

# CORPUS CONSTRUCTION FOR ARABIC ANAPHORA RESOLUTION

PREPARED BY

ABDELHALIM HAFEDH DAHOU AND MOHAMED ABDELMOAZZ

SUPERVISED BY

MR. MOHAMED AMINE CHERAGUI

# OVERVIEW

 1. Introduction

 2. General concepts

 3. Conception

 4. Evaluation

 5. Conclusion

# 1. INTRODUCTION

ANAPHORA RESOLUTION IS SEEN TO BE VERY CHALLENGING AND COMPLEX IN NLP (NATURAL LANGUAGE PROCESSING) SPECIALLY FOR THE ARABIC LANGUAGE DUE TO ITS COMPLEXITY. MAJOR NLP APPLICATIONS NEED A PROPER ANAPHORA RESOLUTION AND IDENTIFICATION.

OUR MAIN OBJECTIVE IS CREATING A TOOL THAT WOULD HELP US ANNOTATING AND CREATING A CORPUS WITH PRONOMINAL AND VERBAL ANAPHORIC LINKS.

## 2. GENERAL CONCEPTS

# NATURAL LANGUAGE PROCESSING (NLP)

A NATURAL LANGUAGE IS ANY LANGUAGE HUMAN LEARN FROM THEIR ENVIRONMENT USED TO COMMUNICATE WITH EACH OTHER.

NATURAL LANGUAGE PROCESSING (NLP) IS A CORPS OF STRATEGIES APPLIED TO REVEAL AND IDENTIFY SENTENCE BOUNDARIES, EXTRACT THEIR GRAMMATICAL STRUCTURE, AND DETECT THE LANGUAGE.

NLP IS BASED ON A NUMBER OF DISCIPLINES, AND COVER SEVERAL FIELDS OF STUDIES SUCH AS TEXT PROCESSING AND TEXT CLASSIFICATION.

# PURPOSE OF NLP

THE OBJECTIVE OF NLP IS PROCESSING THE LANGUAGE AS A HUMAN DOES, MEANS THAT A COMPUTER MUST UNDERSTAND THE INPUT OF A TEXT OR SPEECH AND MANIPULATE IT EASILY AND FASTER.

BY AUTOMATICALLY PROCESSING LINGUISTIC DATA, NLP WOULD SAVE US A HUGE AMOUNT OF TIME AND EFFORT.

# NLP APPLICATIONS

- INFORMATION EXTRACTION.
- INFORMATION RETRIEVAL.
- QUESTION-ANSWERING.
- SUMMARIZATION.
- MACHINE TRANSLATION.
- SENTIMENT ANALYSIS.

# ARABIC NATURAL LANGUAGE PROCESSING (ANLP)

THE ARABIC LANGUAGE HAS MORE THAN 422 MILLION SPEAKER, DUE TO THAT HUGE NUMBER AND ITS STRATEGICAL AND POLITICAL INFLUENCE ON THE WORLD, IT GAINED HUGE INTEREST IN THE NLP FIELD.

BY UTILIZING NLP, DEVELOPERS CAN ORGANIZE AND STRUCTURE KNOWLEDGE TO PERFORM LINGUISTICS TASKS SUCH AS ARABIC TEXT TRANSLATION, SPEECH RECOGNITION, AND TOPIC SEGMENTATION... ETC.

# ANLP CHALLENGES

- ABSENCE OF CAPITALIZATION.
- LETTERS CHANGING SHAPE ACCORDING TO THEIR POSITION IN THE WORD.
- THE LACK OF LETTERS DEDICATED TO VOWELS.
- FREE-WORD-ORDER LANGUAGE.
- WRITTEN FROM RIGHT TO LEFT AND SEMI-CURSIVELY.

# OBJECTIVES OF ANLP

THE MAIN FOCUS OF THE ANLP IS TO ENABLE NON ARABIC SPEAKERS GET THE MEANING OF ARABIC TEXTS. AS FOR THE ARABIC WORLD, THE ANLP APPLICATIONS DEVELOPED THERE HAVE DIFFERENT OBJECTIVES.

THE ANLP GOES THROUGH DIFFERENT LEVELS BUT ALSO THROUGH THE DEVELOPMENT OF DIFFERENT MODULES THAT CAN CONTRIBUTE TO UNDERSTANDING THE LANGUAGE, ONE OF THE IMPORTANT MODULES IS THE ANAPHORA RESOLUTION.

# WHAT IS AN ANAPHORA?

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).

الشجرة لازلت صامدة في مكانها وأوراقها جافة .

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

# DIFFERENT ANAPHORA TYPES

A. PRONOMINAL ANAPHORA ARE CHARACTERIZED BY ANAPHORIC PRONOUNS WHICH ARE DIVIDED INTO 3 SUB-TYPES:

- RELATIVE PRONOUNS: RELATIVE PRONOUNS ARE ALWAYS ANAPHORIC AND GENERALLY INDICATING THE PREVIOUSLY EXISTING NOUN PHRASE.
- PERSONAL PRONOUNS: ARE PRONOUNS THAT ARE ASSOCIATED PRIMARILY WITH A PARTICULAR GRAMMATICAL PERSON, WE CAN FIND IT ISOLATED (هُوَ) OR WITH A SUFFIX (كتابه).
- DEMONSTRATIVE PRONOUNS: IS A PRONOUN THAT IS USED TO POINT TO SOMETHING SPECIFIC WITHIN A SENTENCE.

البسط التي نسجتها يد طبيعة.

The valleys that have been created by nature.

## DIFFERENT ANAPHORA TYPES (CONT.)

B. VERBAL ANAPHORA ANOTHER VARIETY OF ANAPHORA WHICH IS CHARACTERIZED BY THE USE OF THE VERB ( فعل DID).

صرح عسكري على ان الانتصارات المنسودة من التحالف أرغمت الانقلابيين على الانسحاب من المدينة.

A military statement that the victories assigned by the coalition forced the coup to withdraw from the city.

## DIFFERENT ANAPHORA TYPES (CONT.)

C. LEXICAL ANAPHORA: LEXICAL ANAPHORA OCCURRED WHEN THE REFERRING STATEMENT GIVE A DESCRIPTION OF PROPER NAMES.

يُوسُفُ أَيُّهَا الصِّدِيقُ أَفْتَنَا فِي سَبْعِ بَقَرَاتٍ سِمَانٍ يَأْكُلُهُنَّ سَبْعٌ عَجَافٌ

Joseph, O man of righteousness! tell me the meaning of the dream of seven fat cows  
whom seven lean cows are devouring

## DIFFERENT ANAPHORA TYPES (CONT.)

D. COMPARATIVE ANAPHORA: THE INDICATION OF COMPARATIVE ANAPHORA IS WHEN THE ANAPHORIC STATEMENT INTRODUCED COMPARATIVE ADJECTIVES (أَكْبَرُ, أَفْضَلُ) (BIGGER, BETTER). THIS DEFINITION PROVES THE EXISTING OF COMPARISON, SIMILARITY AND COMPLEMENT BETWEEN THE ANAPHOR AND THE ANTECEDENT.

كان لديه قطتين ، واحدة سوداء والآخرى رمادية.

he had two cats, one is black and the other grey.



Rule-based approaches.



Statistical approaches.



Machine-learning  
approaches.

## ANAPHORIC RESOLUTION APPROACHES

# ARABIC ANAPHORA RESOLUTION CHALLENGES



The Arabic morphology.



The complexity of a sentence (Agglutination phenomenon).



The phenomenon of non-vowelized words.



The big lack of Arabic anaphora annotated corpus.

# WHAT IS A CORPUS?

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.

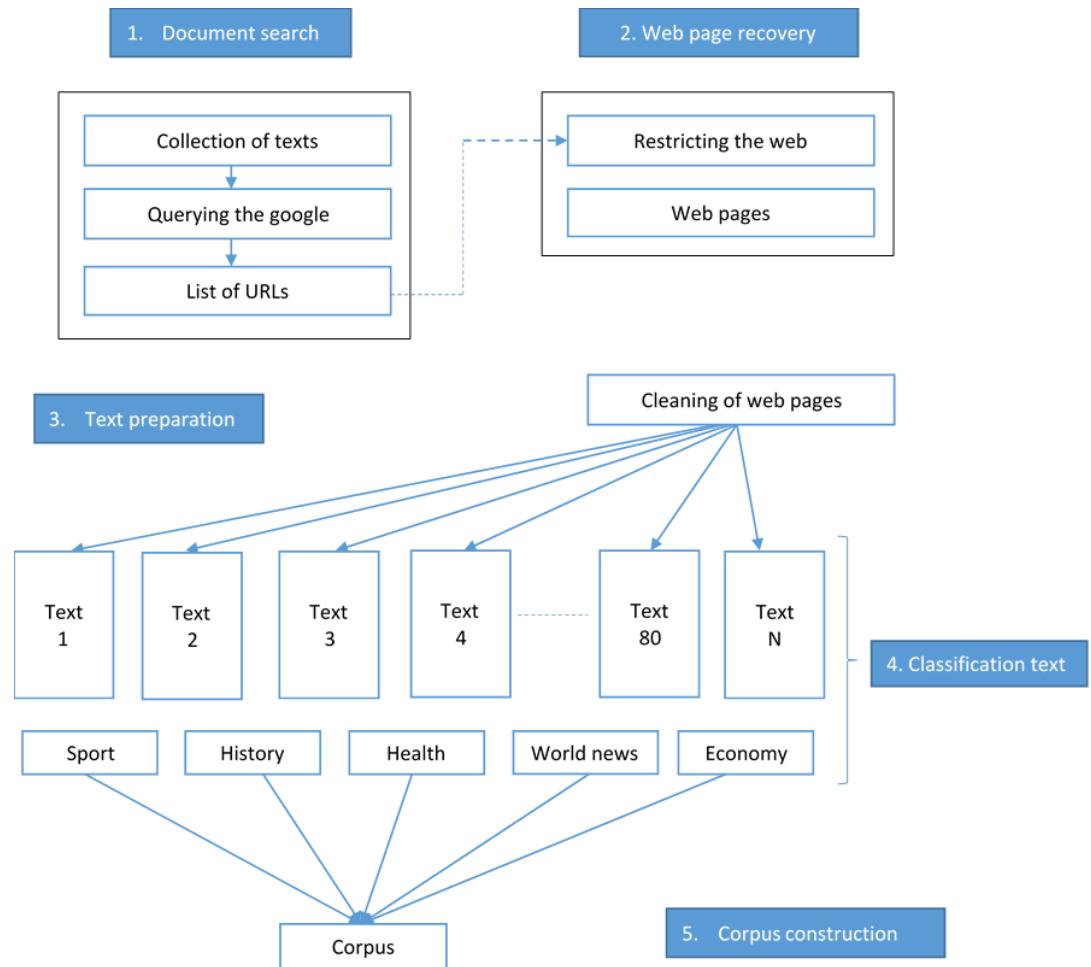
# THE UTILITY OF BUILDING A CORPUS

- EVALUATING AND AMELIORATING APPROACHES.
- USED FOR TRAINING AND TESTING.
- HELP FOR EXTRACTION NEW FEATURES AND RULES.
- RETRIEVING INFORMATION ABOUT WORDS BY APPLYING FREQUENCY AND REPETITION METHODS.

# HOW TO BUILD A CORPUS?

- DATA GATHERING.
- PRIMARY CHECK AND CONVERSION.
- LINGUISTIC ANNOTATION.
- TEXT CATALOGUING AND FINAL CHECKING.

# STEPS OF A CORPUS CREATION FROM THE WEB



# ANAPHORIC CORPUS ANNOTATION

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.

MUC scheme: uses SGML tags to annotate anaphoric expressions.

Extensible Markup Language (XML) scheme: simple and very flexible text format derived from SGML.

Meta scheme: which is divided into two schemes (Core scheme and extensible scheme)

## ANAPHORIC ANNOTATION SCHEMES

# WORKS ON ANAPHORIC ANNOTATED CORPORA

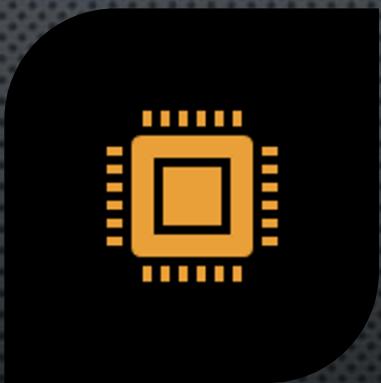
Work on annotated corpora	Scheme	Word segments	Pronouns	Text categories
<b>Arabic corpora annotation (AnAtAr)</b>	XML-based	77.124 word	4.300	Arabic newspapers Computer technical manuals
<b>QurAna Corpus</b>	AQA, MUC-7	128.000 word	24.679	Original Quranic text
<b>Combining QAC and QurAna</b>	MUC-7	128.240 word	29.287	Original Quranic text

# 3. CONCEPTION

# GENERAL ENVIRONMENT ARCHITECTURE



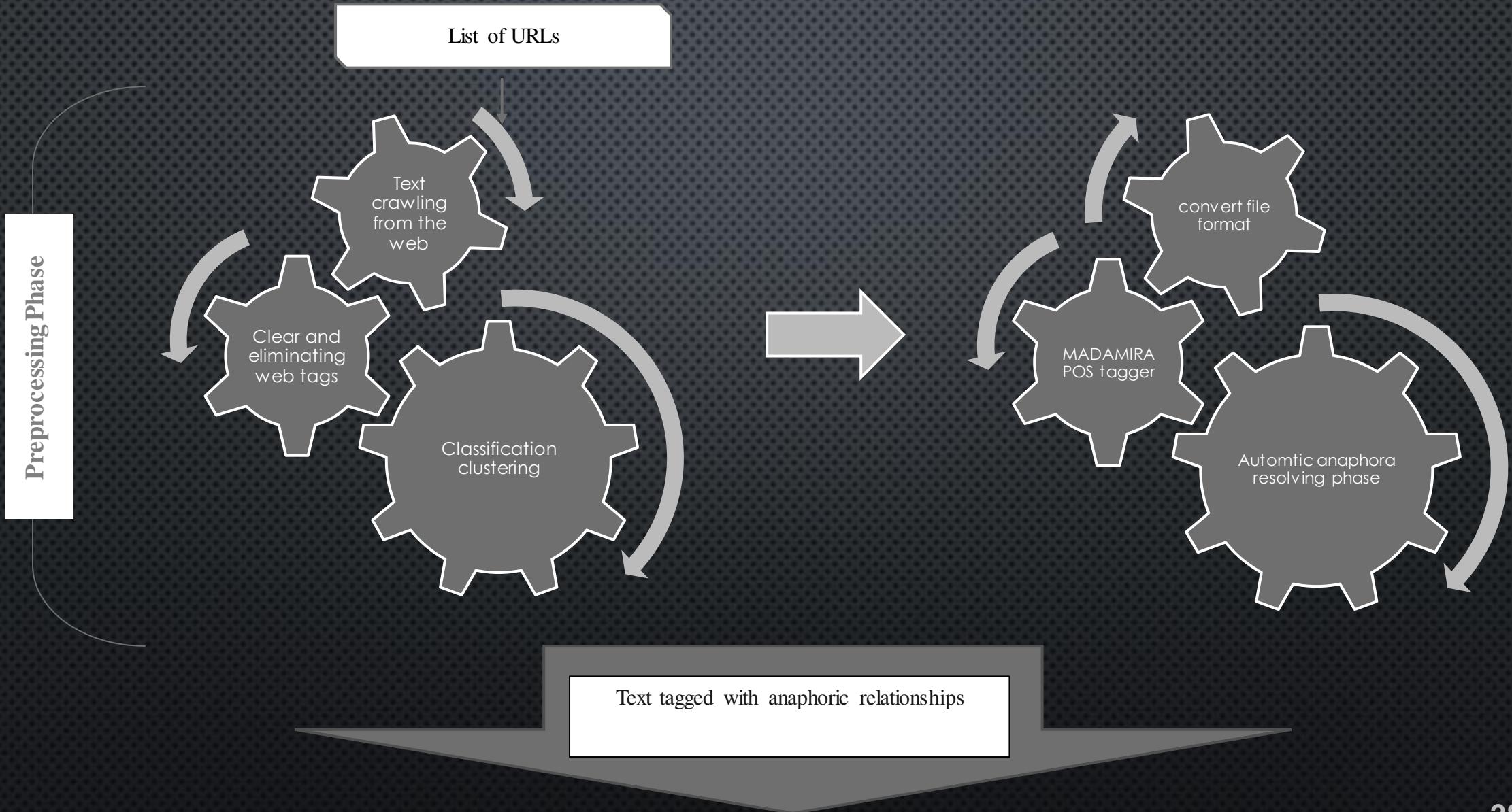
PREPROCESSING  
PHASE.



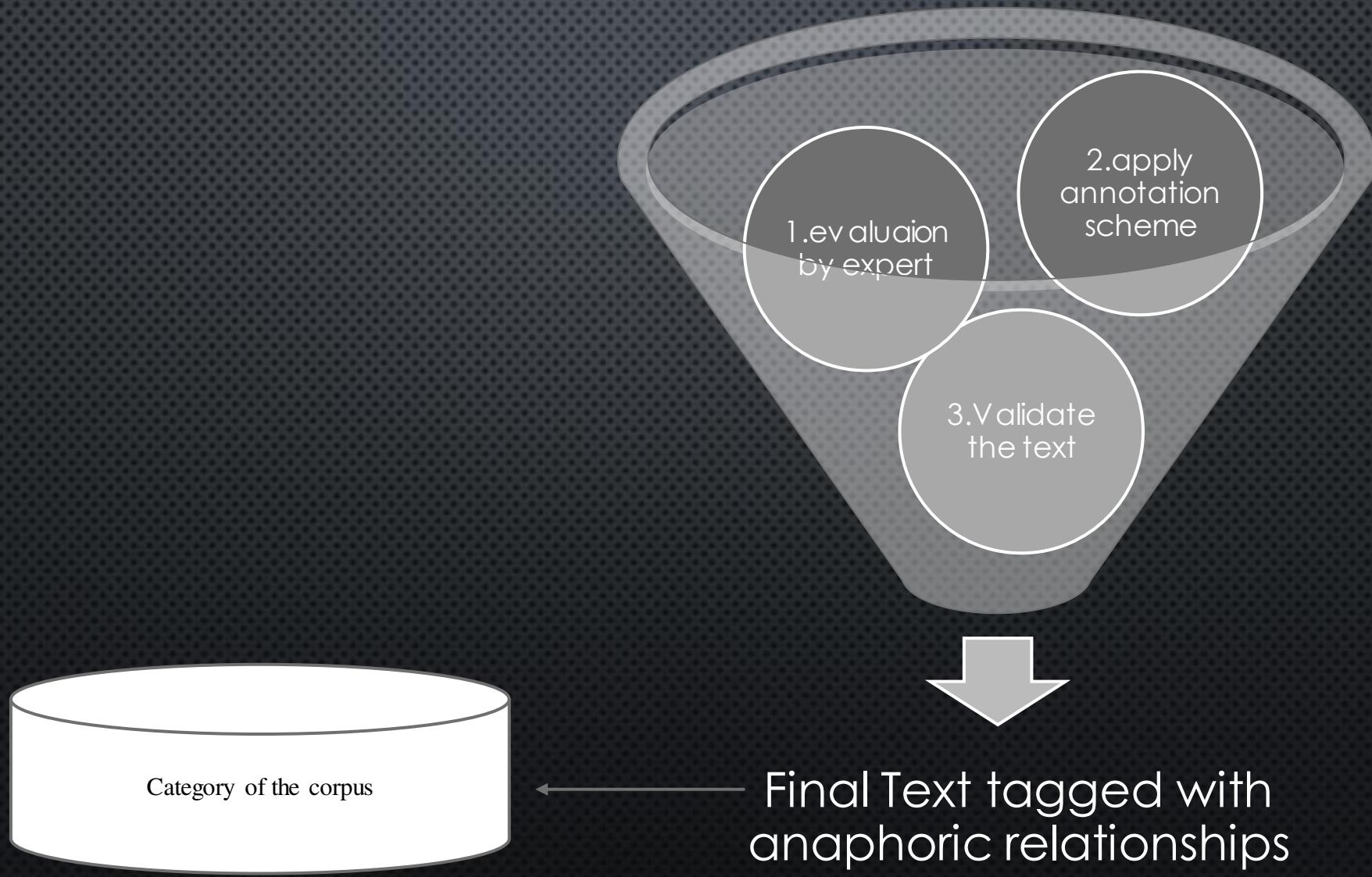
PROCESSING PHASE.



EVALUATION PHASE.

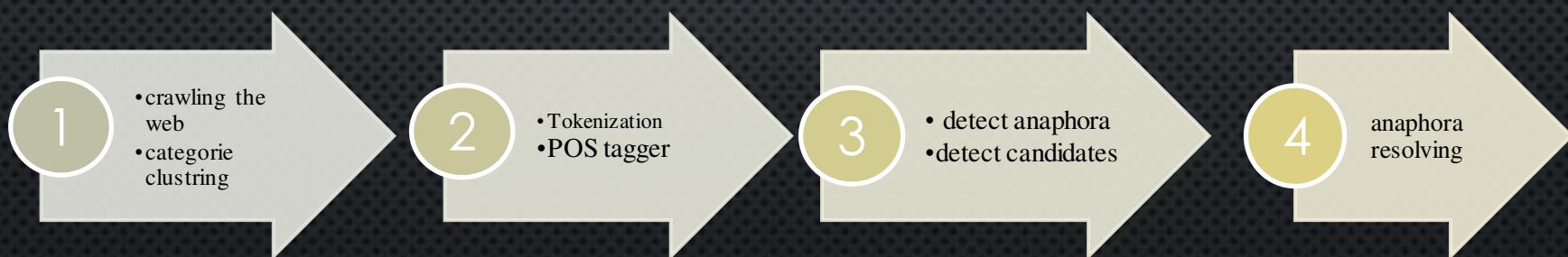


## Processing Phase



# PRE-PROCESSING PHASE:

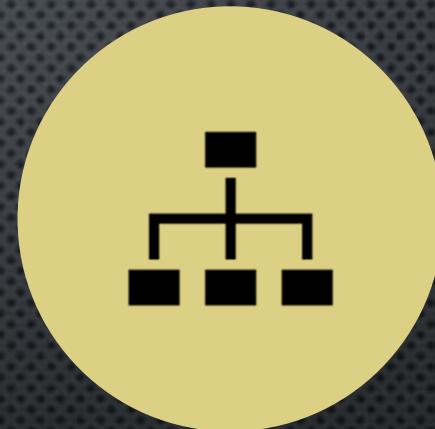
THIS PHASE IS GOING THROUGH 4 SUB-PHASES AND USES SEVERAL MODULE THROUGH THAT.



# CONSTRUCTION OF A RAW CORPUS



CRAWLING TEXT FROM THE  
WEB MODULE.



CATEGORIES  
CLASSIFICATION MODULE.

## CONSTRUCTION OF A RAW CORPUS (CONT.)

- **CRAWLING FROM THE WEB MODULE:** 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.
- **CATEGORIES CLASSIFICATION MODULE:** THE CORPUS COMPOSED OF ARTICLES IN DIFFERENT FIELDS SUCH AS: CULTURE, POLITICS, ECONOMY AND THE LAST CATEGORY ONE MISCELLANY.

# TEXT PROCESSING



TOKENIZATION  
MODULE.



FORMAT CONVERSION  
MODULE.



POS TAGGER  
MODULE.

# TEXT PROCESSING (CONT.)

- *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.
- *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. WE DEVELOPED PYTHON SCRIPT TO DO THAT.
- *POS TAGGER MODULE:* THE PRIMARY TARGET OF PART-OF-SPEECH (POS) TAGGING IS TO IDENTIFY THE GRAMMATICAL GROUP OF A GIVEN 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

# ELEMENT DETECTION



ANAPHORA  
IDENTIFICATION MODULE.



CANDIDATES  
IDENTIFICATION MODULE.

العنوان : تشيلي تحل مكان البرازيل في استضافة قمة الأمم المتحدة للمناخ عام 2019 تحل تشيلي مكان البرازيل في استضافة أعمال قمة الأمم المتحدة للمناخ العام المقبل بعد اعلان برازيليا انسحابها من تنظيم القمة

. وفق **ما** تبلغه المجتمعون في مؤتمر المناخ في بولندا الجمعة

وتحلت البرازيل الشهر الماضي عن استضافة قمة كوب 25 **التي** كانت مقررة أواخر عام 2019

بحجة أن تنظيمها مكلف جدا على الخزينة

وكان الرئيس المنتخب جايير بولسونارو قد أعلن خلال حملته الانتخابية انه ينظر في سحب بلاده من اتفاق باريس التاريخي للمناخ

**ما** أثار قلق أنصار البيئة في بلاده وفي الخارج .

وقالت وزيرة البيئة التشيلية كارولينا شميدت للوفود في مؤتمر الأمم المتحدة للمناخ كوب 24 في مدينة كاتوفيتسي البولندية إن بلادها سوف تعرض استضافة كوب 25

. أما كوستا ريكا فسوف تستضيف الاجتماعات التحضيرية لقمة **التي** تعقد قبل أشهر من المناسبة الرئيسية

ويجتمع ممثلو 200 دولة في بولندا لمناقشة خطة شاملة لسبل تنفيذ تعهدات اتفاق باريس **الذي** يهدف الى الحد من ارتفاع حرارة الارض وتجنب التأثيرات السيئة للتغير المناخي .

تشيلي تحل مكان البرازيل في استضافة قمة الأمم المتحدة للمناخ عام 2019 تحل تشيلي غافة أعمال قمة الأمم المتحدة للمناخ العام المقبل بعد اعلان برازيليا انسحابها من ا

وفق ما تبلغه المجتمعون في مؤتمر المناخ في بولندا الجمعة

وخلت البرازيل **الشهر** الماضي عن استضافة قمة كوب 25 التي كانت مقررة أواخر عام

**حجۃ أن تنظیمها مکلف جداً على الخزینة**

**المنتخب** جاير بولسونارو قد أعلن خلال حملته الانتخابية انه ينوي في سحب بلاده من اتفاق

ما أثار قلق أنصار البيئة في بلاده وفي الخارج

**البيئة التشيلية** كارولينا شميدت **لوفود** في مؤتمر الأمم المتحدة للمناخ كوب 24 فـ  
البولندية إن **بلادها** سوف ت تعرض استضافة كوب 25

الجتماعات التحضيرية للقمة التي تعقد قبل أشهر من المناء

و 200 دولة في بولندا لمناقشة خطة شاملة لسبل تنفيذ تعهدات اتفاق باريس الذي يهدى حرارة الارض وتجنب التأثيرات السيئة للتغير المناخي.

**تشيلي تحل مكان البرازيل في استضافة قمة الأمم المتحدة للمناخ عام 2019** **تحل تشيلي مكان البرازيل في استضافة قمة الأمم المتحدة للمناخ عام 2019**

· وفق ما تبلغه المجتمعون في مؤتمر المناخ في بولندا الجمعة

**وتخلت البرازيل الشهر الماضي عن استضافة قمة كوب 25 التي كانت مقررة أواخر**

حججة أن تنظيمها مكلف جدا على الخزينة

كان الرئيس المنتخب جاير بولسونارو قد أعلن خلال حملته الانتخابية انه ينظر في سحب بلاده من

**ما أثار** قلق أنصار البيئة في بلاده وفي الخارج

ت وزيرة البيئة التشيلية كارولينا شميدت للوفود في مؤتمر الأمم المتحدة للمناخ كوب 24 في مدينة ك

تعرّض استضافة كوب 25

. أما كوستاريكا فسوف تستضيف الاجتماعات التحضيرية للقمة التي تعقد قبل أشهر م

٤ ممثلو 200 دولة في بولندا لمناقشة خطة شاملة لسبل تنفيذ تعهدات اتفاق باريس الذي يهدف الى التأثيرات السلبية للتغير المناخي.

# ANAPHORA RESOLUTION

ANAPHORA RESOLVING MODULE: RESOLVING MODULE IS THE MAIN PART OF THE ANAPHORA RESOLUTION PROCESS. THIS MODULE FOCUSES 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 LINGUISTIC RULES AND THEIR RESPECTIVE SCORES.

LINGUISTIC RULES	DESCRIPTION
<b>Definiteness</b>	A score of 1 is given if an NP is definite and of 0 if not.
<b>Recency</b>	A score of 1 is assigned to the recency NP to the anaphora and 0 if not.
<b>Referential Distance</b>	A score of 2 is assigned to NPs in the current sentence and further than those are given 0.
<b>First Noun Phrases</b>	A score of 1 is issued to the first NP of each sentence and 0 if not.
<b>NPs in the title</b>	A score of 1 is issued to the existing NP in title and 0 if not
<b>Grammatical function</b>	Scores of 1 are given to an NP that has the same syntactic structure as the anaphora and 0 if not.
<b>Frequency of NP in text</b>	A score of 2 is assigned to the most frequent NP in text and 0 if not.

File Edit View Format

Paragraph B I

العنوان : تشيلي تحل مكان البرازيل في استضافة قمة<sup>15</sup> الأمم المتحدة للمناخ عام 2019 تحل تشيلي مكان البرازيل في استضافة أعمال قمة<sup>15</sup> الأمم المتحدة للمناخ العام المقبل بعد اعلان برازيليا انسحابها من تنظيم<sup>23</sup> الـقمة<sup>15</sup> وفق ما<sup>23</sup> تبلغه المجتمعون في مؤتمر المناخ في بولندا الجمعة . وتخليت البرازيل الشهر الماضي عن استضافة<sup>33</sup> قمة<sup>34</sup> كوب 25 التي<sup>34</sup> كانت مقررة أواخر عام 2019 بحجة أن تنظيمها مكلف جدا على الخزينة وكان الرئيس المنتخب<sup>45</sup> جايير بولسونارو قد أعلن خلال حملته الانتخابية انه ينطر في سحب بلاده<sup>49</sup> من اتفاق باريس التاريخي للمناخ<sup>53</sup> ما<sup>53</sup> أثار قلق أنصار البيئة في بلاده وفي الخارج . وقالت وزيرة البيئة التشيلية كارولينا شميدت للوفود في مؤتمر الأمم المتحدة للمناخ<sup>69</sup> كاتوفيتسه البولندية إن بلادها سوف تعرض استضافة كوب 25 أما كوستا ريكا فسوف تستضيف الاجتماعات التحضيرية كوب 24 في مدينة<sup>76</sup> التي<sup>76</sup> تعقد قبل أشهر من المناسبة الرئيسية . ويجتمع ممثلو 200 دولة في بولندا لمناقشة خطة شاملة لسبل تنفيذ تعهدات اتفاق للقمة<sup>90</sup> الذي<sup>90</sup> يهدف الى الحد من ارتفاع حرارة الارض وتجنب التأثيرات السيئة للتغير المناخي . باريس

Powered by Tiny

# PROCESSING PHASE



AUTOMATIC CORPUS  
ANNOTATION.



EXPERT PROCESSING.

# 4. EVALUATION

# DEVELOPMENT ENVIRONMENT



**Development language:**  
**python.**



**IDE:** PyCharm.



**Machine features:**

Processor: i5-4210H.  
RAM: 12Gb.  
Operating system: Windows 10/ Mint Linux.

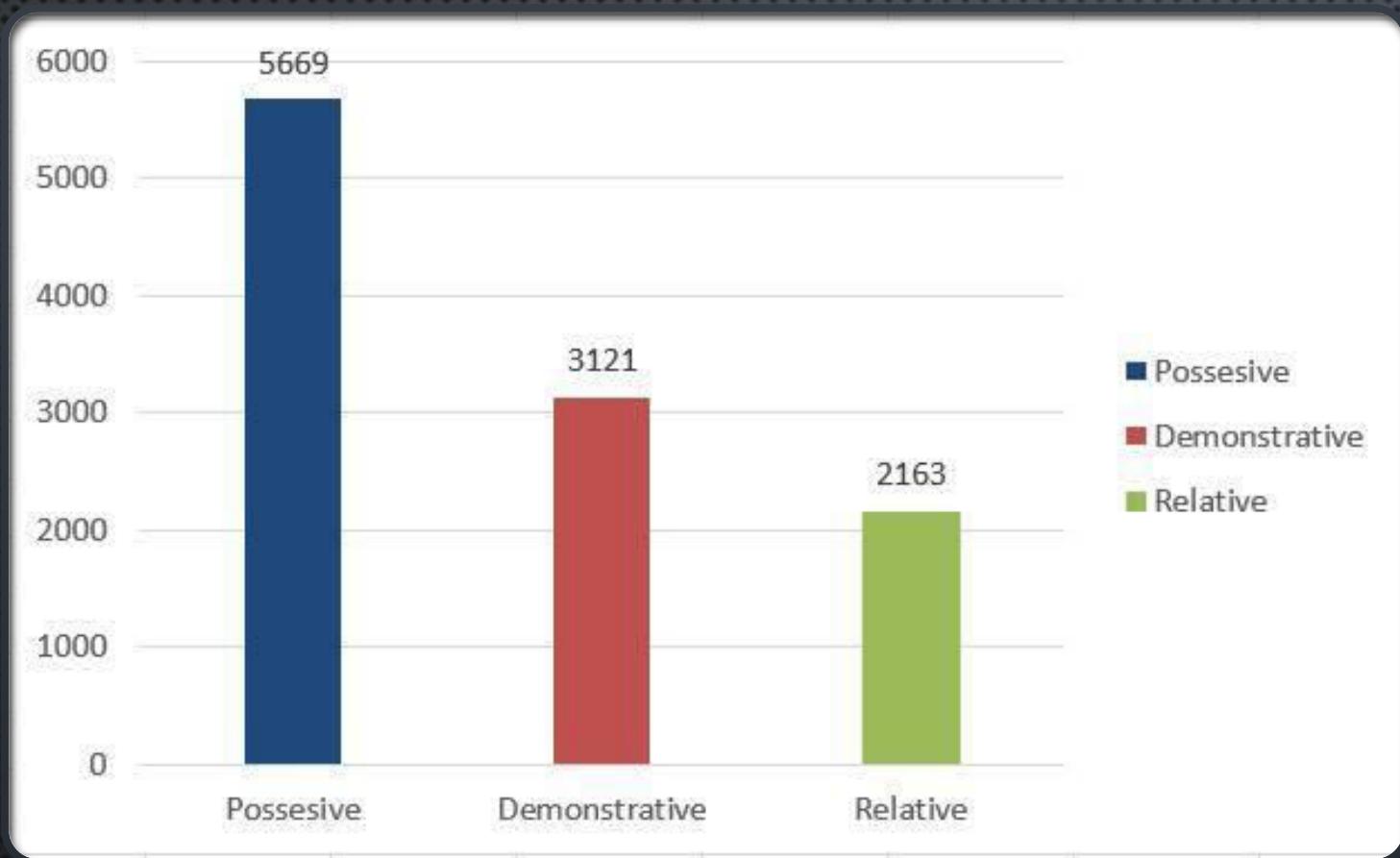
# ARABIC ANAPHORIC ANNOTATED CORPUS (A3C) IN NUMBERS

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

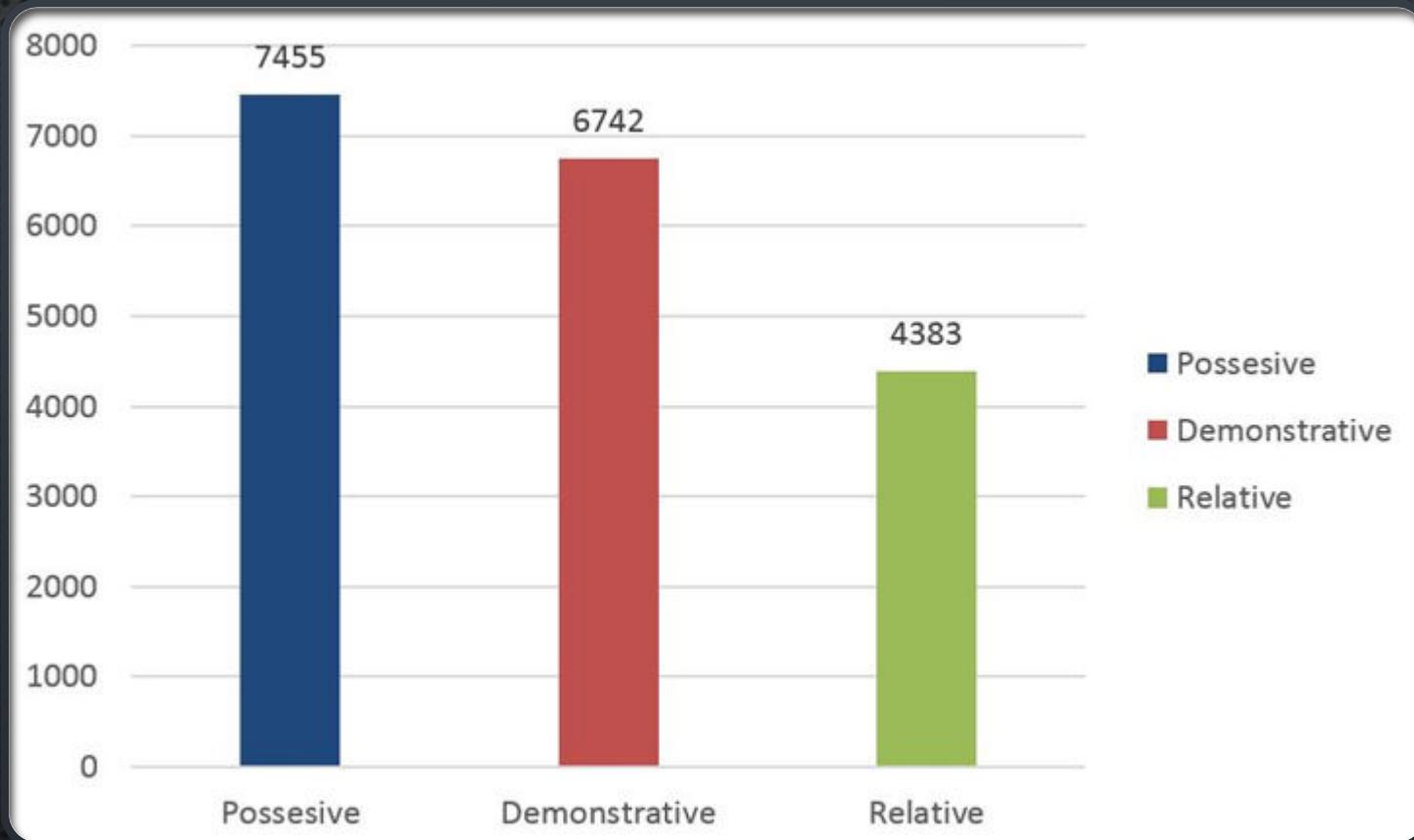
## A3C IN NUMBERS (CONT.)

Category	Pronominal anaphora count	Possessive	Demonstrative	Relative
Economy	18580	7455	6742	4383
Education	13210	5696	3121	2163
Politics	28540	13539	9276	5725
Sport	10953	6437	3572	3201
Miscellany	15069	6597	4853	3619

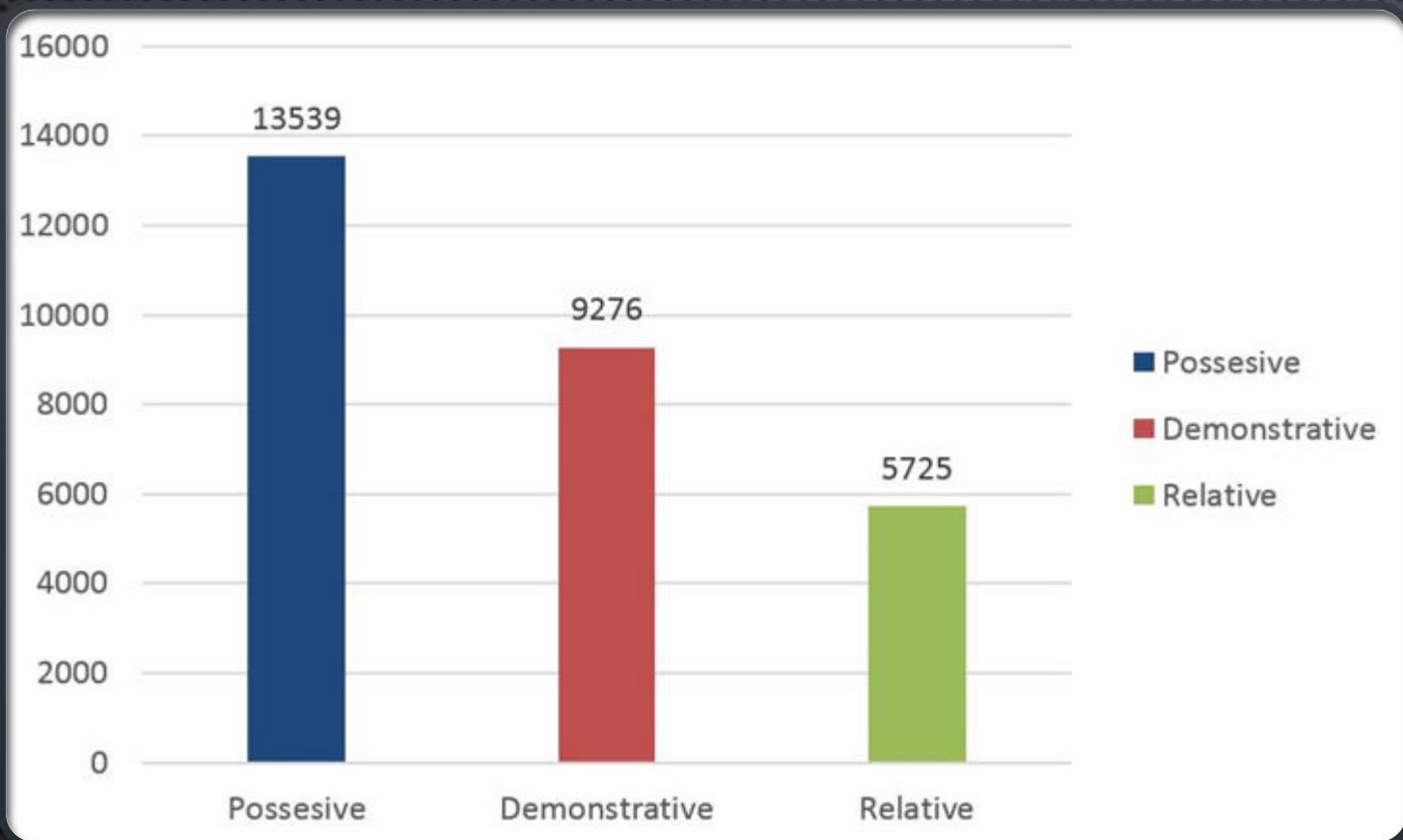
# EDUCATION CATEGORY PRONOMINAL ANAPHORA STATISTICS.



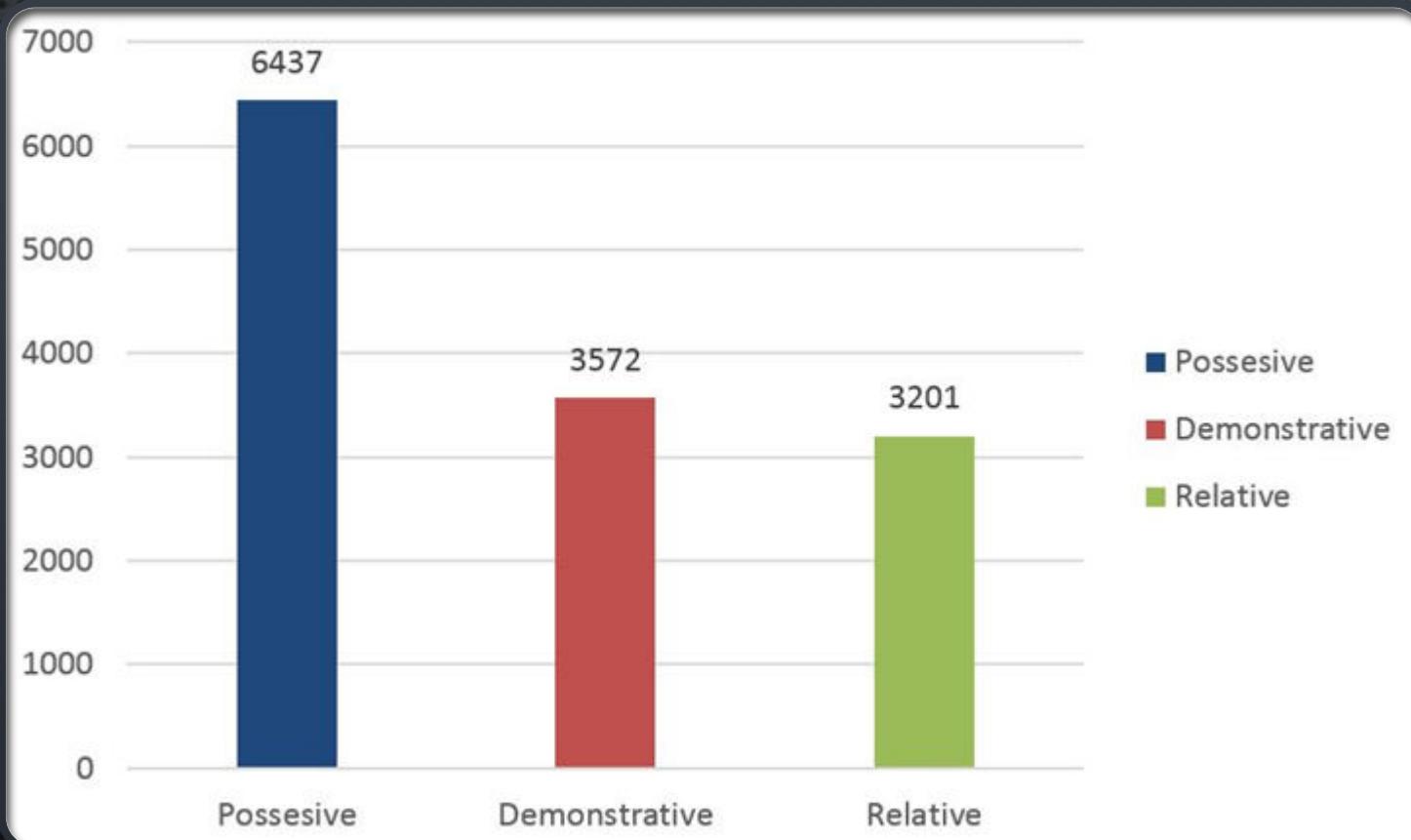
# ECONOMIC CATEGORY PRONOMINAL ANAPHORA STATISTICS.



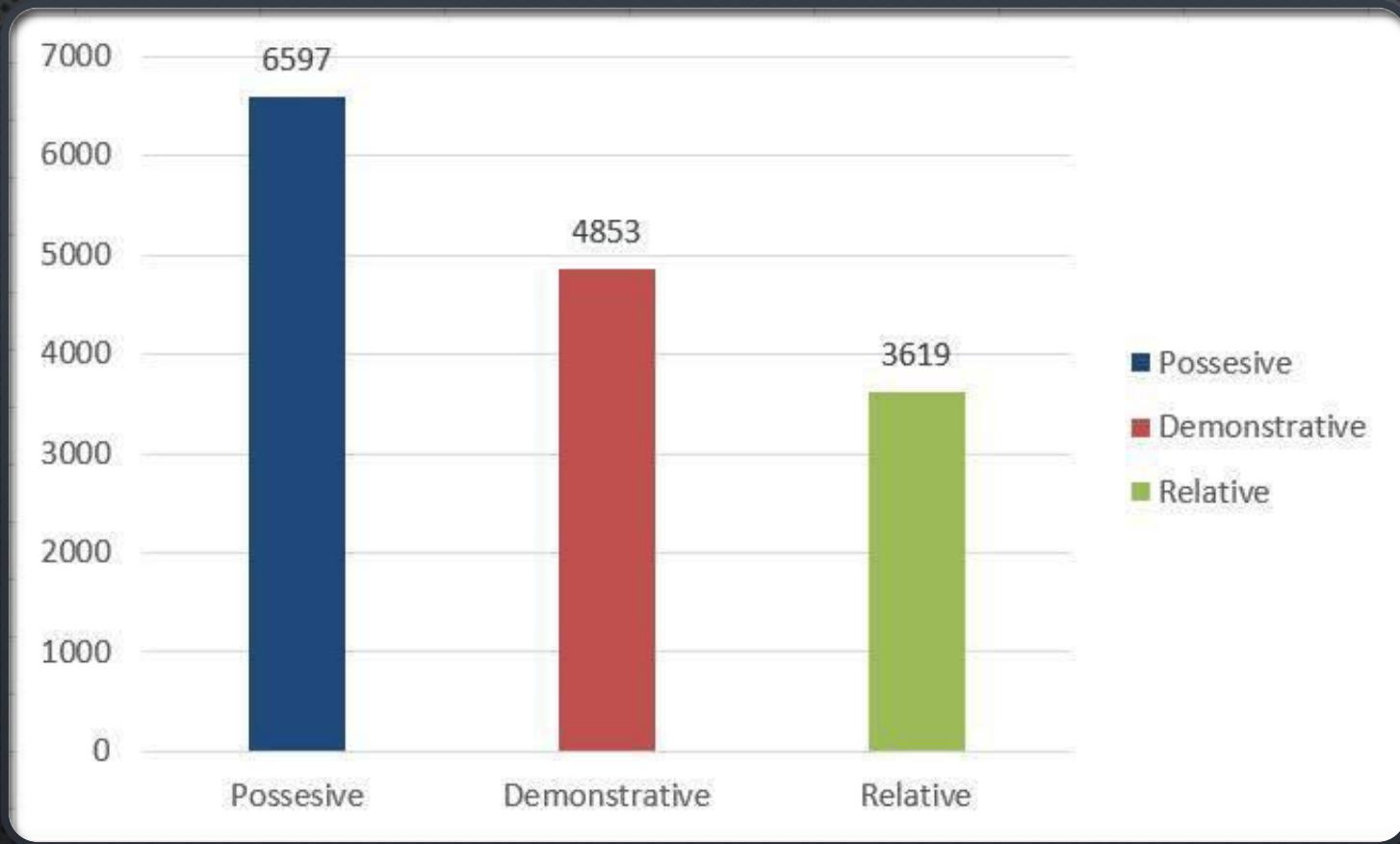
# POLITICS CATEGORY PRONOMINAL ANAPHORA STATISTICS.



# SPORTS CATEGORY PRONOMINAL ANAPHORA STATISTICS.



# MISCELLANY CATEGORY PRONOMINAL ANAPHORA STATISTICS.





ARABIC  
ANAPHORIC  
ANNOTATING  
TOOL (A<sup>3</sup>T)  
INTERFACE

*Original Text*

إصابة 28 شخصاً إثر خروج "ترام" عن القضبان في البرتغال

أصيب 28 شخصاً بجروح في العاصمة البرتغالية لشبونة مساء أمس، إثر خروج ترام عن [القضبان]. وأظهرت صور نشرت على الإنترنت الترام مقلوحاً على جانبه على ناصية شارع، فضلاً عن تجمع الكثير من مركبات الطوارئ. وقالت الشرطة إن الحادث، الذي وقع في منطقة لابا بالعاصمة، لم يسفر عن وقوع إصابات خطيرة، وأعلنت شركة كاريس المشغلة للترام عن فتح تحقيق في سبب الخروج عن القضبان، بينما أفاد شهود عيان، أن السبب ربما [.يعود إلى تعطل مكابح الترام]

ORIGINAL  
TEXT INPUT.

Choose file

Browse

Add

✖ Re-format

*Original Text*

إصابة 28 شخصاً إثر خروج ترام عن القضبان في البرتغال

أصيب 28 شخصاً بجروح في العاصمة البرتغالية لشبونة مساء أمس

إثر خروج ترام عن القضبان

وأظهرت صور نشرت على الإنترنت الترام مقلوباً على جانبه على ناصية شارع

فضلاً عن تجمع الكثير من مركبات الطوارئ

وقالت الشرطة إن الحادث

الذي وقع في منطقة لابا بالعاصمة

# ORIGINAL TEXT SENTENCE SEGMENTATION.

# MADAMIRA PART-OF- SPEECH TAGGING.

Annotated Text

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
  <madamira_output
    xmlns="urn:edu:columbia:ccls.madamira.configuration:0.1">
    <out_doc id="ExampleDocument">
      <out_seg id="SENT1">
        <segment_info>
          إصابة 28 شخصاً إثر خروج ترام عن القضبان في البرتغال
        </segment_info>
        <word_info>
          <word id="0" word="إصابة">
            <svm_prediction>
              <morph_feature_set diac="إصابة!" pos="noun" vox="na"
                mod="na" gen="f" num="s" stt="c"/>
            </svm_prediction>
          </word>
        </word_info>
      </out_seg>
    </out_doc>
  </madamira_output>
```

Apply annotation scheme    apply the correction    Save the Text

*Annotated Text*

إصابة 28 شخصاً إثر خروج ترام عن القضبان في البرتغال  
أصيب 28

<antecedent id="9" name="شخصاً"/>  
يخرج في العاصمة البرتغالية لشبونة مساء أمس  
إثر خروج

<antecedent id="3" name=" ترام"/>  
عن القضبان.

وأظهرت صور نشرت على الإنترنت الترام مقلوطاً على جانب

<anaphor id="12" name="هـ" referent\_id="3"/>  
على ناصية شارع،  
فضلاً عن تجمع الكثير من مركبات

<antecedent id="15" name="الطوارئ"/>  
وقالت الشرطة إن

<antecedent id="21" name="الحادث"/>

<anaphor id="22" name=" الذي" referent\_id="21"/>  
ووقع في منطقة لابا بالعاصمة،  
لم يسفر عن وقوع إصابات خطيرة  
أن السبب ربما يعود إلى تعطل مكابح الترام.

*Proposal solution by the system*

*Proposal solutions*

٥	ترام	٣	<b>+79%</b>
٥	شخص	٩	<b>+50%</b>
الذي	الحادث	21	<b>+90%</b>
الذي	الطوارئ	15	<b>+10%</b>

Amine      Mohamed

Apply annotation scheme
apply the correction
Save the Text

# EXPERT VERIFICATION AND CORRECTION INTERFACE.

# 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.

BY DOING THE VERIFICATION ON THE “ANATAR” CORPUS, OUR TOOL ACHIEVED A SUCCESS RATE OF 83.19%.

# 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:

$$\text{Accuracy} = \frac{\text{NUMBER OF CORRECTLY RESOLVE ANAPHORA}}{\text{NUMBER OF ALL ANAPHORAS}}$$

CONCERNING THE VERBAL ANAPHORA, WE ARRIVE AT A SATISFACTORY RESULT WHOSE RATE OF RECOGNITION AVERAGE IS: 57.23%.

# 5. CONCLUSION

# OUTCOME

IN THE AUTOMATIC PROCESSING OF THE ARABIC LANGUAGE, ANAPHORA PLAY AN IMPORTANT ROLE IN UNDERSTANDING TEXTS BY DETERMINING LINKS BETWEEN THE TEXT ENTITIES.

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<sup>3</sup>T, SUCH AS:

- USE A MORE EFFECTIVE TAGGER.
- DEEPENING CONSTRAINTS: DEFINING OTHER OPERATING RULES.
- EXPLOIT OTHER ANAPHORIC RESOLUTION APPROACHES SUCH AS THE MACHINE LEARNING APPROACH.
- SOLVE OTHER TYPES OF ANAPHORA (LEXICAL AND COMPARATIVE)

AS FOR OUR ARABIC ANAPHORIC ANNOTATED CORPUS (A<sup>3</sup>C), WE MAY LOOK FORWARD FOR A BETTER CORRECTION AND INCREASING IT'S SIZE ALLOWING MORE EFFECTIVENESS IN OTHER RESEARCHES.



THANKS



QUESTIONS?