Ministry of Higher Education and Scientific Research
Ahmed Draia University of ADRAR
Faculty of Science and Technology
Department of Mathematics and Computer Science

Master Thesis in Computer Science

Specialty: Intelligent Systems

Corpus construction for Arabic anaphora resolution

Prepared by

Dahou Abdelhalim Hafedh and Abdelmoazz Mohamed

President: Mr. KOHILI Mohamed
Examiner 1: Mr. CHOGUEUR Djilali
Examiner 2: Mr. MEDIANI Mohamed
Supervisor: Mr. CHERAGUI Mohamed Amine

Academic Year: 2018/2019

Thanks

It is customary to say that a dissertation is not the fruit of the sole work of its author, but the result of numerous and close collaborations; it does not deviate from the rule. We thank God above all for having given us the will to finish this thesis.

This work was born with a lot of help and encouragement from people around us. This short thank you will not be enough to reward their efforts but still ... At the end of two pleasant years in mathematics and computer science department of the University of ADRAR we would like to thank all teachers for their dedication, all our thoughts of gratitude goes to our MR. CHERAGUI Mohamed Amine, who always been there for the help and listening and being available throughout the realization of this thesis.

We would also like to thank the members of the jury who agreed to examine us. We send our sincerest thanks to all our colleagues and friends who share with us the good moments of joy during these two years.

A Huge thanks for our parents for helping us getting here. Hope we'll always make them proud.

We express our gratitude to all our loved ones who have always supported and encouraged us during the realization of this thesis. Finally, all those who have contributed, from near or far to the realization of this memory and which we cannot unfortunately quote, find here the expression of our deep gratitude.

ABSTRACT

Anaphora resolution is seen to be a very challenging and complex problem in the NLP for Arabic language. A majority of the NLP applications used for question answering, information extraction, and text summarization, need a proper resolution and identification of the anaphora. Despite the fact that several authors have published studies for anaphora resolution in many European languages, including English, very few studies have been published for anaphora resolution in the Arabic language and this due the lack of resources.

In our study, we have developed an anaphoric annotating tool for Arabic (A^3T) that can resolve the Arabic pronominal anaphora and verbal anaphora. We have determined the proper rules which can be used for this task. The different linguistic rules depend on the morphological, lexical, syntactic (Using MADAMIRA part of speech tagger) and statistical constraints. Our aim is to build a real corpus (A^3C) which will be used for anaphora resolution (i.e., either for system training or evaluation).

Key words: Arabic language, linguistic rules, pronominal anaphora, Verbal anaphora.

La résolution de l'Anaphore est considérée comme un problème très complexe et complexe dans la TALN pour la langue arabe. Une majorité des applications TALN utilisées pour la réponse aux questions, l'extraction d'informations et la synthèse de texte nécessitent une résolution et une identification appropriées de l'anaphore. Bien que plusieurs auteurs aient publié des études sur la résolution de l'anaphore dans de nombreuses langues européennes, y compris l'anglais, très peu d'études ont été publiées sur la résolution de l'anaphore en langue arabe, ce qui s'explique par le manque de ressources.

Dans notre étude, nous avons développé un outil d'annotation anaphorique pour l'arabe (A^3T) capable de résoudre l'anaphore pronominale arabe et l'anaphore verbale. Nous avons déterminé les règles appropriées pouvant être utilisées pour cette tâche. Les différentes règles linguistiques dépendent des contraintes morphologiques, lexicales, syntaxiques (utilisation de MADAMIRA de la parole) et statistiques. Notre objectif est de construire un corpus réel (A3C) qui sera utilisé pour la résolution de l'anaphore (c'est-à-dire, soit pour la formation du système, soit pour l'évaluation).

Mots-clés: langue arabe, règles linguistiques, anaphore pronominale, anaphore verbale.

TABLE OF CONTENT

Contents

LIST OF FIGURES & TABLES

List of figures

List of Tables

GENERAL INTRODUCTION

Natural Language Processing (NLP) is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation. For achieve the point of understanding, the NLP goes through different levels but also through the development of different modules that can contribute to understanding, one of the important modules is the anaphora resolution.

Anaphora resolution is one of the most frequent linguistic phenomena used in natural language processing that should be solved in order to establish the coherence and correctly understand the text. Anaphora resolution is a process of determining the antecedent of an anaphor and the subsequent replacement of the anaphor by its antecedent. The anaphora can be pronominal or even in some cases verbal that needs to be interpreted in relation to an element appearing before him or after in the speech. The few researches in this module is due the lack of corpora annotated with anaphoric relations especially for Arabic languages although it is very much needed in most NLP systems. The annotation task of anaphoric relations is very time consuming and requires a significant effort from the human annotator.

The work that we present in this memoir aims to develop an anaphor resolution system dedicated to Arabic language touching the pronominal and verbal aspect and generate a corpus that annotated with anaphoric relations. The motivation behind this work is to produce an annotated corpus to encourage works and enables future research in Arabic anaphora resolution

This thesis is organized into four (04) main chapters, as follows:

- ✓ The first chapter: is an overview about the natural language processing as well as the Arabic language.
- ✓ The second chapter: illustrate the anaphoric phenomenon, the different existing types and the approaches of resolution for the language Arab.
- ✓ The third chapter: presents the concept of corpus, the different techniques to build a corpus and Available Annotated Corpora on Arabic language for anaphora resolution.
- ✓ The fourth chapter: it presents the general architecture of the tool to develop "A^3C".
- ✓ The fifth chapter: focus on the presentation of our tool, as well as different tests and results obtained.

And finally, we end with a general conclusion, evoking new research perspectives.

CHAPTER I

Natural language processing

Principles and basic concept

1. Introduction

Natural language is any language that humans learn from their environment and use to communicate with each other. Whatever the form of the communication, natural languages are used to express our knowledge and emotions and to our responses to other people and to our surroundings.

Natural language processing (NLP) is a Corps of strategies applied to reveal and identify sentence boundaries, extract their grammatical structure, and detect the Language. Moreover, Entity extraction (organization, person, place, etc.) and meaning of the input to ameliorate or achieve a useful task (translation, summarization, classification, etc.). It based on a number of disciplines being linguistics, computer and information sciences, mathematics, electrical and electronic engineering, psychology, and robotics, etc. NLP applications cover a several of fields of studies, such as text processing (Tokenization, Normalization, Stemming, Lemmatization, etc.), summarization, text classification and categorization, Named Entity Recognition (NER), Question Answering, Speech Recognition.

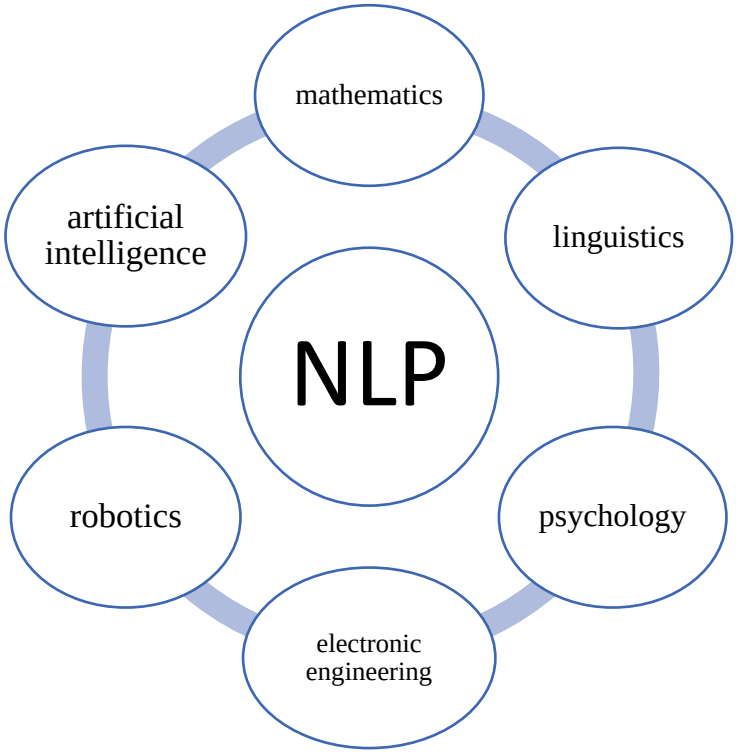


Figure 1: Domains of natural language processing.

This first chapter has been structured in the following sections: Section 2, the goals of NLP. Section 3, Brief history about NLP. Section 4, Different levels of treatments. Section 5, Application of NLP, we have highlighted the most common areas of application. Section 6, we talked about Arabic language, section 7, its Arabic versus NLP, section 8, we talk about the

objectives of Arabic NLP, section 9, it is the challenges of Arabic NLP. Finally, the last section presents the conclusion of the chapter.

2. Purpose of Natural language processing

The objective of NLP is to process the language like human language processing. It means that the computer must understand the input text or speech and manipulate it easily and faster like human [1]. It is based on linguistics, formalisms (representation of information and knowledge in machine interpretable formats) and computer science. To automatically process these data, it is first necessary to explain the rules of the language and then to represent them in operational and calculable formalisms and finally to implement them using computer programs [2].

In addition, the aim of NLP is to design of software or scripts capable of automatically processing linguistic data (written texts, oral dialogues) or the linguistic smaller units (for example: sentences, statements, groups of words or simply isolated words). Linguistic analysis software includes morpho-syntactic and parsers labeling software, which is the basis of most NLP applications (machine translation, speech processing, etc.).

3. History of Natural Language Processing

The goal of this section is to give some important dates in the development of the natural language processing since the late of 1940s [3] [4] [5].

Research in natural language processing has been going on for several decades dating back to the late 1940s. Machine translation (MT) was the first computer-based application related to natural language.

- The old MT projects was started by Weaver and Booth in 1946 specially for breaking enemy codes during World War II.
- In 1949 Weaver's suggested using ideas from cryptography and information theory for language translation. Research began at various research institutions in the United States within a few years.
- In 1957: Chomsky tries to construct a "formalized theory of linguistic structure". This method uses phrase structure rules that break down sentences into smaller parts. The work was published under the name of "Syntactic Structures".
- In 1958: John McCarthy released the programming language LISP: Locator/Identifier Separation Protocol, a language still commonly used today.

- Since 1960 there have been some significant developments, both in production of prototype systems and in theoretical issues.
- In 1962: First international conference on Machine translation of languages.
- In 1964: ELIZA was created and by only rearranging sentences and following some relatively simple grammar rules, impersonated a psychiatrist.
- In 1966: the ALPAC (Automatic Language Processing Advisory Committee) was the committee assigned to evaluate the progress of NLP research
- In 1966: ALPAC and the NRC halted research on machine translation because progress, after twelve years and twenty million dollars, had slowed and machine translation had become more expensive than manual human translation.
- 1970: SHRDLU was an early natural language understanding computer program, developed by Terry Winograd at MIT. That consisted of rearranging blocks, cones, or pyramids.
- In the late 1970's, attention shifted to semantic issues, discourse phenomena, and communicative goals and plans.
- 1971: birth of the system LUNAR which answers questions about samples of rocks brought back from the moon.
- In 1982: the concept of a Chabot was created, and the project Jabberwocky began. The purpose was to create an AI program that could simulate natural human chat.
- By the early 1990s, there was an increasing awareness of the limitations of isolated solutions to NLP problems and a general push towards applications that worked with language in a broad, real-world context.
- Since 1990s to the present times, NLP has swiftly grown. This amelioration caused by the advent of technologies such as: Internet, fast computers with increased memory, increased availability of large amounts of electronic text.

Figure 2: A brief history of Natural Language.

4. Treatment levels of a natural language

The most clarification method for showing what is actually done within a Natural Language Processing system is by means of the 'levels of language' approach. In this section we will

give a description of levels of language. They will be sequentially mentioned: phonology, lexical, morphology, syntactic, semantic and pragmatic.

4.1. Phonology level

This level deals with the interpretation of speech sounds within and across words. There are, in fact, three types of rules used in phonological analysis [6]:

- Phonetic rules: for sounds within words.
- Phonemic rules: for variations of pronunciation when words are spoken together.
- Prosodic rules: for fluctuation in stress and intonation across a sentence.

4.2. Lexical level

Linguistic analysis must go through a first phase of lexical analysis, which consists in testing the belonging of each word of the text to the vocabulary of the language [7]. This phase contains two sub-phases:

- Segmentation phase: for the purpose of breaking down the plain text into paragraphs and sentences.
- Lexical phase: proposes a multitude of word decompositions into a set of affix triads (prefix, affix, and suffix) and roots.

"ذهب", for example, can take the form of a noun " ذَّهب" (Gold) or a verb "ذَهَبَ " (Go) but its part-of-speech and lexical meaning can only be derived in context with other words used in the phrase/sentence.

4.3. Morphology level

Morphology is the beginning of analysis an input. Morphology study the word structures and its formation (Their flexion: case indications, genre, number, mode, time, etc. their derivation (proclitic, prefixes, base, suffixes, enclitics) and their compositions). Moreover, it is focusing on the analysis of the individual components of words. The most important unit in morphology named the morpheme[1]. The process of getting to the morpheme is done by eliminating all derivational forms and retain the root. The following example shows the input and output of this level. [8]

[1] Morpheme: the smallest meaningful unit in a language.

Input	*Output*
الكِتَابُ Book	كُتُب + اسم + مفرد Book + Noun + singular
مدن cities	جمع + اسم + مدينة city + Noun + plural
مَدْمُوج integrated	دَمَجَ + اسم + مفرد integrate + Verb + past simple

Table 1: Examples of morphological processing.

4.4. Syntactic level

Syntax involves applying the rules of the target language grammar, this task is to determine the role of each word in a sentence and organize this data into a structure that is more easily manipulated for further analysis. Semantics are the examination of the meaning of words and sentences [2].

- Grammar: a statement consists of a noun phrase, a verb phrase, and in some cases, a prepositional phrase.
- Parsing: is the process of converting a sentence into a tree that represents the sentence's syntactic structure.

4.5. Semantic level

Semantic analysis plays an important role in the study of natural language, which deals with studying the meaning of the word out of the context of the sentence or text. The semantic level look toward removing the ambiguities which remain after the syntactic treatment.

4.6. Pragmatic level

The aim of this type of treatment is to give meaning to the word in relation to the context in which it is used. In other terms, it is the analysis of the real meaning of a speech in a human language. This is accomplished by removing ambiguities that cannot be eliminated by semantic processing.

5. Application of natural language processing

NLP application are in use today everywhere, we can face them in machines or especially in the websites and social media applications. In this following section, we will give you a brief description about the most common applications of natural language processing.

5.1. Information Extraction (IE)

More recent application area, IE focuses on the recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text. These extractions can then be utilized for a range of applications including question-answering, visualization, and data mining.

5.2. Information Retrieval (IR)

IR is about finding relevant resources to an information need from a humongous collection of resources. IR is useful today everywhere because the availability of large amounts of electronic information. IR is not limited to text; it is applicable to image and video search as well.

5.3. Question-Answering

In contrast to Information Retrieval, which provides a list of potentially relevant documents in response to a user's query, question-answering is concerned by building systems that answer questions asked by humans such as SIRI, chat box.

5.4. Summarization

The process by which the document is reduced to few sentences or phrases to produce a suitable summary. The idea of this application is to collect the most important information and create a subset of data from the whole data set. The summarization is not limited to text, it's can be applied to image (i.e. salient[2]) and video (the important events).

5.5. Machine Translation

It is a form of computational linguistics which use a technique that decode a text or speech from one language to another? The two most common approaches are rule based and statistical, they differ in the way of the input content processing [9].

- Rule based MT systems: works based upon the specification of rules for morphology, syntax, lexical selection and transfer and generation.
- The statistical approach works based up on the statistical models extracted from parallel aligned bilingual text corpora, which takes the assumption that every word in the target language is a translation of the source language words with some probability

[2] Salience: The part that your eyes are first drawn to in the visual

5.6. Sentiment Analysis

Mostly used on the web and social media monitoring, NLP is a great tool to comprehend and analyze the responses. It helps to analyses the attitude and emotional state of the writer (person commenting/engaging with posts). This application is also known as opinion mining. It is implemented through a combination of Natural Language Processing and statistics by assigning values to the text (positive, negative or neutral) and in turn making efforts to identify the underlying mood of the context (happy, sad, angry, annoyed, etc.)

6. The Arabic language

The Arabic language has more than 422 million speakers[3]. It's a language both interesting and challenging caused by its complex linguistic structure. Referred to his history as it is linked to Islam one of the world three monotheistic religions, so it is used by 1.4 billion Muslim in their prayers at least 5 times daily. Also have a strategic importance because his native speakers living in an important region with huge oil reserves critical to the world economy.

Arabic is written semi-cursively (writing the letters are connected to each other) from right to left, using an alphabet of 28 letters. Most Arabic words are extracted from a 3-character root by adding or penetrating letters, which generates new words using schemes. This makes it difficult to control in the field of automatic language processing [10].

Each form of an Arabic writing can correspond to a sequence of one or more prefixes, a root and one or more suffixes. Roots themselves are inflected and derived forms from lemmas.

7. Arabic Natural Language Processing (ANLP)

Over the last few decades, Arabic natural language processing (ANLP) has gained increasing importance, and several state-of-the-art systems have been developed for a wide range of applications, including machine translation, information retrieval and extraction, speech synthesis and recognition…etc. These applications encountered several complex problems pertinent to the nature and structure of the Arabic language. Most ANLP systems developed in the Western world focus on tools to enable non-Arabic speakers get the meaning of Arabic texts. Because the need for such tools was urgent, they were developed using machine learning approaches. And it has as quality not requiring deep linguistic knowledge and fast and inexpensive. Developers of such tools had to deal with difficult issues as we going to talk about in section (9).

[3] https://www.internetworldstats.com/stats19.htm , (Accessed the 20/05/2019)

On the other hand, ANLP applications developed in the Arab World have different objectives and usually employ both rule-based and machine-learning approaches.

8. Objectives of ANLP

The following are some of the objectives of ANLP for the Arab World [10]:

- *Knowledge and technology transfer to the Arabic world:* Due to the use of English in the most recent publication, it's quite impossible for human translators to translate that huge amount of data to Arabic for the Arab readers. ANLP could help reduce the time and cost of those tasks.
- *Arabic language Modernization*: Translating new concepts and terminology into Arabic involves coinage, Arabization, and making use of lexical gaps in the Arabic language. This will positively affect the revitalization of the Arabic language and enable it to fulfill the essential needs for its speakers.
- *Arabic linguistics improvement and modernization:* Arabic NLP needs a more formal and precise grammar of Arabic than the traditional grammar so widely employed today. Innovation is needed as well to preserve the valuable heritage of traditional Arab grammarians.
- *Possibility of information retrieval, extraction, summarization, and translation for the Arab user:* The hope is to bridge the gap between peoples of the Arab world and their peers in more technically advanced countries. By making information available to Arabic speakers in their native language, Arabic NLP tools empower the present generation of educated Arabs.

9. ANLP challenges and solutions

In addition to classical challenges such as ambiguity, coordination, reference, anaphora and ellipsis, existing in Latin languages (Spanish, French, Italian, etc.), there are other problems specific to Arabic language and certain other Semitic languages, like the diglossia phenomenon, the problem agglutination and syntactic structure. [10]

9.1. Arabic diglossia

Diglossia is a phenomenon whereby two or more varieties of the same language exist side-by-side in the same speech community. Each is used for a specific purpose and in a distinct situation. Using the wrong variety in a situation is usually ridiculed. [11]

A single ANLP application can't process data from all Arabic varieties due to the unique characteristics of each verity even though they share some properties. An ANLP application must aim a single variety so it would have a good understanding of the linguistic properties of that variety to insure some good results.

- **Solution**: The most important solution for Arabic Diglossia is building corpora for Arabic (build resources for the various varieties of Arabic). ANLP researchers and developers should focus on one variety at a time in their applications since it is hard to build a system that can handle all the varieties of Arabic simultaneously. Developers must be clear as to which variety of Arabic is appropriate for their specific applications.

9.2. Arabic script

The Arabic script is a challenge on itself to the automatic processing. Arabic is far from being an easy language to read, due to the lack of letters dedicated to vowels (including a set of orthographic symbols called diactrics), the letters changing shape according to their position in the word, the absence of capitalization and minimal punctuation. Another challenging thing in Arabic script is Some Arabic letters share the same shape and are only differentiated by adding certain marks such as a dot, a hamza or a madda placed above or below the letter.

- **Solution**: In this case tokenization is the solution, but due to the complexity and rich morphology of Arabic, tokenization and morphological analysis must be combined in one process. For the case of the absence of capitalization, it can be compensated by a deeper look into the language to detect regularities that could help information extraction, recognizing patterns of Arabic names, dates, addresses, etc., can improve recall of Arabic entity recognition.

Example: For example, the letter (ع) "'ain" has an initial shape (عـ), a median shape (ـعـ), a final connecting shape (ـع), and a final non-connecting shape (ع).

9.3. Syntactic structure of Arabic

Arabic is a free word order language. While the primary word order in Classical Arabic and Modern Standard Arabic is verb-subject-object, it is also allowed to have subject-verb-object and object-verb-subject. All varieties of Arabic allow subject less sentences when the subject is recoverable and allow equational sentences without explicit use of the equivalent of verb "to be" in English. So, "I a student" meaning "I am a student" is perfectly grammatical in Arabic.

- **Solution**: Grammatical descriptions of Modern Standard Arabic are the solution to such problem. Such descriptions are very useful in the processing of contemporary Arabic although they were not written from a computational viewpoint.

Example: As in the case of a prepositional attachment as in "قابلت مدير البنك الجديد" which could mean "I met with the new bank manager" or "I met with the manager of the new bank" depending on the internal analysis of the noun phrase.

10. Conclusion

The Arabic language has its own characteristics that are different from Indo-European languages. These characteristics are in fact the major problems with the work done on the Arabic language in the field of NLP.

In this chapter we've talked about the NLP in general and the most common challenges and solution it has treating the Arabic language while including some Arabic most complex features.

In the next chapter we're going to discuss the Anaphora which is the main language phenomena treated in our work.

CHAPTER II

Anaphora resolution

1. Introduction

Since a very long time, the anaphora resolution is considered as a very challenging but important task for many of the NLP applications, like question-answer, text summarization, and the machine translations. The aim of Anaphora resolution is to detect the suitable noun phrase that precede a certain anaphora phenomenon and matched with it. It is a commonly used phenomenon in NLP and plays an important role in the simplification of the expressions and connecting the words in a context. Several methods and works proposed and tested by many authors in different languages, like English and gives acceptable results. In the other side, the works in Arabic language are still at the developing stage based on the few works existed. This lack due to the following facts: Arabic language resources are very limited (Corpus, POS, parsers), morphological ambiguity, Free Word Order, etc.

Therefore, the resolution of anaphora helps uncover the meaning and role of the anaphor by finding the proper antecedent and it also helps to fully and correctly understand the text.

The chapter has been structured in the following sections: Section 2, what's an Anaphora. Section 3, the difference between anaphora, cataphora and deixis. Section 4, the different anaphoric phenomena in the Arabic language. Section 5, Challenges in Arabic anaphora resolution. Section 6, Anaphoric Resolution Approaches (The Case of the Arabic Language), we have highlighted many of the recent studies about Arabic resolution. Finally, the last section presents the conclusions of our chapter.

2. 9

Anaphora is a Greek expression that means "carrying back". Anaphora is a linguistic relation between two textual entities which is defined when a textual entity (the anaphor) refers to another entity of the text which usually occurs before (the antecedent).

The aim of anaphora resolution is to illustrate the relation between the two expressions in the text, i.e. defining the conditions under which an expression refers to another one which usually occurs before it. An expressions or entity may be a pronoun, a verb, definite descriptions, a lexical modifier, a noun phrase, or a proper noun. The "pointing back" (reference) is called an anaphor and the entity to which it refers is its antecedent. The following sentence has four antecedents and four anaphors. The co-reference relationship is indicated by giving both the antecedent and its referent the same index and underlining both.

The anaphora phenomena appear in the following example below as the anaphor is colored in red and the antecedent in green:

الشجرة لازلت صامدة فى مكانـــها وأوراقـــها جافة .

The tree is still in place and its leaves are dry.

Figure 3: Anaphora example.

3. Difference between anaphora, cataphora and deixis

There are some ambiguities between anaphora, cataphora and deixis in the process of resolution. In this section we will give a brief description for each one and the mains differences between them.

3.1.Anaphora/ Cataphora

Cataphor is matching between two expressions where the second one contains the important information to define or interpret the preceding one. In generally cataphor and the antecedent belong to the same (intra-sentential). The main difference between the cataphor and anaphor is the placement (cataphor placed before the entity that refers to and the anaphor placed after).

هو الله الواحد القهار

He is the god, He is the One the Prevailing

Figure 4: Cataphora example.

3.2.Anaphora/deixis:

Deixis is a Linguistic phenomenon which indicates the interlocutors in input text, or referred to place or time (الأن ,هنا). deixis, on the other hand, is not necessarily related to anaphora.

تم الأن نقل المريض الى المستشفى على جناح السرعة

The patient has now been rushed to the hospital

Figure 5: Deixis example.

4. Different Arabic anaphora types

The diversity of anaphora phenomena creates a challenge in the process of resolution. The most researches concentrate on one type of anaphora named pronominal. In this section we inventory the different classes of anaphora with examples.

4.1. Pronominal anaphora

This type is the most common and treated among the other types due to the facility of identified because it's characterized by anaphoric pronouns. Those pronouns they haven't meaning or semantic without their antecedents. Moreover, Researches in this type found that there are some non-anaphoric pronoun like "أنا ""نحن ""انت".

Pronominal anaphora divides into demonstrative pronouns, personal pronouns and relative pronouns as shown in figure below.

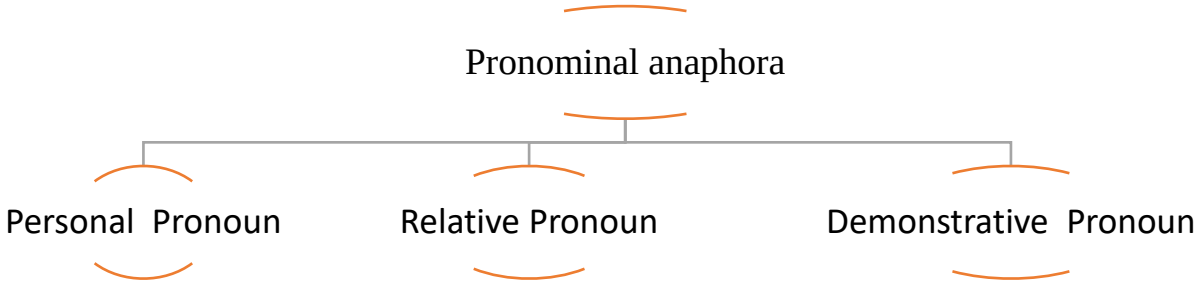


Figure 6: Pronominal anaphora types.

4.1.1. Demonstrative pronouns

Demonstrative pronouns: A demonstrative pronoun is a pronoun that is used to point to something specific within a sentence. Part of this pronouns are deictic (i.e. in the form of the 1st and 2nd person pronouns which indicates the interlocutors in input text, or referred to place or time هنا, الأن). Moreover, can be anaphoric but in generally are cataphoric not like other languages such French and English.

Demonstrative	Pronouns
Anaphoric	هذا , ذالك , هنالك , هذه , هاتان , هذان
Deictic	هنا , الأن

Table 2: Demonstrative types.

4.1.2. Relative pronouns

Relative pronouns are always anaphoric and generally indicating the previously existing noun phrase (الذي, التي, الذان, اللتان, اللاتي, الائي, اللتين, اللواتي, الذين, أولاء, من, ما).

4.1.3. Personal pronouns

Personal pronouns: Personal pronouns are pronouns that are associated primarily with a particular grammatical person we can find it isolated (هو) or with a suffix (كتابه). Generally, they are referential.

In Arabic language, 3rd personal pronouns classified in joint or disjoint pronouns and also in nominative, dative or accusative ones. Thus, we distinguish:

Personal pronouns	
Nominative disjoint personal pronouns	هو, هي, هما, هم, هن hun~a/humo/humaA/hiya/huwa
Accusative disjoint personal pronouns	اياه, اياها, اياهما, اياهم, اياهن Iiy~aAhumo/Iiy~aAhu/Iiy~aAhun~a/Iiy~aAhaA/Iiy~aAhumaA
Dative and accusative joint personal pronouns	ـه, ـها, ـهما, ـهم, ـهن hun~a/humo/humaA/haA/hu
Nominative joint personal pronouns	أ, و, ن Noon/waw/alef

Table 3: Personal Pronouns.

4.2. Verbal anaphora

Verb anaphora is another variety of anaphora which is characterized by the use of the verb (فعل did) as shown in example below. [12]

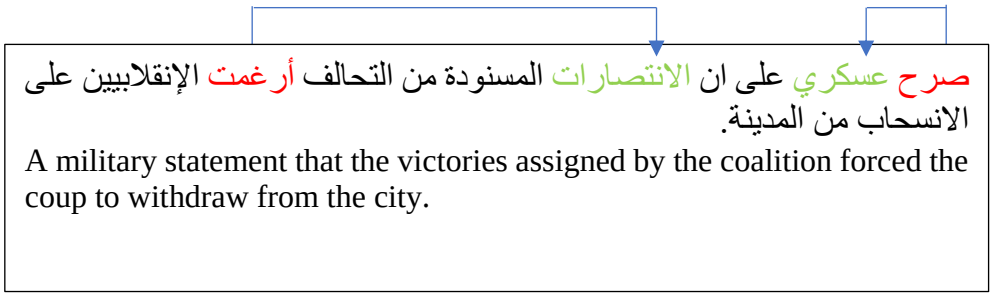


Figure 7: Verbal anaphora example.

4.3. Lexical anaphora

Lexical anaphora occurred when the referring statement give a description of proper names. The referring expression represent a semantically close concepts to the antecedent that referred to (e.g. synonyms, specialization) as shown in example below.

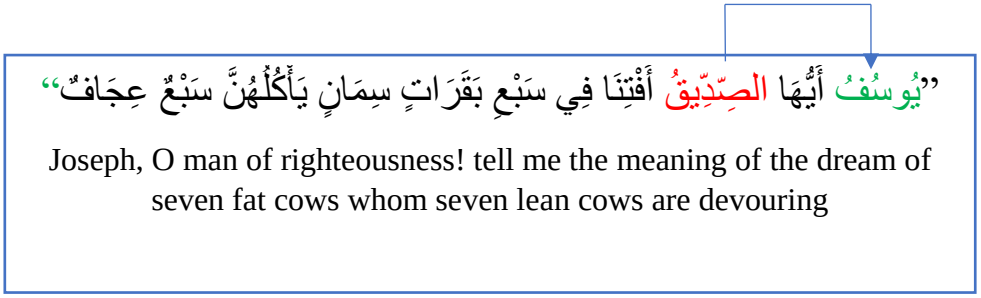


Figure 8: Lexical anaphora example. (Surat Yusaf, aya 46)

4.4. Comparative anaphora

The indication of comparative anaphora is when the anaphoric statement introduced comparative adjectives (أكبر, افضل) (bigger, better). This definition proves the existing of comparison, similarity and complement between the anaphor and the antecedent.

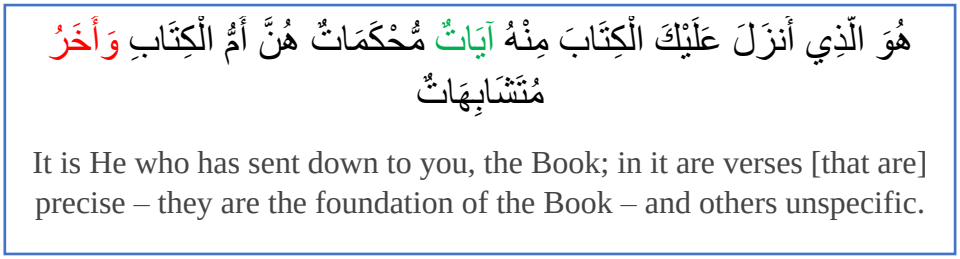


Figure 9: Comparative anaphora example.

The "one anaphora" is a subtype of comparative anaphora which found a lot of attention from many researchers.

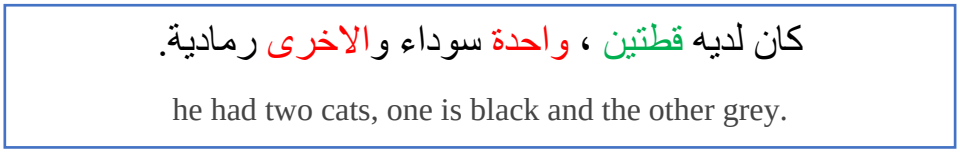


Figure 10: "One" anaphora example.

5. Challenges in Arabic anaphora resolution

Through our analysis and review we noticed that anaphora resolution can be quite tough compared to some other languages (like English) and that for some different reason, that's some of the main reasons we notice:

- First, the Arabic morphology that can be quite difficult for the analysis, like the anaphora could be attached or detached from a word. For example, in the word (يشبه/

look) the letter (ه /hu/) is a part of this word while in the word (لباسه/ His clothes) it represents a pronoun.

- Second, the complexity of sentence (in Arabic language can combine numbers of words in just one as in the given example "سنستدرجهم" (We will gradually lead them). This phenomenon is called agglutination and we have to break up the phrase and start the process of resolution to solve it.
- Third, the big lack of Arabic annotated corpus. That's cause of the big effort, time and complexity it would takes by the human annotator to annotate a big volume corpus.
- Fourth, the phenomenon of non-vowelized words in Arabic language can leads to an ambiguity in the resolution process, like the word "فهم" it can interpreted understand or they.

There are others, but they do depend on the nature of the texts, for example of the hidden antecedent which is a big problem found mostly in the Quran.

6. Anaphoric resolution approaches

In general, resolution systems deal with the following steps:

- **The pre-processing phase:** Part-of-Speech Tagging, NP chunker, Gender and Number filter, and identification of grammatical relation.
- **Anaphora Identification:** The goal of this phase is to identify all pronoun in the passage based to their grammatical parts of-speech and applied a filtering task to eliminate non-anaphora pronouns using a patterns or dictionary.
- **Identifying the search scope:** this process identifies all NPs that occurred before the anaphora and they are considered as probable candidates for antecedents. The construction of the search scope based on the number of sentences exist before the anaphora. After that, the candidates are compared to the anaphora based on number agreements, the gender and a search scope.
- **Extraction of features:** the aim of this module is to extract the properties of either anaphora, its antecedent candidate or their relationship to facility choosing the suitable antecedent.
- **Anaphora Resolving:** each candidate assigned a score value of every feature and so on. Therefore, all candidates have a final score value and the uppermost aggregate point score was suggested as the suitable antecedent. The one nearest is selected if there is any similar score.

Next, we will discuss the approaches to anaphora resolution which may be used in these steps. We can enumerate three main different approaches: rule-based, statistical and machine learning approaches.

6.1. Rule-based approach

Rule based approach focus on the creation of a heavy knowledge base by language experts and computer scientists. The efficiency of this approach is that it is easy to introduce human domain knowledge into the linguistic one. In the other side, this approach requires huge linguistic knowledge, very large number of patterns and there are exceptional cases which don't follow any rules. In addition, there are ambiguous cases where the sentence has more than one meaning and can follow more than one rule or can't follow anyone at all.

The approach initially developed and tested for Arabic language by Lamia Belguith [12]. The approach resolves pronoun anaphora with using linguistic and domain knowledge as show in example below. Instead, it makes use of corpus-based NLP techniques.

فـــهم يلعبون
They are playing

Figure 11: Example.

After the use of syntactic patterns, possible to determine that /هم / is an anaphoric pronoun because the word /يلعبون/ play is a verb (i.e. in Arabic we cannot find two consecutive verbs, except for action verbs " افعال الشروع".[12]

6.2. Statistical Approach

Statistical approach uses large set of correctly annotated corpora for training, and algorithm to analyses those corpora automatically to learn such rules from those trainable models. The most popular techniques used in this approach are MDP and bigrams. The advantage robust is especially about the statistical features like collocation patterns, frequency, repetition and recency. to give more accurate result to this approach is just adding more feature or increasing the training set.in the other side. To increase the accuracy of results for the rule-based approach, the only way is to increase the complexity of the rules, which is very difficult task.

6.3. Machine-learning approach

The machine learning approaches solve the pronoun through training and classification algorithm by using the characteristic (feature) vectors of anaphors and antecedents. This

approach resolves the anaphora identification without using linguistic or domain knowledge. Some researchers have introduced some novel features which were helpful to give a more precise anaphora resolution for Arabic pronouns like [13].

7. previous works in Arabic anaphora resolution:

A lot of works have been done in Arabic anaphora resolution, we did our research and we resumed all the works done we've came through in the following table.

Authors	Approach	Corpus	Success rate
Ruslan Mitkov Lamia Belghith	Rule-based	Technical manuals	95.2%
Khaled elghemry. Rania Al-Sabbagh. Najwa El-Zeiny.	Statistical approach	Web Corpus	87.6%
Souha Mezghani Lamia Belghith Abdelmajid Ben Hamadou	Rule-based	Technical manuals Newspaper articles Literary texts	/
Abdullatif Abolohom Nazlia Omar	Rule-based and machine learning approach	Quran annotated corpus	74.8%
Saoussen Mathlouthi Bouzid Fériel Ben Fraj Trabelsi Chiraz Ben Othmane Zribi	Probabilistic and dynamic approach (Hybrid)	Literary texts	86.23%
Abdullatif Abolohom Nazlia Omar	Rule-based approach (Scoring)	Quran Corpus (Qurpro)	84.48%
Souha Mezghani Hammami Lamia Hadrich Belghith	Machine learning approach	Newspaper Articles	84.5%
		Literary texts	72.1%
		Technical Manuals	86.2%
Feriel Ben Fraj Trabelsi Chiraz Ben Othmane Zribi	Probabilistic and dynamic approach (Hybrid)	Literary texts	81%

Table 4: Existing Arabic anaphora resolution systems.

8. Conclusion

In this chapter we discussed the anaphora phenomenon and its resolution and the biggest challenges we may face while doing it. The major motivation for this work is the lack of Arabic NLP works especially in anaphora resolution and corpora annotation. We conclude that the major challenge in this field and especially for the Arabic language is the low available resources. The next chapter will be about corpus, its purpose and its creation.

CHAPTER III

Corpora

1. Introduction

Corpus (in plural: corpora) is a large collection of linguistic data, compiled either as a transcription of recorded speech or as written texts, used to study all the aspects of language such as syntax, morphological, semantics, pragmatics, speech, and recently in lexicographic studies. [14]

The corpus is not allowed to be defined only formally, as a set of text or a sequence of alphanumeric characters. It verifies three types of conditions: meaning, conditions of acceptability, and exploitability conditions.

- **Significance conditions:** A corpus is constituted for a specific study (relevance),
- **Conditions of acceptability:** The corpus must provide a good representation (representativeness),
- **Operability conditions:** The texts that make up the corpus must be in homogeneity.

Each of these conditions needs to be commented, from the lighting complementary, and remarkably convergent.

The purpose of this chapter is to provide a description about corpora and reports briefly an overview of existing annotated resources, annotation schemes and annotating tools in Arabic language.

This chapter is structured in 5 sections. In section 2, corpus building. Section 3, How to build a corpus (or Corpora)? Section 4, Corpus annotation (case Anaphoric resolution). Section 5 presents the existing anaphoric annotated in Arabic language. Section 6 presents the conclusions.

2. Corpus building

The highlight on the field of building a corpus is due to the recent explosion in technology especially the massive production of computers and software. For the English language, there is the Brown Corpus of Standard American English [14], the British National Corpus (BNC) in 1995. In the other side, electronic Arabic texts did not exist early and still in advanced level caused by the lack of efficient tools such as tokenizers, taggers, morphological analyzers and optical character readers. We can mention the popular corpora that was transcribed from Al-Sharq Al-Awsat newspaper (around 40,000 words) in 1986 named the Buckwalter Arabic Corpus, and of course we can’t talk about Arabic corpora without mentioning the Arabic pen treebank, probably the largest corpus known in Arabic. Developed by the LDC (Linguistics

Data Consortium), the Arabic pen treebank is a corpus containing over 2 million words tree annotated with part-of-speech, morphology, gloss and syntax. [15]

The utility of building a corpus revolve around the following point:

a. Evaluating and ameliorating approaches.
b. Used for training and testing.
c. Help for extraction new features and rules.
d. Retrieving information about words by applying frequency and repetition methods.

Therefore, having a big amount of data stored on the computer provides a good resource to carry out this new view of language analysis.

3. How to build a corpus (or Corpora)?

Sometimes finding a corpus suitable for your work isn't an easy task due the lake of annotated corpora, in this case you may need to create and design your own corpus that would suit your work and research. In order to do that there is some factors you may need to take in consideration, including size, balance and representativeness and will be discussed below.

- Size: The size of the corpus depends very much on the type of questions that are going to be asked of it.
- Balance: sample texts up to a predetermined word limit.
- Representativeness: A corpus can be said to be representative if the findings from that corpus are generalizable to language or a particular aspect of language as a whole

Probably the easiest way of obtaining texts already in electronic format is to download them from the internet. Most web pages can be easily rendered as a text file and that what will be will discussed below.

3.1. Web as a source

Nowadays, we can consider the web as the most important source of information and the biggest one. It has the advantage of being the largest one, free and all time available. With just simple clicks we have access to a huge amount of text in almost all world languages. It may still be a bit difficult to find the right particular text we may need for our research due to the huge amount of data. But it is still considered as the greatest knowledge resource. The web is also an infinite source of resources (textual, graphical and sound) and This allow the fastest corpus constitution. [16]

As we said above, the construction of a corpus of texts from the web was not a simple task. Such constitution has contributed to developing and improving several linguistic tools such as question-answering systems, information extraction systems, machine translation systems, etc. In fact, one of the greatest characteristics of the web source is that it is multilingual. As for the current distribution of languages used on the Web, recent estimates of the top ten languages used in the web (Sep.17 2018) report that English and Chinese are the most used languages, followed by Spanish. The Arabic is the fourth on Internet users by language with more than 168.1 million users, followed by other major languages such as Portuguese, Malaysian, Japanese, Russian, French, German[4]. The process of building a corpus is a cyclical one, means as long as your work or research goes on, the corpus may change, some part may be removed or added depending in how it would help you through the research. You are going to adapt the corpus the more you learn by applying the knowledge you gathered. It is important to keep a detailed record of the data you collect, as information may seem to not be useful at the beginning but through the work it may be crucial atsome point as it can be used in a wider range of research. In the following figure we present the most important phases or keys to collect a corpus from the web.

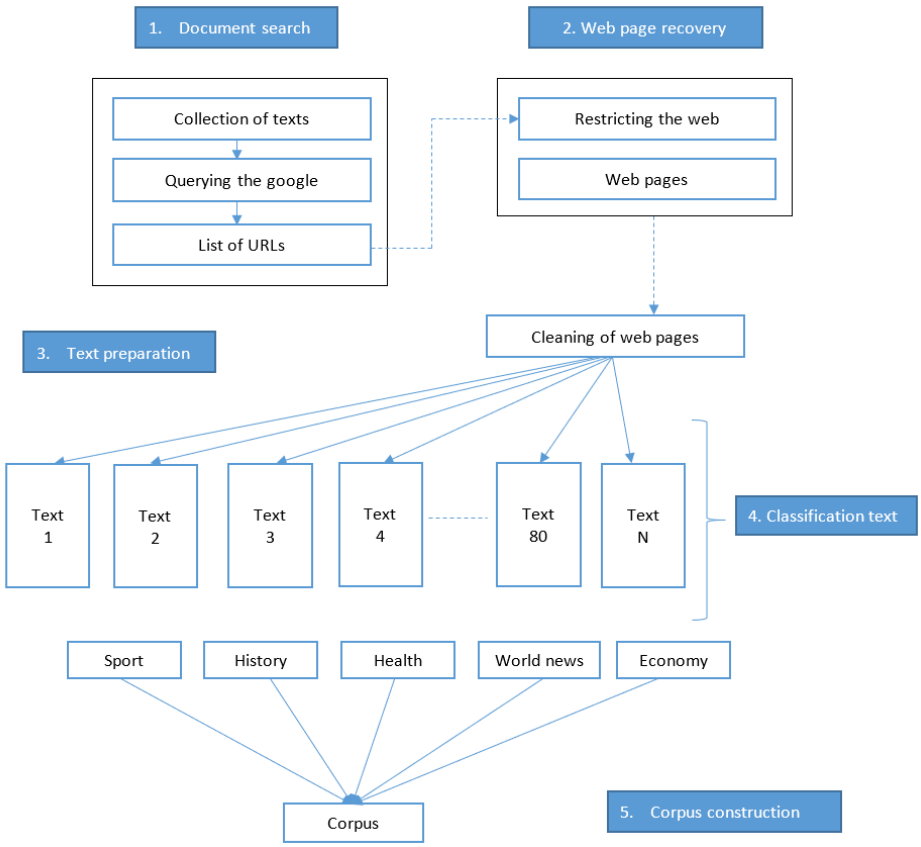


Figure 12: The steps of creating a web corpus.

[4] https://speakt.com/top-10-languages-used-internet/ (Accessed the 09/06/2019)

3.2. Corpus structure

The corpus structure may differ following the study and the aim of creating it. As a text corpus can consist of only one very long line of text, dividing it into smaller parts may be the better option as it makes it possible to include/exclude certain parts when searching. For instance, it allows the user to do a classification of the texts following its genre, type of texts… etc. as we may do some statistics about the corpus as the frequency of a word and know if a word is more limited to one specific topic.

To do this, a corpus must be equipped with marks or labels indicating the beginnings and ends of such parts. They are referred to as structure tags and the parts of a corpus they mark are called structures. The most typical parts are files, paragraphs and sentences.

It's up to the corpus author the choice of the kind of tags he wants to use. He may also include other different tags as POS tags, Gender and number tags… etc. As an example, for a corpus that contains lots of direct speech, it might be useful to mark the beginning and end of the direct speech so that searches and queries can be restricted only to direct speech. Marking the beginning and end of proverbs is definitely going to be useful for a research into their use. It is totally up to the corpus author to decide which structures will be included.

Each structure can, but does not have to, have additional labels giving more specific information about the structure. These are called meta-data or structure attributes. For example, the structure might carry information about the year of publication, the genre, the dialect, the style, author and source, simply anything that the author wants to include.

An example of a corpus consisting of two documents, each document has two paragraphs and paragraph has two sentences with information about the year of publication of the whole document, the style of each paragraph.

```
<doc pub="1970" lang="en">
<p style="informal">
<s> <pers gender="female">Rebecca</pers> has worked with a full range of clients
including <brand sect="automotive">BMW</brand> and <brand
sect="air">Airbus</brand>. </s>
</p>
</doc>
```

Figure 13: Example of corpus structure.

4. Some freely available Arabic corpora

The universe of Arabic corpora is huge touching a various domain, in this section we going to talk about some of the free available corpora in the internet through various categories. [16]

4.1. Raw text corpora

Raw text corpora do not include any kind of annotation, it contains only clean texts. In this type we can enumerate three big corpora. The first one is Adjir corpora which is a monolingual corpus with a word count of 113 million word. The second one is UN corpus which is a multilingual corpus containing over 2,7 million word. The last but not least, containing over 2 million words, the Arabic Multi Dialect Text corpora is, as it names suggest, a multi dialect corpus covering four major Arabic dialects (Egypt, Gulf, North Africa, Levantine).

4.2. Annotated corpora

Annotated corpora may be very beneficial for system training and evaluation depending on the research and the type of annotation. In this category we have the Qatar Arabic language bank containing 2 million word and it an error annotated corpus. Error annotated corpus are much useful for building spelling correction tools. We also have The Quranic Arabic corpus which is a corpus of the holy Quran annotated with both POS (part-of-speech) and syntax tagging.

4.3. Speech corpora

As its name suggest, speech corpora contain audio recording and transcribed data. To our knowledge, there is only one freely available corpus in this category which is Arabic Speech Corpora containing over 67 thousand of files which was compiled by Almeman and Lee.

5. Corpus annotation (case Anaphoric resolution)

The idea behind annotating resources is help the evaluation and the training systems or ameliorating approaches. But the annotation may be difficult due to the huge amount of texts and the anaphora diversity and its production may be challenging and very time-consuming which follows a specific annotation scheme. [14]

Several annotation schemes were proposed and tested for anaphoric and co-referential annotation task. The popular one is MUC scheme [17] that uses Standard Generalized markup

```
<COREF ID="100"> محمد </COREF> قال أنه <COREF ID="101"
TYPE="IDENT" REF="100"> هو </COREF>
```

Figure 14: SGML tagging example.

language (SGML) tags to annotate anaphoric expressions offering a standard format which several schemes are derived from it.

In the above example, the pronoun "هو" is tagged as referring to the entity, "محمد".

- The "**TYPE**" Attribute: indicate the relationship between the anaphor and the antecedent.
- The "**ID**" Attribute: The ID attribute is a unique number which identifies an entity.
- The "**REF**" Attribute: The REF uses that ID to indicate the co-reference link.

Extensible Markup Language (XML) [18] scheme is a simple, very flexible text format derived from SGML. Expressions which are either anaphoric or the antecedent of an anaphoric expression are annotated as <exp> elements. Every <exp> element has an attribute named "id", the value of which is of type ID (i.e. a unique identifier in the document).

```
<exp id="f17"> Baptiste</exp> vend <exp id="f18"> <ptr src="f17"/> sa
</exp>
```

Figure 15: XML-based scheme example.

The above example is in English because this scheme was applied for a French corpus, the pronoun " sa" (in English its) is referring to the phrase "Baptiste".

- The "**PTR**" Attribute: The link between an anaphor and its antecedent which is added to the anaphor.
- The " **SRC**" Attribute: identifying the antecedent of the current anaphor and may have several values separate with white spaces if the anaphoric expression has several antecedents

The Meta scheme [19] developed for the co-reference level and could be useful for different types of applications. The scheme consists of two schemes: a core scheme and an extended scheme.

- Core scheme: deals just with the annotation of co-reference relations similar to the MUC scheme.
- Extended scheme: gives the hand to annotate other kinds of anaphors such as bridging anaphors.in addition, anaphoric relations involving an extended range of anaphoric expressions (such as incorporated clitics) and of antecedents (as in discourse deixis)

In this scheme, each markable NP is annotated with a <de> element with an ID attribute.as shown in the following example:

```
<de ID="de_01"> we </de> are going to take <de ID="de_07">
the engine E3 </de> and shove <de ID="de_08"> it </de>
```

Figure 16: Core scheme example.

In the linking phase, its use the < link > elements that contain:

1. **HREF** pointer to the <de> element that stands in an anaphoric relation with an antecedent
2. **TYPE** attribute specifying the relation (which in the case of the Core Scheme can only be IDENT)

The <link> elements contain then one or more <anchor> elements, with a single <href> pointer to the antecedent.

```
<link href=" corerf.xml#id (de_07)" type =" ident ">
<anchor href="coref.xml#id(de_08)"/>
</link>
```

Figure 17: Linking phase example.

Note: the corerf.xml is the file that contain all markable NP with their annotation scheme.

In the case of Arabic language, as we know, there is not enough studies carried out in the field of anaphorical or co-referential corpus annotation. This lack is due to the challenge of developing an annotating tool. This latter caused by the absence of efficient tools such as tokenizers, taggers, morphological analyzers and optical character readers in the Arabic language. Many annotating tools have been developed in others language such as Callisto, MMAX2 and PALinkA.those tools were tested for the Arabic language and based on [12] they fail to make the annotation in case of joint pronouns. Therefore, Arabic annotated resource is really needed to encourage works on Arabic anaphora resolution.

6. Works on Arabic anaphoric annotated corpora

Some works have been suggested by different authors for building an Arabic corpora annotation with anaphoric links. In addition, the lack of resources leads researches to create an annotating tool or made it manually. As we knew that the majority of works doesn't cover all anaphora types in Arabic but still focusing on the pronominal with several techniques.

In this section we will give a brief description about existing works in the following table:

Work on annotated corpora	Scheme	Word segments	Pronouns	Text categories
Arabic corpora annotation [12]	XML-based	77.124 word	4300	Arabic newspapers Computer technical manuals
QurAna Corpus [20]	AQA, MUC-7	128.000 word	24679	Original Quranic text
Combining QAC and QurAna [21]	MUC-7	128.240 word	29.287	Original Quranic text

Table 5: available corpora for anaphora resolution in Arabic language.

6.1. Arabic Corpora Annotation with Co-referential Links

This corpus was annotated by (AnATAr) tool due to the absence of annotating software in the field of Arabic language previous these work as the author mentioned and this was his motivation to build his anaphoric annotating tool. This tool uses Rule-based approach for the automatic detection of Arabic pronouns exist in the corpus (technical manual, newspaper articles, texts of Tunisian books used for basic education and a novel) and let the user match every pronoun with the correct antecedent. For the facility and to accelerate the detection process of anaphoric pronouns it uses several patterns to solve the ambiguities (pleonastic pronoun, cataphoric pronoun, syntactic issues, etc.) and morphological analysis because the pronoun may occur as suffixes of nouns, verbs or prepositions. after the detection of pronouns, the user selects the correct antecedent of the selected anaphora and the tool automatically colored theme with the same color using SGML markup and the program added the scheme tags for the anaphora and the antecedent.

6.2. QurAna Corpus

Qurana is a large corpus that contain over 24.000 anaphoric pronouns tagged with their antecedents. It's can used to train, optimize, evaluate existing approach and for the resolution of anaphora problems. This corpus based on the annotation of the original Arabic Quran and was done by the first authors and took from him over one year [20]. In addition, we can find in their paper a list of the most frequent concepts in Quran (Allah, mankind, Prophet Mohammed, etc.). The second advantage of this work after the construction of the corpus, is that there is a form developed with PHP and MySQL allows the access to this corpus and applied different

queries like the search of pronoun in several verses with their antecedence or translated to other language and finally give us the ability for downloading this annotated corpus.

6.3. Arabic Anaphora Resolution Corpus of the Holy Quran

Well in this work, we might say that it's more like an amelioration of the QurAna corpus. It contains about 24.653 personal tagged with their antecedent and other anaphoric information like gender, number, distance between the couple anaphora/antecedent… etc. over all it contains information that can be used to implement a feature vector of a statistical anaphora resolution system. Plus, it is a description of a bank of sentences pattern containing 481 antecedent patterns. The idea behind this work was to provide a resource that would help in future researches in Arabic anaphora resolution and of course in analyzing the Quran script. And also, it could be used for training and testing anaphora resolution systems and evaluating them.

7. Conclusion

The annotated corpora proposed a large resource for researchers who are trying to create tools for Arabic resolution systems and evaluating the performance of their works. On the other hand, the available corpora in Arabic language can pay more attention to the data training and more complicated learning algorithms. The lack of resources on the anaphora annotation motivated us to build an annotated corpus that contain anaphora phenomena and their antecedent tagged. We propose a tool designed for text collection from the web (web as corpus) and corpus annotation by automatically detecting anaphora and the correct antecedent with a list of other suitable antecedent. This work will be presented and explained in detail on the next chapter.

CHAPTER IV

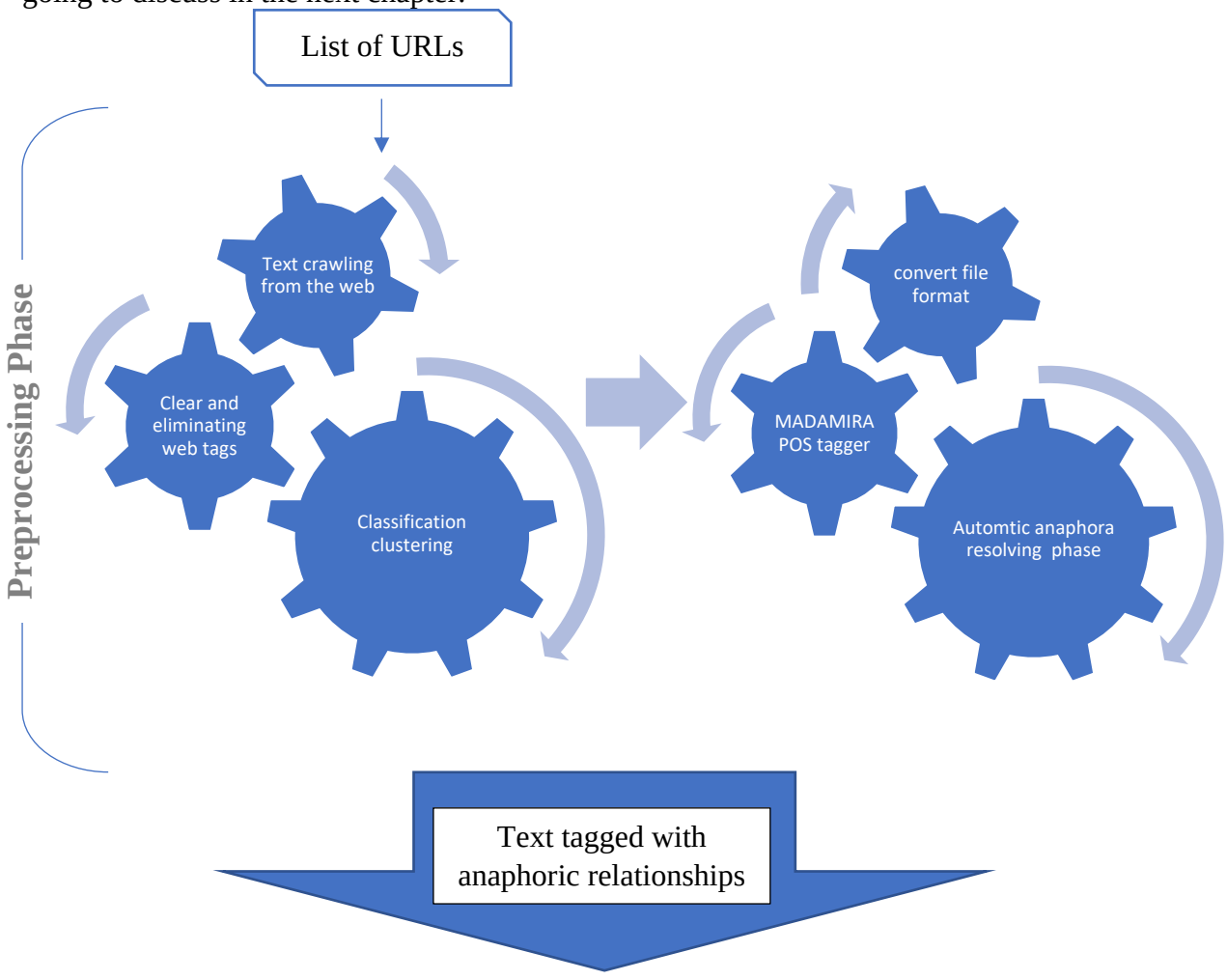
Conception and architecture of Arabic Anaphora Annotated Corpus (A^3C)

1. Introduction

After spending 3 chapters talking about the theory of NLP, Anaphora, and corpora. Here we come to the more interesting chapter, which is the conception of our system of resolution. We have done two things: first, we created a tool we called A^3T (Arabic Anaphora Annotating Tool) to help us doing the text cleaning, the linguistics tagging and automatic annotation, second and the most important part is the creation of our own corpus we named A^3C (Arabic Anaphora Annotated Corpus), a corpus annotated with anaphoric links, which we took from the web, cleaned and processed using our tool (A^3T), and then verified by an expert in linguistics. In this chapter, we going to discuss in detail the general structure of the environment, the modules we used, the different aspects of the tool and of course the process of the A^3C creation and annotation. We are going to start with the general architecture below.

2. General architecture of the system

We considered decomposing the realization of our annotation environment in three (03) phases, which are: Pre-processing phase and processing phase and last the evaluation phase which we going to discuss in the next chapter.



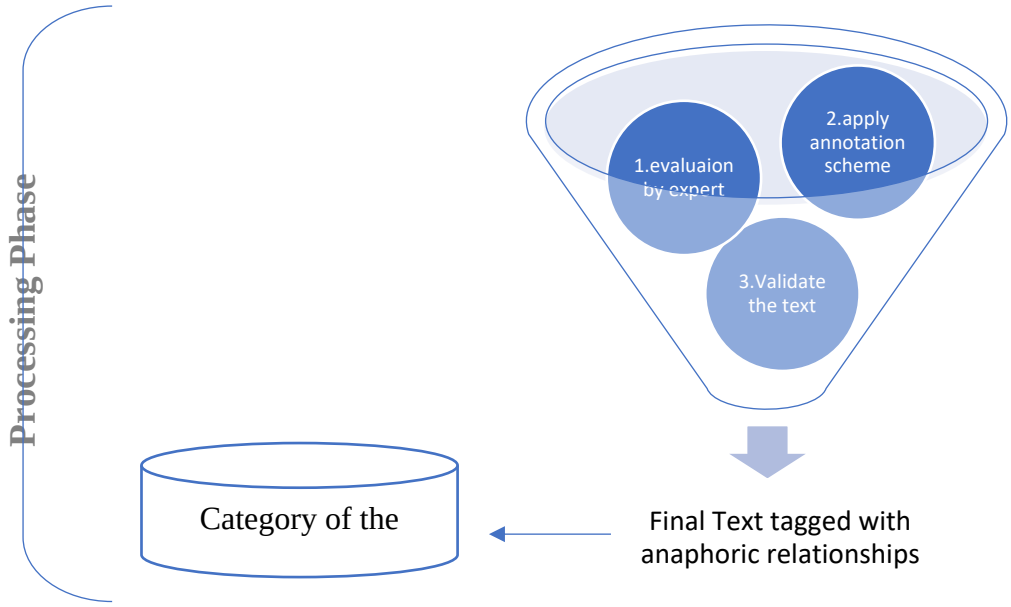


Figure 18: The system general architecture.

2.1. Pre-processing phase

We developed a system to resolve Arabic anaphora (pronominal and verbal anaphora). This system comprises of several modules to achieve this phase and be input to the next phase. These modules are crawling text from the web, NLP pre-processing task (tokenization, POS tagger), and automatic annotation module. The process goes as we can see in the next figure.

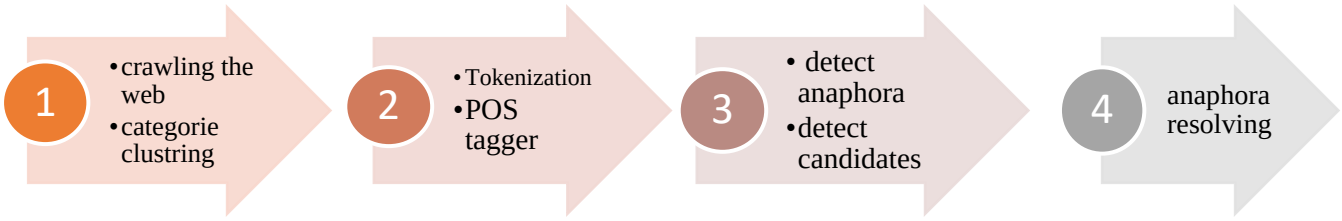


Figure 19: The general architecture of the pre-processing phase.

In pre-processing phase, we may divide it into two steps: construction of the raw corpus and the development of the anaphoric annotation tool A^{3}T which will be detailed below.

2.2.2. Construction of the raw corpus

Construction of a raw corpus contains these two modules:

- *Crawling text from the web module:* A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the pages and adds them to the list of URLs to visit, called the crawl frontier. On this module, we developed a script that requires us to enter the URL of each category and specify the limited number of pages, then it store all links of articles exists in the current pages into a list. Then

the script will acquire access to this links with searching for text tags (<p> or <h>) and store them into another text list. After that, apply the cleaning phase which consist of eliminating html tags from the text (we used a python package named BeautifulSoup or bs4) and save it into (*.txt) file format in the category folder.

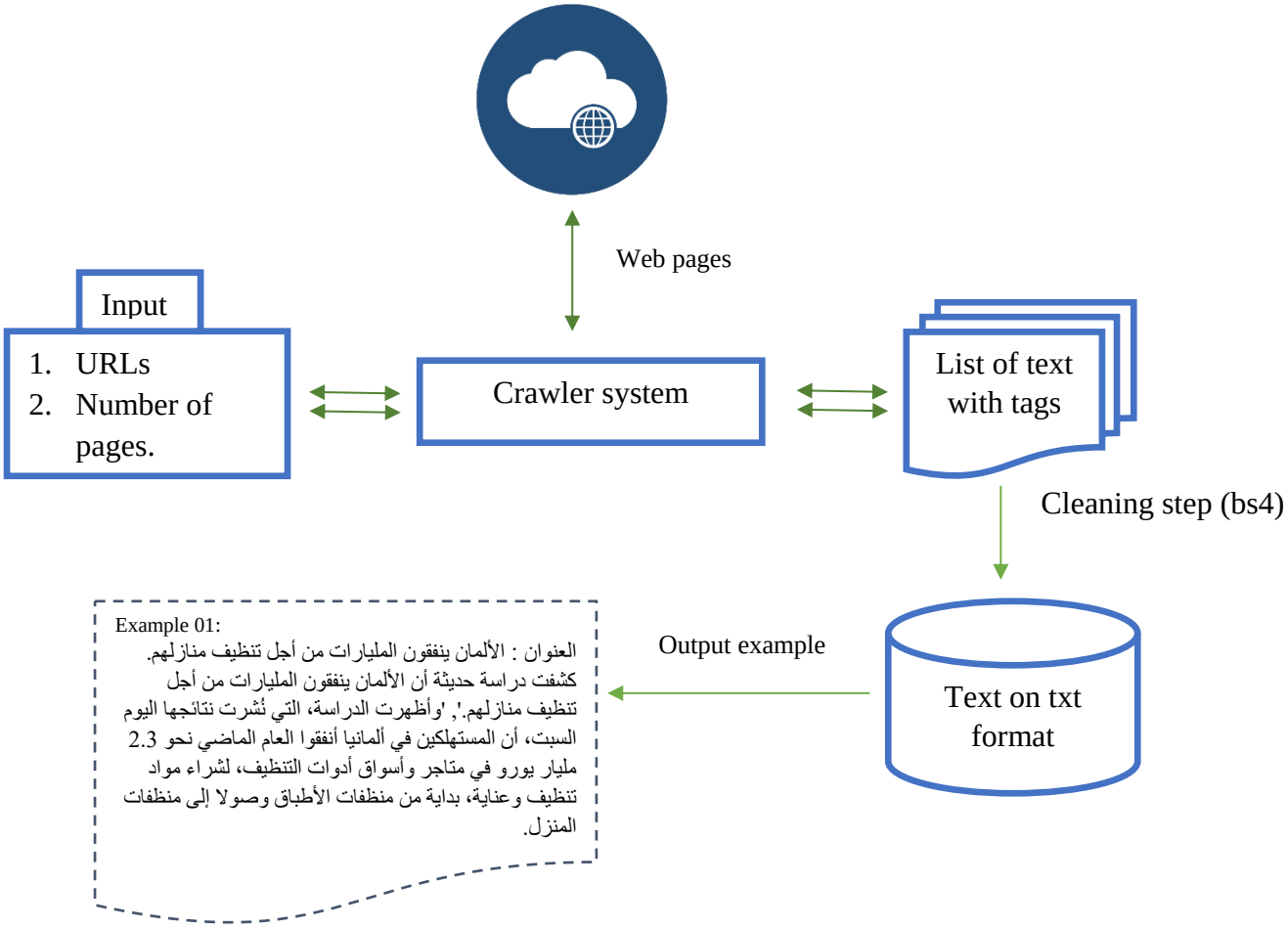


Figure 20: Representation of the crawling system.

- *Categories Classification module:* The corpus composed of articles in different fields such as: culture, politics, economy and the last category one religions. Articles are drawn from daily Arabic newspaper (Alriyadh newspaper[5]) with text size of 1.351.323 words which are going to be included in the corpus. Alriyadh newspaper is the first Arabic newspaper that was published in Riyadh (Saudi Arabia city), it was first published on the 11 May 1965 and it contains a various choice of text categories like religion sports and arts, etc. The choice of categories is due to the quantity of text and number of anaphora phenomena existing in it. We used Alriyadh newspaper because:
  a. It handles a big archive of information.

[1] http://www.alriyadh.com/ (Accessed the 12/01/2019)

b. Good structure of article.
c. Lot of articles existing in each page.
d. Diversity of categories.

2.2.2. Anaphoric Annotation tool

Anaphoric annotating tool for Arabic has the advantage of automatic detection of Arabic anaphora and the most suitable antecedent. Our aim is to build a tool that can help us to annotate a corpus which will be used for anaphora resolution specifically pronominal and verbal anaphora (i.e., either for system training or evaluation). The process of building the tool passed through the following modules:

- *Tokenization module:* Is the first step in text analytics the process of breaking down a text paragraph into smaller chunks such as words or sentence is called Tokenization. Token is a single entity that is building blocks for sentence or paragraph. The following figure shows the separators used in this phase:

‘,’ , ‘ ;’ , ‘ :’ , ‘.’ , ‘ ?’ , ‘ ! ‘: punctuation

Figure 21: Separators used in the tokenization.

In this operation each text file will be represented with sequence of sentences. The figure below shows an example of this operation.

Input:

إصابة 28 شخصًا إثر خروج ترام عن القضبان في البرتغال. أصيب 28 شخصًا بجروح في العاصمة البرتغالية لشبونة مساء أمس, إثر خروج ترام عن القضبان.

Output:

1. إصابة 28 شخصًا إثر خروج ترام عن القضبان في البرتغال. 2. أصيب 28 شخصًا بجروح في العاصمة البرتغالية لشبونة مساء أمس. 3. إثر خروج ترام عن القضبان.

Figure 22: Example of tokenization.

- *Convert from TXT to XML module:* The purpose of this phase is to simplify text and represented on MADAMIRA file to prepare it for the POS tag phase. XML I/O is the default method to passing data into MADAMIRA and receiving results. So, we developed a script that would convert a (*.txt) file to an (*.xml) file structured as the input of MADAMIRA. With its users wrap the data they wish MADAMIRA to process in an XML file (optionally including sections to override the default system configuration) and receive the results in another XML file.

```
<in_seg id="SENT1">  إصابة 28 شخصًا إثر خروج ترام عن القضبان في البرتغال  </in_seg>
<in_seg id="SENT2">  أصيب 28 شخصًا بجروح في العاصمة البرتغالية لشبونة مساء أمس،  </in_seg>
<in_seg id="SENT3">  إثر خروج ترام عن القضبان. </in_seg>
```

Figure 23: Example of TXT to XML Conversion.

- *POS tagger module:* The primary target of Part-of-Speech (POS) tagging is to identify the grammatical group of a given word. Whether it is a NOUN, PRONOUN, ADJECTIVE, VERB, ADVERBS, etc. based on the context. POS Tagging looks for relationships within the sentence and assigns a corresponding tag to the word.

  For our purposes, The MADAMIDA tagger was utilized. It is mentioned that its precision is 95.9%. Provides high-quality word-level disambiguation of Arabic text, and can conduct the following, valuable NLP tasks for Modern Standard Arabic or the Egyptian dialect words: Lemmatization, Discretization, Morphological Analysis, Base Phrase Chunking and Named Entity Recognition.[22]

```
<word id="6" word="خروج">
<morph_feature_set diac="خُرُوجِ" pos="noun" per="na" asp="na" vox="na" mod="na" gen="m" num="s" stt="c" cas="g"/>
</word>
```

Figure 24: MADAMIRA Output.

- *Anaphora identification module:* The identification of anaphora is carried out by referring to their grammatical parts of-speech. Based on MADAMIRA tag set. the identification is easier due to the good presentation and markup of anaphora in each sentence with important features used in the resolution phase such as: Gender, Number and distance… etc. The output of this phase is a list that contain all anaphora in the text using a specific ID (which is a number the MADAMIRA sets to each word to identify it in the text as we can take it as the word position in the text) and its features values we use in the resolution phase.

Features (Label)	Definition	Value
Aspect (asp)	Command	c
	Imperfective	i
	Perfective	p
	Not Applicable	na
Case (cas)	Nominative	n
	Accusative	a
	Genitive	g
	Not applicable	na
	Undefined	u
Gender (gen)	Feminine	f
	Masculine	m
	Not applicable	na
Mood (mod)	Indicative	i
	Jussive	j
	Subjunctive	s
	Not applicable	na
	Undefined	u
Number (num)	Singular	s
	Plural	p
	Dual	d
	Not applicable	na
	Undefined	u
Person (per)	1st	1
	2nd	2
	3rd	3
	Not applicable	na
State (stt)	Indefinite	i
	Definite	d
	Construct/Poss/idafa	c
	Not applicable	na
	Undefined	u
Voice (vox)	Active	a
	Passive	p
	Not applicable	na
	Undefined	u

Table 6: MADAMIRA tag-set

- Pronominal anaphora detection : as we going to see in the next figure, the process of pronominal anaphora detection differs from one type to another, for that the feature we need to extract aren't the same due to the MADAMIRA special tag set for different types specially relative pronouns cause it lakes the gender and number feature in MADAMIRA POS tagging. as the pos tagging for pronominal attached anaphora doesn't have a tag for gender, number and person, but it output them as an attached form (i.e, form1="+POSS_PRON_3MS") we had to create a script that would split the features off and put them each in their proper tag.

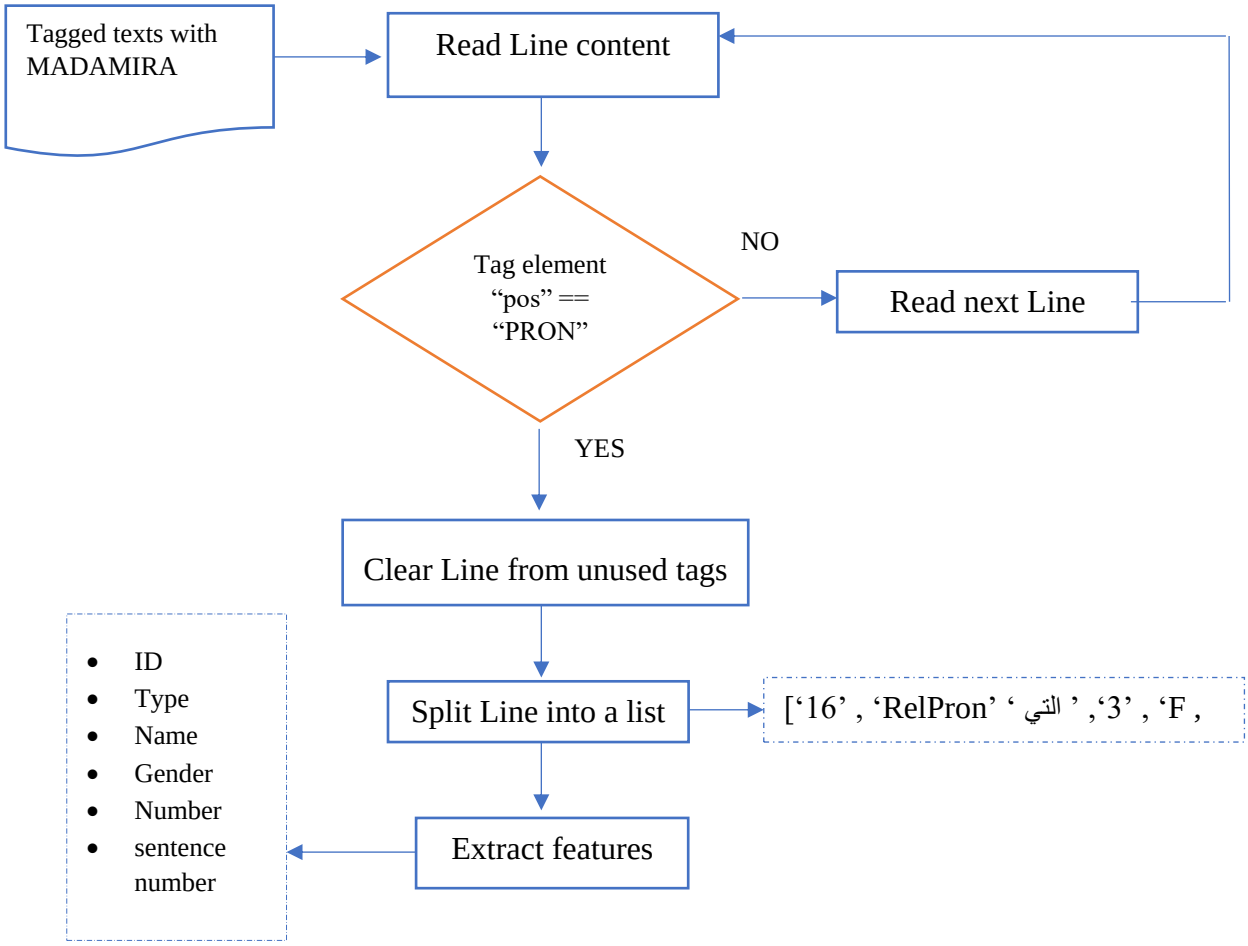


Figure 25: Pronominal anaphora detection process.

- Verbal anaphora detection: in this case we going to take all the features we took for the pronominal anaphora as gender, number…etc. but we add another feature who is going to help us on the resolution of this anaphora which is the voice feature (active or passive form) and we going to explain later in the resolution module why do we need this special feature.

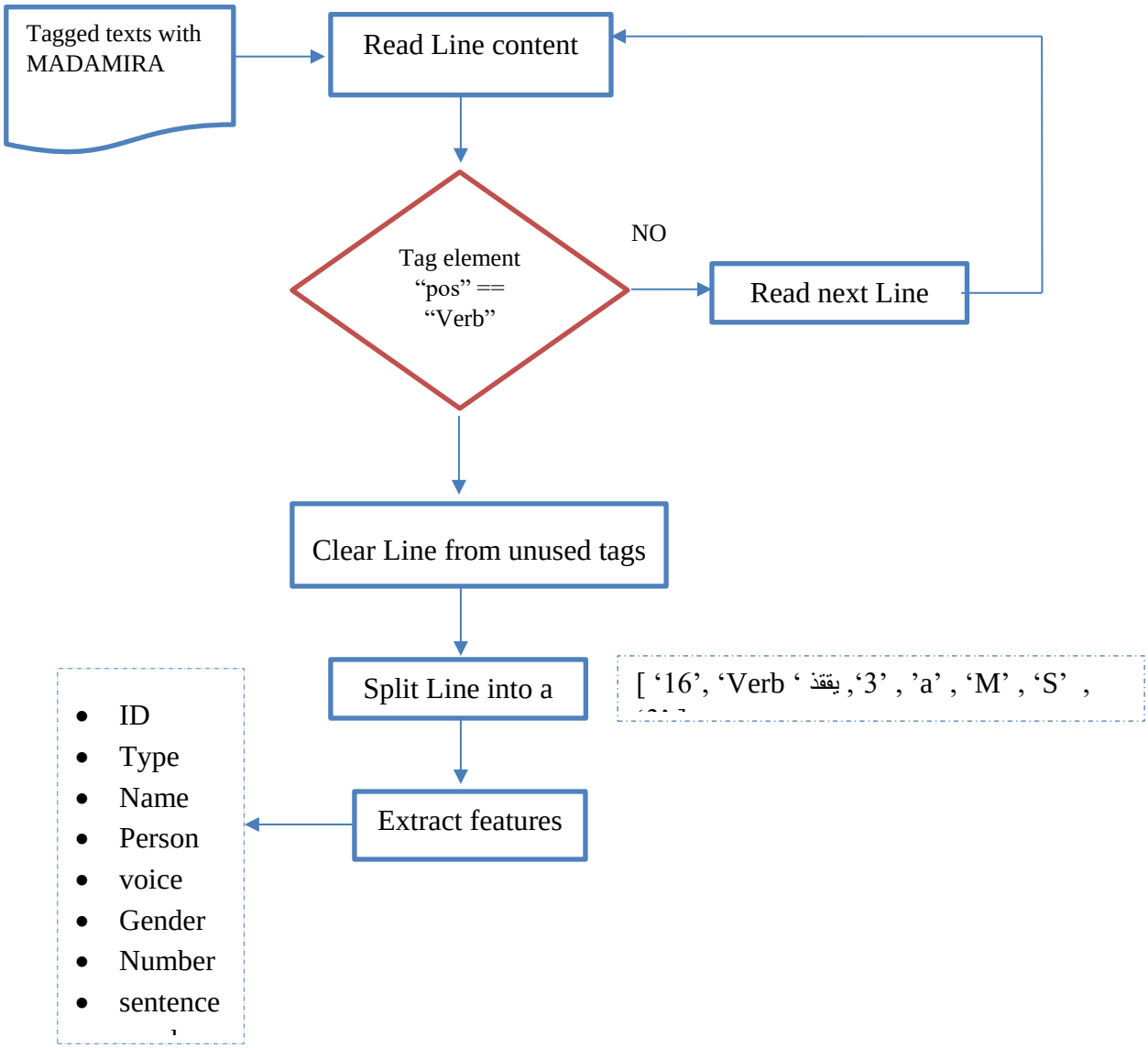


Figure 26: Verbal anaphora detection process.

When these operations are carried out, the program show each anaphora marked using colors following its type.

- *Identification of candidates' module:* Generally, all words that may be a possible candidate for the anaphora as nouns and NPs (Noun Phrase) are chosen as antecedents. In this phase, the search scope was limited with two sentences before each anaphora phenomena based on previous works [23], as for the demonstrative anaphora type, we also took two sentences

after as in some cases the correct antecedent may be after the anaphora. Candidates are selected based on ID attribute and all are stored in a list with theirs features.

The program will show automatically every antecedent in the text after gathering all information about them.

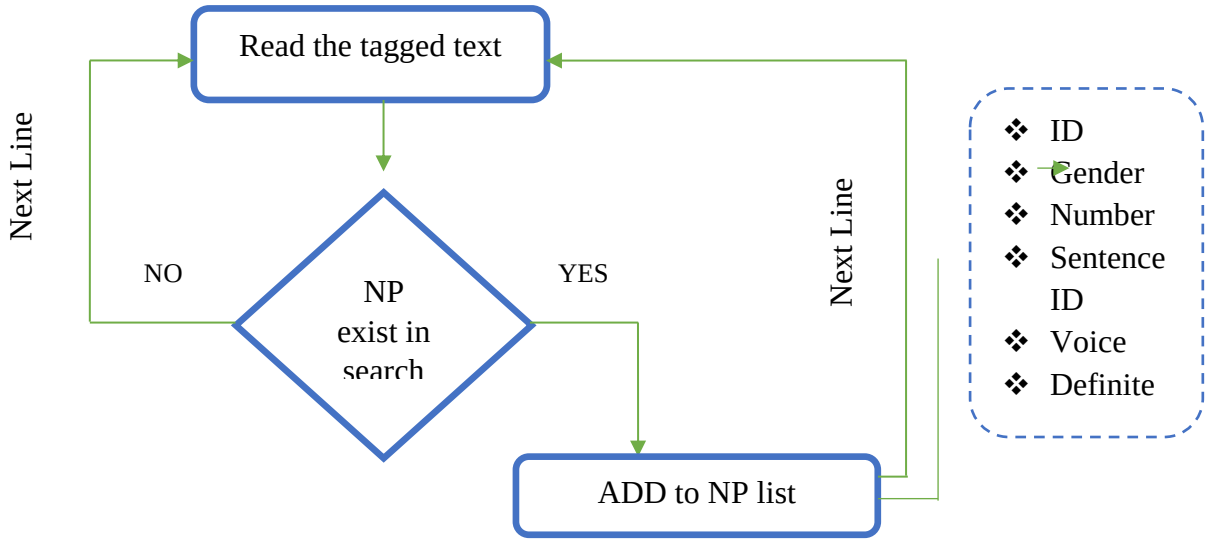


Figure 27: NPs detection process.

- *Anaphora resolving module:* Resolving module is the main part of the anaphora resolution process. This module focusses to resolve two types of anaphora, pronominal and verbal anaphora. After the identification of the anaphors and filtering the candidate list, we chose the most suitable antecedents for each anaphora from the listed most likely candidates. The filtering step using the semantic features of gender, number and existing sentence (search scope), those features are the eliminating factors used in anaphora resolution.
  - Pronominal anaphora: we applied set of linguistic rules, also named the preferential factors that can favor some candidate antecedents more than others in order to find the correct antecedent. All candidates contain a score value for each rule. We determined the scores for each rule and joined them to the previous score of each precursor. All candidates were graded and the one with the uppermost aggregate point score was suggested as the precursor. If several candidates showed similar point scores, the one nearest the anaphor was selected. The following table contain all features used with their score. As for relative pronouns we took the last matching NP to it as the best antecedent as it is a linguistic rule [12].

Linguistic rules	Description
Definiteness	A score of 1 is given if an NP is definite and of 0 if not.
Recency	A score of 1 is assigned to the recency NP to the anaphora and 0 if not.
Referential Distance	A score of 2 is assigned to NPs in the previous sentence or two sentences and further than those are given 0.
First Noun Phrases	A score of 1 is issued to the first NP of each sentence and 0 if not.
NPs in the title	A score of 1 is issued to the existing NP in title and 0 if not
Grammatical function.	Scores of 1 are given to an NP that has the same syntactic structure as the anaphora and 0 if not.
Frequency of NP in text	A score of 2 is assigned to the most frequent NP in text and 0 if not.

Table 7: The Linguistic Rules and their Respective Scores.

The linguistic factors are extended according to the Arabic language characteristics. They are linguistic ascertainments based on a set of Arabic texts such as: "Definiteness", "Distance", " Section header " and "grammatical function". In addition, we include the statistical factors as corpus-based extracted from a tagged Arabic corpus such as: "Frequency", "Referential Distance". After having the score of each feature, we do the sum of score and the antecedent with the higher score is taken as the best antecedent. In the case of having two or more antecedents with the same score we choose the closest one to the anaphor as the best antecedent.

- Verbal anaphora: as for the verbal anaphora we choose the same features as we did in the pronominal anaphora, with adding another feature which is the voice of the verb (passive or active form) as we noticed in the passive form the antecedent comes before the verbal anaphora, but in the active form it occur right after it.

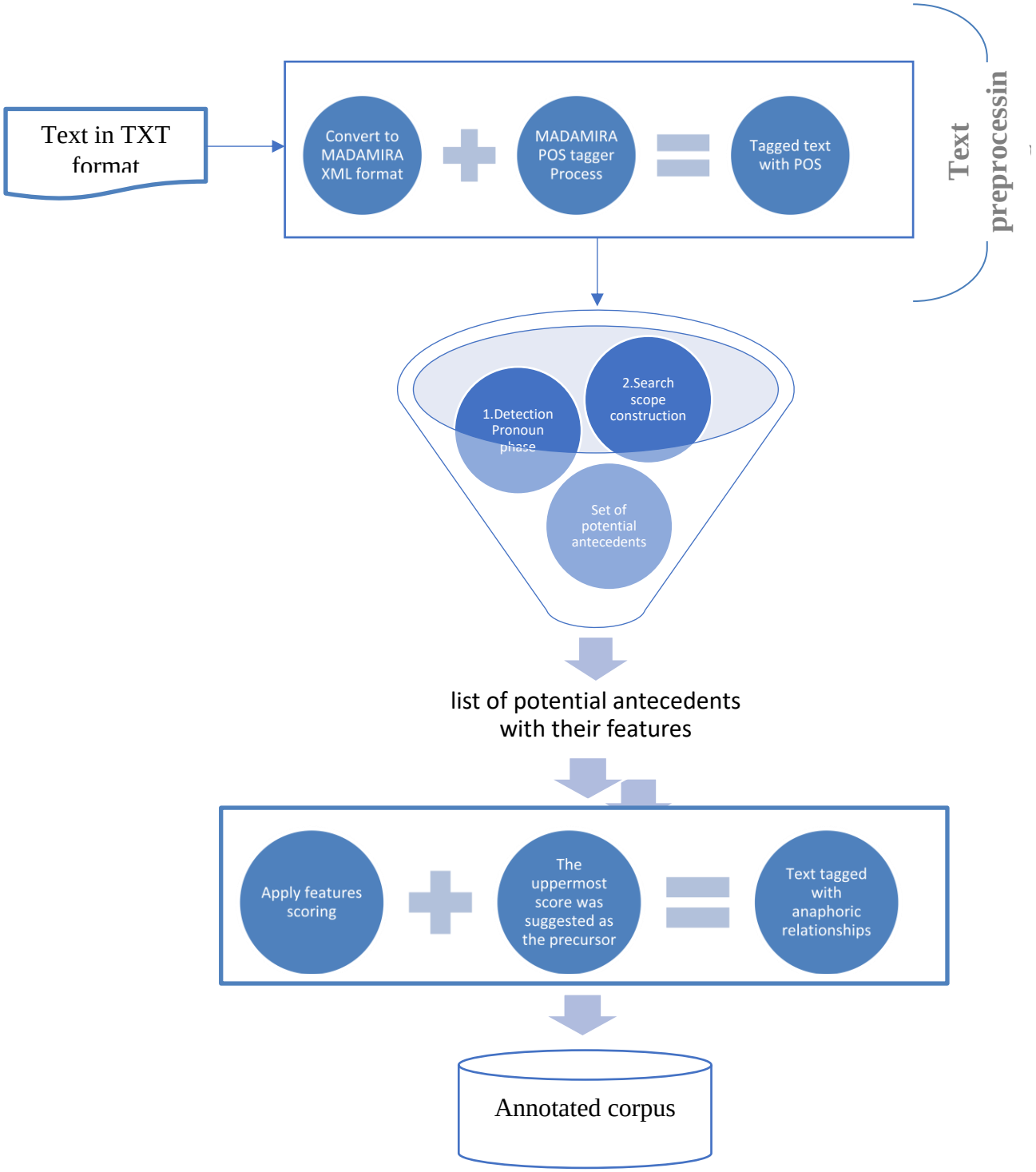


Figure 28: Antecedents detection in interface of our annotating tool.

2.2. Processing phase

2.2.2. Automatic corpus annotation

In order for a human or computer to understand and learn from the data it has been provided, the data has to be prepared in such a way that the computer can more easily find patterns and inferences. This is usually done by adding relevant metadata to a dataset. Any metadata tag used to markup elements of the dataset is called an annotation over the input.

Our target is to produce a Corpora we named A^{3}C that could be used for automatic anaphora resolution process of Arabic. Due to the lack of annotating tools in this language, we developed our tool under the name of A^{3}T to make the annotation process easier and gain time.

In this phase, the input of Tool is text encoded with (*.txt) and the output of the previous phase (a list that contains pairs of anaphora and the suitable antecedent).

When these two inputs are carried out, the program added the following tags: the antecedent is marked with <Antecedent id=”3” >. The remaining elements (anaphors) are marked with <Anaphor id=”5” rfr=”3” >. To make an efficient interaction with the annotated text, the SGML markup is hidden from the human annotator and each anaphora phenomena is marked using different colors.

We have introduced some syntactic information (Part of speech, grammatical functions, etc.) and other information (Definiteness, distance, frequency, in heading of sentence).

- TYPE attribute: indicates the anaphora type (Possessive, Demonstrative, relative pronoun, etc.)
- POS attribute: the antecedent part of speech (definite noun, indefinite noun, etc.).
- FRQ attribute: indicates the number of time that antecedent occurs in text.
- DIST attribute:” indicates the distance between the anaphor and its antecedent in term of sentence.
- GEN attribute: specify the gender of antecedent or anaphora.
- NUM attribute: specify the number of antecedent or anaphora.

أظهرت
< Antecedent id=”1” pos=”d” frq=”3” gen=”f” num=”s”> الدراسة </ Antecedent>
التي نُشرت نتائج
< Anaphor id=”2” rfr=”1” dist=”0” gen=”f” num=”s” type = “poss”> ها < /Anaphor >

Figure 29: Example from the A^{3}C.

2.2.2. Expert processing

For the last part of the application, we needed some experts on linguistics to verify and evaluate the final output of the software to both know the performance of our system and made correction to the A^3C if needed too, that way we would have a trustworthy Corpus that can be used for other researches. For that part, we created a friendly interface for the experts which it is going to give him the possibility to verify and if necessary, modify the links between the anaphora and its antecedent. The interface shows the annotated text in the middle as all the couples anaphora/candidate are shown on the right with the chosen couple by the system highlighted. All do the expert have to do in this case is to verify if the highlighted couple is the correct one, otherwise he can highlight the correct one from the other suggested couples.

3. Conclusion

In this chapter, we discussed the general architecture of the system we developed which we choose to call A^3T (Arabic Anaphora Annotating Tool), plus the modules used and the steps achieved leading from a raw corpus to finally a corpus fully annotated with anaphoric links specifically verbal and pronominal ones which we named A^3C (Arabic Anaphora Annotated corpus).

As our work is done, we going to need to discuss the statistics concerning our corpus plus the evaluation of annotating tool we created done by the linguistics expert, and of course a presentation of the system interface. All of that is going to be done in the next and final chapter.

CHAPTER V

Evaluation and Results

1. Introduction

The crucial role of evaluating a computer application in a scientific context, is to position our results and show the performance and effectiveness of our proposed system.

The purpose of this chapter is to focus on the Evaluation part our tool A^3T (Arabic Anaphoric Annotating tool) and the corpus made with. through four major points: the development environment we used, A presentation of the A^3T interface, statistics about the A^3C (Arabic Anaphoric Annotated Corpus), last point will be the evaluation of the A^3T performance.

2. Development Environment

2.1. Development language

The programming language we have adopted to implement our application is the python, which is an object and multiplatform programming language. It promotes mandatory programming structured, functional and object oriented.

the python allows (without imposing) a modular and object-oriented approach to programming. this language has been developed since 1989 by Guido van Rossum and many volunteer contributors.

- *Why python?* It is important to know, however, that this language adapts well to the application's domain, namely the search for information. Python is a language that can be used in many contexts and can be used in any type of use through specialized libraries.
- *Python Features*

the characteristics of the python are:

- easy to use for beginners.
- It is a generalist language.
- Portable.
- Free, but it can be used without restriction in commercial projects.
- Provide clear and user-friendly error messages.
- Python is a language that continues to evolve, supported by a community of users and managers, most of whom are supporters of free software. Alongside the main interpreter, written in C and maintained by the creator of the language, a second interpreter, written in Java, have been developed.

2.2. Development system

We have developed our system under very specific conditions, the following list clearly indicates the system configuration:

- Programming language version: Python 3.5.2
- IDE: Pycharm 2017.1.4.
- Machine features (computer):
  - ✓ Processor: i5-4210H 3.5Ghz.
  - ✓ Hard drive: 1Tb HDD, 128Gb SSD;
  - ✓ Screen resolution: 1920*1080.
  - ✓ RAM: 12Gb.
  - ✓ GPU: Nvidia GTX 860m.
  - ✓ Operating system: Windows 10 x64/Mint Linux 19.1 x64.

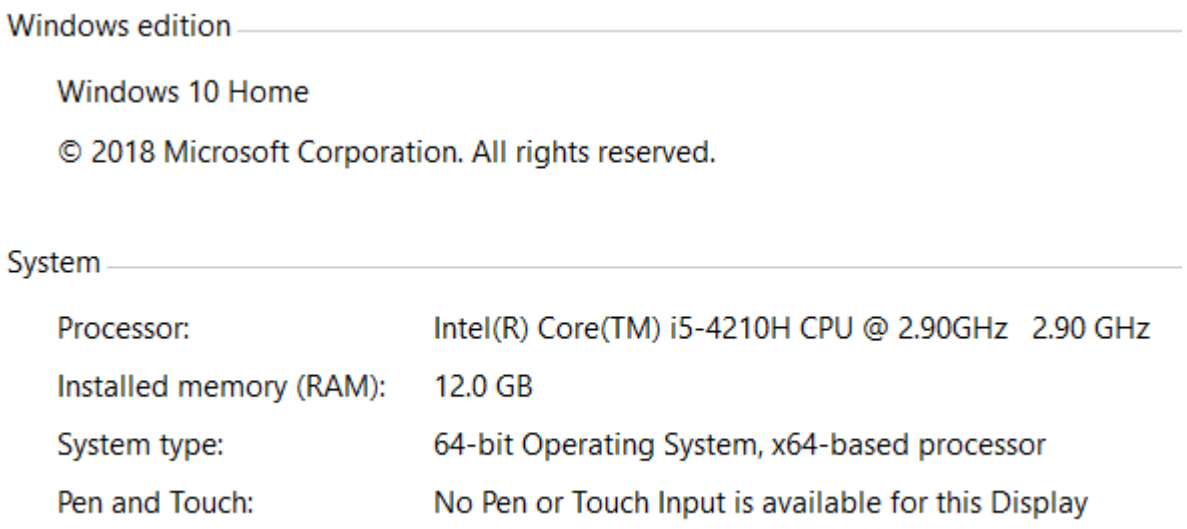


Figure 30: System configuration.

3. A^{3}C in numbers

In this section, we are going to discuss about statistics and numbers extracted from our corpus. As we mentioned in the last chapter, we worked on 5 different categories of newspaper articles (economy, education, politics, sport and miscellany). As shown in the table below, we extracted the words, nouns, verbs and sentences count from each category.

Category	Words count	Nouns count	Verbs count	Sentences count
Economy	302696	153449	23149	24020
Education	290044	134667	33189	29108
Politics	205954	98040	18121	14669
Sport	241374	170632	17685	12782
Miscellany	311255	107958	25896	17871
Total	1351323	511476	118040	98450

Table 8: A^{3}C word statistics.

In the next table we are going to give u the exact count of pronominal anaphora in each category followed by a percentage of each pronominal anaphoric type existing in.

Category	Pronominal anaphora count	Possessive	Demonstrative	Relative
Economy	18580	7455	6742	4383
Education	13210	5669	3121	2163
Politics	28540	13539	9276	5725
Sport	10953	6437	3572	3201
Miscellany	15069	6597	4853	3691
Total	86352	39697	27564	19163

Table 9: A^{3}C pronominal anaphora statistics.

Next, we are presenting all pronominal anaphoric types for each category using graphs so it would be clearer.

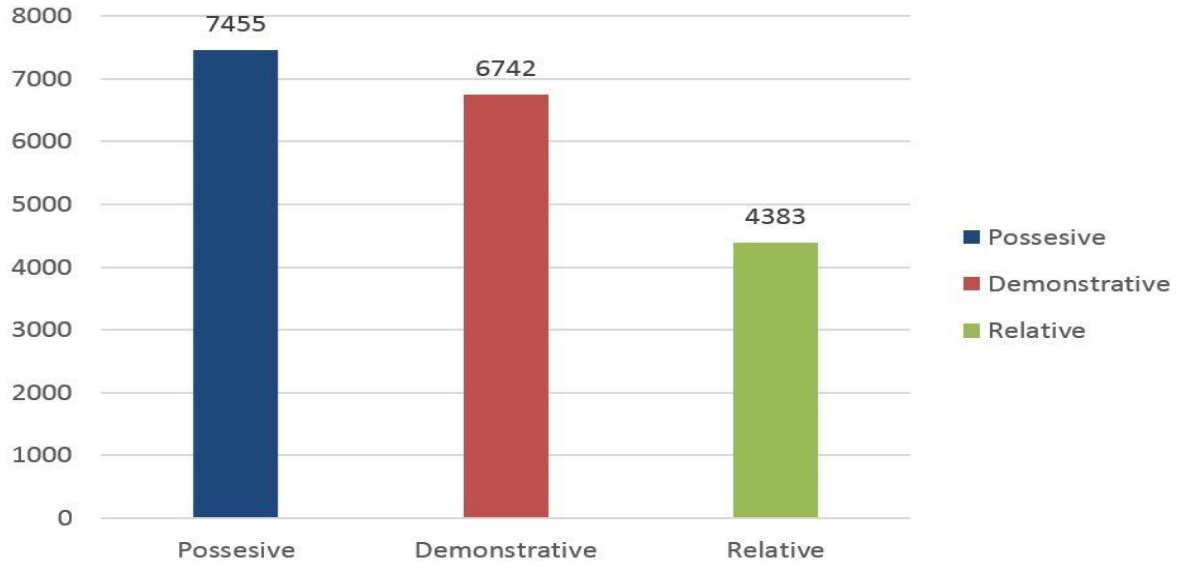


Figure 31: Economic category pronominal anaphora statistics.

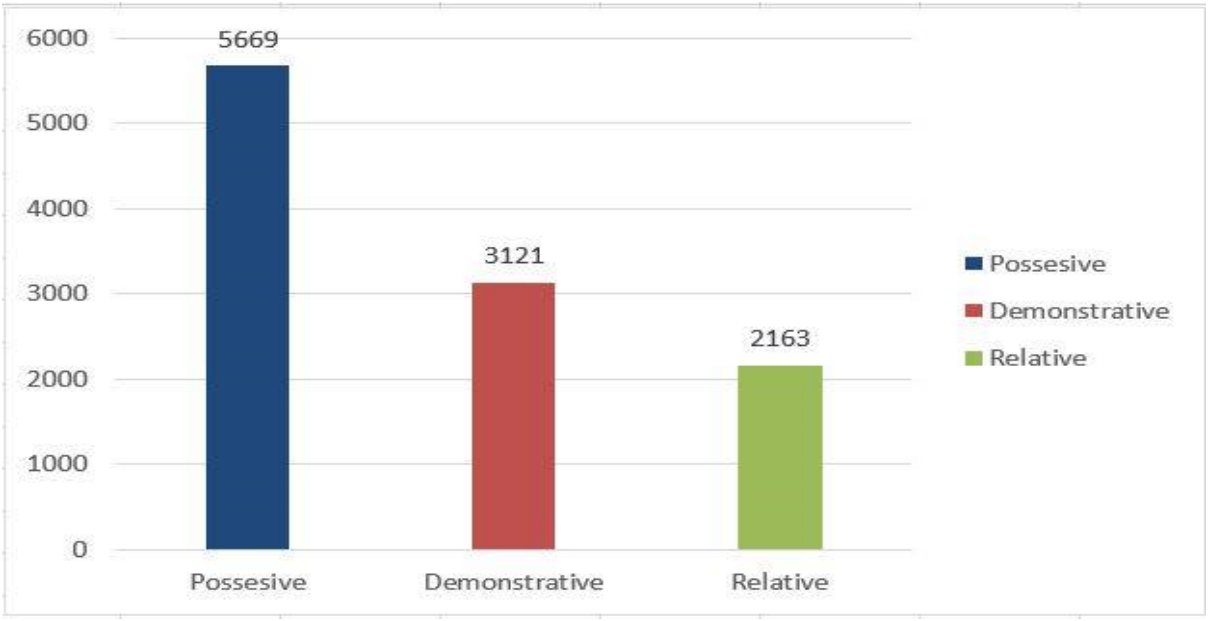


Figure 32: Education category pronominal anaphora statistics.

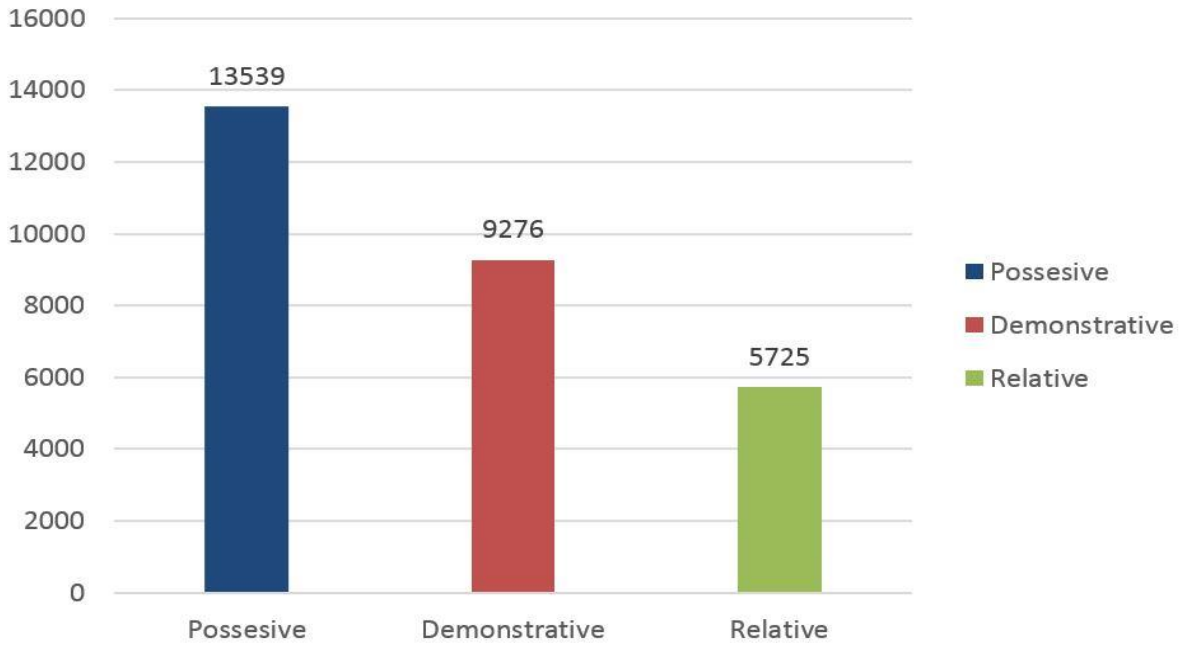


Figure 33: Politcs category pronominal anaphora statistics.

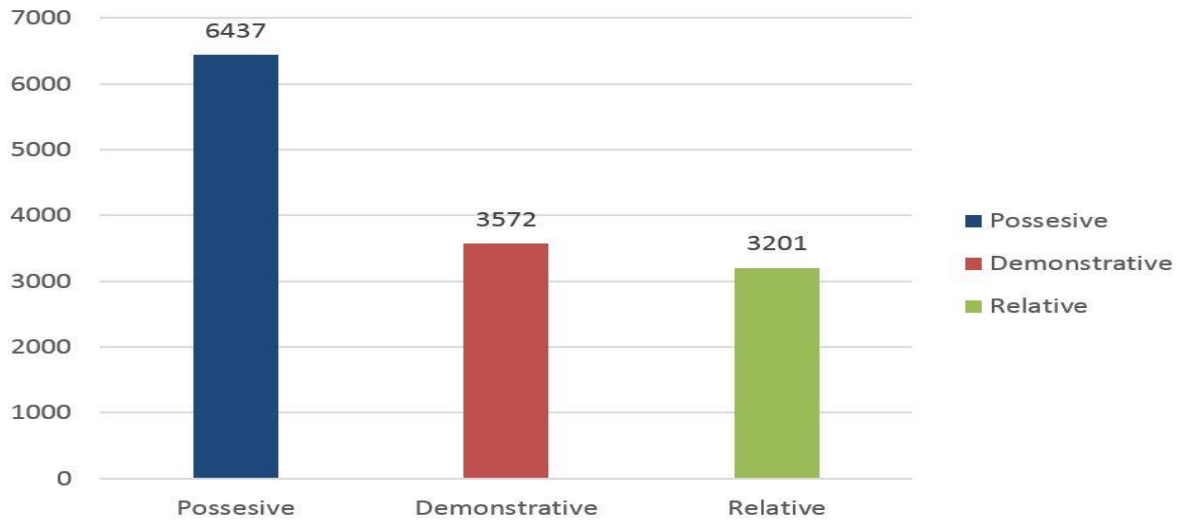


Figure 34: Sports category pronominal anaphora statistics.

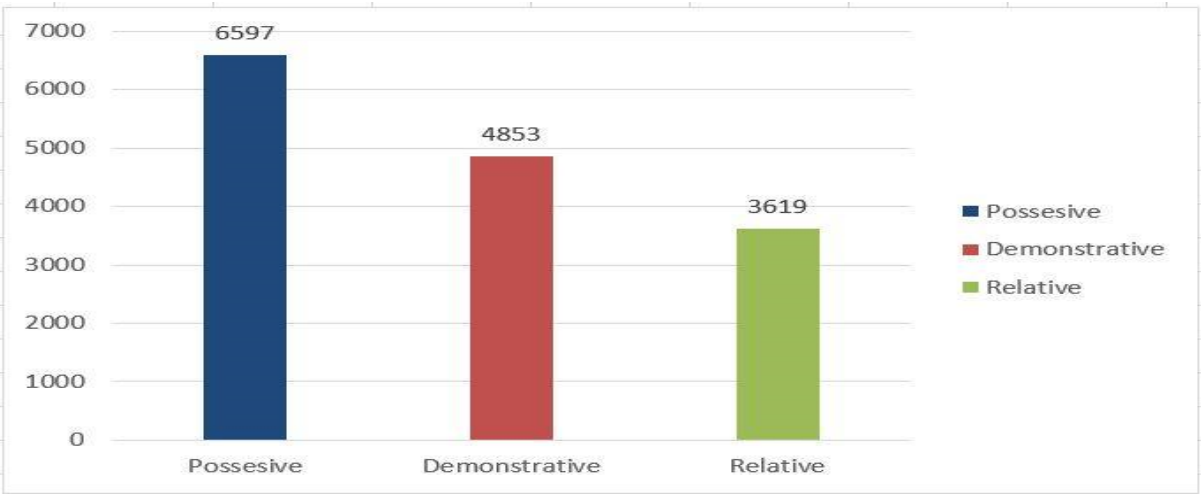


Figure 35: Miscellany category pronominal anaphora statistics.

We notice that in all categories we have chosen, the possessive type is always dominant with highest percentage. Second comes the demonstrative type. And last, the relative part. The sport category contains the highest percentage of possessive type (51.75%). the greater percentage of the demonstrative type is found in the economic category (36.25%). As we perceived that the relative type is way up in the politics than other categories.

4. A^{3}T expert interface

In this section, we found that it is preferable explaining the expert interface using screenshots containing an example.

First, the A^{3}T shows the original text we are going to annotate added by clicking the button add (figure 37).

Figure 36: Original text input.

After clicking in the button re-format, the text is going to split into sentences as shown in figure 38, and another text area appears with the MADAMIRA POS-tag (figure 39)

Figure 37: Original text sentence segmentation.

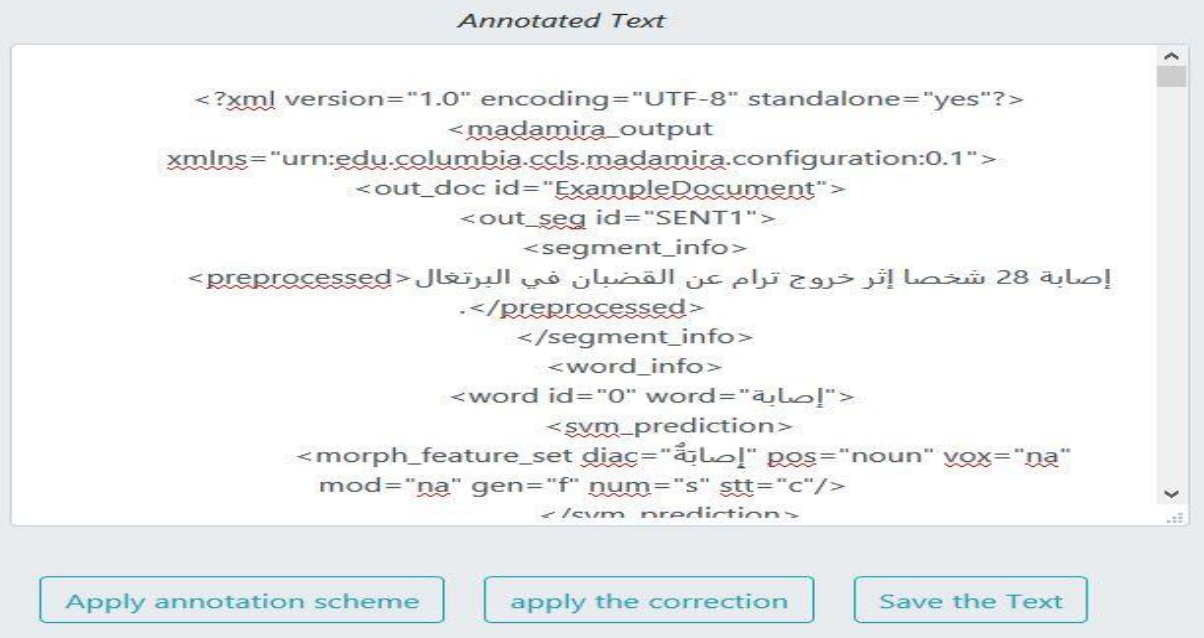


Figure 38: MADAMIRA part-of-speech tagging.

Next step is clicking on apply annotation scheme, the annotated text appears in the annotated text area with the highest probable antecedent linked to the anaphora. As in the right we show all the possible candidates as shown in the with their IDs and its probability of being the right antecedent. In case the expert finds an error, all he has to do is change the referent_id in the text following the suggestion in the right and hit “apply correction”. After finishing all the verification, the experts save the verified text with his name written in the right and clicking the button “save the text”, and he can pass to the next text.

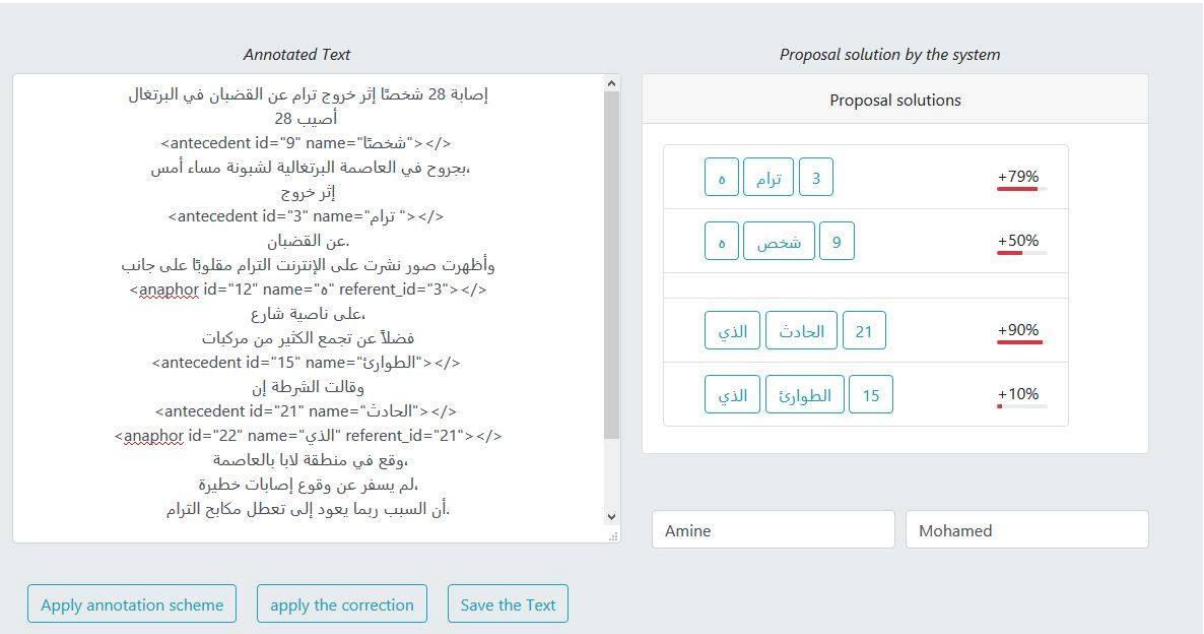


Figure 39: Expert verification and correction interface.

5. Evaluation and discussion

5.1. Pronominal Evaluation

We used the corpora that was annotated using AnATAr represent: a technical manual, newspaper articles, texts of Tunisian books used for basic education and a novel dataset, to evaluate the resolution system's performance. This evaluation passed through several basic steps as show in the following figure.

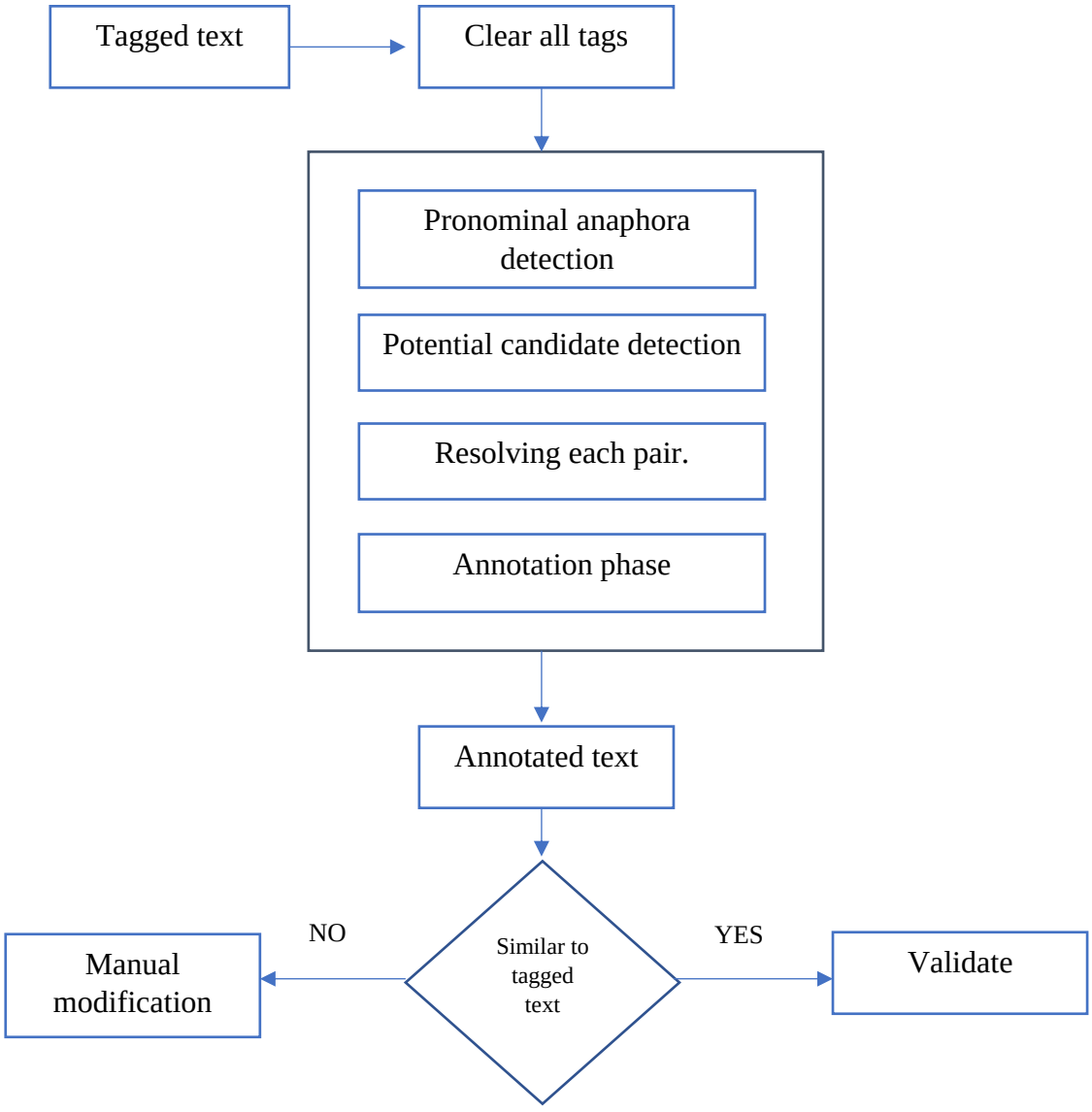


Figure 40: Process of Pronominal evaluation.

By doing the verification on the "AnAtAr" corpus, our tool achieved a success rate of 83.19%.

5.2. Verbal Evaluation

Since we are not aware of any other works done on the verbal anaphora, we going to evaluate our work by seeking the help of an expert in linguistics, who will define if the link between the verbal anaphora and its antecedent is correct or not and give us the right antecedent to see if it was initially included on the list. Through that, the performance of our system was evaluated using the following accuracy metric:

$$\textbf{Accuracy} = \frac{\text{Number of correctly resolve anaphora}}{\text{Number of all anaphoras}}$$

Concerning the verbal anaphora, we arrive at a satisfactory result whose rate of recognition average is: 57.23%.

6. Conclusion

In this chapter, we presented and discussed the result achieved by our application "A^3T" for resolving pronominal and verbal anaphora of Arabic texts. Tests exploited on "AnAtAr" corpus have given encouraging and satisfactory results. As the creation of our new corpus "A^3C" and its annotation with anaphoric links and presented the expert interface that helped us verifying the corpus and correcting it.

GENERAL CONCLUSION

Outcome

It is obvious, that the automatic processing of the Arabic language is an area that requires a lot of effort and deep research to make it able to simulate the human processing of texts. In the automatic processing of the Arabic language, anaphora play an important role in understanding texts by determining links between the text entities. In this study, we developed a tool for solving pronominal and verbal anaphora to select optimal referents among candidates based on linguistic concepts (MADAMIRA POS-Tag). In order to evaluate our application, several tests were carried out, the results obtained on the tests corpus showed a satisfactory average percentage in terms of success rate of 83.19% for pronominal anaphora and 57.23% on the verbal anaphora.

Perspectives

As a perspective, we can identify a few points that can improve the quality of our A^3T, such as:

- Include other knowledge in the resolution: semantic aspects
- use a more effective Tagger.
- Deepening constraints: defining other operating rules.
- Exploit other anaphoric resolution approaches such as the statistical approach.
- Solve other types of anaphora (lexical, nominal and comparative)

As for our corpus A^3C we may look forward for a better correction and increasing it's size allowing more effectiveness in other researches.

BIBLIOGRAPHY

[1] Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

[2] Reshamwala, A., Mishra, D., & Pawar, P. (2013). Review on natural language processing. IRACST Engineering Science and Technology: An International Journal (ESTIJ), 3(1), 113-116.

[3] Joseph, S. R., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. Natural Language Processing: A Review.

[4] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[5] T. Winograd, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, 1971, MIT-AI-TR-235

[6] Guerra, A. "T. Rowe Price to hone in on voice systems," Wall Street and Technology, Vol. 19, No. 3, 2000.

[7] Aloulou, C. (2003). Analyse syntaxique de l'Arabe: Le système MASPAR. *RECITAL, Nantes–France*.

[8] DAHOU Abdelghani, «Acquisition de Connaissances à partir d'un texte Arabe non vocalisé (JEEM BOX) » mémoire de Master 2, Université d'Adrar, juin 2014.

[9] Sreelekha, S. (2017). Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective. *arXiv preprint arXiv, 1708*

[10] Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, *8*(4), 14.

[11] FERGUSON, C. 1996. Epilogue: Diglossia revisited. In *Contemporary Arabic Linguistics in Honor of El-Said Badawi*. The American University in Cairo.

[12] Hammami, S., Belguith, L., & Ben Hamadou, A. (2009). Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links. *International Arab Journal of Information Technology (IAJIT)*, *6*(5).

[13] Bouzid, S. M., Trabelsi, F. B. F., & Zribi, C. B. O. (2017, October). How to combine salience factors for Arabic Pronoun Anaphora Resolution. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)* (pp. 929-936). IEEE.

[14] Al-Sulaiti, L., & Atwell, E. (2004). *Designing and developing a corpus of contemporary Arabic* (Doctoral dissertation, University of Leeds (School of Computing)).

[15] Bakari, W., Bellot, P., & Neji, M. (2017). A preliminary study for building an Arabic corpus of pair questions-texts from the web: AQA-WebCorp. *arXiv preprint arXiv:1709.09404*.

[16] Zaghouani, W. (2017). Critical survey of the freely available Arabic corpora. *arXiv preprint arXiv:1702.07835*.

[17] Hirschman L., MUC-*7* Coreference *Task Definition,* Longman, 1997

[18] Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., and Antoniadis G.,"Annotating a Large Corpus with Anaphoric Links," *in Proceedings of the Discourse Anaphora and Reference Resolution Conference,* pp. 134-137, UK, 2000.

[19] Davies S., Poesio M., Bruneseaux F., and Romary L., "Annotating Coreference in Dialogues: Proposal for a Scheme for MATE, 1998.

[20] Sharaf, A. B. M., & Atwell, E. (2012, May). QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *LREC* (pp. 130-137).

[21] Seddik, K. M., Farghaly, A., & Fahmy, A. A. (2015). Arabic Anaphora Resolution: Corpus of the Holy Qurâ [euro](TM) an Annotated with Anaphoric Information. *International Journal of Computer Applications*, *124*(15).

[22] Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014, May). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC* (Vol. 14, pp. 1094-1101).

[23] Abdullatif Abolohom and Nazlia Omar (January 2017). A Computational Model for Resolving Arabic Anaphora using Linguistic Criteria. Indian Journal of Science and Technology, Vol 10(3),