اسْتِعْدَادٌ، سَرْدٌ	Only the last character may be assigned this diacritic
(7)	عِ	Tanween kasra تنوين كسرة	Kasra + نَ	مَسَاءً، مَلَاقَةً، مَعَانٍ، حَمَامٍ	Only the last character may be assigned this diacritic
(8)	ا، و، ي	Vowel مَدَّة	Long (a), (e), or (o) vowel	كَاتِبٌ، مَغَامٌ، قَالٌ، عِيدٌ، طِينٌ، بُيُوتٌ، كُوفَةٌ، رُوحٌ	—
(9)	ى	Alef leyna أَلِفٌ لَيِّنَةٌ	Long (a) vowel	مَصْطَفَى، مَشْتَى، نَادَى، مَغَالَى	Only a terminal ى may be assigned this diacritic.
(10)	ا	Bypassed character حرف غير منطوق	Not pronounced	أَلْسِمَاءٌ، وَالسَّمَاءُ، قَالُوا، أَوْلِيَاكَ	—
(11)	ع	Hidden alef vowel مَدَّةٌ مُسْتَتِرَةٌ بِأَلِفٍ	Long (a)	هَذَا، ذَلِكَ، الرَّحْمَنُ،	—
(12)	عْ	Shadda شَدَّة	مَعْلَمٌ؛ لٌ = لٌ+لٌ كُتِبَ؛ تٌ = تٌ+تٌ حَقٌّ؛ قٌ = قٌ+قٌ الصُّبْحُ؛ صٌ = صٌ+صٌ	In fact, shadda is not a diacritic but is a mark of doubling the character while pronouncing it. The character with a shadda needs another diacritic (from no.1 to no.7) to determine its vowel.	

An Arabic word may appear in any of three diacritization states:

1. **Full diacritization:** This means the assignment of all the diacritic information for each character in the word including the last one. In Arabic, the diacritization of the last character sometimes depends on the syntactic analysis of the word within its sentence [2].
2. **Half diacritization:** This the same as full diacritization except for that it does not include the diacritic mark of the last character if it depends on the syntactic analysis of the word [2].
3. **Partial diacritization:** Any other diacritization state of the word that provides less diacritic information than half diacritization is called partial diacritization [2].

5. The Prefix-Body-Suffix Structure of the Arabic Word

While all languages allow us to express the same ideas using a variety of sounds, they differ a great deal in the ways they provide for stringing concepts together. One scale on which they differ is the “analytical/agglutinative” scale. An agglutinative language allows the speaker to glue multiple morphemes together into a single word; an analytical language divides them into separate words. English is a rather analytical language, French is even more so; Arabic is much more agglutinative, though not so much as modern Finnish or Turkish [2]. Arabic word may correspond to a single entity but can as well be compounded of more than one entity. In fact it may be a phrase or even a complete sentence. So, the Arabic word is in general a complex. If we study a sufficiently large sample of Arabic text, we can infer the following general simple structure for Arabic words:

- a. The main part of a noun or a verb, occurs in the middle. Let us call this part the *body of the word*.
- b. The body may be prefixed by a definitive article, a preposition, a gender determiner, a tense determiner, etc., or some combination of these. When a *prefix* precedes a body, it may slightly modify its string and also be slightly modified. We should note that the prefix cannot be a standalone word.
- c. The body may also be suffixed by a pronoun, a gender determiner, a tense determiner, etc., or some combination of these. When a *suffix follows* a body, it may slightly modify its string and also be slightly modified. We should also note that the suffix cannot be a standalone word [2].

6. Arabic word categories

Arabic grammarians traditionally classify words into three main categories. These categories are also divided into subcategories, which collectively cover the whole of the Arabic language, these categories are:

1. Nouns

A noun in Arabic is a name or a word that describes a person, thing, or idea.

The linguistic attributes of nouns that have been used in our tagset [14] are:

- **Gender:** Masculine Feminine Neuter
- **Number:** Singular Plural Dual
- **Person:** First Second Third
- **Case:** Nominative Accusative Genitive
- **Definiteness:** Definite Indefinite

2. Verbs

Verbs indicate an action, although the tenses and aspects are different. Verb aspect is divided into three classes: Perfect, Imperfect, and Imperative.

The verbal attributes are [14]:

- **Gender:** Masculine Feminine Neuter
- **Number:** Singular Plural Dual
- **Person:** First Second Third
- **Mood:** Indicative Subjunctive Jussive

3. Particles

The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions and Interjections.

The subcategories of particle are [14]:

- | | | |
|---------------|--------------|--------------|
| Prepositions | Adverbs | Conjunctions |
| Interjections | Exceptions | Negatives |
| Answers | Explanations | Subordinates |

7. Our Lexicon Approach

The objective of our project is to build a Lexicon for the Arabic language by automatic means. This lexicon contains morphological information, part-of-speech tags, linguistic attributes, patterns and affixes for all lexicon entries.

Our new algorithm for constructing a lexicon for the Arabic Language automatically starts by entering a vowelized or non-vowelized Arabic text document taken from the Holy Qur'an and 242 Arabic abstracts chosen from the *Proceedings of the Saudi Arabian National Computer Conference*. It ends with a lexicon for the Arabic Language. Figure 2 shows the main stages for constructing a Arabic language lexicon using our system.

To achieve the objective of the project, we have designed and implemented several processes that carry out separate and well-defined tasks that can be re-used in other natural language processing systems.

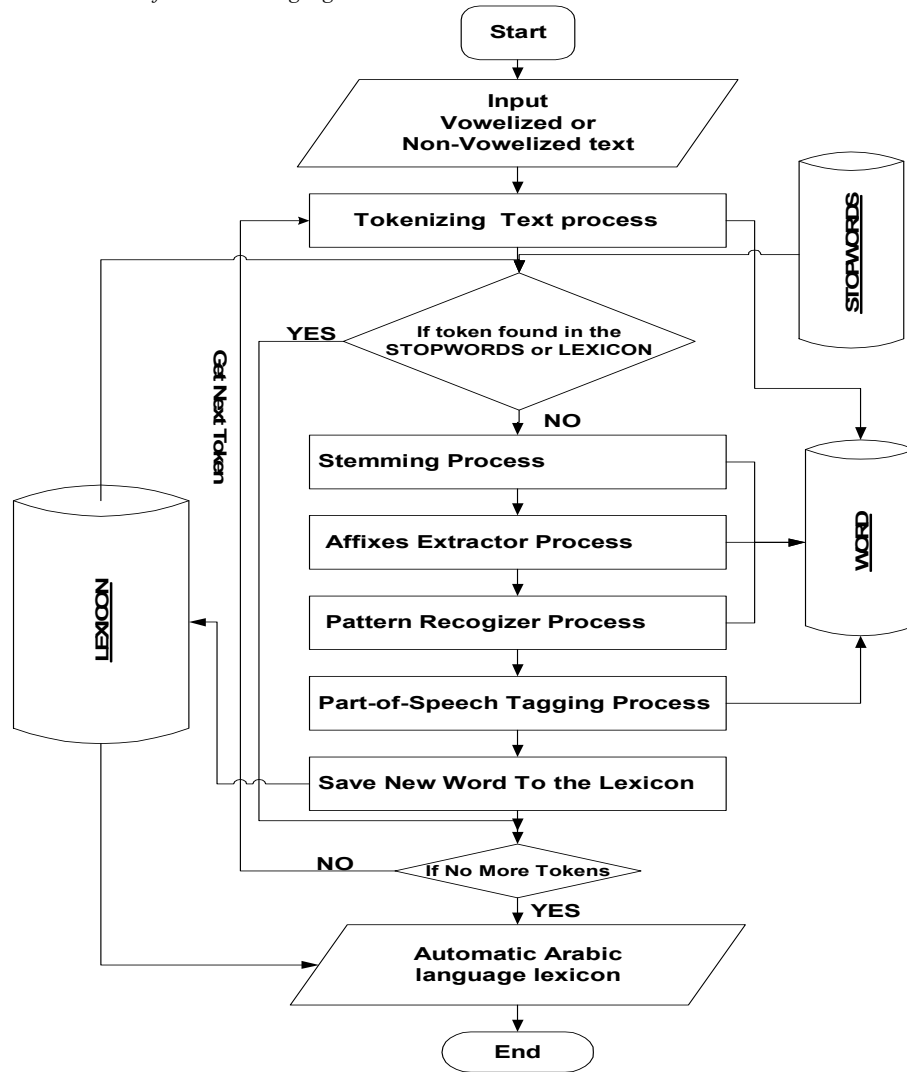


Figure 2. Algorithm for the Automatic Construction of an Arabic Language lexicon

Tokenizing process

This process locates a document and isolates the words (tokens); these tokens are stored in a work table that contains all the information associated with each token in the document. These tokens are compared with the contents of the STOPWORD and LEXICON tables; if they are stopwords or they have been already stored in the lexicon then their linguistic attributes are stored in the WORD table, where each token is given an automatic unique number, it is stored with its word number and document number, and the process continues with the next token.

The Stopwords matcher is constructed of the STOPWORDS table in the project database which contains many Arabic stop words along with their lexical information. These stopwords were compiled by [20].

The Stopwords matcher process compares the words in the document with words in the STOPWORDS table. If it is matched then the lexical information about the word stored previously in the STOPWORDS table is assigned to the words in the document as its tag and other attributes and processes continue with the next token.

The Stemming Process

This process is designed to extracting the root of all of the words in the document. The stemming process extracts roots constructed of three letters and stores the root in the Root attribute in the WORD table. The root of the word is the most important morphological attribute since many processes use the root of the word to accomplish its task. Many morphological systems have been built to extract the roots of the Arabic words, e.g. Al-Fadaghi and Al-Anzi [3]; we used an

algorithm for extracting the root of the Arabic word designed by Al-Shalabi [5].

Affix Extraction Process

This process extracts the affixes from the word. Affixes are of three types: prefixes are the extra characters added to the beginning of the word; infixes are the extra characters added to the middle of the word; suffixes are the extra characters added at the end of the word. By extracting the root of the word, we are specifying the original characters of the word, so all other letters that form the word are extraneous characters. This process determines the affixes of each of the words and stores them in the lexicon.

The Pattern Recognizer Process

This process extracts patterns from the Arabic word documents. The pattern recognizer identifies relative pronouns attached to the end of the verbs and the definiteness letters, progress verb letters, order verb letters, prepositions, conjunctions such as "و", "ف", etc. attached to the beginning of the word. These affixes are not part of the word and should be avoided when the pattern is recognized.

The pattern is constructed by combining the letters "ف", "ع", "ل", "ن" with the affixes of the word according to

their order in the word. Then the patterns are stored in the lexicon with the word after they have been extracted.

The Part-of-Speech Tagging Process

This process assigns the part-of-speech tags for all lexicon entries. We used the full automatic Arabic text tagging system implemented by Kanaan, Al-Shalabi and Sawalha [14]. Then part-of-speech tags are stored in the lexicon.

Storing the Words in the Lexicon

This process is responsible for storing new words in the LEXICON table in the project database. When all of the operations have been finished, all tokens of the document have been processed and stored in the WORD table. Along with each word are stored all its attributes such as part-of-speech tag, root, pattern, affixes, relative pronouns, and conjunctions attached to the token. This process will transfer all new words, not already found, to the lexicon along with all associated lexical attributes.

Once this process has finished, the user can view all words stored in the lexicon on the screen shown in Figure 3.

الكلمة	التصنيف الرئيسي	التصنيف الفرعي	الجنس	العدد	التنحيص	العلامة الاعرابية	التعريف	العلامة الاعرابية	الجزر	الوزن
حذف	فعل	فعل ماضي	مذكر	١/٣/١	غائب				حذف	فعل
دراسة	اسم	اسم معرب	مذكر	مفرد	غائب	نكرة			درس	فعالة
سارعت	فعل	فعل ماضي	مذكر/مؤنث	مفرد				داللي	سرع	فاعل
عمية	اسم	اسم معرب	مذكر	مفرد	غائب	نكرة			عمل	فعلية
فعالية	اسم	اسم معرب	مذكر	مفرد	غائب	مجرور	نكرة		فعل	فعالية
لاسترجاع	اسم	اسم معرب	مذكر	مفرد	غائب	مجرور	نكرة		رجع	استفعال
لعمية	اسم	اسم معرب	مذكر	مفرد	غائب	مجرور	نكرة		عمل	فعلية
للاجرة	اسم	اسم معرب	مذكر	مفرد	غائب	مجرور	نكرة		جهر	فعله
للمدخلات	اسم	اسم معرب	مؤنث	جمع	غائب	مجرور	نكرة		دخل	مفعلات

Figure 3. The Lexicon

Tagging Nouns

We have constructed many morphological rules that identify the words as nouns in the Arabic language. Rules for extracting nouns from documents are constructed according to the special grammar of the Arabic language. This grammar includes the affixes of the word.

- Prefixes such as; "ال", "لل", etc.
- Suffixes such as; "ة", "ة", "ى", etc.
- Diacritic Marks attached to the first and last letters of the word.

The position of the word in the sentence is a good indicator in identifying nouns. Some words are always followed by nouns, such as "كان وأخواتها", "إن", "واخواتها", and some of these words are mainly used in recognizing proper nouns such as "السيد", which means "Mr", "المملكة" which means "kingdom", etc. Thus we can construct a rule to help us identify nouns in the text using these phenomena.

The Arabic Language has many patterns; some of them apply to nouns only; some of them apply to verbs only; and some apply to both nouns and verbs.

We recognize some of those used as noun patterns, such as "فاعل", "مفعول", etc.

Tables 2 to 6 show a complete set of rules for tagging declinable nouns.

Table 2: Rules for Suffixes of Declinable Nouns

Rule #	Rule Description	Example	Noun type	Linguistic attribute
Rule 1	Any word ending with "ون" or "ين" and not beginning with any of these particles (حروف المضارعة " ن ، ي ، ت ، ا ")	مهندسون <i>Engineers</i>	اسم معرب <i>Declinable Noun</i>	جمع مذكر <i>Plural, Masculine</i>
Rule 2	Any word ending with "ات" if the mark before the particle "ا" is "ا"	مدرسات <i>Teachers</i>	اسم معرب <i>Declinable Noun</i>	جمع مؤنث <i>Plural, feminine</i>
Rule 3	Any word ending with "ة" must be a NOUN.	كتابة <i>Writing</i>	اسم معرب <i>Declinable Noun</i>	مفرد مؤنث <i>Singular, feminine</i>
Rule 4	Any word ending with "ة" or "ى" and not beginning with any of these particles (حروف المضارعة " ن ، ي ، ت ، ا ") must be a NOUN.	املاء <i>Dictation</i>	اسم معرب <i>Declinable noun</i>	مفرد مؤنث <i>Singular feminine</i>
Rule 5	Any word with the last letter "ياء ي" and the previous letter "كسرة كسر" must be a NOUN.	حمر اوي <i>Has red color</i>	اسم منسوب <i>Related Noun</i>	

Table 3: Rules for Prefixes of Declinable Nouns

Rule #	Rule Description	Example	Noun type	Linguistic attributes
Rule 6	Any word beginning with "ال" or "أل" or "إل" must be a NOUN.	(الكتاب) <i>The book</i>	اسم معرب <i>Declinable noun</i>	
Rule 7	Any word beginning with "لل" must be a NOUN.	(للعلم) <i>For the science</i>	اسم معرب <i>Declinable noun</i>	مجرور <i>Genitive</i>
Rule 8	Any word beginning with "م" (ميم) if the char. previous to the last letter is "كسرة كسر" must be a NOUN.	(مكرم)	اسم فاعل <i>Agent noun</i>	
Rule 9	Any word beginning with "م" (ميم) if the char. previous to the last letter is "فتحة فتح" must be a NOUN.	(مكرم)	اسم مفعول <i>Patient noun</i>	

Table 4: Words Always Followed by Nouns

Rule #	Rule Description	Example	Noun type	Linguistic attributes
Rule 10	Any word following "يا" or "ها" or "يا" or "ها" must be a NOUN.	(يا محمد)	اسم منادى	منصوب Accusative
Rule 11	Any word following any of (حروف " الجر " من ، عن ، على ...) must be a NOUN.	(على الشجرة)	اسم مجرور	مجرور Genitive
Rule 12	Any word following any of (إن ، أن) must be a NOUN. (إن واخواتها)	(ليت الشباب ...)	اسم إن	منصوب Accusative
Rule 13	Any word following "إلا" must be a NOUN.	(إلا الفشل)	اسم معرب	منصوب Accusative
Rule 14	Any word followed by any of these words [رئيس ، معالي ، مملكة ، مدينة] must be a NOUN.	المملكة الأردنية	اسم علم Proper noun	
Rule 15	[أب ، أخ ، حم ، نو ، فو]	أبو أحمد	الاسماء الخمسة	

Table 5: Rules for Diacritic Marks of Declinable Nouns

Rule #	Rule Description	Example	Noun type	Linguistic attribute
Rule 16	Any word ending with ' or ' or ' must be a NOUN. (التنوين)	(رجل) (رجلاً)	اسم معرب	نكرة Indefinite
Rule 17	Any word in which the mark of the first letter is "ضممة" and the mark of the second letter is "فتحة" & the third letter is "ي" must be a NOUN.	(كُتِيب)	اسم تصغير	

Table 6: Rules for Patterns of Declinable Nouns

Rule #	Rule Description	Example	Noun type	Linguistic attribute
Rule 18	Any word that has the weight [على] (فاعل) => [وزن]	(شامل)	اسم فاعل	
Rule 19	Any word that has the weight [على] (مفعول) => [وزن]	عقار، مقول ، طوال ، عفور ، فاروق ، حميم ، حزير	(صيغ مبالغة) لاسم الفاعل	
Rule 20	Any word that has the weight [على] (مفعول) => [وزن]	(مقتول)	اسم مفعول	
Rule 21	Any word that has the weight [على] (مفعول) => [وزن]	(ملعب)	اسم مكان	
Rule 22	Any word that has the weight [على] (مفعول) => [وزن]	(منزل)،(موعد)	اسم مكان/زمان	
Rule 23	Any word that has the weight [على] (مفعول) => [وزن]	(مخرز)	اسم آلة	

Tagging Verbs

This process is responsible for identifying verbs in the document. A verb is defined as a word that indicates a meaning by itself that is united with a tense or time. Verbs take words or letters as indicators such as the particles "قد", "سوف", or pronouns, or the letters "س", "ن", "ت" [1].

The rule of Arabic morphology are based on patterns, affixes, and combinations of the two.

Pattern

Verbs in the Arabic language have roots that consist of 3 or 4 letters. From a single root many verb forms can be generated according to fixed rules that add letters such as "ا", "أ", "ت", "س", "ل", "م", "ن", "ن", "م", "ل", "س", "ت", "أ", "ا", "ي", "و", "ه", "سالتمونيها". These fixed rules are called patterns. Table 7 lists 15 different essential patterns.

Table 7: Essential Verb Patterns

#	Pattern	Pattern analysis									Added letters	# of added letters
		ل			ع		ف					
1	فعل	ل			ع		ف					0
2	فَعَلَ	ل			ع	ع	ف				ع	1
3	فاعِل	ل			ع	ا	ف				ا	1
4	أفعل	ل			ع		ف			ا	أ	1
5	تفَعَّل	ل			ع	ع	ف	ت			ع، ت	2
6	تفاعِل	ل			ع	ا	ف	ت			ت، ا	2
7	انفعل	ل			ع		ف		ن	ا	ا، ن	2
8	افنعل	ل			ع		ف	ت		ا	ا، ت	2
9	افعل	ل		ل	ع		ف			ا	ا، ل	2
10	استفعل	ل			ع		ف	ت	س	ا	ا، س، ت	3
11	افعوعل	ل			ع	و	ع	ف		ا	ا، ع، و	3
12	فعلل	ل			ع		ف					0
13	تفعلل	ل			ع		ف	ت			ت	1
14	افعلل	ل	ل	ل	ع		ف			ا	ا، ل	2
15	افعللل	ل	ل	ن	ع		ف			ا	ا، ن، ل	3

Affix

Some affixes are used with verbs and some with nouns and some with both verbs and nouns. We have extracted 31 groups of affixes that are used with the essential patterns listed in Table 7; these affixes affect verb semantics, such as verb aspect (perfect, imperfect, imperative), gender (masculine, feminine), number (singular, dual, plural), and person (first, second, third), and mood (indicative, subjunctive, jussive) as shown in Table 8.

The number property of words that have patterns with no suffixes as in rules 1 and 14 cannot be specified

directly. To identify the number, we have to refer to the next word, which is typically the subject of the sentence, since the verb and its subject are identical in number property. For example, "كتب الطالب الدرس" which means "the student wrote the lesson", the verb "كتب" "wrote" in this sentence is a singular verb, while it is dual in the sentence "كتب الطالبان الدرس", which means "the two students wrote the lesson", and plural in the sentence "كتب الطلاب الدرس" which means "the students wrote the lesson". By referring to the subject we can determine the number of the verb.

Table 8: Verb Affixes Rules

#	Rule	Category	Gender	Number	Person	Mood
1	Pattern	1 Perfect	1 Masculine	1+2+3	3 Third	Indicative
2	Pattern + ا	1 Perfect	1 Masculine	2 Dual	3 Third	Indicative
3	Pattern + وا	1 Perfect	1 Masculine	3 Plural	3 Third	Indicative
4	Pattern + ت	1 Perfect	2 Feminine	1 Singular	3 Third	Indicative
5	Pattern + تا	1 Perfect	2 Feminine	2 Dual	3 Third	Indicative
6	Pattern + ن	1 Perfect	2 Feminine	3 Plural	3 Third	Indicative
7	Pattern + ت	1 Perfect	1 Masculine	1 Singular	2 Second	Indicative
8	Pattern + تما	1 Perfect	3 Neuter	2 Dual	2 Second	Indicative
9	Pattern + تم	1 Perfect	1 Masculine	3 Plural	2 Second	Indicative
10	Pattern + ت	1 Perfect	2 Feminine	1 Singular	2 Second	Indicative
11	Pattern + تن	1 Perfect	2 Feminine	3 Plural	2 Second	Indicative
12	Pattern + ت	1 Perfect	3 Neuter	1 Singular	1 First	Indicative
13	Pattern + نا	1 Perfect	3 Neuter	3 Plural	1 First	Indicative
14	ي+Pattern	2 Imperfect	1 Masculine	1+2+3	3 Third	Indicative

15	ان+Pattern+ي	2 Imperfect	1 Masculine	2 Dual	3 Third	Indicative
16	ون+Pattern+ي	2 Imperfect	1 Masculine	3 Plural	3 Third	Indicative
17	ت+Pattern	2 Imperfect	2 Feminine	1 Singular	3 Third	Indicative
18	ان+Pattern+ت	2 Imperfect	2 Feminine	2 Dual	3 Third	Indicative
19	ن+Pattern+ي	2 Imperfect	2 Feminine	3 Plural	3 Third	Indicative
20	ت+Pattern	2 Imperfect	1 Masculine	1 Singular	2 Second	Indicative
21	ان+Pattern+ت	2 Imperfect	3 Neuter	2 Dual	2 Second	Indicative
22	ون+Pattern+ت	2 Imperfect	1 Masculine	3 Plural	2 Second	Indicative
23	ين+Pattern+ت	2 Imperfect	2 Feminine	1 Singular	2 Second	Indicative
24	ن+Pattern+ت	2 Imperfect	2 Feminine	3 Plural	2 Second	Indicative
25	أ+Pattern	2 Imperfect	3 Neuter	1 Singular	1 First	Indicative
26	ن+Pattern	2 Imperfect	3 Neuter	3 Plural	1 First	Indicative
27	ا+Pattern	3 Imperative	1 Masculine	1 Singular	2 Second	Indicative
28	ا+Pattern + ا	3 Imperative	3 Neuter	2 Dual	2 Second	Indicative
29	وا+Pattern + ا	3 Imperative	1 Masculine	3 Plural	2 Second	Indicative
30	ي+Pattern + ا	3 Imperative	2 Feminine	1 Singular	2 Second	Indicative
31	ن+Pattern + ا	3 Imperative	2 Feminine	3 Plural	2 Second	Indicative

Rules

Rules are extracted from the syntax of the Arabic sentence formation; tags are assigned to verbs

حروف الجزم + فعل

حروف النصب + فعل

فعل + حروف العطف + فعل

according to their position in the Arabic sentence, where some types of pronouns, prepositions and letters are affixed to verbs. Some of these rules are:

{ "قد" ، "سوف" } + فعل

{ "س" ، "ت" ، "ن" } + فعل

فعل + ضمير متصل

Lexical Attribute Rules for the Arabic Language

Once the type and major subtype of the word have been identified, another process is needed to obtain the linguistic attributes of the word (Person, Number, Gender, Aspect, and Mood). Each attribute requires special treatment.

1- Gender (Masculine, Feminine)

We assumed that all Arabic words are masculine except those words ending with "ة" , "اء" , "ى" or "ي" , which are feminine.

2- Number (Singular, Plural, Dual)

If a word ends with "ون" or "ين" and does not begin with any of the letters "أ" , "ت" , "ي" , "احرف المضارعة" , "ن" , "أ" , "ت" , "ي" , then the number attribute of the word must be masculine plural "جمع مذكر سالم" and if a word ends with "ات" it must be feminine plural "جمع مؤنث سالم"

Any noun that ends with "ان" or "ين" must have dual number attribute; other words will be assumed to be singular.

3- Person (First, Second, Third)

This lexical attribute is used for pronouns only whether they are attached to the word or separate. Pronouns indicate first, second and third person as follows:

1- First person pronouns : (ت، نا، أنا، نحن، ي)

2- Second person pronouns: (ت، تما، تم، تن، ن، ا،

ك، ي، كما، كم، كن)

3- Third person pronouns: (هو، هما، هم، هي، هن)

4- Case (Nominative, Accusative, Genitive)

The case of any singular, feminine or plural noun is determined according to the following rules:

- *Nominative* "مرفوع": if the word ends with a letter that has the diacritic mark " الضمه "
- *Accusative* "منصوب": if the word ends with a letter that has the diacritic mark " الفتحة "
- *Genitive* "مجرور": if the word ends with a letter that has the diacritic mark " الكسرة "

The case of any masculine and plural noun is determined according to the following rules:

- *Nominative* "مرفوع": if the word ends with "ون"
- *Accusative* "منصوب": if the word ends with "ين" and it is not preceded by any preposition and the previous word does not have the genitive case.
- *Genitive* "مجرور": if the word ends with "ين" and it is preceded by any preposition or the previous word has genitive case.

The case of any dual noun is determined according to the following rules:

- *Nominative* "مرفوع" if the word ends with "ان"
- *Accusative* "منصوب" if the word ends with "ين" and it is not preceded by any preposition and the previous word does not have the genitive case.
- *Genitive* "مجرور" if the word ends with "ين" and it is preceded by any preposition or the previous word has genitive case.

5- Definiteness (Definite, Indefinite)

We assume that the definiteness attribute of a noun is Indefinite (نكرة) except for these types of nouns.

- 1- Proper nouns. "اسم العلم"
- 2- Any noun made definite by "ال" (the).
- 3- Any noun following another definite noun. "مضاف إلى معرفة"
- 4- Any noun following "أحرف النداء", "يا", "التمنيّة" المقصودة بالنداء

Some other nouns are always definite (معرفة) such as:

- 1- Pronouns "الضمائر"
- 2- اسم الإشارة
- 3- الأسماء الموصولة

8. Implementation

We built a program that applies all of the rules described in this paper using MicroSoft Visual Basic 6.0 and MicroSoft Access Database. The figures below show the program screens used to input text and display the results of applying rules to build our lexicon for the Arabic language automatically.

Figure 4 is designed to facilitate the entry of new text by pressing the button labeled "فتح" (1) and selecting the document using the open dialog box displayed; the new text is shown in the form text box, and a unique document number must be entered for each document. The button labeled "تحليل" (2) performs tokenizing, stemming, affix extraction and pattern generation processes and then the Word List screen shown in Figure 5 will be displayed. The button labeled "استعلام عن معلومات وثيقة" (3) allows the user to inquire about any document that has been processed and stored in the project database before; the document number should be entered before pressing this button. Then the Word Display Screen shown in Figure 5 is displayed. The button labeled "عرض محتويات LEXICON" (4) allows users to view the contents of the lexicon stored in the project database, the screen shown in Figure 7 will be displayed. Pressing the last button labeled "انتهاء" (5) ends the program.



Figure 4. The Current Document Scanned by the System

The screen shown in Figure 5 shows all the words (except for the stop words) extracted from the most recent document after applying tokenizing, stemming,

and the affix removal and pattern generation processes. The table on the screen shows the word, its root, extra letters attached to the beginning of the word such as

conjunctions, letters indicating imperfect verbs, prefixes, two groups of infixes, suffixes, pronouns, and the pattern. When the button labeled "التحليل" (1) is pressed the part-of-speech tagging process is applied to the word shown on this screen and the linguistic

attribute extraction process is also applied to these words. This information is displayed on the screen shown in Figure 6. Pressing the last button labeled "انتهاء" (2) ends the program.

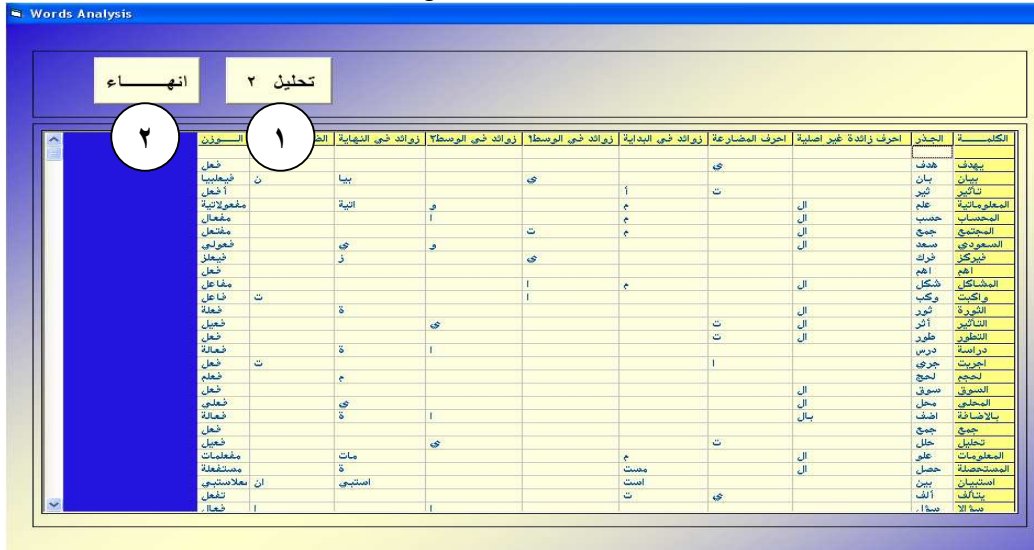


Figure 5. All the Words except the Stop Words

The screen shown in Figure 6 is designed to display linguistic information about the words in the processed document other than stop-words. The screen shows the word, the main part-of-speech category, the subcategory of the part-of-speech tag and the linguistic attributes (gender, number, person, case, definiteness, aspect, and mood). The button labeled "استعلام" (1) displays the lexicon after processing the document as shown in Figure 7. The button labeled "القائمة الرئيسية" (2) shows the title screen. Pressing the button labeled "انتهاء" (3) ends the program.

The screen shown in Figure 7, is designed to display the lexicon s constructed automatically so far. The table in this screen displays the lexicon information about a given word, its linguistic attributes (gender, number, person, case, definiteness and mood). it also displays the root and the pattern. The button labeled "بحث" (1) displays the search screen where we can search for specific words stored in the lexicon database. The button labeled "رجوع" (2) shows the title screen. Pressing the button labeled "انتهاء" (3) ends the program.

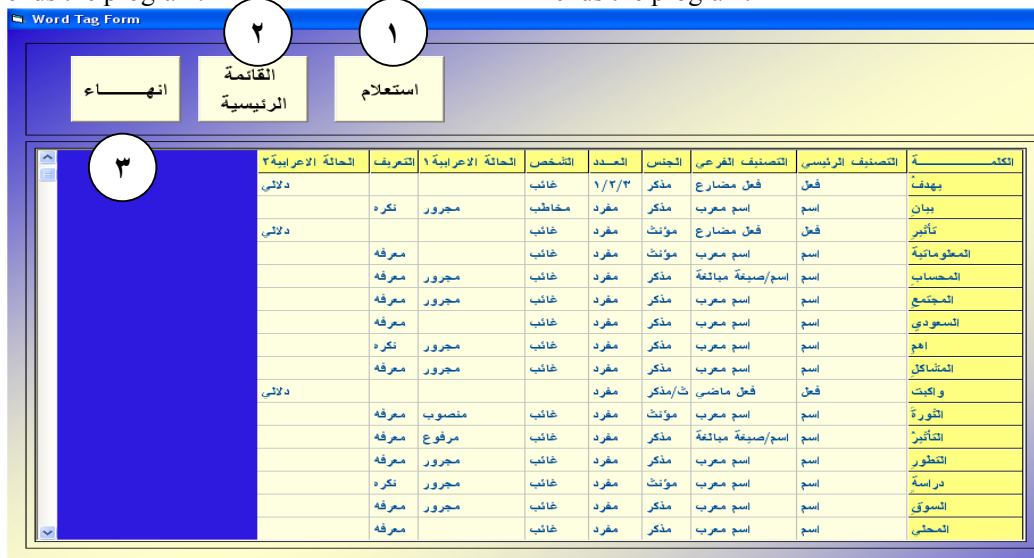


Figure 6. Displaying a Word and Its Attributes



Figure 7. Displaying the Lexicon

The search screen shown in Figure 8 is invoked by pressing the button labeled "بحث" (1) in the fifth screen shown in Figure 7. This screen allows the user to search the lexicon database for an item that matches the word entered in the text box labeled "الكلمة" (1) on the screen. It then displays the information stored in

the lexical entry for that word. To execute the search process the user enters the word and presses the button labeled "بحث" (2). The button labeled "رجوع" (3) displays the previous screen, the fifth screen, shown in Figure 7. The button labeled "انتهاء" (4) terminates the program.



Figure 8. The Search Screen

9. Results

We tested our system using passages from the Holy Qur'an, which are vowelized, and another set of non-vowelized Arabic abstracts chosen from the

Proceedings of the Saudi Arabian National Computer Conference. We run our system on a group of these documents selected randomly; we obtained the results shown in Table 9.

Table 9: System Accuracy Table

System model		# Words	# incorrect Words	# correct Words	% incorrect Words	% correct Words	Total %
Stemming Process		388	69	319	% 17.78	%82.22	%82.22
Pattern Analyzer Process		319	6	313	% 1.88	%98.12	%98.12
Part-of-Speech Tagging		313	11	302	% 3.50	%96.50	%96.50
Lexical Attribute Analyzer Process	Gender	302	27	275	% 9.95	% 91.05	%96.03
	Number	302	18	284	% 5.96	% 94.04	
	Person	302	13	289	% 4.30	% 95.70	
	Case	302	9	293	% 2.98	% 97.02	
	Definiteness	302	1	301	% 0.33	% 99.67	
	Mood	302	4	298	% 1.33	% 98.67	

When we calculate the system's efficiency, we discard the errors coming from the stemming process, since the focus of the research is on constructing an Arabic lexicon automatically. Other essential parts of the system are analyzed and the efficiency of each part is calculated. Faults in the system were caused by some uncontrolled conditions; the stemming algorithm used in our program is designed for extracting roots constructed of three letters, however some roots have four letters, which we don't handle in our system. Another factor that affects the efficiency of the system is the incorrect roots extracted when some of the word's letters are doubled and the doubled letters are marked with shadda "´", which is not a diacritic but is a mark that the character is doubled when it is pronounced. Errors in the number attribute occurred because some plurals in the Arabic language can be formed irregularly and some singular or dual words have the shape as the plural words, which makes detecting this attribute automatically a very hard task.

11. References

- [1] Abuleil, S. and Evens, M. "Discovering Lexical Information by Tagging Arabic Newspaper Text", Workshop on Semitic Language Processing. COLING-ACL.98, University of Montreal, Montreal, PQ, Canada, 1998, pp 1-7.
- [2] Ahmed, Mohamed Attia M. Elaraby 2000. A Large-Scale Computational Processor of the Arabic Morphology, and Applications, M.Sc. Thesis, Cairo University.
- [3] Al-Fadaghi, S, and Al-Anzi, F. "A New Algorithm to Generate Arabic Root-Pattern Forms". 11th National Conference and Exhibition, Proceeding of 11th National Conference and Exhibition. Riyadh, Saudi Arabia, 1989.
- [4] Ali F. and Jean S. , " Intuitive Coding of the Arabic Lexicon".
- [5] Al-Shalabi, R. and Evens, M. "A Computational Morphology System for Arabic". Workshop on Semitic Language Processing. COLING-ACLTM98, University of Montreal, Montreal, PQ, Canada, 1998. pp. 66-72.

10. Conclusion

In this study, we developed a methodology for the automatic construction of an Arabic lexicon. Our implementation achieved an accuracy of about 96%. We also developed many rules for extracting linguistic attributes from text that may be helpful for Arabic language applications in the fields of information retrieval and natural language processing. We have designed and implemented several utility processes; a pattern extractor and a stemming process, and a part-of-speech tagging process that are essential components of our system. The pattern extractor process uses the roots extracted by the stemming process to generate the patterns of the words. The error percentage of about 4% percent comes from faulty roots and patterns. The system can achieve more accuracy by enhancing the root finding algorithm, which would also enhance the pattern extractor process.

- [6] Al-Shalabi, R., Kanaan, G., Muaidi H. New Approach for Extracting Arabic Roots, Proceeding of the International Arab Conference on Information Technology. Alexandria, Egypt, 2003, pp 42-59.
- [7] Amsler, Robert A. 1982. "Computational Lexicology: A Research Program." 657-663. Proceedings of the National Computer Conference, Houston, TX. 657- 663.
- [8] Bonne J., Doug J., "Acquisition of Semantic Lexicon" University of Maryland, 1999.
- [9] Calzolari, Nicoletta. 1984. "Machine-Readable Dictionaries, Lexical Databases and the Lexical System," COLING '84, Stanford, 1984, 460.
- [10] Dafydd G., "Lexicography, lexicology, lexicon theory", <http://coral.lili.unibielefeld.de/Classes/Winter98/ComLex/GibbonElsnet/elsnetbook.dg/node1.html>
- [11] Dorr, Bonnie J. and Douglas Jones. 1996. "Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision," Proceedings of the Workshop on Breadth and Depth of Semantic Lexicons, 34th Annual Conference of the Association for Computational Linguistics, Santa Cruz, CA, pp. 42--50, 1996.
- [12] Grishman, Ralph, Catherine Macleod and Adam Meyers. 1994. "COMLEXSyntax: Building a Computational Lexicon, COLING94. Kyoto.
- [13] Jörg Tiedemann, "Automatical Lexicon Extraction from Aligned Bilingual Corpora", Diploma Thesis, Otto-von-Guericke-Universität Magdeburg, 1997, <http://stp.lig.uu.se/~jeorg/diplom/node1.html>.
- [14] Kanaan G., Al-Shalabi R., Sawalha M., " Full Automatic Arabic Text Tagging System", the proceedings of the International Conference on Information Technology and Natural Sciences, Amman/Jordan, 2003, pp 258-267.
- [15] Mohamed A. and Elaraby A., " A Large-Scale Computational Processor of the Arabic Morphology, and Applications", 2000.
- [16] Naser, R.T. 1967. The Structure of Arabic from Sound to Sentence. Librairie du Liban, Beirut, Lebanon.
- [17] Philip K. and Kevin K. "Learning a Translation Lexicon from Monolingual Corpora".
- [18] Ralph G., Nicoletta C. , " Survey of the State of the Art in Human Language Technology", New University, Istituto di Linguistica Computazionale del CNR.
- [19] Sara R. , " Building a hyponymy lexicon with hierarchical structure", Centre for Speech Technology (CTT), KTH Stockholm, Sweden, GSLT, ACL 2002.
- [20] Salls, Bonnie Glover and Yaser Al- Onaizan, 2003. Arabic Stop Word List, NLP research activities at University of Southern California's Information Science Institute.
- [21] Smadja, F., "Retrieving Collocations from Text," Computational Linguistics, 19 (1), 1993, 143-177.
- [22] Tiedemann, Jörg. 1997. Automatical Lexicon Extraction from Aligned Bilingual Corpora. Diploma Thesis, Otto-von-Guericke-Universität Magdeburg, 1997, <http://stp.lig.uu.se/~jeorg/diplom/node1.html>.
- [23] Walker, D.E., 1985. "Knowledge Resource Tools for Accessing Large Text Files", in G. Johannesen, (ed.), Information in Data: Proc. of the Centre for the New OED, 11-24.