Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links

Souha Hammami, Lamia Belguith, and Abdelmajid Ben Hamadou
LARIS-MIRACL Laboratory, University of Sfax, Tunisia

Abstract: *Annotated resources are much needed for evaluation and training of anaphora resolution systems. The coreferential chain annotation is a difficult task which can not be realised without an appropriate tool. In this paper, we present our work on Arabic corpora annotation with anaphoric links (i.e., the annotation of the identity relation between the anaphors and their antecedents). In particular, we propose an anaphoric annotating tool for Arabic. Anaphoric annotating tool for Arabic has the advantage of automatic detection of Arabic pronouns and allows the human annotator to select several anaphoric pronouns related to the same antecedent. Our aim is to build a real corpus which will be used for anaphora resolution (i.e., either for system training or evaluation).*

Keywords: *Anaphora resolution, Arabic language, corpus annotation tool, pronominal anaphora, lexical anaphora.*

Received December 18, 2008; accepted June 24, 2009

1. Introduction

Corpora (raw and annotated) are very important for the research in anaphora and coreference resolution. Raw corpora are commonly available, but they have so far made only a limited contribution to the process of anaphora resolution (extracting collocation patterns). However, corpora annotated with anaphoric or coreferential links are not widely available, even though they are much needed for different methods in anaphora/coreference resolution systems. These corpora can be used for obtaining empirical data and rules in the building of new anaphora resolution approaches. They can also be used for training, optimization and evaluation of existing approaches.

The production of annotated corpora is a challenging and time-consuming task which follows a specific annotation scheme. In order to facilitate the application of this scheme, some annotating tools are developed making the annotation process easier for the user. With the help of these annotating tools, several corpora with anaphoric or coreferential links has been produced especially for English (MUC7 [12], GNOME [18], *etc.*) and for French (ARCADE [23], FReeBank [21], *etc.*). Nevertheless, this is not the case for the Arabic language which works on anaphora resolution are very few (i.e., to our knowledge there is only one research work on Arabic anaphora resolution done by Mitkov and Belguith [15]). The Mitkov's approach was initially developed and tested for English and was later adapted for Arabic by Mitkov *et al.* [15]. They have used one additional indicator (relative pronoun indicator) and adapted other indicators such as the referential distance. The evaluation of the adapted version for Arabic reported a success rate of 95.8%. However this evaluation was based on 190 anaphors only, taken from an Arabic technical manual. Thus, more experiments should be reported to argue whether the obtained accuracy will be the same in large corpus or not and also to affirm that Mitkov's results do hold-up out of domain or are restricted domain.

We propose, in this paper, a tool designed for annotating Arabic corpus by marking the coreferential chains.

This paper is structured in six sections. In the second section, we define some basic notions such as anaphora, cataphora and deixis. Section 3 presents the typology of Arabic anaphors since they are very specific and different from those of other languages. We present, in section 4, a statistical study of theses different anaphors. Section 5 reports briefly an overview of existing annotation schemes, annotating tools and annotated resources. Section 6 presents the adopted scheme, the annotating tool that we have developed and the requirements for our tool. Section 7 presents the conclusions and further ways of exploiting the annotated corpora.

2. Basic Notions: Anaphora, Cataphora and Deixis

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). When the anaphor refers to an antecedent and both have the same referent in the real world, they are called coreferential. Although, coreference and anaphora are two different concepts, in reality, they most often co-occur except in some cases. Note that, not all varieties of anaphora are based on referring expressions such as verb anaphora in example 1 or bound anaphors in example 2. On the contrary, coreference may occur without anaphora. For example, the use of the same proper name consecutively with each one referring to the same entity.

(1) هيَأ الشَيخ أخاه الصَبي لنومه كما يفعل كلَ ليلة

/hay~aOa Al$~ayoxu OaxaAhu AlS~abıy~a linawomihi kamaA yafoEalu kul~a layolapK/[1] the old man prepared his brother for sleeping as he does each night

(2) وضعت أمي في المزهرية أزهارا صفراء تتوسطها واحدة حمراء

/waDaEato Oum~iy fiy Almizohariy~api OazohaArFA SaforaA'a tatawas~aTuhaA waAHidapN HamoraA'a/ My mother puts in the vase yellow flowers and in the centre a red one.

- Anaphora/ Cataphora: the anaphora is defined as being the resumption of an entity already evoked previously in the text, whereas, the cataphora occurs when a reference is made on an entity mentioned further in the text (e.g., "هو الموت لا يترك حيّا." /huwa Almawotu lAa yatoruku Hay~FA/ It is death that leaves no one alive).
- Anaphora/deixis: the deixis is a linguistic phenomenon which identifies a person, an object, a place, *etc.* in a context or in a specific situation (e.g., "أنت أخذت كتابي" /Oanota Oaxa*ota kitaAbiy/ you took my book).

3. Typology of Arabic Anaphors

3.1. Pronominal Anaphora

Pronouns form a special class of anaphors because of their empty semantic structure; they do not have an independent meaning from their antecedent. In addition, not all pronouns are anaphoric: e.g., deictic pronouns such as "أنا" /OanaA/ *I*, "أنتَ" /Oanota/ *you* and "نحن" /naHonu/ *we* are not anaphoric ones [11].
Pronominal anaphora includes third personal pronouns (ضمير الغائب), demonstrative pronouns (أسماء الإشارة) and relative pronouns (الأسماء الموصولة).

Personal pronouns: in Arabic, third personal pronouns can be classified in disjoint or joint pronouns and also in nominative, dative or accusative ones. Thus, we distinguish:

- Nominative disjoint personal pronouns (الضمائرالمنفصلة في محل رفع)

هنّ	هم	هما	هي	هو
hun~a	humo	humaA	hiya	huwa

(3) روى أشونا لإخوتــه ما شاهده في رحلة الصّيد المثيرة وهم يستمعون إليــه بانتباه وغبطة

/rawaY Oa$uwnaA liIixowati__hi maA $aAhadahu fiy riHolapi AlS~ayodi Almuviyrapi wahumo yasotamiEuwna Iilayo__hi biAnotibaAhK wagiboTapK/ Achouna told his brothers all that he saw in the exciting hunting trip and they were listening to him with interest and joy.

- Accusative disjoint personal pronouns (الضمائر المنفصلة في محل نصب)

Iiy~aAhumo	إيَاهم	Iiy~aAhu	إيَاه
Iiy~aAhun~a	إيَاهنَ	Iiy~aAhaA	إيَاها
		Iiy~aAhumaA	إيَاهما

(4) أعجب أحمد بـاللَوحة الّتي رسمتها فأهديته إيَاها

/OuEojiba OaHomadu biAll~awoHapi Alat~iy rasamotuhaA faOahodayotuhu Iiy~aAhaA/
Ahmed was impressed by the painting which I drew so I offered it to him.

- Dative and accusative joint personal pronouns (الضمائر المتَصلة في محل نصب وجرَ)

ـهنَ	ـهم	ـهما	ـها	ـه
hun~a	humo	humaA	haA	hu

(5) جدّتي عائشة عجوز ليس لها من الأهل سوانا وسوى ولد واحد يعيش في ديار الغربة مع زوجته وأبنائه لم تره منذ سنوات

/jad~atiy EaA}i$apu EajuwzN layosa lahaA mina AlOaholi siwaAnaA wasiwaY waladK waAHidK yaEiy$u fiy diyaAri Algurobapi maEa zawojatihi waOabonaA}ihi lamo tarahu muno*u sanawaAtK/ My grandmother Aicha is an old woman whose family consists only of us and one son who lives abroad with his wife and his children. She hasn't seen him for years.

- Nominative joint personal pronouns (الضمائر المتَصلة في محل رفع)

ن	واو	ا
noon	waw	alef

(6) خرج الأولاد (VSO)

/xaraja AlOawolaAdu/ The children have gone out.

(7) الأولاد خرجوا (SVO)

/AlOawolaAdu xarajuwA/ The children have gone out.
The dative and accusative joint personal pronoun is the pronoun that can not begin a sentence; contrary to the disjoint pronoun (nominative and accusative). So, the dative and accusative joint personal pronoun should be attached to a noun (زوجته /zawojatihi/ his wife), a verb (تره /tarahu/ saw him) or a preposition (لها /lahaA/ to her) as shown in example 5. Nevertheless, the disjoint pronoun can also be attached to some prefixes (e.g., the

[1] For all examples, the transliteration is produced by the buckwalter Arabic transliteration system (http://www. qamus.org/ transliteration. htm).

conjunction of coordination ”ف ” /fa/ and “و” /wa/) such as (وهم /wahumo/ *and they*) in example 3.

The nominative joint pronoun is a particular pronoun which is always suffixed to a radical verb also called the clitic pronoun[2] . This pronoun takes up the position of the subject in a (SVO) sentence [7] as in example 7. Consequently, in example 6 we can not use the clitic pronoun (و) because the subject (الأولاد /AlOawolaAdu/ the children) occurs after the verb.

Pleonastic pronouns are considered non-anaphoric since they are not interpreted as linked to any expression (antecedent). For example, in English the pronoun “it” could be pleonastic (e.g., “It is important”, “It is necessary”, *etc.*). Similarly, in Arabic language, the joint pronouns (“ـه” /hu/) and (“ـها” /haA/) can be non-anaphoric in some cases as in equations 8 and 9.

أرى أنّــه من الواجب أن أعتذر لها (8)

/OaraY Oan~a__hu mina AlwaAjibi Oano OaEota*ira lahaA/ I think that it is obligatory to apologize for her

إنّـها تمطر (9)

/Iin~ahaA tumoTiru/ It’s raining.

- Relative pronouns: the relative pronoun in Arabic is always anaphoric and is referring to the immediate previously mentioned noun phrase. “الذي, التي, اللذان, اللتان, اللذين, اللتين, اللاتي, اللواتي, اللائي, الذين, الآلاء, من, ما” For example, in example 10 the relative pronoun (الّتي) refers to the noun phrase (الكتب /Alkutuba/ the books).

تصفّحت الكتب الّتي اشتريتـها (10)

/taSaf~aHotu Alkutuba Al~tiy Ai$otarayotuhaA/I skimmed the books which I have bought.

- Demonstrative pronouns: some demonstrative pronouns are only deictic (e.g., الآن, هنا) and some others can have different uses (deixis, cataphora or anaphora). But, contrary to other languages such French and English, most of the demonstrative pronouns are cataphoric. For example, the demonstrative pronouns (هذا), (ذاك) and (هنالك) are successively deictic in example 11, cataphoric in example 12 and anaphoric in example 12.

لم يمرّ في تاريخ العائلة حدث كهذا. (11)

/lamo yamur~a fiy taAriyxi AlEaA}ilapi HadavN kaha*aA/ It has never happened an event like that in the family background.

(12) اشترى أحمد كرة من ذاك المتجر، هنالك حيث
يمكن أن نجد جميع أنواع اللّعب.

/Ai$otaraY OaHomadu kurapF mino *aAka Almatojari, hunaAlika Hayovu yumokinu Oano najida jamiyEa OanowaAEi All~uEabi/ Ahmed has bought a ball from that shop where we can find all sorts of toys.

3.2. Lexical Anaphora

Lexical anaphora is realised when the referring expressions are definite descriptions or proper names. These definite expressions increase the cohesiveness of the text and moreover they convey some additional information (synonymy, generalisation, specialization, *etc.*) [11].

(13) ولد ابن خلدون في تونس ثم هاجر العلامة إلى مصر

/wulida Aibonu xuloduwn fiy tuwnisa vum~a haAjara AlEal~aAmapu IilaY miSora/ Ibn Khaldoun was born in Tunisia then the scientist immigrates to Egypt.

3.3. Comparative Anaphora

The comparative anaphora [11, 24] represent anaphora in which the anaphoric expressions are introduced by lexical modifiers (e.g., آخر ,أخرى other, واحدة one) or comparative adjectives (e.g., أكبر greater than, أحسن better than). This type of anaphora specifies the relationship (such as set-complement, similarity and comparison) between the entities invoked by the anaphor and the antecedent. In example 11 taken from the Coran, surat “elomran” the lexical modifier (أخرى) refers to the noun phrase (فئة).

(14) قد كان لكم آية في فئتين التقتا فئة تقاتل في سبيل الله وأخرى كافرة

/qado kaAna lakumo |yapN fiy fi}atayoni AilotaqataA fi}apN tuqaAtilu fiy sabiyli Allahi waOuxoraY kaAfirapN/ There, you people have had an intellectual lesson to comprehend: Two forces met one fighting in favor of God and the other against God.

3.4. Verb Anaphora

Verb anaphora is another variety of anaphora which is characterized by the use of the verb (فعل did).

خلقنا لــنؤدّي واجباتنا وليس لنا بدٌ من تأديتها، (15)
فإن لم نفعل فنحن وحدنا الملومون.

/xuliqonaA linuWad~iy waAjibaAtinaA walayosa lanaA bud~N mino taOodiyatihaA ،faIino lamo nafoEalo fanaHonu waHodunaA Almaluwmuwn/. We live to do our duties and we have to achieve them and if we don’t, we are the only reproachable.

Compared with other language such as English or French, anaphoric expressions in Arabic are almost classified similarly. Although, we noticed some particularities for Arabic language. On the one hand, the third person pronouns can be used as demonstrative pronouns (e.g., هو الرّئيس قادم /huwa Alr~a}iysu qaAdimN/ He, the president, is coming). On the other hand, there is a dual form for the pronouns (i.e., هما

[2] According to [7], a clitic pronoun in Arabic is a pronoun that forms a phonological unit with the verb and it is always attached to it.

/humaA/) and the singular feminine pronoun (e.g., هي /hiya/ she) can refers to a plural non-human unit.

4. Statistical Study

Through statistics, we intend to determine the widespread type of anaphors. For this study, we have used a corpus of 77 457 words composed of:

- 12 articles of the newspaper « الصباح » (economics, culture, sport…) (6 712 words).
- A technical manual of a computer (24 338 words).
- A Tunisian book used for 8th level of basic education (32 871 words).
- A novel «حكايا طائر السمرمر» (13 536 words).

Figure1 illustrates the distribution of the various types of anaphora in our corpora. Note that even when we varied the corpora, we have obtained merely the same results. Indeed, the pronominal anaphora is the most frequent in all corpora. The joint personal pronouns are more frequent than the other type of pronouns (i.e., among the 4139 anaphoric pronouns 3544 are joint pronouns). The comparative and verb anaphora do not occur much in the corpora. This is due to the fact that these anaphora kinds are not frequently used in written Arabic.

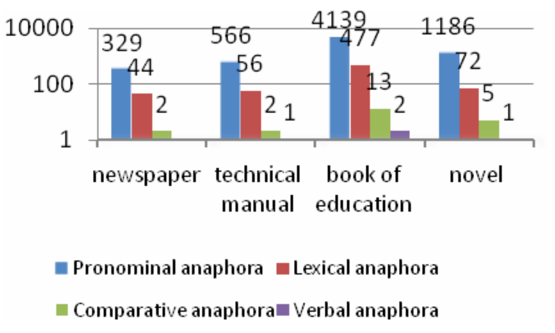


Figure 1. Statistical study of anaphoric expressions.

5. Previous Work

5.1. Annotation Scheme

Several Annotation schemes have been proposed since the early 1990s. The first scheme developed in the Lancaster and IBM project is the UCREL scheme [8]. UCREL allows the annotating of all kinds of anaphoric relations (pronominal and lexical anaphora, ellipsis and the generic uses of pronouns). It delimits the antecedent noun phrases and assigns a uniquely index number (identifier) to each one. The anaphoric relation is coded on the anaphoric expression by means of this identifier related to the antecedent. The scheme can also indicates the direction of pronominal reference by adding special symbols ('<' anaphoric relation, '>' cataphoric relation). This annotation scheme was considered by [23] that it provides a complete markup system for special cases such as multiple, ambiguous or uncertain antecedents, but the proprietary format could not probably be easily transformed in a standard markup language such as XML.

The MUC annotation scheme is proposed in the MUC-7 coreference task Definition [12]. Contrary to the UCREL, MUC scheme uses SGML tags to annotate anaphoric expressions offering a standard format which several schemes are derived from it. This scheme denotes the initial mention of an entity from the coreferential chain with <COREF ID="#">, and the rest of elements from the chain are marked using <COREF ID="#" TYPE="IDENT" REF="#"> tag. The ID attribute is a unique number which identifies an entity. Then, each one has its own ID. The attribute TYPE indicates the type of relation between the anaphor and the antecedent which is always an identity relation (IDENT). Finally, the REF attribute indicates which entity is coreferential with the current one. However, MUC presents two limits, *i*) it marks only the identity relations between two noun phrases, and *ii)* it can not capture relation between plural pronouns and discontinuous antecedents.

The XML-based scheme proposed in [23] uses a set of five semantic relations to type the anaphoric relation such as coreference, set membership, description, sentential antecedent and indefinite relation. It encodes each expression (anaphor and antecedent) with the <exp> tag. As in MUC scheme, a unique ID number is inserted in <exp>. The link between an anaphor and its antecedent is indicated with the <ptr> element which is added to the anaphor. The *src* attribute allows identifying the antecedent of the current anaphor. This scheme can also encode special cases (double anaphoric links, ambiguity of anaphors and conjoined antecedent NPs, *etc.*).

The MATE scheme [5] consists of two schemes: a core scheme and an extended scheme. The core scheme deals only with coreference relations similar to the MUC scheme. The extended scheme offers the possibility to annotate other kinds of anaphors such as bridging anaphors. The MATE meta-scheme [2] adopts a stand-off annotation style allowing the parallel annotation for arbitrary number of linguistic levels (morphologic, syntax, discours, *etc.*).

5.2. Annotating Tools

The annotation task of anaphoric or coreferential relations require a considerable effort from the human annotator. Therefore, many annotating tools have been developed to offer an efficient interaction with the annotated text. In this section, we present some annotating tools such as Callisto[3], MMAX2 [16] and PALinkA [17].

[3] http://callisto.mitre.org

These tools are written in Java, taking advantage of its portability[4].

Callisto is a new annotating tool recently developed at MITRE which represents a different design from the Alembic Workbench system [6]. It has been built with a modular design by the use of stand-off annotation. Callisto can be extended by means of integration of the domain specific extension components. MMAX2 is developed at EML Heidelberg [16]. As Callisto, MMAX2 uses a stand-off annotation allowing it to keep the database separate from annotations on the file level. So, it supports arbitrarily many levels of annotation where each one consists of a separate file.

PALinkA [17] enables the annotation of coreference relations, automatic summarization and centering theory. The user can specify the annotation task in an external file. For the coreference annotation task, the text is displayed without the XML tags in the main screen of the tool. In addition, coreferential links can be identified in the right window by using the entities tree. PALinkA was tested for English, French, Romanian and Spanish language.

5.3. Annotated Corpora

The availability of annotated corpora with coreferential links is vital for automatic anaphora resolution systems [13]. As a consequence, the number of corpora annotated both anaphorically and coreferentially have increased. For English, there are some resources such as the Lancaster Anaphoric Treebank [8] annotated with the UCREL scheme, the MUC coreference task (MUC-6 and MUC-7), a part of the Penn Treebank [9] and the corpus of the University of Worlverhampton [14]. For French, there are the ARCADE corpus containing one million words annotated with anaphoric links [23] at ELRA and other corpus with free access in the FReeBank base [21]. There are also some resources for other languages, such as the VENEX [19] corpus of anaphora and deixis in spoken and written Italian and the Eus3LB [1] corpus of Basque which is a part of the corpus of the 3LB project. However, as far as we know, there is not any study carried out in the field of anaphorical or coreferential corpus annotation for Arabic. In fact, the work of Mitkov and Belguith [15] was based on 190 anaphors detected from a technical manual and Haddar [10] and Chalabi [4] were mainly interested in the resolution of elliptic forms (zero Anaphora).

6. Developing our Annotated Corpora

Our target is to produce an annotated resource that could be used for automatic anaphora resolution process of Arabic. As we have mentioned above, in our knowledge there is no available resource for Arabic. So, an Arabic annotated resource is really needed to encourage works on Arabic anaphora resolution. Thus, we tried to develop an operational tool which allows building such resource easily either by computer scientists or linguists.

First, we started by testing and evaluating some tools like PALinkA and Callisto for Arabic even though they are not developed for Arabic language.

Our aim was to check out whether these systems could work for Arabic or not. In fact, we can display the text and annotate it with anaphoric links. However, for the joint pronouns, as they require the selection of a part of the word and not the entire word, these systems fail to make the annotation. This is one of the reasons that conducted us to develop our own annotating tool. Section 6.2 develops other reasons.

6.1. Adopted Scheme

For our annotation, we adopted the XML-based scheme proposed by [23]. There are three main reasons that lead us to choose this scheme. The first one is its compatibility with the MUC annotation scheme, widely used in evaluation tasks. The second reason relies on the fact that it allows capturing relations between plural pronouns and discontinuous antecedents which is impossible to do with MUC scheme. Also, it supports the annotation of a variety of anaphoric relations (coreference, set membership, *etc.*) contrary to the MUC which only covers the identity (IDENT) relation. The third reason is that the adoption of this scheme is practical because it is coded in SGML which considerably facilitates the process of annotation.

Note that we have examined other schemes such as MATE which allows the markup of further varieties of anaphoric relations and adopts the stand-off style to link the expressions. The problem of a Stand-off markup in XML documents is that it may be unreadable for a human annotator without the help of some user interface [23]. In addition, since we are focusing only on anaphoric annotations we do not need to use this scheme.

6.2. AnATAr

6.2.1. Description

In this section, we present our tool for annotating the coreferential links in an Arabic corpus. The input of Anaphoric Annotating Tool for Arabic (AnATAr) is encoded with XML and is segmented into paragraphs and sentences. The segmentation is accomplished by the the STAr tokenizer for Arabic texts [2] which is integrated in our annotating tool. Segmenteur de Textes Arabe (STAr)is a tokenizer for Arabic texts which is based on a contextual analysis of punctuation

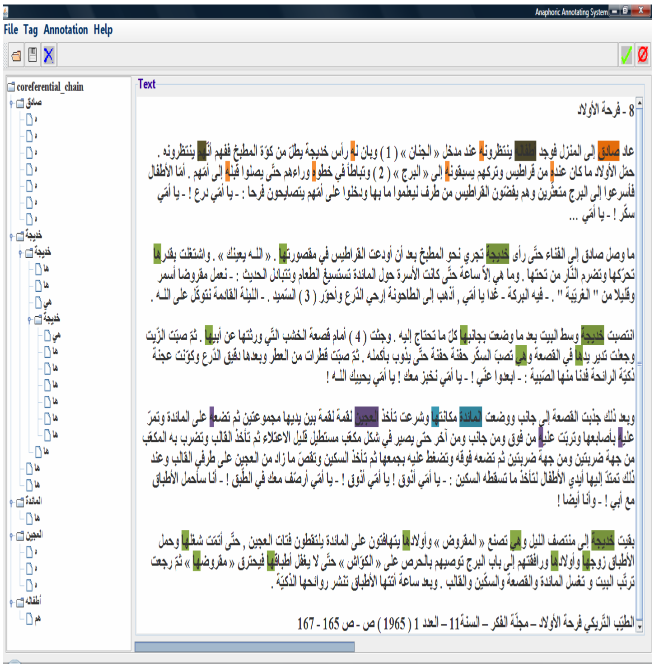
[4] We present only these tools because they are the most known ones and also because we have evaluated them.

marks and a list of particles, such as the coordination conjunctions.

AnATAr allows the user to select the text to be annotated. Once selected, the text is displayed at the right side of the screen as shown in Figure 2. The annotation process is kept as simple as possible. This process is done by the following operations:

- The anaphor selection: the user selects the anaphors using the mouse. She or he right clicks in order to choose the anaphor from the menu.
- The antecedent selection: the user selects the unit corresponding to the antecedent of the selected anaphor and chooses the antecedent. If this antecedent is already inserted in the tree which is displayed in the left screen as shown in Figure 2, the user can select it from the tree.

When these two operations are carried out, the program added the following tags: the initial mention of a chain (antecedent) is marked with <exp id=”e3” >. The remaining elements from the chain (anaphors) are marked with <exp id=”e1”><ptr src=”e3”/ >. To make an efficient interaction with the annotated text, the SGML markup is hidden from the human annotator and each coreferential chain is marked using different colours as shown in Figure 2.



8 - فرحة الأولاد

عاد صادق إلى المنزل فوجد أطفاله ينتظرونه عند مدخل « الجنان » (1) وبان له رأس خديجة يطلّ من كوّة المطبخ ففهم أنّهم ينتظرونه .
حمّل الأولاد ما كان عنده من قراطيس وتركهم يسبقونه إلى « البرج » (2) وتباطأ في خطوه وراءهم حتّى يصلوا قبله إلى أمّهم . أمّا الأطفال
فأسرعوا إلى البرج متعثّرين وهم يفضّون القراطيس من طرف ليعلموا ما بها ودخلوا على أمّهم يتصايحون فرحا : - يا أمّي درع ! - يا أمّي
سكّر ! - يا أمّي ...

ما وصل صادق إلى الفناء حتّى رأى خديجة تجري نحو المطبخ بعد أن أودعت القراطيس في مقصورتها . « اللـه يعينك » . واشتغلت بقدرها
تحرّكها وتضرم النّار من تحتها . وما هي إلاّ ساعة حتّى كانت الأسرة حول المائدة تستسيغ الطعام وتتبادل الحديث : - نعمل مقروضا أسمر
وقليلا من " الغربيّة " . - فيه البركة - غدا يا أمّي , أذهب إلى الطاحونة إرحي الدّرع وأحوّر (3) السّميد . - الليلة القادمة نتوكّل على اللـه .

انتصبت خديجة وسط البيت بعد ما وضعت بجانبها كلّ ما تحتاج إليه . وجثت (4) أمام قصعة الخشب التّي ورثتها عن أبيها . ثمّ صبّت الزّيت
وجعلت تدير يدها في القصعة وهي تصبّ السكّر حفنة حفنة حتّى يذوب بأكمله . ثمّ صبّت قطرات من العطر وبعدها دقيق الدّرع وكوّنت عجنة
ذكيّة الرائحة فدنا منها الصّبية : - ابعدوا عنّي ! - يا أمّي نخبز معك ! يا أمّي يحييك اللـه !

وبعد ذلك جذبت القصعة إلى جانب ووضعت المائدة مكانتها وشرعت تأخذ العجين لقمة لقمة بين يديها مجموعتين ثم تضعه على المائدة وتمرّ
عليه بأصابعها وتربّت عليه من فوق ومن جانب ومن آخر حتى يصير في شكل مكعّب مستطيل قليل الاعتلاء ثم تأخذ القالب وتضرب به المكعّب
من جهة ضربتين ومن جهة ضربتين ثم تضعه فوقه وتضغط عليه بجمعها ثم تأخذ السكين وتقصّ ما زاد من العجين على طرفي القالب وعند
ذلك تمتدّ إليها أيدي الأطفال لتأخذ ما تسقطه السكين : - يا أمّي أذوق ! يا أمّي أذوق ! - يا أمّي أرصّف معك في الطّبق ! - أنا سأحمل الأطباق
مع أبي ! - وأنا أيضا !

بقيت خديجة إلى منتصف الليل وهي تصنع « المقروض » وأولادها يتهافتون على المائدة يلتقطون فتات العجين , حتّى أتمّت شغلها وحمل
الأطباق زوجها وأولادها ورافقتهم إلى باب البرج توصيهم بالحرص على « الكوّاش » حتّى لا يغفل أطباقها فيحترق « مقروضها » ثمّ رجعت
ترتّب البيت و تغسل المائدة والقصعة والسكّين والقالب . وبعد ساعة أتتها الأطباق تنشر روائحها الذكيّة .

الطيّب التّريكي فرحة الأولاد – مجلّة الفكر – السنة11 – العدد 1 (1965) ص - ص 165 - 167

Figure 2. The interface of our annotating tool.

In addition to these requirements which were highlighted by several researchers [13, 17], we describe in the following section other ones specific to Arabic language.

6.2.2. AnATAr: Main Advantages

When we started to develop our annotating tool, we kept in mind some specifities of Arabic anaphoric expressions. As, for examples, the density of pronouns in texts, the successive use of pronouns in one or many sentence(s) and the fact that the pronoun in Arabic can often appear as a suffix of a noun, a verb or a preposition.

The density of pronouns in Arabic texts is widely higher than that in other languages such as French. Tutin [22] reports that the anaphoric density[5] varies depending on the texts genre and the maximum value is (4.32) in narrative ones. The value of the anaphoric density in Arabic is widely higher in different genre of corpus (e.g., in narrative texts this density can reach 13.69). These observations lead us to think that the annotation of written texts in Arabic is more difficult and time consuming than that in other language such as French or English. Therefore, in order to accelerate the identification process of anaphoric entities by the annotator, we decided to integrate a module which could automatically detect the anaphoric pronouns. This module is based on *i)* a morphological analyser MOPH2 [3], *ii)* a set of syntactic patterns to solve some ambiguities, *iii)* pleonastic patterns to identify pleonastic pronouns and *iv)* cataphoric patterns to identify cataphoric pronouns as shown in Figure 3.

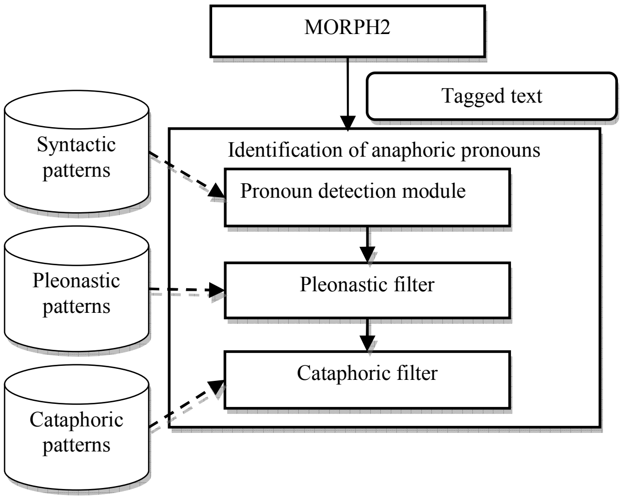


Figure 3. Automatic anaphoric pronoun identification.

In Arabic, the identification of pronouns is not always straightforward. Indeed, Arabic is an agglutinative language, the pronouns can occur as suffixes of nouns, verbs or prepositions. Thus, the morphological analysis is very essential to solve the ambiguities. For example, in the word (وجه /wajohu/ face) the letter (ـه /hu/) is a part of this word while in the word (كتابه /kitaAbahu/ his book) it represents a pronoun. In other cases the morphological analysis is insufficient to solve these ambiguities, so we use some syntactic patterns. For example, in the sentence “ فهم الولد الدَّرس” /fahima Alwaladu Ald~arosa/ *The boy understood the lesson,* the word “فهم” /fahima/

[5] The anaphoric density is the ratio number of anaphoric expressions/number of words.

understood is a verb, however in the sentence "فهم يلعبون" /fahumo yaloEabuwna/ *and they play* it is an anaphoric pronoun "هم" attached to the coordination conjunction "ف" /fa/.

So, in the latter case, we can determine that "هم" /humo/ is an anaphoric pronoun because the word "يلعبون" /yaloEabuwna/ *play* is a verb (i.e., in Arabic we can not find two consecutive verbs, except for action verbs "أفعال الشَرَوع").

The following step consists of identifying the pleonastic pronouns which do not refer to any antecedent. For this step, we use a simple pattern matching procedure. If the matching succeeds, the pronoun is considered as pleonastic. The set of pleonastic patterns is constructed basing on the corpus which contains 224 instances of the pronoun "ـه". Among these pronouns 93 are pleonastic.

The last step determines the cataphoric pronouns based on a list of patterns. In Arabic, a cataphor is generally intra-sentential (i.e., this cataphor and its subsequent belong to the same sentence) such as in example 16.

(16) ها هو الرَّئيس قادم

/ haA huwa Alr~a}iysu qaAdimN/ Here is the president coming.

If the pronoun is preceded by a demonstrative pronoun "ها" /haA/ and followed by a definite noun phrase, then it is a cataphoric pronoun.

Consequently, when the annotator selects the option "Automatic Identification of Anaphoric Pronouns" in the top menu, all identified anaphoric pronouns will be marked by using different colours depending on their type (anaphoric, cataphoric or pleonastic).

In Arabic, we frequently find the successive use of pronouns (especially the joint pronouns) in one or many sentences which refer to the same antecedent. In the following example there are seven anaphoric pronouns (ها /haA/). All these pronouns refer to the same noun phrase (العرب).

احتاجت العرب إلى الغناء بمكارم أخلاقها وطيب أعراقها
(17) وذكر أيّامها الصّالحة وأوطانها النّازحة وفرسانها الأمجاد
وسمحائها الأجواد لتهتزّ أنفاسها إلى الكرم

/AiHotaAjato AlEarabu IilaY AlginaA'i bimakaArimi OaxolAaqihaA waTiybi OaEoraAqihaA wa*ikori Oay~aAmihaA AlS~aAliHapi waOawoTaAnihaA Aln~aAziHapi wafurosaAnihaA AlOamojaAdi wasamoHaA}ihaA AlOajowaAdi litahotaz~a OanofaAsuhaA IilaY Alkarami/.

Consequently, in our system, the annotator can select several anaphoric pronouns then she or he links them with their antecedent by a simple click. This can speed up the process of annotation and minimize the number of click carried out by the human annotator.

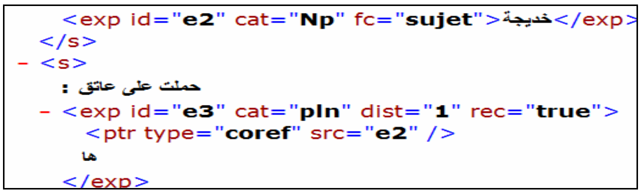
Compared to PalinkA, AnATAr has the advantage to be specific for annotating Arabic corpora with anaphoric links and easily used by linguist annotators who generally do not have knowledge about annotation schemes and computers. Note that, since the main objective of our tool is to develop annotated corpora for anaphoric resolution, it can not be used for other annotation tasks such as annotating centering and annotating for automatic summarisation.

6.3. Annotating Corpora

The corpora that we have annotated using AnATAr represent: a technical manual, newspaper articles, texts of Tunisian books used for basic education and a novel as shown in section 4. In the first stage of the annotation of our corpora, we have decided to focus on the annotation of the identity relation between the anaphors (pronouns, definite descriptions or proper names) and their antecedents (noun phrases).

However, we have excluded the demonstrative pronouns from pronominal anaphora because most of these pronouns are cataphoric and when they are anaphoric, they do not refer generally to a simple noun phrase but to a sentence or a paragraph. This choice could be justified by two reasons: the first one consist of the fact that the process of annotation is a very difficult task and the second one consist of the fact that our target is to produce an annotated corpora for the most frequent types of anaphors in Arabic. Consequently, the task of the human annotator becomes easier and more rapid.

Anaphoric relations have been annotated in the texts of the book for 8th of basic education level; we have introduced some syntactic information (Part of speech, grammatical functions, *etc.*) and other information (Definiteness, distance, *etc.*) in order to test them in Arabic language. For example, the attribute *"cat"* indicates the pronoun type (Joint pronoun to a verb, joint pronoun to a noun, relative pronoun, *etc.*) or the antecedent part of speech (proper name, definite noun, indefinite noun, *etc.*). The attribute "dist" indicates the distance between the anaphor and its antecedent. So, when (dist=1) it implies that the antecedent occurs in the precedent sentence as shown in Figure 4. The corpus has been annotated in our laboratory and validated by our linguists.[6]



```
    <exp id="e2" cat="Np" fc="sujet">خديجة</exp>
  </s>
- <s>
    حملت على عاتق :
  - <exp id="e3" cat="pln" dist="1" rec="true">
      <ptr type="coref" src="e2" />
      ها
    </exp>
```

Figure 4. An extract of the annotated corpus.

7. Conclusion

We proposed in this paper an annotating tool for Anaphoric relations which can be used to annotate

[6] The annotation process was performed by three linguists which are professors at the faculty of letters.

Arabic corpora. These corpora are used for anaphora resolution. AnATAr takes many advantages from existing systems for anaphoric annotation such as Java environment, XML language and the easy interface for annotating. In addition, it offers other advantages like the automatic detection of Arabic pronouns and the selection of several anaphoric pronouns related to the same antecedent.

As perspectives, we intend to annotate the totality of corpora by at least two human annotators in order to ensure that the corpora are marked up correctly, as much as possible. Then, we consider developing and adding to our tool a program which offers the possibility to compare the different annotations given by the linguists. We also intend to examine the usefulness of some parameters used in anaphora resolution systems (e.g., distance, indefiniteness, focus, *etc.*) for Arabic language. Finally, we believe that our annotating tool and annotated corpora will be very profitable to the NLP community and will contribute to the development of an automatic anaphora resolution tool for Arabic. We are planning to publish these resources in order to make them available.

References

[1] Aduriz I., Ceberio K., and Diaz A., “Pronominal Anaphora in Basque: Annotation of a Real Corpus,” *Computer Journal of NLP*, vol. 37, no. 2, pp. 99-104, 2006.

[2] Belguith L., Baccour L., and Mourad G., “Segmentation de Textes Arabes Basée Sur L'analyse Contextuelle des Signes de Ponctuations et de Certaines Particules,” *in Processions of 12ème Conférence sur le Traitement Automatique des Langues Naturelles*, France, pp. 451-456, 2005.

[3] Belguith L. and Chaâben N., “Analyse et Désambiguïsation Morphologiques de Textes Arabes non Voyellés,” *in Processions of 13ème Conférence sur le Traitement Automatique des Langues Naturelles*, Belgique, pp. 493-501, 2006.

[4] Chalabi A., “Elliptic Personal Pronoun and MT in Arabic,” *in Proceedings of Arabic Language Processing,* Morocco, pp. 19-22, 2004.

[5] Davies S., Poesio M., Bruneseaux F., and Romary L., “Annotating Coreference in Dialogues: Proposal for a Scheme for MATE,” http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_ manual.html, 1998.

[6] Day D., Aberdeen J., Caskey S., Hirschman L., Robinson P., and Vilain M., “Alembic Workbench Corpus Development Tool,” *in Proceedings of the 1st International Conference on Language Resource and Evaluation*, Spain, pp. 1021-1028, 1998.

[7] Fassi A., *Distributing Features and Affixes in Arabic Subject Verb Agreement Paradigms*, John Benjamins Publishing Company, 1996.

[8] Garside R., Fligelstone S., and Botley S., “Discourse Annotation: Anaphoric Relations in Corpora,” *in Proceedings of Corpus Annotation Linguistic Information from Computer Text Corpora*, London, pp. 66-84, 1997.

[9] Ge N., “Annotating the Penn Treebank with Coreference Information,” *Internal Report*, Brown University, 1998.

[10] Haddar K., “Caractérisation Formelle des Ellipses de la Langue Arabe et Processus de Recouvrement de la Langue Arabe,” *Thèse de Doctorat*, Université de Tunis II Faculté des Sciences de Tunis, 2000.

[11] Hechiri C., “ الضمير ودوره في الجملة,” *PhD Thesis*, Faculty of Letters and Humanities of Sfax, 1998.

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

[13] Mitkov R., *Anaphora Resolution,* Longman, 2002.

[14] Mitkov R., Evans R., Orasan C., Barbu C., Jones L., and Sotirova V., “Coreference and Anaphora: Developing Annotating Tools Annotated Resources and Annotation Strategies,” *in Proceedings of DAARC2000,* UK, pp. 49-58, 2000.

[15] Mitkov R., Belguith L., and Malgorzata S, “Multilingual Robust Anaphora Resolution,” *in Proceedings of the 3rd International Conference on Empirical Methods in Natural Language Processing*, Grenade, pp. 7-16, 1998.

[16] Müller C., “Representing and Accessing Multi-Level Annotation in MMAX2,” *in Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing,* Italy, pp. 73-76, 2006.

[17] Orasan C., “PALinkA: A Highly Customisable Tool for Discourse Annotation,” *in Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue,* Japan, pp. 39-43, 2003.

[18] Poesio M., “The MATE/GNOME Scheme for Anaphoric Annotation,” *in Proceedings of SIGDIAL*, Boston, pp. 168-175, 2004.

[19] Poesio M., Delmonte R., Bristot A., Chiran L., and Tonelli S., “The VENEX Corpus of Anaphora and Deixis in Spoken and Written Italian,” http: //cswww.essex.ac.uk/staff/poesio /publications/VENEX04.pdf.

[20] Salmon-Alt S., “Entre Corpus Et Théorie: L’annotation Coréférentielle,” *Computer Journal of Traitement Automatique des Langues Nouveaux Corpus Nouvelles Pratiques Nouveaux Concepts*, vol. 42, no. 2, pp. 459-486, 2001.

[21] Salmon-Alt S., Bick E., Romary L., and Pierrel M., “La FREEBANK: Vers Une Base Libre De

Corpus Annotés," *in Processions of TALN 2004*, Morocco, pp. 19-21, 2004.

[22] Tutin A., "A Corpus Based Study of Pronominal Anaphoric Expressions in French," *in Proceedings of the Discourse Anaphora and Reference Resolution Conference*, UK, pp. 265-277, 2002.

[23] 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.

[24] Webber B., Stone M., Joshi A., and Knott A., "Anaphora and Discourse Structure," *in Proceedings of Computational Linguistics*, UK, pp. 545-587, 2003.

Souha Hammami received a Master degree in computer science in 2004 from the National Engineering School of Sfax, Tunisia. Since 2006, she is preparing her PhD in computer science at the Faculty of Economic Sciences and Management of Sfax University, Tunisia. Her research is focused on anaphora resolution in Arabic texts.

Lamia Belguith received a PhD degree in computer science in 1999 from the Higher Institute of Management in Tunis, Tunisia. She is an associate professor in the Department of Computer Science at the Faculty of Economic Sciences and Management of Sfax, Tunisia.

Abdelmajid Ben Hamadou is a professor in the Department of Computer Sciences at the Higher Institute of Computer Sciences and Multi-media of Sfax, Tunisia. His research interests include multimedia applications, formal approaches, intelligent information systems, natural language processing, semantic web, video indexing, and multi-agent system.