Using Text Mining to Identify Crime Patterns from Arabic Crime News Report Corpus

PhD Thesis

Meshrif Fayad Alruily

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Software Technology Research Laboratory
Faculty of Technology
De Montfort University

September 2012

DEDICATION

I would like to dedicate this thesis to my parents,

my brothers and sisters,

my wife,

and my children (lovely twins) Abdulaziz and Talal.

ABSTRACT

Most text mining techniques have been proposed only for English text, and even here, most research has been conducted on specific texts related to special contexts within the English language, such as politics, medicine and crime. In contrast, although Arabic is a widely spoken language, few mining tools have been developed to process Arabic text, and some Arabic domains have not been studied at all. In fact, Arabic is a language with a very complex morphology because it is highly inflectional, and therefore, dealing with texts written in Arabic is highly complicated.

This research studies the crime domain in the Arabic language, exploiting unstructured text using text mining techniques. Developing a system for extracting important information from crime reports would be useful for police investigators, for accelerating the investigative process (instead of reading entire reports) as well as for conducting further or wider analyses. We propose the Crime Profiling System (CPS) to extract crime-related information (crime type, crime location and nationality of persons involved in the event), automatically construct dictionaries for the existing information, cluster crime documents based on certain attributes and utilise visualisation techniques to assist in crime data analysis.

The proposed information extraction approach is novel, and it relies on computational linguistic techniques to identify the abovementioned information, i.e. without using predefined dictionaries (e.g. lists of location names) and annotated corpus. The language used in crime reporting is studied to identify patterns of interest using a corpus-based approach. Frequency analysis, collocation analysis and concordance analysis are used to perform the syntactic analysis in order to discover the local grammar.

Moreover, the Self Organising Map (SOM) approach is adopted in order to perform the clustering and visualisation tasks for crime documents based on crime type, location or nationality. This clustering technique is improved because only refined data containing meaningful keywords extracted through the information extraction process are inputted into it, i.e. the data is cleaned by removing noise. As a result, a huge reduction in the quantity of data fed into the SOM is obtained, consequently, saving memory, data loading time and the execution time needed to perform the clustering. Therefore, the computation of the SOM is accelerated. Finally, the quantization error is reduced, which leads to high quality clustering. The outcome of the clustering stage is also visualised and the system is able to provide statistical information in the form of graphs and tables about crimes committed within certain periods of time and within a particular area.

The proposed model architecture is validated through experiments using a corpus collated from different sources; it was not used during system development. Precision, recall and F-measure are used to evaluate the performance of the proposed information extraction approach. Also, comparisons are conducted with other systems. In order to evaluate the clustering performance, four parameters are used: data size, loading time, execution time and quantization error.

DECLARATION

I declare that the work described in this thesis is original work undertaken by me for the degree of Doctor of Philosophy, at the Software Technology Research Laboratory (STRL), De Montfort University, United Kingdom. No part of the material described in this thesis has been submitted for the award of any other degree or qualification in this or any other university or college of advanced education.

Meshrif Fayad Alruily

ACKNOWLEDGEMENTS

First and foremost, I thank **Allah the Almighty** for giving me the ability to complete this research.

I would also like to express my deepest gratitude and appreciation to my supervisor **Dr. Aladdin Ayesh**, for his expertise, guidance, ideas, encouragement and constructive comments throughout my research; I had countless discussions with him, either face to face or online, discussing my work, and without him, this thesis would not have been possible.

My sincere thanks also go to **Prof. Hussein Zedan**, the head of STRL, for giving me the motivation to start this research. Also, I am thankful to him for his support and valuable suggestions during time of research.

Special thanks also to my parents, my brothers and sisters for their endless support, love and prayers throughout my life. Without them, I would never have been able to achieve my goals.

Last but not least, my deepest thanks go to my wife **Amal**, for her love, patience and personal support and for taking full responsibility for caring for our twins (Abdulaziz and Talal).

Thank you all.

Leicester, England, 2011

Meshrif Alruily

PUBLICATIONS

1. **Alruily, Meshrif and Ayesh, Aladdin and Zedan, Hussein, *Crime Type Document Classification from Arabic Corpus*, In Proceedings of the 2009 Second International Conference on Developments in eSystems Engineering, DESE '09, IEEE Computer Society, pp 153-159, 2009.**

2. **Alruily, Meshrif and Ayesh, Aladdin and Zedan, Hussein, *Automatically Constructing Dictionaries for Extracting Meaningful Crime Information from Arabic Text.* In Proceeding of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, IOS Press, 2010.**

3. **Alruily, Meshrif and Ayesh, Aladdin and Al-Marghilani, Abdulsamad, *Using Self Organizing Map to Cluster Arabic Crime Documents*, In Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT), 2010**

4. **Alruily, Meshrif and Ayesh, Aladdin and Zedan, Hussein, *Automated dictionary construction from Arabic corpus for meaningful***

crime information extraction and document classification, In Proceeding of the Computer Information Systems and Industrial Management Applications (CISIM), 2010.

5. **Alruily, Meshrif and Ayesh, Aladdin, and Zedan, Hussien, _Arabic Language in the Context of Information Extraction Task_, International Journal of Computational Linguistics Research (IJCLR), vol. 2, issue 2, pp 83-90, 2011.**

6. **Alruily, Meshrif and Ayesh, Aladdin and Zedan, Hussein, _Crime Profiling for Arabic Language Using Computational linguistic Techniques_, International Journal of Information Processing and Management, Elsevier, 2011. (Submitting)**

Contents

List of Figures

Listings

List of Tables

List of Acronyms

ACE Automatic Content Extraction

ACNRC Arabic Crime News Report Corpus

ANERsys Arabic Named Entity Recognition System

ANNs Artificial Neural Networks

BAMA Buckwalter Arabic Morphological Analyse

BMU Best Matching Unit

BoW Bag of Word

CoNLL Conference on Natural Language Learning

CPS Crime Profiling System

CRF Conditional Random Fields

DARPA Defence Advance Research Project Agency

Esri Environmental Systems Research Institute

F Frequency

GB Government and Binding

G.Corpus General Corpus

GIS Geographic Information System

HMM Hidden Markov Model

KML Keyhole Markup Language

KNN K-Nearest Neighbour

L Left

LSI Latent Semantic Indexing

IDF Inverse Document Frequency

IE Information Extraction

IR Information Retrieval

KDT Knowledge-Discovery in Text

ME Maximum Entropy

MLTextMAES Multilingual Text Mining for Arabic-English Scripts

MMA Multilingual Morphological Analysis

MUC Message Understanding Conferences

MT Machine Translation

NER Name Entity Recognition

NLP Natural Language Processing

NP Noun Phrase

O Object

PNAES Persons Names Arabic Extraction System

PP Prepositional Phrase

PoS Part-of-Speech

QA Question Answering

R right

S Subject

S.Corpus Special Corpus

SOM Self Organising Map

SVM Support Vector Machines

SVO Subject Verbs and Object

V Verb

VSM Vector Space Mode

VSO Verb, Subject and Object

UAE United Arab Emirates

T Translation

TF-IDF Term Frequency - Inverse Document Frequency

TF Term Frequency

Chapter 1

Introduction

Objectives

- Provide the motivation for undertaking this research.
- Present the research question.
- Highlight the research contribution.
- Explain the research methodology.
- Provide the thesis structure.

1.1 Introduction

Nowadays the volume of data in electronic form is increasing rapidly, whether it is structured or unstructured data; according to McKnight [2], between 85 and 90% of data is held in unstructured form. Therefore, text mining is necessary for managing and extracting useful information from unstructured sets of data, such as web pages, news reports and emails, using a variety of text mining techniques. Hence, text mining has become an important and active research field. It is well known that text mining techniques have mostly been developed for the English language because most electronic data is in English. However, the recent expansion in the publication of non-English electronic text makes it more imperative than ever to develop techniques to process other languages, such as Arabic.

Moreover, in the Arabic language, as with other languages, there is a multitude of specialist texts in areas, such as the biomedical sciences, finance, politics and crime, some of which have not been investigated by researchers. This study focuses on the crime domain in the Arabic language for which no information system can be found in the literature. Therefore, there is a need to investigate the Arabic crime context using a text mining approaches.

In this current research, information extraction based on computational linguistic techniques is used for extracting specific, predefined entities from the text. These extracted entities are keywords that reveal the essential idea of the document or article; keywords play an important role in many text mining tasks, such as clustering, summarization and document retrieval [3]. One of our aims is to develop a system that is able to recognize phrases that contain information related to crime in a given document in order to extract the type of crime, the nationality of the perpetrator and other actors, and the crime location. The extracted keywords will be utilized, through a Self Organising Map (SOM), to perform the clustering and visualisation tasks. Moreover, during the information extraction stage, dictionaries containing the crime type, crime location and nationality will be automatically created. Furthermore, statistical information will be produced to clarify the picture relating to the status of crime in a particular area.

1.2 Motivation

According to Chau et al. [4], most criminal-justice data are stored in structured form, in databases. Accordingly, it should be relatively easy for crime investigators and detectives to search for particular information within this type of data; the reason of this simplicity is because the data are represented in tables, i.e. the data are organised and easy to access. However, using databases that were designed to store data in predefined fields and attributes might not be sufficient because a great deal of potentially useful information cannot be stored in such a rigid manner. On the other hand, data can also be stored in unstructured form, i.e. free text, such as police narrative and crime news reports. These reports contain highly valuable information but it is difficult to access. It would be useful to automatically identify meaningful crime information, such as type of crime, crime location, nationalities of persons involved, from crime textual reports. Therefore, the first motivation reflects the need to developing an information extraction system for the Arabic crime domain in order to mine unstructured text, thereby extracting meaningful crime-related information. Although Arabic is a very widely spoken language, very few mining tools have been developed to exploit the data that lies within bodies of Arabic text in general, and within the crime domain in particular. There are, to the author's knowledge, no systems specifically developed for the Arabic crime context.

As well known, text mining research relies heavily on the availability of a suitable corpus. Because no information systems have been applied to the crime domain in the Arabic language, the major problem to be faced is the lack of data. This issue will be solved by compiling news reports on crime incidents, published by some Arabic newspapers. The reason for exploiting newspapers is that it is difficult, in Arab countries, to obtain official reports or narrative reports from police stations, however, the news reports contain the information that the police reports would normally include. Therefore, collecting these data to make them available for other researchers is the second motivation. Moreover, one of clustering techniques will be used for grouping similar documents based on particular attributes, such as crime type or location, and this is the third motivation. The fourth motivation is related to assisting in the analysis of crime by applying certain visualisation techniques.

1.3 Research Question

As discussed above, crime related information is hidden within bodies of unstructured text documents, and thus, the buried information needs to be extracted in order for it to be used in crime analysis. Accordingly, there is a need to apply information extraction techniques in order to identify pertinent crime information, and so the research question is:

How can we develop an automated crime profiling system for the Arabic language that is able to extract crime-related information from textual documents?

In doing so, other questions need answering, such as:

How can we automatically build dictionaries for crime type, location and nationality?

How can we cluster similar crimes based on crime type, location or nationality?

How can we generate statistical information about crimes as well as utilize visualisation techniques to cluster and present these data?
However, there are many and various challenges to be overcome in this endeavour. The texts to be investigated in this project are written in Arabic, and dealing with this language is complicated because, as well known, Arabic is a language with a very complex morphology owing to the fact that it is highly inflectional. Also, prepositions and conjunctions can be attached to words as prefixes. Moreover, Arabic does not have capital letters (as many other languages do in order to recognize proper nouns), and therefore, there is no apparent clue to denote the presence of a name reflecting a crime location or a nationality. Furthermore, most studies in Arabic have focused on extracting certain types of entities, such as enamex (e.g. proper names: names of persons, locations and organisations), timex (e.g. date and time types) and numex (e.g. money and percent types) [5, 6], but they have paid very little attention to extracting event types. Consequently, recognising crime events, e.g. murder or theft, is one of the key challenges of this research. Moreover, this domain has not been investigated in the literature in the context of the Arabic language. Therefore, there are no available Arabic resources, such as corpora for analysing this domain. Furthermore, the presence of varying writing styles in the many different news reports makes identifying useful information all the more difficult.

1.4 Research Contribution

The main aim of this thesis is the development of an automatic Crime Profiling System (CPS) for the Arabic crime domain to assist in crime analysis. The main original contributions are as follows:

- Automatically Constructing Dictionaries

  Crime type, location and nationality dictionaries are automatically constructed using the computational linguistic techniques developed from an unannotated corpus. These dictionaries will assist in the extraction of relevant crime information, instead of using manually built dictionaries, which are time consuming.

- Keyword-based Clustering

  The clustering technique used in this work is improved because only refined data containing meaningful keywords (gathered through the information extraction process) are inputted into it. As a result, a huge reduction in the quantity of data is achieved, consequently saving memory and reducing both the data loading and the execution time needed to perform the clustering. Therefore, the computational speed of the clustering technique is accelerated. Finally, the quantization error is also reduced, which results in higher clustering quality.

- Arabic Crime Corpus

  An Arabic Crime News Report Corpus (ACNRC) containing a huge collection of crime news reports, describing various types of crime incidents, is created. These news reports are collected from various sources in the Arab world, and as a result, there is diversity in terms of the writing styles employed for describing the incidents therein. This corpus will be made available online for other researchers to conduct further studies on the crime domain.

1.5 Research Methodology

The research methodology employed in this research is described and summarised in the following points:

- Background

To begin this research, the background, the broader context and related researches were considered by studying a great deal of the relevant literature. This aided in clarifying the motivations and in establishing the roadmap for the research as well as in exploring and determining the approaches used. For collecting the appropriate materials needed to cover this phase, several sources were used, such as Google Search, books, IEEE Xplore, ACM, CiteSeerX and SpringerLink.

- Data Collection

  Data related to crime incidents are collected in the early phase in order to investigate the crime domain and, later, to apply the text mining techniques. This step is achieved by compiling crime news reports from various Arabic newspapers published in different Arabic courtiers.

- Data Analysis Techniques

  Three data analysis techniques (frequency, collocation and concordance) are used in this research in order to better comprehend the language of the crime domain. This will assist in identifying the computational linguistic techniques needed for the information extraction task.

- Architecture

  The main aim of the early phase of this work is to design a model architecture that is appropriate for achieving the objectives of this research, which include, extracting meaningful crime information, clustering crime news reports based on different attributes, automatically building dictionaries, and using different types of visualization techniques. Therefore, the components of the proposed architecture are specified.

- Experiments

  The proposed architecture is tested on the collected data by performing real experiments in order to show how it is able to extract crime-related information, automatically build dictionaries, cluster similar documents, and carry out visualization tasks.

- Evaluation

  Precision, recall and F-measure are used to evaluate the efficiency of the proposed information extraction approach. Also, comparisons are conducted with other systems. For evaluating the clustering performance, four parameters are used: data size, loading time, execution time and quantization error.

1.6 Success Criteria

To measure the success of the proposed system, the research question will be answered through conducting experiments; these will measure the following points:

- The ability to extract crime-related information (crime type, location and nationality) from within Arabic crime news reports.
- The ability to automatically build dictionaries for crime type, crime location and nationality.
- The ability of the system to cluster crime reports with a high degree of performance.
- The ability to assist in crime data analysis by producing statistical information about crimes.
- A user can visualize useful information generated from unstructured textual data.

1.7 Thesis Structure

The rest of the thesis is organised into 6 chapters, as follows:

Chapter 2 provides an introduction to the Arabic language. Also, it contains descriptions of the well-established text mining techniques (information extraction and clustering) that are used in this research. Moreover, a literature review on research related to this work is provided; it justifies the approaches adopted in this research for the information extraction and clustering tasks.

Chapter 3 contains the data analysis phase, which includes frequency analysis, collocation analysis and concordance analysis, to investigate the crime context, i.e. discovering the behaviour of the words used in the crime language for describing crime incidents. Also, this chapter describes how the computational linguistic techniques are developed (based on special syntactic constructions).

Chapter 4 provides an overview of the proposed model architecture with descriptions of its stages, namely, the initial preprocessing, the information extraction, the intermediate preprocessing, the clustering and the visualization stages.

Chapter 5 presents the implementation of the proposed model using the newly collated data in order to test its effectiveness in terms of its ability to extract relevant information, automatically build dictionaries, and perform clustering and visualization tasks.

Chapter 6 provides the performance evaluation results as well as the limitations of the model.

Chapter 7 summaries this research and suggests future work.

Chapter 2

Background and Related Research

Objectives

- Provide an introduction to the Arabic language.
- Provide an overview of information extraction and clustering text mining techniques.
- Discuss related research and current approaches.
- Provide an overview of crime analysis.

2.1 Introduction

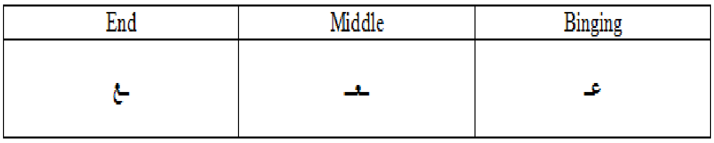
This chapter provides an introduction to the Arabic language and text mining field. A general background to the Arabic language, focusing on the syntactic information that is related to this research, is presented in section 2.2. The Arabic part of speech (noun, verb, particle), the case assignment methods and the types of sentence in the Arabic language, are explained in this section, concentrating on the transitive construction and genitive case. Section 2.3 provides an introduction to process of text mining, which includes data gathering, text preprocessing, text mining techniques and presenting the results. Also, the preprocessing components, i.e. tokenization, filtering, normalisation, stemming and document representation, as well as text mining applications are discussed in this section. Section 2.4 introduces the information extraction technologies available and discusses the different approaches adopted in literature. Furthermore, it critically reviews related works in the Arabic language within the context of information extraction, and related research in the crime domain. The difference between information extraction and information retrieval explained in this section. Section 2.5 provides an introduction to Artificial Neural Networks (ANNs), including its architecture (feed-forward and feedback networks) and learning paradigms (supervised and unsupervised learning) as well as an introduction to document clustering and related works using the Self Organisation Map (SOM) neural network. The structure of SOM and its algorithm is also explained.

2.2 Arabic Language

The Arabic language is a Semitic language, which is the native tongue in 22 Arab countries. Arabic consists of 29 letters that can be used to form words; other languages, such as Farsi and Urdu, also use Arabic characters [7]. Arabic words can be divided into three classes: noun (اسم / esm), verb (فعل / fe'al) and particle (حرف / hrf) [8, 9, 10]. However, when working with the Arabic language, some other important characteristics need to be taken into account [11]:

1. A character may have up to three different forms, each form correspond to the position of that character in the word (beginning, middle or end), such as the letter "ع / Ayn " in Table 2.1.

Table 2.1: Position of the character in the word



End	Middle	Binging
ـع	ـعـ	عـ

2. Arabic does not have capital letters; this characteristic represents a considerable obstacle to the Named Entity Recognition (NER) task because in other languages capital letters represent a very important feature.

3. Finally, it is a language with a very complex morphology because it is highly inflectional.

A linguistic study of Arabic words and grammatical structures will be required before extracting the most appropriate structures for common Arabic sentence forms within the crime domain. Hence, the use of the linguistic internal structures of Arabic sentences will allow us to identify logical sequences of words. As previously mentioned, the structure of Arabic can comprise of three categories: noun, verb and particle, and they are explained in the following sections.

2.2.1 Noun

This category in Arabic includes any word that describes a thing, idea, person or location, and is not related to tense (time) [8]. There are certain signs that can help in identifying nouns in Arabic, as described in one particular rhyme or verse by the grammarian Ibn Malik [12]:

بَالجر وَالتنوين وَالندَاء وَال *** ومسند لّئِسم تميّز حصل

Ibn Malik explained that if one of them is present, it is possible to classify that word as a noun; these signs are as follows:

- If a word accepts to be assigned the genitive case; the genitive case is achieved through a prepositional phrase or a 'construct state' (Idaafah structure) [13, 14]. Consider the following examples:

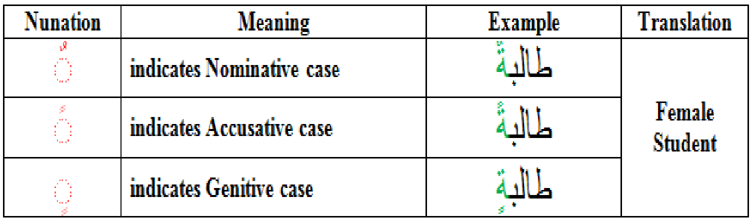
  1. ذهب الطّالب الَى المدرسة / dhab altalb-u ila almadrst-i.
     Went the student (nominative) to the school (genitive).
     The student went to the school.

  2. كتّاب المعلم / ktab-u almoalm-i.
     Book (nominative) the teacher (genitive).
     The teacher's book.

  Sentence 1 represents the first case (prepositional phrase), where the noun "المدرسة / almadrst / the school" (governee) follows the preposition "الَى / ila / to" (governor). As a result, the object of the preposition is assigned the genitive case. On the other hand, Sentence 2 represents the construct state (possessive case), which is comprised of two nouns. The first noun is called the construct head, annexe or possessor (in Arabic Mudaf), and it should not take nunation or definiteness signs, such as "ال / al / the". The second is called possessee or annexor (in Arabic Mudaf ilayh), and it is assigned the genitive case by the construct head (governor). The whole construct state inherits the (in)definiteness value of the possessee (mudaf ilayh) [15, 16]. In Sentence 2, the governor (construct head / mudaf) "كتَاب / ktab / book " assigns the genitive case to "almoalm / المعلم / the teacher". As a consequence, the construct head "كتَاب / ktab / book " inherits definiteness from "almoalm / المعلم / the teacher".

- If a word accepts a diacritic called nunation (the Arabic term is تنوين / tanwyn) at its end [15], then that word must be a noun; this type of diacritic is only accepted by nouns and it is not at all associated with verbs. However, the majority of the texts currently being published on the Internet, for news reports in particular, have no diacritics. Nevertheless, nouns in Arabic accept three types of nunation diacritic, as in the following Table 2.2.

Table 2.2: The three types of nunation diacritic



Nunation	Meaning	Example	Translation
◌ٌ	indicates Nominative case	طالبةٌ	Female Student
◌ً	indicates Accusative case	طالبةً	
◌ٍ	indicates Genitive case	طالبةٍ	

- If a word is preceded by the vocative particle "يَا / yea / O ", such as يَا محمد / O, Mohammad.

- If a word is fused at its beginning with the definite article ال / al / the.

2.2.1.1 Types of Noun

The Arabic language has two types of noun. The first type is the primitive noun (non-derived nouns); these are not derived, as in Table 2.3. The second type is the derivative noun, these are derived from verbs or other nouns, as in Table 2.4. For example, the derivative word "السَارق / alsarq / the thief" is derived from the root of the verb ”سرق / srq / thieve”. So, this word ”السَارق / alsarq / the thief” denotes an action (act of stealing) with no time specification.

Table 2.3: Samples of primitive nouns

Primitive noun	**English translation**
شمس	Sun
قمر	Moon
رجل	Man
بَاب	Door

Table 2.4: Samples of derivative nouns

Derivative noun	**Translation**	**Derived from**	**Translation**
المشمس	the sunny	شمس	Sun
مكنسة	sweeper	كنس	sweep
لَاعب	player	يلعب	play
السَارق	thief	سرق	thieve

Arabic nouns are inflected for gender (masculine and feminine) and number (singular, dual and plural) [17]. Also, nouns are either definite, which start with the article ”ال / al / the” or indefinite, having no ”ال / al / the” article at the beginning. Moreover affixes and clitics, such as some prepositions, conjunctions and possessive pronouns, can be attached to them. Clitics are subdivided into proclitic (located at the beginning of a stem) and enclitics (located at the end of a stem). For example, Table 2.5 shows the different morphological segments for the word ”وبدرجَاتهم” which means ”and by their grades”.

Table 2.5: Example for morphological segments

	enclitic	**affix**	**stem**	**proclitic**	**proclitic**
Arabic	هم	ات	درج	ب	و
pronunciation	hm	at	drj	be	wa
Gloss	their	s	grade	by	and

2.2.1.2 Case Assignment

The notion of case assignment has been discussed by researchers under the theory of Government and Binding (GB), which was coined by Chomsky. The concept of 'government' is defined as, "a particular structural relationship which may hold between two nodes in a tree. It plays a crucial role in GB in the assignment of case and in containing the distribution of empty categories" [13]. Habash et al. [14] defined case assignment as, "a relationship between two words: one word (the case governor or assigner) assigns a case to the other word (the case assignee)". In other words, the word has the power to affect the case of another word; this is called 'the governor', and which in Arabic is called "العَامل / alamil". The affected word is called 'the governee' or "المعمول / almamul". According to Habash et al. [14], there are different types of syntactic governors that assign cases to their governees. Consequently, all nouns and adjectives can be one of the following cases: nominative (مرفوع / mrfwo'a), accusative (منصوب / mnsob) or genitive (مجرور / mjrwr). Buckly [18], explained these three cases in detail, clarifying the conditions for each case; 7 conditions for nominative, 25 for accusative and 2 for the genitive case.

2.2.2 Verb

This word type points out an event or action. Arabic verbs have two tenses: perfect and imperfect. Verbs are inflected in terms of number (singular, plural and dual), gender (masculine and feminine), person (1st, 2nd, 3rd), voice (active and passive) and mood (subjunctive, indicative, jussive and imperative) [19].

Moreover, Arabic verbs are divided into two types: transitive or intransitive [20]. With respect to transitive verbs, a verb needs one object or more as well as the subject, in order for its meaning to be completed, e.g. "قطفت التفَاحة / qtft altfaht / I picked up the apple" [20, 21]. On the other hand, a sentence that contains an intransitive verb has no object, e.g. "مرض خَالد / mrd khaled / khaled got sick", i.e. this type of verb does need an object to complete its meaning. However, intransitive verbs can be transformed into transitive verbs either by changing the form of the word or by adding a preposition after the verb [20, 21]. For example, in "ذهبت إِلَى دبي / thhbt ela dubai / I went to Dubai", the verb is converted to transitive by adding the preposition "إِلَى / ila / to" after the verb. Alahmadi [22] described these types of verbs, such as "قَام / qam / did", "ذهب / dhb / went", "تورط / twr / involved", "عثر / athr / found", "تخصّص / tksas / specialised", "ادين / adyn / convicted", "شرع / shraa / commenced", "تعرض / taard / subjected", "اعترف / aatraf / confessed", "اقدم / aqdm / conducted" as 'transitive verbs by preposition'. Most Arab linguists state that most intransitive verbs cannot refer to the object of the sentence but they can be strengthened by certain prepositions, which are called transitive prepositions, such as "ب / bi / by", "ل / li / of", "من / min / from", "علَى / ala / on", "في / fi / in", "إِلَى / ila / to", in order to refer to the object. As a result, these prepositions can be described as governors because they control the status of the verbs, i.e. they assign the transitive case to verbs. This research concentrates on 'transitive verbs by preposition' in terms of exploiting them to extract patterns of interest as will be shown later in Chapter 3.

2.2.3 Particle

This class includes prepositions, conjunctions, interrogative particles, exceptions, and interjections. In other words, it includes the words that are not nouns or verbs, and sometimes these words are called function words. Detailed information about prepositions are here provided.

Prepositions are words that are used to connect other words in order to form meaningful sentences. Curme [23] defined a preposition as, "a word that indicates a relation between the noun, or pronoun it governs and another word, which may be a verb, an adjective, or another noun or pronoun". Quirk and Greenbaum [24] defined a preposition thus, "in the most general terms, a preposition expresses a relation between two entities, one being that represented by the prepositional complement". Prepositions are defined in 'classical' Arabic (the language of the Holy Qur'an) by Fitehi [25], as, "words that form with the noun phrases (or other linguistic entities) they govern exocentric constructions functioning as adjunct, prepositional object, predicate, postmodifier or as conjunctive of a relative pronoun".

The Arabic language has twenty prepositions, most of which are short. Most of them are formed from three letters, such as "علَى / ala / on" or from two, such as "في / fi / in", but they can be formed with only one Arabic letter and fused as a prefix with nouns, such as "ل / li / of" or "ب / bi / by" [26]. So, a preposition is either separate or attached. The separable prepositions are words that precede nouns, such as "في المدرسة / fi (preposition) almdrst (noun) / in the school". On the other hand, the inseparable prepositions are letters that are attached to the following noun, such as "ذهب ليلعب / dhb li-ylab / went for playing", where the preposition "ل / li / of" is fused with "يلعب / ylab / playing". Table 2.6 lists the prepositions with their pronunciations and meanings in the English language [17].

Table 2.6: List of the prepositions in the Arabic language

Preposition	**Pronunciation**	**Transliteration**
ب	bi	by, with, at, in
إِلَى	ila	to, for
ك	ka	like, as
ت	ta	for oath
و	wa	for oath
من	min	from
عن	'an	about, away from
مع	ma'	with
في	fi	in, at
عَلَى	ala	on, unto
منذ	munthu	since
حتَى	hatta	until
ل	li	of
عدَا	'ada	except
حَاشَا	hasha	except
خلَا	khla	except
متَى	mta	when
لعل	la'alla	perhaps

Sometimes, the prepositions are called genitive words because they add a verb's meaning to the noun, and they are also called in Arabic "الجر / aljr / dragging" because they drag the meaning of a verbs to the noun [27]. As a result, the task of the preposition in the language is to convey the meaning of a verb to the word that follows the preposition because some verbs are unable to reach nouns by themselves, such as intransitive verbs. For example; "ذهب الرجل الَى دبي/ dhab alrjl ila dubai / The man went to Dubai". In this example, the verb has a subject "الرجل / alrjl / the man" and a quasi- sentence (the prepositional phrase) "إِلَى دبي / ila dubai / to Dubai", which is the complement of the sentence "ذهبت / thhbt / I went"; this helps in determining the meaning of the whole sentence. Also, in the example "I went by car", the object of the preposition 'car' reveals the manner of movement. As a result, it can be said that prepositions are tools used to indicate meanings that are latent or implicit in verbs [27]. In other words, the verb's latent meaning is stimulated by the presence of a preposition [28]. Therefore, prepositions can work as links between some verbs and nouns, whether they are adjacent or not. Fitehi [25] stated that a preposition can occur in different positions within a sentence, as follows:

- Verb Noun Noun Preposition Noun.
- Verb Noun Preposition Noun.
- Verb Preposition Noun Noun.
- Preposition Noun Noun.

The structure of a prepositional phrase (PP) in Arabic, as in English, is composed of two parts: preposition and noun-phrase [29]. For example' "الولد في البيت / alwalad fi albyt" means "the boy in the house". In this example, the preposition is "في / fi / in" and the noun is "البيت / albyt". Also, Fitehi [25] stated that a phrase that is governed by a preposition in a sentence is called a prepositional complement and together they form a prepositional phrase. The prepositional complement can be one of the following types:

- a member of the noun class, e.g. "المدرسة / almdrast / the school" in "ذهب الَى المدرسة / dhab ila almdrast / went to the school" or a construction whose head is a member of the noun class, e.g. "مدرسة كبيرة / mdrst kbyrt / big school" in "درس في مدرسة كبيرة / drs fi mdrst kbyrt / studied in big school".
- a pronoun, e.g. "نَا / naa / us" in "الرجَال كذبوَا علينَا / alrjal kdbu alina / the men lied to us".

- maa clause, e.g. "ذهب بمَا لديه من مَال / dhab bi-maa ldyh min mal / He went with his money".
- anna clause, e.g. "ادعوَا بَانهم فقرَاء / adawo bi-annhm fuqraa / they pretended that the they are poor".
- a demonstrative pronoun, e.g, "hada / this / هذَا" in " استعرت الكتَاب من هذَا الرجل astart alktab min hada alrjul / I borrowed the book from this man".

According to Dukes et al. [30], a prepositional phrase must always be linked to a head node, either a verb or noun. Figure 2.1 shows an example of a prepositional phrase in a verbal sentence. The prepositional phrase "في سرقة / fi sarqt / in theft" is the object of the verb "تورط / twart / involved". The verbal sentence is comprised of three elements: Verb, Subject and Object (VSO), as can be seen in Figure 2.1.

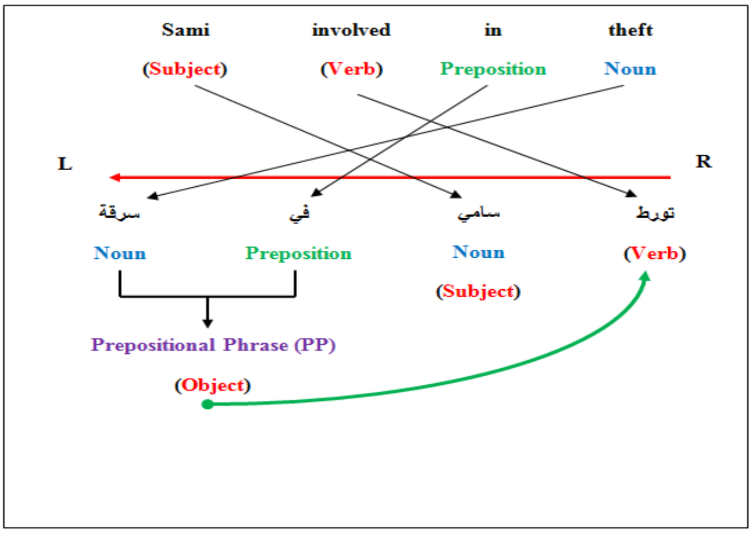


Figure 2.1: Prepositional phrase in a verbal sentence

On the other hand, a nominal sentence contains two parts, the first of which is called the subject phrase, in Arabic "mubtada / initial", and the second part is predication (or khabar / reporter in Arabic); the predication can be a prepositional phrase [31], as in Figure 2.2. In this example, the predication "في المدرسة / fi almdrst / in the school" comes after "الطَالب / altabl / the student" in order to clarify it.

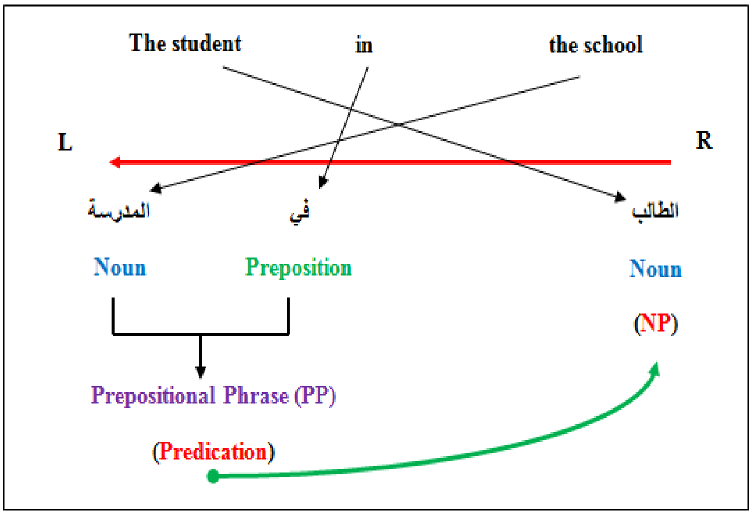


Figure 2.2: Prepositional phrase in a nominal sentence

Accordingly, prepositions play an important role in forming meaningful sentences and syntactic constructions for nouns and verbs. Prepositions govern verbs in terms of transitive or intransitive construct as well as governing nouns in terms of assigning them the genitive case.

2.2.4 Arabic sentence structure

Sentences in Arabic are divided into two types, as follows:

- Nominal sentence; this is also known as a verbless sentence, i.e. there is no explicit verb within the sentence. The structure of the sentence is comprised of two parts: a subject phrase and a predicate phrase [32]. According to Hadj et al. [10], a nominal sentence can start with a noun or a particle. Consider the following nominal sentence:

  البيت كبير / albyt kabyr / The house big

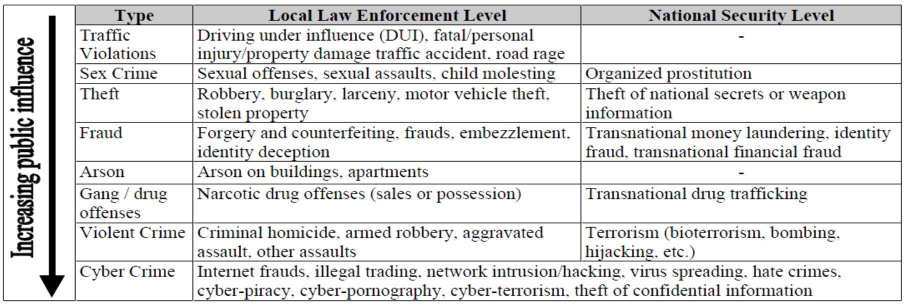
  The house is big

- Verbal sentence; a verbal sentence is one that contains a verb. The basic word order in a verbal sentence is Verb, Subject and Object (VSO), such as ”أَكل الولد التفَاحة” / akal (V) alwald (S) altufaht (O) / eat the boy the apple”. However, other orders can occur, such as SVO, e.g. ”الولد اكل التفَاحة / alwald (S) / akal (V) altufaht / the boy eat the apple” [33].

2.3 Crime Analysis

Crime is defined as ”an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law” [34]. The crime domain includes several types of crime, starting from civic crimes, such as drinking and driving, to international crimes, for instance, homicide by terrorists and these types could be changed over time [1, 35]. Table 2.7 presents different types of crime.

Table 2.7: Types of Crime [1]


Increasing public influence

Type	Local Law Enforcement Level	National Security Level
Traffic Violations	Driving under influence (DUI), fatal/personal injury/property damage traffic accident, road rage	-
Sex Crime	Sexual offenses, sexual assaults, child molesting	Organized prostitution
Theft	Robbery, burglary, larceny, motor vehicle theft, stolen property	Theft of national secrets or weapon information
Fraud	Forgery and counterfeiting, frauds, embezzlement, identity deception	Transnational money laundering, identity fraud, transnational financial fraud
Arson	Arson on buildings, apartments	-
Gang / drug offenses	Narcotic drug offenses (sales or possession)	Transnational drug trafficking
Violent Crime	Criminal homicide, armed robbery, aggravated assault, other assaults	Terrorism (bioterrorism, bombing, hijacking, etc.)
Cyber Crime	Internet frauds, illegal trading, network intrusion/hacking, virus spreading, hate crimes, cyber-piracy, cyber-pornography, cyber-terrorism, theft of confidential information	

2.3.1 Crime Analysis Definition

Crime analysis is new field in law enforcement, which has no standard definitions because the methodologies employed in crime analysis differ from agency to agency. In some police departments, crime analysis is used for producing crime statistics and for mapping crime for the benefit of command staff. In other places, crime analysis is used to help crime investigators by analysing various police reports and other important information. Osborne and Wernicke [36] stated, "the objective of most crime analysis is to find meaningful information in vast amounts of data and disseminate this information to officers and investigators in the field to assist in their efforts to apprehend criminals and suppress criminal activity". Gottlieb [37] defined crime analysis as "a set of systematic, analytical processes directed at providing timely and pertinent information relative to crime patterns and trend correlations to assist operational and administrative personnel in planning the deployment of resources for the prevention and suppression of criminal activities, aiding the investigative process, and increasing apprehensions and the clearance of cases". Within this context, crime analysis supports a number of departmental functions, including patrol deployment, special operations and tactical units, investigations, planning and research, crime prevention, and administrative services (budgeting and programme planning). The following are some important points that clarify the importance of crime analysis [37]:

- Crime analysis assists in providing useful information to law enforcement strategists, e.g. general and specific crime trends.
- Crime analysis can help both in detecting and preventing crime in advance.
- Crime analysis is necessary in order to exploit the untapped substantial information that is available in law enforcement agencies, the criminal justice system and the public domain.
- Crime analysis is required to meet the budgetary and logistical needs of law enforcement.
- Crime analysis helps in addressing crime problems locally and globally within and between law enforcement agencies.

2.3.2 Crime Analysis Types

There are various types of crime analysis, as described below, but the first three types are considered the main ones in the field of crime analysis. All types are as follows [37]:

1. Tactical Crime Analysis

   This type of analysis provides information for crime investigators in order for them to identify specific and immediate crime problems, and therefore, it offers police officers quick response times.

2. Strategic Crime Analysis

   This type of analysis is known as 'crime trends', in which frequencies of crime in terms of any increase or decrease within a particular area can be obtained. Thus, this type is used to study specific types of crime occurring in specific periods of time within a specified area.

3. Administrative Crime Analysis

   Administrative crime analysis is concerned with providing summary data, statistics and general trend information to departmental administrators, police officers, command staff, local and national government personnel, and the public. Consequently, exploiting official reports through this type of analysis assists in measuring crime. Moreover, administrative crime analysis is able to facilitate comparisons with regard to crime statistics in a particular city, such as the average for a particular type of crime relative to a previous year; this analysis can also be used to compare the crime ratios and trends of similar cities. Such analyses can be automatically implemented through modern technology, which serves to further motivate us to build a system that can be applied to reports written in the Arabic language.

4. Investigative Crime Analysis

   Investigative crime analysis is sometimes called 'criminal investigative analysis'; this is carried out by crime analysts to profile perpetrators, eliminate suspects (or otherwise), and identify potential victims through mining the available information on personal histories, behavioural patterns and criminal characteristics.

5. Intelligence Analysis

   Intelligence analysis concentrates on organised crime, terrorism and supporting specific investigations with information analysis.

6. Operations Analysis

   Operations analysis focuses on law enforcement agencies in terms of how they manage their resources, such as the allocation of grants or other sources of funding, personnel assignments, and day-to-day budgeting issues.

2.4 Text Mining

In recent years, the considerable growth in electronic free-text has led to massive volumes of unstructured textual data being available. According to Gupta and Lehal [38], over 80% of information is stored as text. A significant challenge has arisen as a result of this, affecting organisations and individuals, which is how to process this unstructured textual data to extract useful information. Consequently, text mining techniques and tools have now in great demand.

Text mining is also known as text data mining or Knowledge-Discovery in Text (KDT)[38, 39, 40]. It was defined by Delen and Crossland [39] as "a semi-automated process of extracting knowledge from a large amount of unstructured data". Also, Yuen-Hsien et al. [41] described text mining thus, "text mining, like data mining or knowledge discovery, is often regarded as a process to find implicit, previously unknown, and potentially useful patterns from a large text repository". Dörre et al. [42] described text mining as, "the same analytical functions of data mining to the domain of textual information, relying on sophisticated text analysis techniques that distil information from free-text documents".

It seems that text mining approaches are similar to data mining techniques; both of them mine in data [43]. The main difference between them is that data mining only deals with structured data, such as databases. Whereas text mining is applied to textual data, such as books and articles [38, 39, 44].

Text mining technology nowadays is important for most companies, organisations and governments, as it assists in handling their textual data. However, people communicate using natural language and they use it for recording information, but the difficulty then is: how can computers understand the natural language? Also, how can computers overcome simple language problems, such as spelling mistakes, which are so easy for humans to solve? The next section describes the processes of text mining that makes dealing with natural language possible.

2.4.1 Process of Text Mining

Figure 2.3 depicts the process of text mining, which can be divided into four phases, as follows [41, 42, 38]:

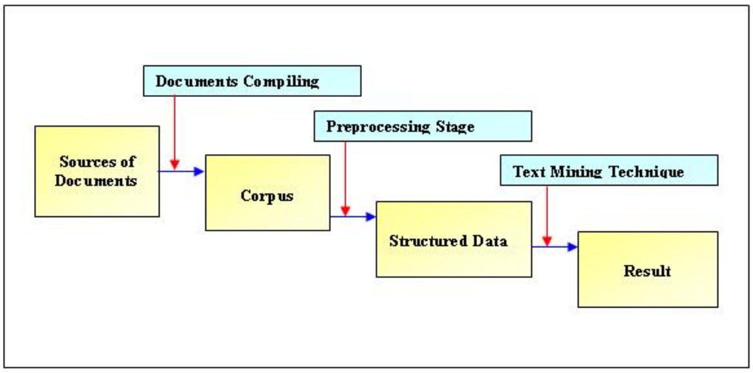


Figure 2.3: Process of text mining

- Document gathering

  This phase is responsible for collecting documents related to the context (specific domain) being studied from different sources. Therefore, the aim of this step is to generate a corpus. This phase is not so important for companies because their data (emails and files) are already available and continuously archived. Compiling documents can be performed manually or by using web crawlers. A web crawler, or sometimes called a web spider, is a program utilized to browse web pages in order to retrieve particular documents containing data of interest [45, 46]. Therefore, using these types of program saves a great deal of time and effort. The Wget program is example of web crawlers, available under GNU public license [47].

- Document Preprocessing

  In this phase, the textual documents that are obtained from the text gathering phase are prepared to make them in a standardised format for the text mining process. This phase consists of several stages, which include tokenization, filtering, normalisation and stemming; detailed information about each will be provided shortly.

- Text Mining Techniques

  In this phase, the text mining techniques, namely information extraction, topic tracking, summarization, categorization, clustering, concept linkage and question answering are applied.

- Results

  Knowledge is extracted from the previous phase, and this is then stored in formats ready for usage. Visualisation is implemented in this phase; a user can visualise useful information generated from unstructured textual data.

2.4.2 Text Preprocessing

Text preprocessing is required in order to optimise the performance of the text mining techniques [48]. Text preprocessing is comprised of several components, specifically, tokenization, filtering, normalisation, stemming and document representation [48, 49, 50, 51].

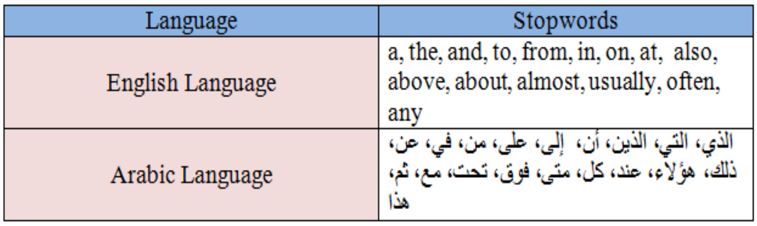
2.4.2.1 Tokenization

Text can be seen as a block or stream of characters. This process is necessary for other processes, such as part of speech tagging process to be able to deal with unstructured documents. According to Skiba [52], a token can be a single word, such as 'school', or a sequence of words, such as a named entity (e.g. United Kingdom). Also, it can be a sentence or even a paragraph. The Arabic language is similar to English in that they are both classified as segmented languages, i.e. they have some signs that assist in indicating segmentation boundaries; the words in both languages are separated by blank spaces, and therefore, this process can be directly applied to this type of language by exploiting the blank spaces between words as the 'delimiter'.

2.4.2.2 Filtering

Punctuation marks, commas, diacritics (in Arabic) and stopwords are all removed through this process. Stopwords (non-useful words) were introduced in 1958 by Hans Peter Luhn; they usually include prepositions, articles and conjunctions [53, 54]. These types of stopword do not have any impact on text mining and they could hinder the process. Thus, they are removed from the document representation, prior to any processing. Moreover, words that frequently occur in all documents in a given corpus are removed because they have no significant role to play in distinguishing the documents from each other. Also, words that appear in only a few documents are removed as well [55]. This process leads to reducing the dimensionality and size of the dataset, which would otherwise represent major obstacles to the text mining process [56]. Table 2.8 shows some of stopwords in English and Arabic Languages that are usually removed through this process.

Table 2.8: Some of stopwords in the Arabic and English languages



Language	Stopwords
English Language	a, the, and, to, from, in, on, at, also, above, about, almost, usually, often, any
Arabic Language	الذي، التي، الذين، أن، إلى، على، من، في، عن، ذلك، هؤلاء، عند، كل، متى، فوق، تحت، مع، ثم، هذا

2.4.2.3 Normalisation

In order to avoid or reduce data sparseness problem in the data being processed this process is implemented before applying the text mining techniques. In Arabic, it is possible to write 'America' in two different ways "أَمريكَا " (with hamza above) or "امريكَا" (without hamza); this case is a typographic variant. Also, there are spelling variants in Arabic, e.g. 'Beijing' can be written as "بكين" or "بيجين" [57]. Therefore, to make the data more consistent, this process is applied.

2.4.2.4 Stemming

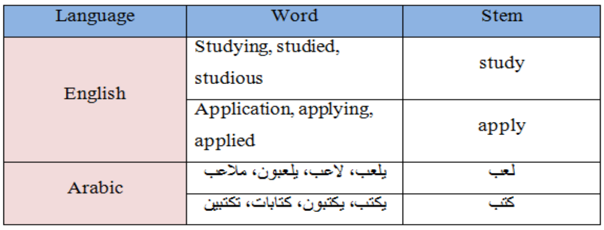
The stem of a word (the basic form of a word) can be obtained by eliminating the word's affixes. In other words, it is the process of reducing a word to its root or radix, and therefore, it reduces the total number of possible words. Inflectional languages, such as Arabic, contain words that have some morphological variants. Consequently, text analysis encounters problems when it deals with words that have the same or approximately similar meaning but with different spelling, e.g. play, played and playing. Therefore, all words that have similar meaning but different forms are transformed into their stem. According to Benajiba et al. [11], the common form of an Arabic word is:

Prefix(es) + Stem + Suffix(es)

In some cases, a word has no prefixes or suffixes and is therefore a stem, such as ”كتب / ktb / wrote” (singular), and in other cases, a prefix is added to the stem (e.g. stem كتب + prefix ي) to form the word ”يكتب / yktb / writing”. In the case of added a suffix to this stem (suffix وَا + stem كتب), the form of the word is now ”كتبوَا / ktbow / wrote” (plural), and in the case of an infix being added to the stem (infix ا + stem كتب), the word is changed to ”كتَاب / ktab / book”.

One well-known stemmer algorithm is the Porter stemmer, which is used on the English language [58]; another is the Lovins stemmer [59]. Also, there are some stemmers that have been specifically developed for the Arabic language, such as the Buckwalter stemmer [60] and the Khoja stemmer [61]. Table 2.9 shows some words in English and Arabic together with their roots.

Table 2.9: Some words in the Arabic and English languages with their stems



Language	Word	Stem
English	Studying, studied, studious	study
	Application, applying, applied	apply
Arabic	يلعب، لاعب، يلعبون، ملاعب	لعب
	يكتب، يكتبون، كتابات، تكتبين	كتب

Clearly, the word ’applying’ has been converted to its root ’apply’ and its suffix ”ing” has been removed. Also, in the Arabic language, one can see that the word ”يلعب / yla’ab / palying”, has been transformed into its stem ”لعب” by eliminating its prefix ”ي”.

According to Neto et al. [50], there is another method that can be used instead of filtering and stemming, which is called N-gram. N-grams are defined as ”sequences of characters or words extracted from a text” [62]. Also, Mhamdi [63] described an

N-gram as "a sub-sequence of N characters from a given sequence of characters". N-gram was firstly used to recognize the language of a document but it is a useful approach in many fields, such as computational linguistics [64]. It has been used in a wide variety of applications, such as spelling error detection, stemming, Optical Character Recognition (OCR), error correction, language identification, automatic text classification, protein classification and information retrieval [62, 65]. N-gram is categorised into two levels or types [62]:

- Character N-gram

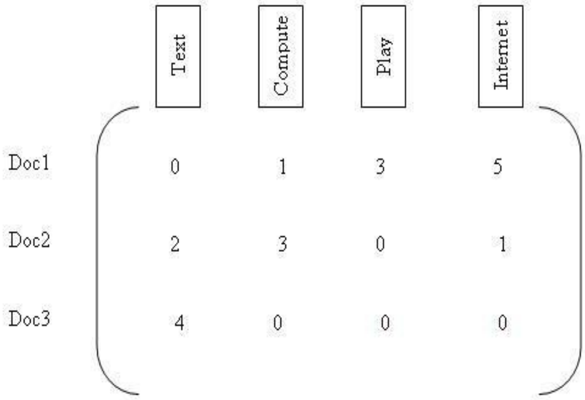
  Character N-gram is a set of N continuous characters taken from a given string, which have different lengths. The following are some types of character N-gram based on the size of N [62]:
  Unigram, N= 1
  Bigram, N= 2
  Trigram, N= 3
  Quadrigram, N= 4
  Depending on the aim of the application, the size of N is determined. For example, the bigram for the word 'crime' is as follows:
  -c cr ri im me e- , '-' sign refers to space.
  It can be seen that it moves character by character through the entire text. Through each move, a slice of character is extracted from the text.

- Word N-gram

  Mitra and Chaudhuri [62] defined word N-grams as "sequences of N consecutive words extracted from text". Therefore, word N-gram is appropriate for modelling language statistically.

2.4.2.5 Document Representation

Because the methods of classification or clustering are not able to directly process unstructured data, a document representation method is required to assist these techniques, so that textual documents can be handle effectively [40, 66]. Therefore, texts need to be transformed into an appropriate form, one that computers can process. There are many methods for representing free texts, such as Vector Space Model (VSM), Bag of Phrase, N-gram, and ontology based representation [66]. However, in the preprocessing phase, a typical document clustering process uses VSM [67]. Accordingly, this model is investigated because the clustering technique in this research uses VSM for document representation. VSM was introduced by Salton and McGill [68]. They used this model in experiments related to information retrieval. In this model, all documents are transformed into vectors, and this process is called the indexing or encoding phase, i.e. a set of words that are in each document, which can then be used to represent them. These vectors are grouped into one matrix. Therefore, VSM provides an efficient way of analysing a huge collection of documents as well as simple data structure [53]. Figure 2.4 depicts how each document is represented in the VSM.



$$\begin{array}{c|cccc} & \text{Text} & \text{Compute} & \text{Play} & \text{Internet} \\ \hline \text{Doc1} & 0 & 1 & 3 & 5 \\ \text{Doc2} & 2 & 3 & 0 & 1 \\ \text{Doc3} & 4 & 0 & 0 & 0 \end{array}$$

Figure 2.4: Vector space model

It can be seen that each single column in the matrix represents a word (also known as a term or feature). Thus, number of words in the whole document collection defines the size of the vector. Once all the textual documents have been represented by their words, all the frequencies of the words are generated. The next step is to calculate the weight of each word. In fact, this step is very important because the weight of a word reflects its importance in a document. Term Frequency - Inverse Document Frequency (TF-IDF) is used for weight calculation. The Term Frequency (TF) refers to the number of times a word occurs in a document. The Inverse Document Frequency (IDF) indicates the number of documents in which the word occurs. The following is a TF-IDF equation [66]:

$$TFxIDF(t_k, d_j) = Occ(t_k, d_j)xLog\frac{Nb\_doc}{Nb\_doc(t_k)}$$

Where:

- $Occ(t_k, d_j)$ represents the number of times the term t_k appears in the document d_j.

- Nb_doc indicates the total number of documents.

- $Nb\_doc(t_k)$ is a variable that refers to the number of documents in which the t_k term occurs.

In addition, there are many methods for calculating the similarity of documents. As already established, each document is considered to be a vector of weights. Consequently, the documents' similarity can be obtained by measuring the distances between these vectors. The following are two equations for calculating document similarity [66, 69]:

- The Euclidean distance: used for measuring the distance of the straight line between two points. It is based on the Pythagoras theorem. For example, if

A(x_1, y_1) and B(x_2, y_2) are two points, the Euclidean distance is obtained by:

$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

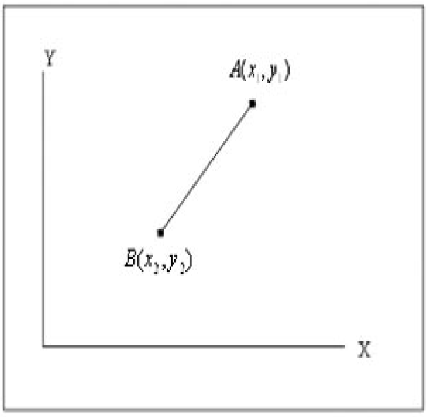


Figure 2.5: Euclidean distance

- The Manhattan distance: calculates the distance between two points, for instance, X(x_1, y_1),Y(x_2, y_2), which can be calculated as follows:

$d(A, B) =$ (x_1-x_2) + (y_1-y_2)

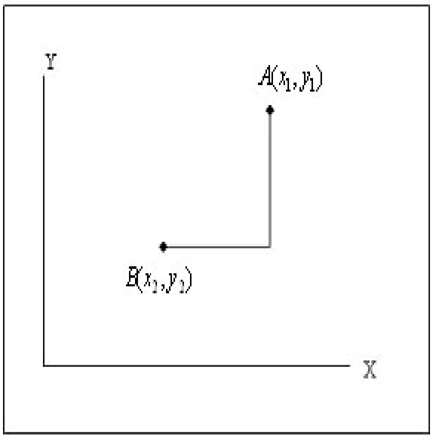


Figure 2.6: Manhattan distance

2.4.3 Text Analysis

A concordancer tool is type of program that assists researchers in analysing texts, i.e. it is able to present of all examples that contain a particular word or phrase being studied. It becomes important for lexicographers to investigate word different meanings. Also, it is useful for language learners when investigating the grammatical structure by studying output of concordance [70, 71].

The first concordance tool developed for Arabic language is called aConcorde created by Roberts et al. [72], which helps in analysing and presenting Arabic concordances in a proper manner . Also, LOLO which is a system for extracting statistical information from Arabic or English corpora developed by Almas and Ahmad [73], is used for conducting the analysis phase, i.e. frequency analysis, collocation analysis and concordance analysis, because it is able to manage, process and visualise collections of multilingual texts. However, there are another tools used to perform task of concordance, such as MonoConc [74], WordSmith [75] and Xaira [76] and for more information about performance of these tools against aConcorde in terms of retrieving concordance output from Arabic corpora can be found in [71]. Moreover, a web-based program called Sketch Engine is used to study the behaviour of words because it provides a number of language-analysis functions, such as concordancer [77].

2.4.4 Applications of Text Mining

There are many practical applications for text mining; business, security and so on. Governments apply text mining techniques for homeland security and law enforcement. After September 11, 2001, governments have endeavoured to detect and prevent any terrorist or criminal acts, and so they are strengthening their capabilities through improved data analysis [44, 78]. Coplink is a text mining tool that was developed by the University of Arizona Artificial Intelligence Lab for fighting crime and terrorism [78]. Also, the field of medicine, text mining has been used in designing drugs as well as in health management systems; it is also in managing and accessing patients' electronic medical records [44, 79, 80]. SAS Enterprise Miner and SAS Text Miner are examples of applications that are used for examining medical records [79]. Additionally, text mining technologies are used in the education environment, such as in citation analysis and plagiarism detection [44, 81]. Beagle is a tool for plagiarism detection that is based on text mining methods [81].

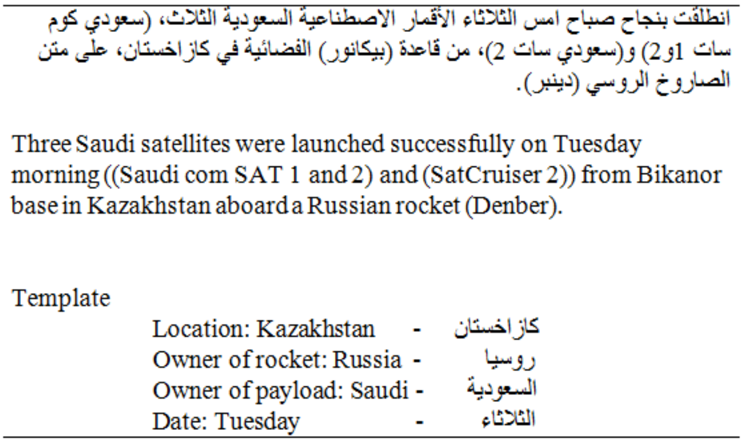
2.5 Information Extraction

As mentioned previously, with the Internet and the ongoing revolution in technology, the volumes of data in electronic form have become huge. As a result, a great deal of the information published on the Internet remains effectively hidden within large bodies of unstructured data. The challenge that needs addressing is how to find specific information about a particular subject inside these data. According to Noklestad [82], there are three different strategies that are suitable for dealing with the issues relevant to this subject, and they are as follows: information retrieval, information extraction and question answering. In this section, the information extraction approach is discussed in terms of its role in the text mining field.

According to Fan et al. [44], the basic process in analysing textual data is information extraction, and it is particularly useful when dealing with vast volumes of text. The beginning for extracting specific types of information from particular domains was in the late 1980s; the Defence Advance Research Project Agency (DARPA) initiated a series of Message Understanding Conferences (MUC). DARPA's MUC described information extraction as "a task involving the extraction of specific, well-defined types of information from natural language texts in restricted domains, with the specific objective of filling pre-defined template slots and databases" [83]. In MUC-1 (1987) and MUC-2 (1989), messages about naval operations were studied. MUC-3 (1991) and MUC-4 (1992) focused on news reports about terrorist events, and MUC-5 (1993) and MUC-6 (1995) investigated news articles about joint ventures and management changes [84]. In MUC-7 (1997), the task was to fill templates by identifying missile and rocket launch events from news articles published by the New York Times [84, 85, 86]. This task is called scenario template extraction. For example, from the following Arabic text about a rocket in Figure 2.7, the task is to extract some information in order to fill a template that includes three slots: for the location from where the rocket was launched, the rocket's owner, the owner of the payload, and the date. Table 2.10 presents the domains that have been studied by the MUCs.

Table 2.10: Domains studied by message understanding conferences

MUC	**Domain**	**year**
1	Naval operations	1987
2	Naval operations	1989
3	Terrorist events	1991
4	Terrorist events	1992
5	Management changes	1993
6	Joint ventures	1995
7	Missile and rocket launch events	1997



انطلقت بنجاح صباح امس الثلاثاء الأقمار الاصطناعية السعودية الثلاث، (سعودي كوم
سات 1و2) و(سعودي سات 2)، من قاعدة (بيكانور) الفضائية في كازاخستان، على متن
الصاروخ الروسي (دينبر).

Three Saudi satellites were launched successfully on Tuesday morning ((Saudi com SAT 1 and 2) and (SatCruiser 2)) from Bikanor base in Kazakhstan aboard a Russian rocket (Denber).

Template

Location: Kazakhstan - كازاخستان
Owner of rocket: Russia - روسيا
Owner of payload: Saudi - السعودية
Date: Tuesday - الثلاثاء

Figure 2.7: The outcome of information extraction for filling template slots

Borthwick [85] described a simpler information extraction task, called template relationships. For example, from the sentence ”US President Barack Obama”, the system should be able to extract ’Barack Obama’ as a person’s name, ’US’ as a country’s name and that Barack Obama is ’President’ of the US. However, the first step before identifying any relationship between the two entities is to categorise them into predefined classes, such as a person’s name or location. This task is called Named Entity Recognition (NER), which was introduced in MUC-6 [87, 88]. NER falls under the information extraction domain, and its mission is to identify entities from a text and to classify them into predefined categories, which include: personal names, locations, organisational names, dates, times, money and percentages [87]. So, the NER task can be divided into two sub-tasks: entity detection and entity classification [89]. Recently, two other disciplines related to NER have emerged: Automatic Content Extraction (ACE) and the Conference on Natural Language Learning (CoNLL). The former focuses on entity types, including person, organisation, location, facility, weapon, vehicle and geo-political entities, whereas, the latter concentrates on only four entity types: person, organisation, location, and miscellaneous entities [90]. According to Cunningham [91], the information extraction can be divided into five types: Named Entity recognition (NE), Coreference resolution

(CO), Template Element construction (TE), Template Relation construction (TR) and Scenario Template production (ST).

2.5.1 Information Extraction Approaches

Several approaches have been proposed for performing information extraction tasks, they can be divided into three categories, as follows:

- Hand-crafted rules, known as linguistic approaches.

  This approach is usually used for extracting information from specific domains. It is simple to build because its goal is only to fill out templates, i.e. it is not document understanding. Moreover, it can be trained on annotated or unannotated corpora [92]. Although it can achieve reasonable levels of success, there are some disadvantages [93, 94]. For instance, it is time consuming because it is slow to build, and it is difficult to scale to new domains [94].

  Toral [93] explained that this approach is applied using rules and gazetteers. The Gate system was developed at Sheffield University, and is a type of software that follows this approach. The task of this system is to extract named entities [95].

- Machine learning approaches

  Because of difficulties that faced researchers when building hand crafted rules, a need for automatically learning extraction rules emerged [96]. There are three types of machine learning approach: supervised, semi-supervised and unsupervised. According to Nadeau [87], most developed systems have been designed based on handcrafted rule based systems or supervised learning based systems. In both approaches, a corpus must be studied and analysed by hand to gain sufficient pertinent knowledge to build the rules or to feed the machine learning algorithms. However, the supervised learning techniques, which include Hidden Markov Models (HMM), Maximum Entropy (ME), Support Vector Machines (SVM) and Conditional Random Fields (CRF) need a large annotated corpus for designing the systems. As a result, this disadvantage of supervised machine learning led to the emergence of semi-supervised and unsupervised machine learning [87]. The semi-supervised (weakly supervised) is implemented with little supervision. The idea of this type of machine learning is to use a set of seeds to provide a system with a little external support to start learning how to extract. For example, finding names of the diseases to extract can be done by providing the system with seeds, such as five disease names. First, the system seeks out the sentences that contain these seeds in order to understand the contexts in which they appear. Then, the system tries to find other disease names that exist in the same context [97]. Nadeau [87], developed a semi-supervised NER technique that learned to recognize 100 entity types with little supervision. With regards to unsupervised machine learning, clustering is considered the most typical approach where, for example, named entities can be gathered based on the similarity of context from clustered groups [87, 97].

- Hybrid approach

  This approach combines the hand crafted recognition and machine learning methods.

2.5.2 Information Extraction vs. Information Retrieval

Information Retrieval (IR) is different from Information Extraction (IE). With regards to IR, this technique is intended to retrieve the most relevant documents from a set of documents for a specific query. In contrast, IE is the task of extracting specific types of information from documents. Figure 2.8 shows the difference between

IR and IE.

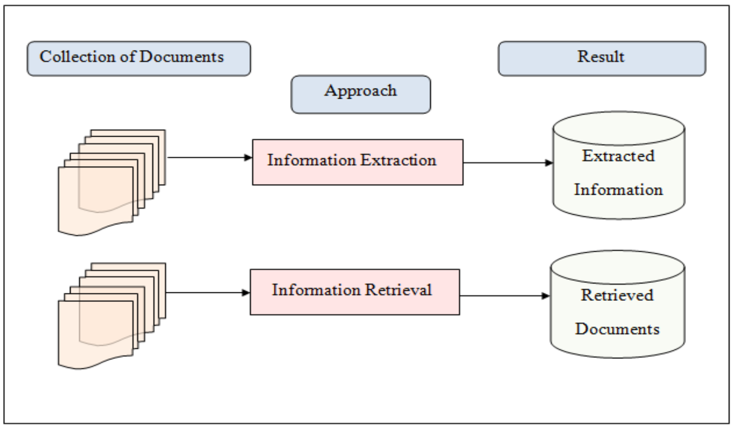


Figure 2.8: The difference between information extraction and information retrieval

2.5.3 Arabic in the Context of Information Extraction Task

Named Entity Recognition (NER) has had an influential role in developing various types of Natural Language Processing (NLP) systems, such as text clustering, Information Retrieval (IR), Question Answering (QA), Machine Translation (MT) and text summarisation, and has served to improve their performance [5, 98, 99, 100]. Although most research has been devoted to the English language, in the past few years, researchers have started paying attention to the Arabic language. Arabic textual data in electronic form has rapidly increased with more than 20,000 Arabic websites on the Internet and more than 300 million users [101]. Hence, there is a need for tools to deal with this type data, i.e. extracting useful information from Arabic text.

Most of the systems developed for Arabic in the literature focus on NER [5, 6]. The majority of systems developed for Arabic in this field rely on predefined proper name gazetteers. Maloney and Niv [98] developed a system called TAGARAB in order to recognize information relating to names, dates, times, and numerics within Arabic text. A combination of a pattern matching engine and morphological analysis as well as a words list is used to achieve the recognition task. The morphological analysis is used to assist the system in recognizing the various morphological word-shapes and to provide part of speech information for each token before entering the data into the pattern matching engine, i.e. a high precision morphological analysis is employed. The performance achieved for the aforementioned entities is presented in Table 2.11. Good results were achieved but it is an expensive process as each token is examined. Also, Mesfar [101] developed a system for the Arabic language to recognize proper names (person, location, organisation names), dates and numerics through a combination of morphological analysis, syntactic grammar and rules. The role of the morphological analyser is to strip off affixes from inflected words to assist in the matching process. Also, this system relies on gazetteers that contain the names of persons, locations and organisations, together with trigger words that indicate entities of interest, such as person's title. As in the TAGARAB system, this is an expensive process because the whole text must enter the morphological analyser. Furthermore, huge knowledge resources are used in this system, which include many predefined gazetteers, such as a personal names list that contains 12,400 Arabic first names and a locations list that consists of 5,038 entries, as well as a list of keywords that has 872 trigger words for indicating to these entities. The evaluation results of this system are listed in Table 2.11. Likewise, Shaalan and Raza [102] developed a system called Name Entity Recognition for Arabic (NERA) to extract 10 named entities from Arabic text; person name, location, company, date, time, price, measurement, phone number, ISBN and file name. They use a rule-based approach that relies on various fixed predefined dictionaries, such as for personal names (263,598 complete names, 175,502 first names and last names with 33,517 names), locations (4,900 names) and organisations (273,491 names of companies). Also, there is a dictionary containing trigger words (indicator words) for helping to identify entities, such as using job titles to indicate persons' names. Moreover, a dictionary called Blacklist is used to reject unwanted entities in order to filter the result. It is noticeable that gazetteers are extensively used but building them is time consuming because Arabic resources (corpora, gazetteers) are generally not free and can be hard to access; they are also relatively few in number. Table 2.11 presents the performance results achieved for this system in terms of precision, recall, and f-measure.

Moreover, Al-Shalabi et al. [8] presented an algorithm for extracting proper nouns from Arabic texts. They use a set of keywords and special verbs together with some specific rules. Firstly, they use predefined keywords to mark phrases that may contain proper nouns. Secondly, the proposed rules are applied to extract the proper nouns that directly follow the keywords and then the extracted words are classified into one of these categories, based on the type of the keyword: people, locations, organisations, events and products. However, it should be noted that the system is not able to extract proper nouns that do not appear directly after the keywords or the special verbs. Although they reported that they could extract 86.1% of proper nouns in a text, they evaluated their developed system using only 20 documents, which is a very small set of data and is perhaps insufficient for determining the effectiveness of the system. Nevertheless, the performance evaluation is presented in Table 2.11.

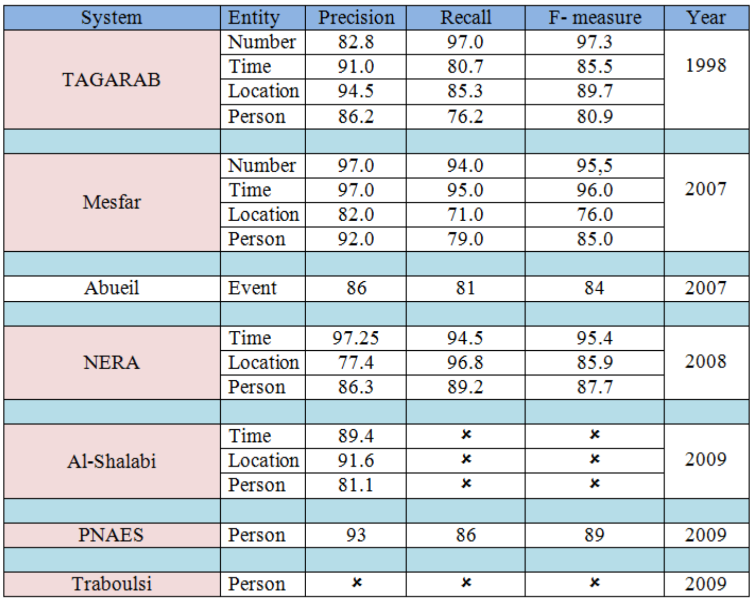
Elsebai et al. [103] adopted a rules-based approach that makes use of the outputs generated by the Buckwalter Arabic Morphological Analyser (BAMA) for developing the Persons Names Arabic Extraction System (PNAES). Their system uses a set of keywords (introductory verbs and words) to indicate the phrases that might contain person names, i.e. there is no predefined person names gazetteer. However, lists containing Arabic person names that start with the definite article "al / the", organisations and location names are used. The reason for using organisation and location names lists is to match the extracted word with words in the lists, and once the matching occurs, the word is discarded, otherwise it is classified as a person name, i.e. it is similar to the Blacklist used in [102] to perform filtering. Although this system is able to deal with names that appear not necessarily next to a keyword, unlike the above system developed by Al-Shalabi et al. [8], quite complicated rules are created to cover all the probabilities of a person's name occurring in a text. Finally, the performance achieved for the persons names recognition task was good but it was tested only on one resource, and therefore, the efficiency of this system was not rigorously examined. Table 2.11 presents the performance results.

Abulei and Evens [104] developed an events extraction and classification system for Arabic information retrieval systems. Their developed system is based on a predefined keyword lists and a parser to extract the events, the dates and related proper nouns. The lists include event words (such as assassinated), valid dates and proper nouns (people, organisations, locations). The system relies on 'direct look-up' to recognize the type of an event, i.e. if an event type is not in the list, it will not be identified.

Additionally, Abuleil [105] proposed an algorithm for scanning and understanding events (natural disasters, bombings and deaths) in Arabic text in order to extract information related to them, such as event locations, event types and dates. The system uses different lists that are described as event elements. For example, in order for an event to be marked in the text, special words (keywords), including nouns (e.g. earthquakes, hurricanes and massacres) and verbs, are used. A proper name list, consisting of person, location and organisation names, is used to extract entities from the text. Moreover, some particles and nouns are used as link tools when they occur in the text next to proper nouns and noun phrases, linking the event and its location. Also, some particles and nouns are used as relationship tools between more than one event in the text when they occur before the keyword of the second event. The evaluation results of the system were good, as in Table 2.11, however, the system was only tested on a corpus compiled from one source. Furthermore, the above lists were manually built, based on reading the texts, which is a key disadvantage of this system. In order to confirm this drawback, the system was tested twice; the first experiment showed that some keywords were missing, and some particles and nouns had to be added to the lists. It was reported that after updating the lists, in the second experiment, the system was not able to extract 28 events due to 9 missing keywords. This problem is caused by a lack of any deep contextual analysis, which must be done on huge Arabic texts (within the data collection phase). Also, the authors mentioned that once all the elements (keywords, noun phrases, special nouns and particles, and proper nouns) were evident appear in the description of the event, the system worked well, but this means that the system fails to detect and understand the event if one of element missing.

Piskorski et al. [106] developed a multilingual news event extraction system at the Joint Research Centre of the European Commission (JRCEC). The system was able to extract violent and natural disaster events from online news. They use pattern matching engine and a set of lexicons, which means that a rule-based approach is adopted. The event extraction grammar was originally designed to be applied on the English language but the technique has been extended to work on other languages, such as French, Italian and Arabic. With regards to the Arabic language, the Arabic news articles were translated into English using translation systems, and then they implemented the event extraction grammar. The evaluation results were not reported for this system.

Table 2.11: Precision, recall and F-measure of above systems



System	Entity	Precision	Recall	F- measure	Year
TAGARAB	Number	82.8	97.0	97.3	1998
	Time	91.0	80.7	85.5	
	Location	94.5	85.3	89.7	
	Person	86.2	76.2	80.9	
Mesfar	Number	97.0	94.0	95,5	2007
	Time	97.0	95.0	96.0	
	Location	82.0	71.0	76.0	
	Person	92.0	79.0	85.0	
Abueil	Event	86	81	84	2007
NERA	Time	97.25	94.5	95.4	2008
	Location	77.4	96.8	85.9	
	Person	86.3	89.2	87.7	
Al-Shalabi	Time	89.4	×	×	2009
	Location	91.6	×	×	
	Person	81.1	×	×	
PNAES	Person	93	86	89	2009
Traboulsi	Person	×	×	×	2009

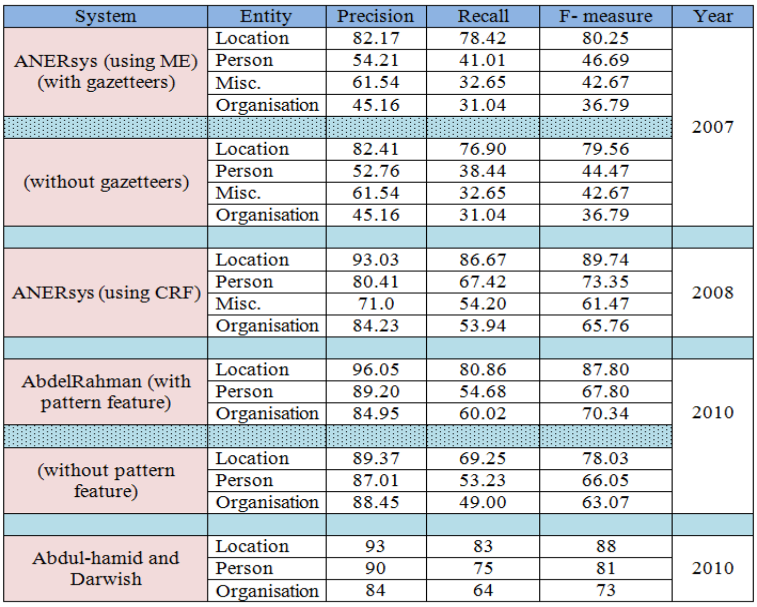
Clearly, all the above systems use the rule-based approach. Also, common entities (person names, locations, organisations, dates and numbers) have been investigated by these systems except the two systems developed by Abulei and Evens [104] and by Abuleil [105]; they tried to recognize events in texts. Additionally, the described systems did not mention the data sparseness problem in the Arabic language, except [102]. Furthermore, it seems that most dictionaries (gazetteers), especially the keywords lists in the aforementioned systems, were built based on authors' observations or knowledge. In other words, there is no objective explanation or analysis phase carried out on the data being studied for identifying the keywords. However, this phase (data analysis phase) was discussed by Traboulsi [107] in order to identify patterns of person names in Arabic texts. Three types of analysis (frequency, collocation and concordance) were conducted on huge corpora to identify the most important keywords, to discover the most frequent words collocating with keywords, and to obtain the concordance of the keywords. Consequently, the most frequent named entity structures were discovered, which led to the construction of a local grammar for recognizing person names, i.e. the other structures that may contain person names are discarded or neglected. As a result, the system might not be able identify of some named entities. Also, no performance evaluation was conducted for this system.

On the other hand, machine learning approaches have also been adopted in Arabic NER research. Benajiba et al. [11] designed the Arabic Named Entity Recognition System (ANERsys) based on Maximum Entropy (ME). Annotated corpora were used in this work as well as external resources such as dictionaries. Three different gazetteers were manually built: a location gazetteer consisting of 1,950 names of continents, countries, cities, rivers and mountains, a person gazetteer containing 2,309 names and an organisation gazetteer consisting of a list of 262 names of companies, football teams and other organisations. As mentioned earlier, it is time consuming to build gazetteers. Also, in order for the system to be tested, several experiments must be first performed to train it and to derive a set of features for assisting in the recognition process. Moreover, Benajiba and Rosso [108], changed the probabilistic model by using Conditional Random Fields (CRF) instated of ME. Although they achieved promising results through their improvement of ANERsys, their system still relies on an annotated corpus and the same predefined manually built gazetteers that were used with ME. Preprocessing, such as stemming as well as a part of speech feature is used. Also, they use a nationality feature for marking nationalities in the input text because nationalities are utilized in detecting the named entities; they are used as precursors to recognizing them. This feature relies on a dictionary of 334 different nationalities, which was manually built. Table 2.12 shows the evaluation results for ANERsys using ME and CRF. Also, Benajiba et al. [99] compared three machine learning approaches: Support Vector Machines (SVM), ME and CRF. The latter was the best and it yielded an overall f-measure of 83.

AbdelRahman et al. [57] integrated two machine learning techniques (bootstrapping semi-supervised pattern recognition and Conditional Random Fields (CRF)) for identifying 10 named entities: person, location, organisation, job, device, car, cell phone, currency , date, and time classes. However, the developed system, as with the above systems, relies on predefined gazetteers (person (3,228), location (2,183), organisation (403), job (70), device (253), car (223) and cell phone (184) to assist in recognizing the entities. Also, 16 different features are employed for implementing CRF as well as 232 different seeds. Table 2.12 lists the performance results of this system.

Abdul-Hamid and Darwish [6], created a system that is able to recognize named entities (person, location and organisation names) for the Arabic language, based on a set of features without using morphological or syntactic analysis or gazetteers. For implementing this work, Conditional Random Fields (CRF) was used. This technique was trained on a large set of surface features in order to avoid using Arabic morphological and syntactic features. Table 2.12 presents the evaluation results.

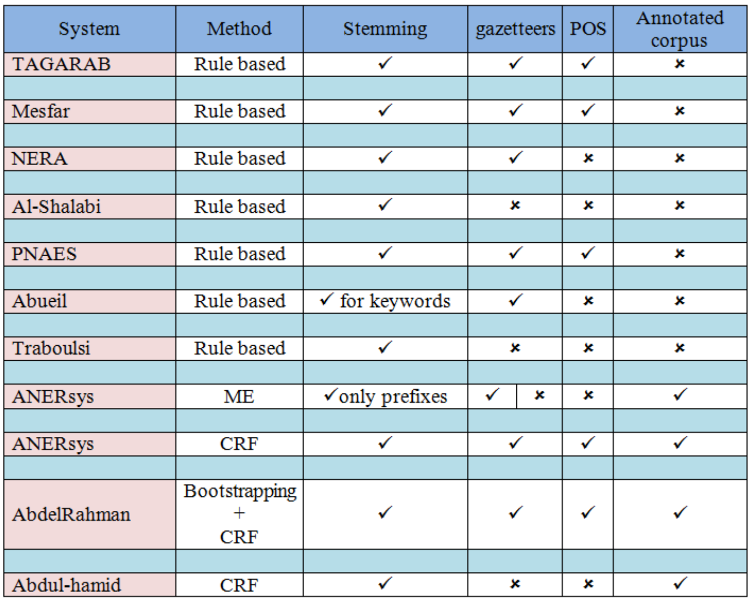
Table 2.12: Precision, recall and F-measure for machine learning systems



System	Entity	Precision	Recall	F- measure	Year
ANERsys (using ME) (with gazetteers)	Location	82.17	78.42	80.25	2007
	Person	54.21	41.01	46.69	
	Misc.	61.54	32.65	42.67	
	Organisation	45.16	31.04	36.79	
(without gazetteers)	Location	82.41	76.90	79.56	
	Person	52.76	38.44	44.47	
	Misc.	61.54	32.65	42.67	
	Organisation	45.16	31.04	36.79	
ANERsys (using CRF)	Location	93.03	86.67	89.74	2008
	Person	80.41	67.42	73.35	
	Misc.	71.0	54.20	61.47	
	Organisation	84.23	53.94	65.76	
AbdelRahman (with pattern feature)	Location	96.05	80.86	87.80	2010
	Person	89.20	54.68	67.80	
	Organisation	84.95	60.02	70.34	
(without pattern feature)	Location	89.37	69.25	78.03	
	Person	87.01	53.23	66.05	
	Organisation	88.45	49.00	63.07	
Abdul-hamid and Darwish	Location	93	83	88	2010
	Person	90	75	81	
	Organisation	84	64	73	

It is noticeable that the above machine learning approaches need an annotated corpus for them to be implemented. According to Ku [109], they also need large training data sets. Moreover, they often rely on different predefined gazetteers. Table 2.13 presents a comparison of all the above systems in terms of method, type of corpus and whether or not they use gazetteers, POS and/or stemming.

Table 2.13: A comparison between the system applied to Arabic text

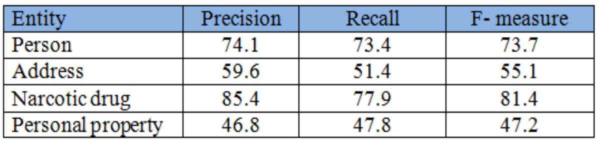


System	Method	Stemming	gazetteers	POS	Annotated corpus
TAGARAB	Rule based	✓	✓	✓	✗
Mesfar	Rule based	✓	✓	✓	✗
NERA	Rule based	✓	✓	✗	✗
Al-Shalabi	Rule based	✓	✗	✗	✗
PNAES	Rule based	✓	✓	✓	✗
Abueil	Rule based	✓ for keywords	✓	✗	✗
Traboulsi	Rule based	✓	✗	✗	✗
ANERsys	ME	✓only prefixes	✓ ✗	✗	✓
ANERsys	CRF	✓	✓	✓	✓
AbdelRahman	Bootstrapping + CRF	✓	✓	✓	✓
Abdul-hamid	CRF	✓	✗	✗	✓

With regards to the crime domain, there have been several efforts to develop information extraction systems for automatically extracting meaningful crime-related information. Data mining techniques have been used in this domain and a comprehensive survey of the effectiveness of the various methods for crime data analysis is provided by Thongtae and Srisuk [35]. However, these techniques are beyond the remit of this research.

In the text mining field, Chau et al. [4] studied police narrative reports to extract five meaningful entities, namely, person, address, vehicle, narcotic drugs and personal property in order to facilitate crime investigation. The developed system is comprised of hand-crafted lexicons, rule-based and machine learning. It begins with extracting noun phrases from documents based on linguistic rules. The extracted noun phrases are processed by a finite state machine to generate binary values. This process is achieved by conducting a matching process using predefined dictionaries for each word in the phrase and for the words that immediately precede and follow the noun phrase. Also, other sets of features (e.g. upper and lower case) are used in the process. The binary values are sent to the feedforward/backpropagation neural network component as input in order for it to predict the most likely entity type. The system was evaluated only on 36 documents collected from one source. Table 2.14 presents the evaluation results.

Table 2.14: Evaluation results in terms of precision, recall and F-measure

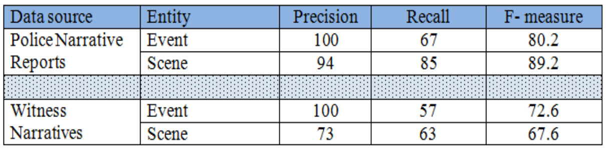


Entity	Precision	Recall	F- measure
Person	74.1	73.4	73.7
Address	59.6	51.4	55.1
Narcotic drug	85.4	77.9	81.4
Personal property	46.8	47.8	47.2

Also, Chau et al. [110] integrated the above system (entity extraction) with some data mining techniques, such as association and prediction methods to develop a crime data mining system.

Ku et al. [109] developed an information extraction system to extract crime-related information from different resources, such as police reports, newspaper articles and witness narrative reports written in the English language. Their system relies on rule-based and lexical look-up approaches. They manually built eighty-eight gazetteer lists. Also, they employed the Gate open-source framework, which includes several modules, such as tokenizer, sentence splitter, part-of-speech (PoS) tagger, noun chunk, and JAPE rules for pattern matching. Table 2.15 shows the evaluation results for extracting crime events and crime scenes.

Table 2.15: Evaluation results of extracting crime event and crime scene



Data source	Entity	Precision	Recall	F- measure
Police Narrative Reports	Event	100	67	80.2
	Scene	94	85	89.2
Witness Narratives	Event	100	57	72.6
	Scene	73	63	67.6

Moreover, Riloff [94] developed a program called AutoSlog for extracting information in the terrorism domain. The system relies on concept nodes in order to extract information about terrorist incidents. AutoSlog is comprised of 13 concept nodes, and each one is triggered by predefined keywords (terrorist action words), such as 'bombed' and 'kidnapped', and is activated within a specific linguistic grammar (e.g. in passive form) in order to extract information, such as targets, perpetrators and victims. Also, it relies on a corpus tagged by POS tagger.

Also, in order to support crime investigation, other attempts have been made by governments and companies. The European Commission funded the AVENTINUS project which can be described as a multilingual information extraction system for identifying these entities: persons, narcotics, location, organisations, transportation means, communication means and places and dates. AVENTINUS is designed for multilingual drug enforcement authorities, improving multilingual communication and information processing [111], and the rule-based approach is used for implementing the information extraction. Also, the Scene Of Crime Information System (SOCIS) was developed by a team from Sheffield and Surrey Universities in collaboration with four UK police forces (Surrey Police, Hampshire Constabulary, Kent County Constabulary and South Yorkshire Police) for crime scene photograph indexing and retrieval [112]. Moreover, the system performs information extraction using rule-based technique to extract all the names entities that might appear up in a caption: address, age, conveyance-make, date, drug, gun type, identifier, location, measurement, money, offence, organisation, person and time. The system achieved 80% precision and 95% recall. Both systems (AVENTINUS and SOCIS) rely on predefined gazetteers. Finally, the Locard Company developed an evidence tracking system (commercial software), which offers management for all crime-related exhibits; their website (www.locard.co.uk) provides more information about this software.

It can be inferred that each one of the above systems is focused on extracting specific entities. Also, different approaches have been used to implement the entity extraction, for example, Chau et al. [4] utilize machine learning, Ku et al. [109] use rule-based and lexical look-up, and the system developed by Riloff [94] is based on linguistic grammar.

2.6 Artificial Neural Networks for Clustering

Artificial Neural Networks (ANNs) are related to field of artificial intelligence [113]. ANNs seem to function as biological neurons; they are composed of a large number of interconnected artificial neurons, and as a result, they form a network of processing elements (nodes). Furthermore, these nodes are connected by weighted connections that look like synapses in the nervous system [114]. Malik [115] defined an ANN as ”a functional imitation of a simplified model of the biological neurons, and their goal is to construct useful ’computers’ for real-world problems and reproduce intelligent data evaluation techniques like pattern recognition, classification and generalization by using simple, distributed and robust processing units called artificial neurons”. Generally, the main aim of developing ANN algorithms is to mimic the biological neuron system in the human brain in terms of information processing and knowledge acquisition. In the human brain neurons are connected via sets of synapses, and these neurons can send and receive signals to each other via those synapses. These signals are central to the functioning of the brain, which thus depends on the data sent between the neurons [116].

2.6.1 Architectures of Neural Networks

The topologies of ANNs can be divided into two types, as follows [117]:

- Feed-forward Networks

  Figure 2.9 depicts an ANN with three layers. The first layer is the input layer, which has five input nodes for feeding data into the network. The second layer is a hidden layer, and the last layer is the output layer, which contains two output neurons. These layers are organised, and there are connections between the input layer and hidden layer, and between the hidden layer and the output layer. Thus, the nodes in the hidden layer are a link between the neurons in the input layer and the neurons in the output layer. Furthermore, the data is transmitted in only one direction, from input to output. The Perceptron network and the Adaline network are examples of feed-forward networks.

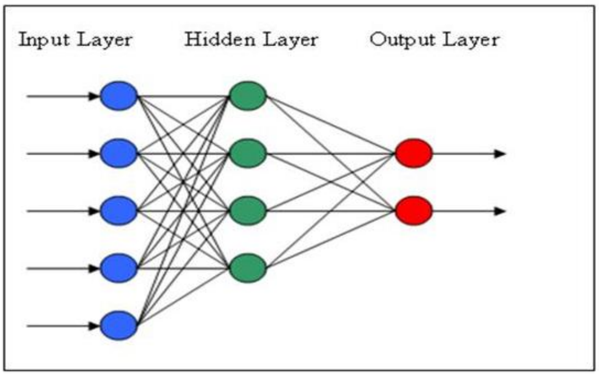
  

  Figure 2.9: Feed-forward network

- Feedback Networks

  These networks have feedback connections. It can be seen in Figure 2.10 that there is a connection between the output layer and the input layer. Moreover, signals can be sent in many directions, not in just one, as in the feed-forward topology. A node can be connected to any other node in previous layer or the next layer. The Jordan network, the Elman network, the Hopfield network and Kohonen nets are examples of feedback networks.

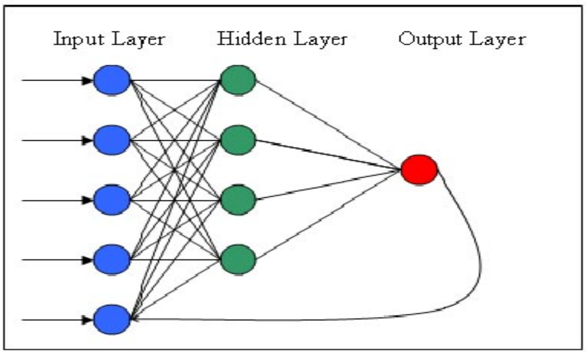


Figure 2.10: Feedback network

2.6.2 Paradigms of Learning

The learning paradigms (or machine learning) can be classified into two types, as follows [117, 118, 119, 120]:

- Supervised Learning

  In the case of supervised learning, the output is known beforehand. Consequently, the network requires an external teacher to train it to obtain the desired output. This process takes place by informing the output units of the most suitable response to the input signals. This type of machine learning is based on training data [109]. Many algorithms have been developed for supervised learning including Support Vector Machines (SVM), Latent Semantic Indexing (LSI), naïve Bayesian classifier and K-Nearest Neighbour (KNN).

- Unsupervised Learning

  With regard to unsupervised learning, there is no need to have an external teacher to train the network; the network learns by itself. In other words, it learns to identify patterns from a collection of a data without external intervention. Kohonen learning is known as Self Organising Map (SOM), and is an example of unsupervised learning.

2.6.3 Document Clustering

Fung et al. [121] defined document clustering as "an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters". Initially, it was developed to improve the precision and recall of information retrieval systems, and as a good method for locating the closest neighbours of a document [122]. Recently, it has been employed in other areas, such as in search engine for organising a user's query results, in web mining, and in document browsing [122, 123]. As mentioned earlier, a document representation model is required in order to assist in the clustering task. Once all documents are represented in the VSM, the clustering techniques can be applied. Clustering has been defined as the "unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)" [124]. Also Iiritano and Ruffolo [49] described the clustering methods as "techniques for partitioning a set of objects in non-overlapped groups (clusters) on the base of suitable similarity measures". Clustering techniques can be categorised into two types: partitional, where documents are divided into a collection of flat clusters (e.g. K-means) and hierarchical techniques that use hierarchical structures for document clustering (e.g. the Single-Link, and the Complete-Link methods) [67, 124]. Furthermore, the hierarchical techniques can be sub-divided into two types: agglomerative clustering (bottom-up) and divisive clustering (top-down) [51].

Moreover, the Self Organising Map (SOM) neural network is another technique for clustering documents. This technique is popular and widely used for many reasons;

it is simple to train the algorithm and it preserves the topology of the original input data. Also, it can reduce the dimensionality of the input data by mapping high dimension data onto a low dimension map. Hence, the high dimension data spaces can be visualized. Additionally, SOM has the ability for exploratory data analysis and it is able to perform data clustering without specifying the number of clusters [67, 125, 126, 127]. More information about the SOM clustering technique is provided in the next section.

2.6.4 Self Organising Map

Self Organising Map (SOM), Self Organising Feature Map (SOFM) or Kohonen map is a type of artificial neural network that was first proposed by Tuevo Kohonen [2, 128]. It can be described as a data visualization technique that reduces the dimensions of the data through the use of self-organising neural networks [129, 130]. SOMs have been used in various applications, including clustering, feature extraction, pattern recognition, data fusion, computer vision, biometric analysis, video coding, electromagnetism, forecasting tasks, image processing and visualization of complex data [131, 132, 133]. Lee et al. [134] applied SOM in the field of public health for improving a patient care and reducing health care costs. They developed a system for cancer diagnosis by identifying the relationships between cancer types and external factors from medical records. The results were satisfactory and it has the power to extract patterns and re-categorise clinical medical records. Senjyu et al. [126] proposed a new prediction scheme using SOM for next day load curve forecasting.

With regards to document clustering, Phuc and Hung [135] developed a graph clustering system using SOM to group similar documents and extract the main ideas from within them. Also, Freeman et al. [67] used Hierarchical SOM for document clustering as a 'tree view' of document topics. Kohonen et al. [136] used the WEB-

SOM method which is another type of SOM that is able to organise a collection of documents onto a graphical map display; this allows the user to explore and browse the documents.

In the crime domain, Chen et al. [137] used SOM to cluster and visualize crime-related information to assist crime analysis. On the other hand, in the literature, SOM has not been used in the Arabic crime domain. However, for Arabic texts, Al-Marghilani et al. [138] developed Multilingual Text Mining for Arabic-English Scripts (MLTextMAES), based on SOM in order to cluster similar documents in both languages.

2.6.4.1 Structure of Self Organising Map

SOM has many different structures but the popular architecture is composed of two layers of processing units [139]. These layers are the input layer and the output layer, as shown in Figure 2.11. These two layers are fully connected, in other words, all units in the input layer are connected to all output nodes (neurons).

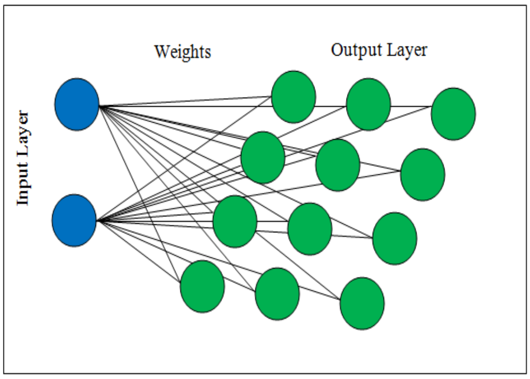


Figure 2.11: The architecture of the SOM

2.6.4.2 Algorithm of Self Organising Map (SOM)

The idea behind SOM is that it performs mapping for similar input vectors to similar areas of the output grid [126]. The following is the SOM algorithm:

1. Initialize weight randomly
2. Initialize neighbourhood ratio
3. Set input pattern
4. Calculate Euclidean distance
5. Find the winner neuron (smallest distance)
6. Update winner and neighbour weight neurons
7. Repeat Steps 3 to 7 until the convergence criterion is satisfied

As can be seen, this algorithm is iterative. The first step is to randomly initialize the weight vectors of the output map. At each iteration (training), a sample vector is randomly chosen from the input data. This phase is called the learning process or competitive learning. Through competitive learning, the Euclidean distance is calculated for choosing the Best Matching Unit (BMU). The wining neuron or BMU is the one most similar to the input pattern, that is, its weight is close to the input pattern. As a result, all neurons on the output layer enter into competition with each other. The neuron on the output layer that has the smallest distance to the input pattern is the winner. Once the winning neuron has been selected, its weight and the weight of its neighbours are both updated in order to make them more similar to the input pattern. This process is repeated with other documents until accurate results are obtained or the maximum number of iterations (epochs) is reached. [126, 127, 140, 141].

2.7 Summary

A background to the Arabic language has been presented along with a discussion of the challenges facing text mining techniques when dealing with Arabic text. Also, detailed information was provided about the text mining pre-processes that might have to be applied to Arabic text before uitlise any text mining techniques. Moreover, researches related to this thesis were also examined. The current approaches used for information extraction (applied to text in general and to the crime domain in particular) were reviewed in this chapter. The most common problems inferred from the previous works are as follows:

- The Arabic crime domain has not been studied, and therefore, there is no system developed for crime investigation.

- There is no available Arabic corpus in the literature for the crime context.

- Using predefined dictionaries containing person names, locations, and so on; it is well known that manually building dictionaries is time consuming.

- Lack of a deep analysis phase for extracting keywords because most authors use only their own experience to define the keyword list.

- Machine learning systems must use an annotated corpus for achieving information extraction. As a result, a PoS tagger must be used or the corpus must be manually tagged.

- Building and testing a system based on one data source is considered not sufficient because the system must be tested on a variety of datasets (collected from different sources) to evaluate the system's performance.

Also, related works that have used Self Organising Map neural networks for document clustering were described in this chapter.

Chapter 3

Syntactic Analysis for Crime Domain

Objectives

- Present how the data is analysed.
- Describe computational linguistic techniques.
- Show how the local grammar is constructed.

3.1 Introduction

This chapter describes the development of computational linguistic techniques for recognising and extracting crime-related information (crime type, location (scene) and nationality). A syntactic analysis for the Arabic crime domain is performed in order to identify the behaviour of the words used in the crime context. For implementing this intensive analysis, section 3.2 presents a large corpus, which contains news reports on various crime incidents collected from different sources. Section 3.3 presents the frequency analysis for the corpus in order to identify the most significant words within the frequency distribution of the first 100 words (after excluding certain types of words, such as prepositions and conjunctions). As a result, section 3.4 explores the crime action words that were identified by the frequency analysis through applying further analysis (achieved by using concordance and collocation analyses). Also, a transitive verb analysis is presented in this section as well describing the crime type local grammar. Section 3.5 describes the nationality local grammar (after analysing the nationality word) using collocation and concordance analyses. Finally, section 3.6 provides the location local grammar that was generated from analysing the words that are most often used for stating locations, using the same analysis techniques (concordance and collocation). LOLO, which is a system for extracting statistical information from Arabic or English corpora developed by Almas and Ahmad [73], is used for conducting the analysis phase, i.e. frequency analysis, collocation analysis and concordance analysis, because it is able to manage, process and visualise collections of multilingual texts.

3.2 Analysis Phase Corpus

For conducting the syntactic analysis phase, Arabic Crime News Report Corpus (ACNRC) were built by collecting a large data of Arabic news reports, describing various types of incidents that happened in different Arabic countries. We have compiled the corpus from the Alriyadh [142], Sabq [143], Okaz [144], Al-jazirah [145], Ahram [146], Almessa [147], Alrai [148], Alraimedia [149], Alqabas [150], Alamalyawm [151], Alwatan [152], Al-seyassah [153], Echoroukonline [154], Albayan [155], Raya [156] and Al-sharq [157] newspapers. The crime reports from these newspapers were saved in plain files with UTF-8 encoding. This ACNRC contains 502,609 tokens.

3.3 Frequency Analysis

All languages can be divided into two types: general language and special language (or sub-language) [73]; they can also be called open domain and restricted domain [158]. With regard to specialist language, every context or sector has its own, and so each specific text domain has its own special vocabulary and idiosyncratic syntactic structures. Hirschman and Sager [159] described sub-language as "the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical specialty or science field) in which the authors of the documents share a common vocabulary and common habits of word usage". Thus, recurrent expressions regularly appear in restricted language, and these repeated patterns are the key to processing textual documents in any specific domain [159]. According to Diekema et al. [158], systems that work on domain-specific texts have to use specific extraction methods.

Much like the English language, Arabic texts are comprised of two types of linguistic units. Firstly 'closed class' words such as prepositions, determiners and conjunctions, which indicate the natural language, and secondly, 'open class' words such as verbs, nouns and adjectives, which indicate the topic [160]. More open class words are used in specialist texts; these words are often distinct and occur more frequently. The use of open class words together with well-defined words may reveal sentences governed by local grammar, which is described as syntactic restriction [107].

Frequency analysis is able to present the most frequent words in the corpus being studied. Table 3.1 lists the ten most frequent words in the Arabic Crime News Report Corpus (ACNRC); most of these belong to the closed class and the first four most frequent words are prepositions. Additionally, the list includes these words: "شرطة / shurtat / police", " تم / tm / completed" and "القبض / alqbd / the arrest", which belong to the open class. A comparison between our special corpus (ACNRC) and a general corpus [73] containing around two million tokens was performed, and this comparison shows that six tokens, namely "من / min / from", "في / fi / in", "علَى / ala / on", "الَى / ila / to", "ان / anna / that" and "عن / aan / about", are common to both. Moreover, it shows that the closed words class is prevalent, whether the corpus is for the general domain or a special domain. Nevertheless, some prepositions, such as "ب / bi / by" and "ل / li / of" are extensively used in Arabic text but they did not appear through this analysis, whether in the special corpus or in the general corpus, because they are fused or attached to nouns; such words cannot be discovered in this step of the analysis.

Table 3.1: The frequency distribution of the first 10 most frequent tokens in both the special crime domain corpus (ACNRC) and the general corpus

ACNRC	**Translation**	**F**	**G.Corpus**	**Translation**	**F**
من	from	14914	في	in	119549
في	in	11473	من	from	73659
علَى	on	8896	ان	that	50048
الَى	to	5356	علَى	on	50022
ان	that	5152	الَى	to	28289
تم	completed	2406	وقَال	and said	25780
شرطة	police	2165	يوم	day	21633
القبض	arrested	2095	التي	which	20752
عليه	on	1944	عن	about	16428
عن	about	1871	اسرَاءيّل	Israel	14958

The open class words are rich in terms of the information that they carry because they reveal or indicate the topic of the document. Therefore, the first hundred words are selected after removing prepositions and conjunctions (closed class). Accordingly, some new words entered the list and became among the first 100 words. Table 3.2 lists the closed class words that were removed. Table 3.3 shows the most frequent words that are either in a noun or verb form.

Table 3.2: The removed closed class words

Word	**Translation**	**Word**	**Translation**
من	from	ذلك	that
في	in	كمَا	as
علَى	on	ثم	then
الَى	to	هذه	this
ان	that	اثنَاء	during
عن	about	وقد	may
احد	one	و	and
التي	which	لم	not
حيث	where	خلَال	through
مع	with	الَا	only
الذي	who	بين	between
بعد	after	غير	but
قبل	before	عدّ	number
مَا	what	او	or
حتَى	until	دون	without
آخر	another	اثر	after

Table 3.3: Frequency distribution of the first 100 words after removing closed class words in the ACNRC

Word	Translation	F	Word	Translation	F
تم	completed	2406	التحقيقاَت	the probes	469
شرطة	police	2165	القضية	the case	464
القبض	the arrest	2095	منزل	house	462
المتهم	the accused	1635	مبلغ	amount	461
رجَال	men	1634	تمكنت	able	460
الاَمن	security	1562	سياَرة	car	452
منطقة	area	1175	طريق	road	452
كاَن	was	1172	جدة	Jeddah	441
مدير	manager	1136	المحكمة	the court	438
التحقيق	the investigation	1038	الحاَدث	the Incident	435
الاَمنية	the security	1189	الفور	immediately	428
امس	yesterday	974	المنطقة	the area	426
قاَم	did	900	قضية	case	426
البحث	the search	887	اشخاَص	persons	421
الشرطة	the police	874	العقيد	the colonel	419
داَخل	inside	871	العصاَبة	the gang	416
النياَبة	public persecution	774	قتل	murder	412
محمد	Mohammad	733	افرَاد	members	407
التحرياَت	the investigations	726	وقاَل	and said	403
كاَنت	was	719	بسرقة	in theft	402
ضبط	detect	680	الجناَءيّة	the criminal	401
الجنسية	nationality	678	الوَاقعة	the incident	400
الرياَض	Riyadh	678	الاَمر	the issue	398
احدَى	one	672	عاَما	year	393
الوّاء	Major-General	664	الموقع	the location	392
الجناَءيّ	the Criminal	659	اعترف	confessed	391
الجاَني	the Criminal	653	الاَعلاَمي	the media	391
مركز	centre	635	رءيّس	head	391
العاَمة	the general	633	سنة	year	385
المجني	victim	615	تعرض	subjected	383
المباَحث	the investigator	607	بشرطة	in police	382
المتهمين	the accused	594	موَاطن	citizen	382
الاَول	the first	592	عثر	found	379
امن	security	587	زوجته	his wife	377
العاَم	the general	576	الناَطق	spokesman	374
محاَفظة	province	574	دورياَت	patrols	367
الف	thousand	558	الجهاَت	officials	359
رياَل	Riyal (currency)	558	وجود	presence	359
الجناَة	the criminals	532	الداَخلية	the interior	356
بمنطقة	in area	511	السياَرة	the car	356
محكمة	court	501	مستشفَى	hospital	351
تبين	clarified	500	قسم	section	348
مباَحث	criminal investigation	495	تمكن	could	339
العمر	the age	489	سنوَات	years	339
المخدرَات	the drugs	472	حي	district	338
شخص	person	333	مصدر	source	327
فريق	team	333	الاَشخاَص	the persons	319
معلومَات	information	332	وَالبحث	and the investigation	318
عملية	operation	329	الناَر	the fire	314
السرقة	the theft	327	مدينة	city	313

Some words within Table 3.3, are considered highly informative. These words can be divided into three groups based on the type of information that they convey or indicate. Table 3.4 lists crime action words. Table 3.5 presents the word ’nationality’, which is usually used to illustrate a person’s nationality in Arabic texts. Finally, Table 3.6 shows the words that are often used for stating locations.

Table 3.4: The crime actions words that are found in the 100 most frequent words

Word	**Translation**
قتل	murder
المخدرات	the drugs
بسرقة	in theft
السرقة	the theft

Table 3.5: The word that is used to illustrate a person’s nationality

Word	**Translation**
الجنسية	the nationality

Table 3.6: The words that are often used to illustrate a place name in a text

Word	**Translation**
منطقة	area
محافظة	province
بمنطقة	in area
مدينة	city

The nature of the event can be obtained directly through certain words, such as ” قتل / qtl / murder” and ” بسرقة / bi-srqt / in theft” in Table 3.4. Moreover, the word ” الجنسية / aljensyt / nationality” in Table 3.5 and the words related to location, such as ” منطقة / mntqt / area”, ” مدينة / madint / city” and ” محافظة / mohfdt/ province” in Table 3.6 are particularly useful for recognising and extracting nationality and crime location because these two entities often occur in the context of the previous words.

The most significant result from this frequency analysis is that the way crime reports are written indicates that the type of crime most usually relies on using nouns instead of verbs, such as 'theft', which appeared twice in Table 3.4; in the first case "بسرقة / bi-srqt / in theft", the preposition "ب / bi / in" is attached to it, and in the second, "السرقة / alsrqt / the theft", the definite article "ال / al /the" attached to it. Moreover, place names are in noun form because they are proper nouns. With regards to nationality, in the Arabic language there is type of adjective called 'nisba', which is used to denote pertinence, such as origin and nationality [161]. It is derived from a noun by adding "ي / iyy" in the masculine case or "ية / iyyt" in the feminine case, as a suffix to the noun [161, 162]. This type of adjective also exists in English, such as 'from Kuwait', from which the adjective 'Kuwaiti' can be derived, and from 'sun' the adjective 'sunny' is derived. In the Arabic, the nisba for "بريطانيا / brytanya / Britain" becomes "بريطاني / brytany / British" in the masculine case, and in the feminine case, it is "بريطانية / brytaniyyt/ British".

Therefore, these words can be considered as seeds for discovering the syntactic context of the event type, event location and nationality in order to identify the local grammar and then to build the indicator nodes for each case.

The following sections present the collocation analysis for the words in Tables 3.4, 3.5 and 3.6, i.e. identifying the most frequent collocation pairs for these words. This step is important because it can assist in determining the behaviour of these words within sentences. As a result, the most frequent words that collocate with the words in these tables are discovered as well as their syntactic construction. Moreover, concordance analysis is carried out in order to identify the structural patterns that contain crime type, crime location and nationality. Consequently, the dominant patterns used for stating the crime event, crime location and nationality are obtained.

3.4 Crime Type

This analysis presents the collocation of the crime action words in Table 3.4. These words often occur in the form of prepositional phrases and sometimes in the form of noun phrases. Because of the fact that a prepositional phrase is considered to be the complement of a head node (noun or verb), we investigated the head nodes that precede these prepositional phrases. It was found that the head nodes are transitive verbs, which are prevalent as will be shown. Moreover, with regards to the noun phrases that contain a crime type, the transitive verbs are also head nodes for them, as will be seen later. Therefore, the indicator nodes for extracting the type of crime are constructed based on transitive verbs. These verbs are used as keywords for triggering the indicator nodes when they are followed by specific prepositions, i.e. syntactic constraint is applied. As a result, the contexts of these verbs are analysed in order to identify the prepositions that always collocate with them. The analysis phase for the crime action words and transitive verbs is presented in the following sections.

3.4.1 Analysis of Crime Action Words

The following are the collocation and concordance analyses for the crime action words identified in the frequency analysis phase.

- سرقة / srqt / theft

Table 3.7 shows the collocation results for "سرقة / srqt / theft". It can be seen that the most frequent words are prepositions, namely "في / fi / in", "عن / 'an / about" and "من / min / from". These prepositions are considered the head nodes (governors) for this word because they occur immediately before it (Position -1), i.e. they assign the genitive case to the word "سرقة / srqt / theft", and consequently, they form prepositional phrases. Table 3.8 shows the results of another analysis, where only the most frequent words that directly precede the word ”سرقة / srqt / theft” are selected.

Table 3.7: Collocation results for ”سرقة / srqt / theft”

Theft		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	390	173	217	39	34	43	27	30	14	56	77	40	30
في	in	323	221	102	12	8	14	32	155	10	16	31	27	18
عن	about	109	86	23	3	8	3	8	64	6	4	6	2	5
علَى	on	86	67	19	15	14	11	5	22	0	2	5	5	7
بعد	after	46	28	18	4	6	10	1	7	0	4	4	8	2

Table 3.8: Collocation results of ”سرقة / srqt / theft” at Position -1

Theft		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	323	221	102	12	8	14	32	155	10	16	31	27	18
عن	about	109	86	23	3	8	3	8	64	6	4	6	2	5
قضية	case	32	31	1	0	0	0	0	31	0	0	0	1	0
من	from	390	173	217	39	34	43	27	30	14	56	77	40	30
علَى	on	86	67	19	15	14	11	5	22	0	2	5	5	7
جرَاءيم	crimes	24	23	1	0	0	0	1	22	0	0	0	0	1
جريمة	crime	23	23	0	0	0	1	0	22	0	0	0	0	0
عملية	operation	20	20	0	0	0	0	0	20	0	0	0	0	0

It can be seen that not only prepositions precede this word; there are also nouns, for example, ”قضية / qdyt / case”, ” جرَاءيم / jraam / crimes”, ”جريمة / jrymt / crime” and ”عملية / amlyt / operation”, and they together form noun phrases, such as ”جرَاءيم سرقة / jra’am srqt / theft crimes”, ”جريمة سرقة / jrymt srqt / theft crime ” and ”عملية سرقة / amlyt srqt / theft operation”. According to these results, the word ”سرقة / srqt / theft” often occurs in the form of a prepositional phrase or noun phrase because it is often preceded by a preposition or a head noun.

- قتل / qtl / murder

Table 3.9 shows the collocation analysis of the word ”قتل / qtl / murder”. The initial results are similar to the results for the above word (سرقة / srqt / theft) because the most frequent words again are prepositions, such as ”علَى / ala / on” and ”في / fi / in”; they occur directly before it (Position -1) 111 and 25 times, respectively. Also, the noun ”جريمة / jrymt / crime” appears to collocate frequently with ”قتل / qtl / murder”; 74 times at Position -1.

Table 3.9: Collocation results for ”قتل / qtl / murder”

murder		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
علَى	on	147	124	23	2	6	3	2	111	6	4	0	8	5
في	in	112	61	51	7	5	6	18	25	7	17	8	8	11
من	from	102	49	53	19	13	8	3	6	0	15	5	14	19
جريمة	crime	77	76	1	0	0	2	0	74	0	0	1	0	0
بن	son	61	1	60	0	1	0	0	0	0	26	4	25	5

Table 3.10 presents a further analysis, where only the most frequent words that immediately occur before the word ”قتل / qtl / murder” are chosen. It confirms the above analysis in that the prepositions ”علَى / ala / on” and ”في / fi / in” often precede this word (Position -1). As a result, they form prepositional phrases because the prepositions are the heads of these phrases. Moreover, the words ”جريمة / jrymt / crime of” and ”قضية / qdyt / case of”, already seen to collocate with ”سرقة / srqt / theft”, also precede ”قتل / qtl / murder”, and thus they form noun phrases.

Table 3.10: Collocation results for ”قتل / qtl / murder” at Position -1

murder		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
علَى	on	147	124	23	2	6	3	2	111	6	4	0	8	5
جريمة	crime	77	76	1	0	0	2	0	74	0	0	1	0	0
في	in	112	61	51	7	5	6	18	25	7	17	8	8	11
فَى	in	25	22	3	2	0	0	1	19	0	1	1	0	1
قضية	case	17	17	0	2	0	2	2	11	0	0	0	0	0

- بسرقة / bi-srqt / by theft

The word بسرقة / bi-srqt / by theft” is in reality a prepositional phrase because the preposition ”ب / bi / by” is fused with the word ”سرقة / srqt / theft”, and together they form a single word. Clearly, this word is always in prepositional phrase form. As a result, the collocations for this word in Table 3.11 reveal that no other prepositions can affect it because it is already influenced by the attached preposition. It can be seen that the three prepositions ”من / min / from”, ”علَى / ala / on” and ”في / fi / in” do not occur in Position -1.

Table 3.11: Collocation results for ”بسرقة / bi-srqt / by theft”

By theft		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	225	72	153	25	24	16	7	0	1	57	53	29	13
علَى	on	65	55	10	11	17	27	0	0	0	0	3	5	2
في	in	51	17	34	8	3	4	2	0	0	4	13	11	6
قَام	did	37	37	0	2	2	0	10	23	0	0	0	0	0
احد	one	28	12	16	3	3	3	3	0	2	3	2	6	3

It can be deduced that the three crime action words ”سرقة / srqt / theft”, ”قتل / qtl / murder” and ”بسرقة / bi-srqt / by theft” mostly appear in the form of prepositional phrases and sometimes in the form of noun phrases. This confirms our observation regarding the style in which these texts are written, describing event types by using nouns instead of verbs. Therefore, the occurrence of these words, mostly in the form of prepositional phrases, is exploited. As mentioned in Chapter 2, the prepositional phrase is preceded by a head node, which is a noun or verb. As a consequence, the verbs that occur in collocation with the previous words were identified, and Table 3.12, Table 3.13 and Table 3.14 show the most frequent verbs appearing in collocation with ”سرقة / srqt / theft”, ”قتل / qtl / murder” and ”بسرقة / bi-srqt / by theft”, respectively.

Table 3.12: Most frequent verbs found in collocation results for ”سرقة / srqt / theft”

theft		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
تخصّصت	specialised	34	34	0	0	0	1	33	0	0	0	0	0	0
اعترف	confessed	18	18	0	4	9	4	1	0	0	0	0	0	0
تخصّص	specialised	17	17	0	0	1	2	14	0	0	0	0	0	0
تعرض	subjected	14	12	2	1	5	4	2	0	1	0	1	0	0
اعترفوَا	confessed	13	13	0	4	2	3	4	0	0	0	0	0	0
تورط	involved	13	12	1	2	3	3	4	0	0	1	0	0	0

Table 3.13: Most frequent verbs found in collocation results for ”قتل / qtl / murder”

murder		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
اقدم	conducted	41	40	1	0	3	10	27	0	0	0	1	0	0
قَام	did	12	2	10	0	1	0	1	0	0	1	5	4	0
شهدت	had	17	17	0	7	6	3	1	0	0	0	0	0	0
اقدمت	conducted	9	8	1	0	1	0	7	0	0	1	0	0	0

Table 3.14: Most frequent verbs found in collocation results for "بسرقة / bi-srqt / by theft"

by theft		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
قَام	did	37	37	0	2	2	0	10	23	0	0	0	0	0
اعترف	confessed	25	24	1	1	0	5	7	11	0	0	0	0	1
قَاموَا	did	17	17	0	0	0	0	0	17	0	0	0	0	0
اعترفوَا	confessed	17	17	0	0	2	1	0	14	0	0	0	0	0
قَامَا	did	16	16	0	0	0	1	0	15	0	0	0	0	0
وقَام	and did	15	15	0	0	0	0	1	14	0	0	0	0	0

Table 3.12 and Table 3.13 show that the verbs mostly occur on the right side of the tokens "سرقة / srqt / theft" and "قتل / qtl / murder", occurring most often at Position -2. This result indicates that there is separation between these verbs and "سرقة / srqt / theft" and "قتل / qtl / murder". Contrarily, all the verbs in Table 3.14 occur immediately before "بسرقة / bi-srqt / by theft", i.e. at Position -1, because the preposition "ب / bi / by" is fused with the word. As a result, there is no separation. However, all these verbs are transitive (with prepositions) except the verb "شهدت / shhdat / had". This fact clarifies that the separation that occurs at Position -1 in Table 3.12 and Table 3.13 in front of the verbs is because Position -1 is for the prepositions that link these verbs to their complements. Also, the results of this analysis show that the collocations of the crime action words share the same verbs, such as the verb "اعترف / aatraf / confessed", which occurs in Table 3.12 and Table 3.14. As a consequence, this result provides a strong motivation to further study the context of these transitive verbs.

3.4.2 Analysis of Transitive Verbs

As shown above, our study of the Arabic Crime News Report Corpus (ACNRC) has led us to identify the characteristics of the language used (crime language). The most significant feature is that prepositional phrases are used for describing the type of offence, and a secondary one is that transitive verbs are used. In the Arabic language, each transitive verb (transitive verb by preposition) has a dependency relationship with specific prepositions, and together they form a grammatical structure called a transitive construction, as already discussed in Chapter 2. Moreover, transitive verbs do not accept certain prepositions; of the prepositions they do accept, the context decides the most appropriate one. The verb is considered the head node of the prepositional phrase, and the prepositional phrase is considered the complement of the verb. As mentioned before, the role of the preposition is to convey the meaning of the verb to the words that follow because some verbs are unable to reach the nouns (object/complement) by themselves. However, as we mentioned earlier, the aim of this analysis is to determine the linguistic behaviour of each verb in order to identify its associated preposition or prepositions in this current domain (crime domain).

Table 3.15 shows the transitive verbs that are to be analysed. The verb ”قَام / qam / did”, ”اعترف / aatraf / confessed”, ”تعرض / taard / subjected” and ”عثر / athar / found” already appear in the list of the 100 most frequent tokens in the ACNRC. Moreover, ”قَام / qam / did”, ”اعترف / aatraf / confessed”, ”تعرض / taard / subjected”, ”اقدم / agdama / conducted”, ”تورط / twrt / involved” and تخصّصت / takasasat / specialised” have already been seen in the collocation analysis above; these verbs were addressed in terms of their occurrence in the general corpus. It is apparent from the table below that 6 prepositional verbs out of the 9 occur in the special corpus (ACNRC) more frequently than they do in the general one. This result reflects their importance and influence in the crime domain (crime language). Alahmadi [22] identified these verbs together with the prepositions most often associated with them.

Table 3.15: Frequency distribution of the transitive verbs by prepositions in the special corpus (ACNRC) and General corpus

Verb	Pronunciation	Translation	ACNRC	G.Corpus
قَام	qam	did	901	321
اعترف	aatraf	confessed	391	68
تعرض	taarrada	subjected	384	433
عثر	athar	found	381	332
اقدم	agdama	conducted	179	43
تورط	twrt	involved	64	84
تخصّصت	takasasat	specialised	47	0
أَدين	adyn	convicted	14	26
شروع	shoroaa	commenced	41	5

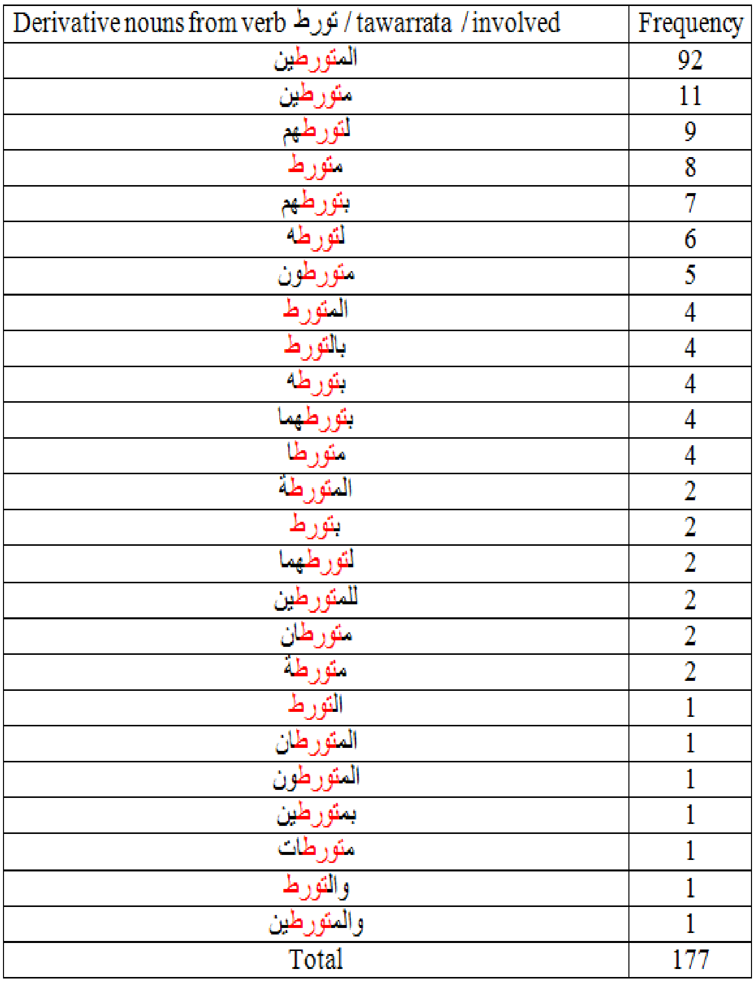
As mentioned earlier, Arabic verbs are also inflected in terms of number (singular, plural and dual), gender (masculine and feminine), person (1st, 2nd, 3rd), voice (active and passive) and mood (subjunctive, indicative and jussive). These verbs are generally in the past tense but they can be written in different forms, based on the context, as in the Table 3.16.

Table 3.16: A sample of different forms for the verbs

Verb	Translation	Inflected word
اعترف	confessed	اعترفوَا ، اعترفت ، اعترفن ، اعترفَا
تورط	involved	تورطت ، تورطوَا ، تورطن ، تورطهن
تخصّص	specialised	تخصّصَا ، تخصّصوَا ، تخصّصن ، تخصّصهن

Moreover, nouns can be derived from verbs and these are called derivative nouns. They too are inflected for gender (masculine and feminine) and number (singular, dual and plural). Accordingly, it has been found that there are indeed some nouns derived from the above prepositional verbs. For example, the verb "تورط / twrt / involved" can be transformed into a noun in different forms. Table 3.17 shows the different forms of the nouns that are derived from this verb, together with their frequency in the corpus. The total frequency of these nouns is 177, so the occurrence of all the forms of the verb "تورط / twrt / involved", which includes the different forms of nouns and verbs, is 378. The reason for involving derivative nouns in this research is because in the Arabic language each derivative noun follows its original verb, i.e. if a verb has a dependency relationship with a preposition, its noun also has a relationship with the same preposition. In consequence, this assists in extracting offence types from the texts.

Table 3.17: Frequency distribution of the inflected form derived from the verb "تورط / twrt / involved"



Derivative nouns from verb تورط / tawarrata / involved	Frequency
المتورطين	92
متورطين	11
لتورطهم	9
متورط	8
بتورطهم	7
لتورطه	6
متورطون	5
المتورط	4
بالتورط	4
بتورطه	4
بتورطهما	4
متورطا	4
المتورطة	2
بتورط	2
لتورطهما	2
للمتورطين	2
متورطان	2
متورطة	2
التورط	1
المتورطان	1
المتورطون	1
بمتورطين	1
متورطات	1
والتورط	1
والمتورطين	1
Total	177

In the following, the context of the selected prepositional verbs is investigated. The analysis of verbs will examine the words that follow the verb being studied. Thus, the most frequent words occurring on the left side of the verb, i.e. following the verb, are the target of this analysis to see if they contain nouns that might indicate a type of crime. Additionally, this analysis includes investigating which preposition most often follows each transitive verb, i.e. the associated preposition that has a dependency relationship with the verb. The length of the analysis (collocation span) from the verb is 5 words to the right and 5 words to the left. The reason for this is to distinguish between common words that may appear on both sides and significant words that mostly occur on the left side of the verb, i.e. following it. The analysis for these verbs are as follows:

- قَام / qam / did

According to Alahmadi [22], this verb can associate with many different prepositions, and Table 3.18 shows the collocation results for "قَام / qam / did". The five most frequent words collocating with this verb are selected. As can be seen, three words are prepositions: "عَلَى / ala / on", "من / min / from" and "فِي / fi / in". The highest occurrence appearing immediately after the verb (Position +1) is the preposition "عَلَى / ala / on". However, this type of verb can also associate with the preposition "ب / bi / by". This preposition is attached to the noun and they form a prepositional phrase as a single word. In other words, it is fused to nouns and it becomes their prefix. So, the most frequent words that start with this preposition are selected and presented in Table 3.19.

Table 3.18: Collocation results for "قَام / qam / did"

Did		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	236	96	140	26	23	23	13	11	4	36	37	30	33
علَى	on	153	65	88	17	14	17	17	0	18	19	18	17	16
حيث	where	120	112	8	6	3	3	1	99	0	0	2	3	3
ان	that	119	109	10	6	13	19	30	41	0	0	0	6	4
في	in	103	44	59	13	13	15	3	0	2	12	13	12	20

Table 3.19: Most frequent words associated with preposition "ب / bi / by" found in the collocation results for "قَام / qam / did"

Did		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
بقتل	killing	43	2	41	1	1	0	0	0	16	10	10	2	3
بسرقة	theft	37	0	37	0	0	0	0	0	23	10	0	2	2
بهَا	with	31	2	29	0	0	0	2	0	22	4	0	3	0
بَاطلَاق	firing	25	0	25	0	0	0	0	0	13	4	3	4	1
بَابلَاغ	report	23	3	20	2	0	1	0	0	11	5	1	3	0

It can be seen in Table 3.19 that three words (prepositional phrases) "بقتل / bi-qtl / by killing", "بسرقة / bi-srqt / by theft" and "بَاطلَاق / bi-etlaq / by firing", which appeared in the collocation list of this verb, are crime action words, indicating types of crime in the form of prepositional phrases. Moreover, the results denote that the preposition "ب / bi / by" has highest frequency with this verb. As a result, it can be inferred that there is dependency relationship between them and consequently, the proposition "ب / bi / by" is a candidate for being the companion to this verb in the crime domain.

In addition, the position of the words that are types of crime is varied e.g. "بقتل / bi-qtl / by killing" occurs at Positions +1, +2, +3, +4 and +5. Figure 3.1 depicts this case. Thus, the occurrence of the preposition decides the place (position) of the prepositional phrase in the sentence. This condition reveals that sometimes there is a separation between the verb and its following preposition in the sentence i.e. it is not necessary the associated preposition that immediately follows the verb.

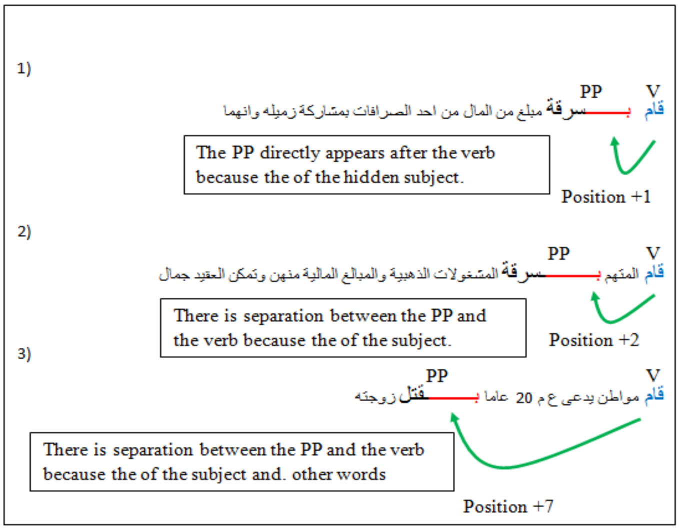
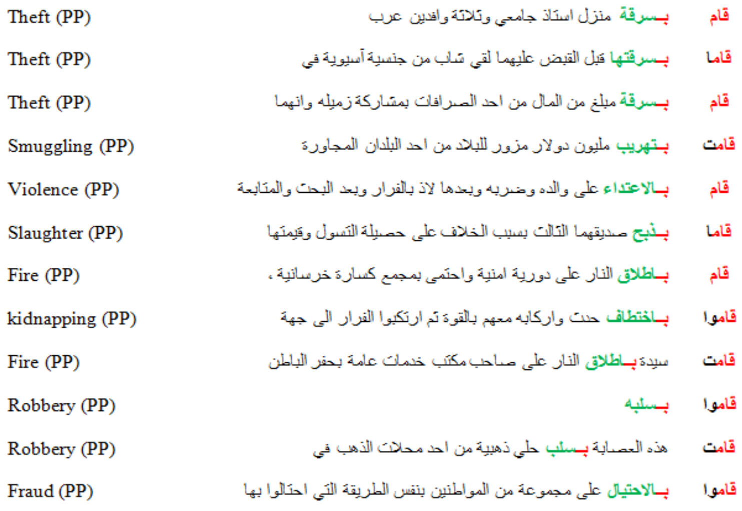

1)
PP
V
قام ب سرقة مبلغ من المال من احد الصرافات بمشاركة زميله وانهما
The PP directly appears after the verb because the of the hidden subject.
Position +1
2)
PP
V
قام المتهم ب سرقة المشغولات الذهبية والمبالغ المالية منهن وتمكن العقيد جمال
There is separation between the PP and the verb because the of the subject.
Position +2
3)
PP
V
قام مواطن يدعى ع م 20 عاما ب قتل زوجته
There is separation between the PP and the verb because the of the subject and. other words
Position +7

Figure 3.1: Different examples of occurrence of the preposition after the verb

The following Figure 3.2 presents a sample of the concordance analysis results of this verb in different forms, such as the verb "قَاما / qama / did" (masculine dual), "قَامت / qamt / did" (feminine singular) and "قَاموا / qamow / did" (masculine plural) as well as "قَام / qam / did" (masculine singular). It can be seen that this verb is always followed by the preposition "ب / bi / by", and that the type of crime is in the form of a prepositional phrase.



Theft (PP)	بـسرقة منزل استاذ جامعي وثلاثة وافدين عرب	قام
Theft (PP)	بـسرقتها قبل القبض عليهما لقي شاب من جنسية آسيوية في	قاما
Theft (PP)	بـسرقة مبلغ من المال من احد الصرافات بمشاركة زميله وانهما	قام
Smuggling (PP)	بـتهريب مليون دولار مزور للبلاد من احد البلدان المجاورة	قامت
Violence (PP)	بـالاعتداء على والده وضربه وبعدها لاذ بالفرار وبعد البحث والمتابعة	قام
Slaughter (PP)	بـذبح صديقهما الثالث بسبب الخلاف على حصيلة التسول وقيمتها	قاما
Fire (PP)	بـاطلاق النار على دورية امنية واحتمى بمجمع كسارة خرسانية ،	قام
kidnapping (PP)	بـاختطاف حدث واركابه معهم بالقوة ثم ارتكبوا الفرار الى جهة	قاموا
Fire (PP)	سيدة بـاطلاق النار على صاحب مكتب خدمات عامة بحفر الباطن	قامت
Robbery (PP)	بـسلبه	قاموا
Robbery (PP)	هذه العصابة بـسلب حلي ذهبية من احد محلات الذهب في	قامت
Fraud (PP)	بـالاحتيال على مجموعة من المواطنين بنفس الطريقة التي احتالوا بها	قاموا

Figure 3.2: The concordance lines of "قَام / qamt / did", and its inflected forms

- اعترف / aatraf / confessed

The verb "اعترف / aatrafa" (masculine singular) means confessed in English, meaning that the accused person or the criminal informs the investigators about his/her actions. Thus, probability of having a type of crime in the same context is relatively high. However, the verb analysis results below not only reveal the type of preposition that most often follows this verb, but also provides a new case regarding the occurrence of the word that represents the type of crime. As seen above, with the verb "قَام / qam / did", the analysis shows that the crime type occurs immediately after the preposition. In contrast, the new case here is that the type of crime not only occurs directly after the preposition but can also occur one or two tokens away from the preposition, i.e. in the form of a noun phrase. Table 3.20 lists the collocation results for "اعترف / aatrafa / confessed".

Table 3.20: Collocation results for ”اعترف / aatraf / confessed”

Confessed		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
المتهم	the accused	73	47	26	6	3	5	15	18	25	1	0	0	0
من	from	73	27	46	15	7	5	0	0	1	4	10	13	18
علَى	on	70	40	30	5	15	17	3	0	6	1	6	3	14
في	in	60	27	33	7	9	6	5	0	3	9	4	9	8
معه	with him	58	42	16	1	1	3	3	34	0	2	5	4	5

We have already seen that this verb appears in collocation with ” بسرقة / bi-srqt / by theft”, which is comprised of two parts: the preposition ”ب / bi / by” and ”سرقة / srqt / theft”, as in Table 3.14. Also, Position +1 in Table 3.20 shows that there is no significant preposition frequently occurring directly after the verb. This indicates that the preposition most often associated with this verb is ”ب / bi / by”. Therefore, another collocation analysis on this verb was carried out, and the most frequent words with the first letter being ”ب / bi / by” are selected. Table 3.21 shows the result of the second collocation analysis of this verb.

Table 3.21: Most frequent words associated with preposition ”ب / bi / by” found in the collocation results for ”اعترف / aatraf / confessed”

Confessed		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
بَارتكَابه	committing	41	1	40	0	1	0	0	0	32	7	1	0	0
بسرقة	theft	25	1	24	1	0	0	0	0	11	7	5	0	1
بمَا	including	18	6	12	0	4	0	0	2	8	3	1	0	0
بقيَامه	did	17	0	17	0	0	0	0	0	14	1	2	0	0
بمشَاركة	Participation	13	2	11	0	2	0	0	0	1	2	2	4	2

The words in Table 3.21 mostly appear to the left, especially at Position +1. The type of crime appears explicitly through the word ” بسرقة / bi-srqt / by theft”. Hence, ”ب / bi / by” is elected to be the most frequent preposition associating with this verb.

Moreover, based on this analysis, the word " بَارتكَابه / bi-ertkabh" which means committing appears many times with this verb. However, the combination of the verb "اعترف / aatrafa / confessed" and the prepositional phrase " بَارتكَابه / bi-ertkab / by committing" do not have meaning, as in the following sentence:

- هو اعترف بَارتكَاب / hwa aatraf bi-ertkab

  He confessed committing.

The first question that comes to mind is: what did he commit? As a result, a new case emerges because there is a lack of meaning. Thus, there is a need to know the word that follows the prepositional phrase in order for the meaning to be completed. Table 3.22 presents the collocation for the word (prepositional phrase) " بَارتكَابه / bi-ertkabh / to committing" in order to investigate the five most frequent words that most often immediately follow it (Position +1). The results show that the words " جريمة / jrymt / crime" and " عملية / amlyt / operation", which already appear in collocation with " سرقة / srqt / theft" and "قتل /qtl / killing", also occur in collocation with this word. For this reason, these two words are analysed as well.

Table 3.22: Collocation results for " بَارتكَابه / bi-ertkabh / to committing"

Committing		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
الوَاقعة	the incident	13	0	13	0	0	0	0	0	12	0	0	1	0
الجريمة	the crime	11	0	11	0	0	0	0	0	9	1	0	1	0
جريمة	crime	7	0	7	0	0	0	0	0	7	0	0	0	0
لّوَاقعة	for incident	5	0	5	0	0	0	0	0	5	0	0	0	0
عملية	case	5	0	5	0	0	0	0	0	5	0	0	0	0

The collocations for the word " جريمة / jrymt / crime" and " عملية / amlyt / operation" are presented in Table 3.23 and Table 3.24, respectively. Clearly, types of crime, such as "قتل / qtl / murder" or "القتل / alqtl / the murder", " سرقة / srqt / theft

" or "السرقة / alsrqt / the theft ", "تزوير / tzwyr / forgery" and "تهريب / thryb / smuggling" occur in the contexts of both words. This case is explained through the two examples in Figure 3.3.

Table 3.23: Collocation results for " جريمة / jrymt / crime"

Crime		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
قتل	murder	77	1	76	0	0	1	0	0	74	0	2	0	0
القتل	the murder	27	0	27	0	0	0	0	0	26	0	1	0	0
سرقة	theft	23	0	23	0	0	0	0	0	22	0	1	0	0
اخرَى	other	5	0	5	0	0	0	0	0	4	1	0	0	0
تزوير	forgery	4	0	4	0	0	0	0	0	4	0	0	0	0

Table 3.24: Collocation results for " عملية / amlyt / operation"

Operation		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
القبض	the arrest	57	3	54	2	1	0	0	0	47	2	1	1	3
الضبط	the arrest	27	0	27	0	0	0	0	0	27	0	0	0	0
تهريب	smuggling	25	0	25	0	0	0	0	0	25	0	0	0	0
السرقة	the theft	24	0	24	0	0	0	0	0	24	0	0	0	0
سرقة	theft	20	0	20	0	0	0	0	0	20	0	0	0	0

It can be seen in both sentences 1 and 2 the type of crime "القتل / alqtl / the killing" and " السرقة / alsrqt / the theft" occurred within the noun phrase that followed the preposition "ب / bi / by" and it is two tokens away from this preposition. This result provides that this transitive verb can be a head node for a crime type in form of noun phrase that follows its preposition.

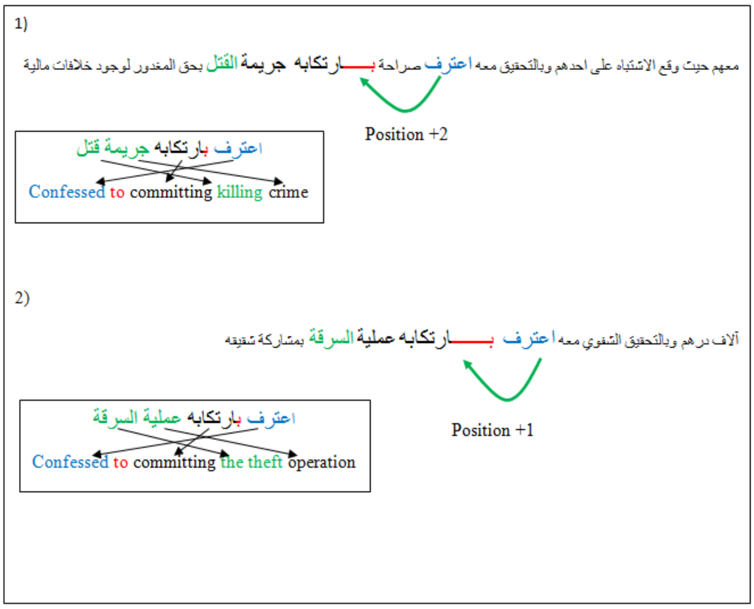
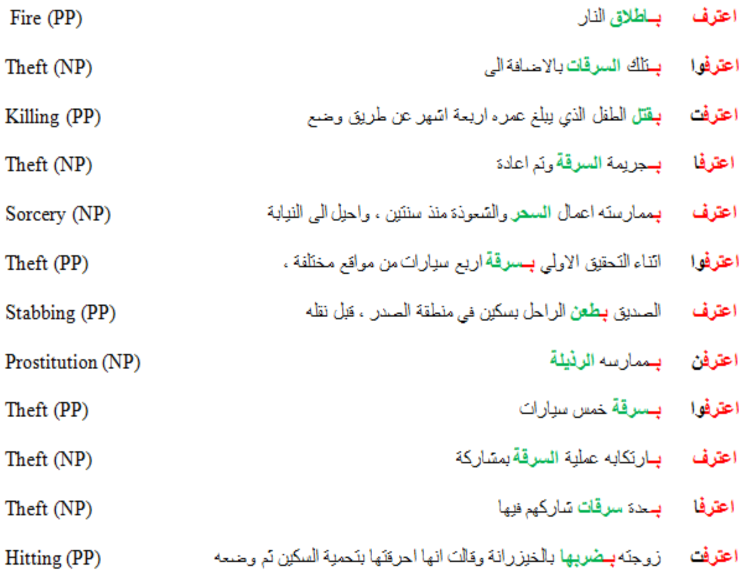

1)
معهم حيث وقع الاشتباه على احدهم وبالتحقيق معه اعترف صراحة بـــــارتكابه جريمة القتل بحق المغدور لوجود خلافات مالية
Position +2
اعترف بارتكابه جريمة قتل
Confessed to committing killing crime
2)
آلاف درهم وبالتحقيق الشفوي معه اعترف بـــــارتكابه عملية السرقة بمشاركة شقيقه
Position +1
اعترف بارتكابه عملية السرقة
Confessed to committing the theft operation

Figure 3.3: The verb is the head node of the noun phrases that contain a crime action word

Figure 3.4 presents the concordance analysis results for this verb in order to visually examine them. Clearly, there is a dependency relationship between the verb and the preposition ”ب / bi / by”. Furthermore, the crime type can be seen in the context of this verb, either in the form of a prepositional phrase or a noun phrase. Noun phrases that contain crime types (crime action word) occur after the preposition. Also, different words, such as ”اعترفت / aatrft / confessed” (feminine singular), ”اعترفن / aatrfn / confessed” (feminine plural), ”اعترفَا / aatrfa / confessed” (masculine dual) and ”اعترفوَا / aatrfow / confessed” (masculine plural), which are derived from the verb ”اعترف”, (masculine singular), can be seen in these results, and they are also linked to their complements by the same preposition.



Fire (PP)	بـاطلاق النار	اعترف
Theft (NP)	بـتلك السرقات بالاضافة الى	اعترفوا
Killing (PP)	بقتل الطفل الذي يبلغ عمره اربعة اشهر عن طريق وضع	اعترفت
Theft (NP)	بـجريمة السرقة وتم اعادة	اعترفا
Sorcery (NP)	بممارسته اعمال السحر والشعوذة منذ سنتين ، واحيل الى النيابة	اعترف
Theft (PP)	اثناء التحقيق الاولي بـسرقة اربع سيارات من مواقع مختلفة ،	اعترفوا
Stabbing (PP)	الصديق بطعن الراحل بسكين في منطقة الصدر ، قبل نقله	اعترف
Prostitution (NP)	بـممارسه الرذيلة	اعترفن
Theft (PP)	بـسرقة خمس سيارات	اعترفوا
Theft (NP)	بـارتكابه عملية السرقة بمشاركة	اعترف
Theft (NP)	بـعدة سرقات شاركهم فيها	اعترفا
Hitting (PP)	زوجته بـضربها بالخيزرانة وقالت انها احرقتها بتحمية السكين ثم وضعه	اعترفت

Figure 3.4: The concordance lines for "اعترف / aatrf / confessed", and its inflected forms

- تعرض / taarrad / subjected

The word "تعرض / taarrad" means subjected. This verb can reach its object using the prepositions "ل / li / of" or "الَى / ila / to", as will be shown. The preposition "ل / li / of" is like the preposition "ب / bi / by" in that both of them are fused with the word as a prefix. So, this type of preposition cannot appear on its own, as most other prepositions do. Table 3.25 presents the results of the analysis. It can be seen that the preposition "الَى / ila / to" is the only word that occurs directly after the word "تعرض / taarrad / subjected" (Position +1); this is good evidence which informs us that this word does not accept the other prepositions in the list, even though they appear in the same context.

Table 3.25: Collocation results for ”تعرض / taarrad / subjected”

Subjected		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	148	37	111	3	15	15	2	2	0	33	28	25	25
في	in	105	32	73	4	11	11	5	1	0	21	17	12	23
عن	about	83	79	4	1	4	2	5	67	0	1	3	0	0
الَى	to	80	25	55	8	5	6	2	4	15	13	11	5	11
ان	that	56	53	3	5	7	11	17	13	0	0	0	1	2

Another analysis is here performed in order to select the most frequent words whose first letter is the preposition ”ل / li / of”. The results are presented in the following Table 3.26.

Table 3.26: Most frequent words associated with preposition ”ل / li / of” found in the collocation results for ”تعرض / taarrad / subjected”

Subjected		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
لّلسرقة	of theft	48	0	48	0	0	0	0	0	14	17	5	6	6
له	of him	21	3	18	0	1	2	0	0	17	0	0	0	1
لعملية	of operation	18	0	18	0	0	0	0	0	11	2	3	2	0
لسرقة	of theft	16	0	16	0	0	0	0	0	5	7	0	2	2
لّضرب	of hitting	15	0	15	0	0	0	0	0	9	3	1	1	1

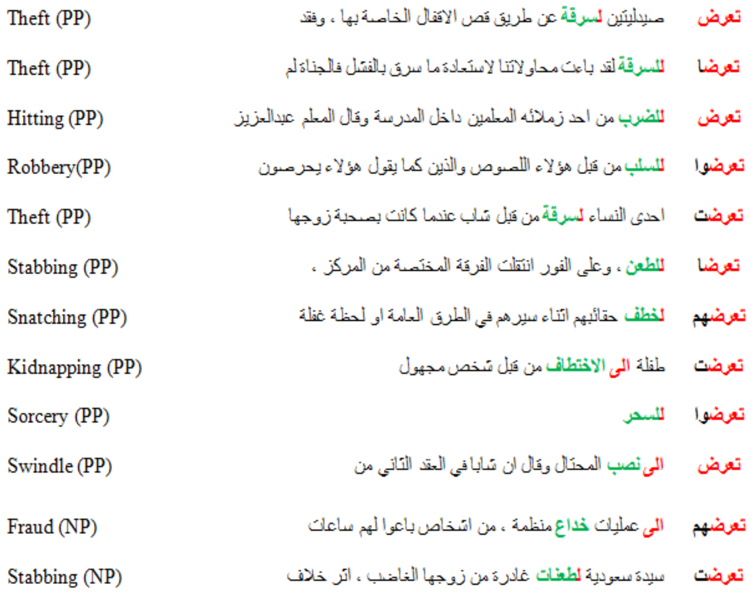
The analysis shows that all the words that are fused with the preposition ”ل / li / of” occur after the word ”تعرض / taarrad / subjected”, except the word ” له / lh / for him”. The type of crime appears explicitly through the words in the first, fourth and fifth rows in the form of a prepositional phrase.

With regards to the word (prepositional phrase) ” لعملية / li-amlyat / of operation”, it is like the word ” عملية / amlyt / operation”, which was previously analysed but here it is attached to the preposition ”ل / li / of”; they form a single word. As a result, this word (preposition phrase) is also analysed. Table 3.27 shows the results for this word. The five most frequent words collocating with "لعملية / li-amlyat / to operation" appear to its left and they are crime action words representing different crime types.

Table 3.27: Collocation results for "لعملية / li-amlyat / of operation"

To operation		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
نصب	swindle	14	0	14	0	0	0	0	0	14	0	0	0	0
احتيَال	fraud	9	0	9	0	0	0	0	0	9	0	0	0	0
سرقة	theft	5	0	5	0	0	0	0	0	5	0	0	0	0
سطو	burglary	3	0	3	0	0	0	0	0	3	0	0	0	0
النصب	the swindle	2	0	2	0	0	0	0	0	2	0	0	0	0

Figure 3.5 presents a sample of the concordance analysis results for the verb "تعرض / taard / subjected" (masculine singular) in different forms, e.g. the verbs "تعرضَا / taarda / subjected" (masculine dual), "تعرضوَا / taardow / subjected" (masculine plural) and "تعرضت / taaradt / subjected" (feminine singular). Also, it can be seen that these verbs are followed by either "ل / li / of" or "إلَى / ila / to" to reach their complements that contain a crime type. Moreover, all types of crime (crime action words) occurring in the form of prepositional phrase are preceded by the above prepositions. As with the above verb, the preposition also does not necessary occur directly after the verb.



Theft (PP)	صيدليتين لسرقة عن طريق قص الاقفال الخاصة بها ، وفقد	تعرض
Theft (PP)	للسرقة لقد باءت محاولاتنا لاستعادة ما سرق بالفشل فالجناة لم	تعرضا
Hitting (PP)	للضرب من احد زملائه المعلمين داخل المدرسة وقال المعلم عبدالعزيز	تعرض
Robbery(PP)	للسلب من قبل هؤلاء اللصوص والذين كما يقول هؤلاء يحرصون	تعرضوا
Theft (PP)	احدى النساء لسرقة من قبل شاب عندما كانت بصحبة زوجها	تعرضت
Stabbing (PP)	للطعن ، وعلى الفور انتقلت الفرقة المختصة من المركز ،	تعرضا
Snatching (PP)	لخطف حقائبهم اثناء سيرهم في الطرق العامة او لحظة غفلة	تعرضهم
Kidnapping (PP)	طفلة الى الاختطاف من قبل شخص مجهول	تعرضت
Sorcery (PP)	للسحر	تعرضوا
Swindle (PP)	الى نصب المحتال وقال ان شابا في العقد الثاني من	تعرض
Fraud (NP)	الى عمليات خداع منظمة ، من اشخاص باعوا لهم ساعات	تعرضهم
Stabbing (NP)	سيدة سعودية لطعنات غادرة من زوجها الغاضب ، اثر خلاف	تعرضت

Figure 3.5: The concordance lines for "تعرض / taard / subjected", and its inflected forms

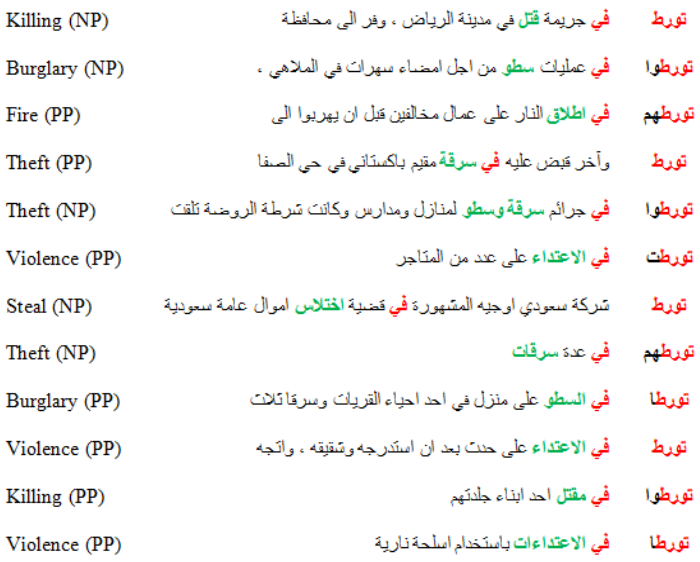
- تورط / twrt / involved

According to Alahmadi [22], the word "تورط / twrt / involved" associates with the preposition "في / fi / in". The outcome of the analysis in Table 3.28 proves that this preposition is the most frequent word, and that it can immediately follow the verb. Moreover, syntactically, the other words, such as the prepositions "من / min / from", "علَى / ala / on" and "عن / aan / about" cannot follow, even if they occur in the verb's context. The collocation results show that a crime action word ("سرقة / srqt / theft") collocates with this verb (see Table 3.28).

Table 3.28: Collocation results for ”تورط / twrt / involved”

Involved		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	61	11	50	5	2	1	2	1	19	9	11	4	7
من	from	23	9	14	0	2	2	3	2	0	5	4	5	0
علَى	on	13	9	4	0	1	2	6	0	0	0	2	1	1
سرقة	theft	13	1	12	0	0	0	1	0	0	4	3	3	2
عن	about	11	11	0	0	2	1	0	8	0	0	0	0	0

The results of the concordance analysis for this verb are presented in Figure 3.6. The preposition ”في / fi / in” often follows this verb. Moreover, different forms, such as ”تورطَا / twrt / involved” (masculine dual), ”تورطوَا / twrtow / involved” (masculine plural) and ”تورطت / twrtt / involved” (feminine singular), which are derived from the verb ”تورط / twrt / involved”, are evident. Types of crime occur in the form of noun phrase, as in sentences 1, 2, 5, 7 and 8, and in the form of prepositional phrase, as in all the other sentences.



Killing (NP)	في جريمة قتل في مدينة الرياض ، وفر الى محافظة	تورط
Burglary (NP)	في عمليات سطو من اجل امضاء سهرات في الملاهي ،	تورطوا
Fire (PP)	في اطلاق النار على عمال مخالفين قبل ان يهربوا الى	تورطهم
Theft (PP)	وآخر قبض عليه في سرقة مقيم باكستاني في حي الصفا	تورط
Theft (NP)	في جرائم سرقة وسطو لمنازل ومدارس وكانت شرطة الروضة تلقت	تورطوا
Violence (PP)	في الاعتداء على عدد من المتاجر	تورطت
Steal (NP)	شركة سعودي اوجيه المشهورة في قضية اختلاس اموال عامة سعودية	تورط
Theft (NP)	في عدة سرقات	تورطهم
Burglary (PP)	في السطو على منزل في احد احياء القريات وسرقا ثلاث	تورطا
Violence (PP)	في الاعتداء على حدث بعد ان استدرجه وشقيقه ، واتجه	تورط
Killing (PP)	في مقتل احد ابناء جلدتهم	تورطوا
Violence (PP)	في الاعتداءات باستخدام اسلحة نارية	تورطا

Figure 3.6: The concordance lines for ”تورط / twrt / involved”, and its inflected forms

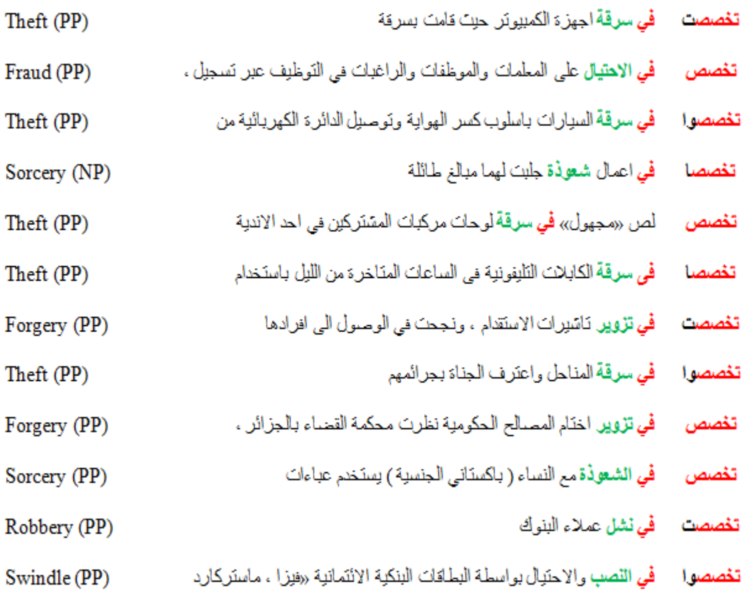
- تخصّصت / tksasat / specialised

Table 3.29 shows that the most frequent word collocating with "تخصّصت / tksasat / specialised" is the preposition "في / fi / in". It occurs 56 times, most of them directly after the verb (47 times at Position +1). Moreover, the other prepositions have no relationship with this verb, and as a result, it is chosen to be the associated preposition for this verb in this domain. The word "سرقة / srqt / theft" occurs 34 times to the left (following the verb), and it is a type of crime that appears within the context of this verb.

Table 3.29: Collocation results for "تخصّصت / tksasat / specialised"

Specialised		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	56	5	51	1	3	1	0	0	47	0	0	0	4
سرقة	theft	34	0	34	0	0	0	0	0	0	33	1	0	0
عصَابة	gang	29	29	0	2	3	0	13	11	0	0	0	0	0
من	from	27	17	10	5	5	6	1	0	0	0	0	4	6
علَى	on	9	8	1	2	1	3	2	0	0	0	1	0	0

Figure 3.7 depicts the results of the concordance analysis. It shows that this verb can take different forms. At the same time, the inflected forms (derived from the main verb) have dependency relationships with the same preposition. It can be seen that a type of crime (crime action word) follows the preposition in all sentences, except in sentence number 4 where it is preceded by a noun. Although the crime action word in sentence 4 occurs in the form of a noun phrase, it is preceded by the transitive verb and its preposition.



Theft (PP)	في سرقة اجهزة الكمبيوتر حيث قامت بسرقة	تخصصت
Fraud (PP)	في الاحتيال على المعلمات والموظفات والراغبات في التوظيف عبر تسجيل ،	تخصص
Theft (PP)	في سرقة السيارات باسلوب كسر الهواية وتوصيل الدائرة الكهربائية من	تخصصوا
Sorcery (NP)	في اعمال شعوذة جلبت لهما مبالغ طائلة	تخصصا
Theft (PP)	لص «مجهول» في سرقة لوحات مركبات المشتركين في احد الاندية	تخصص
Theft (PP)	في سرقة الكابلات التليفونية فى الساعات المتاخرة من الليل باستخدام	تخصصا
Forgery (PP)	في تزوير تاشيرات الاستقدام ، ونجحت في الوصول الى افرادها	تخصصت
Theft (PP)	في سرقة المناحل واعترف الجناة بجرائمهم	تخصصوا
Forgery (PP)	في تزوير اختام المصالح الحكومية نظرت محكمة القضاء بالجزائر ،	تخصص
Sorcery (PP)	في الشعوذة مع النساء (باكستاني الجنسية) يستخدم عباءات	تخصص
Robbery (PP)	في نشل عملاء البنوك	تخصصت
Swindle (PP)	في النصب والاحتيال بواسطة البطاقات البنكية الائتمانية «فيزا ، ماستركارد	تخصصوا

Figure 3.7: The concordance lines for "تخصّصت / tksasat / specialised", and its inflected forms

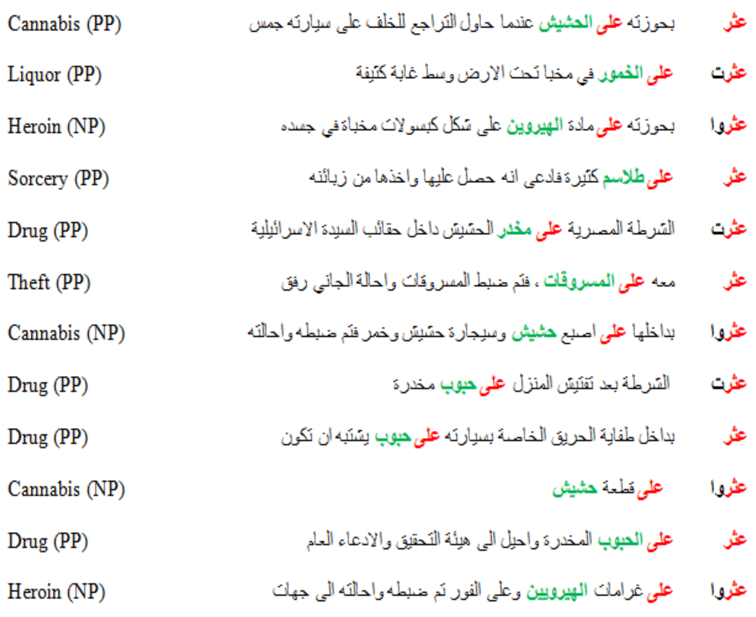
- عثر / athr / found

Through examining the crime reports, we notice that the word "عثر / athr / found" usually appears in crimes relating to drugs or alcohol smuggling. Ahmadi [22] stated that this transitive verb associates with the preposition "علَى / ala / on". After performing collocation analysis, as in Table 3.30, it has been found the preposition "علَى / ala / on" is the most frequent word in collocation with this verb. It occurs 306 times to the left of this verb, and so, clearly, this preposition has a dependency relationship with it. As a result, the other preposition, "في / fi / in", is ignored. Also, the words "كمَا / kma / as" and "و بتفتيش / wbitftish / and inspecting" are discarded because they do not occur after the verb.

Table 3.30: Collocation results for "عثر / athr / found"

Found		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
عَلَى	on	328	22	306	11	5	2	4	0	111	123	50	11	11
من	from	109	36	73	5	16	11	4	0	0	0	15	26	32
في	in	93	35	58	18	11	4	2	0	10	15	14	5	14
كَمَا	as	49	49	0	0	0	0	0	49	0	0	0	0	0
و بتفتيش	and inspecting	41	41	0	5	1	3	32	0	0	0	0	0	0

The results of the concordance analysis of the verbs "عثر / athr / found" (masculine singular), "عثروَا / athrow / found" (masculine plural), and "عثرت / athrt / found" (feminine singular) are presented in Figure 3.8. As can be seen, the verbs can reach their complements, which contain a crime type, through the preposition "عَلَى / ala / on". As can also be seen, the type of crime is mostly in the form of a prepositional phrase but it could be in noun phrase. As with the above verbs, the head noun of the noun phrase follows the preposition that associates with the verb.



Cannabis (PP)	بحوزته على الحشيش عندما حاول التراجع للخلف على سيارته جمس	عثر
Liquor (PP)	على الخمور في مخبا تحت الارض وسط غابة كثيفة	عثرت
Heroin (NP)	بحوزته على مادة الهيروين على شكل كبسولات مخباة في جسده	عثروا
Sorcery (PP)	على طلاسم كثيرة فادعى انه حصل عليها واخذها من زبائنه	عثر
Drug (PP)	الشرطة المصرية على مخدر الحشيش داخل حقائب السيدة الاسرائيلية	عثرت
Theft (PP)	معه على المسروقات ، فتم ضبط المسروقات واحالة الجاني رفق	عثر
Cannabis (NP)	بداخلها على اصبع حشيش وسيجارة حشيش وخمر فتم ضبطه واحالته	عثروا
Drug (PP)	الشرطة بعد تفتيش المنزل على حبوب مخدرة	عثرت
Drug (PP)	بداخل طفاية الحريق الخاصة بسيارته على حبوب يشتبه ان تكون	عثر
Cannabis (NP)	على قطعة حشيش	عثروا
Drug (PP)	على الحبوب المخدرة واحيل الى هيئة التحقيق والادعاء العام	عثر
Heroin (NP)	على غرامات الهيروين وعلى الفور تم ضبطه واحالته الى جهات	عثروا

Figure 3.8: The concordance lines for "عثر / athr / found", and its inflected forms

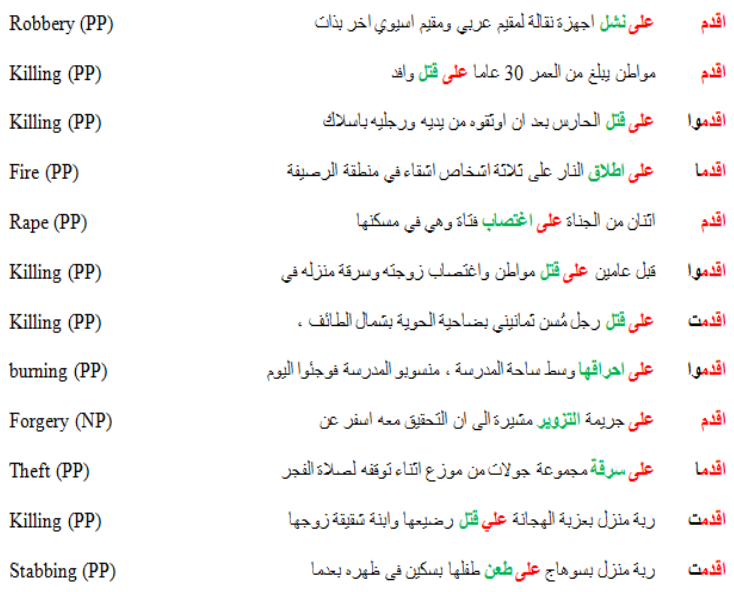
- اقدم / aqdm / conducted

The verb "اقدم / aqdm / conducted", as can be seen in Table 3.31, has a dependency relationship with the preposition "علَى / ala / on". This preposition occurs 106 times, distributed between Positions +1 and +5. Additionally, this preposition is only able to directly follow this verb among the most frequent words (see Position +1). As a result, the other preposition are ignored. Also, a type of crime, represented by the crime action word "قتل / qtl / murder" appears in the context of this verb.

Table 3.31: Collocation results for "اقدم / aqdm / conducted"

Conducted		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
علَى	on	117	11	106	6	0	0	5	0	55	27	8	9	7
بن	son of	54	7	47	4	1	2	0	0	0	16	6	20	5
من	from	43	13	30	3	6	3	1	0	0	6	6	10	8
قتل	killing	41	1	40	0	1	0	0	0	0	27	10	3	0
في	in	22	6	16	1	2	1	2	0	0	7	3	1	5

The concordance analysis results for this verb (and its inflected forms) are presented in the following Figure 3.9. It shows that the complement of the verb always comes after the preposition "علَى / ala / on". Furthermore, all the types of crime occurring here are in the form of a prepositional phrase, except in the ninth sentence "جريمة التزوير / jrymt altzwyr / crime of the forgery", which occurs in the form of a noun phrase. As usual, the head noun of the noun phrase occurs immediately after the preposition.



Robbery (PP)	على نشل اجهزة نقالة لمقيم عربي ومقيم اسيوي اخر بذات	اقدم
Killing (PP)	مواطن يبلغ من العمر 30 عاما على قتل وافد	اقدم
Killing (PP)	على قتل الحارس بعد ان اوثقوه من يديه ورجليه باسلاك	اقدموا
Fire (PP)	على اطلاق النار على ثلاثة اشخاص اشقاء في منطقة الرصيفة	اقدما
Rape (PP)	اثنان من الجناة على اغتصاب فتاة وهي في مسكنها	اقدم
Killing (PP)	قبل عامين على قتل مواطن واغتصاب زوجته وسرقة منزله في	اقدموا
Killing (PP)	على قتل رجل مُسن ثمانيني بضاحية الحوية بشمال الطائف ،	اقدمت
burning (PP)	على احراقها وسط ساحة المدرسة ، منسوبو المدرسة فوجئوا اليوم	اقدموا
Forgery (NP)	على جريمة التزوير مشيرة الى ان التحقيق معه اسفر عن	اقدم
Theft (PP)	على سرقة مجموعة جولات من موزع اثناء توقفه لصلاة الفجر	اقدما
Killing (PP)	ربة منزل بعزبة الهجانة علي قتل رضيعها وابنة شقيقة زوجها	اقدمت
Stabbing (PP)	ربة منزل بسوهاج على طعن طفلها بسكين فى ظهره بعدما	اقدمت

Figure 3.9: The concordance lines for "اقدم / aqdm / conducted", and its inflected forms

- ادين / adyn / convicted

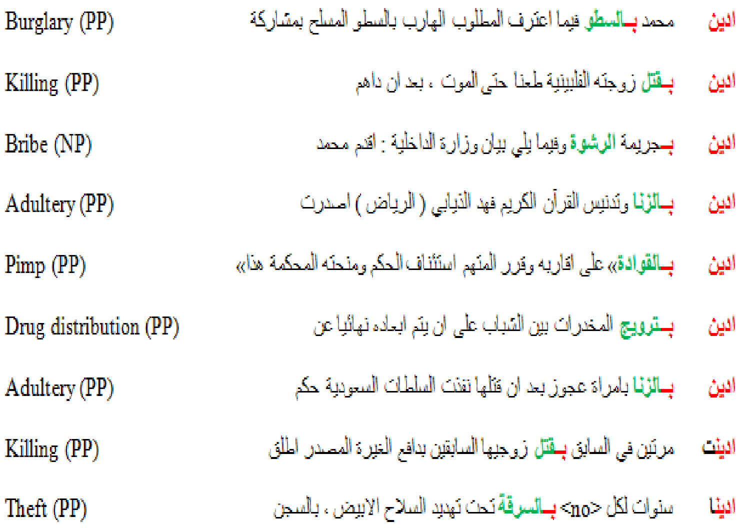
According to the following results in Table 3.32, the word "ادين / adyn / convicted" has association with the preposition "ب / bi / by". The preposition "ب / bi / by" is attached to all the words, except the word in the first row. Moreover, the crime action words "بقتل / bi-qtal / by killing", "بَالزنَا / bi-alzna / by adultery" and "بَالسطو / bi-alstw / by robbery" appear in the context of this verb.

Table 3.32: Collocation results for "ادين / adyn / convicted"

Convicted		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
علَى	on	7	3	4	0	0	0	3	0	0	1	0	2	1
بقتل	by killing	4	0	4	0	0	0	0	0	4	0	0	0	0
بَالسجن	by jail	2	1	1	1	0	0	0	0	0	0	0	1	0
بَالزنَا	by adultery	2	0	2	0	0	0	0	0	2	0	0	0	0
بَالسطو	by burglary	2	0	2	0	0	0	0	0	0	2	0	0	0

Figure 3.10 presents the concordance analysis results for "ادين / adyn / convicted"

(masculine singular), "ادينت / adynt / convicted" (feminine singular) and "ادينَا / adyna / convicted" (masculine dual). These verbs have a dependency relationship with the preposition "ب / bi / by". As in above verbs, the preposition sometimes immediately follows the verb and sometimes not. Additionally, most of the crime action words occur in the form of a prepositional phrase, i.e. they are governed by preposition. The third sentence shows that the head noun that forms the noun phrase with the second noun (crime action word) follows the preposition.



Burglary (PP)	محمد بـالسطو فيما اعترف المطلوب الهارب بالسطو المسلح بمشاركة	ادين
Killing (PP)	بـقتل زوجته الفلبينية طعنا حتى الموت ، بعد ان داهم	ادين
Bribe (NP)	بـجريمة الرشوة وفيما يلي بيان وزارة الداخلية : اقدم محمد	ادين
Adultery (PP)	بـالزنا وتدنيس القرآن الكريم فهد الذيابي (الرياض) اصدرت	ادين
Pimp (PP)	بـالقوادة» على اقاربه وقرر المتهم استئناف الحكم ومنحته المحكمة هذا»	ادين
Drug distribution (PP)	بـترويج المخدرات بين الشباب على ان يتم ابعاده نهائيا عن	ادين
Adultery (PP)	بـالزنا بامراة عجوز بعد ان قتلها تفذت السلطات السعودية حكم	ادين
Killing (PP)	مرتين في السابق بـقتل زوجيها السابقين بدافع الغيرة المصدر اطلق	ادينت
Theft (PP)	سنوات لكل <no> بـالسرقة تحت تهديد السلاح الابيض ، بالسجن	ادينا

Figure 3.10: The concordance lines for "ادين / adyn / convicted", and its inflected forms

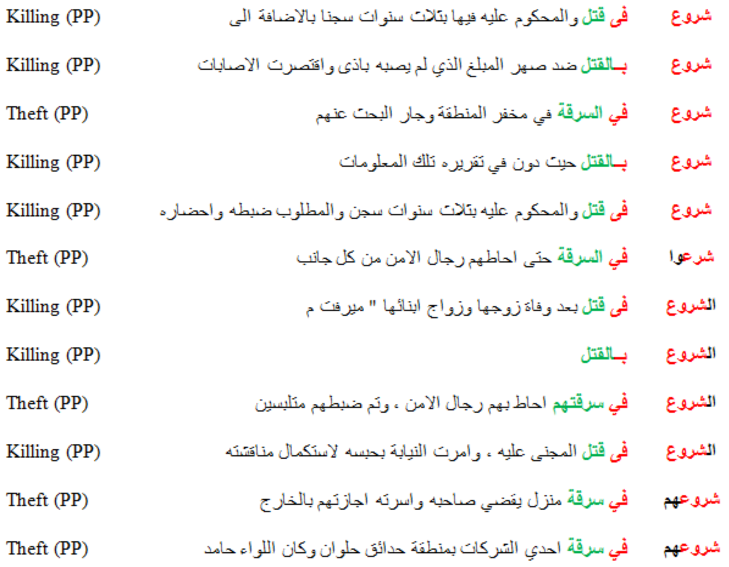
- شروع / shoroaa / commenced

Table 3.33 shows the collocation results for "شروع / shoroaa / commenced". The most frequent preposition collocating with this verb is "في / fi / in". Hence, it is selected to be the associated preposition for this verb. In addition, the preposition "ب / bi / by" is also chosen because the crime action word (prepositional phrase) "بَالقتل / bi-alqtl / by killing" occurs in the context of this verb. Moreover, the word already seen with other verbs, "قضية / qdyt / case", collocates with this verb. This word indicates that a type of crime might occur in the form of a noun phrase in the context of this verb.

Table 3.33: Collocation results for "شروع / shoroaa / commenced"

Commenced		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
قضية	case	6	6	0	0	0	2	0	4	0	0	0	0	0
فَى	in	5	0	5	0	0	0	0	0	5	0	0	0	0
قتل	killing	5	0	5	0	0	0	0	0	0	5	0	0	0
بَالقتل	by killing	4	0	4	0	0	0	0	0	4	0	0	0	0
عليه	on him	3	1	2	0	0	0	0	1	0	0	0	2	0

Figure 3.11 shows the concordance analysis results. The complements (crime action words) of the head nodes (verb) can all be reached by either the "في / fi / in" or the "ب / bi / by" prepositions. Also, it can be seen that the types of crime are in the form of a prepositional phrase because they are directly preceded by these two prepositions.



Killing (PP)	فى قتل والمحكوم عليه فيها بثلاث سنوات سجنا بالاضافة الى	شروع
Killing (PP)	بـــالقتل ضد صهر المبلغ الذي لم يصبه باذى واقتصرت الاصابات	شروع
Theft (PP)	في السرقة في مخفر المنطقة وجار البحث عنهم	شروع
Killing (PP)	بـــالقتل حيث دون في تقريره تلك المعلومات	شروع
Killing (PP)	فى قتل والمحكوم عليه بثلاث سنوات سجن والمطلوب ضبطه واحضاره	شروع
Theft (PP)	في السرقة حتى احاطهم رجال الامن من كل جانب	شرعوا
Killing (PP)	فى قتل بعد وفاة زوجها وزواج ابنائها " ميرفت م	الشروع
Killing (PP)	بـــالقتل	الشروع
Theft (PP)	في سرقتهم احاط بهم رجال الامن ، وتم ضبطهم متلبسين	الشروع
Killing (PP)	فى قتل المجنى عليه ، وامرت النيابة بحبسه لاستكمال مناقشته	الشروع
Theft (PP)	في سرقة منزل يقضي صاحبه واسرته اجازتهم بالخارج	شروعهم
Theft (PP)	في سرقة احدي الشركات بمنطقة حدائق حلوان وكان اللواء حامد	شروعهم

Figure 3.11: The concordance lines for "شروع / shoroaa / commenced", and its inflected forms

As already seen in the above analysis, crime types occur within the context of some transitive verbs i.e. at the complement of these verbs. Therefore, this linguistic phenomenon (transitive verbs with their associated prepositions) will be exploited in order to identify types of crime from Arabic texts.

3.4.3 Crime Type Local Grammar

From the above analysis, we may deduce that types of crime words (crime action words) occur within the context of certain verbs in transitive constructions. These types of crime occur after the associated prepositions of verbs, whether in the form of prepositional phrases or noun phrases. Furthermore, it has been found that each verb associates with a maximum of two prepositions. Moreover, the occurrence of the associated preposition, whether directly after the verb or not, is not important. As a result, the transitive construction can be an effective and useful tool for identifying and extracting types of crime. Hence, the following indicator nodes are identified based on this construction. The verbs are used as keywords to trigger the indicator nodes. Also, a syntactic constraint is applied in order for the extraction task to be achieved. For enabling this condition, these verbs must be in transitive construction.

The developed indicator nodes for extracting crime types from a given text based on the above verbs are presented as follows:

- Name: قَام / qam / did-transitive-ب / bi / by
  Trigger: قَام / qam / did
  Target: crime type
  Constraint: transitive
  Activation: trigger is followed by preposition "ب / bi / by"

- Name: اعترف / aatraf / confessed-transitive-ب / bi / by
  Trigger: اعترف / aatraf / confessed

Target: crime type

Constraint: transitive

Activation: trigger is followed by preposition ”ب / bi / by”

- Name: تعرض / tarad / subjected-transitive-ل / li / of - الَى / ila / to

  Trigger: تعرض / tarad / subjected

  Target: crime type

  Constraint: transitive

  Activation: trigger is followed by preposition ”ل / li / of” or ”الَى / ila / to”

- Name: تورط / twrt / involved-transitive-في / fi / in

  Trigger: تورط / twrt / involved

  Target: crime type

  Constraint: transitive

  Activation: trigger is followed by preposition ”في / fi / in”

- Name: تخصّص / tkasas / specialised-transitive-في / fi / in

  Trigger: تخصّص / tkasas / specialise

  Target: crime type

  Constraint: transitive

  Activation: trigger is followed by preposition ”في / fi / in”

- Name: عثر / athr / found-transitive-علَى / ala / on

  Trigger: عثر / athr / found

  Target: crime type

Constraint: transitive

Activation: trigger is followed by preposition "عَلَى / ala / on"

- Name: اقدم / aqdm / conducted-transitive-في / fi / in

  Trigger: اقدم / aqdm / conducted

  Target: crime type

  Constraint: transitive

  Activation: trigger is followed by preposition "عَلَى / ala / on"

- Name: ادين / adyn / convicted-transitive-ب / bi / by

  Trigger: ادين / adyn / convicted

  Target: crime type

  Constraint: transitive

  Activation: trigger is followed by preposition "ب / bi / by"

- Name: شروع / shoroaa / commenced-transitive-في / fi / in - ب / bi / by

  Trigger: شروع / shoroaa / commenced

  Target: crime type

  Constraint: transitive

  Activation: trigger is followed by preposition "في / fi / in" or "ب / bi / by"

Moreover, based on the above analysis, the local grammar for describing the dominant crime type patterns can be built. Figure 3.12 shows this local grammar.

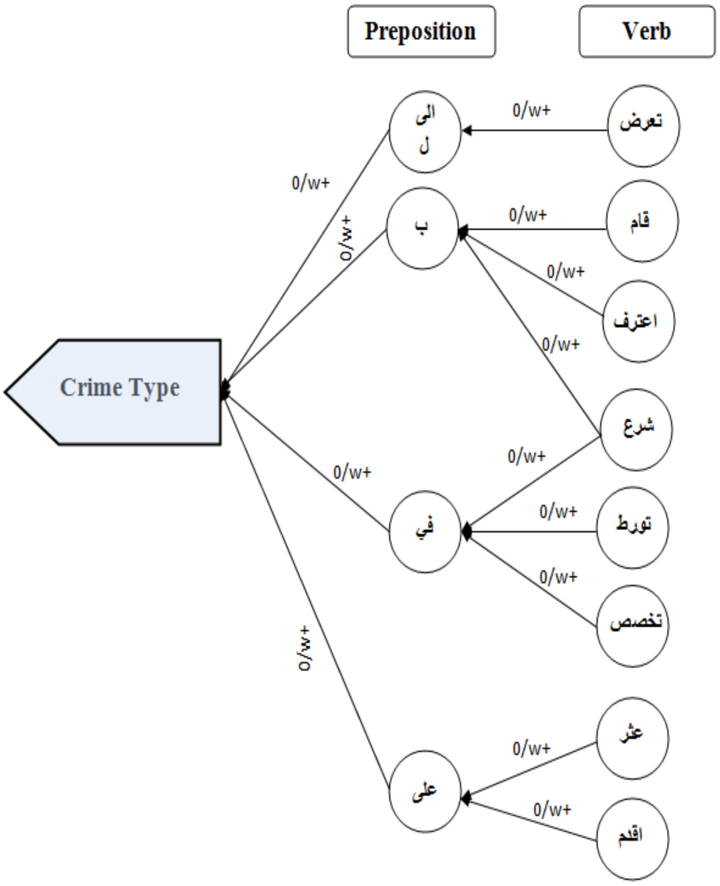


Figure 3.12: Crime type local grammar

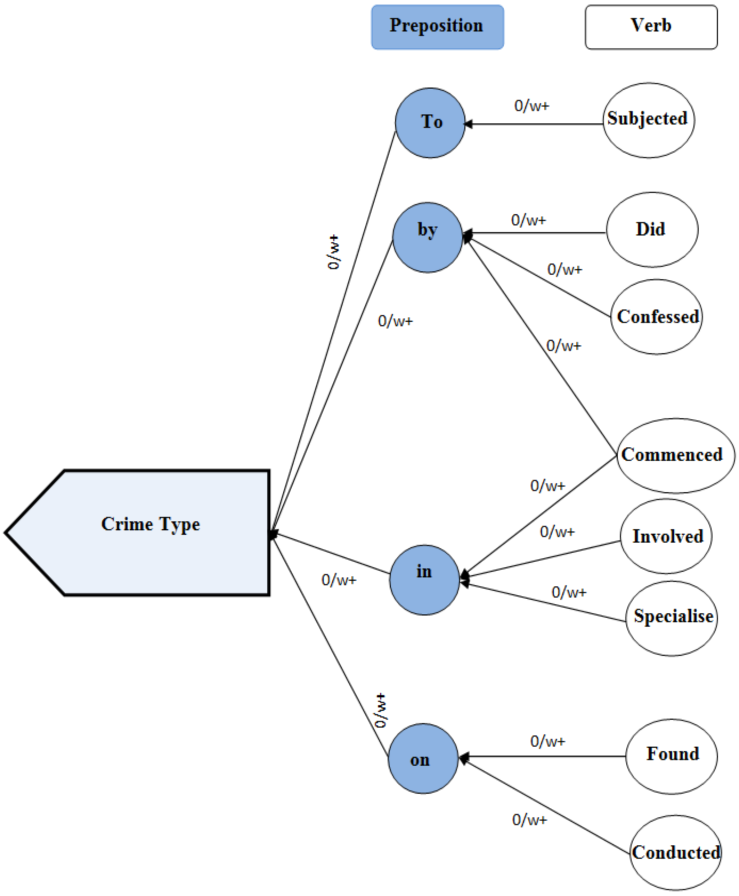


Figure 3.13: Crime type local grammar (English Version)

3.5 Nationality

As mentioned previously, the nationality type is one of the targets that need to be recognized and then extracted. It is has been found that the word " الجنسية / aljensyt / the nationality", which is the word usually used to illustrate a person's nationality, is within the list of 100 tokens (see Table 3.3).

The frequency of " الجنسية / aljensyt / the nationality" (singular with the definite article "ال / al / the") occurs 676 times in the special corpus (ACNRC) but only 91 times in the general corpus. Other forms of this word were found, and their frequencies were also counted. For example, the words " جنسية / jensyt / nationality" (singular without definite article) and "جنسيّات / jensyat / nationalities" (plural) occur 198 and 53 times in the ACNRC but only 42 and 18 times in the general corpus, respectively. On the other hand, the word "الجنسيّات / aljensyat / the nationalities" (plural with definite article "ال / al / the") occurs 33 times in the ACNRC but 239 times in the general corpus. Nevertheless, the general occurrence of this word is greater in the ACNRC than in the general corpus.

The collocation and concordance analyses of all these words were performed in order to study their contexts from two perspectives: words order and syntactic structure. The result shows that nationality is often represented by a word that immediately follows the above words with a syntactic constraint. The syntactic condition is that the words "الجنسية / aljensyt / the nationality", "جنسية / jensyt / nationality", "جنسيّات / jensyat / nationalities" and "الجنسيّات / aljensyat / the nationalities" must be assigned the genitive case by the preposition (governor) "من / min / from". However, there is one exception: "الجنسية / aljensyt / the nationality"; when this word is not preceded by the preposition "من / min / from", the word that occurs instead of the preposition is considered to indicate nationality, as will be shown. The collocation and concordance analyses of each word is provided in the following section.

3.5.1 Analysis of Nationality Word

The collocation and concordance analyses for the words "الجنسية / aljensyt / the nationality", "جنسية / jensyt / nationality", "الجنسيَات / aljensyat / the nationalities" and "جنسيَات / jensyat / nationalities" are as follows:

- الجنسية / aljensyt / the nationality

Table 3.34 shows that the most frequent word collocating with "الجنسية / aljensyt / the nationality" is the preposition "من / min / from". This preposition occurs to the right (preceding) 406 times in various positions, and in 296 out of these 406 cases, it immediately precedes "الجنسية / aljensyt / the nationality". Consequently, there is a dependency relationship between them. The other tokens, such as the prepositions "علَى / ala / on", "في / fi / in" and the word "بن / bin / son of", are neglected because they have no strong relationship (i.e. Position -1) with the word being investigated.

Table 3.34: Collocation results for "الجنسية / aljensyt / the nationality"

The nationality		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	528	406	122	25	33	21	31	296	1	19	32	33	37
علَى	on	191	95	96	17	32	42	3	1	17	22	22	26	9
في	in	101	12	89	8	2	2	0	0	15	26	21	12	15
بن	son	79	63	16	33	27	3	0	0	0	0	0	9	7
سعودي	Saudi	74	74	0	0	1	1	0	72	0	0	0	0	0

On the other hand, the nationality word "سعودي / saudi / Saudi" occurs 72 times immediately before "الجنسية / aljensyt / the nationality". As a result, another analysis was carried out to investigate the most frequent words immediately occurring before the word "الجنسية / aljensyt / the nationality".

The results in Table 3.35 show that some nationalities appear with the suffix ”ي / ya / y”, such as ”سعودي / saudi / Saudi”, ”هندي / hndi / Indian” and ”بَاكستَاني / bakstani / Pakistani”. However, the preposition ”من / min / from” is still at the top of the list.

Table 3.35: Collocation results for ”الجنسية / aljensyt / the nationality” at Position -1

The nationality		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	528	406	122	25	33	21	31	296	1	19	32	33	37
سعودي	Saudi	74	74	0	0	1	1	0	72	0	0	0	0	0
هندي	Indian	31	31	0	0	0	0	0	31	0	0	0	0	0
بَاكستَاني	Pakistani	27	27	0	0	0	0	0	27	0	0	0	0	0
نفس	same	17	14	3	0	0	0	0	14	0	0	0	0	3

The results of our investigation into the words that directly follow the word ”الجنسية / aljensyt / the nationality” can be seen in Table 3.36. All these words represent different nationalities.

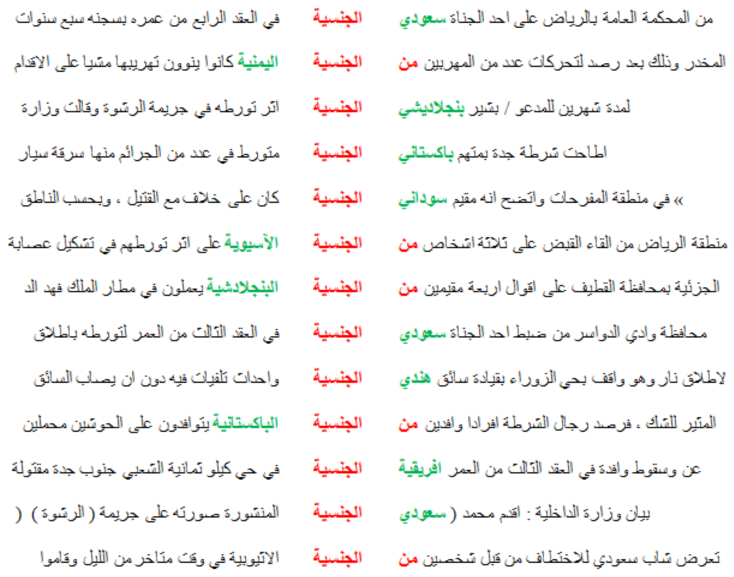
Table 3.36: Collocation result for ”الجنسية / aljensyt / the nationality” at Position +1

The nationality		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
اليمنية	the Yemeni	44	3	41	0	1	2	0	0	39	0	0	0	2
الآسيوية	the Asian	39	0	39	0	0	0	0	0	39	0	0	0	0
البَاكستَانية	the Pakistani	29	0	29	0	0	0	0	0	29	0	0	0	0
الهندية	the Indian	33	0	33	0	0	0	0	0	29	0	0	0	4
البنغَالية	the Bangladeshi	22	2	20	0	0	2	0	0	20	0	0	0	0

Therefore, we may deduce from these two analyses that a nationality can occur before or after the main word ”الجنسية / aljensyt / the nationality”. It can be noticed that the nationalities in Table 3.35 have suffixes ending with ”ي / iyy” but in Table

3.36 they end with " ية / iyyt". In addition, according to the concordance analysis, it is found that the word that directly follows the word " الجنسية / aljensyt / the nationality" is a nationality word when " الجنسية / aljensyt / the nationality" is assigned the genitive case by the preposition " من / min / from". In contrast, if it is not preceded by a preposition, the word that precedes it instead of that preposition is a nationality word.

Figure 3.14 below presents a sample of the occurrence of the word " الجنسية / aljensyt / the nationality". Both cases can be seen, and the occurrence of nationality type before or after is controlled by the preposition " من / min / from", i.e. by the syntactic construction of " الجنسية / aljensyt / the nationality", whether or not it is in the genitive case.



في العقد الرابع من عمره بسجنه سبع سنوات	الجنسية	من المحكمة العامة بالرياض على احد الجناة سعودي
اليمنية كانوا ينوون تهريبها مشيا على الاقدام	الجنسية	المخدر وذلك بعد رصد لتحركات عدد من المهربين من
اثر تورطه في جريمة الرشوة وقالت وزارة	الجنسية	لمدة شهرين للمدعو / بشير بنجلاديشي
متورط في عدد من الجرائم منها سرقة سيار	الجنسية	اطاحت شرطة جدة بمتهم باكستاني
كان على خلاف مع القتيل ، وبحسب الناطق	الجنسية	» في منطقة المفرحات واتضح انه مقيم سوداني
الآسيوية على اثر تورطهم في تشكيل عصابة	الجنسية	منطقة الرياض من القاء القبض على ثلاثة اشخاص من
البنجلادشية يعملون في مطار الملك فهد الد	الجنسية	الجزئية بمحافظة القطيف على اقوال اربعة مقيمين من
في العقد الثالث من العمر لتورطه باطلاق	الجنسية	محافظة وادي الدواسر من ضبط احد الجناة سعودي
واحدات تلفيات فيه دون ان يصاب السائق	الجنسية	لاطلاق نار وهو واقف بحي الزوراء بقيادة سائق هندي
الباكستانية يتوافدون على الحوشين محملين	الجنسية	المثير للشك ، فرصد رجال الشرطة افرادا وافدين من
في حي كيلو ثمانية الشعبي جنوب جدة مقتولة	الجنسية	عن وسقوط وافدة في العقد الثالث من العمر افريقية
المنشورة صورته على جريمة (الرشوة) (	الجنسية	بيان وزارة الداخلية : اقدم محمد (سعودي
الاثيوبية في وقت متاخر من الليل وقاموا	الجنسية	تعرض شاب سعودي للاختطاف من قبل شخصين من

Figure 3.14: The concordance lines for " الجنسية / aljensyt / the nationality"

- جنسية / jensyt / nationality

With regard to the word " جنسية / jensyt / nationality", Table 3.37 shows that the preposition " من / min / from" is most frequent word collocating with it. The preposition appears 190 to the right, and it occurs immediately before " جنسية /

jensyt / nationality” as often as 165 times. Furthermore, no other words occur at Position -1. Also, it can be seen that the word ”عربية / arabyt / Arab”, which is the second most frequent word, appears to the left 86 times but does not appear at all to the right; its occurrence is always directly after the word ”جنسية / jensyt / nationality” (Position +1). Another nationality, ”آسيوية / asywyt / Asian”, can be seen at Position +1.

Table 3.37: Collocation results for ”جنسية / jensyt / nationality”

Nationality		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	218	190	28	15	4	5	1	165	3	2	8	3	12
عربية	Arab	86	0	86	0	0	0	0	0	86	0	0	0	0
في	in	55	16	39	10	3	1	2	0	0	11	12	9	7
علَى	on	50	41	9	2	17	20	2	0	0	1	4	1	3
آسيوية	Asian	38	1	37	0	1	0	0	0	34	0	2	0	1

More two analyses were undertaken in order to obtain the most frequent words that immediately precede and follow the word ”جنسية / jensyt / nationality”. The results can be seen in Table 3.38 and Table 3.39. Table 3.38 shows that the word most often occurring directly before ”جنسية / jensyt / nationality” is the preposition ”من / min / from”. Therefore, the preposition governs this word and assigns the genitive case to it. Table 3.39 shows that the nationality words only occur directly after ”جنسية / jensyt / nationality”.

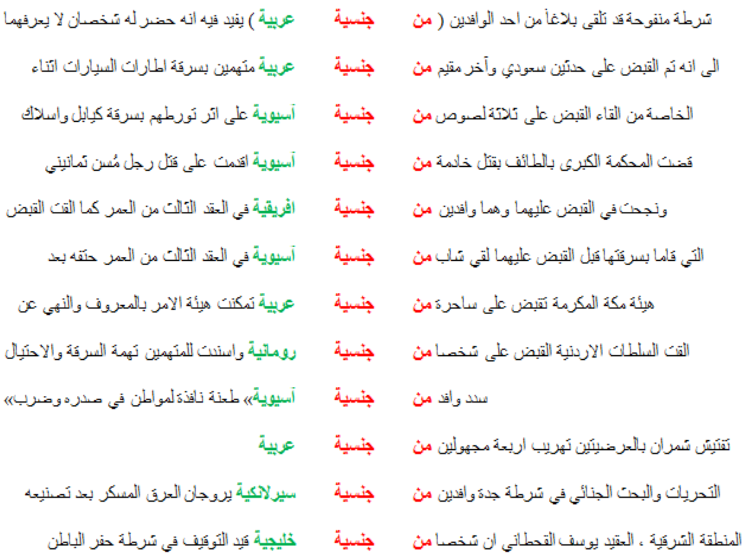
Table 3.38: Collocation results for "جنسية / jensyt / nationality" at Position -1

Nationality		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	218	190	28	15	4	5	1	165	3	2	8	3	12
اعمَال	works	8	4	4	0	0	0	0	4	0	0	3	1	0
يحمل	carry	5	3	2	0	0	0	0	3	0	1	1	0	0
(من	from	3	3	0	0	0	0	0	3	0	0	0	0	0
مشَاهد	scenes	3	3	0	0	0	0	0	3	0	0	0	0	0

Table 3.39: Collocation results for "جنسية / jensyt / nationality" at Position +1

Nationality		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
عربية	Arab	86	0	86	0	0	0	0	0	86	0	0	0	0
آسيوية	Asian	38	1	37	0	1	0	0	0	34	0	2	0	1
افريقية	African	23	1	22	0	0	1	0	0	22	0	0	0	0
بَاكستَانية	Pakistani	4	1	3	0	0	1	0	0	3	0	0	0	0
من	from	218	190	28	15	4	5	1	165	3	2	8	3	12

Furthermore, a sample result of the concordance analysis of this word confirms that if it is preceded by the preposition "من / min / from" (i.e. assigned the genitive case), then the following word is a nationality. Different nationalities can be seen in the following Figure 3.15, such as "عربية / arabyt / Arab", "آسيوية / asywyt / Asian", "افريقية / afryqyt / African" and "رومَانية / rwmanyt / Romanian".



شرطة منفوحة قد تلقى بلاغاً من احد الوافدين (من جنسية عربية) يفيد فيه انه حضر له شخصان لا يعرفهما

الى انه تم القبض على حدثين سعودي وآخر مقيم من جنسية عربية متهمين بسرقة اطارات السيارات اثناء

الخاصة من القاء القبض على ثلاثة لصوص من جنسية آسيوية على اثر تورطهم بسرقة كيابل واسلاك

قضت المحكمة الكبرى بالطائف بقتل خادمة من جنسية آسيوية اقدمت على قتل رجل مُسن ثمانيني

ونجحت في القبض عليهما وهما وافدين من جنسية افريقية في العقد الثالث من العمر كما القت القبض

التي قاما بسرقتها قبل القبض عليهما لقي شاب من جنسية آسيوية في العقد الثالث من العمر حتفه بعد

هيئة مكة المكرمة تقبض على ساحرة من جنسية عربية تمكنت هيئة الامر بالمعروف والنهي عن

القت السلطات الاردنية القبض على شخصا من جنسية رومانية واسندت للمتهمين تهمة السرقة والاحتيال

سدد وافد من جنسية آسيوية» طعنة نافذة لمواطن في صدره وضرب»

تفتيش شمران بالعرضيتين تهريب اربعة مجهولين من جنسية عربية

التحريات والبحث الجنائي في شرطة جدة وافدين من جنسية سيرلانكية يروجان العرق المسكر بعد تصنيعه

المنطقة الشرقية ، العقيد يوسف القحطاني ان شخصا من جنسية خليجية قيد التوقيف في شرطة حفر الباطن

Figure 3.15: The concordance lines for ” جنسية / jensyt / nationality”

- جنسيَات / jensyat / nationalities and الجنسيَات / aljensyat / the nationalities

Table 3.40 and Table 3.41 show the results of the collocation analyses for ” جنسيَات / jensyat / nationalities” and ” الجنسيَات / aljensyat / the nationalities”, respectively. Clearly, the preposition ” من / min / from” is the most frequent word collocating with both words, and it immediately occurs before them (Position -1).

Table 3.40: Collocation results for ” جنسيَات / jensyat / nationalities”

Nationalities		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	67	62	5	3	7	1	1	50	0	1	1	3	0
مختلفة	different	28	0	28	0	0	0	0	0	28	0	0	0	0
علَى	on	17	15	2	5	6	4	0	0	0	0	0	1	1
في	in	14	0	14	0	0	0	0	0	0	4	3	4	3
اشخَاص	persons	8	8	0	0	1	0	7	0	0	0	0	0	0

Table 3.41: Collocation results for "الجنسيَات / aljensyat / the nationalities"

The Nationalities		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	43	38	5	2	4	2	25	5	0	0	4	0	1
الآسيوية	Asian	8	1	7	1	0	0	0	0	7	0	0	0	0
علَى	on	8	4	4	1	3	0	0	0	0	0	0	2	2
مختلف	different	7	7	0	0	0	0	0	7	0	0	0	0	0
مجموعة	groups	4	4	0	3	0	1	0	0	0	0	0	0	0

Moreover, another two analyses were carried out to obtain the most five frequent words that immediately precede (Position -1) and follow (Position +1) both words. The results can be seen in Table 3.42 and Table 3.43. They show that there is no occurrence for nationality words preceding "جنسيَات / jensyat / nationalities" and "الجنسيَات / aljensyat / the nationalities".

Table 3.42: Collocation results for "جنسيَات / jensyat / nationalities" at Position -1

Nationalities		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
من	from	67	62	5	3	7	1	1	50	0	1	1	0	0
يحملون	carrying	1	1	0	0	0	0	0	1	0	0	0	0	0
عدة	several	1	1	0	0	0	0	0	1	0	0	0	0	0
ومن	And from	1	1	0	0	0	0	0	1	0	0	0	0	0
وَافدين	expatriates	4	4	0	0	0	1	3	0	0	0	0	0	0

Table 3.43: Collocation results for "الجنسيَات / aljensyat / the nationalities" at Position -1

The nationalities		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
احدَى	one	16	16	0	0	0	0	0	16	0	0	0	0	0
مختلف	different	7	7	0	0	0	0	0	7	0	0	0	0	0
من	from	43	38	5	2	4	2	25	5	0	0	4	0	1
ابنَا	sons	1	1	0	0	0	0	0	1	0	0	0	0	0
في	in	3	1	2	0	0	0	0	1	0	1	0	1	0

On the other hand, nationality words can be seen in Table 3.44 and Table 3.45, and they only occur to the left at (Position +1), i.e. immediately following "جنسيَات / jensyat / nationalities" and "الجنسيَات / aljensyat / the nationalities".

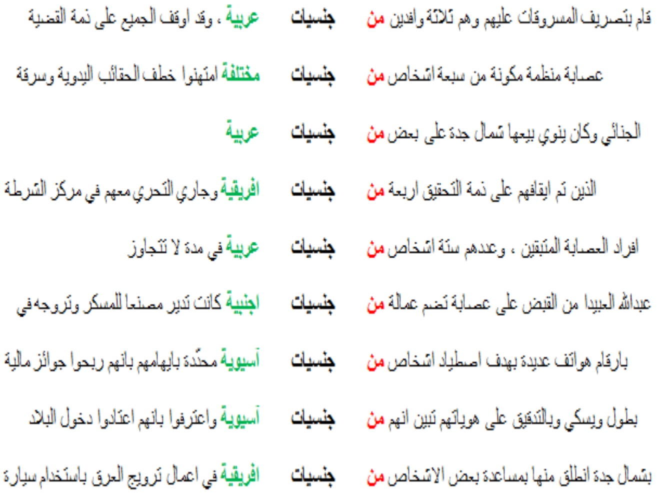
Table 3.44: Collocation results for "جنسيَات / jensyat / nationalities" at Position +1

Nationalities		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
مختلفة	different	28	0	28	0	0	0	0	0	28	0	0	0	0
افريقية	African	7	0	7	0	0	0	0	0	7	0	0	0	0
عربية	Arab	6	0	6	0	0	0	0	0	6	0	0	0	0
آسيوية	Asian	6	0	6	0	0	0	0	0	4	2	0	0	0
امَارتيه	Emirates	3	0	3	0	0	0	0	0	3	0	0	0	0

Table 3.45: Collocation results for "الجنسيَات / aljensyat / the nationalities" at Position +1

The nationalities		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
الآسيوية	the Asian	8	1	7	1	0	0	0	0	7	0	0	0	0
العربية	the Arab	6	0	6	0	0	0	0	0	6	0	0	0	0
الَافريقية	the African	3	0	3	0	0	0	0	0	3	0	0	0	0
الَاجنبية	the foreign	2	0	2	0	0	0	0	0	2	0	0	0	0
الشرق	the east	2	0	2	0	0	0	0	0	2	0	0	0	0

Additionally, the concordance analysis results for both words can be seen in Figure 3.16 and Figure 3.17, respectively. This analysis proves that when the both of these words are the assigned genitive case, the word that follows them is a nationality.



قام بتصريف المسروقات عليهم وهم ثلاثة وافدين من جنسيات عربية ، وقد اوقف الجميع على ذمة القضية

عصابة منظمة مكونة من سبعة اشخاص من جنسيات مختلفة امتهنوا خطف الحقائب اليدوية وسرقة

الجنائي وكان ينوي بيعها شمال جدة على بعض من جنسيات عربية

الذين تم ايقافهم على ذمة التحقيق اربعة من جنسيات افريقية وجاري التحري معهم في مركز الشرطة

افراد العصابة المتبقين ، وعددهم ستة اشخاص من جنسيات عربية في مدة لا تتجاوز

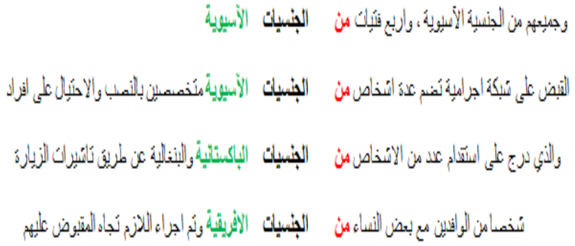
عبدالله الحبيدا من القبض على عصابة تضم عمالة من جنسيات اجنبية كانت تدير مصنعا للمسكر وتروجه في

بارقام هواتف عديدة بهدف اصطياد اشخاص من جنسيات آسيوية محدّدة بايهامهم بانهم ربحوا جوائز مالية

بطول ويسكي وبالتدقيق على هوياتهم تبين انهم من جنسيات آسيوية واعترفوا بانهم اعتادوا دخول البلاد

بشمال جدة انطلق منها بمساعدة بعض الاشخاص من جنسيات افريقية في اعمال ترويج العرق باستخدام سيارة

Figure 3.16: The concordance lines for "جنسيَات / jensyat / nationalities"



وجميعهم من الجنسية الأسيوية ، واربع فتيات من الجنسيات الأسيوية

القبض على شبكة اجرامية تضم عدة اشخاص من الجنسيات الأسيوية متخصصين بالنصب والاحتيال على افراد

والذي درج على استقدام عدد من الاشخاص من الجنسيات الباكستانية والبنغالية عن طريق تاشيرات الزيارة

شخصا من الوافدين مع بعض النساء من الجنسيات الافريقية وتم اجراء اللازم تجاه المقبوض عليهم

Figure 3.17: The concordance lines for "الجنسيَات / aljensyat / the nationalities"

The above analysis shows that a nationality entity occurs within the context of the word "جنسية / jensyt / nationality" and its inflected form when they are assigned genitive case by the preposition "من / min / from". Hence, this syntactic construction will be used in order to recognize this type of entity.

3.5.2 Nationality Local Grammar

The above analysis shows that there is strong relationship between the word "جنسية / jensyt / nationality" (in any inflected form) and the preposition "من / min / from". As a result, from the syntactic point of view, the word is often in the genitive case. Accordingly, the words "جنسية / jensyt / nationality" (singular), "الجنسية / aljensyt / the nationality" (definite singular), "جنسيَات / jensyat / nationalities" (plural) and "الجنسيَات / aljensyat / the nationalities" (definite plural) can all be used for building the indicator nodes. They can be described as triggering keywords; they trigger the indicator nodes into recognising and extracting the nationality (with syntactic constraint). This linguistic condition is activated when the above words are assigned the genitive case by the preposition "من / min / from". As mentioned early, there is one exception: "الجنسية / aljensyt / the nationality"; when this word is not preceded by the preposition "من / min / from", the word that occurs instead of the preposition is considered to indicate nationality.

Thus, the indicator nodes for extracting nationality from text are as follows:

- Name: الجنسية / aljensyt / the nationality-genitive-من / min / from

  Trigger: الجنسية / aljensyt / the nationality

  Target: nationality

  Constraint: genitive

  Activation: trigger is preceded by the preposition ”من / min / from”

- Name: الجنسية / aljensyt / the nationality-notgenitive-من / min / from

  Trigger: الجنسية / aljensyt / the nationality

  Target: nationality

  Constraint: not in genitive case

  Activation: trigger is not preceded by the preposition ”من / min / from”

- Name: جنسية / jensyt / nationality-genitive-من / min / from

  Trigger: جنسية / jensyt / nationality

  Target: nationality type

  Constraint: genitive

  Activation: trigger is preceded by the preposition ”من / min / from”

- Name: الجنسيّات / aljensyat / the nationalities-genitive-من / min / from

  Trigger: الجنسيّات / aljensyat / the nationalities

  Target: nationality

  Constraint: genitive

  Activation: trigger is preceded by the preposition ”من / min / from”

- Name: جنسيّات / jensyat / nationalities-genitive-من / min / from

  Trigger: جنسيّات / jensyat / nationalities

  Target: nationality

  Constraint: genitive

  Activation: trigger is preceded by the preposition "من / min / from"

The following Figure 3.18 depicts the nationality local grammar that was generated based on the syntactic analysis.

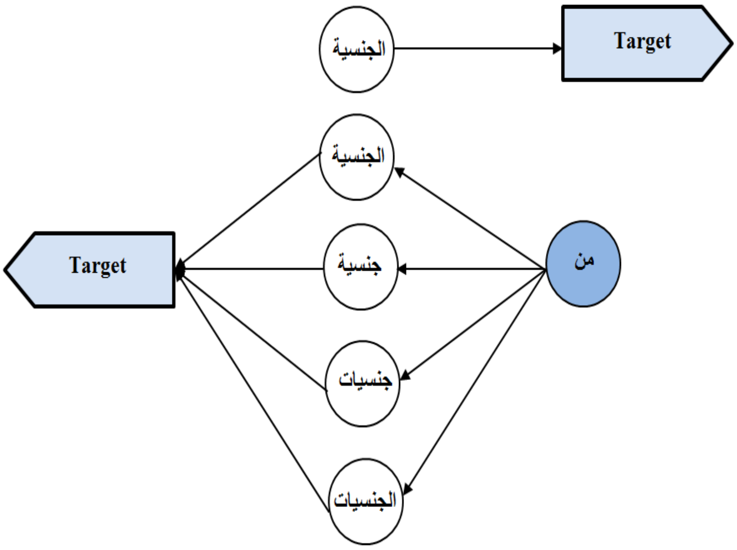


Figure 3.18: Nationality local grammar

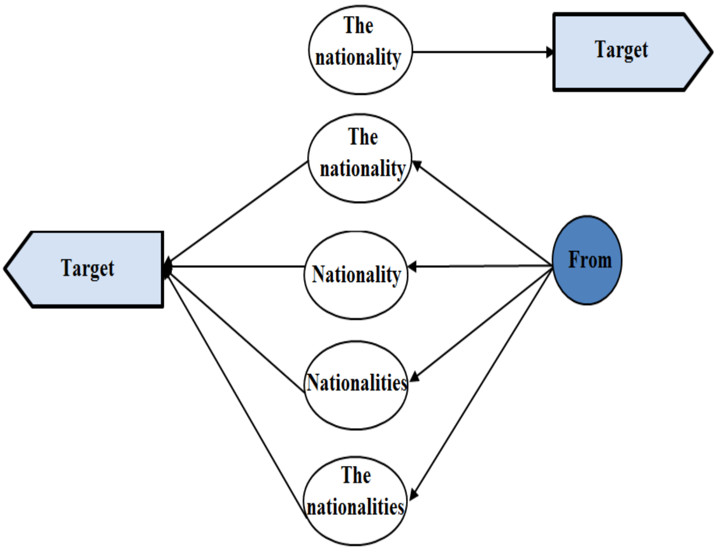


Figure 3.19: Nationality local grammar (English Version)

3.6 Location

The words in Table 3.46, which are often used for stating a place name, are analysed in this section. The words "منطقة / mantqt / area" and "بمنطقة / bi-mantqt / in area" are the same in terms of meaning but the word "بمنطقة / bi-mantqt / in area" is attached to the preposition "ب / bi / in". This composition, as already described, is a prepositional phrase, which is a type of the genitive case. Table 3.46 presents the comparison results for these words between the special corpus (ACNRC) and the general one.

Table 3.46: Occurrence of location words in the ACNRC and general Corpora

Verb	**Pronunciation**	**Translation**	**ACNRC**	**G.Corpus**
منطقة	mntqt	area	1175	1816
محافظة	muhafdt	province	574	480
بمنطقة	bi-mntqt	in area	511	133
مدينة	mdinat	city	313	2565

Likewise, collocation and concordance analyses for the above words were performed in order to investigate their contexts and to obtain a fuller picture. The results show that place names (crime locations) are often represented by words that immediately follow the above words. This is similar to the English compositions 'city of', 'province of' and 'region of', all of which constitute a construct state in terms of grammar. Accordingly, these words are chosen as keywords for triggering the indicator nodes in order to recognize and extract a crime location. Although these words are often assigned the genitive case as objects of specific prepositions in prepositional phrases, or as the second noun of specific construct heads in construct sates, there is no need to apply a syntactic constraint.

3.6.1 Analysis of Location Words

The collocation and concordance analyses for the words in Table 3.46 are as follows:

- منطقة / mantqt / area

Table 3.47 presents the results of the collocation analysis. The collocation analysis clearly shows that the word most frequently associating with "منطقة / mantqt / area" is the preposition "في / fi / in"; it is to the right (preceding) and it occurs 381 times directly before the word "منطقة / mantqat / area". Also, another preposition, "من / min / from", can be seen, and it occurs 36 times in Position -1. The word "شرطة / shurtat / police", which is within the list of 100 tokens, appears 323 times out of 329 directly preceding the word "منطقة / mantqat / area". Moreover, the last row has the word "بشرطة / bi-shurtat / in police", which is in reality a prepositional phrase because the preposition "ب / bi / in " is attached to the word "شرطة / shurtat / police", forming one word ("بشرطة / bi-shurtat / in police"). This word is also within the 100 words list, and it occurs 2,166 times in the ACNRC. On the other hand, the word "الريّاض / alriyad / Riyadh", which is a city name, occurs 286 times immediately after the word being investigated.

Table 3.47: Collocation results for " منطقة / mantqt / area"

Area		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	622	527	95	30	15	20	81	381	0	30	23	14	28
شرطة	police	335	329	6	1	2	3	0	323	0	0	1	2	3
من	from	316	142	174	31	37	23	15	36	2	42	33	40	57
الريَاض	Riyadh	302	7	295	5	0	1	1	0	286	2	0	3	4
بشرطة	in police	140	139	1	0	0	1	0	138	0	0	1	0	0

Also, another two analyses were carried out to obtain the most frequent words at Position -1 and Position +1. The results are listed in Table 3.48 and Table 3.49. Position -1 in Table 3.48 shows the first three words (in, police and from) already seen in the above Table 3.47, and two new words, which are the preposition " الَى / ila / to" and the word " لشرطة / li-shurtat / to police". The word " لشرطة / li-shurtat / to police" is a prepositional phrase because the preposition " ل / li / of" is attached to it, forming a single word. From the syntactic point of view, the words "bi-shurtat" and "li-shurtat" are considered governors and assign the genitive case to the word " منطقة / mantqt / area".

Table 3.48: Collocation results for " منطقة / mantqt / area" at Position -1

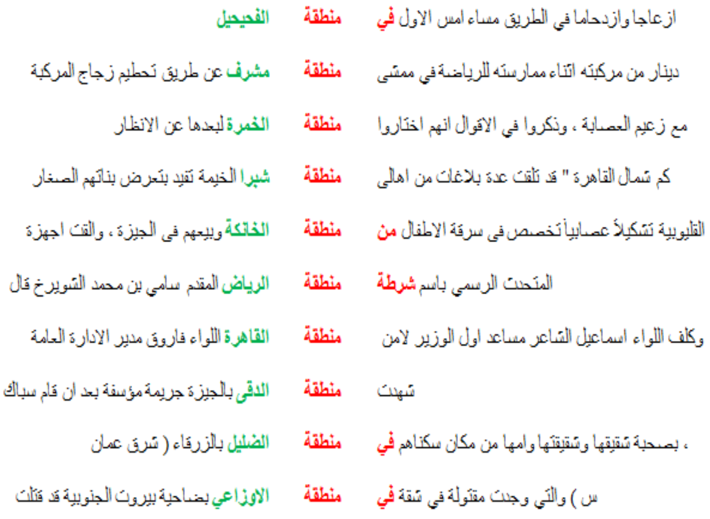
Area		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	622	527	95	30	15	20	81	381	0	30	23	14	28
شرطة	police	335	329	6	1	2	3	0	323	0	0	1	2	3
بشرطة	in police	140	139	1	0	0	1	0	138	0	0	1	0	0
الَى	to	145	98	47	12	8	7	12	59	0	8	13	13	13
لشرطة	to police	43	43	0	1	0	0	0	42	0	0	0	0	0

With regards to the words in Table 3.49 that occur at Position +1, i.e. immediately following " منطقة / mantqt / area", all of them are the names of cities. This is similar to English where the phrases 'area of' and 'region of' are used before what we expect to be a place name. Moreover, in terms of grammar, they form a construct state because the word " منطقة / mantqt / area" inherits its definiteness from the name of the city that follows it.

Table 3.49: Collocation results for " منطقة / mantqt / area" at Poistion +1

Area		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
الريَاض	Riyadh	302	7	295	5	0	1	1	0	286	2	0	3	4
عسير	Asir	78	0	78	0	0	0	0	0	78	0	0	0	0
نجرَان	Najran	54	2	52	2	0	0	0	0	52	0	0	0	0
البَاحة	Al Baha	44	0	44	0	0	0	0	0	44	0	0	0	0
جَازَان	Jazan	41	1	40	1	0	0	0	0	40	0	0	0	0

The following Figure 3.20 presents a sample of the concordance analysis. The results show that all the words immediately following " منطقة / mantqt / area" are the names of cities, even though the word " منطقة / mantqt / area" is sometimes not preceded by the common words already seen, such as " بشرطة / bi-shurtat / in police", " لشرطة / li-shurtat / to police" and " في / fi / in".



ازعاجا وازدحاما في الطريق مساء امس الاول في منطقة الفحيحيل

دينار من مركبته اثناء ممارسته للرياضة في ممشى منطقة مشرف عن طريق تحطيم زجاج المركبة

مع زعيم العصابة ، وذكروا في الاقوال انهم اختاروا منطقة الخمرة لبعدها عن الانظار

كم شمال القاهرة " قد تلقت عدة بلاغات من اهالي منطقة شبرا الخيمة تفيد بتعرض بناتهم الصغار

القليوبية تشكيلاً عصابياً تخصص فى سرقة الاطفال من منطقة الخانكة وبيعهم فى الجيزة ، والقت اجهزة

المتحدث الرسمي باسم شرطة منطقة الرياض المقدم سامي بن محمد الشويرخ قال

وكلف اللواء اسماعيل الشاعر مساعد اول الوزير لامن منطقة القاهرة اللواء فاروق مدير الادارة العامة

شهدت منطقة الدقى بالجيزة جريمة مؤسفة بعد ان قام سباك

، بصحبة شقيقها وشقيقتها وامها من مكان سكناهم في منطقة الضليل بالزرقاء (شرق عمان

س) والتي وجدت مقتولة في شقة في منطقة الاوزاعي بضاحية بيروت الجنوبية قد قتلت

Figure 3.20: The concordance lines for " منطقة / mantqt / area"

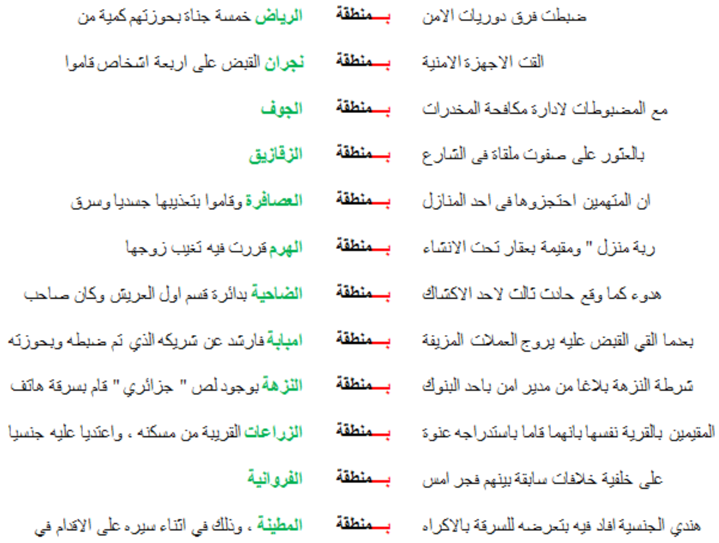
- بمنطقة / bi-mantqt / in area

With regard to the word ”بمنطقة / bi-mantqt / in area”, this word is a prepositional phrase because it is a combination of the preposition ”ب / bi / in” and the word ”منطقة / mantqt / area”. As a result, this word is governed and assigned the genitive case by the preposition ”ب / bi / in”. Clearly, because this composition ”بمنطقة / bi-mantqt / in area” is prepositional phrase, there is no need to investigate its collocation from the right. Table 3.50 shows the results of the collocation analysis of the word (prepositional phrase) ”بمنطقة / bi-mantqt / in area” for the words that directly follow it (at Position +1). It can be seen that all the words are the names of areas.

Table 3.50: Collocation results for ”بمنطقة / bi-mantqt / in area”

In area		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
الريَاض	Riyadh	52	2	50	0	0	0	0	2	49	0	0	0	1
نجرَان	Najran	24	0	24	0	0	0	0	0	24	0	0	0	0
مكة	Mecca	22	1	21	0	0	0	0	1	20	0	1	0	0
البَاحة	Al-Baha	18	0	18	0	0	0	0	0	18	0	0	0	0
السَالمية	Al-Salmit	11	0	11	0	0	0	0	0	11	0	0	0	0

Figure 3.21 shows the concordance analysis for this word. It can be seen words such as "الهرم / alhram / Pyramid", "الريَاض / alriyad / Riyadh" and "الجوف /aljouf / Jouf" are the names of areas, and that they occur directly after the prepositional phrase "بمنطقة / bi-mantqt / in area".



ضبطت فرق دوريات الامن بمنطقة الرياض خمسة جناة بحوزتهم كمية من

القت الاجهزة الامنية بمنطقة نجران القبض على اربعة اشخاص قاموا

مع المضبوطات لادارة مكافحة المخدرات بمنطقة الجوف

بالعثور على صفوت ملقاة فى الشارع بمنطقة الزقازيق

ان المتهمين احتجزوها فى احد المنازل بمنطقة العصافرة وقاموا بتعذيبها جسديا وسرق

ربة منزل " ومقيمة بعقار تحت الانشاء بمنطقة الهرم قررت فيه تغيب زوجها

هدوء كما وقع حادث ثالث لاحد الاكشاك بمنطقة الضاحية بدائرة قسم اول العريش وكان صاحب

بعدما القي القبض عليه يروج العملات المزيفة بمنطقة امبابة فارشد عن شريكه الذي تم ضبطه وبحوزته

شرطة النزهة بلاغا من مدير امن باحد البنوك بمنطقة النزهة بوجود لص " جزائري " قام بسرقة هاتف

المقيمين بالقرية نفسها بانهما قاما باستدراجه عنوة بمنطقة الزراعات القريبة من مسكنه ، واعتديا عليه جنسيا

على خلفية خلافات سابقة بينهم فجر امس بمنطقة الفروانية

هندي الجنسية افاد فيه بتعرضه للسرقة بالاكراه بمنطقة المطينة ، وذلك في اثناء سيره على الاقدام في

Figure 3.21: The concordance lines for "بمنطقة / bi-mantqt / in area"

- محَافظة / mohafdat / province

The word "محَافظة / mohafdat" means province in English. To investigate its context, a collocation analysis was preformed, and the results are in Table 3.51. As can be seen, the words "في / fi / in", "من / min / from", "شرطة / shurtat / police" and "الَى / ila / to" that appear in this result are already in the collocation results for "منطقة / mantqt / area", i.e. they share the same context.

Table 3.51: Collocation results for ” محافظة / mohafdat / province”

Province		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	294	249	45	10	15	12	31	181	0	12	13	10	10
شرطة	police	189	186	3	1	7	2	0	176	0	0	1	0	2
من	from	189	96	93	21	23	13	9	30	0	21	19	21	32
علَى	on	86	12	74	7	2	3	0	0	0	11	25	21	17
الَى	to	71	44	27	5	5	5	6	23	1	6	4	7	9

Moreover, two collocation analyses were accomplished in order to investigate the words most frequently occurring immediately before and after the word ” محافظة / mohafadt / province”. Table 3.52 presents the results of the words most frequently occurring directly before the word ” محافظة / mohafdat / province”. These words are the same as those that appeared in the collocation of ” منطقة / mantqt / area”. Also, they assign the genitive case to the word after them (” محافظة / mohafdat / province”) because they are either prepositions or construct heads of the construct state.

Table 3.52: Collocation results for ” محافظة / mohafdat / province” at Position -1

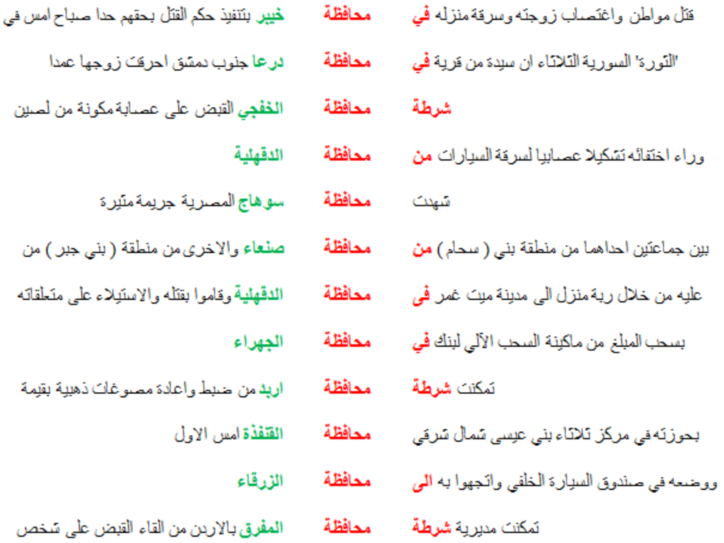
Province		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	294	249	45	10	15	12	31	181	0	12	13	10	10
شرطة	police	189	186	3	1	7	2	0	176	0	0	1	0	2
من	from	189	96	93	21	23	13	9	30	0	21	19	21	32
الَى	to	71	44	27	5	5	5	6	23	1	6	4	7	9
بشرطة	in police	13	13	0	0	0	1	0	12	0	0	0	0	0

On the other hand, Table 3.53 provides the words most frequently occurring at Position +1. The same results as those already discussed with the words ” منطقة / mantqt / area” and ” بمنطقة / bi-mantqt / in area” were obtained. All the words in Position +1 are province names.

Table 3.53: Collocation results for " مُحَافظة / mohafdat / province" at Position +1

Province		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
جدة	Jeddah	48	1	47	0	1	0	0	0	46	0	0	1	0
الطَاءيف	Taif	32	1	31	1	0	0	0	0	30	0	0	0	1
الَاحسَاء	Al-Ahsa	22	0	22	0	0	0	0	0	22	0	0	0	0
الجهرَاء	Al-Jahra	22	0	22	0	0	0	0	0	21	1	0	0	0
بلجرشي	Baljurashi	17	0	17	0	0	0	0	0	17	0	0	0	0

Figure 3.22 shows the concordance analysis results for this word. It can be seen that all the words that occur directly after the prepositional phrase " مُحَافظة / mohafdat / province" are the names of provinces.



قتل مواطن واغتصاب زوجته وسرقة منزله في محافظة خيبر بتنفيذ حكم القتل بحقهم حدا صباح امس في

'الثورة' السورية الثلاثاء ان سيدة من قرية في محافظة درعا جنوب دمشق احرقت زوجها عمدا

شرطة محافظة الخفجي القبض على عصابة مكونة من لصين

وراء اختفائه تشكيلا عصابيا لسرقة السيارات من محافظة الدقهلية

شهدت محافظة سوهاج المصرية جريمة مثيرة

بين جماعتين احداهما من منطقة بني (سحام) من محافظة صنعاء والاخرى من منطقة (بني جبر) من

عليه من خلال ربة منزل الى مدينة ميت غمر في محافظة الدقهلية وقاموا بقتله والاستيلاء على متعلقاته

بسحب المبلغ من ماكينة السحب الآلي لبنك في محافظة الجهراء

تمكنت شرطة محافظة اربد من ضبط واعادة مصوغات ذهبية بقيمة

بحوزته في مركز ثلاثاء بني عيسى شمال شرقي محافظة القنفذة امس الاول

ووضعه في صندوق السيارة الخلفي واتجهوا به الى محافظة الزرقاء

تمكنت مديرية شرطة محافظة المفرق بالاردن من القاء القبض على شخص

Figure 3.22: The concordance lines for " مُحَافظة / mohafdat / province"

- مدينة / mdynt / city

Table 3.54 lists the most frequent words in this collocation analysis. The prepositions "في / fi / in", "من / min / from" and "الَى / ila / to" are the most frequent words occurring immediately before this word. Table 3.55, which contains only the words most frequently occurring directly before word " مدينة / mdynt / city", confirms that these prepositions are most frequent words at Position -1 as well as the word " شرطة / shurtat / police". In Table 3.54, the word "الريَاض /alriyadh / Riyadh", which is the name of a city, occurs immediately after the " مدينة / mdynt / city" 43 times. Moreover, Table 3.56 confirms that all the words that directly follow " مدينة / mdynt / city" are city names.

Table 3.54: Collocation results for " مدينة / mdynt / city"

City		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	165	145	20	9	27	7	6	96	0	11	3	3	3
من	from	98	70	28	14	12	8	7	29	0	3	2	10	13
الريَاض	Riyadh	45	2	43	1	0	0	1	0	43	0	0	0	0
علَى	on	40	17	23	5	7	5	0	0	0	8	2	6	7
الَى	to	34	26	8	0	1	4	4	17	0	1	1	1	5

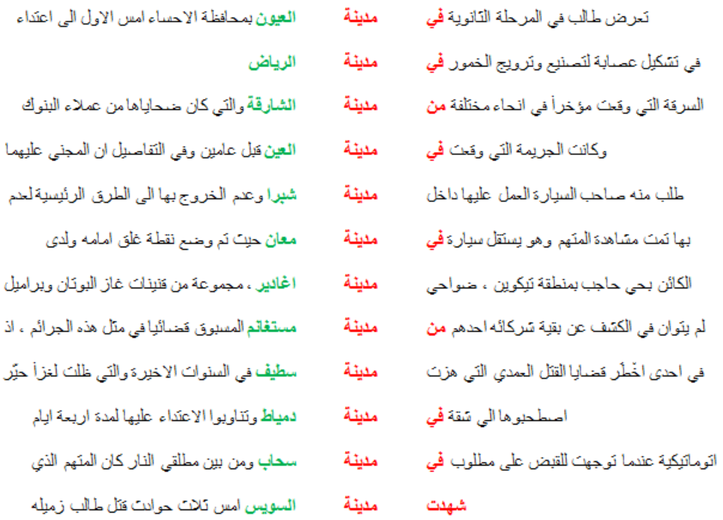
Table 3.55: Collocation results for " مدينة / mdynt / city" at Position -1

City		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	165	145	20	9	27	7	6	96	0	11	3	3	3
من	from	98	70	28	14	12	8	7	29	0	3	2	10	13
شهدت	had	18	18	0	0	0	0	0	18	0	0	0	0	0
الَى	to	34	26	8	0	1	4	4	17	0	1	1	1	5
شرطة	police	16	16	0	1	3	1	0	11	0	0	0	0	0

Table 3.56: Collocation results for ” مدينة / mdynt / city” at Position +1

City		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
الرِّيَاض	Riyadh	45	2	43	1	0	0	1	0	43	0	0	0	0
نصر	Nasr	21	0	21	0	0	0	0	0	21	0	0	0	0
جدة	Jeddah	21	0	21	0	0	0	0	0	20	1	0	0	0
البَاحة	Al-Baha	8	0	8	0	0	0	0	0	8	0	0	0	0
الدمَام	Al-Dammam	7	0	7	0	0	0	0	0	7	0	0	0	0

The results of the concordance analysis for ” مدينة / mdynt / city” are presented in Figure 3.23. All the words directly occurring after ” مدينة / mdynt / city” are the names of cities, even when they are not preceded by the prepositions or by the word ” شرطة / shurtat / police”, as in sentences 5, 7 and 9.



تعرض طالب في المرحلة الثانوية في مدينة العيون بمحافظة الاحساء امس الاول الى اعتداء

في تشكيل عصابة لتصنيع وترويج الخمور في مدينة الرياض

السرقة التي وقعت مؤخراً في انحاء مختلفة من مدينة الشارقة والتي كان ضحاياها من عملاء البنوك

وكانت الجريمة التي وقعت في مدينة العين قبل عامين وفي التفاصيل ان المجني عليهما

طلب منه صاحب السيارة العمل عليها داخل مدينة شبرا وعدم الخروج بها الى الطرق الرئيسية لعدم

بها تمت مشاهدة المتهم وهو يستقل سيارة في مدينة معان حيث تم وضع نقطة غلق امامه ولدى

الكائن بحي حاجب بمنطقة تيكوين ، ضواحي مدينة اغادير ، مجموعة من قنينات غاز البوتان وبراميل

لم يتوان في الكشف عن بقية شركائه احدهم من مدينة مستغانم المسبوق قضائيا في مثل هذه الجرائم ، اذ

في احدى اخطر قضايا القتل العمدي التي هزت مدينة سطيف في السنوات الاخيرة والتي ظلت لغزا حيّر

اصطحبوها الى شقة في مدينة دمياط وتناوبوا الاعتداء عليها لمدة اربعة ايام

اتوماتيكية عندما توجهت للقبض على مطلوب في مدينة سحاب ومن بين مطلقي النار كان المتهم الذي

شهدت مدينة السويس امس ثلاث حوادث قتل طالب زميله

Figure 3.23: The concordance lines for ” مدينة / mdynt / city”

- وِلَاية / wlayt / state

Despite the word ” وِلَاية / wlayt / state” not being within the list of 100 token, it has the same characteristics as all the above words. Table 3.57 and Table 3.58 present the words most frequently occurring immediately before and after ” وِلَاية / wlayt /

state". In Table 3.57, the words "في / fi / in", "شرطة / shurtat / police", "من / min / from " and "الَى / ila / to" have all been seen in collocation with the above words ("منطقة / mantqt / area", "محَافظة / mohafdat / province" and "مدينة / mdynt / city"). As can be seen in Table 3.58, all the words that follow the word "ولَاية / wlayt / state" at Position +1 are state names.

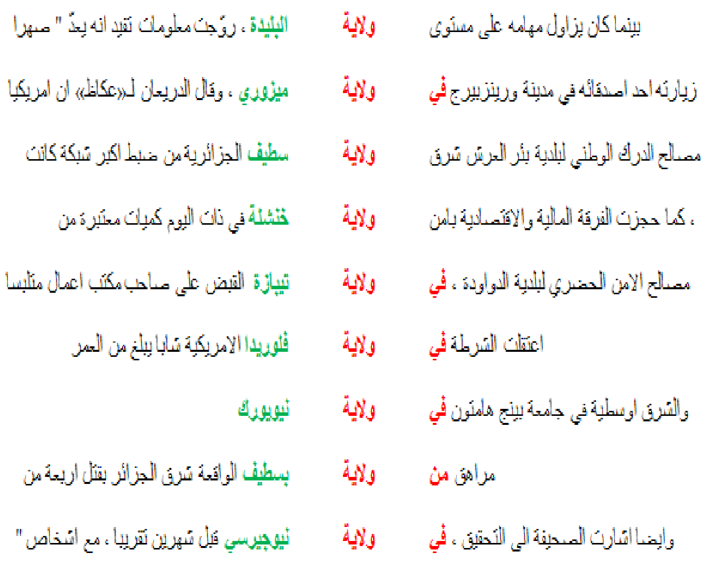
Table 3.57: Collocation results for "ولَاية / wlayt / state"

State		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
في	in	10	8	2	1	1	0	0	6	0	1	1	0	0
شرطة	police	2	2	0	0	0	0	0	2	0	0	0	0	0
من	from	6	2	4	0	0	0	0	2	0	2	1	0	1
الَى	to	4	4	0	0	2	1	0	1	0	0	0	0	0
شرق	east	2	1	1	0	0	0	0	1	0	0	1	0	0

Table 3.58: Collocation results for "ولَاية / wlayt / state" at Position +1

State		Positions												
Collocate	T	F	R	L	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
سطيف	Setif	3	0	3	0	0	0	0	0	3	0	0	0	0
نيويورك	New York	3	0	3	0	0	0	0	0	3	0	0	0	0
ميزوري	Missouri	2	0	2	0	0	0	0	0	2	0	0	0	0
تيبَازة	Tipaza	2	0	2	0	0	0	0	0	2	0	0	0	0
سيدي	Sidi	2	0	2	0	0	0	0	0	2	0	0	0	0

Figure 3.24 presents the results of the concordance analysis for the word "ولَاية / wlayt / state". As with the above words, all the words that immediately follow "ولَاية / wlayt / state of" are the names of sates, even when not preceded by prepositions such as "في / fi / in", "من / min / from" and "الَى / ila / to" or the word "شرطة / surtat / police".



البليدة ، روّجت معلومات تفيد انه يدّ " صهرا	ولاية	بينما كان يزاول مهامه على مستوى
ميزوري ، وقال الدريعان لـ«عكاظ» ان امريكيا	ولاية	زيارته احد اصدقائه في مدينة ورينزبيرج في
سطيف الجزائرية من ضبط اكبر شبكة كانت	ولاية	مصالح الدرك الوطني لبلدية بئر العرش شرق
خنشلة في ذات اليوم كميات معتبرة من	ولاية	، كما حجزت الفرقة المالية والاقتصادية بامن
تيبازة القبض على صاحب مكتب اعمال متلبسا	ولاية	مصالح الامن الحضري لبلدية الدواودة ، في
فلوريدا الامريكية شابا يبلغ من العمر	ولاية	اعتقلت الشرطة في
نيويورك	ولاية	والشرق اوسطية في جامعة بينج هامتون في
بسطيف الواقعة شرق الجزائر بقتل اربعة من	ولاية	مراهق من
نيوجيرسي قبل شهرين تقريبا ، مع اشخاص "	ولاية	وايضا اشارت الصحيفة الى التحقيق ، في

Figure 3.24: The concordance lines for "وِلَاية / wlayt / state"

The result of the analysis shows that place names (crime locations) are often represented by words that immediately follow the words "منطقة / mantqt / area", "بمنطقة / bi-mantqt / in area", "محَافظة / mohafdat / province", "مدينة / mdynt / city" and "وِلَاية / wlayt / state". This type of structure will be used for identifying crime locations.

3.6.2 Location Local Grammar

The analyses of the words "منطقة / mantqt / area", "بمنطقة / bi-mantqt / in area", "محَافظة / mohafdat / province", "مدينة / mdynt / city" and "وِلَاية / wlayt / state" provide a full picture of their behaviour within sentences. Furthermore, they show that their respective contexts contain place names. These names always immediately follow them, and therefore, these words are employed for constructing indicator nodes for extracting crime locations from a given text, by using them as triggering keywords with no syntactic constraint. The indicator nodes for extracting crime location are as follows:

- Name: منطقة / mntqt /area

  Trigger: منطقة / mntqt /area

  Target: place name

- Name: مدينة / mdynt / city

  Trigger: مدينة / mdynt / city

  Target: place name

- Name: محَافظة / mohafdat / province

  Trigger: محَافظة / mohafdat / province

  Target: place name

- Name: ولَاية / wlayt / state

  Trigger: ولَاية / wlayt / state

  Target: place name

The following Figure 3.25 depicts the local grammar for location.

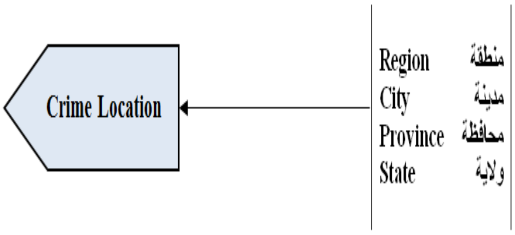


Figure 3.25: Location local grammar

3.7 Summary

This chapter has presented how the data were analysed, using frequency, collocation and concordance analysis approaches. These analyses have led to identifying the most important words (see Table 3.4, Table 3.5 and Table 3.6) within the huge collection of data, and have identified the behaviour of those words within text. As shown above, the results of the analyses have assisted in discovering the computational linguistic techniques (transitive and genitive constructions) for extracting crime type, location and nationality. As a result, this has led to the building of their local grammars, represented by indicator nodes for each case, which, to the author's knowledge, these special syntactic structures have not been used in the literature for extracting the aforementioned entities in Arabic text. These indicator nodes are triggered by predefined keywords and activated in a specific linguistic context. Accordingly, in order for the indicator nodes to be activated, the syntactic constraints must be achieved, except when extracting crime location because there is no need for any syntactic constraint. For example, in extracting a type of crime, the transitive construction is used, e.g. the indicator nodes "تورط / twrt / involved-transitive-في / fi / in"; this indicator node is triggered by the keyword "تورط / twrt / involved", and the enabling condition that allows the indicator node to be activated is that the word *tawarat*, which must be in a transitive construction. In other words, it must be followed by the preposition "في / fi / in" in order to form the following transitive structure:

في + تورط / twrt + fi

Involved + in

The indicator node, in the case of extracting the crime type, is considered the head node of the prepositional phrase. As a result, the complement of the indicator node that starts with a preposition is usually a crime type. The following is an example of this case.

- تورط في سرقة / twrt fi srqt-i
- Involved-transitive in-governor theft-genitive.

In this example, the indicator node is triggered by the transitive verb ”تورط / twrt / involved”, and is activated by the preposition ”في / fi / in”. The crime action word ”سرقة / srqt / theft”, which follows the preposition, is extracted as the crime type. Generally, all the indicator nodes for extracting the type of crime must be transitive verbs (by specific prepositions).

With regards to extracting a nationality from a given text, the genitive case is utilized. The indicator node here is triggered by the word ”جنسية / jensyt / nationality” or its other inflected forms, such as ”جنسيَات / jensyat / nationalities”. Also, it has a syntactic constraint. So, in order for the indicator node to be activated, the trigger word must be assigned the genitive case. Consider the following example:

- المجرم من جنسية سعودية / almjrm min jensyt saudit
- The offender from-governor nationality-genitive of Saudi-genitive.

The indicator node here is assigned the genitive case by the preposition (governor) ”من / min / from”. Moreover, the word ”سعودية / saudit / Saudi”, which follows the indicator node, is a nationality.

Finally, in order for a crime location to be extracted, the indicator nodes are triggered by words such as ”مدينة / mdynat / city of”, ”محَافظة / mohafdat / province of” and ”ولاَية / wlayt / state of”. There is no need for syntactic constraint to be applied because the target, which is the name of the place, often occurs directly after the triggering keywords. Consider the following example:

- منطقة الريَاض / mantqt alriyad

- Area of Riyadh

In the above example, the indicator node is triggered by the keyword "منطقة / mantqt / area". The word "الريَاض / alriyad / Riyadh", which follows the indicator node, is extracted as a place name. Table 3.59 provides a summarisation of the three cases, upon which is based the proposed framework, as will be seen in the following chapter.

Table 3.59: General description for the indicator node state for crime type, nationality and crime location

Entity - Event	**Indicator node state**
Crime type	Verbs in transitive construction by specific prepositions
Nationality	Specific noun in genitive case
Crime location	Specific nouns without any syntactic condition

Chapter 4

Architecture of Crime Profiling System

Objectives

- Present detailed information about the proposed architecture phases.
- Describe the components of each phase.
- Show how the components of the system's architecture interact.

4.1 Introduction

The syntactic analysis performed on data relating to the Arabic crime domain in Chapter 3 has revealed the requirements needed to design a framework for the Crime Profiling System (CPS). The proposed framework is presented in this chapter. Section 4.2 provides an overview of the framework's architecture. Section 3.4 describes the initial preprocessing stage, which is comprised of four components: data gathering, tokenization, normalisation and early filtering. Section 4.4 presents the information extraction stage, which is dedicated to extracting crime information from crime news reports in order to generate a summary and automatically construct dictionaries. Section 4.5 provides the intermediate preprocessing stage, which consists of post filtering, stemming, frequency analysis, generating a word index and document representation. Section 4.6 provides a brief description of the clustering stage. Finally, section 4.7 presents the visualisation stage.

4.2 Framework Overview

An overview of the framework's five stages, as depicted in Figure 4.1, is presented in this section. The five stages of our framework are as follows:

- Initial Preprocessing Stage

  This stage contains four components: data gathering, tokenization, normalisation and early filtering.

- Information Extraction Stage

  The information extraction process is conducted by utilising the proposed computational linguistic techniques that were presented in the previous chapter. Automatically building dictionaries is achieved in this stage as well as generating a summary.

- Intermediate Preprocessing Stage

  This stage consists of the post-filtering, stemming, indexing and document representation processes.

- Clustering Stage

  The clustering tasks, for grouping similar crimes together, are performed in this stage.

- Visualisation Stage

  In order to assist in analysing the collection of documents, a visualisation process is employed.

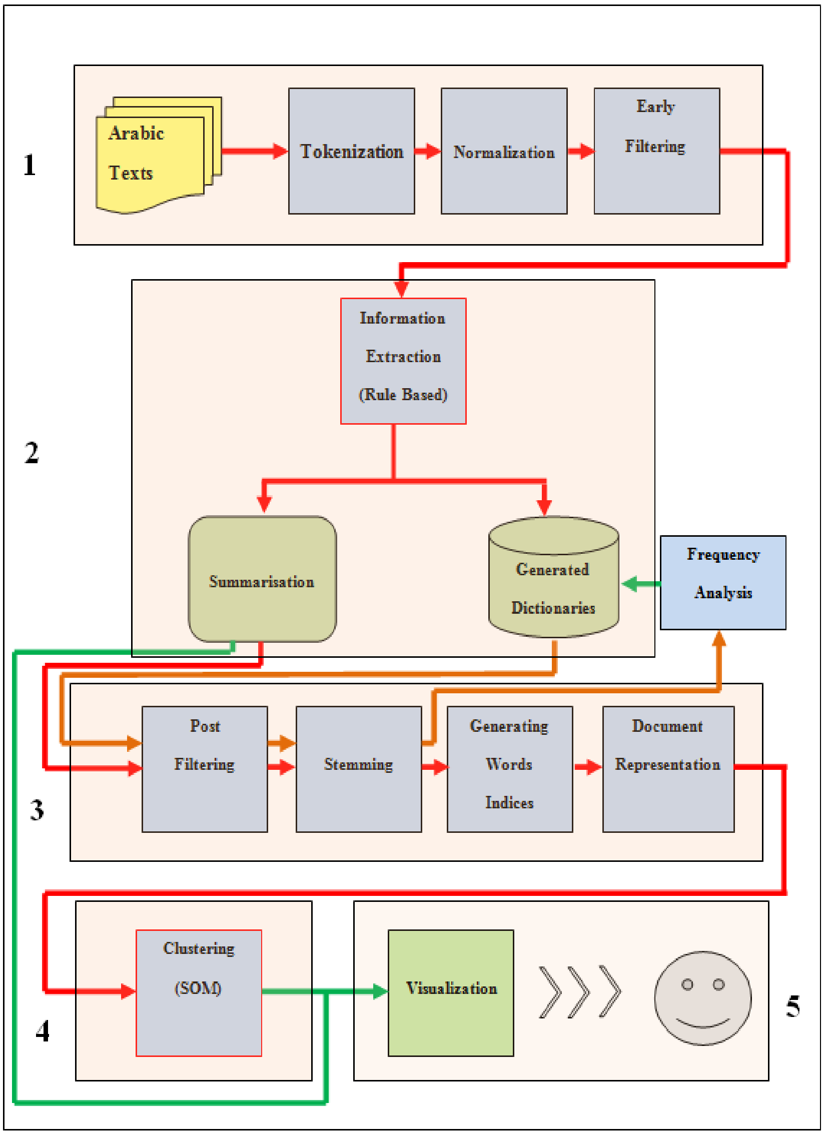


Figure 4.1: News-based Crime Profiling System (CPS) architecture

It can be seen that our proposed framework contains two text mining techniques: information extraction and clustering (together with other processes). Detailed explanations for each stage are provided in the following sections.

4.3 Initial Preprocessing Stage

This stage is comprised of four components: data gathering, tokenization, normalisation and early filtering. Each component is described as follows:

4.3.1 Data Gathering

It is well known that text mining research relies heavily on the availability of a suitable corpus. A corpus is a collection of documents used by text mining techniques for performing various tasks. It can be seen in Figure 4.1 that the first step is document gathering, conducted in order to build a corpus. In this research, the study is conducted on a specific context (the crime domain) but because there is no compilation of Arabic news reports about crime readily available in the literature, we need to collect our own specialised data, and so for this research project, news reports relating to crime, published in Arabic newspapers and available over the Internet, are our target for collection. As mentioned in Chapter 1, there are difficulties in accessing official police data or police narrative reports directly but newspaper reports usually contain the same information. Also, the content of web forums that are specialised and/or interested in crime news will be used in this work.
To collect crime news report (published in different Arabic countries) from the newspapers' websites, we will copy each crime incident from the crime section and save it in a plain file; the same step is done for web forums. As a result, these files are cleansed from HTML tags, and because we are compiling a specific data, this step is manually performed. The reason for compiling a corpus from different sources is

to avoid the problem of bias, which could occur if the system were developed and tested on documents collected from only one country. The result of this process can be seen in Chapter 5.

4.3.2 Tokenization

An important step in the processing of textual documents, which takes place before information extraction and clustering, is tokenization. Here the words in the documents are separated out into individual words that are identified by the blank spaces between them. Thus, tokenization allows the unstructured text to be split into tokens, which assists the system in processing the text. As a result, each textual file is represented through one vector. Figure 4.2 illustrates this process.

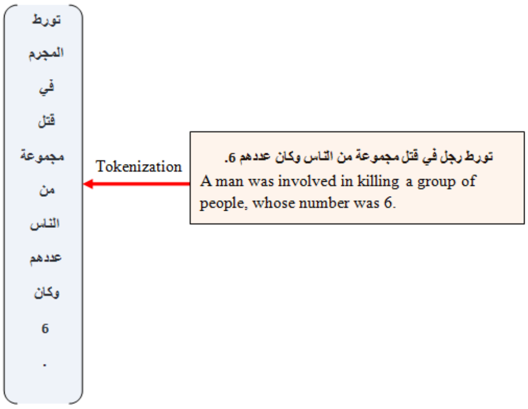

تورط
المجرم
في
قتل
مجموعة
من
الناس
عددهم
وكان
6
.
Tokenization
تورط رجل في قتل مجموعة من الناس وكان عددهم 6.
A man was involved in killing a group of people, whose number was 6.

Figure 4.2: Example of how the tokenization process works

4.3.3 Normalisation

Because there are spelling variations in the Arabic language, and some letters that mainly perform the same function are written in different forms, it is necessary to employ a normalisation strategy to avoid the data sparseness problem. This process plays important role in this work and cannot be ignored because of the different writing styles used by journalists, which may affect the proposed approach. For example, the verb ”اقدم / aqdm / proceed” in this research is used to identify a crime type when it is followed by the preposition ”علَى / ala / on”, which means that the verb is in a transitive construction. However, some newspapers do not adopt the official Arabic writing style, and instead write in classical Arabic. Therefore, the above verb ”اقدم / aqdm / proceed” is sometimes written as ”أقدم”, i.e. the letter ”أ” (with *hamza* above) is used instead of ”ا” (with no *hamza* above). Therefore, this process is devoted to performing three normalisation strategies, as follows:

- The first is related to the letter ”ا”; this may appear in the texts as ”أ” (with *hamza* above), ”إ” (with *hamza* below) or ”آ” (with *maad* above). These will all be normalised to ”ا”.

- The second letter is ”ة”, which may appear as ”ه”; this will be normalised to ”ة”.

- The third letter is ”ى”, which may be written with two dots under it (”ي”), and this will be normalised to ”ى”.

As a result, this process makes the corpus more consistent.

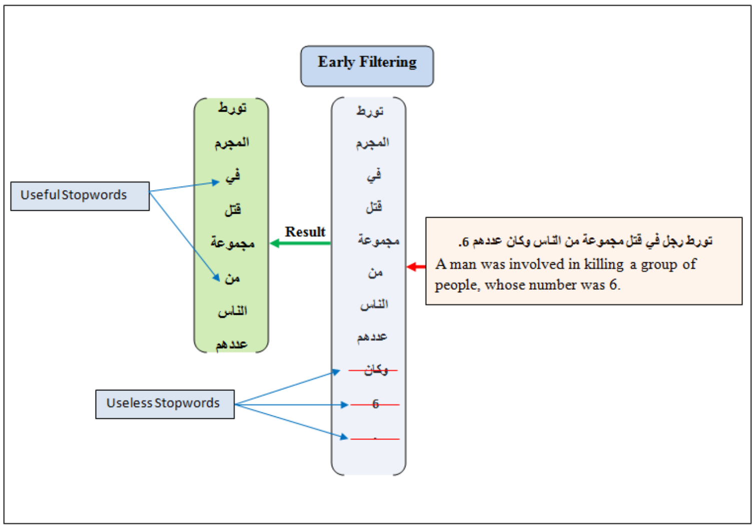
4.3.4 Early Filtering

In most text mining systems, the functional words, such as prepositions and conjunctions, as well as punctuation marks and diacritics, are usually removed during the filtering process. In this system, there are two filtering processes: early filtering (in this stage) and advanced filtering (in the third stage, the intermediate preprocessing stage). In the early filtering, the system will remove all the stopwords, punctuation marks, numbers and diacritics from each text document with the exception of prepositions, as shown in Table 4.1; these are retained for the process that follows. These remaining stopwords, i.e. the prepositions, are considered in this research to be central to the operation of the system in that they play a significant role in the syntactic constructions that are tapped to perform the next stage (information extraction stage).

Table 4.1: List of prepositions that are kept in the early filtering stage

Arabic Preposition	Pronunciation	English Translation
علَى	ala	on
في	fi	in
إلَى	ela	to
ب	bi	by
ل	li	to

Figure 4.3 is an illustration of the process whereby the unwanted stopwords are removed. The result of this component will be the same text document but with no stopwords (except prepositions), punctuation marks, numbers or diacritics.



.تورط رجل في قتل مجموعة من الناس وكان عددهم 6

A man was involved in killing a group of people, whose number was 6.

Figure 4.3: Early removal process, removing all types of stopwords except prepositions

4.4 Information Extraction Stage

Information extraction is the process whereby relevant information is extracted from a document. This is achieved in this research by utilising the proposed computational linguistic techniques, which are represented by the indicators nodes described in Chapter 3. This process is designed to identify three entities, which are the event type (e.g. murder, theft or assault), the location (e.g. Cairo), and the nationalities of the people involved in the crime, in order to automatically build dictionaries and generate a summary for each news report. These summaries are comprised of data that are of higher quality and better suited to the clustering process; according to Bacao et al. [163], the quality of the data is important, as it can have consequences for the quality of the clustering results.

During the extraction process, the proposed indicator nodes are used in the identification of the aforementioned entities, with some syntactic constraints. The specific processes for extracting event type, location and nationality are explained below.

1. Type of Crime

   As shown above, our study of the crime domain corpus has led us to identifying the characteristics of the language used (crime language). The types of offence usually occur within the transitive grammatical structure.

   For extracting crime types, the system looks for words in a text that match the words in the verb list, and when a match occurs, the system will look for the first possible associated preposition that follows that verb, i.e. transitive construction is achieved. As a result, the system is able to avoid using an annotated corpus by using prepositions to achieve the syntactic constraint. After that, the three words that immediately follow the preposition are extracted, and within these three words, the crime word should be present. Moreover, these extracted words are used to describe the document during the clustering stage that will follow. The reason that the three words after the preposition are extracted is to increase the probability of the crime word being identified. In Arabic, the crime action word sometimes does not appear immediately after the preposition as seen in the previous chapter, i.e. it is preceded by a noun. The following are representations of the possible position of a crime action word. In the first case, shown in Figure 4.4, the type of crime noun appears immediately after the preposition. Figure 4.5 illustrates where the type of crime noun does not appear immediately after the preposition because there is a word between, and Figure 4.6 illustrates the case where there are two words between the preposition and crime action word.

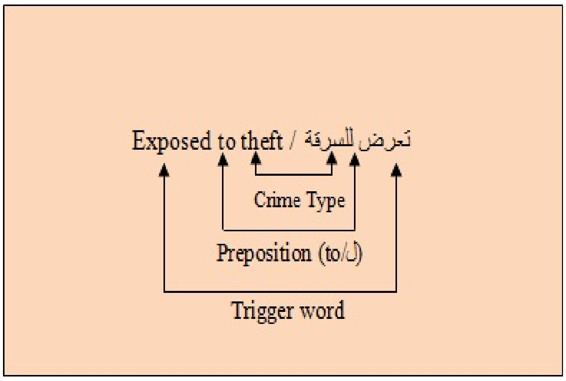


Figure 4.4: First case: crime type immediately follows the preposition

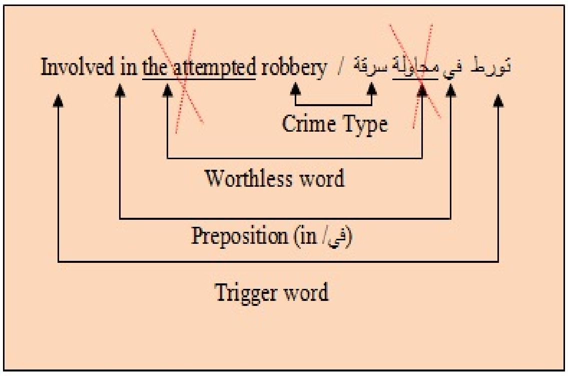


Figure 4.5: Second case: crime type after one word from the preposition

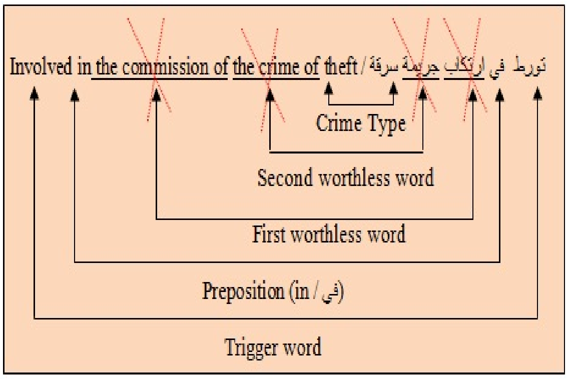


Figure 4.6: Third case: crime type after two words from the preposition

Another point to consider in this grammatical construct is that sometimes the preposition does not always immediately follow the verb. An example of this is illustrated in Figure 4.7, where the preposition, which is followed by the type of crime noun, appears nine words after the verb. If these nine words are removed, effectively abbreviating the sentence, the meaning can still be inferred from the head node (verb) and the preposition and noun (prepositional phrase), therefore, removing the nine words shown in the example does not detract from the key meaning of the sentence.

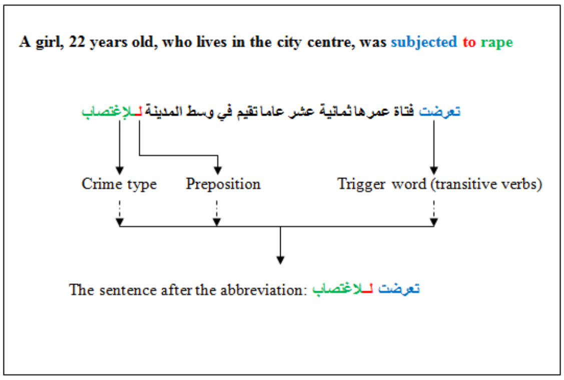
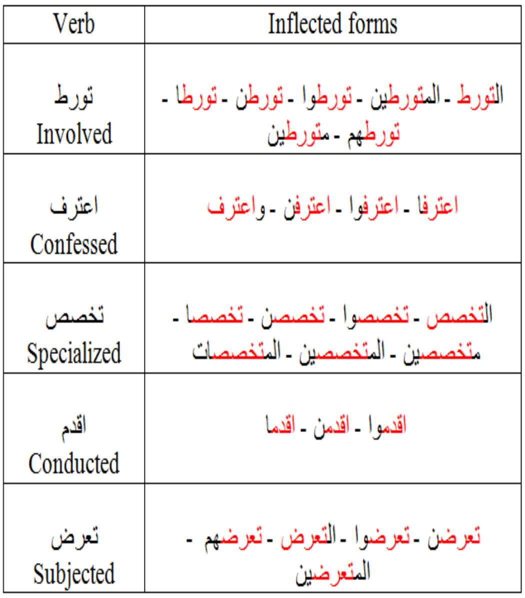

A girl, 22 years old, who lives in the city centre, was subjected to rape
تعرضت فتاة عمرها ثمانية عشر عاما تقيم في وسط المدينة لـلاغتصاب
Crime type
Preposition
Trigger word (transitive verbs)
The sentence after the abbreviation: تعرضت لـلاغتصاب

Figure 4.7: The associated preposition not directly following its verb

As well known, the Arabic language is an agglutinant language, and thus we must consider some agglutination cases. In Arabic, the preposition can either be separate from the noun, or fused together with the noun to form a single word, as previously illustrated in Chapters 2 and 3. Some verbs, such as "تعرض / taard / subjected" and "اعترف / aatrf / confessed" (in the transitive construction form) come with the prepositions " ل / li / to" and "ب / bi / by", which must be attached to the following noun. Accordingly, this stage is also designed to deal with the case of clitics.

Moreover, the Arabic language is rich in terms of morphology, whereby a word can be broken down into its base form and affixes, and usually it is the base, or root, of the word that is kept in dictionaries for extraction purposes. Examples in Arabic include verbs that have suffixes or prefixes attached to denote gender or plurality, and some of these affixes change the word into a noun. However, the proposed framework maintains a list of verbs in the past tense instead of the base or root form; this is because most news reports about crime are written in the past tense. The system uses N-gram to recognize many of the inflected forms of the verb by identifying the keyword, i.e. the past tense of the verb, from which the inflected forms are derived. These verbs can also produce nouns, such nouns in Arabic, derived from the verb, are often followed by the same preposition that the verb takes. Examples of the keywords (past tense of verbs) and their associated inflected forms are shown in Table. 4.2. The advantages of this approach are, firstly, not all of the words in the document need to be stemmed, as with other text mining systems, and secondly, it reduces the amount of required keywords in the list. Also, there is no need for an annotated corpus, in other words, the framework has no linguistic components, such as PoS taggers. Instead, lists of intransitive verbs and their prepositions are provided to the system in order to extract the desired patterns.

Table 4.2: Keywords and their inflected forms



Verb	Inflected forms
تورط Involved	التورط - المتورطين - تورطوا - تورطن - تورطا - تورطهم - متورطين
اعترف Confessed	اعترفا - اعترفوا - اعترفن - واعترف
تخصص Specialized	التخصص - تخصصوا - تخصصن - تخصصا - متخصصين - المتخصصين - المتخصصات
اقدم Conducted	اقدموا - اقدمن - اقدما
تعرض Subjected	تعرضن - تعرضوا - التعرض - تعرضهم - المتعرضين

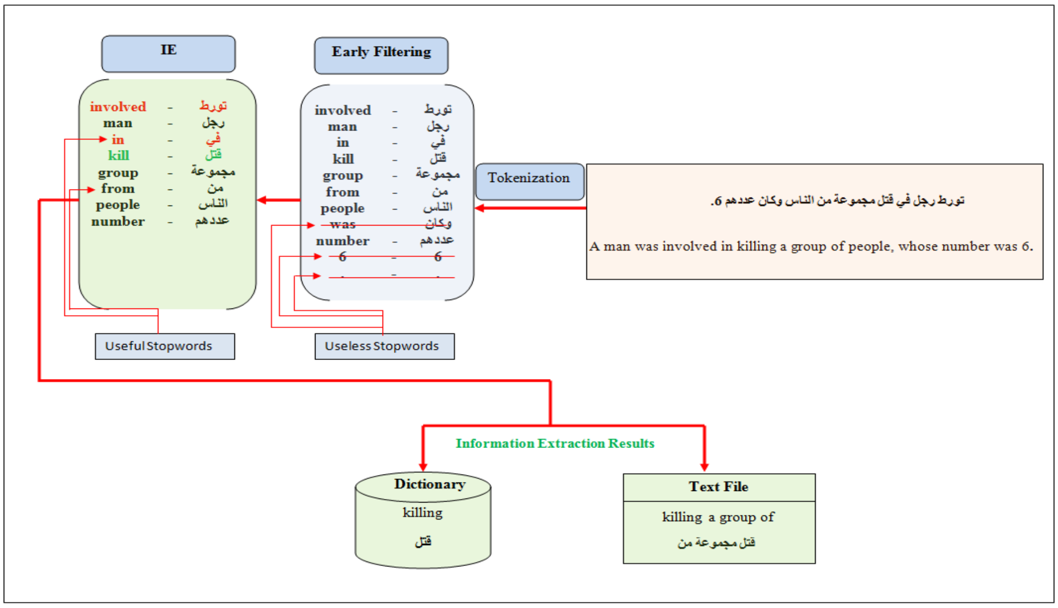
2. Nationality

As mentioned earlier in Chapter 3, the word ”جنسية / jnsyt / nationality” in the form of singular or plural (”جنسيّات / jnsyat / nationalities”) is usually used to illustrate a person’s nationality in crime news reports, and its linguistic context is usually the genitive construction. As a result, for extracting nationality patterns from Arabic crime texts, the indicator nodes that were constructed based on the nationality local grammar in the previous chapter are utilized to perform this task. The system looks for words in a text that match the words in the nationality keyword list, and when a match occurs, the system checks the syntactic construction of the word to determine whether or not it is in a genitive construction. Therefore, the system will check whether the word is preceded by the preposition ”من / min / from” in order to achieve the syntactic constraint. If the condition is achieved, then the word that immediately follows the keyword is extracted as a nationality. However, there is one exception; the word ”الجنسية / aljensyah / the nationality” in singular form with the article ”ال / al / the” attached. When there is no preposition before it, the word that occurs instead of the preposition is identified as the nationality.

3. Location

   In order for the system to extract the place names, the indicator nodes proposed in Chapter 3 are used without syntactic constraint. Therefore, once matching occurs between any trigger word and a word in the contents of the file, the word that follows the keyword is extracted and classified as a location name. Using this linguistic technique serves to overcome the lack of any capital letter feature; this is not available in the Arabic language, so it cannot be used as a clue for extracting proper names, as in the English language.

The following Figure 4.8 presents the whole process of the initial preprocessing stage with the information extraction stage. It shows the transformation phases for the text as it proceeds through each step.



تورط رجل في قتل مجموعة من الناس وكان عددهم 6.

A man was involved in killing a group of people, whose number was 6.

Figure 4.8: Initial preprocessing stage and information extraction stage

Figure 4.9 shows example of how the Crime Profiling System (CPS) extracts the crime-related information based on the techniques proposed in Chapter 3.

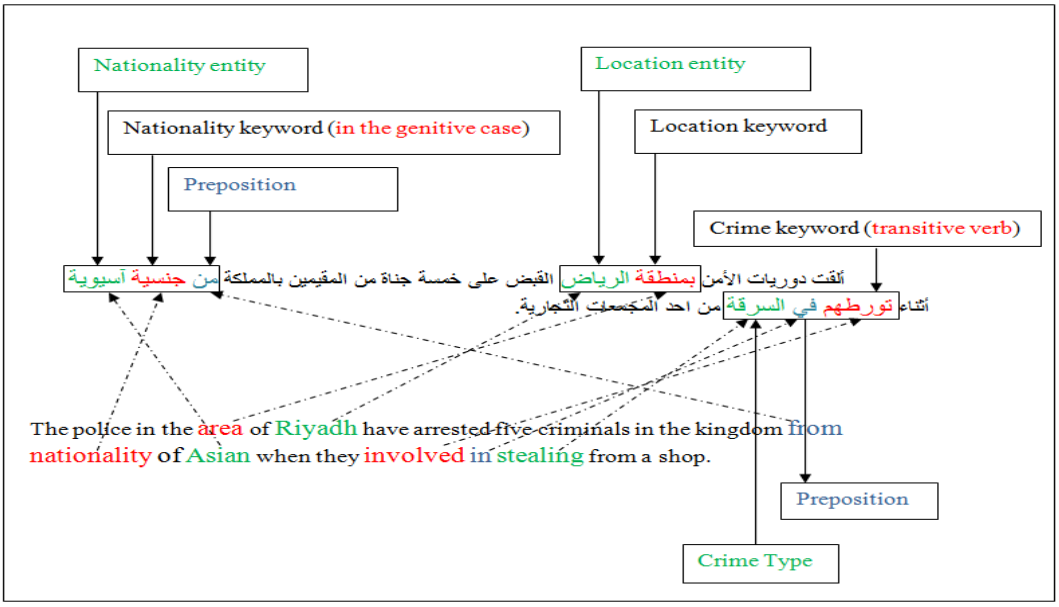

Nationality entity
Location entity
Nationality keyword (in the genitive case)
Location keyword
Preposition
Crime keyword (transitive verb)
ألقت دوريات الأمن بمنطقة الرياض القبض على خمسة جناة من المقيمين بالمملكة من جنسية آسيوية
أثناء تورطهم في السرقة من احد المجمعات التجارية.
The police in the area of Riyadh have arrested five criminals in the kingdom from
nationality of Asian when they involved in stealing from a shop.
Preposition
Crime Type

Figure 4.9: Initial preprocessing stage and information extraction stage

The above example contains three cases as follows:

- Crime Type

  In above example the indicator nodes "تورط / twrt / involved-transitive-في / fi / in" described in Chapter 3 is used to extract the crime type; this indicator node is triggered by the keyword (transitive verb) "تورط / twrt / involved" because a match had occurred when the system looked for words in a text that match the words in the transitive verb list. The indicator node is considered the head node of the prepositional phrase. As a result, the complement of the indicator node that starts with a preposition is usually a crime type. As mentioned in Chapter 3, the prepositions "في / fi / in" or "ب / bi / in" are the associated prepositions of the verb "تورط / twrt / involved". Therefore, the system will look for the associated preposition that follows that verb in order to achieve the syntactic constraints (transitive construction). After that, the word that immediately follow the preposition are extracted as a crime type. In the above example, the crime action word ” سرقة / srqt / theft”, which follows the preposition ”في / fi / in”, is extracted as the crime type.

- Nationality

  With regards to extracting a nationality from the given example, the indicator node ” جنسية / jensyt / nationality-genitive-من / min / from” was utilized. The indicator node here is triggered by the word ” جنسية / jensyt / nationality” because the match occurred. In order for the indicator node to be activated to extract a nationality entity, the trigger word must be assigned the genitive case. In this example, this syntactic constrain was achieved because the keyword ”جنسية / jensyt / nationality” is assigned the genitive case by the preposition (governor) ”من / min / from” that preceded it. As a result, the word ” سعودية / Asyawy / Asian”, which immediately follows the keyword ” جنسية / jensyt / nationality” is a nationality.

- Location

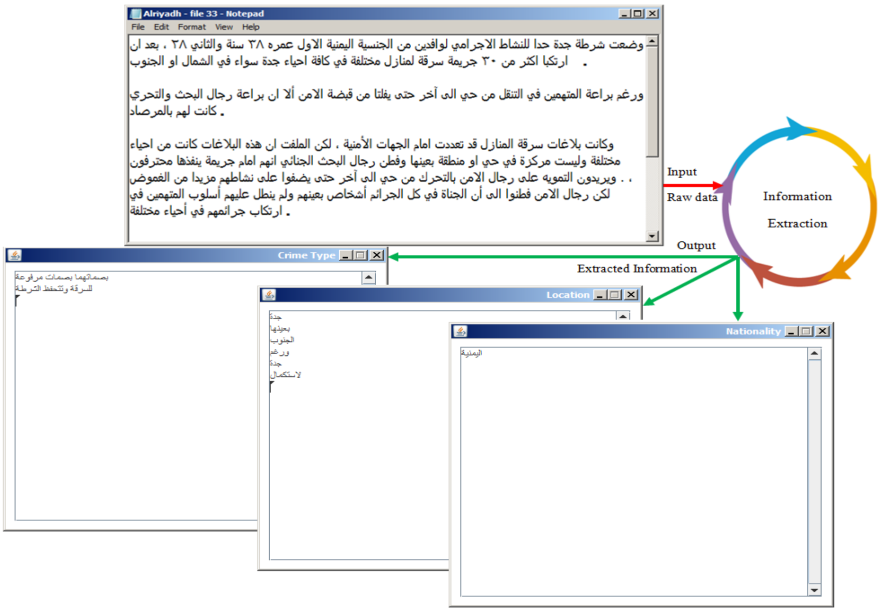
  Finally, in order for a crime location to be extracted in the above example, the indicator node ”منطقة / mntqt / area” mentioned in Chapter 3 was used without syntactic constraint. After the matching occurred between one of the location keywords and a word in the example, the word ”الريَاض / Alriyadh / Riyadh” that follows the keyword ”منطقة / mntqt / area is extracted and classified as a location name.

4.4.1 Summarisation

A summary is a condensed copy of the original document containing, only the essential information. The idea of summarisation is to reduce the length of the document, retaining only key information and the overall meaning. The information extraction method for this work has been already explained.

For the crime profiling system, upon completion of the information extraction, a summary is produced about each crime news report, which contains information relating to crime type, location and nationality. Figure 4.10 illustrates this generated summary, whereby individual windows contain the different types of information about the crime (crime type, location and nationality).

It is from these summaries that the clustering will be produced, according to the three types of information, i.e. type of crime, location and nationality. These words are considered to have meaningful information about the document required for the clustering process.



وضعت شرطة جدة حدا للنشاط الاجرامي لوافدين من الجنسية اليمنية الاول عمره ٣٨ سنة والثاني ٣٨ ، بعد ان
ارتكبا اكثر من ٣٠ جريمة سرقة لمنازل مختلفة في كافة احياء جدة سواء في الشمال او الجنوب .

ورغم براعة المتهمين في التنقل من حي الى آخر حتى يفلتا من قبضة الامن ألا ان براعة رجال البحث والتحري
كانت لهم بالمرصاد .

وكانت بلاغات سرقة المنازل قد تعددت امام الجهات الأمنية ، لكن الملفت ان هذه البلاغات كانت من احياء
مختلفة وليست مركزة في حي او منطقة بعينها وفطن رجال البحث الجنائي انهم امام جريمة ينفذها محترفون
. . ويريدون التمويه على رجال الامن بالتحرك من حي الى آخر حتى يضفوا على نشاطهم مزيدا من الغموض
لكن رجال الامن فطنوا الى أن الجناة في كل الجرائم أشخاص بعينهم ولم ينطل عليهم أسلوب المتهمين في
ارتكاب جرائمهم في أحياء مختلفة .

بصماتهما بصمات مرفوعة
للسرقة وتتحفظ الشرطة

جدة
بعينها
الجنوب
ورغم
جدة
لاستكمال

اليمنية

Figure 4.10: Result of the information extraction (IE) process in the system

4.4.2 Generating Dictionaries

It is an important aspect of this research that the dictionary generated for the system is done so automatically, i.e. that there is no manual building of the dictionary. It is necessary, first of all, to explain the function of the dictionary in the crime profiling system followed by an explanation of how the dictionary is generated automatically. As explained in the previous section, the summaries are created to be used in the clustering process that follows. When these summaries are created, in some cases the summary box may return an empty result, for example, the system may not identify a nationality from the article using the indicator nodes, however, a word denoting nationality may exist in the text but not be identified as its context does not fall under the system's rule. In order to overcome this failing, the system will check the document for a word that, for example, denotes nationality by matching with the generated dictionary. There are three dictionaries, one for each information type, i.e. crime type, location and nationality.

As mentioned above, the three dictionaries are generated automatically, which is a distinctive feature of our system. For the location and nationality categories, the words that are in the summary are automatically sent to their respective dictionaries; these words include both relevant and irrelevant words. The way that the dictionary determines whether or not specific words should be used in the matching, i.e. that they are relevant words, is by checking the frequency of each word; it is expected that relevant words, e.g. Jeddah, will occur more frequently, and when a word reaches a certain frequency threshold, it will be used in matching. In order to generate the dictionary for type of crime, the system only extracts from the summary the word that immediately follows the preposition.

4.5 Intermediate Preprocessing Stage

The intermediate preprocessing stage is comprised of five components: post-filtering, stemming, frequency analysis, generating word index, and document representation. The main goal of this stage is to prepare the extracted data from the summarising stage in an adequate form in order to be processed in the clustering stage.

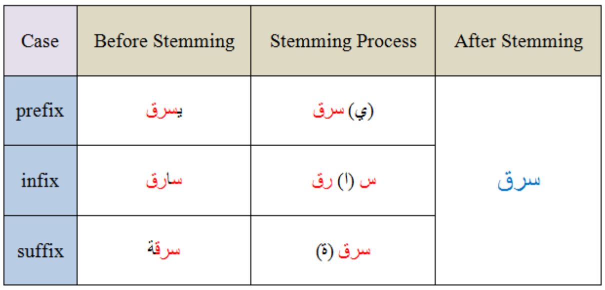
4.5.1 Post Filtering

The post-filtering process of the intermediate preprocessing stage is designed to remove the prepositions (see Table 4.1) that were retained for the information extraction stage but were not removed in the early filtering process. Therefore, the size of the data is reduced.

4.5.2 Stemming

Once all crime reports are summarised and the dictionaries are automatically constructed, the summarised files and dictionaries are ready to be stemmed. As already explained, in order to obtain the root of a word, all suffixes, prefixes and/or infixes are removed. Table 4.3 shows three cases with the word ”سرق / srq / steal” in Arabic, and how the system deals with them.

Table 4.3: Stemming process for removing prefixes, infixes and suffixes for the word ’سرق / srq / steal”



Case	Before Stemming	Stemming Process	After Stemming
prefix	يسرق	(ي) سرق	سرق
infix	سارق	س (ا) رق	
suffix	سرقة	سرق (ة)	

Thus, this process removes all the affixes from the words, reducing them to their stems. Stemming in the proposed system is required because it makes it easier for the system to allocate numbers for generating the words' indices, which are used in the clustering process. Furthermore, it assists in applying the frequency analysis process in an efficacious manner.

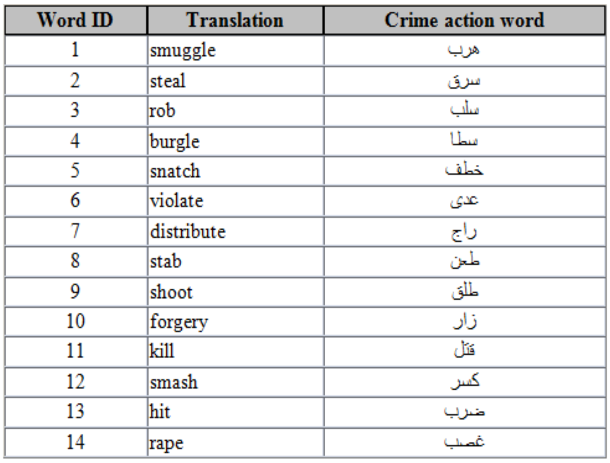
4.5.3 Frequency Analysis

Once the crime type, location and nationality dictionaries are built, and have passed through the post-filtering and stemming phases, the frequency analysis is applied. This process is necessary in order to cleanse the dictionaries of useless words by calculating the number of times the extracted words appear. Thus, only words with high frequencies will be chosen to be included in a dictionary, all others are discarded.

4.5.4 Generating Word Index

The clustering process only has the ability to process numerical data, not text, and therefore it is necessary to allocate to each word a specific number. Once the summarised files have been stemmed, each file is assigned a set of numbers, whereby each number corresponds to just one word within the file. Table 4.4 presents a sample of the crime action words with their unique numbers.

Table 4.4: Each word of interest is assigned a unique number in order to convert textual data into numerical form in order to facilitate the clustering process



Word ID	Translation	Crime action word
1	smuggle	هرب
2	steal	سرق
3	rob	سلب
4	burgle	سطا
5	snatch	خطف
6	violate	عدى
7	distribute	راج
8	stab	طعن
9	shoot	طلق
10	forgery	زار
11	kill	قتل
12	smash	كسر
13	hit	ضرب
14	rape	غصب

4.5.5 Document Representation

Once the content of each document has been encoded (in the previous process), these numerical data are transformed into vectors. The Vector Space Model (VSM) that was discussed in Chapter 2 is used to implement this task. Following this, the newly created file, which only contains numerical data, is sent to the clustering phase.

4.6 Clustering Stage

Once the crime news reports have been represented in VSM, the Self Organising Map (SOM) is used to cluster and visualise the crimes, based on crime type, location or nationality. As previously mentioned, these reports only contain text extracted in information extraction stage and then pruned in the intermediate prepocessing stage, until they were encoded and presented by VSM in a single file. The SOM algorithm (and how it works) was reviewed in Chapter 2.

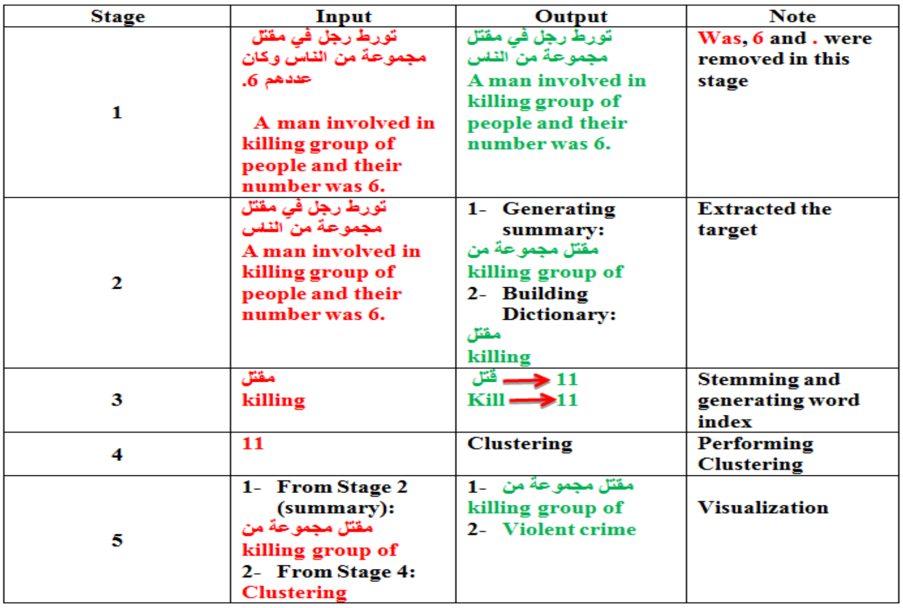
4.7 visualization Stage

To assist in analysing the collection of documents and to better understand the data, visualization techniques are required. visualization is considered to be useful in drawing conclusions, as it assists in revealing the bigger picture, which could not be achieved by sifting through so many documents. This framework is designed to provide the user with the ability to visualize the summary results of each news report separately, as in Figure 4.10. As a result, important crime-related information hidden within bodies of textual crime reports is easily revealed, and therefore, the visualization assists in accelerating the crime investigation process.

Furthermore, the output of the clustering stage is visualized through SOM, i.e. large amounts of textual data can be visualized on the map following the clustering process. Also, statistical information about crimes committed within certain periods of time (in the form of tables or graphs) is produced.

The Following Table 4.5 shows how a crime news report is processed by the above architecture.

Table 4.5: Role of each Stage in the Architecture to Process a Crime News Report



Stage	Input	Output	Note
1	تورط رجل في مقتل مجموعة من الناس وكان عددهم 6. A man involved in killing group of people and their number was 6.	تورط رجل في مقتل مجموعة من الناس A man involved in killing group of people and their number was 6.	Was, 6 and . were removed in this stage
2	تورط رجل في مقتل مجموعة من الناس A man involved in killing group of people and their number was 6.	1- Generating summary: مقتل مجموعة من killing group of 2- Building Dictionary: مقتل killing	Extracted the target
3	مقتل killing	قتل → 11 Kill → 11	Stemming and generating word index
4	11	Clustering	Performing Clustering
5	1- From Stage 2 (summary): مقتل مجموعة من killing group of 2- From Stage 4: Clustering	1- مقتل مجموعة من killing group of 2- Violent crime	Visualization

4.8 Summary

This chapter has presented an overview of the architecture of the proposed system. The five stages: initial preprocessing, information extraction, intermediate preprocessing, clustering and visualization, which constitute this system, were all described. As seen earlier, each stage is comprised of components to perform specific tasks. The role of each component in each stage was explained, and how these components interact was described. The implementation of these components of proposed system is detailed in the following chapter.

Chapter 5

EXPERIMENTS

Objectives

- Explore the Crime Profiling System (CPS).
- Implement the information extraction process to generate summarisation and to construct dictionaries.
- Implement the clustering task based on the results of the information extraction process.

5.1 Introduction

The experiments that were carried out to test the proposed system are presented in this chapter. Some experiments are concerned with applying the proposed local grammars that were explained in Chapter 3, in order to generate a summary for each document, i.e. to extract the crime-related information (crime type, crime location and nationality) and to automatically build dictionaries for them. The other experiments involve clustering the crime reports using the Self Organising Map (SOM) technique, based on the three different attributes (crime location, nationality and crime type). The clustering experiments were conducted with and without the aid of the information extraction stage in order to highlight the effects of this stage on the SOM.

In this chapter, the new and untouched corpus for implementing and testing the Crime Profiling System (CPS) is presented in section 5.2. The system implementation is described in section 5.3. The information extraction running procedure is explained in section 5.4. The experiments' results after implementing the CPS are provided in section 5.5, which contains the experiments for extracting the aforementioned information (in order to generate summaries and automatically build dictionaries) as well as two experiments (comparative experimentation) for clustering crime news reports with/without using the CPS information extraction approach. Moreover, a crime analysis (through presenting statistical information on the highest and/or lowest type of crime) and a visualisation of the results is also provided. Finally, the summary for this chapter is presented in section 5.6.

5.2 Arabic Crime News Report Corpus

The structure of the Arabic Crime News Report Corpus (ACNRC) is depicted in Figure 5.1. The main folder consists of sub-folders representing the different sources of the collected data, i.e. the countries where the reports were published. These sub-folders also contain other folders that represent the newspapers that published the reports, and inside these are the collected news reports. These reports are comprised of textual data, describing various crimes. The name of each report is composed of five parts, which are as follows:

- The first part is a letter referring to the language. For instance, for Arabic it is A and for English it is E.
- The second part is a letter referring to the name of the country. For example, Saudi Arabia is S, Kuwait is K, and so on.
- The third part is a letter taken from the name of the newspapers, such as Albayan is B.
- The fourth part gives the serial number of the document inside the folder.
- The last part refers to the number of tokens in a file.

The total number of tokens is 587,212. A sample of the corpus is in Appendix A.

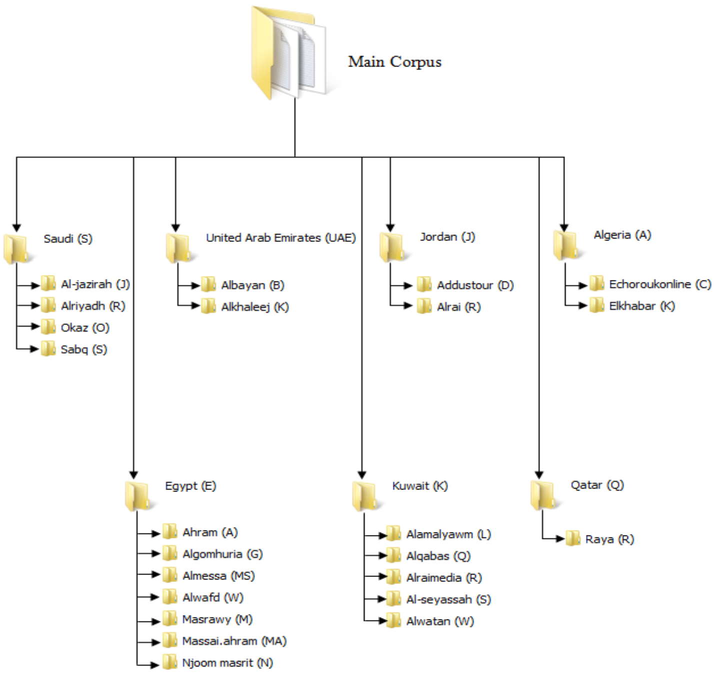


Figure 5.1: Structure of the Arabic Crime News Report Corpus (ACNRC)

The data used in the experiments is comprised of 80,943 tokens, published by Alriyadh [142], Sabq [143], Okaz [144] and Al-jazirah [145] from Saudi Arabia, Ahram [146] and Massai Ahram [164] from Egypt, Addustour [165] from Jordan, Alraimedia [149], Alqabas [150], Alamalyawm [151] and Alwatan [152] from Kuwait, Elkhabar [166] and Echoroukonline [154] from Algeria, and Albayan [155] and Alkhaleej [167] from the United Arab Emirates (UAE). These reports were saved in plain files with UTF-8 encoding. It is important to note that this data is new and was not used during the system's development phase.

5.3 System Implementation

The first two stages of the Crime Profiling System (CPS), i.e. the initial preprocessing stage (tokenization, normalisation and early filtering) and the information extraction stage are implemented using Java. Moreover, the Multilingual Morphological Analysis (MMA) software, developed by Al-Marghilani [168], is used to implement the intermediate preprocessing stage, which includes post-filtering, stemming, generating a word index and document representation. For implementing the clustering stage and for visualising the outcome of the clustering technique, a software package designed for clustering multilingual (Arabic and English) documents using the Self Organising Map (SOM) is adapted to perform these tasks [168]. The Matlab software package is used for implementing the SOM algorithm [169]. For generating tables and graphs (in order to achieve the strategic and administrative crime analysis), the Excel program is used. Moreover, we make use of a 'spatial analyst' software package called ArcGIS Explorer developed by the Environmental Systems Research Institute (Esri) for performing spatial analyses by displaying the crime locations on the map [170].

5.4 Information Extraction Running Procedure

In order for the user to extract crime-related information, the CPS is designed with a graphical user interface, shown in Figure 5.2.

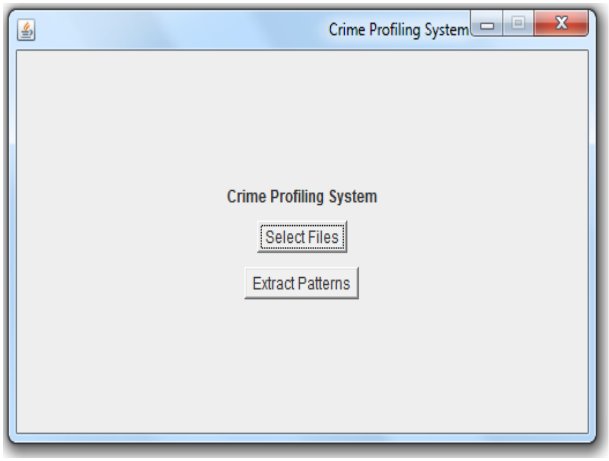
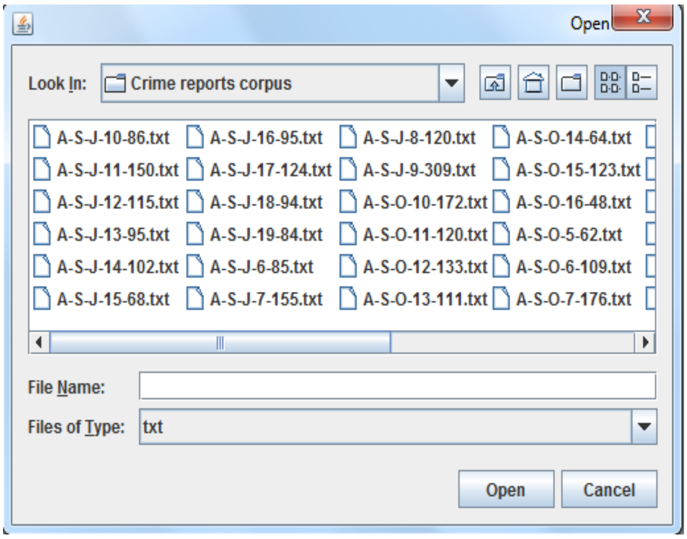


Figure 5.2: The main window of the Crime Profiling System (CPS)

The main window contains two buttons: to select a file from the main folder as in

Figure 5.3, and to extract patterns to summarise and construct dictionaries.



Look In: Crime reports corpus

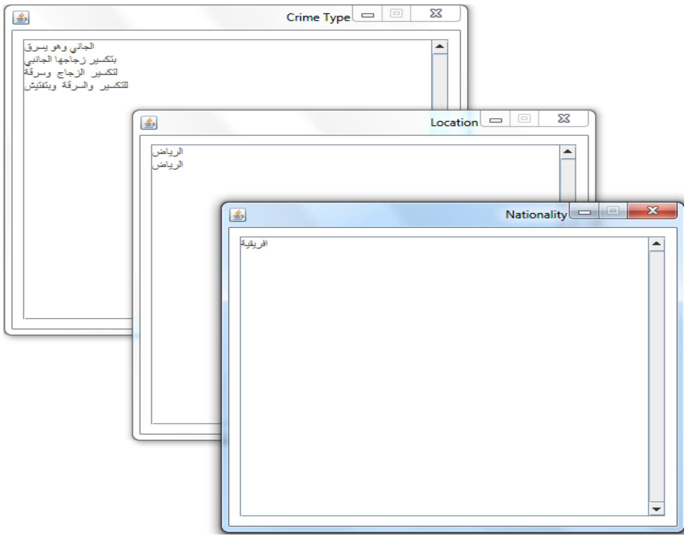
A-S-J-10-86.txt
A-S-J-11-150.txt
A-S-J-12-115.txt
A-S-J-13-95.txt
A-S-J-14-102.txt
A-S-J-15-68.txt
A-S-J-16-95.txt
A-S-J-17-124.txt
A-S-J-18-94.txt
A-S-J-19-84.txt
A-S-J-6-85.txt
A-S-J-7-155.txt
A-S-J-8-120.txt
A-S-J-9-309.txt
A-S-O-10-172.txt
A-S-O-11-120.txt
A-S-O-12-133.txt
A-S-O-13-111.txt
A-S-O-14-64.txt
A-S-O-15-123.txt
A-S-O-16-48.txt
A-S-O-5-62.txt
A-S-O-6-109.txt
A-S-O-7-176.txt

File Name:

Files of Type: txt

Figure 5.3: CPS File Open window

Figure 5.4 shows how the system generates three windows, each of which contains specific information about crime (type of crime, crime location and nationality).



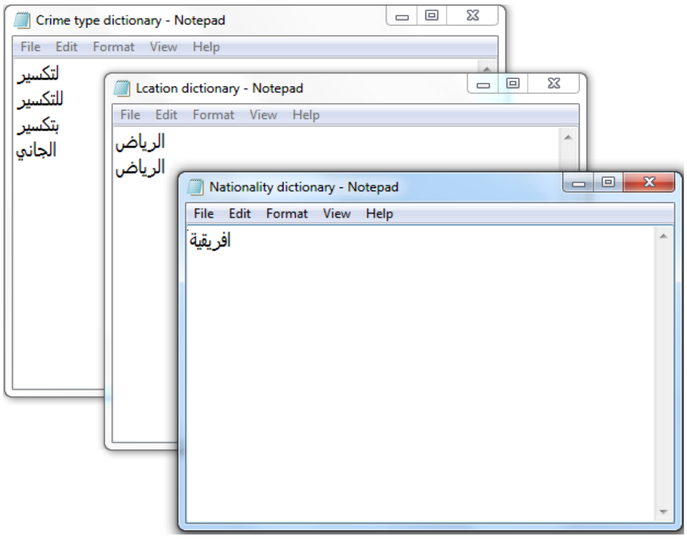
الجاني وهو يسرق
بتكسير زجاجها الجاني
لتكسير الزجاج وسرقة
للتكسير والسرقة وبتفتيش

الرياض
الرياض

افريقية

Figure 5.4: Summarisation results (crime type, location and nationality windows)

Also, the extracted information is directly sent and saved in one of the three plain files, for the dictionary construction task (see Figure 5.5).


Crime type dictionary - Notepad
File Edit Format View Help
Lcation dictionary - Notepad
File Edit Format View Help
Nationality dictionary - Notepad
File Edit Format View Help

لتكسير
للتكسير
بتكسير
الجاني

الرياض
الرياض

أفريقية

Figure 5.5: Three files representing the crime type, location and nationality dictionaries

The following section presents the experiments that were performed to implement the CPS.

5.5 Experiments

In the following experiments, 401 crime reports collected from various online news sources and comprising of 57,595 tokens are used to test the performance of the Crime Profiling System (CPS) for extracting the aforementioned entities in order to generate summaries for each report and to automatically build dictionaries. It was found that, after conducting the information extraction experiments there were 79 reports have some entities were not extracted directly by the local grammars, and therefore, these reports were removed and new other 79 reports were added instead in order to use them in the clustering experiments. As a result, the size of the above corpus changed to 71,882 tokens. The clustering experiments were performed in order to show how the proposed information extraction approach guides the Self Organising Map (SOM) to gain improvement in clustering quality. Furthermore, based on extracting crime-related information, graphs and tables are generated for providing statistical information about the crime status in the locations.

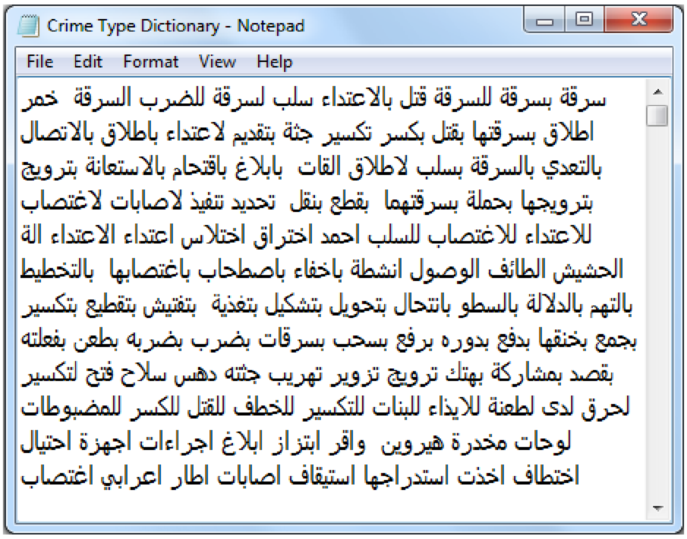
5.5.1 Crime Type

In this experiment, the CPS crime type local grammar was assessed for its ability to recognize and extract the crime type from each report, and to generate a summary for each report, as already shown in Figure 5.4. Table 5.1 shows the results of this experiment, which includes a number of ’true’ and ’false’ extracted entities from each dataset, together with their targets.

Table 5.1: Crime type extraction results using crime type local grammar

Dataset	True	False	Goal
Riyadh	37	47	40
Sabq	70	76	76
Okaz	96	84	140
Ahram	49	47	64
Alwatan	67	52	80
Alamalyawm	62	36	72
Gokarsat	17	18	24
Total	398	360	496

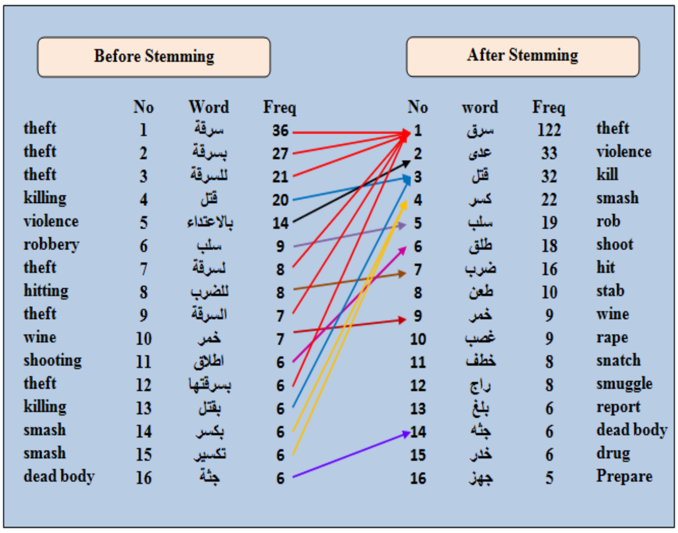
The extracted patterns are sent to crime type dictionary file. The system was able to automatically build a crime type dictionary with 794 tokens. Figure 5.6 shows a sample of the result.



سرقة بسرقة للسرقة قتل بالاعتداء سلب لسرقة للضرب السرقة خمر
اطلاق بسرقتها بقتل بكسر تكسير جثة بتقديم لاعتداء باطلاق بالاتصال
بالتعدي بالسرقة بسلب لاطلاق القات بابلاغ باقتحام بالاستعانة بترويج
بترويجها بحملة بسرقتهما بقطع بنقل تحديد تنفيذ لاصابات لاغتصاب
للاعتداء للاغتصاب للسلب احمد اختراق اختلاس اعتداء الاعتداء الة
الحشيش الطائف الوصول انشطة باخفاء باصطحاب باغتصابها بالتخطيط
بالتهم بالدلالة بالسطو بانتحال بتحويل بتشكيل بتغذية بتفتيش بتقطيع بتكسير
بجمع بخنقها بدفع بدوره برفع بسحب بسرقات بضرب بضربه بطعن بفعلته
بقصد بمشاركة بهتك ترويج تزوير تهريب جثته دهس سلاح فتح لتكسير
لحرق لدى لطعنة للايذاء للبنات للتكسير للخطف للقتل للكسر للمضبوطات
لوحات مخدرة هيروين واقر ابتزاز ابلاغ اجراءات اجهزة احتيال
اختطاف اخذت استدراجها استيقاف اصابات اطار اعرابي اغتصاب

Figure 5.6: Sample of the crime type dictionary

The dictionary was filtered to remove certain stopwords, and its size was accordingly reduced to 676 tokens. The stopwords here include the most frequent words that appear in Chapter 2 Table 3.3, not crime action words. Furthermore, a frequency analysis was carried out after the filtering process, and this led to reducing the number of tokens to 380. In order to obtain the words only in their base form, their affixes were removed, and consequently, the size of the dictionary became 228 words. Figure 5.7 shows a comparison of the content of the dictionary before and after the stemming process. It can be seen that the frequencies of the various forms of the word ”سرق / srqt / theft” were collated into one frequency, occurring 122 times. Also, it can be noticed that the most frequent words either before or after the stemming are crime action words. Moreover, four crime action words (”غصب / gsb / rape”, ”خطف / ktf / snatch”, ”رَاج / raj / smuggle” and ”خدر / kdr / drug”) rose to within the list of the 16 most frequent crime words following the stemming process.



Before Stemming

	No	Word	Freq
theft	1	سرقة	36
theft	2	بسرقة	27
theft	3	للسرقة	21
killing	4	قتل	20
violence	5	بالاعتداء	14
robbery	6	سلب	9
theft	7	لسرقة	8
hitting	8	للضرب	8
theft	9	السرقة	7
wine	10	خمر	7
shooting	11	اطلاق	6
theft	12	بسرقتها	6
killing	13	بقتل	6
smash	14	بكسر	6
smash	15	تكسير	6
dead body	16	جثة	6

After Stemming

No	word	Freq	
1	سرق	122	theft
2	عدى	33	violence
3	قتل	32	kill
4	كسر	22	smash
5	سلب	19	rob
6	طلق	18	shoot
7	ضرب	16	hit
8	طعن	10	stab
9	خمر	9	wine
10	غصب	9	rape
11	خطف	8	snatch
12	راج	8	smuggle
13	بلغ	6	report
14	جثه	6	dead body
15	خدر	6	drug
16	جهز	5	Prepare

Figure 5.7: Sample of the crime type dictionary before and after the stemming process

The crime type dictionary was tested to see if it could identify any crime types that the crime type local grammar had failed to extract. Table 5.2 presents the results of this experiment.

Table 5.2: Crime type extraction results after utilising crime type dictionary

Dataset	True	False	Goal
Riyadh	40	47	40
Sabq	74	82	76
Okaz	133	105	140
Ahram	64	60	64
Alwatan	80	59	80
Alamalyawm	68	41	72
Gokarsat	22	22	24
Total	481	416	496

Accordingly, using the crime type dictionary assisted in recognising more crime types, and therefore, the number of entities correctly identified increased to 481 (from 398) entities. Consequently, only 15 types of crime were not extracted. However, the types of crime that were wrongly extracted also increased to 416 (from 360).

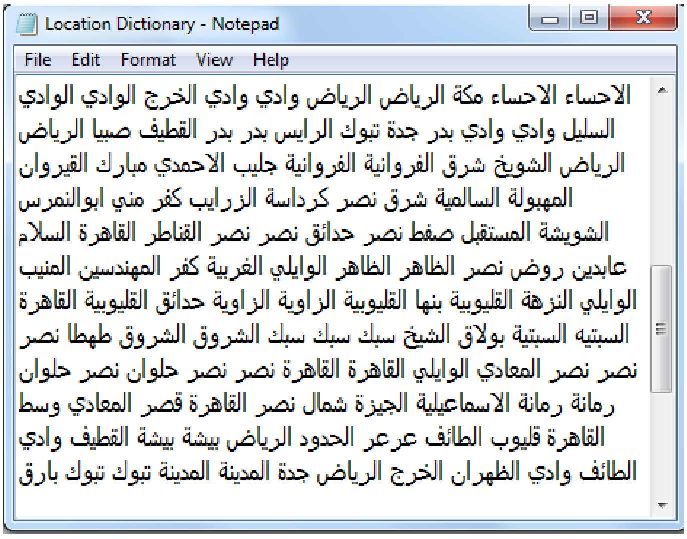
5.5.2 Location

The experiment here is dedicated to testing the CPS location local grammar in order to assess its ability to extract crime location from the given reports (to generate summaries) and to automatically build the location dictionary. Table 5.3 shows the number of extracted patterns, whether true or false, for each dataset, and the third column refers to the number of patterns that should be recognized in each dataset.

Table 5.3: Location extraction results using location local grammar

Dataset	True	False	Goal
Riyadh	34	0	41
Sabq	50	0	68
Okaz	97	3	145
Ahram	53	5	74
Alwatan	59	0	67
Alamalyawm	50	0	53
Gokarsat	25	0	27
Total	368	8	475

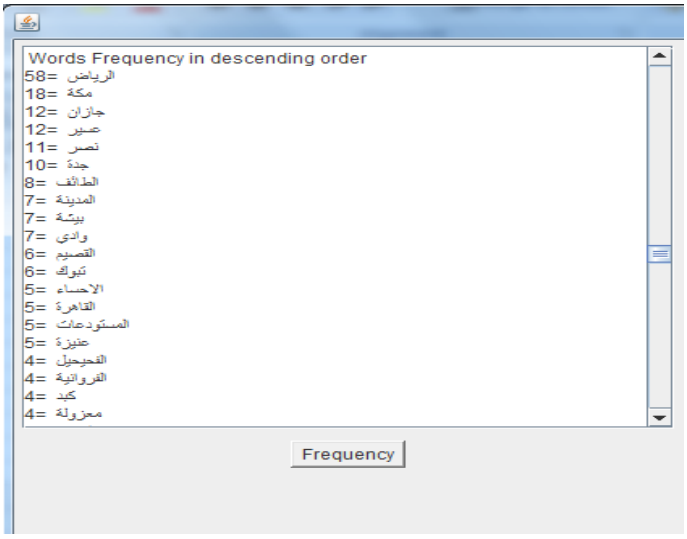
The CPS location local grammar was initially able to extract 376 tokens, which form the location dictionary (see Figure 5.8).



الاحساء الاحساء مكة الرياض الرياض وادي وادي الخرج الوادي الوادي
السليل وادي وادي بدر جدة تبوك الرايس بدر بدر القطيف صبيا الرياض
الرياض الشويخ شرق الفروانية الفروانية جليب الاحمدي مبارك القيروان
المهبولة السالمية شرق نصر كرداسة الزرايب كفر مني ابوالنمرس
الشويشة المستقبل صفط نصر حدائق نصر نصر القناطر القاهرة السلام
عابدين روض نصر الظاهر الظاهر الوايلي الغربية كفر المهندسين المنيب
الوايلي النزهة القليوبية بنها القليوبية الزاوية الزاوية حدائق القليوبية القاهرة
السبتيه السبتية بولاق الشيخ سبك سبك سبك الشروق الشروق طهطا نصر
نصر نصر المعادي الوايلي القاهرة القاهرة نصر نصر حلوان نصر حلوان
رمانة رمانة الاسماعيلية الجيزة شمال نصر القاهرة قصر المعادي وسط
القاهرة قليوب الطائف عرعر الحدود الرياض بيشة بيشة القطيف وادي
الطائف وادي الظهران الخرج الرياض جدة المدينة المدينة تبوك تبوك بارق

Figure 5.8: Sample of the location dictionary construction

The frequency analysis process was also applied here, and the results can be seen in Figure 5.9. Consequently, the location dictionary contains only 131 different location names. As already mentioned, the dictionary is used when the CPS location local grammar fails to extract the crime location. As can be seen, different cities names from different countries appear in this result, such as "القَاهرة / alqahrt / Cairo", "الريَاض / alryad / Riyadh", "الفروَانية / alfrwanyt / Al Farwaniyah" that located in Egypt, Saudi Arabia and Kuwait, respectively.



```
Words Frequency in descending order
58= الرياض
18= مكة
12= جازان
12= عسير
11= نصر
10= جدة
8= الطائف
7= المدينة
7= بيشة
7= وادي
6= القصيم
6= تبوك
5= الاحساء
5= القاهرة
5= المستودعات
5= عنيزة
4= الفحيحيل
4= الفروانية
4= كبد
4= معزولة
```

Figure 5.9: Frequency analysis results for the location dictionary

In the following experiment, the location dictionary was tested to extract crime locations that had not been identified directly by the location local grammar. Table 5.4 lists the new results after utilizing the location dictionary.

Table 5.4: Location extraction results after utilizing location dictionary

Dataset	True	False	Goal
Riyadh	38	2	41
Sabq	62	1	68
Okaz	137	4	145
Ahram	66	11	74
Alwatan	65	0	67
Alamalyawm	53	1	53
Gokarsat	26	0	27
Total	447	19	475

As can be seen, the assistance of the location dictionary has led to increasing the number of location entities that were correctly identified to 447 (from 368). As a result, 28 crime locations were not recognized either by the local grammar or the dictionary. Also, using the dictionary increased the incorrectly recognized entities to 19 (from 8).

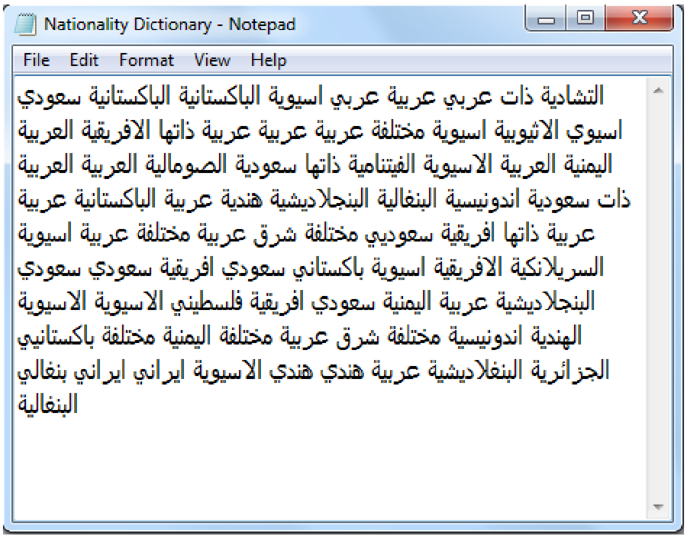
5.5.3 Nationality

The CPS nationality local grammar was tested in this experiment in order to examine its ability to extract nationality entities from the same datasets used in the above experiments. Table 5.5 shows the results of this experiment. The CPS was able to recognize 88 entities; 80 correct out of 210 entities. As a result, 8 entities were wrongly identified.

Table 5.5: Nationality extraction results using nationality local grammar

Dataset	True	False	Goal
Riyadh	11	1	17
Sabq	18	1	45
Okaz	28	5	65
Ahram	0	0	0
Alwatan	11	1	50
Alamalyawm	11	0	32
Gokarsat	1	0	1
Total	80	8	210

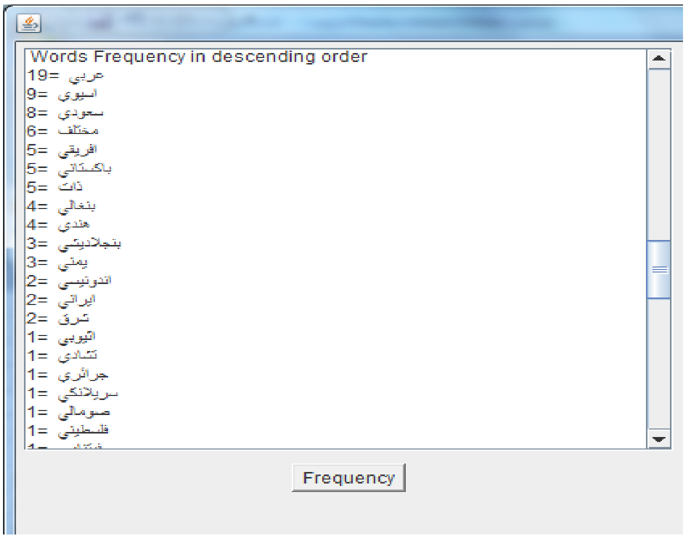
The system was able to generate the nationality dictionary; here, the number of tokens extracted was 88. The results from building this dictionary are presented in Figure 5.10.



```
التشادية ذات عربي عربية عربي اسيوية الباكستانية الباكستانية سعودي
اسيوي الاثيوبية اسيوية مختلفة عربية عربية عربية ذاتها الافريقية العربية
اليمنية العربية الاسيوية الفيتنامية ذاتها سعودية الصومالية العربية العربية
ذات سعودية اندونيسية البنغالية البنجلاديشية هندية عربية الباكستانية عربية
عربية ذاتها افريقية سعوديي مختلفة شرق عربية مختلفة عربية اسيوية
السريلانكية الافريقية اسيوية باكستاني سعودي افريقية سعودي سعودي
البنجلاديشية عربية اليمنية سعودي افريقية فلسطيني الاسيوية الاسيوية
الهندية اندونيسية مختلفة شرق عربية مختلفة اليمنية مختلفة باكستانيي
الجزائرية البنغلاديشية عربية هندي هندي الاسيوية ايراني ايراني بنغالي
البنغالية
```

Figure 5.10: Sample of the nationality dictionary

The processes of removing affixes and a frequency analysis were then applied, and the results can be seen in Figure 5.11. As a result, the number of words that form this dictionary is only 21.



```
Words Frequency in descending order
19= عربي
9= اسيوي
8= سعودي
6= مختلف
5= افريقي
5= باكستاني
5= ذات
4= بنغالي
4= هندي
3= بنجلاديشي
3= يمني
2= اندونيسي
2= ايراني
2= شرق
1= اثيوبي
1= تشادي
1= جزائري
1= سريلانكي
1= صومالي
1= فلسطيني
```

Figure 5.11: Frequency analysis results for the nationality dictionary

Likewise, the nationality dictionary was tested to extract the missing entities that had not been identified by the nationality local grammar. The results of this experiment are presented in Table 5.6.

Table 5.6: Nationality extraction results after utilizing nationality dictionary

Dataset	True	False	Goal
Riyadh	17	3	17
Sabq	41	6	45
Okaz	63	10	65
Ahram	0	0	0
Alwatan	36	3	50
Alamalyawm	27	2	32
Gokarsat	1	0	1
Total	185	24	210

Clearly, the number of nationality entities that were correctly identified after using the nationality dictionary improved to 185 (from 80). Also, it can be noticed that the total number of nationality entities that were incorrectly recognized increased to 24.

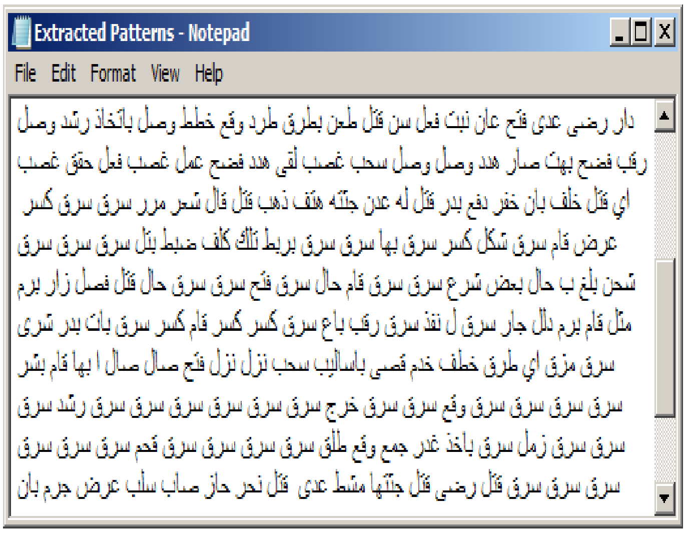
5.5.4 Clustering

Two experiments were carried out on 401 documents in order to show how the information extraction process guides the Self Organising Map (SOM) toward delivering acceptably accurate results. The corpus contains 71,882 tokens. The SOM was trained on the same documents, obtaining good results; the best learning rate, radius and iteration are 0.5, 30 and 1000, respectively. The size of the map is 6 x 6.

- Clustering with utilizing the CPS information extraction stage

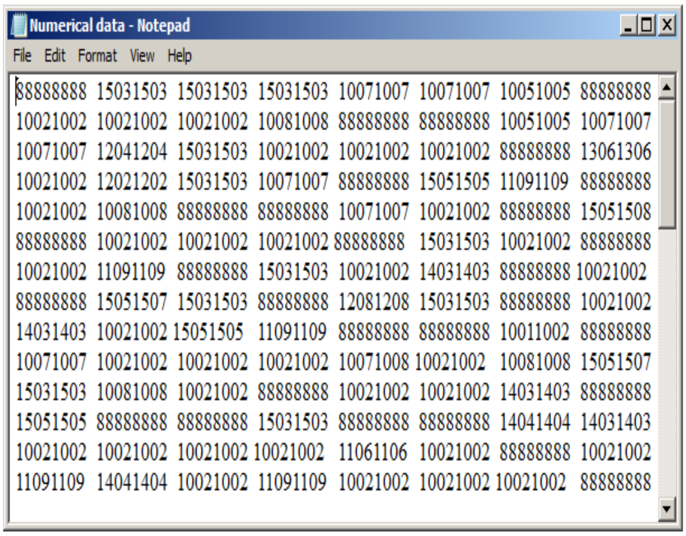
  As explained earlier, the information extraction process was employed to extract the types and locations of the crimes as well as the nationalities, and then a summary for each file as well as three dictionaries were generated. In this experiment, we focus on the type of crime. The extracted crime type patterns from each document are used by the SOM to perform the clustering, instead of processing the whole of each document's content. Figure 5.12 shows the extracted patterns after they have passed the stemming process. Also, Figure

5.13 presents the results of converting the extracted patterns into numbers in order to be sent to the clustering process.



دار رضى عدى فتح عان نبت فعل سن قتل طعن بطرق طرد وقع خطط وصل باتخاذ رشد وصل
رقب فضح بهت صار هدد وصل وصل سحب غصب لقى هدد فضح عمل غصب فعل حقق غصب
اي قتل خلف بان خفر دفع بدر قتل له عدن جثته هتف ذهب قتل قال شعر مرر سرق سرق كسر
عرض قام سرق شكل كسر سرق بها سرق سرق بربط تلك كلف ضبط بثل سرق سرق سرق
شحن بلغ ب حال بعض شرع سرق سرق قام حال سرق فتح سرق سرق حال قتل فصل زار برم
مثل قام برم دلل جار سرق ل نفذ سرق رقب باع سرق كسر كسر قام كسر سرق بات بدر شرى
سرق مزق اي طرق خطف خدم قصى باساليب سحب نزل نزل فتح صال صال ا بها قام بشر
سرق سرق سرق سرق وقع سرق سرق خرج سرق سرق سرق سرق سرق رشد سرق
سرق سرق سرق زمل سرق باخذ غدر جمع وقع طلق سرق سرق سرق سرق فحم سرق سرق سرق
سرق سرق سرق قتل رضى قتل جثتها مشط عدى قتل نحر حاز صاب سلب عرض جرم بان

Figure 5.12: The extracted patterns from the crime news reports



```
88888888 15031503 15031503 15031503 10071007 10071007 10051005 88888888
10021002 10021002 10021002 10081008 88888888 88888888 10051005 10071007
10071007 12041204 15031503 10021002 10021002 10021002 88888888 13061306
10021002 12021202 15031503 10071007 88888888 15051505 11091109 88888888
10021002 10081008 88888888 88888888 10071007 10021002 88888888 15051508
88888888 10021002 10021002 10021002 88888888 15031503 10021002 88888888
10021002 11091109 88888888 15031503 10021002 14031403 88888888 10021002
88888888 15051507 15031503 88888888 12081208 15031503 88888888 10021002
14031403 10021002 15051505 11091109 88888888 88888888 10011002 88888888
10071007 10021002 10021002 10021002 10071008 10021002 10081008 15051507
15031503 10081008 10021002 88888888 10021002 10021002 14031403 88888888
15051505 88888888 88888888 15031503 88888888 88888888 14041404 14031403
10021002 10021002 10021002 10021002 11061106 10021002 88888888 10021002
11091109 14041404 10021002 11091109 10021002 10021002 10021002 88888888
```

Figure 5.13: Results of the word indexing process

Accordingly, after extracting the type of crime, the new size of the corpus is now 4,043 tokens (13KB), which is much smaller than the original size of 71,882 tokens (40KB). Figure 5.14 shows a sample of the result of the document clustering based on type of crime, using the extracted patterns obtained from the previous processes.

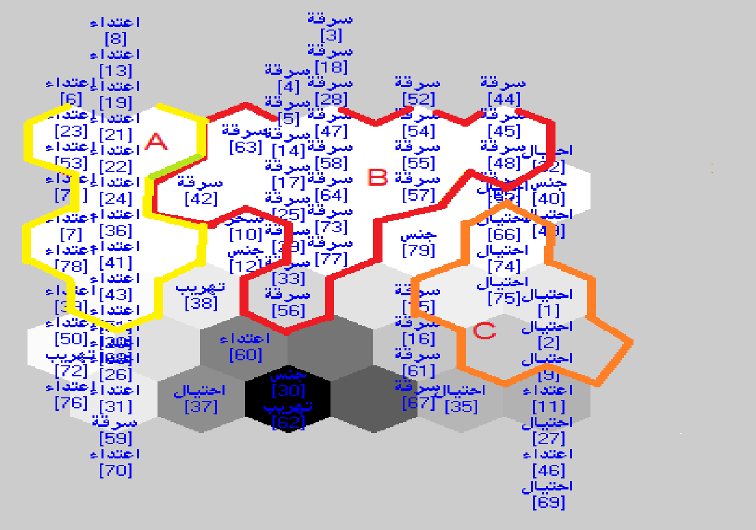


Figure 5.14: Clustering results with aid of the CPS information extraction (A: Violence, B: Theft and C: Fraud)

- Clustering without using the CPS information extraction stage

  For assessing this work in terms of the effectiveness of the clustering, another experiment was carried out on the same corpus, but this one did not rely on the information extraction process. The whole content of each file was stemmed and used for the clustering process through the SOM. A sample of the result of this experiment can be seen in Figure 5.15.

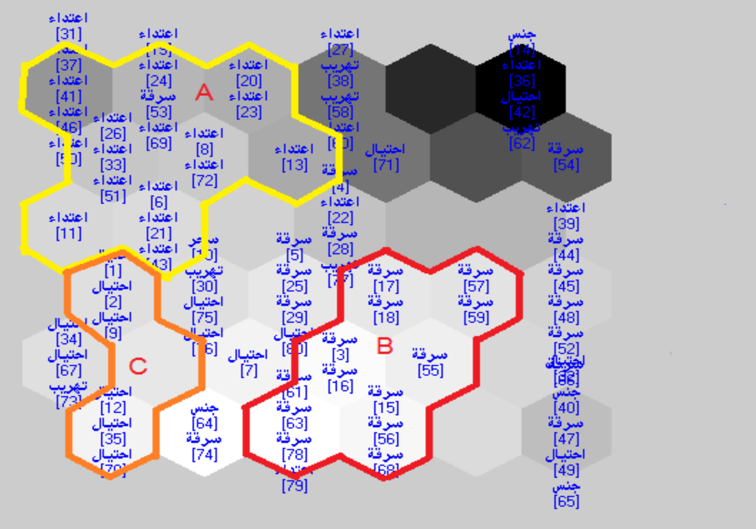
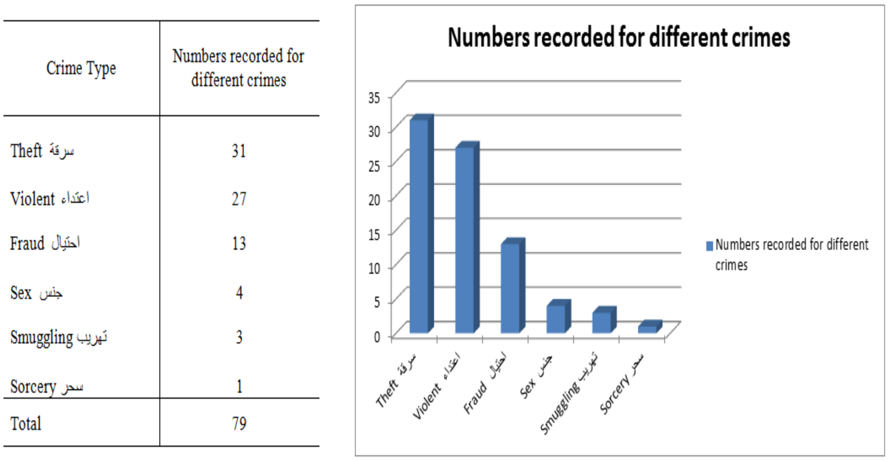


Figure 5.15: Clustering results without aid of the CPS information extraction (A: Violence, B: Theft and C: Fraud)

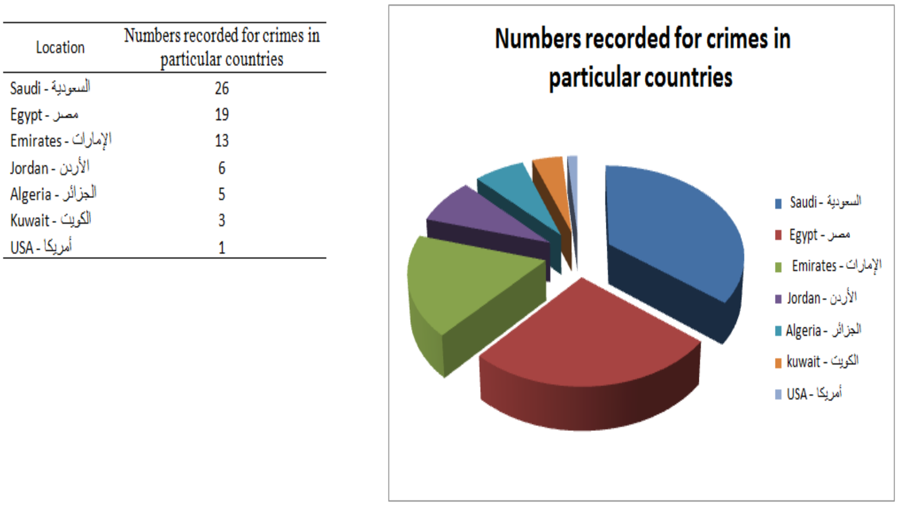
5.5.5 Crime Analysis

An additional benefit of this work is that this current system can be easily adapted to provide crime profiling for regions. In other words, it can be used to present a general picture about the security status of any area, based on local news reports. The system can offer statistical information about the highest and/or lowest type of crime. Figure 5.16 shows that the crime of theft is the most common crime occurring in the Arab region; it is reported 31 times in our corpus (79 crime news reports). In addition, extracting the crime location can assist in identifying how safe a particular area is, and through combining such statistics, this system is able to provide information about the number of crimes occurring in a specific location. Figure 5.17 depicts the number of crimes and their location in a pie-graph. It shows the numbers of crimes that happened in Saudi Arabia, Egypt, the United Arab Emirates, Jordan, Algeria, Kuwait and the USA.



Crime Type	Numbers recorded for different crimes
Theft سرقة	31
Violent اعتداء	27
Fraud احتيال	13
Sex جنس	4
Smuggling تهريب	3
Sorcery سحر	1
Total	79

Figure 5.16: Averages for the different crime types, reported across the Arab countries mentioned above



Location	Numbers recorded for crimes in particular countries
Saudi - السعودية	26
Egypt - مصر	19
Emirates - الإمارات	13
Jordan - الأردن	6
Algeria - الجزائر	5
Kuwait - الكويت	3
USA - أمريكا	1

Figure 5.17: Numbers of crimes reported by location

5.6 Summary

This chapter has presented a real implementation of our Crime Profiling System (CPS) for extracting meaningful crime information, i.e. crime type, crime location and nationality, by utilizing the proposed computational linguistic techniques. Moreover, it has been shown that the system is able to extract meaningful crime information from an unannotated corpus, to automatically construct dictionaries, to generate summarisations and to cluster Arabic crime texts (employing the Self Organising Map (SOM) technique). Also, the developed system can assist in providing other useful information, e.g. general and specific crime trends (frequencies of crime within a particular area) to law enforcement bodies or the general public. Thus, the strategic and administrative crime analysis mentioned in Chapter 1 has been almost fully achieved in this research. The results of the performance evaluation of the CPS are presented in the following chapter.

Chapter 6

Evaluation

Objectives

- Discuss the results of the experiments with the error analysis.
- Present the performance evaluation of Crime Profiling System.

6.1 Introduction

As already seen in Chapter 5, the experiments were performed in order to reveal the extent to which the system is effective in extracting meaningful patterns, generating summaries, automatically building dictionaries, clustering crime reports, producing statistical information about crime, and displaying crime locations on maps (using the GIS tool).

This chapter is dedicated to evaluating the performance of the Crime Profiling System (CPS), i.e. to assessing the efficacy of the crime type, location and nationality local grammars used in this research as well as to evaluating the effect of utilising dictionaries on the performance of the CPS. The system is evaluated using precision, recall and F-measure, as in section 6.2. Additionally, the efficacy of the information extraction approach with respect to the performance of the Self Organising Map (SOM) is evaluated through four parameters: data size, loading time, computation time and quantization error, as will be seen in section 6.3.

6.2 Evaluating the Information Extraction

The performance evaluation of IE systems is carried out by comparing the answer file (result) that was automatically generated by the system against the same texts that were manually produced by humans (gold standard file) [86]. Accordingly, each of 401 crime news reports was read to identify each type of crime, crime location and nationality. We manually annotated all the aforementioned information in the crime reports and then we counted the total number of relevant entities that should be extracted by our system in order to evaluate it in terms of Precision (P) and Recall (R). Although several evaluation metrics could be found in the literature, the most popular and official metrics in series of Message Understanding Conferences

MUS3, MUC4 and MUC6 are P and R [82, 171]. For more details about evaluation metrics see [171] and [172].

Gaizauskas and Wilks [171] defined recall as ”a measure of the fraction of the required information that has been correctly extracted” and precision as ”a measure of the fraction of the extracted information that is correct”. These two metrics are combined using F-measure, which is the weighted harmonic mean of precision and recall, because there is usually trade off of precision against recall [173, 174]. The precision, recall and F-measure are formulated as follows:

$$Precision = \frac{number\,of\,correctly\,recognised\,entities}{total\,number\,of\,recognised\,entities}$$

$$Recall = \frac{number\,of\,correctly\,recognised\,entities}{total\,number\,of\,correct\,entities}$$

$$F - measure = \frac{2\,\times\,recall\,\times\,precision}{recall\,+\,precision}$$

The performance results obtained after evaluating the CPS in terms precision, recall and F-measure for the crime type, location and nationality extraction processes are presented in the following sections.

6.2.1 Type of Crime

The system was able, directly (using the crime type local grammar) and through using the crime type dictionary, to extract a total number of 834 entities. The number of entities that were correctly recognised is 481 out of a total number of 496 relevant entities. The performance results achieved for the crime type extraction process through using the crime type local grammar and the crime type dictionary are presented in Tables 6.1 and 6.2.

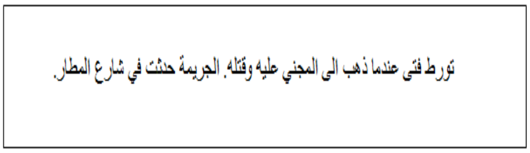
Table 6.1: The CPS evaluation results using crime type local grammar

Dataset	Precision (%)	Recall (%)	F-measure (%)
Riyadh	44	93	60
Sabq	48	92	63
Okaz	53	68	60
Ahram	51	77	61
Alwatan	56	84	67
Alamalyawm	63	86	73
Gokarsat	49	71	58
Overall	53	80	63

Table 6.2: The CPS evaluation results after using crime type dictionary

Dataset	Precision (%)	Recall (%)	F-measure (%)
Riyadh	46	100	63
Sabq	47	97	64
Okaz	56	95	70
Ahram	52	100	68
Alwatan	58	100	73
Alamalyawm	62	94	75
Gokarsat	50	92	65
Overall	54	97	69

The results derived from the 401 crime news reports show that using the crime type dictionary has enabled the CPS to perform better (F-measure 69%). These results indicate that the CPS was able to build a relatively reliable crime type dictionary. This means that the performance of the crime local grammar seems satisfactory, either for building the dictionary or correctly recognising the type of crime (by obtaining a recall score of 80%). The remaining unidentified entities result from the crime action words being outside the local grammar, i.e. they were used for describing the incident in the form of a verb. As can be seen, the assistance of the dictionary has led to improving the recall result to 97%. However, the precision value seems low, because the sentences' boundaries were neglected. In other words, removing commas and full stops caused some confusion for the system while the text was being processed. As a result, when the system was searching for the preposition that should follow the verb (although not necessarily in all cases) in order to achieve the syntactic constraint (for the transitive construction), it proceeded into the following sentence in order to find that preposition, and this led to incorrect pattern extraction. Figure 6.1 shows an example explaining this case.



تورط فتى عندما ذهب الى المجني عليه وقتله. الجريمة حدثت في شارع المطار.

Figure 6.1: Sentence containing verb in intransitive construction

The verb "تورط / tawarat / involved" in the first sentence is in the intransitive construction because its companion preposition "في / fi / in" does not follow it. Therefore, the system should ignore this verb because it is not a transitive verb but because the full stop between the two sentences is removed, the system carries on searching for the preposition "في / fi / in". As a result, the system selects the preposition "في / fi / in" located in the second sentence. This preposition, performing a special role within the second sentence, has no relationship with the first sentence. Hence, a false transitive verb is used to extract a type of crime, which leads to an incorrectly extracted pattern. As a consequence, commas and full stops should be retained, and the text should be split into sentences using the sentence splitter process.

The evaluation results vis-a-vis extracting crime type with or without the crime type dictionary have been compared with system developed by Abuleil [105], which, to our knowledge, is the only system that has been developed for the Arabic language to extract events from within text (although not specific to any particular domain). It was chosen for comparison with the CPS because there is no system available that has been specifically developed for the Arabic crime domain. However, to overcome this problem, we compared our CPS with the system created for the English crime domain by Ku et al. [109]; Both systems were discussed in Chapter 2. Figure 6.2

shows the comparison between the three systems in terms of their performance in extracting events.

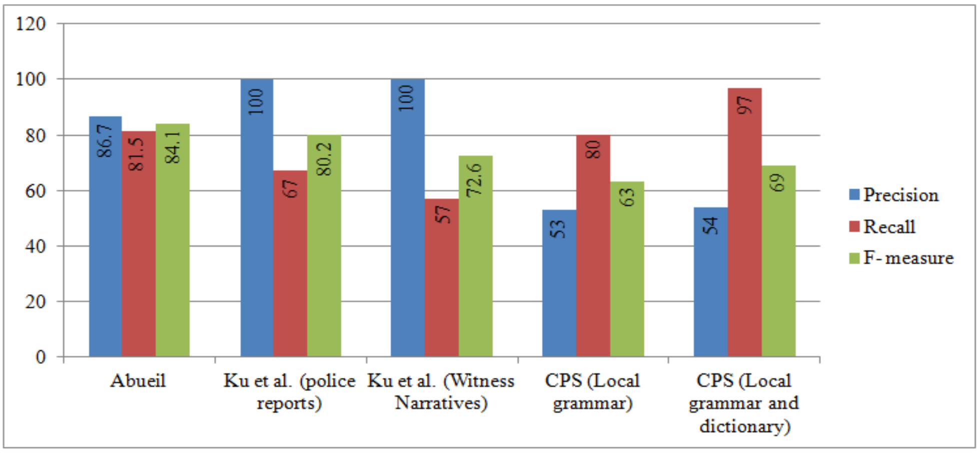


Figure 6.2: Performance of CPS compared with other systems in terms of extracting crime type

Although the performance comparison in terms of F-measure shows that systems developed by Abuleil [105] and Ku et al. [109] obtained results better than our system (CPS), both systems use external predefined event dictionaries, which leads to obtaining high precision scores. On the other hand, the CPS does not utilise any external event list, rather it makes use of the automatically built crime type dictionary. However, the CPS outperforms the others in terms of the recall score.

6.2.2 Crime Location

As already seen in the previous chapter, the CPS was able to correctly recognise 447 entities out of 475. Tables 6.3 and 6.4 show the evaluation results for the location extraction process, using the location local grammar and the location dictionary.

Table 6.3: The CPS evaluation results using location local grammar

Dataset	Precision (%)	Recall (%)	F-measure (%)
Riyadh	100	83	91
Sabq	100	74	85
Okaz	97	67	79
Ahram	91	72	80
Alwatan	100	88	94
Alamalyawm	100	94	97
Gokarsat	100	93	96
Overall	98	77	86

Table 6.4: The CPS evaluation results after using location dictionary

Dataset	Precision (%)	Recall (%)	F-measure (%)
Riyadh	95	93	95
Sabq	98	91	95
Okaz	97	94	96
Ahram	86	89	87
Alwatan	100	97	98
Alamalyawm	98	100	99
Gokarsat	100	96	95
Overall	96	94	95

The precision and recall results show that the location local grammar was able to extract 77% of the location entities, with a 98% precision rate. On the other hand, using the location dictionary has led to increasing in the recall result to 94% with precision a rate of 96%. It can be noticed that, after utilising the location dictionary, the precision is slightly decreased; this is because of a lack of semantics. It is found that the location dictionary contains personal name entities; these were extracted and classified as locations by the location local grammar because they are also location names. As a result, incorrect recognition occurs when the dictionary, which neglects semantics, is used by the CPS. Moreover, sometimes crime locations cannot be discovered by the location local grammar, and therefore the CPS fails to capture them. For example, in the sentence ”الجريمة حدثت في دبي / aljrimt hdthat fi dubai / the crime happened in Dubai”, the city name ”Dubai” cannot be identified because it does not follow these words: " منطقة / mantqt / area", " بمنطقة / bi-mantqt / in area", " محَافظة / mohafdat / province", " مدينة / mdynt / city" or " ولَاية / wlayt / state". However, the location dictionary overcomes this problem, and it increased the recall value to 94% (from 77%). Consequently , the F-measure score increased to 95% (from 86%).

The CPS performance in terms of extracting location entities with and without (i.e. only using the location local grammar) utilising the location dictionary was compared with systems that utilise a rule-based approach (TAGARAB [98]; Mesfar [101]; NERA [102] and Ku et al. [109]), as in Figure 6.3.

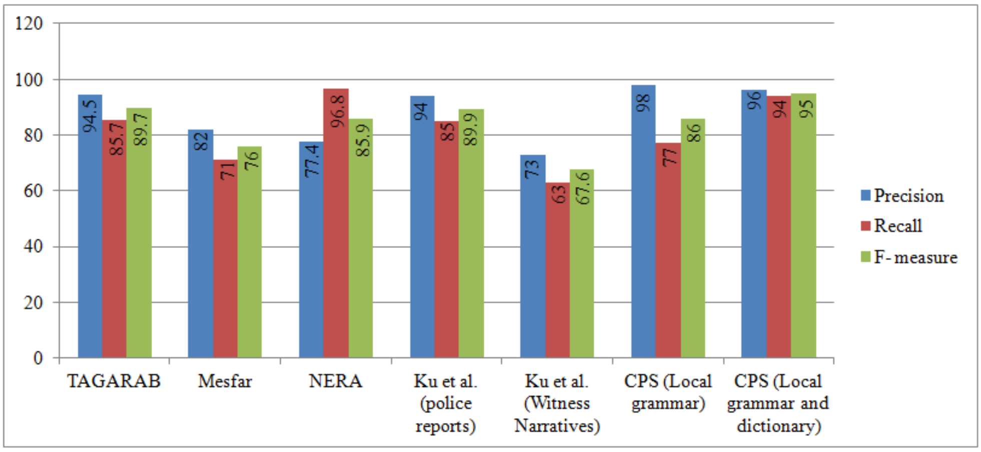


Figure 6.3: Comparison between CPS and other systems that used rule-based method in terms of extracting location

Also, Figure 6.4 provides a comparison between the CPS (with and without using the location dictionary) with other systems that were developed based on machine learning approaches (ANERsys using Maximum Entropy (ME) [11]; ANERsys using Conditional Random Field (CRF) [108] and AbdelRahman [57]).

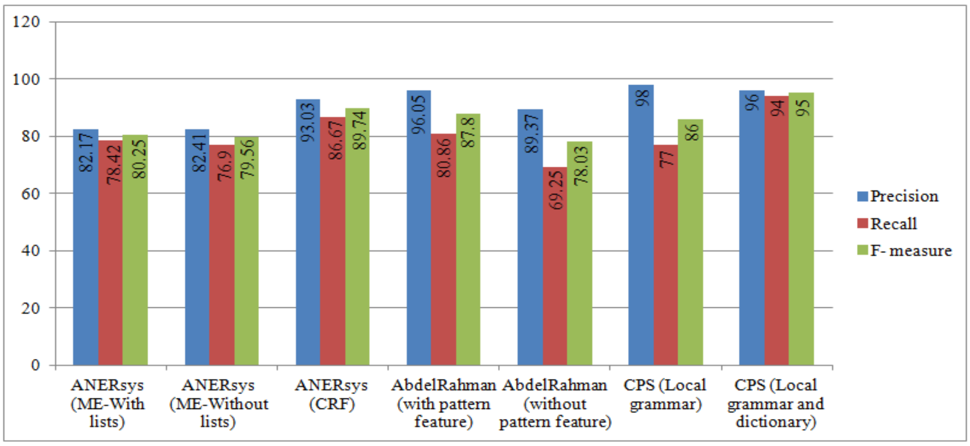


Figure 6.4: Comparison between CPS and other systems that used machine learning in terms of extracting location

A comparison of the accuracy of the CPS with these other systems reveals that, with the assistance of its dictionaries, the CPS is the second best system (after NERA [102]) in terms of recall. However, a predefined location dictionary containing 4,900 names was utilised in NERA [102]. Moreover, the CPS is approximately equal to the top system AbdelRahman [57] in terms of precision, although AbdelRahman [57] uses a predefined location dictionary (of 2,183 names). However, the CPS achieved the best performance result (F-measure 95%).

6.2.3 Nationality

As already seen, the nationality local grammar and nationality dictionary together were able to correctly recognise 185 entities out of 210, with 24 entities wrongly extracted. The results of the performance evaluation for the CPS with using the nationality local grammar and nationality dictionary are listed in Tables 6.5 and 6.6.

Table 6.5: The CPS evaluation results using nationality local grammar

Dataset	Precision (%)	Recall (%)	F-measure (%)
Riyadh	92	65	76
Sabq	95	40	56
Okaz	85	43	57
Alwatan	92	22	35
Alamalyawm	100	34	51
Gokarsat	100	100	100
Overall	91	38	54

Table 6.6: The CPS evaluation results after using nationality dictionary

Dataset	Precision (%)	Recall (%)	F-measure (%)
Riyadh	85	100	92
Sabq	87	91	89
Okaz	86	97	91
Alwatan	78	54	64
Alamalyawm	92	72	81
Gokarsat	100	100	100
Overall	86	88	88

The results for precision and recall obtained by applying only the nationality local grammar in recognising nationality entities in the above dataset are 91% and 38%, respectively. Although a high precision value is obtained, the rate for recall is too low, i.e. many nationality entities were not identified. This means that certain entities appear to be outside the nationality local grammar. In some newspapers, it is found that a nationality word (e.g. Saudi, Indian or British) is coupled with the word "وَافد / wafd / expatriate", e.g. "وَافد هندي / wafd hndy / Indian expatriate" instead of using the word "الجنسية / aljnsyt / nationality", e.g. "هندي الجنسية / hndy aljnsyt / Indian nationality". This has led to obtaining a low recall score. However, the dictionary plays a crucial role here, and the results of utilising the nationality dictionary show that the recall rate is increased to 88%. Therefore, the average F-measure value improved to 88% (from 54%). There is, to the author's knowledge, no system developed for extracting this type of entity in Arabic text.

6.2.4 The CPS Overall Performance Results

Table 6.7 lists the overall performance results for the CPS. As can be seen, the performance of the CPS is improved through utilising the dictionaries (in terms of precision, recall and F-measure). Although the CPS does not utilise an annotated corpus or predefined dictionaries, as other systems do, it achieved relatively satisfactory results.

Table 6.7: The overall CPS evaluation results

	Precision (%)	Recall (%)	F-measure (%)
CPS (local grammar)	69	72	70
CPS (with dictionary)	71	94	81

6.2.5 The CPS with Sport Domain

In the following experiment, the performance of the Crime Profiling System (CPS) was tested with another domain (sport domain) in order to reveal the extent to which the system is effective in identifying crime types committed within sport context from sport news reports. In this experiment, the dataset containing 34 sport news reports and comprising 6,008 tokens were used to conduct this test. The CPS crime type local grammar was able to extract 38 entities. The number of entities that were correctly recognised is 24 out of a total number of 34 relevant entities. The performance result achieved for the crime type extraction process through using the crime type local grammar is presented in Table 6.8.

Table 6.8: The CPS evaluation result for extracting crime type from sport news reports

	Precision (%)	Recall (%)	F-measure (%)
CPS	63	71	67

It can be seen that, the system was relatively able to identify a crime type from sport news reports (sport domain) using the same computational linguistic technique (transitive verbs) that was derived from crime domain.

6.3 Clustering Performance

The Self Organising Map (SOM) was used for the clustering and visualisation tasks, and for assessing the effectiveness of the proposed approach on the SOM outputs (i.e. in terms of its clustering performance). As already seen, the SOM was able to cluster 401 texts and to visualise them. The evaluation phase here is performed based on four parameters, as follows:

- Data size
- Loading time
- Execution time
- Quantization error

6.3.1 Data Size and Loading Time

With regards to the size of data, the significant point here is that using the CPS led to a huge reduction in the quantity of data fed into the SOM. Although our system reduced the size of the corpus from 71,882 tokens (40KB) to only 4,043 tokens (13KB), the most important data (that the SOM used for the clustering task) were not affected. Thus, these 4,043 tokens can be considered as effectively representing the original 71,882 tokens. The clustering experiment (utilising the CPS information extraction stage) was assessed by comparing it with a clustering of the same documents but without the CPS information extraction stage (see Section 5.5.4), i.e. where the SOM processed the whole of each document's content. Figure

6.5 shows the loading time after using the CPS, which was 1.51 seconds, and in the other experiment (without the CPS) the loading time was 3.99 seconds, i.e. the loading time was reduced by more than a half with the CPS.

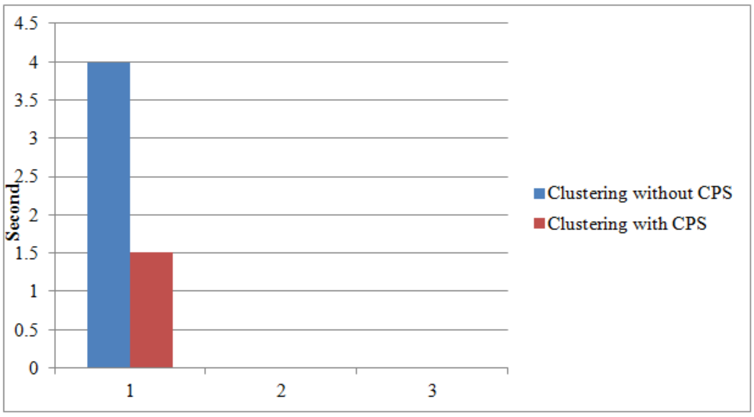


Figure 6.5: A comparison in terms of the loading time with/without using the CPS information extraction

6.3.2 Execution time

As evident in Figure 6.6, the time spent in executing both experiments was measured and the two experiments (with/without the CPS information extraction stage) were repeated five times to ensure the validity of the result. It can be noticed that the second experiment (without the CPS information extraction stage) took longer than the first to process the data in order to perform the clustering. Therefore, the CPS increased the speed of the clustering process.

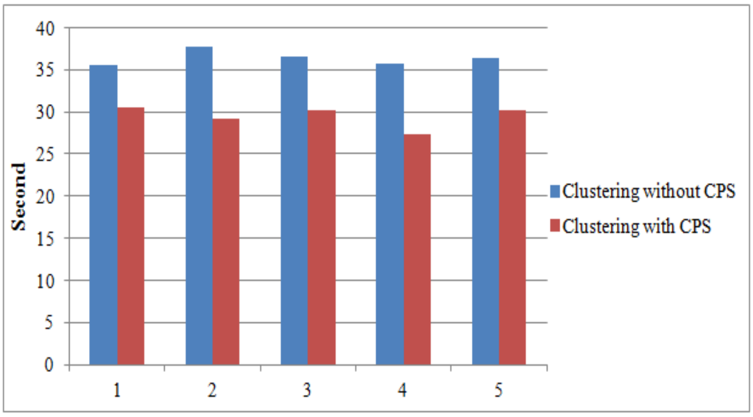


Figure 6.6: The time taken by the SOM to perform the clustering tasks

6.3.3 Quantization Error

The average distance between each data vector and its BMU (quantization error) in the experiment that was supported by the CPS was between 0.462 and 0.47, but in the second experiment (without the information extraction process), the quantization error was between 0.59 and 0.594. Therefore, the performance of the SOM in the experiment that relied on the CPS information extraction represents an improved technique in terms of the quality of clustering; the information extraction process has a strongly positive effect on the performance of the SOM.

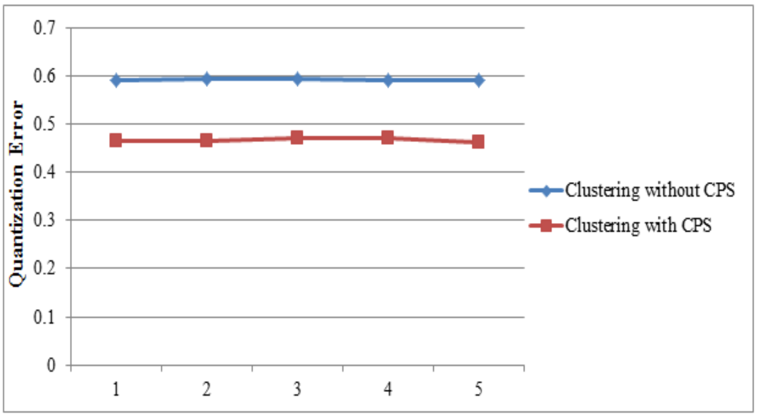


Figure 6.7: The average distance between each data vector and its BMU (Quantization Error) in the two experiments

6.4 Summary

The main aim of this work was to develop an information extraction technique for extracting crime type, crime location and nationality from Arabic news reports. This chapter has presented the evaluation of the proposed system. The performance of Crime Profiling System (CPS) for extracting the aforementioned crime-related information using both the local grammars and dictionaries were evaluated using standard precision, recall and F-measure. Moreover, the use of the automatically constructing dictionaries in improving the performance of the CPS was discussed. Also, the Self Organising Map (SOM), as a clustering technique, was used to evaluate the effectiveness of employing the CPS information extraction approach, based on four parameters; size of data, loading time, execution time and quantization error. This evaluation, performed through comparative experiments, was conducted between the SOM clustering technique with and without using the CPS information extraction approach.

Chapter 7

Conclusion and Future Work

Objectives

- Present the summary of the thesis.
- Provide future work.

7.1 Summary

Text Mining can be described as the process of extracting particular information from within unstructured data, thereby facilitating access to potentially valuable information for use in a wide variety of fields. The importance of this field has grown because a great deal of data is stored as free text, which is difficult to mine, and so, currently, there is an urgent need for effective tools to deal with such data. Text mining techniques have been applied to many languages but although Arabic is a widely spoken language, few mining tools have been developed to process Arabic text.

Accordingly, this thesis has presented the Crime Profiling System (CPS), which has been developed for the Arabic crime domain to extract meaningful crime information (crime type, crime location and nationality), automatically construct dictionaries for the above information, cluster crime documents based on their similarity, and utilise visualisation techniques to assist in crime data analysis.

As already seen through this thesis, the proposed system has answered the research question, which was described in Chapter 1. We have shown that the CPS is able to achieve the following tasks:

- Extract meaningful crime information (crime type, location and nationality) from unstructured Arabic text within the crime domain in order to assist in crime analysing in terms of accelerating the investigative process. Users (e.g. police investigators) are automatically provided with the most significant information instead of reading entire reports. Moreover, the CPS exploits the peculiarity and the nature of the Arabic language in achieving the information extraction task by utilising computational linguistic techniques based on transitive and genitive constructions. Three analyses: frequency, collocation and concordance were used for developing the information extraction approach. Also, the N-gram was employed in identifying the predefined keywords when they are inflected.

- The development of the above information extraction technique leads to generating a summary for each crime news report and to building three dictionaries automatically (for crime type, location and nationality). These dictionaries assist in the information extraction process, and therefore, improve the performance of the CPS.

- The CPS is able to cluster crime reports with a satisfactory degree of performance. The Self Organising Map (SOM) neural network was used for performing the clustering and visualisation tasks. The SOM received high quality data (report summaries), extracted during the information extraction process. As a result, high quality clustering results have been achieved.

- The strategic and administrative crime analysis mentioned in Chapter 1 has been almost fully achieved in this research. The CPS is able to generate statistical data about various crime types, benefitting law enforcement bodies and the general public.

- Various visualisation methods are used for presenting the outputs of the CPS. The results of the information extraction process, the statistical data generated in the form of tables and graphs, the clustering results and spatial analyses for displaying the crime locations on the map can all be visualised.

7.2 Contribution

Developing an automatic Crime Profiling System (CPS) for the Arabic language within the crime domain has been the main aim of this research. The main contributions of this thesis are as follows:

- Automatically Constructing Dictionaries

  Traditional systems usually rely on manually built dictionaries, which is time consuming. The CPS is able to automatically generate crime type, location and nationality dictionaries from unlabelled data to assist in extracting patterns from texts.

- Keyword-based Clustering

  The CPS has improved the performance of the Self Organising Map (SOM) in terms of quality of clustering, loading time and computational time. This is because the patterns extracted from each document are derived only from keywords that reflect the essential meaning of the document. These keywords are used by the SOM to perform the clustering, instead of having to process the entire contents of each document.

- Arabic Crime Corpus

  Due to the fact that there is no available dataset that can be used to develop applications in the context of crime, and due to the difficulties in obtaining official data or narrative reports from police stations (this is especially so in Arab countries), a special corpus called Arabic Crime News Report Corpus (ACNRC) has been compiled from a wide collection of Arabic newspapers.

7.3 Future Work

Many ideas have been generated by the work presented in this thesis which they will lead to extend our work. These ideas are as follows:

- Currently, the developed system is only able to extract crime type, crime location and nationality, and therefore, there is a need to improve this system in order to extract more information about crime. This could be done to answer questions such as to who the perpetrator is and who the victim is as well as other attributes (e.g. age and gender). Also, extracting the instruments used in crime incidents, such as weapons and vehicles, would assist in the investigation process by obtaining pertinent information more speedily. However, it seems clear that the same approach would be used to extract the above information, i.e. the transitive construction. In the frequency analysis, the words "قبض / qbd / arrested" and "عثر / athr / found" appeared, and they may assist in extracting the crime perpetrator and instrument entities if they are followed by the preposition"علَى / ala / on".

- The current system could be extended to predict crime through the analysis of crime trends.

- The Geographic Information System (GIS) will be employed in this research, which helps to more effectively perform spatial analyses and to visually locate crime 'hot-spots'.

- In this research, a tokenization process was implemented on the word level. Future work might include a sentence splitter process to split the text into sentences in order to increase level of the accuracy.

- Instead of using a single information extraction approach, future work might combine the current approach with a machine learning approach.

- Another future direction could involve the application of our information extraction approach to other domains, such as sport or politics, to extract particular information, such as event type, location of the event and the nationalities of people who participated.

Bibliography

[1] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, and Homa Atabakhsh. Crime data mining: an overview and case studies. In *Proceedings of the 2003 annual national conference on Digital government research*, dg.o '03, pages 1–5. Digital Government Society of North America, 2003.

[2] William McKnight. Building business intelligence: text data mining in business intelligence. In *DM Review.*, pages 21–22, 2005.

[3] Jakkrit TeCho, Cholwich Nattee, and Thanaruk Theeramunkong. A corpus-based approach for keyword identification using supervised learning techniques. In *Proceedings of the 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, volume 1, pages 33 –36, May 2008.

[4] Michael Chau, Jennifer J. Xu, and Hsinchun Chen. Extracting meaningful entities from police narrative reports. In *Proceedings of the annual national conference on Digital government research*, pages 271–275. Digital Government Society of North America, 2002.

[5] Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini, and Josef van Genabith. An automatically built named entity lexicon for arabic. In *LREC 2010. Valletta, Malta*, pages 3614–3621, 2010.

[6] Ahmed Abdul-Hamid and Kareem Darwish. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop, ACL 2010*, pages 110–115, 2010.

[7] Atallah Mahmoud Al-Shatnawi and Khairuddain Omar. Methods of arabic language baseline detection the state of art. *Arab Research Institute in Sciences and Engineering (ARISER)*, 4:158–193, 2008.

[8] Riyad Al-Shalabi, Ghassan Kanaan, Bashar Al-Sarayreh, Khaled Khanfer, Ali Al-Ghonmein, Hamed Talhouni, and Salem Al-Azazmeh. Proper noun extracting algorithm for arabic language. In *International Conference on IT, Thailand*, 2009.

[9] Shereen KHOJA. Apt: Arabic part-of-speech tagger. In *Proc. of the Student Workshop at NAACL*, 2001.

[10] Y.O. Mohamed El Hadj, I.A. Al-Sughayeir, and A.M. Al-Ansari. Arabic part-of-speech tagging using the sentence structure. In *Proceeding of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt*, pages 241–245, 2009.

[11] Yassine Benajiba, Paolo Rosso, and Jose Ruiz. Anersys: An arabic named entity recognition system based on maximum entropy. In *CICLing*, pages 143–153, 2007.

[12] Ibn Alnazim Abu Abdullah Badralddin. *Explanation Ibn Alnazim on Alfiyat Ibn Malik*. Dar Alkotob Alalmyt, 2000.

[13] Moheiddin A. Homeidi. The notion of governor in modern standard arabic (msa). *Journal of King Saud University. Languages & Translation*, 15(1):49–62, 2003.

[14] Nizar Habash, Ryan Gabbard, Owen Rambow, Seth Kulick, and Mitch Marcus. Determining case in arabic: Learning complex linguistic behavior requires complex linguistic features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, page 10841092, 2007.

[15] Mohammed Jiyad. *A Hundred and One Rules. A Short Reference for Arabic Syntactic, Morphological & Phonological Rules for Novice & Intermediate Levels of Proficiency.* 2006.

[16] Daphna Daphna. On the construct state, uniqueness and genitive relations. In *Proceedings of IATL 18*, 2002.

[17] Naglaa Ghali. *Arabic Grammar Unravelled.* Fun With Arabic, 2007.

[18] Ron Buckley. *Modern Literary Arabic: A Reference Grammar.* Beirut : Librairie du Liban Publishers, 1st ed edition, 2004.

[19] Aitao Chen and Fredric Gey. Building an arabic stemmer for information retrieval. In *Proceedings of TREC*, pages 631–639, 2002.

[20] Moulana Abdus Sattar Khan. *Arabic Tutor*, volume 2. Madrasah Inamiyyah, Camperdown, South Africa, first edition edition, 2007.

[21] Reima Al-Jarf. *A Contrastive Analysis of English and Arabic Morphology for Translation Students.* AL-Obeikkan Printing Press, Riyadh, Saudi Arabia, 2000.

[22] Mosa Mohammad Almlyani Alahmadi. *Lexicon of Transitive Verbs by Preposition.* Dar Alalm Lelmalayyn, 1986.

[23] George O. Curme. *Grammar of the English Language: Parts of Speech and Accidence*, volume 2. Heath and Company, New York., 1935.

[24] Randolph Quirk and Sidney Greenbaum. *A University Grammar of English.* Longman, Great Britain., 1973.

[25] Muhammad Soliman Ibrahim Fiteih. *Preposition and Prepositional Verbs in Classical Arabic.* PhD thesis, Department of Linguistics and Phonetics, The University of Leeds, 1983.

[26] Karin C. Ryding. *A Reference Grammar of Modern Standard Arabic.* Cambridge University Press, 2005.

[27] Marina Najjar. The arabic prepositions: their original meanings and their contemporary use. Master's thesis, American University of Beirut, 1986.

[28] Salahalddin Alzablawy. Grammarians and prepositions. *Arab Heritage Journal*, 32, 1988.

[29] Arnold C. Satterthwait. Computational research in arabic. *Mechanical Translation*, 17(2):62–70, 1963.

[30] Kais Dukes, Eric Atwell, and Abdul-Baquee M. Sharaf. Syntactic annotation guidelines for the quranic arabic dependency treebank. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1822–1827, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[31] Allan Ramsay and Hanady Mansour. Towards including prosody in a text-to-speech system for modern standard arabic. *Comput. Speech Lang.*, 22:84–103, January 2008.

[32] Mohammed Attia. Report on the introduction of arabic to pargram. In *The ParGram Fall Meeting 2004, The National Centre for Language Technology, School of Computing, Dublin City University, Ireland*, 2004.

[33] Islam Al-Momani. Case-assignment under government in modern literary arabic. *LANGUAGE IN INDIA*, 10:24–64, 2010.

[34] merriam webster. Definition of crime. http://www.merriam-webster.com/dictionary/crime, Accessed [05/06/2012].

[35] P. Thongtae and S. Srisuk. An analysis of data mining applications in crime domain. In *Proceedings of the IEEE 8th International Conference on Computer and Information Technology Workshops*, pages 122–126, Washington, DC, USA, 2008. IEEE Computer Society.

[36] Deborah Osborne and Susan Wernicke. *Introduction to Crime Analysis Basic Resources for Criminal Justice Practice*. The Haworth Press, Inc. 10 Alice Street, Binghamton, NY 13904-1580., 2003.

[37] Steven Gottlieb, Sheldon I. Arenberg, and Raj Singh. *Crime Analysis: From First Report To Final Arrest*. Montclair, CA: Alpha Publishing., 1994.

[38] Vishal Gupta and Gurpreet S. Lehal. A survey of text mining techniques and applications. *Journal of Emerging Technology in Web Intelligence*, 1:60–76, 2009.

[39] Dursun Delen and Martin D. Crossland. Seeding the survey and analysis of research literature with text mining. *Expert Syst. Appl.*, 34:1707–1720, 2008.

[40] Li-Ping Jing, Hou-Kuan Huang, and Hong-Bo She. Improved feature selection approach tfidf in text mining. In *Proceedings of the First International Conference on Machine Learning Cybernetics,Beijing*, pages 944–946, 2002.

[41] Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. Text mining techniques for patent analysis. *Inf. Process. Manage.*, 43:1216–1247, September 2007.

[42] Jochen Dörre, Peter Gerstl, and Roland Seiffert. Text mining: finding nuggets in mountains of textual data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 398–401, New York, NY, USA, 1999. ACM.

[43] Vidhya. K. A and G. Aghila. Text mining process, techniques and tools : an overview. *International Journal of Information Technology and Knowledge Management*, 2:613–622, 2010.

[44] Weiguo Fan, Linda Wallaceand Stephanie Rich, and Zhongju Zhang. Tapping into the power of text mining. In *Communications of ACM*, volume 49, pages 76–82, 2006.

[45] Radha Shakarmani, Nikhil Kedar, and Khandelwal. Performance assessment using text mining. *International Journal of Computer Applications*, 1(12):1–5, 2010.

[46] Junghoo Cho and Hector Garcia-Molina. Parallel crawlers. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

[47] Mauro Tortonesi. Gnu wget 1.10. http://lists.gnu.org/archive/html/info-gnu/2005-06/msg00003.html, 2005.

[48] B. Mathiak and S Eckstein. Five steps to text mining in biomedical literature. In *Proc. of the 2 European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa. Italy.*, pages 47–50, 2004.

[49] S. Iiritano and M. Ruffolo. Managing the knowledge contained in electronic documents: A clustering method for text mining. In *Proceedings of the 12th International Workshop on Database and Expert Systems Applications*, DEXA '01, pages 454–458, Washington, DC, USA, 2001. IEEE Computer Society.

[50] J. L. Neto, A. D. Santos, C. A.A. Kastner, and A. A. Freitas. Document clustering and text summarization. In *Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD)*, pages 41–55, 2000.

[51] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11:586–600, 2000.

[52] Michal Jan Skiba. Text preprocessing in programmable logic. Master's thesis, Electrical and Computer Engineering, Waterloo, Ontario, Canada, 2010.

[53] Andreas Hotho and Andreas Nurnberger. A brief survey of text mining. *Journal for Language Technology and Computational Linguistics (JLCL)*, 20(1):19–62, 2005.

[54] Antoine Blanchard. Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information*, 29(4):308–316, 2007.

[55] A. Selamat and H.H. Ismail. Finding english and translated arabic documents similarities using ghsom. In *International Conference on Computer and Communication Engineering (ICCCE)*, pages 460 –465, may 2008.

[56] Motaz K. Saad and Wesam Ashour. Arabic text classification using decision trees. In *Proceedings of the 12th international workshop on computer science and information technologies (CSIT), Moscow , Saint-Petersburg, Russia*, volume 2, pages 75–79, 2010.

[57] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy, and Aly Fahmy. Integrated machine learning techniques for arabic named entity recognition. *International Journal of Computer Science Issues (IJCSI)*, 7(3):27–36, July 2010.

[58] M. F. Porter. An algorithm for suffix stripping. *Program*, 4:130–137, 1980.

[59] Jolie Beth Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[60] Tim Buckwalter. *Buckwalter Arabic Morphological Analyzer Version 1.0*. Linguistic Data Consortium, Philadelphia, 2002.

[61] Shereen Khoja and Roger Garside. Stemming arabic text. In *Computer Science Department, Lancaster University, Lancaster, UK*, 1999.

[62] P. Majumder, M. Mitra, and B. B. Chaudhuri. N-gram: a language independent approach to ir and nlp. In *Proceedings of the International Conference on Universal Knowledge and Language (ICUKL),Goa, India*, 2002.

[63] Faouzi Mhamdi, Ricco Rakotomalala, and Mourad Elloumi. A hierarchical n-grams extraction approach for classification problem. In *SITIS*, pages 211–222. 2006.

[64] U. Y. Nahm and R. J. Mooney. Using information extraction to aid the discovery of prediction rules from texts. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Text Mining, Boston,MA*, page 5158, 2000.

[65] Tomovic Andrija, Janicic, P, and Keselj V. N-gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedicine*, 81(2):137–153, 2006.

[66] A. Amine, Z. Elberrichi, M. Simonet, and M. Malki. Evaluation and comparison of concept based and n-grams based text clustering using som. *INFOCOMP Journal of Computer Science*, 7:27–35, 2008.

[67] R. Freeman, Hujun Yin, and N.M. Allinson. Self-organising maps for tree view based hierarchical document clustering. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN).*, volume 2, pages 1906 –1911, 2002.

[68] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[69] L. Khreisat. Arabic text classification using n-gram frequency statistics a comparative study. In *Proceedings of the international conference on data mining (DMIN), Nevada, USA*, pages 78–82, 2006.

[70] Jian-Cheng Wu, Kevin C. Yeh, Thomas C. Chuang, Wen-Chi Shei, and Jason S. Chang. Totalrecall: A bilingual concordance for computer assisted translation and language learning. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 201–204, Sapporo, Japan, July 2003. Association for Computational Linguistics.

[71] Andrew Roberts, Latifa Al-Sulaiti, and Eric Atwell. aconcorde: Towards an open-source, extendable concordancer for arabic. *Corpora journal*, 1:39–57, 2006.

[72] Andrew Roberts, Dr Latifa Al-sulaiti, and Eric Atwell. aconcorde: towards a proper concordance for arabic, 2005.

[73] Yousif Almas and Khurshid Ahmad. Lolo: a system based on terminology for multilingual extraction. In *Proceedings of the Workshop on Information Extraction Beyond The Document (IEBeyondDoc)*, pages 56–65, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[74] John Lawler. Review of monoconc pro 2.0 concordancing software. LINGUIST List 11.1411, 2000.

[75] Mike Scott. Wordsmith tools 4.0. http://www.lexically.net, [Accessed: 2012].

[76] Lou Bernard. Bnc-baby and xaira. In *Proceedings of the Sixth Teaching and Langauge Corpora conference*, 2004.

[77] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. The sketch engine. proc euralex 2004, lorient, france; pp 105-116,. http://www.sketchengine.co.uk, 2004.

[78] Hsinchun Chen. Homeland security data mining using social network analysis. In *Proceedings of the 1st European Conference on Intelligence and Security Informatics (EuroISI)*, pages 4–4, Berlin, Heidelberg, 2008. Springer-Verlag.

[79] P. Cerrito and J. C. Cerrito. Data and text mining the electronic medical record to improve care and to lower costs. *SAS Institute.*

[80] D. T. Heinze, M. L. Morsch, and J. Holbrook. Mining free-text medical records. In *Proceeding of the American Medical Information Association (AMIA)*, 2001.

[81] J.J.G. Adeva, N.L. Carroll, and R.A. Calvo. Applying plagiarism detection to engineering education. In *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 722 –731, 2006.

[82] Anders NØklestad. *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection.* PhD thesis, University of Oslo, 2009.

[83] Patrick Ruch, Laura Perret, and Jacques Savoy. Features combination for extracting gene functions from medline. In *Proceedings of ECIR*, pages 112–116, 2005.

[84] Ian H. Witten. *Text Mining.* Practical handbook of Internet computing, Boca Raton, FL: CRC Press, 2004.

[85] Andrew Borthwick. *A Maximum Entropy Approach to Named Entity Recognition.* PhD thesis, New York University, 1999.

[86] Jim Cowie and Wendy Lehnert. Information extractinon. *Communication of the ACM*, 39(1):81–91, 1996.

[87] David Nadeau. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision.* PhD thesis, University of Ottawa, 2007.

[88] Ionas Michailidis, Konstantinos Diamantaras, Spiros Vasileiadis, and Yannick Frre. Greek named entity recognition using support vector machines, maximum entropy and onetime. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 47–52, 2006.

[89] Zornitsa Kozareva. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL)*, pages 15–21, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[90] Luke Nezda, Andrew Hickl, John Lehmann, and Sarmad Fayyaz. What in the world is a shahab? wide coverage named entity recognition in arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Genoa, Italy*, 2006.

[91] Hamish Cunningham. *Information Extraction, Automatic.* Encyclopedia of Language and Linguistics, 2nd edition, 2005.

[92] Robert Gaizauskas. An information extraction perspective on text mining: Tasks, technologies and prototype applications. http://www.itri.brighton.ac.uk/projects/euromap/Textizauskas.pdf.

[93] Antonio Toral. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of the EACL Workshop on New Text*, 2006.

[94] Ellen Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the eleventh national conference on Artificial intelligence (AAAI)*, pages 811–816, 1993.

[95] Jim Cowie and Yorick Wilks. Information extraction. In *Communications of the ACM*, 1996.

[96] Claire Nedellec. Machine learning applied to information extraction in specific domains - an example, gene interaction extraction from bibliography in genomics. 2002.

[97] Satoshi Sekine. Named entity: History and future. Technical report, Proteus Project Report, 2004.

[98] John Maloney and Michael Niv. Tagarab: A fast, accurate arabic name recognizer using high-precision morphological analysis. In *Proceedings of the Workshop on Computational Approaches to Semetic Languages, COLING-ACL98, University of Montreal*, pages 8–15, 1998.

[99] Yassine Benajiba, Mona Diab, and Paolo Rosso. Arabic named entity recog-

nition: A feature-driven study. *IEEE Transactions on Audio, Speech and Language Processing*, 17(5):926–934, July 2009.

[100] Saleem Abuleil. Extracting names from arabic text for question-answering systems. In *RIAO*, 2004.

[101] Slim Mesfar. Named entity recognition for arabic using syntactic grammars. In *Natural Language Processing and Information Systems*, volume 4592 of *Lecture Notes in Computer Science*, pages 305–316. Springer Berlin / Heidelberg, 2007.

[102] Khaled Shaalan and Hafsa Raza. Arabic named entity recognition from diverse text types. In *Proceedings of the 6th international conference on Advances in Natural Language Processing*, pages 440–451, Berlin, Heidelberg, 2008. Springer-Verlag.

[103] Ali Elsebai, Farid Meziane, and Fatma Zohra Belkredim. A rule based persons names arabic extraction system. In *Proceedings of the IBIMA,Cairo, Egypt*, volume 11, pages 53–59, 2009.

[104] Saleem Abuleil and Martha Evens. Events extraction and classification for arabic information retrieval systems. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 769–770, Washington, DC, USA, 2004. IEEE Computer Society.

[105] Saleem Abuleil. Using nlp techniques for tagging events in arabic text. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, volume 2, pages 440–443, Washington, DC, USA, 2007. IEEE Computer Society.

[106] Jakub Piskorski, , Hristo Tanev, Martin Atkinson, Erik Van, and Der Goot. Cluster-centric approach to news event extraction. In *Proceedings of the con-*

ference on New Trends in Multimedia and Network Information Systems, volume 18, pages 276–290. IOS Press, 2008.

[107] Hayssam Traboulsi. Arabic named entity extraction: A local grammar-based approach. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 139–143, 2009.

[108] Yassine Benajiba and Paolo Rosso. Arabic named entity recognition using conditional random fields. In *Proceedings of the Arabic Language and Local Languages Processing Workshop (LREC), Marrakech, Morocco*, 2008.

[109] Chih Hao Ku, Alicia Iriberri, and Gondy Leroy. Natural language processing and e-government: crime information extraction from heterogeneous data sources. In *Proceedings of the international conference on Digital government research*, pages 162–170. Digital Government Society of North America, 2008.

[110] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. Crime data mining: A general framework and some examples. *Computer*, 37:50–56, 2004.

[111] Thomas Schneider. Multilingual information processing: the aventinus project. In *First International Conference on Language Resources & Evaluation, Granada, Spain*, pages 775–779, 1998.

[112] K. Pastra, H. Saggion, and Y. Wilks. Intelligent indexing of crime scene photographs. *Intelligent Systems, IEEE*, 18(1):55 – 61, jan-feb 2003.

[113] Michael I. Jordan and Christopher M. Bishop. Neural networks. *ACM Comput. Surv.*, 28:73–75, March 1996.

[114] S. S. Cross, R. F. Harrison, and R. L. Kennedy. Introduction to neural networks. *Lancet*, 346:1075–1079, 1995.

[115] N. Malik. Artificial neural networks and their applications. *IEEE PES special publication*, page 6, 2005.

[116] I. A. Basheer and M. Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43(1):3 – 31, 2000.

[117] B. Kröse and P. van der Smagt. *An Introduction to Neural Networks*. University of Amsterdam, eighth edition, 1996.

[118] Chung-Hong Lee and Hsin-Chang Yang. Construction of supervised and unsupervised learning systems for multilingual text categorization. *Expert Syst. Appl.*, 36:2400–2410, March 2009.

[119] E. Come, L. Oukhellou, T. Denoeux, and P Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334 – 348, 2009.

[120] Callon M, Courtial J.P, Turner W.A., and Bauin S. From translations to problematic networks: an introduction to co-word analysis. *Social Science Information*, 22:191–235, 1983.

[121] Benjamin C. M. Fung, Martin Ester, and Simon Fraser. Hierarchical document clustering. In *The Encyclopedia of Data Warehousing and Mining, John Wang (ed.), Idea Group*, 2005.

[122] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.

[123] Benjamin C.M. Fung, Ke Wang, and Martin Ester. Hierarchical document clustering using frequent itemsets. In *Proceedings of SIAM International Conference on Data Mining(SDM)*, 2003.

[124] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31:264–323, September 1999.

[125] A. Naseri and A. R. Soroush. Combined use of unsupervised and supervised learning for daily peak load forecasting. In *Energy Conversion and Management*, pages 1302–1308, 2008.

[126] T. Senjyu, Y. Tamaki, and K. Uezato. Next day load curve forecasting using self organizing map. In *Proceedings of the International Conference on Power System Technology*, volume 2, pages 1113–1118, 2000.

[127] Su-Hsien Huang, Hao-Ren Ke, and Wei-Pang Yang. Structure clustering for chinese patent documents. *Expert Syst. Appl.*, 34:2290–2297, May 2008.

[128] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464–1480, 1990.

[129] Hayoung Oh. Attack classification based on data mining technique and its application for reliable medical sensor communication. *International Journal of Computer Science and Applications*, 6(3):20–32, 2009.

[130] A. Samecka-Cymerman, A. Stankiewicz, K. Kolon, and A.J. Kempers. Self-organizing feature map (neural networks) as a tool in classification of the relations between chemical composition of aquatic bryophytes and types of streambeds in the tatra national park in poland. *Chemosphere*, 67(5):954 – 960, 2007.

[131] T. Fu, f. Chung, V. Ng, and R. Luk. Pattern discovery from stock time series using self-organizing maps. In *The 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Temporal Mining.San Francisco. California*, pages 27–37, 2001.

[132] C. K. Chow and S. Y. Yuen. Signal self organizing map. In *Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA,*, pages 213–218, 2007.

[133] T. Honkela, T. Leinonen, K. Lonka, and A. Raike. Self organizing map and constructive learning. In *Proceedings of ICEUT, IFIP, Beijin*, pages 339–343, 2000.

[134] Chung-Hong Lee, Chih-Hong Wu, and Hsin-Chang Yang. Text mining of clinical records for cancer diagnosis. In *Proceedings of the Second International Conference on Innovative Computing, Informatio and Control(ICICIC)*, pages 172–, Washington, DC, USA, 2007. IEEE Computer Society.

[135] Do Phuc and Mai Xuan Hung. Using som based graph clustering for extracting main ideas from documents. In *Proceeding of the IEEE International Conference on Research, Innovation and Vision for the Future (RIVF)*, pages 209 –214, 2008.

[136] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojarvi, Vesa Paatero, and Antti Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11:574–585, 2000.

[137] Hsinchun Chen, Homa Atabakhsh, Tim Petersen, Jenny Schroeder, Ty Buetow, Luis Chaboya, Chris O&apos;Toole, Michael Chau, Tom Cushna, Dan Casey, and Zan Huang. Coplink: Visualization for crime analysis. In *Proceedings of The National Conference on Digital Government Research*, pages 1–6, 2003.

[138] Abdulsamad Al-Marghilani, Hussein Zedan, and Aladdin Ayesh. Text mining based on the self-organizing map method for arabic-english documents. In

Proceedings of the Nineteenth Midwest Artificial Intelligence and Cognitive Science Conference, 2008.

[139] Samuel Eyassu and Björn Gambäck. Classifying amharic news text using self-organizing maps. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Semitic '05, pages 71–78, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[140] Dino Isa, V. P. Kallimani, and Lam Hong Lee. Using the self organizing map for clustering of text documents. *Expert Syst. Appl.*, 36:9584–9591, July 2009.

[141] G. J. Tsekouras, P. B. Kotoulas, C. D. Tsirekis, E. N. Dialynas, and N. D. Hatziargyriou. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. *Electric Power Systems Research*, pages 1–17, 2008.

[142] Alriyadh. Crime reports. http://www.alriyadh.com/, [Accessed: 2008-2011].

[143] Sabq. Crime reports. http://sabq.org/sabq/user/portal.do, [Accessed: 2010].

[144] Okaz. Crime reports. http://www.okaz.com.sa, [Accessed: 2008-2010].

[145] Aljazirah. Crime reports. http://www.al-jazirah.com/, [Accessed: 2008-2010].

[146] Ahram. Crime reports. www.ahram.org.eg, [Accessed: 2010].

[147] Almessa. Crime reports. www.almessa.net.eg/, [Accessed: 2010].

[148] Alrai. Crime reports. http://www.alrai.com, [Accessed: 2010].

[149] Alraimedia. Crime reports. http://www.alraimedia.com/Alrai/, [Accessed: 2008].

[150] Alqabas. Crime reports. http://www.alqabas.com.kw/, [Accessed: 2008].

[151] Alamalyawm. Crime reports. http://www.alamalyawm.com/default2.aspx, [Accessed: 2010].

[152] Alwatan. Crime reports. http://alwatan.kuwait.tt/, [Accessed: 2010].

[153] Alseyassah. Crime reports. http://www.al-seyassah.com/, [Accessed: 2010].

[154] Echoroukonline. Crime reports, [Accessed: 2010].

[155] Albayan. Crime reports. http://www.albayan.ae/, [Accessed: 2008-2010].

[156] Raya. Crime reports. http://www.raya.com/, [Accessed: 2010].

[157] Alsharq. Crime reports. http://www.al-sharq.com/, [Accessed: 2010].

[158] Anne R. Diekema, Ozgur Yilmazel, and Elizabeth D. Liddy. Evaluation of restricted domain question-answering systems. In *Proceedings of the ACL Workshop on Question Answering in Restricted Domains*, 2004.

[159] L. Hirschman and N. Sager. Automatic information formatting of a medical sublanguage. In *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin: Walter de Gruyter, 1982.

[160] Hayssam N. Traboulsi. *Named Entity Recognition: A Local Grammar-based Approach*. PhD thesis, School of Electronics and Physical Sciences, University of Surrey, 2006.

[161] Jack Halpern. Word stress and vowel neutralization in modern standard arabic. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009. The MEDAR Consortium.

[162] Sherri Condon, Gregory A. Sanders, Dan Parvaz, Alan Rubenstein, Christy Doran, John Aberdeen, , and Beatrice Oshika. Normalization for automated

metrics: English and arabic speech translation. In *Proceedings of Machine Translation Summit XII, Ottawa, Canada*, 2009.

[163] Fernando Bação, Victor Lobo, and Marco Painho. Clustering census data: comparing the performance of self-organising maps and k-means algorithms. In *Proceedings of KDNet (European Knowledge Discovery Network of Excellence) Symposium: Knowledge-Based Services for the Public Sector, Workshop 2: Mining Official Data*, 2004.

[164] Massai.ahram. Crime reports. http://massai.ahram.org.eg/, [Accessed: 2010].

[165] Addustour. Crime reports. http://www.addustour.com/, [Accessed: 2010].

[166] Elkhabar. Crime reports. www.elkhabar.com, [Accessed: 2010].

[167] Alkhaleej. Crime reports. http://www.alkhaleej.ae/, [Accessed: 2008-2010].

[168] Abdulsamad Al-Marghilani. *Application of Self-Organizing Maps to Multilingual Text Mining (Arabic-English)*. PhD thesis, De Montfort University, 2008.

[169] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. Som toolbox for matlab 5. http://www.cis.hut.fi/projects/somtoolbox/, [Accesed: 15-08-2010].

[170] Esri. Arcgis explorer software, 2011.

[171] Robert Gaizauskas and Yorick Wilks. Information extraction: Beyond document retrieval. *Journal of Documentation*, 54:70–105, 1998.

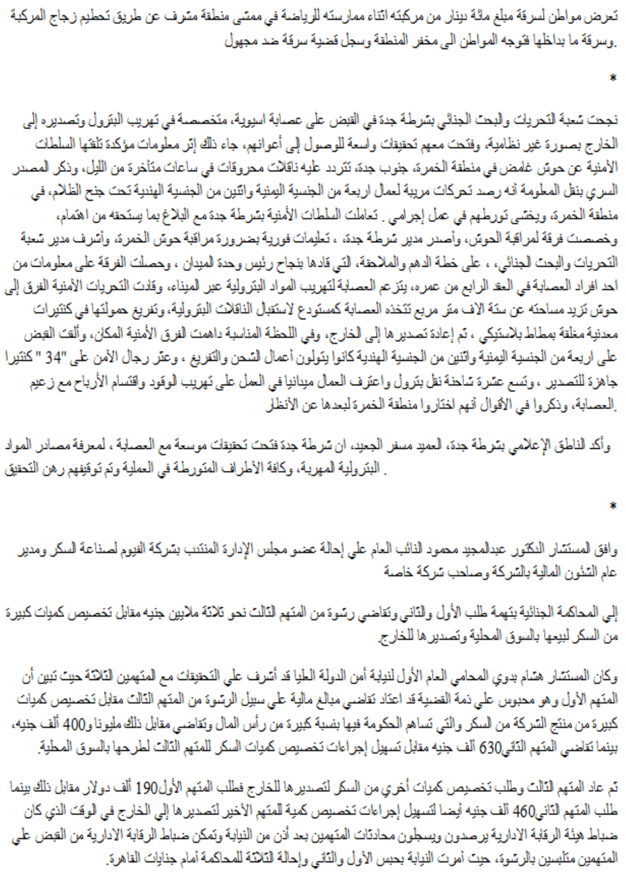
[172] Nancy Chinchor. Muc-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding*, MUC4 '92, pages 22–29, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[173] C. J. van RIJSBERGEN. *INFORMATION RETRIEVAL*. London: Butterworths, 2nd edition, 1979.

[174] Khaled Shaalan and Hafsa Raza. Person name entity recognition for arabic. In *Proceedings of the 5th Workshop on Important Unresolved Matters*, pages 17–24, 2007.

Appendix A

Sample of Corpus



تعرض مواطن لسرقة مبلغ مائة دينار من مركبته اثناء ممارسته للرياضة في ممشى منطقة مشرف عن طريق تحطيم زجاج المركبة وسرقة ما بداخلها فتوجه المواطن الى مخفر المنطقة وسجل قضية سرقة ضد مجهول.

*

نجحت شعبة التحريات والبحث الجنائي بشرطة جدة في القبض على عصابة اسيوية، متخصصة في تهريب البترول وتصديره إلى الخارج بصورة غير نظامية، وفتحت معهم تحقيقات واسعة للوصول إلى أعوانهم، جاء ذلك إثر معلومات مؤكدة تلقتها السلطات الأمنية عن حوش غامض في منطقة الخمرة، جنوب جدة، تتردد عليه ناقلات محروقات في ساعات متأخرة من الليل، وذكر المصدر السري بنقل المعلومة انه رصد تحركات مريبة لعمال اربعة من الجنسية اليمنية واثنين من الجنسية الهندية تحت جنح الظلام، في منطقة الخمرة، ويخشى تورطهم في عمل إجرامي . تعاملت السلطات الأمنية بشرطة جدة مع البلاغ بما يستحقه من اهتمام،
وخصصت فرقة لمراقبة الحوش، وأصدر مدير شرطة جدة، ، تعليمات فورية بضرورة مراقبة حوش الخمرة، وأشرف مدير شعبة التحريات والبحث الجنائي، ، على خطة الدهم والملاحقة، التي قادها بنجاح رئيس وحدة الميدان ، وحصلت الفرقة على معلومات من احد افراد العصابة في العقد الرابع من عمره، يتزعم العصابة لتهريب المواد البترولية عبر الميناء، وقادت التحريات الأمنية الفرق إلى حوش تزيد مساحته عن ستة الاف متر مربع تتخذه العصابة كمستودع لاستقبال الناقلات البترولية، وتفريغ حمولتها في كنتيرات معدنية مغلقة بمطاط بلاستيكي ، ثم إعادة تصديرها إلى الخارج، وفي اللحظة المناسبة داهمت الفرق الأمنية المكان، وألقت القبض على اربعة من الجنسية اليمنية واثنين من الجنسية الهندية كانوا يتولون أعمال الشحن والتفريغ ، وعثر رجال الأمن على "34 " كنتيرا جاهزة للتصدير ، وتسع عشرة شاحنة نقل بترول واعترف العمال ميدانيا في العمل على تهريب الوقود واقتسام الأرباح مع زعيم العصابة، وذكروا في الأقوال انهم اختاروا منطقة الخمرة لبعدها عن الأنظار.

وأكد الناطق الإعلامي بشرطة جدة، العميد مسفر الجعيد، ان شرطة جدة فتحت تحقيقات موسعة مع العصابة ، لمعرفة مصادر المواد البترولية المهربة، وكافة الأطراف المتورطة في العملية وتم توقيفهم رهن التحقيق .

*

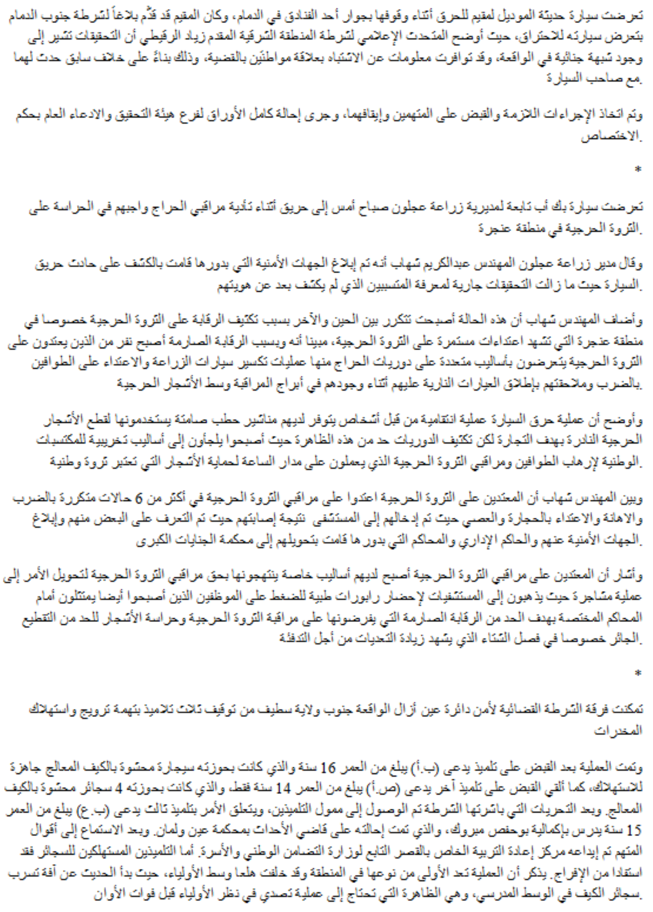
وافق المستشار الدكتور عبدالمجيد محمود النائب العام علي إحالة عضو مجلس الإدارة المنتدب بشركة الفيوم لصناعة السكر ومدير عام الشئون المالية بالشركة وصاحب شركة خاصة

إلي المحاكمة الجنائية بتهمة طلب الأول والثاني وتقاضي رشوة من المتهم الثالث نحو ثلاثة ملايين جنيه مقابل تخصيص كميات كبيرة من السكر لبيعها بالسوق المحلية وتصديرها للخارج.

وكان المستشار هشام بدوي المحامي العام الأول لنيابة أمن الدولة العليا قد أشرف علي التحقيقات مع المتهمين الثلاثة حيث تبين أن المتهم الأول وهو محبوس علي ذمة القضية قد اعتاد تقاضي مبالغ مالية علي سبيل الرشوة من المتهم الثالث مقابل تخصيص كميات كبيرة من منتج الشركة من السكر والتي تساهم الحكومة فيها بنسبة كبيرة من رأس المال وتقاضي مقابل ذلك مليونا و400 ألف جنيه، بينما تقاضي المتهم الثاني630 ألف جنيه مقابل تسهيل إجراءات تخصيص كميات السكر للمتهم الثالث لطرحها بالسوق المحلية.

ثم عاد المتهم الثالث وطلب تخصيص كميات أخري من السكر لتصديرها للخارج فطلب المتهم الأول190 ألف دولار مقابل ذلك بينما طلب المتهم الثاني460 ألف جنيه أيضا لتسهيل إجراءات تخصيص كمية للمتهم الأخير لتصديرها إلي الخارج في الوقت الذي كان ضباط هيئة الرقابة الادارية يرصدون ويسجلون محادثات المتهمين بعد أذن من النيابة وتمكن ضباط الرقابة الادارية من القبض علي المتهمين متلبسين بالرشوة، حيث أمرت النيابة بحبس الأول والثاني وإحالة الثلاثة للمحاكمة أمام جنايات القاهرة.

Figure A.1: Sample of Corpus



تعرضت سيارة حديثة الموديل لمقيم للحرق أثناء وقوفها بجوار أحد الفنادق في الدمام، وكان المقيم قد قدّم بلاغاً لشرطة جنوب الدمام بتعرض سيارته للاحتراق، حيث أوضح المتحدث الإعلامي لشرطة المنطقة الشرقية المقدم زياد الرقيطي أن التحقيقات تشير إلى وجود شبهة جنائية في الواقعة، وقد توافرت معلومات عن الاشتباه بعلاقة مواطنَين بالقضية، وذلك بناءً على خلاف سابق حدث لهما مع صاحب السيارة.

وتم اتخاذ الإجراءات اللازمة والقبض على المتهمين وإيقافهما، وجرى إحالة كامل الأوراق لفرع هيئة التحقيق والادعاء العام بحكم الاختصاص.

*

تعرضت سيارة بك أب تابعة لمديرية زراعة عجلون صباح أمس إلى حريق أثناء تأدية مراقبي الحراج واجبهم في الحراسة على الثروة الحرجية في منطقة عنجرة.

وقال مدير زراعة عجلون المهندس عبدالكريم شهاب أنه تم إبلاغ الجهات الأمنية التي بدورها قامت بالكشف على حادث حريق السيارة حيث ما زالت التحقيقات جارية لمعرفة المتسببين الذي لم يكشف بعد عن هويتهم.

وأضاف المهندس شهاب أن هذه الحالة أصبحت تتكرر بين الحين والآخر بسبب تكثيف الرقابة على الثروة الحرجية خصوصا في منطقة عنجرة التي تشهد اعتداءات مستمرة على الثروة الحرجية، مبينا أنه وبسبب الرقابة الصارمة أصبح نفر من الذين يعتدون على الثروة الحرجية يتعرضون بأساليب متعددة على دوريات الحراج منها عمليات تكسير سيارات الزراعة والاعتداء على الطوافين بالضرب وملاحقتهم بإطلاق العيارات النارية عليهم أثناء وجودهم في أبراج المراقبة وسط الأشجار الحرجية.

وأوضح أن عملية حرق السيارة عملية انتقامية من قبل أشخاص يتوفر لديهم مناشير حطب صامتة يستخدمونها لقطع الأشجار الحرجية النادرة بهدف التجارة لكن تكثيف الدوريات حد من هذه الظاهرة حيث أصبحوا يلجأون إلى أساليب تخريبية للمكتسبات الوطنية لإرهاب الطوافين ومراقبي الثروة الحرجية الذي يعملون على مدار الساعة لحماية الأشجار التي تعتبر ثروة وطنية.

وبين المهندس شهاب أن المعتدين على الثروة الحرجية اعتدوا على مراقبي الثروة الحرجية في أكثر من 6 حالات متكررة بالضرب والاهانة والاعتداء بالحجارة والعصي حيث تم إدخالهم إلى المستشفى نتيجة إصابتهم حيث تم التعرف على البعض منهم وإبلاغ الجهات الأمنية عنهم والحاكم الإداري والمحاكم التي بدورها قامت بتحويلهم إلى محكمة الجنايات الكبرى.

وأشار أن المعتدين على مراقبي الثروة الحرجية أصبح لديهم أساليب خاصة ينتهجونها بحق مراقبي الثروة الحرجية لتحويل الأمر إلى عملية مشاجرة حيث يذهبون إلى المستشفيات لإحضار رابورات طبية للضغط على الموظفين الذين أصبحوا أيضا يمثلون أمام المحاكم المختصة بهدف الحد من الرقابة الصارمة التي يفرضونها على مراقبة الثروة الحرجية وحراسة الأشجار للحد من التقطيع الجائر خصوصا في فصل الشتاء الذي يشهد زيادة التحديات من أجل التدفئة.

*

تمكنت فرقة الشرطة القضائية لأمن دائرة عين أزال الواقعة جنوب ولاية سطيف من توقيف ثلاث تلاميذ بتهمة ترويج واستهلاك المخدرات

وتمت العملية بعد القبض على تلميذ يدعى (ب.أ) يبلغ من العمر 16 سنة والذي كانت بحوزته سيجارة محشوة بالكيف المعالج جاهزة للاستهلاك، كما ألقي القبض على تلميذ آخر يدعى (ص.أ) يبلغ من العمر 14 سنة فقط والذي كانت بحوزته 4 سجائر محشوة بالكيف المعالج. وبعد التحريات التي باشرتها الشرطة تم الوصول إلى ممول التلميذين، ويتعلق الأمر بتلميذ ثالث يدعى (ب.ع) يبلغ من العمر 15 سنة يدرس بإكمالية بوحفص مبروك، والذي تمت إحالته على قاضي الأحداث بمحكمة عين ولمان. وبعد الاستماع إلى أقوال المتهم تم إيداعه مركز إعادة التربية الخاص بالقصر التابع لوزارة التضامن الوطني والأسرة. أما التلميذين المستهلكين للسجائر فقد استفادا من الإفراج. يذكر أن العملية تعد الأولى من نوعها في المنطقة وقد خلفت هلعا وسط الأولياء، حيث بدأ الحديث عن آفة تسرب سجائر الكيف في الوسط المدرسي، وهي الظاهرة التي تحتاج إلى عملية تصدي في نظر الأولياء قبل فوات الأوان.

Figure A.2: Sample of Corpus