The Fifteenth Conference on Language Engineering (ESOLEC'2015)

December 9-10, 2015

Organized by

Egyptian Society of Language Engineering (ESOLE)

Under the Auspices of

PROF. DR. HUSSEIN EISSA
President of Ain Shams University

PROF. DR. MOHAMED AYMAN ASHOUR
Dean, Faculty of Engineering, Ain Shams University

Conference Chairperson
PROF. DR. M. A. R. GHONAIMY

Conference Cochairperson
PROF. DR. SALWA ELRAMLY

Faculty of Engineering –Ain Shams University
Cairo, Egypt

http://esole-eg.org

Conference Chairman

Conference Sponsors

Prof. Dr. M. A. R. Ghonaimy

Technical Program Committee:

Prof. Taghrid Anber, **Egypt**
Prof. I. Abdel Ghaffar, **Egypt**
Prof. M. Ghaly, **Egypt**
Prof. M. Z. Abdel Mageed, **Egypt**
Prof. Khalid Choukri, ELDA, **France**
Prof. Nadia Hegazy, **Egypt**
Prof. Christopher Ciri, LDC, **U.S.A**
Prof. Mona T. Diab, Stanford U., **U.S.A**
Prof. Ayman ElDessouki, **Egypt**
Prof. Afaf AbdelFattah, **Egypt**
Prof. Y. ElGamal**, Egypt**
Prof. M. Elhamalaway**, Egypt**
Prof. S. Elramly, **Egypt**
Prof. H. Elshishiny**, Egypt**
Prof. A. A. Fahmy**, Egypt**
Prof. I. Farag, **Egypt**
Prof. Magdi Fikry, **Egypt**
Prof. Wafaa Kamel, **Egypt**
Prof. S. Krauwer, **Netherlands**
Prof. Bente Maegaard, CST, **Denmark**
Prof. A. H. Moussa, **Egypt**
Prof. M. Nagy**, Egypt**
Prof. A. Rafae, **Egypt**
Prof. Mohsen Rashwan**, Egypt**
Prof. H. I. Shaheen**, Egypt**
Prof. S. I. Shaheen**, Egypt**
Prof. Hassanin M. AL-Barhamtoshy**, Egypt**
Prof. M. F. Tolba, **Egypt**
Dr. Tarik F. Himdi**, Saudi Arabia**

Organizing Committee

Prof. I. Farag
Prof. S. Elramly
Prof. M. Elhamalawy
Prof. Sameh El Ansary
Dr. Mona Zakaria
Prof. Hani Mahdi
Prof. M. Z. Abdelmegeed
Prof. H. Shahein
Dr. A. Passant Elkafrawy
Dr. Bassant Abdelhamid

Conference Secretary General

Prof. Dr. Salwa Elramly

The Fifteenth Conference on Language Engineering
Final Program

Wednesday 9 December 2015

9.00 - 10.00 Registration

10.00 - 10.30 Opening Session: Seminar Room, Biblioteque Building

10.30 - 11.30 **Session 1:** Seminar Room: **Invited Paper 1: Syntax, Semantics and Grammar**

Chairman: Prof. Dr. Ibrahim Farag

A Tutorial on Sentence Semantics Using Lambek Pregroup Grammar and Categorical Quantum Protocols

Prof. M. Adeeb Ghonaimy

Professor Emeritus, Faculty of Engineering, Ain Shams University, Cairo, Egypt.

11.30 - 12.30 Coffee break (Conference Center – Main Building)

12.30 - 13.30 **Session 2** : Seminar Room: **Invited Papers: Natural Language Analysis**

Chairman: Prof. Dr. Nadia Hegazy

1. **BASMA: BibAlex Standard Arabic Morphological Analyzer**
   Sameh Alansary
   Arabic Computational Linguistics Center, Bibliotheca Alexandrina
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University
2. **Part-of-Speech Tagging and Disambiguation for Arabic Language Understanding**
   Sameh Alansary
   Arabic Computational Linguistics Center, Bibliotheca Alexandrina
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University

13.30 - 15.00 **Session 3: Room A: Ontology**

Chairman: Prof. Dr. Aly Aly Fahmy

1. **Developing an Approach for Solving Ambiguity in Requirements Specification to UML Conversion**
   Somaia Osama, Safia Abbas, Mostafa Aref
   Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt
2. **Case Based Reasoning of Semantic Knowledge on Medical System**
   Passent ElKafrawy[*], Rania A. Mohamed[**]
   [*]Mathematics and CS Department, Faculty of Science, Menofia University, Shebin Elkom, Menofia, Egypt
   [**]Faculty Computer Science, Modern University for Technology & Information, Cairo, Egypt
3. **Automatic Part-of-Speech Tagging of Arabic-English Dictionary Senses through WordNet**
   Diaa M. Fayed[*], Aly A. Fahmy[*], Mohsen A. Rashwan[**], Wafaa K. Fayed[***]
   [*]Computer Science, Faculty of Computers and Information, Giza, Egypt
   [**]EECE, Faculty of Engineering, Giza, Egypt
   [***]Arabic Language and Literatures, Faculty of Arts, Giza, Egypt

13.30 - 15.00 **<u>Session 4:</u> Room B: Corpora**

Chairman: Prof. Dr. M. Mohsen Rashwan

1. **كيفَ نبني مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًّا للُّغة العربيَّة بطريقة نصف آليَّة؟**
   المُعتزّ بالله السَّعيد
   كُلِّيَّة دار العُلوم، جامعة القاهرة، مصر
2. **Building a POS-Annotated Corpus for Egyptian Children**
   Heba Salama, Sameh Alansary
   Phonetics and linguistics Department, Faculty of Arts Alexandria University
3. **Discourse Tagging of Political Speeches: A Corpus-based Study**
   Marwa Adel Abu El Wafa*, Sameh Alansary**, Shadia El Soussi***
   **Language and Translation Department, College of Language and Communication, Institute for Language Studies, Arab Academy for Science, Technology and Maritime Transport, Miami, Alexandria, Egypt*
   ***Phonetic and Linguistics Department, Faculty of Arts, University of Alexandria,*
   ElShatby, Alexandria, Egypt
   Bibliotheca Alexandrina, Alexandria, Egypt
   ****Institute of Applied Linguistics, Faculty of Arts, University of Alexandria, ElShatby, Alexandria, Egypt*

15.00 - 16.00 Lunch (Conference Center – Main Building)

<u>Thursday 10 December 2015</u>

10.00 - 11.00 **<u>Session 5</u>: Room A: Social Nets**

Chairman: Prof. Dr. M. Younis Elhamalawy

1. **Classification of Text Images on Social Network Using Linguistic and Behavioral Features**
   Ahmad M. Abd Al-Aziz*, Mervat Gheith**, Ahmed Sharf Eldien***
   **The British University in Egypt (BUE), Cairo, Egypt*
   ***Institute of Studies and Statistical Researches, Cairo University, Cairo, Egypt*
   ****Faculty of Computers and Information, Helwan University, Helwan, Egypt*
2. **NLP in Social Media: An Overview**
   Soha S. Ibrahim, Mostafa M. Aref
   Department of Computer Science, Faculty of Computer Science and information System, Cairo, Egypt

11.00 - 11.30 **<u>Session 6:</u> Room A: Speaker recognition**

Chairman: Prof. Dr. Sameh Al-Ansary

1. **Speaker Identification Based on Temporal Parameters**
   Eman M. Yousri, Mervat Fashal
   Phonetics & Linguistics Dep., Faculty of Arts, University of Alexandria, Alexandria, Egypt

11.30 - 12.00 **<u>Session 7:</u> Room A: Word Sense Disambiguation**

Chairman: Prof. Dr. Sameh Al-Ansary

1. **معالجة الالتباس الدلالي في نتائج تحليل المحلل الصرفي العربي تيم باكولتر**
   أحمد عبد الغني، سامح الأنصاري
   قسم اللسانيات والصوتيات، كلية الآداب، جامعة الإسكندرية

12.00 - 13.00 Coffee Break (Conference Center – Main Building)

13.00 - 14.00 **Session 8**: **Room A: Syntax, Semantics and Grammar**
Chairman : Prof. Dr. M. Hany Kamal

1. **Semantic-Based Approaches for XML Summarization**
Hassan A. Elmadany, Marco Alfonse, Mostafa Aref
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

2. **Automatic Diacritization for Modern Standard Arabic**
Amany Fashwan, Sameh Alansary
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
3. **Syntax-Semantics Classification of Arabic Verbs for Semantic Annotation**
Israa Elhosiny, Sameh Alansary
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

13.00 - 14.00 **Session 9**: **Room B: NLP for Information Retrieval**
Chairman: Prof. Dr. M. Fahmi Tolba

1. **Text mining model using a hybrid of SOM and LSI Techniques**
Abdelfattah ELsharkawi [*], Ali Rashed [**], Hosam Eldin Fawzan[*]
[*]*Department of Systems and Computer Engineering, Al-Azhar University, Egypt*
[**]*Department of Electrical and Computer Engineering, Faculty of Engineering Science, Sinai University, Egypt*
2. **CMET: A Semantic Framework for Comparing and Merging Entities and Terms and its Application in Answer Selection**
Mahmoud A. Wahdan, Safia Abbas, Mostafa Aref
Computer Science Department, Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

3. **Graph Matching Based Technique for Words Segmentation in Arabic Sign Language**
A. S. Elons, M. F. Tolba
Scientific Computing Department, Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

14.00 - 15.00 **Session 10**: **Room A: Round Table**
Chairman: Prof. Dr. Nadia Hegazy

The Value of Arabic Language Engineering in the conflict of the World

15.00 - 16.00 Lunch (Conference Center – Main Building)

16.00 - 16.30 Closing Session

Program Summary

	Day	Time	Location	Subject	Chairman
Session 1	Wednesday	10:30 – 11:30	Seminar Room	Syntax, Semantics and Grammar	Prof. Dr. Ibrahim Farag
Coffee Break	Wednesday	11:30 – 12:30	Conference Center - Main building		
Session 2	Wednesday	12:30 – 13:30	Seminar Room	Natural Language Analysis	Prof. Dr. Nadia Hegazy
Session 3	Wednesday	13:30 - 15:00	Room A	Ontology	Prof. Dr. Aly Aly Fahmy
Session 4	Wednesday	13:30 - 15:00	Room B	Corpora	Prof. Dr. M. Mohsen Rashwan
Lunch	Wednesday	15:00 - 16:00	Conference Center - Main building		
Session 5	Thursday	10:00 – 11:00	Room A	Social Nets	Prof. Dr. M. Younis Elhamalawy
Session 6	Thursday	11:00 – 11:30	Room A	Speaker Recognition	Prof. Dr. Sameh Al-Ansary
Session 7	Thursday	11:30 – 12:00	Room A	Word Sense Disambiguation	Prof. Dr. Sameh Al-Ansary
Coffee Break	Thursday	12:00 – 13:00	Conference Center - Main building		
Session 8	Thursday	13:00 – 14:00	Room A	Syntax, Semantics and Grammar	Prof. Dr. M. Hany Kamal
Session 9	Thursday	13:00 – 14:00	Room B	NLP for Information Retrieval	Prof. Dr. M. Fahmi Tolba
Session 10	Thursday	14:00 – 15:00	Room A	Round Table	Prof. Dr. Nadia Hegazy
Lunch	Thursday	15:00- 16:00	Conference Center - Main building		
Closing Session	Thursday	16:00- 16:30	Conference Center - Main building		

Seminar Room: Library building
Room A: Conference Center - Main building
Room B: Conference Center - Main building

الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY OF LANGUAGE ENGINEERING

أعضاء الجمعية من المؤسسات

1- مركز نظم المعلومات – كلية الهندسة – جامعة عين شمس

2- معهد الدراسات والبحوث الإحصائية – جامعة القاهرة

3- مركز الحساب العلمى – جامعة عين شمس

4- الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحرى

5- أكاديمية أخبار اليوم

6- معهد بحوث الإلكترونيات

7- معهد تكنولوجيا المعلومات

8- مكتبة الإسكندرية

9- المعهد القومى للاتصالات (NTI)

10- الشركة الهندسية لتطوير نظم الحاسبات (RDI)

11- الهيئة القومية للاستشعار من بعد و علوم الفضاء

12- كلية الحاسبات و المعلومات جامعة قناة السويس

13- دار التأصيل للبحث و الترجمة

أهداف الجمعية

1- الاهتمام بمجال هندسة اللغويات مع التركيز على اللغة العربية بصفتها لغتنا القومية والتركيز على قواعد البيانات المعجمية وصرفها ونحوها ودلالتها بهدف الوصول إلى أنظمة ألية لترجمة النصوص من اللغات الأجنبية إلى اللغة العربية والعكس, وكذلك معالجة اللغة المنطوقة والتعرف عليها وتوليدها, ومعالجة الأنماط مع التركيز على اللغة المكتوبة بهدف إدخالها إلى الأجهزة الرقمية.

2- متابعة التطور فى العلوم والمجالات المختصة بهندسة اللغة

3- التعاون مع الجمعيات العلمية المماثلة على المستوى المحلى والقومى والعالمى.

4- إنشاء قواعد بيانات عن البحوث التى سبق نشرها والنتائج التى تم التوصل إليها فى مجال هندسة اللغة بالإضافة إلى المراجع التى يمكن الرجوع إليها سواء فى اللغة العربية أو اللغات الأخرى.

5- إنشاء مجلة علمية دورية للجمعية ذات مستوى عال لنشر البحوث الخاصة بهندسة اللغة وكذلك بعض النشرات الدورية الإعلامية الأخرى بعد موافقة الجهات المختصة.

6- عقد ندوات لرفع الوعى فى مجال هندسة اللغة

7- تنظيم دورات تدريبية يستعان فيها بالمتخصصين وتتاح لكل من يهمه الموضوع. وذلك من أجل تحسين أداء المشتغلين فى البحث لخلق لغة مشتركة للتفاهم بين الأعضاء

8- إنشاء مكتبة تتاح للمهتمين بالموضوع تشمل المراجع وأدوات البحث من برامج وخلافه.

9- خلق مجال للتعاون وتبادل المعلومات وذلك عن طريق تهيئة الفرصة لعمل بحوث مشتركة بين المشتغلين فى نفس الموضوعات.

10- تقييم المنتجات التجارية أو البحثية والتى تتعرض لعملية ميكنة اللغة.

11- رصد الجوائز التشجيعية للجهود المتميزة فى مجالات هندسة اللغة.

12- إنشاء فروع للجمعية فى المحافظات.

الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

المؤتمر الخامس عشر لهندسة اللغة

9-10 ديسمبر 2015

جمهورية مصر العربية-القاهرة

ينظم المؤتمر

الجمعية المصرية لهندسة اللغة

تحت رعاية

الأستاذ الدكتور/ حسين عيسى

رئيس جامعة عين شمس

الأستاذ الدكتور/ محمد أيمن عاشور

عميد كلية الهندسة - جامعة عين شمس

رئيس المؤتمر

الأستاذ الدكتور/ محمد أديب رياض غنيمى

مقرر المؤتمر

الأستاذ الدكتور / سلوى حسين الرملى

مكان عقد المؤتمر : كلية الهندسة - جامعة عين شمس

Table of Contents

Page

I. Syntax, Semantics and Grammar

1. **A Tutorial on Sentence Semantics Using Lambek Pregroup Grammar and Categorical Quantum Protocols** 1

Prof. M. Adeeb Ghonaimy
Professor Emeritus, Faculty of Engineering, Ain Shams University, Cairo, Egypt

2. **Semantic-Based Approaches for XML Summarization** 13

Hassan A. Elmadany, Marco Alfonse, Mostafa Aref
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

3. **Syntax-Semantics Classification of Arabic Verbs for Semantic Annotation** 18

Israa Elhosiny, Sameh Alansary
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

4. **Automatic Diacritization for Modern Standard Arabic** 32

Amany Fashwan, Sameh Alansary
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

II. NLP for Information Retrieval

5. **Text mining model using a hybrid of SOM and LSI Techniques** 40

Abdelfattah ELsharkawi [*], Ali Rashed [**], Hosam Eldin Fawzan[*]
[*]*Department of Systems and Computer Engineering, Al-Azhar University, Egypt*
[**]*Department of Electrical and Computer Engineering, Faculty of Engineering Science, Sinai University, Egypt*

6. **CMET: A Semantic Framework for Comparing and Merging Entities and Terms and its Application in Answer Selection** 51

Mahmoud A. Wahdan[*1], Safia Abbas[*2], Mostafa Aref[*3]
Computer Science Department, Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

7. **Graph Matching Based Technique for Words Segmentation in Arabic Sign Language** 58

A. S. Elons, M. F. Tolba
Scientific Computing Department, Faculty of Computers and Information Sciences, Ain Shams University, Cairo, Egypt

III. Social Networks

8. **Classification of Text Images on Social Network Using Linguistic and Behavioral Features** 66

Ahmad M. Abd Al-Aziz[*],Mervat Gheith[**], Ahmed SharfEldien[***]
[*]*The British University in Egypt (BUE), Cairo, Egypt*
[**]*Institute of Studies and Statistical Researches, Cairo University, Cairo, Egypt*
[***]*Faculty of Computers and Information, Helwan University, Helwan, Egypt*

9. **NLP in Social Media: An Overview** 74

Soha S. Ibrahim, Mostafa M. Aref
Department of Computer Science, Faculty of Computer Science and information System, Cairo, Egypt

IV. Corpora

10. **كيفَ نبني مُدوَّنةَ لُغَويَّةَ مُوَسَّمةَ تركيبيًّا للّغة العربيَّة بطريقة نصف آليَّة؟** 79

المُعتزّ بالله السَّعيد
كُلِّيَّة دار العُلُوم، جامعة القاهرة، مصر

11. **Discourse Tagging of Political Speeches: A Corpus-based Study** 90

Marwa Adel Abu El Wafa[*], Sameh Alansary[**], Shadia El Soussi[***]
[*]*Language and Translation Department, College of Language and Communication, Institute for Language Studies, Arab Academy for Science, Technology and Maritime Transport, Miami, Alexandria, Egypt*
[**]*Phonetic and Linguistics Department, Faculty of Arts, University of Alexandria ElShatby, Alexandria, Egypt*
Bibliotheca Alexandrina, Alexandria, Egypt
[***]*Institute of Applied Linguistics, Faculty of Arts, University of Alexandria, ElShatby, Alexandria, Egypt*

12. **Building a POS-Annotated Corpus for Egyptian Children** 104

Heba Salama, Sameh Alansary
Phonetics and linguistics Department, Faculty of Arts Alexandria University

V. Ontology

13. **Automatic Part-of-Speech Tagging of Arabic-English Dictionary Senses through WordNet** 120

Diaa M. Fayed[*], Aly A. Fahmy[*], Mohsen A. Rashwan[**], Wafaa K. Fayed[***]
[*]*Computer Science, Faculty of Computers and Information, Giza, Egypt*
[**]*EECE, Faculty of Engineering, Giza, Egypt*
[***]*Arabic Language and Literatures, Faculty of Arts, Giza, Egypt*

14. **Developing an Approach for Solving Ambiguity in Requirements Specification to UML Conversion** 130

Somaia Osama, Safia Abbas, Mostafa Aref
Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

15. **Case Based Reasoning of Semantic Knowledge on Medical System** 135

Passent ElKafrawy[*], Rania A. Mohamed[**]
[*]*Mathematics and CS Department, Faculty of Science, Menofia University, ShebinElkom Menofia, Egypt*
[**]*Faculty Computer Science, Modern University for Technology & Information, Cairo, Egypt*

VI. Natural Language Analysis

16. **BASMA: BibAlex Standard Arabic Morphological Analyzer** 149

Sameh Alansary
Arabic Computational Linguistics Center, Bibliotheca Alexandrina
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University

17. **Part-of-Speech Tagging and Disambiguation for Arabic Language Understanding** 158

Sameh Alansary
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
Bibliotheca Alexandrina, Alexandria, Egypt

VII. Word Sense Disambiguation

18. **معالجة الالتباس الدلالي في نتائج تحليل المحلل الصرفي العربي تيم باكولتر** 172

أحمد عبد الغني، سامح الأنصاري
قسم اللسانيات والصوتيات، كلية الآداب، جامعة الإسكندرية

VIII. Speaker Recognition

Speaker Identification Based on Temporal Parameters 193

Eman M. Yousri, Mervat Fashal

Phonetics & Linguistics Dep., Faculty of Arts, University of Alexandria, Alexandria, Egypt

A Tutorial on Sentence Semantics Using Lambek Pregroup Grammar and Categorical Quantum Protocols

Prof. M. Adeeb Ghonaimy

Professor Emeritus, Faculty of Engineering.

Ain Shams University, Cairo, Egypt

adeebghonaimy@gmail.com

adeeb.ghonaimy@eng.asu.edu.eg

Abstract - Sentence semantics depends mainly on two basic principles: the principle of compositionality [Partee et al, 1990] (sometimes called Frege's principle), and the distributional principle. Briefly, the compositionality principle states that the meaning of a complex expression is a function of the meaning of the parts and the syntactic rules by which they are combined. The distributional principle is that words that occur in a similar context tend to have similar meaning [Turney and Pentel, 2010].

In this tutorial, the syntax used in compositionality is Lambek pregroup grammar [Lambek, 2006]. In order to integrate the above concepts together, categorical quantum protocols were used [Abramsky, and Coecke; 2004] to develop a categorical compositional distributional model of meaning [Grefenstette and Sadrzadeh, 2011] [Coecke, et al, 2010] [Kartsaktis, 2014]. This model is sometimes abbreviated as DisCoCat model. This tutorial gives outline for this model explaining the basic elements of the principles involved including Lambek pregroup grammar and categorical quantum protocols.

1 INTRODUCTION

Sentence semantics depends on two basic principles: the compositionality principle [Partee et al., 1990] and the distributional principle [Turney and Pentel,2010]. The first one is attributed to Frege’s principle that the meaning of a sentence is a function of the meaning of its parts. The second is related to Wittgenstein's philosophy of "meaning in use", where meanings of words can be determined from their context. In 2010, [Coecke et al, 2010] used high-level concepts from categorical quantum protocols to combine compositional and distributional models. The grammar used is Lambek's pregroup grammar [Lambek, 2008] [Lambek, 2006]. An introduction to categories is given by [Coceke and Paquette, 2011]. This combined model is abbreviated as DisCoCat [Grefenstette, and Sadrzadeh, 2011].

In order to give an idea about this model, a number of topics will be presented in the following sections. Section 2 will deal with Lambek pregroup grammar with some definitions and simple examples, section 3 will deal with Categorical Quantum Protocols [Abramsky and Coecke, 2004] with definitions of Category Theory.

Section 4 will deal with compositional and distributional models of meaning. Section 5 discusses the unification of compositional distributional categorical models of meaning (the DisCoCat model) together with experimental support for it [Grefenstette and Sadrzadeh, 2011]. Section 6 is the conclusion.

2 Lambek Pregroup Grammar

Let us fist define the **pregroup**. A pregroup is a partially ordered monoid (a semigroup with unity element). Each element *a* has **a left adjoint** a^l and **a right adjoint** a^r such that

$$a^l a \rightarrow 1 \rightarrow a a^l \quad , \quad a a^r \rightarrow 1 \rightarrow a^r a$$

Here the arrow denotes **partial order**. A relation that is **reflexive**, **antisymmetric**, and **transitive** is called a partial order [Epp,1993].

i. e., for all a, b, and c, in P where P is a set and that ≤ is a relation on P, we have that.

$a \leq a$ (reflexivity)

if $a \leq b$ and $b \leq a$ then $a = b$ (antisymmetry)

if $a \leq b$ and $b \leq c$ then $a \leq c$ (transitivity).

A set with a partial order on it is called a **partially ordered set, poset**. Lambek considered **Free Pregroups** and posetsof *basic types*, which may differ from one language to another, and which is meant to express certain elementary grammatical concepts. From the basic types one forms *simple types* by repeated adjunction. Thus a *simple type* has one of the following forms:

$$\ldots a^{ll}, a^{l}, a, a^{r}, a^{rr}, \ldots$$

where a is a basic type. A compound type is a string of basic types. The types form a monoid under concatenation (1 being the empty string). The partially ordered monoid of types is a **pregroup** with adjunctions defined inductively thus:

$$1^{l} = 1 = 1^{r},\ (xy)^{l} = y^{l}x^{l},\ (xy)^{r} = y^{r}x^{r}$$

The resulting pregroup is the free pregroup generated by the given poset of basic types.

Let us now consider the pregroup of types freely generated by a poset of basic types for a small fragment of English.

π_j = j^{th} personal subject pronoun, where j = 1, . . ., 6 denotes the three persons singular followed by the three persons plural. In modern English, the original second person singular has disappeared and was replaced by the second person plural. Moreover, there is no morphological distinction between the three plural verb forms.

s_k = declarative sentence in the *k*th simple tense (k = 1, 2) where they stand for the present and past indicative respectively.

q_k = yes-or-no questions in the k^{th} simple tense.

o = direct object.

p_2 = past participle of intransitive verb.

i = infinitive of intransitive verb.

Both of the last-mentioned types may also apply to compound verb phrases.

$\boldsymbol{\pi}$ = subject when the person is irrelevant.

q = yes-or-no question when the tense is irrelevant.

Let us now consider a small fragment of English:

He has type $\boldsymbol{\pi}_3$(= third person subject)

Her has type o (=direct object)

Sees has type $\boldsymbol{\pi}_3^r \mathbf{s}_1 \mathbf{o}^l$ to indicate that we require a third person subject on the left and a direct object on the right.

Now look at the sentence

he sees her

$$\underline{\boldsymbol{\pi}_3\ (\boldsymbol{\pi}_3^r}\mathbf{s}_1\underline{\mathbf{o}^l)\mathbf{o}}\rightarrow\mathbf{s}_1$$

We calculate in two steps

$$\underline{\boldsymbol{\pi}_3(\boldsymbol{\pi}_3^r}\mathbf{s}_1\mathbf{o}^l) = (\boldsymbol{\pi}_3\boldsymbol{\pi}^{\mathbf{r}}_3)\mathbf{s}_1\mathbf{o}_l \rightarrow \mathbf{1s}_1\mathbf{o}^l = \mathbf{s}_1\mathbf{o}^l$$

$$(\mathbf{s}_1\underline{\mathbf{o}^l)\mathbf{o}} = \mathbf{s}_1\ \underline{(\mathbf{o}^l\mathbf{o})} \rightarrow \mathbf{s}_1 1 = \mathbf{s}_1.$$

It is convenient to indicate contraction by underlines.

Similarly, we have

I saw her

$$\underbrace{\pi_1 (\pi^r} \mathbf{s}_2 \underbrace{\mathbf{o}^l) \mathrm{o}} \to \mathbf{s}_2$$

where the first underline represents the generalized contraction

$$\boldsymbol{\pi}_1 \boldsymbol{\pi}^r \to \boldsymbol{\pi}\boldsymbol{\pi}^r \to 1$$

In the next example we make use of two further type assignments:

*Has*h as type $\boldsymbol{\pi}_3^r \mathbf{s}_1 \mathbf{p}_2^l$

Seen has type $\mathbf{p}_2 \mathbf{o}^l$

The former requires one complement on each side, the latter only a simple complement on the right to give

He has seen her

$$\underbrace{\boldsymbol{\pi}_3 (\boldsymbol{\pi}_3^r} \mathbf{s1} \underbrace{\mathbf{p}_2^l)(\mathbf{p}_2} \underbrace{\mathbf{o}^l) \mathbf{o}} \to \mathbf{s}_1$$

Note in contrast

I have seen her

$$\underbrace{\boldsymbol{\pi}_1 (\boldsymbol{\pi}_1^r} \mathbf{s}_1 \underbrace{\mathbf{p}_2^l) (\mathbf{p}_2} \underbrace{\mathbf{o}^l) \mathbf{o}} \to \mathbf{s}_1$$

You had seen her

$$\underbrace{\boldsymbol{\pi}_2 (\boldsymbol{\pi}_2^r} \mathbf{s}_2 \underbrace{\mathbf{p}_2^l) (\mathbf{p}_2} \underbrace{\mathbf{o}^l) \mathbf{o}} \to \mathbf{s}_2$$

Unfortunately, *has* must be assigned a different type in direct questions, namely

Has: $\mathbf{q}_1 \mathbf{p}_2^l \ \boldsymbol{\pi}_3^l$

To obtain

Has he seen her?

$$(\mathbf{q}_1 \underbrace{\mathbf{p}_2^l \underbrace{\pi_3^l) \pi_3} (\mathbf{p}_2} \underbrace{\mathbf{o}^l) \mathrm{o}} \to \mathbf{q}_1$$

In Lambek's book a detailed presentation of English grammar is given including:

Nouns, adjectives, verbs, adverbs.

Negative and interrogative sentences

Indirect questions.

Doubly transitive verbs.

He gave also a list of the posets of basic types.

It should be noted finally that there are aspects of the English language that were not considered. I give here one example" the irregular verbs" in which Steven Pinker considered in his book *Words and Rules* [Pinker, 1999].

Regarding other languages, Pregroups have been used to analyze the sentence structure of many languages, For example, French, German, Italian, Polish, Arabic, Japanese, and Persian. Therefore, it is possible to use it to study comparative structures of different languages.

3 Categorical Quantum Protocols

The tools available for developing quantum algorithms and protocols until 2004 were low-level. However it was learned from computer Science the importance of compositionality, types, and abstractions [Abramsky and Coecke, 2004, 2005, 2008]. A simple exposition was given by [Coecke, 2005] with an exposition for Categories given by [Coecke and Paquette, 2010]. In this tutorial, I am going to use the simple exposition given by Coecke.

Let us start by defining a category. Consider a system of type *A* and perform an operation *f* on it. Then, we have,

$$A \xrightarrow{f} B$$

where A is the initial type of the system, B is the resulting type and *f* is the operation. One can also perform an operation

$$B \xrightarrow{g} C$$

and write $g \circ f$ for the consecutive application of these two operations. Clearly we have

$$(h \circ g) \circ f = h \circ (g \circ f)$$

If we further set

$$A \xrightarrow{1_A} A$$

For the operation "do nothing on a system of type A" we have

$$\mathbf{1_B \circ f = f \circ 1_{\mathit{A}} = f}$$

Hence, we can define a Category C as consisting of:

- Objects A, B, C, …
- Morphisms *f, g, h*, … ϵ C(A, B) for each pair A, B.
- Associative composition, i. e

$f \epsilon C(A, B), g \epsilon C(B, C) \Rightarrow g \circ f \epsilon C(A, C)$

With $(b \circ g) \circ f = h \circ (g \circ f)$

- An identity morphism $1_A \epsilon C(A, A)$ for each A, i. e.
- $f \circ l_A = 1_B \circ f = f$

When in addition we want to be able to conceive two systems A and B as one whole $A \otimes B$ and also to consider compound operations $f \otimes g$: $A \otimes B \rightarrow C \otimes D$, then we pass from ordinary categories to a (2- dimensional) variant called **monoidal categories**.

Let us now consider what is called the **language of pictures** which has some primitive data (lines, boxes, triangles, and diamonds) in which we have two kinds of composition, namely parallel (conceiving two systems as a compound single one) and sequential (concatenation in time) and which will obey a certain axiom. Then we derive some results using this picture calculus, e. g. teleportation, logic gate teleportation and entanglement swapping

The primitive data of our formalism consists of:

(1) Boxes with an input and an output which we call "operation" or "channel"

(2) Triangles with only an output which we call "state"or "preparation" procedures" or "ket"
(3) Triangles with only an input which we call "co-state" or" measurement branch "or "bra"
(4) Diamonds without inputs or output, which we call "values" or "probabilities" or "weights"
(5) Lines which might carry a symbol to which we refer as the "type" or the "kind of system, and the A-labeled line itself will be conceived as "doing nothing to a system of type A" or the "identity of A"

Figure (1) shows the primitives of the language of pictures.

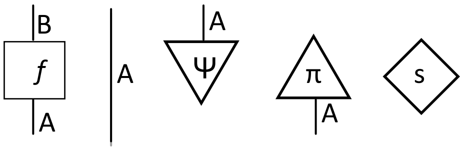


Figure 1 Primitives of the language of pictures

Figure (2) shows examples of combinations of different picture primitives.

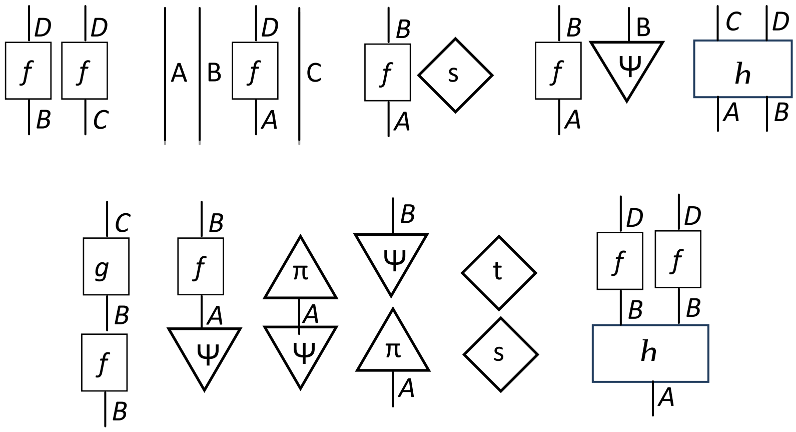


Figure 2 Examples of combinations of different picture primitives

If we connect up a state and costae (i. e. we produce a bra-ket) we obtain a diamond shape since no inputs or outputs remain. Thus we obtain what we called probability. On the other hard if we connect up a costate and a state (i. e. produce a ket-bra) we obtain a square shape with a genuine input and a genuine output.

Fig. (3) and Fig. (4) show also some useful identities.

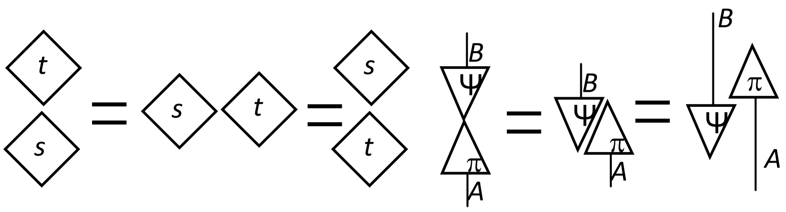


Figure 3 Some useful identities

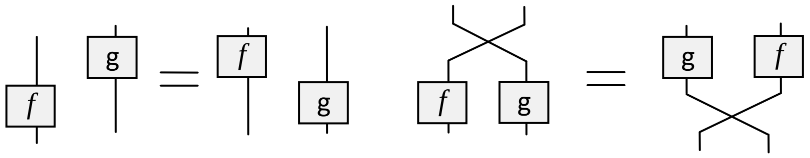


Figure 4 Some more identities

It is also assumed that lines carry an orientation which means that there exists an operation on types which sends each type *A* to a type **A*** with the opposite orientation. We refer to **A*** as A's *dual*. We also assume that for each box *f: A ⇢ B* there exits one upside down box *f†: B ⇢ A* called *f†*, s*adjoint*. These situations are shown in fig. (5).

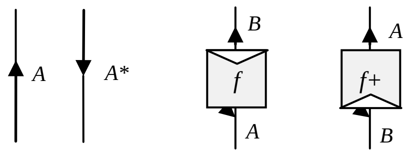


Fig (5) Dual and Adjoints

4 Compositional and Distributional Semantics

Let us first give a brief idea about vector-based models of word meaning. One of the early methods was gained from the field of information retrieval [Salton, G. and McGill, M. J., 1984].

Also, a more modern exposition is given by [Van Rijsbergen, 2004] which concentrates on the geometry of information retrieval and gives an introduction of its relation to quantum mechanics. The idea of Vector Space Models (VSM) is to represent each document as a point in a space (a vector in a vector space).

Points that are close together in this space are semantically similar and points that are far apart are semantically distant. VSM performs well on tasks that involve measuring the similarity of meaning between words, phrases, and documents. They are also related to the distributional hypothesis which means that words that occur in similar contexts tend to have similar meaning. In order to apply this abstract hypothesis leads to vectors, matrices, and higher order tensors [Turney, P. D. and Pantel, P. 2010].

The principle of compositionality is the principle that states that the meaning of a complex expression is a function of the meaning of its parts and the way these parts are syntactically combined. A number of researches have tried to reconcile the frameworks of distributional semantics with the principle of compositionality. Let us first consider composition models: [Mitchel and Lapata M., 2008] . Some researchers formulate semantic composition as a function of two vectors **u** and **v**. They assume that individual words are represented by vectors acquired from corpus. The word's vector typically represents it co-occurrence with neighbouring words .A hypothetical semantic space is illustrated in Fig. (6).

	animal	**stable**	**village**	**gallop**	**jokey**
Horse	o	6	2	10	4
Run	1	8	4	4	0

Figure 6 A hypothetical semantic space for horse and run.

Here, the space has only five dimensions and the matrix cells denote the co-occurrence of the target words (*horse* and *run*) with the context words (*animal, stable*, and so on).

Let **p** denote the composition of two vectors **u** and **v**, representing a pair of constituents which stand in some syntactic relation ***R***. We can thus define a general class of models for this process of composition as:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}, \boldsymbol{R})$$

If we consider ***R*** is fixed to a single well defined linguistic structure, for example the verb-subject relation, then we can write:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v})$$

This still leaves *f*unspecified.

If we assume that **p** lies in the same space as **u** and**v**, avoiding the issues of dimensionality associated with tensor products, and that *f* is a linear function, then we generate a class of*additive* models:

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$$

where **A** and **B** are matrices which determine the contribution made by **u** and **v** to produce **p**.

In contrast, if we assume that *f* is a linear function of the tensor product of **u** and **v**, then we obtain *multiplicative* models

$$\mathbf{p} = \mathrm{C}\mathbf{u}\mathbf{v}$$

where C is a tensor of rank 3 which projects the tensor product of **u** and **v** into the space of**p**.

Further constraints can be introduced to reduce the free parameters in these models leading finally to:

$$\mathbf{p}_i = u_i + v_i$$

and $$\mathbf{p}_i = u_i \, .v_i$$

For example, the addition of two vectors representing **horse** and **run** in Fig. (6) would yield

Horse + **run** = [1 14 6 14 4]

whereas their product is given by

Horse . **run** = [0 48 8 40 0]

As a result of the assumption of symmetry, both these models are "bag of words" models and word –order insensitive. Relaxing the assumption of symmetry in the case of simple additive model produces a model which weighs the contribution of the two components differently as

$$p_i = \alpha u_i + \beta v_i$$

The previous reference contains more details about this approach using for evaluation the British National Corpus (BNC) together with some parsed versions of it.

Another research in this direction was given by [Van de Cruys, T. et al, 2014]. In this paper the authors modeled compositionality as a multi-way interaction between latent factors which are automatically constructed from corpus data. Here, they used the UKWAC corpus which is a 2 billion word corpus automatically harvested from the Web, also together with a parsed version of it. Also, they used a tensor-based factorization model. They obtained better results than the previous paper.

One of the main problems in the previous approach that uses simple addition and multiplication is the commutativity of the operators: they treat the sentence as a "bag of words" where the word order does not matter, for example equating the meaning of the sentence "dog bites man" with that of "man bites dog". This fact motivated researchers to seek solutions based on noncommutative operators, such as the tensor product between vector spaces [Kartsaklis, D. 2014]. Thus the composition of two words is achieved by a structural mixing of the basis vectors that result in an increase of dimensionality:

$$\vec{w}_1 \otimes \vec{w}_2 = \sum_{i,j} c_i^{w1} \, c_j^{w2} (\vec{n}_i \otimes \vec{n}_j)$$

The meaning of a word is then represented as the tensor product of the word's context vector with another vector that denotes the grammatical relationships. As an example, the meaning of the sentence "dog bites man" is:

$$\overrightarrow{dog\ bites\ man} = (\overrightarrow{dog} \otimes \overrightarrow{subj}) \otimes \overrightarrow{bites} \otimes (\overrightarrow{man} \otimes \overrightarrow{obj})$$

Thus the bag of word problem is solved at the expense of increasing the dimensionality. This new problem was solved in this paper using some categorical concepts. Other papers tackled also this issue [Clark, S. et al., 2008] and [Coecke, B. et al, 2010] where the last paper gave the mathematical foundations for the compositional distributional model of meaning.

Due to the role of Categorical Quantum Protocol in the unification of the different models for sentence semantics we devote the following section to research in this direction and finally discuss two papers for evaluating these models.

5 Distributional Compositional Categorical (DisCoCat) Model of Meaning

In this section we will see how to combine distributional and compositional semantics together with the axiomatic framework for dealing with quantum information processes [Abramsky, S. and Coecke, B., 2004] [Clark, S. et al, 2013] which admits purely diagrammatic calculus [Coecke, B., 2010]. The teleportation protocol in quantum mechanics [Nielsen, M. A, and Chuang, I. L., 2000] [Benenti, G, et al, 2004] provides a cornerstone for the diagrammatic reasoning techniques. In Fig. (7) we show the derivation of the general teleportation protocol where the *f*-label represents both the measurement outcome and the corresponding correction performed by Bob [Coecke, B., 2010]

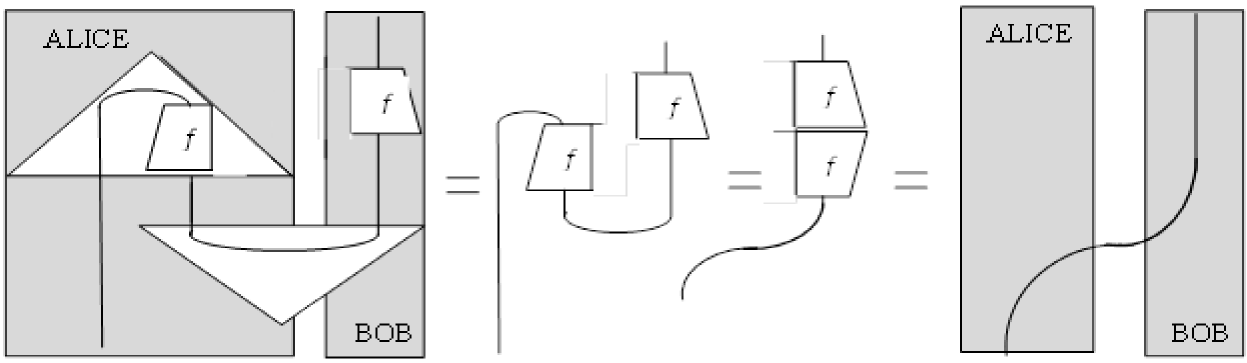


Figure 7 General Transportation Protocol

The main conceptual idea behind these diagrams is that besides these operational physical meaning, they also admit a "logical meaning" in terms i*nformation flow*. Referring to Fig. (8), the dashed line represents the logical flow which indicates the state incoming at Alice's side first gets acted upon by an operation *f*, and then by its adjoint $f^{\dagger}$which in the case that *f* unitary results in the outgoing state at Bob's side being identical to the incoming state at Alice's side.

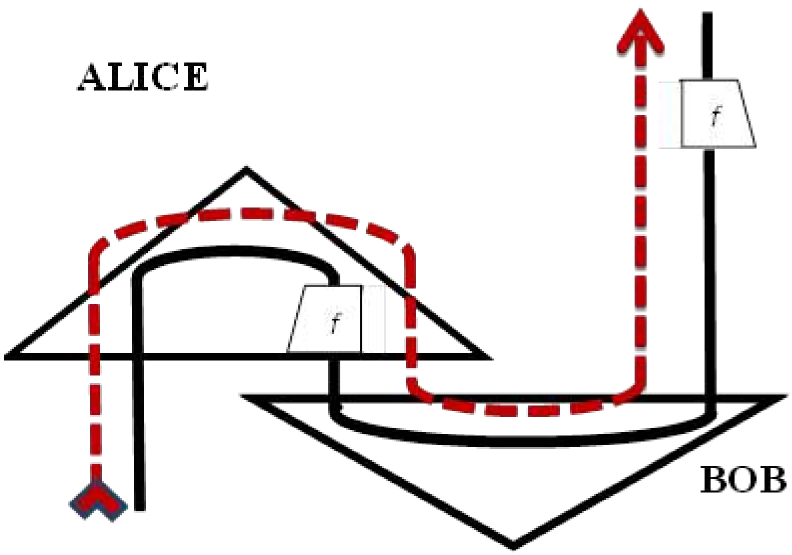


Figure 8 Logical Information Flow and Physical Flow

When interpreted in Hilbert space, the key ingredients of this formalism are "cups" and "Caps":

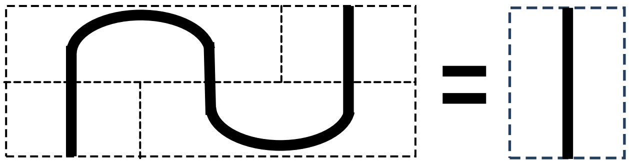
$:= |00\rangle + |11\rangle$

$:= \langle 00| + \langle 11|$

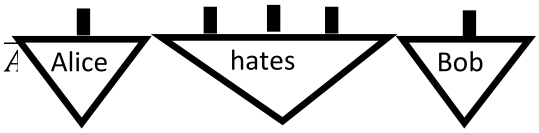
and the equation that governs them is:

$$((\langle \mathbf{00}| + \langle \mathbf{11}|) \otimes \mathbf{I}_d)\ (\mathbf{I}_d \otimes (|\mathbf{00}\rangle + |\mathbf{11}\rangle)) = \mathbf{I}_d$$

which diagrammatically depicts as:



To apply the above concepts to Natural Language Processing, let us consider an example for transitive verbs. A transitive verb requires both an object and a subject to yield a grammatically correct sentence. Consider the sentence "Alice hates Bob": Assume that the words in it are represented by vectors which we denote by triangles:



How do these words interact to produce the meaning of the sentence. We feed the meaning of vectors $\overrightarrow{Alice}$ and $\overrightarrow{Bob}$ into the verb $\overrightarrow{hates}$ which then output the meaning of the sentence Fig. (9) shows how to achieve this

Figure 9 How the transitive verb interacts with the subject and object

Let us see how this example is represented using Lambek pregroup grammar.

Alice hates Bob

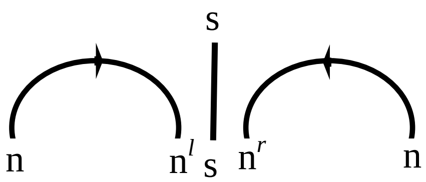
$n \quad n^l s n^r \quad n \quad = (n n^l)\ s (n^r n)$

$\leq 1\ s\ 1$

$= s$

Thus, this is a valid grammatical structure for a sentence.

The inequalities using $n^l n \leq 1$, and $n^r n \leq 1$ can also be represented with "directed" caps:



In category theoretic language, both the diagrammatic language for quantum axiomatic and pregroups are called *compact closed categories*, while the quantum language is *symmetric*, pregroups have to be *non-symmetric* given the importance of word order in sentences. As another example, consider the following sentence:

"Alice does not like Bob"

where" does" and "not" are assigned only "logical" meanings.

Fig. (10) Shows the details of this example.

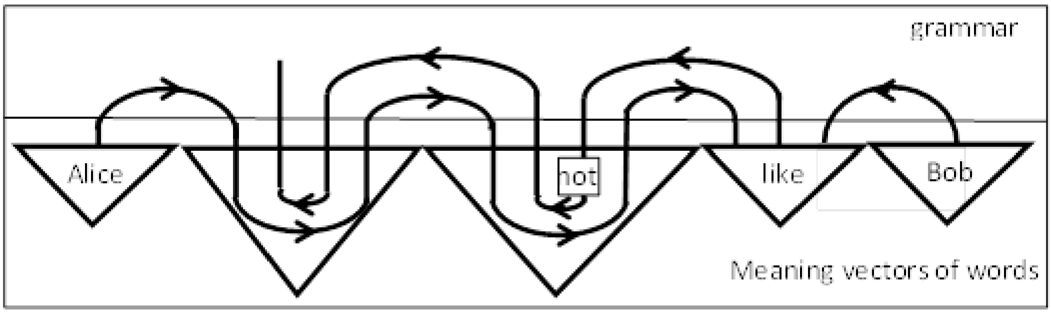


Figure 10 The sentence (Alice does not like Bob)

Some authors have given experimental support for this Distributional Compositional Categorical model of meaning (DisCoCat) [Grefenstette, E. and Sadrzadeh, M., (2011) (1)] [Grefenstette, E. and Sadrzadeh, M., (2011) (2)], and indicated that it is a promising approach. However, more experimentation is still needed and this represents only the beginning.

6 Conclusions

In this paper the different basic models for sentence semantics were presented and then how to unify them together. The first model was the compositional model which needs a grammar that has to be checked first to make sure the grammaticality of the sentence. The grammar chosen was Lambek pregroup grammar. This choice is related to the categorical model for describing the high level quantum protocol that is needed for unifying the different semantic models.

The second model was the distributional semantic model which is an emperical model and needs a large corpora. The corpora considered were the British National Corpus and another one harvested from the Web, and is called UKWAC.

The unification model was called DisCoCat, where the name reflects that this model unifies the Distributional Compositional Categorical models together. Although the grammar used is for the English language, Lambek pregroup grammar could be used for other languages. A reference was given that compared the grammars for a number of languages, among them the Arabic language.

It should be noted that high-level quantum models started to appear in the literature in 2004 and its application for linguistics started to appear in 2010. Since then, intensive applications in linguistics attracted a large number of researchers.

REFERENCES

[1] Abramsky, S. and Coecke, B. "A Categorical Semantics of Quantum Protocols", Proc.Of the 19th Annual IEEE Symposium on Logic in Computer Science, PP. 415 – 425. IEEE Computer Science Press, 2004.
[2] Benenti, G, et al "Principles of Quantum Computational and Information" World Science, 2004,
[3] British National Corpus, Wikipedia, 30/10/2015.
[4] Clark, S, et al. "A Quantum Teleportation Inspired Algorithm Produces Sentence Meaning From Word Meaning And Grammatical Structure". ArXiv: 1305.0556.Oct. 2013.
[5] Clark, S. et al, "A Compositional Distributional Model of Meaning", Proc. Conf. On Quantum Interactions, University of Oxford, 2004, PP. 133 -140.
[6] Cocke, B. and Paquette, E. O. "Categories for the Practicing physicist, In Coecke, B. editor, "New structures for physics", Lecture Notes in Physics, PP. 167 – 271. Springer, 2010.ArXiv: 0905.3010.
[7] Coecke, B. et al, "Mathematical Foundation for a Compositional Distributional Model of Meaning". Linguistic Analysis 36, Lambek et al, 2010.

[8] Coecke, B. et al. "Mathematical Foundations for a Compositional Distributional Model of Meaning". Linguistic Analysis 36, Lambek et al, 2010.
[9] Coecke, B. Quantum Picturism. "Contemporary Physics 51,59 -83. ArXiv; 0908.1787.
[10] Coecke, B.”Kindergarten Quantum Mechanics". Lecture Notes. In Khrennikov, editor, Quantum Theory Reconsideration of the Foundations III, PP. 81 – 98. AIP Press, 2005 ArXiv: quant-ph 0510032.
[11] Epp, S. "Discrete Mathematics with Applications "PWS Publishing Company, 1993.
[12] Ferraresi A, et al. "Introducing and Evaluating UKWAC, a Very large Web – derived Corpus of English", 2008.
[13] Grefenstette, E. and Sadrzadeh, M. "Experimental Support for a Categorical Compositional Distributional Model of Meaning". Proc.Of the 2011 Conference on Emperical Methods of Natural Language Processing. PP. 1394 – 404,Edinburgh, Scotland, U. K. July 2011(1) , Association for Computational Linguistics.
[14] Grefenstette, E. and Sadrzadeh, M. "Experimenting with Transitive Verbs in a DisCoCat" arXiv: 1107, 3119 v2. July 2011(2).
[15] Kartsaklis, P. "Compositional Operators in Distributional Semantics" arXiv:1401. 5327 [cs. CL.]21, Jan. 2014.
[16] Lambek, J. "From Word to Sentence" (A Computational algebraic approach to Grammar". Polimetrico International Scientific Publisher, 2008.
[17]Lambek, J. "Pregroups and Natural Language Processing" The Mathematical Intelligencer, Vol. 28, No. 2,2006, PP. 41 – 48.
[18]Lambek, J” The Mathematics of Sentence Structure, "The American Mathematical Monthly, Vol.65, No3, PP.154-170, 1958.
[19]Mitchel, J. and Lapata, M. "Vector-based Model of Semantic Composition", Proc. Of ACL- 08: PP. 236 – 244. 2008.
[20]Nielson, M. A. and Chuang, I. L. “Quantum Computation and Quantum Information". Cambridge University press, 2000.
[21]Partee, B. H. et al. "Mathematical Methods in Linguistics", Kluwer Academic Publishers, 1995.
[22]Pinker, S. "Words and Rules." (The Ingredients of Language).Basic Books, 1999.
[23]Sadrzadeh, M. "High Level Quantum Structures in Linguistics and Multi Agent Systems", 2007 American Association for Artificial Intelligence.
[24]Salton, G. and McGill, M. j. "Introduction to Modern Information Retrieval", McGraw – Hill, 1984.
[25]Turney, P. D. and Pantel, P. "From Frequency to meaning: Vector Space Models of Semantics". Journal of Artificial Intelligence Research 37 (2010).PP. 141 -188.
[26]Van de Cruys, T. et al. "A Tensor-based Model of Semantic Compositionality". Conference of the North American Chapter of the Association of Computational Linguistics. May 2014.
[27]Van Rijsbergen, C. J. "The Geometry of Information Retrieval", Cambridge University Press. 2004.

BIOGRAPHY

Prof. M. Adeeb Ghonaimy was born on the 28th of December 1936. He currently holds the position of Professor Emeritus at the Faculty of Engineering , Ain Shams University.
In 1999, he was awarded the State Prize of Appreciation in Advanced Engineering Technologies.
From 1987 to 1997, he was the Director of the Egyptian Universities Network.
From 1979 to 1981 and from 1985 to 1997, he was the Director of the Information Systems Center , Faculty of Engineering , Ain Shams University.
From 1979 to 1981 and from 1989 to 1991 he was the Chairman of the Electronics and Computer Department , Faculty of Engineering , Ain Shams University.
He got his M.A.Sc and Ph.D degrees from the University of Toronto, Canada in 1961 and 1965 respectively. He got his B.Sc from the Electrical Engineering Department (Communication Section), Faculty of Engineering , Cairo University in 1958.

مقالة تعليمية عن دلالة الجمل باستخدام نحو(Lambek) المعتمد على (Pregroups) والبروتوكولات الكمية التى تستخدم نظرية التصنيف

أ. د/ محمد اديب رياض غنيمى
الأستاذ المتفرغ بكلية الهندسة
جامعة عين شمس
adeebghonaimy@hotmail.com
adeeb.ghonaimy@eng.asu.edu.eg

تعتمد دلالة الجملة على نموذجين أساسيين : الأول يسمى نموذج التركيب والثانى يسمى نموذج التوزيع . فكرة التركيب تنص على ان معنى الجملة الكبيرة تكون دالة فى مكونات هذه الجملة التى تربطها قواعد النحو . وفكرة التوزيع تقول أن الكلمات التى توجد فى سياق واحد يكون لها معانى متشابهة . ونظرا لأن فكرة التركيب تعتمد على قواعد النحو التى تربط مكونات الجملة فقد تم اختيار نحو Lambek الذى يعتمد على (Pregroups). و طريقة التوزيع تتطلب وجود حصائل (corpora) كبيرة مثل British National Corpus الذى يحتوى على 100 مليون كلمة أو UKWAC الذى تم تحصيله من خلال شبكة الإنترنت عن طريق محتويات الشبكة المعرفية (Web) ويحتوى على أكثر من 2 بليون كلمة . ونظراً لأن لكل نموذج مميزاته وعيوبه فقد إبتدأ فى الأونة الأخيرة (بعد سنة 2010) محاولات دمج النموذجين وذلك عن طريق بروتوكولات كمية تستخدم نظرية التصنيف. وأحد هذه النماذج يتم إختصاره إلى (DisCoCat) لتعكس دمج ثلاثة نماذج فى نموذج واحد . وقد تمت بعض وسائل التقييم التى بينت الإضافة التى نتجت عن هذا الدمج .

Semantic-Based Approaches for XML Summarization

Hassan A. Elmadany [*1], Marco Alfonse [*2], Mostafa Aref [*3]

[*] *Computer Science Department, Faculty of Computer and Information Sciences Ain Shams University, Cairo, Egypt*

[1] hassanelmadany@cis.asu.edu.eg

[2] marco@fcis.asu.edu.eg

[3] mostafa.aref@cis.asu.edu.eg

***Abstract*— eXtensible Markup Language (XML) is one of the standard data representations used in various applications. The need to summarize XML document to generate concise, readable summary that provides all important information is very noble as it saves both time and effort. This paper presents Main approaches for summarizing XML documents based on both its structural and data contents.**

1 INTRODUCTION

eXtensible Markup Language (XML) represents different data in efficient way due to its flexibility as it can be supported in various applications. With the increasing uses of XML in data exchange and representation and difficulty to read and understand large and complex XML documents. It is necessary to provide approaches that summarize XML document in a semantic manner. XML summarization has challenges due to [1]:

- *Informativeness:* a unit of information, e.g. tags and text must be informative to the user as its importance in the document as it must be presented concisely to the user.
- *Non-redundancy:* a tag could occur multiple times in a document and each tag is associated with a distinct value. Clearly, it is not important to repeat all occurrences of the tag in the generated summary, but represent it concisely using a single tag.
- *Coverage:* referring to the amount of information rather than data in the XML summary.
- *Coherence:* the context of a tag in terms of its parents or siblings may be important.

There are two kinds of summaries that can be generated: (1) Generic Summarization based on the entire contents of the XML documents ", a generic summary summarizes the entire contents of the XML document" [1]. (2) A Query-Based summarization which summarizes the parts of the document which are relevant to what the user types in his query [1]. We classify the approaches for summarizing XML documents into two main categories according to the coverage degree of the generated summary: 1) XML structural summarization, 2) XML content and structure summarization [2].

However, XML structural summarization approaches focus on generating a summary of XML document based on its structural, XML content and structure summarization approaches focus on generating XML summary based on the content features of the logical structure of the XML document to provide a semantic summary from the original XML document. In this paper, we focus on Content and structures summaries approaches for XML documents. We categorized the XML Content and structures summaries approaches into three main categories

1) *Ranking Approach*

2) *Schema Approach*

3) *Compression Approach*

This paper is organized as follows: section 2 presents Ranking approaches, section 3 presents Schema approaches, section 4 presents Compression approaches, section 5 presents a comparison between the approaches discussed in the paper, and finally the conclusion is reported in section 6.

2 RANKING APPROACH

Generate an XML summary that it is concise and readable with respect to memory budget. The generated summary contains the important information from the XML document in small size. In this approach we rank both tags and text values due to its importance to the XML document. They will be included in the summary based on their ranking. Then we rewrite them to make it readable with respect to memory budget. First, Tags and text values are ranked due to some features. Second step, select only the top-ranked to be in the summary.

There are three main methods used to rank, text values [1]:

- *Centroid query method:* first, this method generates a centroid query based on a popular and known keywords. Second step is to calculate a score for each text value with respect to centroid query.
- *Diverse text value:* this method ranks the text values due to its importance in the document according to their occurrence or frequency that is how many times the value has been occurred.
- *Correlated Samples:* if the text units are correlated which means they are dependent on each other. First, we rank the first text value due to its importance in the document. Second, we look to the dependent text value and rank it also.

The important tag unit ranks according to the given concept that "the popularity of a tag within the document does not correlate directly with its importance" [1]. First, we enumerate all paths of the XML document. Second rank the tags according to its frequency /occurrence.

Maya Ramanath and Kondreddi Sarath Kumar [1] used this approach to develop a framework for summarizing XML document with respect to memory budget. They were able to generate a concise and readable summary. Figure 1 represents the algorithm.

They follow the following steps: First, rewrite text units using one of the following methods:

- *Enumerate:* the system will enumerate the text units if for each text, its length is short and the number of unique values is small.
- *Sample:* the system will sample the text units if for each text, its length is short but, the number of unique values is large.
- *Shorten/Enumerate:* the system will enumerate the shortened text units in case we need a shortened text, but the number of unique values is small.
- *Sample/Shorten:* the system will sample the shortened text units in case we need a shortened text, but the number of unique values is large.

Second, rank the text values and tags. Third, construct a summary based on the ranked tags. Finally, if the size of the generated summary is larger than a given size repeat the algorithm.

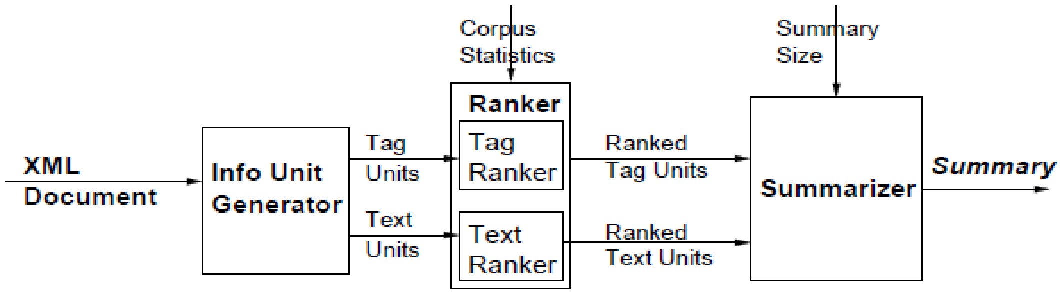


Figure 1: Rewrite-Ranking Summarization [1]

3 Schema Approach

This approach aims to present the important schema elements that are used to make large schemas of XML readable easily. Schema summary provides an overview for the schema and explores its relevant component in depth only. It tries to achieve two main goals: First, it presents an important element in the schema. Second, it achieves information coverage.

The schema can be considered as a labeled-directed-graph with a root node in both relational and hierarchical. So for relational schema each node refers to Relation/attribute column. For hierarchal schema each node refers to hierarchical schema. Each edge represents a structural and value links, e.g. constraint of the foreign key, parent-child link… etc. It contains abstract elements and abstract links. Each abstract element represents a group of original elements in the schema, then choose a single element as a representative of each group. Each abstract link represents one or more links between schema elements with those abstract elements.

There are two contradictory types of schema summary: (1) full summary: this is a type of schema summary that contains only abstract elements and neglect root and (2) Expanded summary [3].
Schema summary has some properties such as

- *Summary complexity:* this property refers to the number of elements in the summary.
- *Summary importance:* refers to the ratio between the total importances of elements in the summary versus the total element's importance in the original schema.
- *Summary coverage:* refers to is the ratio between the total coverage of all schema elements by elements in the summary and the total coverage of all schema elements by the original schema.

Jakub Marciniak [4] developed a framework that summarizes the schema summary as we first extract the schema for a given XML document. Second step, we generate schema summary where its size can be obtained by the number of its nodes with respect to memory budget. Finally the text value is summarized to generate XML summary by presenting the data from the original document corresponding to the schema summary. Figure 2 illustrates these steps discussed above.

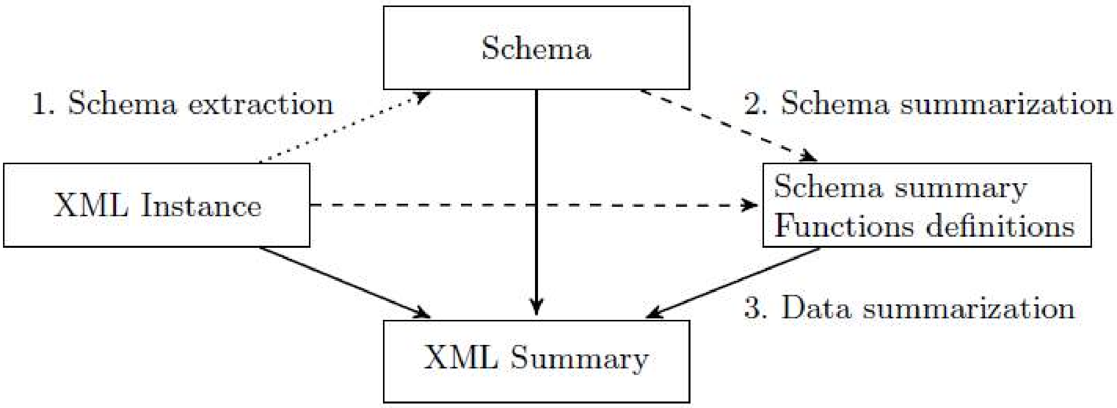


Figure 2: Jakub's Framework [4]

Ten and Ping [5] developed a framework based on both the schema and the XML document itself. They first, remove redundant data from an XML document based on both the abnormal functional dependencies and the schema structure. Second, they summarize tags. In case of key tag it will remain as it is in the included summary. Other tags will be extracted due to their frequency in the document. Third, summarize the text value for multiple values in the same tag we summarize only the first tag, but for long values we are summarizing according to fixed length defined before. So we use schema summarization because it provides important semantic and structural information. As Schema provides the structural and semantic of the XML document that refers to the important information in the document.

4 COMPRESSION APPROACH

This approach summarizes XML documents by compressing its structure and data texts. Most of the tools uses this approach, focus on query processing rather than generate a readable version for the user. It tries to generate XML summary with both speed and effectiveness. XML compressors can be categorized into two main categories:

- *Non-Queriable:* tries to get the highest compression ratio, e.g. XMill, NRCX… etc.
- *Queriable:* provides an evaluation of queries on their compressed formats, e.g. Xgrind, IEX… etc.

However a good compression rate can be achieved using XMill [6], the decompression is required before doing a query. So the time of query response will be increased. It does not need schema Information. It summarizes the XML document as the following: First, compress both data and structure independently of each other. Second, for data text with similar meaning grouping them into one container, e.g. all tags such as <name> can be grouped into one container. The next step is to compress these containers separately. Finally, apply different semantic compressors for the containers for other data items such as numbers. It uses the container expression to group data items into the container. It is a concise language, it is used to select semantic compressors. XMill is considered a semi-automated technique that needs a user assistance to get a good compressed summary."The compression rate can be expressed as bit per byte"[6] e.g. 2 bits/byte means that the compression rate is 25%. They did the XMill on different resources for the data and compression rate for XMill reach to 45%-60% [6].

NRCX [7] stands for Non Redundant Compact XML storage. First, a path index is created to allow XPath queries that contain parent-child relationship. Second, it stores all unique paths in the XML documents regarding the value of this path. Finally, consider that the path id refers as a link between both the path and its content. This path is stored only once. They did experiments for different databases such as Mondial, orders, Shakespeare and Lineitem. They found the compression rate (in MB) 0.189, 1.1821, 3.7 and 7.021 respectively [7].

Grind [8] compresses the text data using a simple context free compression schema based on Huffman coding in a semi adaptive way. It proposed a great factor that retains the structure of the document so we can parse the compressed document using the same techniques. It provides useful features that try to achieve information utilization in the schema in order to improve compression rate. It needs two scans for XML documents so its compression time will be larger. Comparing Xgrind with XMill technique discussed above, they found that the compression rate is lower than the one in XMill, but the average of its compression ratio is about77% of XMill. Also, in the worst case, it equals 68% of XMill [8].

5 DISCUSSION AND COMPARISON

In this section we provide a discussion and comparison between the approaches discussed above in this paper as is illustrated in Table 1. This comparison is based on four criteria: Memory budget concept, User Assistance, Popular Keywords and Query processing.

The Memory Budget Concept means summarize XML documents in a small size stored in the memory. *The User Assistance* Indicates how the user can assist the approach to generate the XML summary. *The Popular Keyword* indicates if the approach uses the popular keywords concept in generating XML summary.

TABLE I
COMPARISON BETWEEN APPROACHES

	Ranking Approach	**Schema Approach**	**Compression Approach**
Memory Budget Concept	Yes	Yes	Yes
User Assistance	No	Yes	Yes
Popular Keyword	Yes	Yes	No
Query processing	No	No	Yes

6 CONCLUSIONS

We presented this paper to highlight XML summarization to generate a semantic summary based on both its structure and data content. The approaches discussed in this paper try to fit the available memory in small size with respect to the size of the original one. The XML Summarization process helps the user to understand the large and complex XML documents by generating a concise summary in less size. We hope that this initial attempt to increase and improve ways to generate summarized XML documents.

REFERENCES

[1] Ramanath, M., & Kumar, K. S. (2008, April). "A rank-rewrite framework for summarizing XML documents". In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (pp. 540-547). IEEE.

[2] Moraes Filho, José de Aguiar. (2010). "Summarizing XML Documents: Contributions, Empirical Studies, and Challenges".

[3] Yu, C., & Jagadish, H. V. (2006, September). "Schema summarization". In *Proceedings of the* 32nd *international conference on Very large data bases* (pp. 319-330). VLDB Endowment.

[4] Marciniak,J.(2010). "XML schema and data summarization". In *Artificial Intelligence and Soft Computing* (pp.556-565). Springer Berlin Heidelberg.

[5] Lv, T., & Yan, P. (2013). "A framework of summarizing XML documents with schemas". *Int. Arab J. Inf. Technol.*, 10(1), 18 27.

[6] Liefke, H., & Suciu, D. (2000, May). "XMill: an efficient compressor for XML data". In *ACM Sigmod Record* (Vol. 29, No. 2, pp. 153-164). ACM.

[7] Atique, M. & Raut, A. D.(2012). "A Non redundant compact XML storage for efficient indexing and querying of XML documents". In *Global Trends in Computing and Communication Systems* (PP. 109-113). Springer Berlin Heidelberg.

[8] Tolani, P. M., & Haritsa, J. R. (2002). "XGRIND: A query-friendly XML compressor". In *Data Engineering, 2002. Proceedings. 18th International Conference on* (pp. 225-234). IEEE.

BIOGRAPHY

Hassan Abdelsabour Elmadany is a Software Engineer and a Researcher in the field of NLP. He got BSc. Of Computer Science since August 2012 and currently a Master student at the Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt. His research interests: Text and XML Summarization, Semantic Web, Ontological Engineering, and Artificial Intelligence.

Dr. Marco Alfonse Tawfik is a Lecturer at the Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt. He got Ph.D. of Computer Science since August 2014, University of Ain Shams. His research interests: Semantic Web, Ontological Engineering, Medical Informatics, and Artificial Intelligence. He has more than 18 publications in refereed international journals and conferences.

Prof.Mostafa Aref is a professor of Computer Science and Vice Dean for Society and the Environment, Ain Shams University, Cairo, Egypt. Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. of Electrical Engineering - Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.

تلخيص مستندات (XML) باستخدام اساليب الدلالات اللفظية

حسن عبدالصبور عبدالحليم محمد المدني*1 ، ماركو الفونس توفيق*2 ، مصطفى محمود عارف *3
*قسم علوم الحاسب، كلية الحاسبات و تكنولوجيا المعلومات، جامعة عين شمس، القاهرة، مصر

[1]hassanelmadany@cis.asu.edu.eg
[2]marco@fcis.asu.edu.eg
[3]mostafa.aref@cis.asu.edu.eg

ملخص
تعتبر مستندات (XML) واحدة من اهم مصادر عرض البيانات حيث يمكن إستخدامها فى العديد من التطبيقات لذا الحاجة إلى تلخيص هذه المستندات ضرورية للحصول على ملخص ذو معنى و قابل للقراءة حيث سيساهم فى توفير كلا من الوقت و الجهد المبذول . وقد قمنا هنا بعرض الاساليب المتاحه لعمليه تلخيص مستندات (XML) اعتمادا على ما تحتويه من بيانات وكذلك الهيكل الخارجى لها .

Syntax-Semantics Classification of Arabic Verbs for Semantic Annotation

Israa Elhosiny[*1], Sameh Alansary[*2]

[*] *Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt*
[1] israa.elhosiny@bibalex.org
[2] sameh.alansary@bibalex.org

***Abstract*—The semantic annotation of verbal predicates implies the systematic mapping between syntax and semantics. Therefore, the prime purpose of the study is to classify the Arabic verbs according to their syntactic and semantic behaviour to facilitate the semantic annotation given the syntactic representation. A manual linguistic analysis has been achieved for a compiled corpus to conclude the syntactic specifications of verbs arguments, and to map each syntactic argument with the suitable semantic representation. This result classification has been applied by building a computational lexicon to automatically analyse the corpus syntactically and semantically. The grammar using the proposed classification displayed a high level of success and component performance; accuracy of results amount to 92% of the total number of the mapped syntactic structures. That is, merely 8% of the corpora fail to be correctly mapped to the semantic graph.**

1 INTRODUCTION

Computational verb lexicon is an important key which supports NLP systems aiming to achieve semantic interpretation. Verbs are usually the components that contain the bulk of meaning of the sentence. Besides, verbs are highly variable, displaying a wide range of semantic and syntactic variability. Verb classifications help NLP systems to deal with this complexity by organizing verbs into groups that share core semantic and syntactic properties. There are many lexical databases that have been built. All of these lexical databases depend on theoretical views about verbs that were formulated in the past. One of the most widely known views concerning the lexicon is that articulated by Bloomfield [1], who wrote, "The lexicon is really an appendix of the grammar, a list of basic irregularities". A lexicon that contains the minimum information is necessary. Therefore, Bloomfield proposes that, the lexicon has to provide a record of precisely the idiosyncratic information associated with each lexical item. However, this view of the lexicon provide an incomplete picture of lexical knowledge as a whole, as the knowledge that a speaker demonstrates with respect to lexical items suggests that there is more to lexical knowledge than knowledge of idiosyncratic word-specific properties.

Reference [2] shows, for a large set of English verbs (about 3200), the correlations between the semantics of verbs and their syntactic behavior. More precisely, she shows that some facets of the semantics of verbs have strong correlations with the syntactic behavior of these verbs and with the interpretation of their arguments. She first precisely delimits the different forms of verb syntactic behavior. Each of these forms is described by one or more alternation. Alternations describe passive forms, there insertions and reflexive forms). She proposes an analysis of English verbs according to these alternations. Moreover, verb is associated with the set of alternations it undergoes. Her preliminary investigation showed that, there are sufficient correlations between some facets of the semantics of verbs and their syntactic behavior to allow for the formation of classes. Beth Levin has then defined about 200 verb semantic classes from her observation, where, in each class, verbs share a certain number of alternations. This very important work emerged from the synthesis of specific investigations on particular sets of verbs (e.g. movement verbs), on specific syntactic behaviors and on various types of information extracted from corpora. Other authors have studied in detail the semantics conveyed by alternations e.g. [3] and the links between them [4].

Of course, these alternations are language specific. They are not universal, even though some are shared by several languages (e.g. the passive alternation). The characteristics of the language, such as case marking, are also an important factor of variation of the form, the status and the number of alternations. English seems to have a quite large number of alternations; this is also the case e.g. for ancient languages such as Greek. French and Romance languages in general have much fewer alternations; their syntax is, in a certain way, more rigid. The number of alternations also depends on the way they are defined; in particular the degree of generality via constraints imposed on context elements is a major factor of variation.

Verb semantic classes are constructed from verbs, which undergo a certain number of alternations. From these alternations, a set of verb semantic classes is organized. We have, for example, the classes of verbs of putting, which include Put verbs, Funnel Verbs, Verbs of putting in a specified direction, Pour verbs, Coil verbs, etc. Other sets of classes include Verbs of removing, Verbs of Carrying and Sending, Verbs of Throwing, Hold and Keep verbs, Verbs of contact by impact, Image creation verbs, Verbs of creation and transformation, Verbs with predicative complements, Verbs of perception, Verbs of desire, Verbs of communication, Verbs of social interaction, etc. As can be noticed, these classes only partially overlap with the classification adopted in WordNet. This is not surprising since the classification criteria are very different. There are some other aspects which may weaken the practical use of this approach from both theoretical and practical points of views. The first is, verbs can exist in multiple lists, sometimes with conflicting structure. The second is, Levin explicitly states the syntax for each class, but falls short of assigning semantic components to each. And syntax alone is not enough.

Thematic relations have a vital role in the classification of verbs. They express generalizations on the types of lexical functions that are established between the verb and its arguments in the predication. There is a consensus among researchers that assignment of thematic roles to the arguments of the predicate imposes a classification on the verbs of the language. Since the type of thematic roles and their number are determined by the meaning of the verb, the lexical decomposition of verb meanings seems to be a prerequisite for semantic classification of verbs. The close relationship between the compositional and relational lexical meanings plays a central role in the classifications of verbs.

The existent verb classifications were developed within the frameworks of Case Grammar and Role and Reference Grammar (RRG). Works of Chafe [5], Cook [6] and Longacre [7] address the issues of verb classification with regard to thematic roles within the framework of the Case Grammar model. RRG, a structural-functionalist theory of grammar, is presented in works of Foley & Van Valin [8] and Van Valin [9]. Characteristic of RRG is that it accounts for a detailed treatment of lexical representation that proves to be instrumental in describing the thematic relations in typologically different languages. It also incorporates the insights of Dowty's and Jackendoff's theories. There is, however, an important difference in the treatment of thematic relations within those two frameworks. In Case Grammar, they have two functions: the first is to serve as a partial semantic representation of the lexical meaning. Second, they are considered an input to the syntactic operations, such as subjectivization, objectivization and rising. In the latter, the RRG model, thematic relations have only the second function.

There is no doubt that the model of semantic roles from the seventies, and in particular its repertory of roles and definitions, has to be replaced by a more stringent semantic model to suit the needs of NLP. The combination of the Dowty [10] model of proto-roles with the model of thematic sorts proposed by Poznansky & Sanfilippo [11] seems to be a very interesting proposal or solution.

The theory of verb classes occupies a central position in the system of lexical representation in the Role and Reference Grammar (RRG). It starts with the Vendler classification [12] of verbs into states (e.g. have, know, believe), achievements (e.g. die, realise, learn), accomplishments (e.g. give, teach, kill) and activities (e.g. swim, walk, talk). It utilizes a modified version of the representational scheme proposed in Dowty to capture the distinctions between these verb classes.

Dowty explains the differences between the verb classes in terms of lexical decomposition system in which stative predicates (e.g. know, be, have) are taken as basics and other classes are derived from them. Thus achievements which are semantically inchoative are treated as states plus a BECOME operator, e.g. BECOME know' "learn". Accomplishments which are inherently causative are represented by the operator CAUSE linked to the achievements operator BECOME, e.g. CAUSE [BECOME know'] "teach". Activities are marked by the operator DO for agentive verbs. These de-compositional forms are termed Logical Structures (LS) by Dowty. In RRG, they are interpreted in the following way as in table (1):

TABLE I
THE ROLE AND REFERENCE GRAMMAR VERB SCHEMA

Verb Class	**Logical Structure**
STATE	predicate'(x) or (x,y)
ACHIEVEMENT	BECOME predicate' (x) or (x,y)
ACTIVITY	(DO (x) [predicate' (x) or (x,y)])
ACCOMPLISHMENT	CAUSE, where is normally an activity predicate and an achievement predicate.

Many works in corpus and computational linguistics have been carried out following the approach of Levin, such as VerbNet [13], a lexicon with lexical semantic, argument and diathesis information for English predicates and adopts Prop Bank semantic annotation [14]. VerbNet identifies semantic roles and syntactic patterns characteristic of the verbs in each class and makes explicit the connections between the syntactic patterns and the underlying semantic relations that can be inferred for all members of the class. It is a lexicon of approximately 5800 English verbs, and groups verbs according to shared syntactic behaviors, thereby revealing generalizations of verb behavior. VerbNet is a domain-independent verb lexicon consisting of over 270 such verb classes, and is inspired by Beth Levin’s classification of verb classes and their syntactic alternations [2]. According to Levin’s work, members within a single verb class participate in shared types of alternations, such as locative alternation (spray verbs,) or the causative alternation (wrinkle verbs,) etc., because of an underlying shared semantic meaning. Thus, although the basis of VerbNet classification is syntactic, the verbs of a given class do share semantic regularities as well because as Levin hypothesized, the syntactic behavior of a verb is largely determined by its meaning.

AnCora also adopted Levin classification. It is a multilingual corpus annotated at different linguistic levels consisting of 500,000 words in Catalan (AnCora-Ca) and in Spanish (AnCora-Es). At present AnCora is the largest multilayer annotated corpus of these languages freely available. The two corpora consist mainly of newspaper texts annotated at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses). All resulting layers are independent of each other, thus making easier the data management. The annotation was performed manually, semi-automatically, or fully automatically, depending on the encoded linguistic information. The development of these basic resources constituted a primary objective, since there was a lack of such resources for these languages. Ancora defined 24 Lexical semantic structures (LSS). They are described and grouped around the 4 general event classes of Vendler. These 24 LSS derive from the analysis of the 50795 verbs in AnCora 2.0 corpora [15].

Arabic language needs a similar lexical resource to be utilized in the Arabic language processing field to support Arabic Natural Language applications with the semantic interpretation for the Arabic texts which enable information extraction and retrieval, machine translation, question answering systems to work efficiently. For this purpose the current study has been achieved to introduce the idea of verbs classification based on their syntax-semantic behavior.

This paper is divided into three sections; section 2 exhibits the bases of corpus compilation and linguistic analysis; syntactic and semantic analysis, section 3 presents the verbs syntax-semantic classification and extraction, section 4 discusses the implementation of the classification by means of building a computational lexicon and mapping grammar. Finally, section 5 concludes the paper.

2 CORPUS COMPILATION AND ANALYSIS

In order to classify the Arabic verbs, three main stages are required: 1) Selecting representative verbs for the syntactic verb classes in Arabic; representative of the different types of transitivity. 2) Gathering representative contexts for the selected verbs which enables verbs arguments to occur. 3) Analyzing the gathered sentences syntactically and mapping the syntactic representation to the semantic representation which enable recognizing which syntactic function can be mapped to which semantic relation. The following sub-sections discuss the three aforementioned requirements.

A. Verbs selection

The most common Arabic verbs in the Arabic UNL enumerative dictionary [16] have been selected. Each verb in the UNL Arabic dictionary has a transitivity attribute, which is used to describe the syntactic behavior of the verb. The Arabic lexicon classifies verbs according to transitivity into two main classes, intransitive verbs and transitive verbs. The intransitive verbs are in turn classified into unaccusative verb whose syntactic subject is not the semantic agent; but a semantic object, as in the sentence "انكسر الزجاج" 'the glass was broken', and unergative verb whose subject is the agent, as in the sentence "أكل الولد" 'the boy ate'. Transitive verbs are further classified into four types, direct transitive; a verb which takes a subject and a single direct object, as in "أحضر الولد الطعام" 'the boy brought the food' indirect transitive; a verb which takes a subject and a single indirect object, such as the verb"وافق" 'agree' in "وافق الأب على الذهاب" 'the father agreed on going', di-transitive; a verb which takes a subject and two objects, such as the verb "أعطى" 'gave' in " أعطى الأستاذ هدية للتلميذ" 'the teacher gave a present to the student'. Some other verbs are without transitivity as copula verb such as “كان” ‘was’ and “أصبح” ‘became’. Figure (1) shows the environment of the UNL dictionary, this dictionary contains 1433 indirect transitive verbs. The verb "وافق" 'agree' is one of the search results for the indirect transitive verbs; 'TRA=TSTI', it takes the preposition "على" in its sub-categorization frame; 'FRA=Y17'.

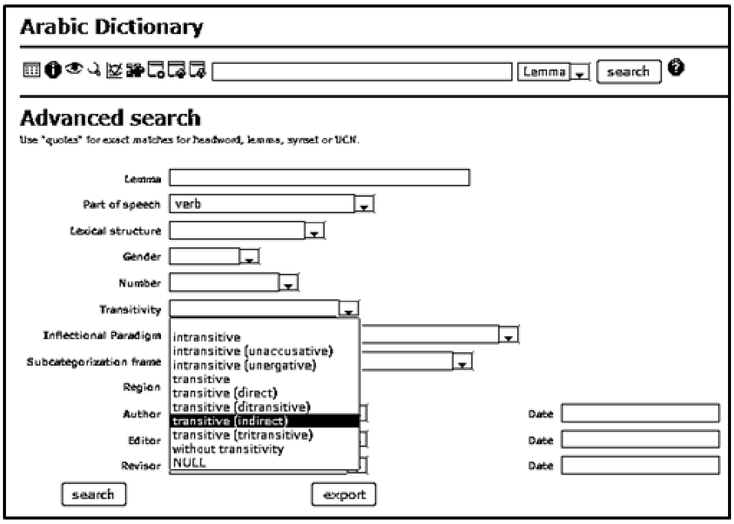


Figure 1: The UNL Arabic dictionary environment

Figure (2) shows the search results for the indirect transitive verbs in the UNL Arabic dictionary, it contains 1,433 verbs. The verb "وافق" 'agree', is one of the selected verbs with their synonyms and its sub-categorization frame.

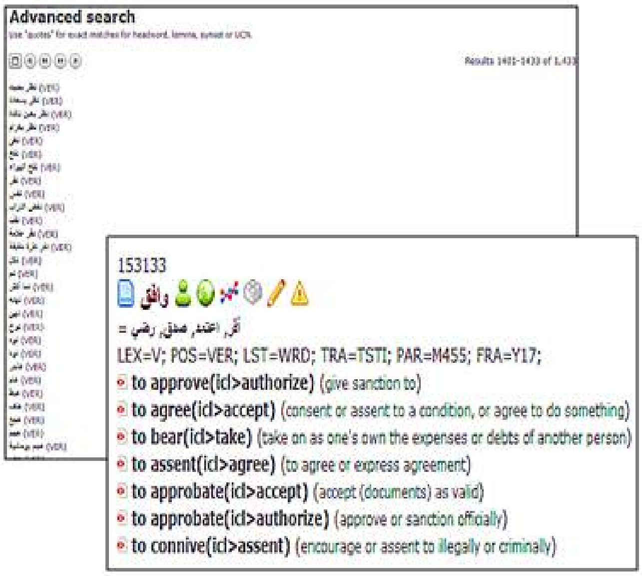


Figure 2: The search results and the verb "وافق" in the UNL Arabic dictionary

After extracting the verb with its syntactic behavior, its semantic class should be recognized by applying the Role and Reference Grammar (RRG) verb schema. RRG used by a wide range of syntax-semantics systems, therefore, the researcher uses the same classification. Passing through the four semantic classes of the Role and Reference Grammar verb schema with the scenario drawn above, the UNL semantic relations are taken into consideration. UNL doesn't contain the "CAU"; causer semantic relation as it can be expressed instead by "agt"; agent relation.

TABLE II
THE SYNTAX-SEMANTICS VERB COMPATIBILITY FOR VERBS COLLECTION

syntax / Semantics	Intransitive	Direct transitive	Indirect transitive	di-transitive	Copula
Action	X	أكل – كسر	وافق – شارك - سافر - ذهب	أقنع – طلب – فصل – ربط	X
State	نام – سكت	أحب – خسر – بلغ – استغرق	احتوى - اشتمل	X	كان - أصبح
Achievement	انتهى – انكسر	X	تحول - تسبب	انفصل – ارتبط	X

There is a consensus among researchers that assignment of thematic roles to the arguments of the predicate imposes a classification on the verbs of languages. Since the type of thematic roles and their number are determined by the meaning of the verb, the lexical decomposition of verb meanings seems to be a prerequisite for syntax-semantic classification of verbs of languages. Since the type of thematic roles and their number are determined by the meaning of the verb, the lexical decomposition of verb meanings seems to be a prerequisite for syntax-semantic classification of verbs. AnCora, the practical syntax-semantics system, characterizes verbs, by means of a limited number of LSS and Event Structure Patterns, according to the four basic event classes: states, activities, accomplishments, and achievements. The general classes can be split into subclasses. It is observed that there are three groups of verbs were not considered in the initial collection of verbs, which explains why some cells in table (2) above, contains four verbs instead of two. For example, the di-transitive action verbs "فصل" 'separate' and "ربط" 'connect', their second objects are always mapped to the 'cao' ;co-object semantic relation, and not 'gol' goal semantic relation (this issue will be discussed in detail in section 2.4).

B. *Corpus Compilation*

In data collection, data are collected from the Egyptian newspapers; Al-Ahram 1999 as it is representative for the Egyptian modern standard Arabic. The pages of Al-Ahram are collected on the Arabicorpus website; a website that allows the researcher to search large, untagged Arabic corpora.

In order to collect an appropriate size of data for linguistic analysis, the size of the corpus to be analyzed has to be precisely estimated; it should not be too small because it would raise the risk of not containing enough data. On the other hand, the corpus should not be too large either, since the time needed for analysis has to be also taken into account when planning corpus building. This corpus is 400 sentences covering 40 Arabic verbs. Sentences are 12 words long to contain all verbs arguments with their modifiers. The average number of sentences is 7 sentences for each verb which differ according to the nature of the verb itself; if it is an intransitive verb, the number of its arguments is less than that of transitive verb.

C. Corpus Analysis

Manual linguistic analysis has been achieved for the corpus for two reasons: first, to conclude the syntactic specifications of verbs arguments, second, to map each syntactic argument with the suitable semantic relation. Therefore, the dependency approach has been adopted for representing Arabic syntactically, as it is suitable for free word order languages and its representation is very close to the semantic relations (both represented by binary relations) which facilitate the semantic representation. The words of the sentences in the corpus are linked using the Quranic Arabic Dependency Treebank tag set. The sentence in (1) is analyzed syntactically to the dependency graph in figure (3):

(1) شاهد الفنان هذا الفيلم في مهرجان القاهرة السينمائي منذ أربعة أشهر

'the artist watched this film in Cairo film festival four months ago'

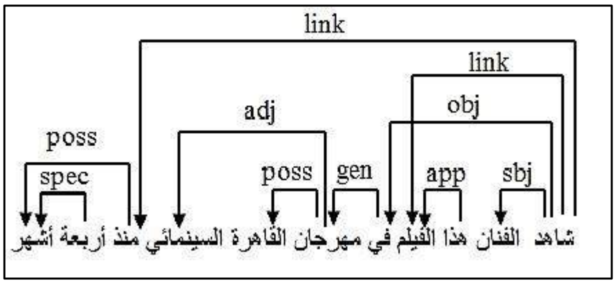


Figure3: The syntactic representation for "شاهد الفنان هذا الفيلم في مهرجان القاهرة السينمائي منذ أربعة أشهر"

All the dependency syntactic relations in the sentence are going to be mapped to the semantic relations of UNL as in figure (4). UNL is using two methods to represent the relations. First, the UNL attributes: which are arcs linking a node to itself. They correspond to one-place predicates, i.e., functions that take a single argument. In UNL, attributes have been normally used to represent information conveyed by natural language grammatical categories (such as tense, mood, aspect, number, etc). In figure (4), the UNL attributes '@proximal' and '@since' are assigned respectively to the nodes of "الفيلم" and "أشهر". Second, the UNL relations; they are labeled arcs that connect a node to another node in a semantic graph. The valency bound semantic relations in the graph are the 'agt'; the agent or doer of the event of watching which is mapped to the syntactic relation subject; sbj, and the obj; the object or the affected by the event which is mapped to the syntactic object semantic relation. The valency free relations are the plc; place and tim; time semantic relations.

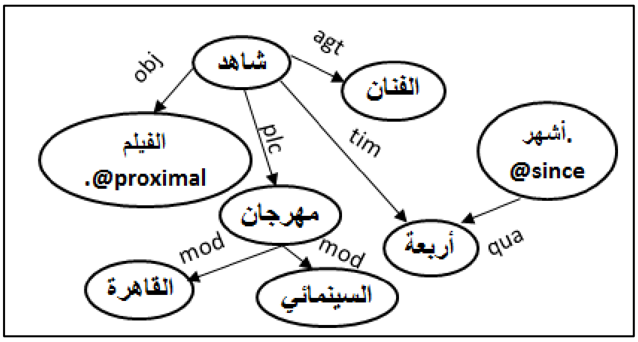


Figure 4: The semantic representation for "شاهد الفنان هذا الفيلم في مهرجان القاهرة السينمائي منذ أربعة أشهر"

All the corpus sentences are analyzed syntactically and the syntactic relations are mapped to the corresponding semantic relations to enable verb grouping according to their syntactic and semantic behavior as will be discussed in the following section.

3 VERBS SYNTAX-SEMANTIC CLASSIFICATION

The distinction between complements and modifiers is often defined, in terms of valency. Valency is considered a central notion in the theoretical tradition of dependency grammar which means that the verb imposes requirements on its syntactic dependents that reflect its interpretation as a semantic predicate. Dependents that correspond to arguments of the predicate can be obligatory or optional in surface syntax but can only occur once with each predicate instance. By contrast, dependents that do not correspond to arguments can have more than one occurrence with a single predicate instance and tend to be optional.

(1) **V**$_{\text{شاهد}}$ **- N** $_{\textbf{head (sbj) -}}$ **N** $_{\textbf{head (obj).}}$
(2) **V**$_{\text{شاهد}}$ **- N** $_{\textbf{agt –}}$ **N** $_{\textbf{obj.}}$

Returning to Figure 4, the subject “الفنان” ‘artist’ and the object ”الفيلم” ‘film’ would be normally treated as valency-bound dependents of the verb "شاهد" ‘watch’, while all the other sentence elements are considered as valency-free dependents. Accordingly, the sub-categorization frame of the verb “شاهد” in (2) can be concluded. The valency-bound dependents syntactic relations have been mapped to the UNL semantic relations as in (3). The subject is mapped to the agent or the doer of an action (agt), and the syntactic object is mapped to the semantic object or the affected thing by the event. Similarly, all the corpus has been analyzed to group its verbs according to their syntax-semantics behavior.

The valency-bound dependents in the sentence in (4) are different from those in (2) as they contain a preposition as in (5). The verb “وافق” ‘agree’ is an indirect transitive verb and accordingly, its syntactic behavior is different from that of the verb “شاهد” ‘watch’. Therefore, such different syntactic behavior will be classified under another syntactic classification while being under the same semantic classification.

(3) وافق مجلس الكليةعلى عقد ندوات
'The faculty board agreed on holding symposia’

(4) **V**$_{\text{وافق}}$ **- N** $_{\textbf{head (sbj) -}}$**PREP** $_{\textbf{head -}}$ **N** $_{\textbf{dependent.}}$
(5) **V**$_{\text{وافق}}$ **- N** $_{\textbf{agt -}}$**N** $_{\textbf{Obj.}}$

The following sub-sections exhibit the proposed classification as a coarse grained classification. We have only considered the productive Arabic verbs and there general syntactic structures. Verb ambiguity phenomenon has been left out as it needs more research based on this present research.

Lexical semantic structure (LSS) determines the number of arguments that a verbal predicate requires and the thematic role of these arguments, and describes the syntactic function of the mapped arguments. We will present the specific LSS derived from the general semantic event classes discussed in section (2). These LSS are the result of combining the general class with the argument structure and the thematic roles that can fill each argument slot. There are 12 LSS compiled and described, grouped around the 3 general event classes.

A. *LSS (A): Action Verbs*

This general event structure is sub-divided into six sub-classes: agentive-object (A1), transitive-agentive-P-object (A2), di-transitive-agentive-object-goal (A3), di-transitive-object-co-object (A4), transitive-agentive-place (A5). As can be seen, action verbs are related to the agentive subjects.

LSS A1: Agentive-object verbs
Sbj=agt

Obj=obj

Arabic verbs: “كسر” `break`, “فتح” `open`, “كتب ”`write', . . .

In this class, the second argument 'obj' is assigned to object thematic relation. Its syntactic function is always as direct object. The first argument is syntactically the subject, and is assigned to agent thematic relation 'agt'.

(6) كسرالرجلالزجاج 'the man broke the glass'

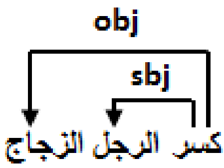


Figure 5:Agentive object verb in the syntactic graph

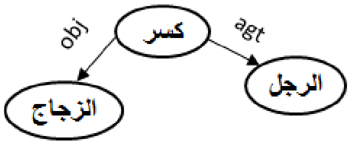


Figure 6: Agentive-object verb in the semantic graph

LSS A2 : Agentive-indirect object
Sbj = agt

Prepositional dependent = obj

Arabic verbs: "وافق" ’agree’, "شارك" 'participate',"أصر" 'insist'…

In this class, the prepositional argument is assigned to object thematic relation. Its syntactic function is always as a prepositional object, and may be introduced by a variety of prepositions. The prepositional object is realized in the dependency grammar as a 'gen' relation between the preposition and the following head noun. The head noun is the semantic object of the verb in class A2. As for the rest of the verbs in A class, the first argument is syntactically the subject, and is assigned to agent thematic role.

(7) وافق المجلس على الاقتراح 'the board agreed on the suggestion'

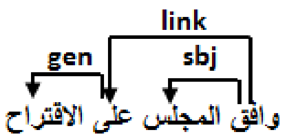


Figure 7. Agentive-indirect object verb in the syntactic graph

Figure 8. Agentive- Indirect object verb in the semantic graph

LSS A3: Agentive- object – goal
Sbj=agt

Obj=obj

Prepositional dependent =gol

Arabic verbs: “أقنع ” `persuade', “أحث” `motivate', …

This type of verbs requires two arguments in addition to the agent. The object syntactic argument in A3 class is assigned to object thematic role and is always the direct object. The prepositional argument is always assigned to the goal thematic role.

(8) يحث الرئيس المواطنين على المشاركة 'the president encourages citizens to participate'

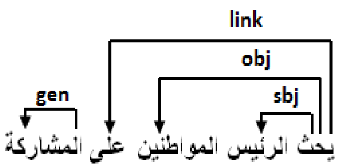


Figure 9. Agentive-object-indirect object verb in the syntactic graph

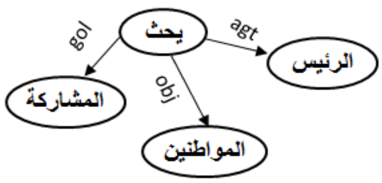


Figure 10. Agentive-object-indirect object verb in the semantic graph

LSS A4: Agentive-object-co-object
sbj=agt

obj=obj

Prepositional dependent=co-obj

Arabic verbs: “فصل” `separate', “ربط” `to connect', “حاذى” ‘ally',

This type of verbs; commutative verbs, requires two arguments in addition to the agent. The object thematic role is assigned to the syntactic object in A4 class and is always the direct object. The co-object thematic role is assigned to the prepositional argument.

(9) يفصل الفندق الأهلي عن المصري 'the hotel separate El-Ahly from El-Masry '

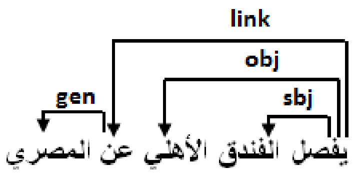


Figure 11. Agentive-object-co-object verb in the syntactic graph

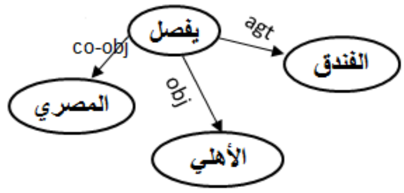


Figure 12. Agentive-object-co-object verb in the semantic graph

LSS A5: Agentive –place

Sbj=agt

Prepositional dependent=plc

Arabic verbs: “ذهب” `go', “سافر” ` exit', “وقف” ` stop',

The subject is associated to the agent thematic role. The prepositional argument is considered as an optional argument and associated to the location thematic role; 'plc' semantic relation and the preposition UNL attribute should be assigned to the location words. For example, @to should be assigned to "نيوجيرسي" ‘New Jersey’ to specify exactly that place; " إلى نيوجيرسي" ‘to New Jersey’ and not "في نيوجيرسي" ‘in New Jersey’

(10) سافر كلينتون إلى نيوجرسي 'Clinton travelled to New Jersey'

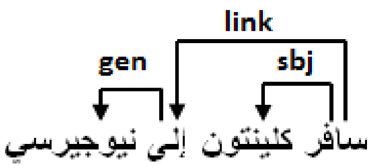


Figure 13. Agentive-place verb in the syntactic graph

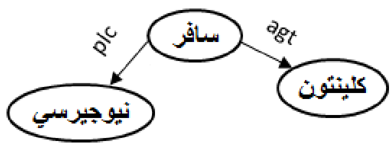


Figure 14. Agentive-place verb in the semantic graph

B. *LSS (B): State Verbs*

This general event structure is subdivided into four classes: state- experiencer (B1), state-experiencer-P-object (B2), state-experiencer-amount (B3), state-existential or attributive (B4). They take an internal argument, which appears as syntactic subject bearing the semantic role of an experiencer.

LSS B1: inergative – experiencer

sbj =exp

Arabic verbs: “نام” `sleep, “بكى” `to cry, . . .

Arg0 is syntactically the subject, and its thematic role is experiencer.

(11) نام الأطفال 'the children slept'

Figure 15. inergative - experiencer verb in the syntactic graph

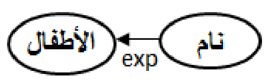


Figure 16. inergative - experiencer verb in the semantic graph

LSS B2: experiencer-object

Sbj=exp

Obj=obj

Arabic verbs: أحب `to love, كره `to dislike', خسر`to lose', . . .

In this class, the thematic object relation is assigned to the syntactic object. It is always a direct object. The experience thematic role is assigned to the subject.

(12) أحب المصري الأرض 'the Egyptian liked the land'

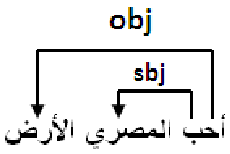


Figure 17.experiencer-object verb in the syntactic graph

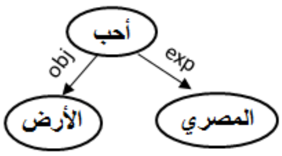


Figure 18. experiencer-object verb in the semantic graph

LSS B3: experiencer- Indirect object

Sbj=exp

Prepositional dependent=obj

Arabic verbs: "احتوى" `to contain, "اشتمل" `to include`.

In this class, the object thematic relation is assigned to the prepositional dependent. Its syntactic function is always a prepositional object, and may be introduced by a variety of prepositions. The first argument is syntactically the subject, and is assigned experiencer thematic role.

(13) يحتوي القصر على استراحتين 'the palace contains two Lounges'

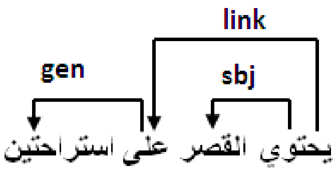


Figure 19. experiencer- indirect object verb in the syntactic graph

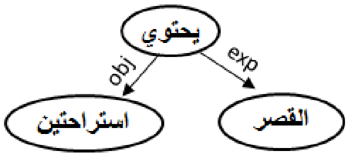


Figure 20. experiencer- indirect object verb in the semantic graph

LSS B4: Experiencer-amount

Sbj=exp

Obj=ext

Arabic verbs: استغرق ` to last', يزن `to weigh'.

The syntactic object may either be a direct object, an adjunct or a prepositional object. It maps with extension thematic role, an argument referring to some sizable and measurable magnitude such as length, weight, time, price, etc. It is observed that, verbs in this class may accept a direct object complement, but passive alternation is not possible.

(14) استغرقت الرحلة يومين 'the journey took two days'

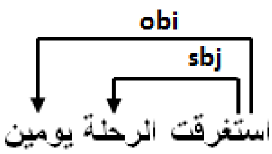


Figure 21. experiencer-amount verb in the syntactic graph

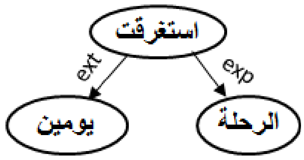


Figure 22. experiencer-amount verb in the semantic graph

LSS B5: state-attributive

Arabic verbs: "كان" ` be', "أصبح" `become',

The verbs "Being, becoming, and remaining" in Arabic have a special status since these verbs resemble each other in meaning and in syntactic effect. They describe states of existence (e.g., being, inception, duration, continuation) and each of them requires the accusative marker on the predicate or complement (xabarkann-a خبر كان), e.g. "كان أستاذاً". This kind of verbs indicates time only as opposed to main or real verbs like "أكل" 'eat' which indicate both meaning and time.

Ibn Jinnii, al-Jurjani, and Ibn al-Sarraj describe them as unreal verbs (copula), with no real function or a significant contribution to the meaning unlike main verbs that have two significances: significance of time configured in its form and significance of the event (concept of doing or taking an action). While "ʔal-ʔafʕaalʔal-naasixa" indicate only the time of the event expressed. Ibn al-Sarraj stated that real (main) verbs indicate both meaning and time as opposed to auxiliary verbs like "كان"(kaana) that indicate time only and that are dependent on the main verb.

Verbs of seeming or appearing also mark their complements with the accusative case, but they are not usually classified among the "sisters" of kaan-a. They do not have syntactic arguments (subject and object do not exist), but they have subject and predicate, they are not mapped to thematic relations as the tense is expressed via tense attribute.

(15) كانبوشرئيسا 'Bush was a president'

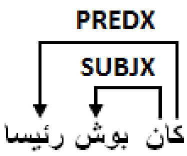


Figure 23. state-attributive verb in the syntactic graph

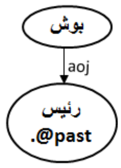


Figure 24. state-attributive verb in the semantic graph

C. *LSS (C): Achievement Verbs*

This general event structure is subdivided into two sub-classes: un-accusative (C1) and un-accusative-state (C2), depending on the constant they associate with (either place or state). Un-accusative verbs are basically monadic in terms of their LSS and in terms of their argument structure, taking a single internal argument (Arg1). Unaccusativity, it is related to the fact that the grammatical subject of an unaccusative verb behaves as the direct object of a transitive verb, consequently, the subject of an unaccusative verb and the object of a transitive verb bear the same semantic role: object for passives.

LSS C1:un-accusative –object

Sbj=obj

Arabic verbs: "ارتفع" `high, "هبط" `to exit', "توقف" `to stop', . . .

The object thematic -role is assigned to the subject of the verb.

(16) سقط المبنى 'the building has fallen'

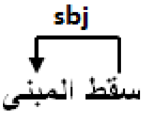


Figure 25. un-accusative -object verb in the syntactic graph

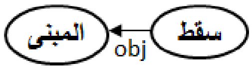


Figure 26. un-accusative -object verb in the semantic graph

LSS C2: un-accusative-state

Sbj=obj

Prepositional dependent=gol

Arabic verbs: "تلون" `cause', "تحول" 'convert', "ترقى" 'promote'

The object thematic role is assigned to the subject. C2 class is characterized with its prepositional dependent, which may be optional or mandatory and mapped to the final state thematic role 'gol' or, alternatively to initial state role 'src'.

(17).تحول الثلج إلى ماء 'the ice transformed into water'

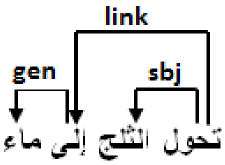


Figure 27. un-accusative - state verb in the syntactic graph

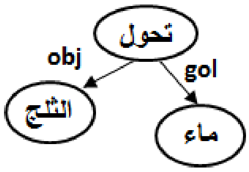


Figure 28. un-accusative -state verb in the semantic graph

LSS C3: un-accusative-co-object

Sbj=obj

Prepositional dependent =co-obj

Arabic verbs: "انفصل" `to separate', "ارتبط" `to associate'....,

In this case, the class is formed by verbs from the A5 class which has undergone the co-object, and shares arguments and thematic-roles with it: arg1 is the subject, with object thematic role, and arg2 is a prepositional object with co-object thematic role. It is not possible to have the Arg0 expressed in this LSS.

(18)انفصلت فنلندة عن روسيا 'Finland separated from Russia'

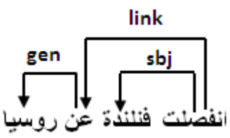


Figure 29. un-accusative –co-object verb in the syntactic graph

Figure 30. un-accusative –co-object verb in the semantic graph

4 IMPLEMENTING CLASSIFICATION

Verbs are coded according to the syntax-semantics grouping in the dictionary as in table (3) below. A, B, and C represent the semantic classes of the verbs, in other words which semantic verb class requires its subject to be mapped to an agent, experiencer, or object. Ax (A1, A2,…, A5) represent the description of the syntactic structure of the semantic class. For example, A4 requires a subject, direct object (N), and indirect second object (PP).

TABLE III

SYNTAX-SEMANTICS CLASSIFICATION FOR VERBS IN THE DICTIONARY

Semantic class	Syntactic structure	Mapping schema	example
Action verbs (A)	subject-direct object (A1)	Agent – object	فتح الطالب الباب
	subject-prepositional dependent (A2)	Agent – object	وافق المجلس على المشاركة
	subject- object – prepositional dependent (A3)	Agent – object – goal	يحث الاتحاد المواطنين على الموافقة
	Subject- object – prepositional dependent (A4)	Agent – object – co-object	تفصلهم الحدود عن قراهم
	Subject- prepositional dependent (A5)	Agent – place	سافر كلينتون إلى نيوجيرسي
State verbs (B)	subject (B1)	Experiencer	بكى الأطفال
	subject- object (B2)	Experiencer – object	أحب المصري الأرض
	subject-prepositional dependent (B3)	Experiencer- object	يحتوي القصر على استراحتين
	subject- object [amount] (B4)	Experiencer- extension	استغرقت الرحلة ثلاث ساعات
	Subject– predicate (B5)	Attribute (aoj)	كان بوش رئيس المخابرات
Achievement verbs (C)	subject (C1)	Object	هبطت القاعدة
	subject- prepositional dependent (C2)	Object-goal	تحول الثلج إلى ماء
	subject prepositional dependent (C3)	Object-co-object	انفصلت فنلندة عن روسيا

This classification has been applied to automatically analyze the corpus syntactically and semantically. Grammar modules have been developed in the integrated analysis environment; IAN analyzer. The grammar has common modules such as; the tokenization, morphological, syntactic, and syntax-semantic mapping modules. Figure (31) shows the result of the dependency syntactic representation and the semantic interpretation output for the verb "يحتوي" 'contains'.

```
[S: 3 ]
   { org }
    يحتوي المركز أيضا على ورشة كاملة مجهزة
   {/ org }
   { unl }
sbj ( 15: المركز ، 01: يحتوي )
link ( 06: أيضا ، 01: يحتوي )
link ( 08: على ، 01: يحتوي )
gen ( 10: ورشة ، 08: على )
adj ( 12: كاملة ، 10: ورشة )
adj ( 14: مجهزة ، 10: ورشة )
   {/ unl }
[/S]
```

Figure 31. The output syntactic representation for "يحتوي المركز أيضا على ورشة كاملة مجهزة"

```
[S: 3 ]
   { org }
    يحتوي المركز أيضا على ورشة كاملة مجهزة
   {/ org }
   { unl }
exp ( 15: المركز ، 01: يحتوي )
man ( 06: أيضا ، 01: يحتوي )
obj ( 10: ورشة ، 01: يحتوي )
mod ( 12: كاملة ، 10: ورشة )
mod ( 14: مجهزة ، 10: ورشة )
   {/ unl }
[/S]
```

Figure 32. The output semantic representation for "يحتوي المركز أيضا على ورشة كاملة مجهزة"

5 CONCLUSIONS

Considering the importance of studying the syntax – semantic interface in natural language understanding, the researcher suggests further researches to be conducted in this domain especially testing the proposed Arabic based syntax-semantics verb classification using more verbs. Besides, the researcher advocates the application of the proposed mapping system in this study to other linguistic registers and genres, such as newspapers articles, magazines, movies, etc. in pursuit of new observations, conclusions, and possibly findings that might enrich the semantic mapping. The grammar using the proposed classification displayed a high level of success and component performance; accuracy of results amount to 92% of the total number of the mapped syntactic structures. That is, merely 8% of the corpora fail to be correctly mapped to the semantic graph.

REFERENCES

[1] L.Bloomfield, *Language*, New York: Henry Holt. **1933**.
[2] B. Levin, *English verb classes and alternations : a preliminary investigation*. Univ. of Chicago Press. **1993.**
[3] S. Pinker, *Learnability and Cognition: The acquisition of argument structure*, MIT Press, **1989**.
[4] A.Goldberg, *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press, **1994**.
[5] W. Chafe , *Meaning and the structure of languag*. University Press, Chicago, **1970**.
[6] W. A.Cook, *Case Grammar: development of the Matrix Model* , Georgetown University Press, **1979**.
[7] R. E.Longacre,., *An anatomy of speech notions*, Peter de Ridder Press, **1976**.
[8] W. Foley, Van valin, R., *Functional syntax and universal grammar*. Cambridge: Cambridge University Press, **1984**.
[9] R. D.Van Valin, *A Synopsis of Role and Reference Grammar, in Advances in Role and Reference Grammar*. Van Valin (ed.), John Benjamins Publishing Company, Amsterdam., **1993**.
[10] D. Dowty, " *On the Semantic Content of the Notion of 'Thematic Role'*", in Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, eds., Properties, Types, And MeaningII, Kluwer, Dordrecht, **1998**.
[11] A.Sanfilippo , V.Poznanski, *The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, **1992**.
[12] Z.Vendler, *Adjectives and Nominalization*. La Haye:Mouton, **1968**.
[13] K.Kipper , M.Palmer, O. Rambow , *Extending PropBank with VerbNet Semantic Predicates.* In Workshop on Applied Interlinguas, **2002**.
[14] M.Palmer, P.Kingsbury, and D. Gildea, *The Proposition Bank: An Annotated corpus of semantic roles*. Computational Linguistics, 31, 71-106, **2005**.
[15] A.Peris, and M.Taule, , *AnCora-Nom: A Spanish lexicon of deverbal nominal- izations*. Procesamiento del Lenguaje Natural, **2011**.
[16] S.Alansary, *MUHIT: A Multilingual Lexical Database*", 13th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt, **2013**.

BIOGRAPHY

Israa Elhosiny

Principal Grammar Developer of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. She is working in the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.
She obtained her BA, Department from Phonetics and Linguistics 2004.
She obtained her M.A., Department from Phonetics and Linguistics 2015.

She has an experience in morphological analysis and generation, text tokenization, POS tagging and disambiguation. She participated in building grammars using UNL for library information system (LIS) and Knowledge Extraction sYStem (Keys).

She is a member in the following scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Sameh Alansary

Director of Arabic Computational Linguistics Center Bibliotheca Alexandrina

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

نحو تصنيف تركيبي- دلالي للأفعال العربية من أجل الوسم الدلالي

إسراء الحسيني[1], سامح الأنصاري[2]

مكتبة الإسكندرية ، الإسكندرية ، مصر
قسم الصوتيات واللسانيات، كلية الآداب جامعة الإسكندرية
[1]israa.elhosiny@bibalex.org
[2]sameh.alansary@bibalex.org

ملخص – إن الوسم الدلالي للمركبات الفعلية يتطلب ربطا نمطيا بين التركيب والدلالة. لذلك، كان تصنيف الأفعال تبعا لسلوكها النحوي والدلالي هدفا أساسيا لإجراء هذه الدراسة، حيث أن هذا التصنيف- في وجود التمثيل النحوي- يجعل الوسم الدلالي أكثر سهولة ومرونة. فمن أجل إجراء هذا البحث، قام الباحث بجمع عينة لمدونة عربية وتحليلها على المستوى النحوي والدلالي يدويا من أجل تحديد السمات التركيبية والدلالية لكل فعل من أفعال العينة ومن ثم بناء التصنيف المقترح بُناء على الصفات المشتركة لهذه الأفعال. وقد تم استخدام التصنيف بعد ذلك عن طريق بناء معجم حاسوبي حتى يساهم وبناء قواعد التحليل الآلي في اختبار التصنيف ذاته في الوصول للوسم الدلالي الآلي. وقد تبين من مراجعة نتائج التحليل أن التصنيف المقترح حقق نسبة صحة 92% في التعرف على الأدوار الدلالية لأفعال المدونة ، أي نحو 8% نسبة خطأ في الوصول للدور الدلالي.

Automatic Diacritization for Modern Standard Arabic

Amany Fashwan[1], Sameh Alansary[2]

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

[1]amany.fashwan@bibalex.org

[2]sameh.alansary@bibalex.org

***Abstract*—Arabic language is receiving growing attention in the NLP community. Modern Standard Arabic is written with an orthography that includes optional diacritical marks (henceforth, diacritics). Diacritics are extremely useful for readability and understanding. The issue of diacritization in Arabic arises as the result of a mismatch between the orthographic conventions that have developed for written MSA and the Arabic language itself, including spoken MSA, with respect to the amount of linguistic information represented. The main objective of this paper is to build a system that would be able to diacritize the Arabic text automatically. In this system the diacritization problem will be handled through two levels; morphological and syntactic processing levels. This will be achieved depending on an annotated corpus for extracting the Arabic linguistic rules, building the language models and testing system output. The adopted technique for building the language models is "Bayes', Good-Turing Discount, Back-Off" Probability Estimation.Precision and Recall are the evaluation measures used to evaluate the diacritization system. At this point, precision measurement was 89.1% while recall measurement was 93.4%on the full-form diacritization including case ending diacritics.These results are expected to be enhanced by extracting more Arabic linguistic rules and implementing the improvements while working on larger amounts of data.**

1 INTRODUCTION

Arabic is currently the sixth most widely spoken language in the world with estimated 422 million native speakers. As the language of the Qur'an (the holy book of Islam), it is also widely used throughout the Muslim world. It belongs to the Semitic group of languages, which also include Hebrew and Amharic (the main language of Ethiopia). It is considered a member of a highly sophisticated category of natural language, which has a very rich morphology, where one root can generate several words having different meanings.

Arabic language is receiving growing attention in the NLP community, due toits socio-political importance and the NLP challenges presented by its dialect differences, diglossia[1], complex morphology, and non-transparent orthography. But like most languages, Arabic is lacking in annotated resources and tools. Fully automated fundamental NLP tools such as tokenizers, part of speech taggers, parsers, and semantic role labelers are still unavailable for Arabic. [1]

Arabic is a language of rich morphology, both derivational and inflectional. Due to the fact that the Arabic script usually does not encode short vowels and omits some other important phonological distinctions, the degree of morphological ambiguity would be very high. In addition to this complexity, Arabic orthography prescribes to concatenate certain word forms with the preceding or the following ones, possibly changing their spelling and not just leaving out the white space in between them. This convention makes the boundaries of lexical or syntactic units, which need to be retrieved as tokens for any deeper linguistic processing, obscure, for they may combine into one compact string of letters and be no more the distinct 'words'.

Modern Standard Arabic is written with an orthography that includes optional diacritical marks (henceforth, diacritics). Diacritics are extremely useful for readability and understanding. Their absence in Arabic text adds another layer of lexical and morphological ambiguity. Naturally occurring Arabic text has some percentage of these diacritics present depending on genre and domain. They are there to aid the reader disambiguate the text or simply to articulate it correctly. For instance, religious text such as the Quran is fully diacritized to minimize the chances of reciting it incorrectly [3].

Diacritization is even more problematic for computational systems, adding another level of ambiguity to both analysis and generation of text. For example, full vocalization is required for text-to-speech applications, and has been shown to improve speech- recognition perplexity and error rate. [4]

The issue of diacritization in Arabic arises as the result of a mismatch between the orthographic conventions that have developed for written MSA and the Arabic language itself, including spoken MSA, with respect to the amount of linguistic information represented. [5]

1Diglossia refers to a situation in which two dialects or languages are used by a single language community.

Predicting the correct diacritization of the Arabic words elaborates the meaning of the words and leads to better understanding of the text, which in turn is much useful in several real life applications.

Automatic words diacritization (aka vowelization, diacritic/vowel restoration) is one of the NLP challenges with languages having diacritics unveiling the phonetic transcription of their words. Arabic is an example of such languages where different diacritics over for the same spelling produce different words with may be of different meanings (e.g. عِلْم "science", عَلَم "flag", عَلَّم "taught", عَلِمَ "knew" … etc.). [6]

There are many challenges that make the task of building a reliable Arabic diacritizer is a hard one. It is found that Modern Standard Arabic texts are typically written without diacritics and are commonly written with many common spelling mistakes such as (ا - أ), (ا - إ), (ه - ة), (ى - ي) ... etc. In addition, it is very difficult to have a training corpus that covers all of (or even most of) full diacritized word forms; however large the corpus will be, it will not be able to cover all full-word diacritized forms. About two thirds of Arabic text words have a syntactically dependent case-ending which invokes the need to a syntax analyzer which is a hard problem.

Although undiacritized Arabic text is sufficient for Arabic speakers to use in writing and reading, this is not the case when dealing with software systems. For example, an Arabic text-to-speech system would not produce speech from undiacritized Arabic text, because there is more than one way of saying the same undiacritized written Arabic word. Moreover, when searching for an Arabic word, many unrelated words would be included in the results.

This suggests the need to diacritize Arabic text. Another reason for the diacritization is to permit the use of dictionaries and machine translation from and to Arabic. For these reasons and many others, software companies that deal with Arabic realize the importance of developing a system for diacritizing the Arabic text. There are a few systems that are available on the market. However, they are not open source and usually integrated with other systems. [7]

Diacritic restoration has been receiving increasing attention and has been the focus of several studies. Different methods such as rule-based, data-driven techniques [8], example-based, hierarchical[9], morphological and contextual-based [10], [11], [3], [12], [13]as well as methods with Hidden Markov Models (HMM)[14], weighted finite state machines, machine learning techniques [2], SVM-statistical prioritized techniques [15] and other statistical techniques [16]have been applied for the diacritization of Arabic text.

In addition, there are some software companies that have developed commercial products for the automatic diacritization of Arabic. However, these products used only text based information, such as the syntactic context and possible morphological analyses of words, to predict diacritics [17]. Examples for the most representative commercial Arabic morphological processors; Sakhr Arabic Automatic Diacritizer [18], [19], Xerx's Arabic morphological processor [20] and RDI's Automatic Arabic Phonetic Transcriptor (Diacritizer/Vowelizer) [21], [22].

In addition to the previous commercial products there are some trials for producing free products for the automatic diacritization of Arabic. For example, Meshkal Arabic Diacritizer [23], Harakat Arabic Diacritizer [24] and Google Tashkeel which is no longer working where the tool is not available now.

After reviewing the existing Arabic diacritized systems it can be noticed that the diacritics classification can be divided into syntactical diacritization, caring about case ending and morphological diacritization, and caring about the rest of the word diacritics. So far, the morphological part of the problem is almost solved, leaving a marginal error of around 3-4%. On the other hand, syntactical diacritization errors are still high, hitting a ceiling that is claimed to be asymptotic and cannot be squeezed any further [25]. The following section will describe tagged corpus that the researcher used while building the diacritization system.

In this paper, the researcher will present an implemented system that takes any raw MSA text and generate its diacritized form. Section 2 details the description and processing of used corpus. Section 3 details the built Arabic diacritization system on both two processing levels; morphological and syntactic processing levels. Section 4 evaluates the output. Finally, section 5 concludes the paper.

2 CORPUS DESCRIPTION AND PROCESSING

There are two kinds of data sets are used here; the training data set which helps in building the diacritizer system and testing data set for evaluating that system. These data sets contain Modern Standard Arabic words, each word associated with its morphological features that uniquely specify the suitable internal diacritics of that word and the case ending diacritics. These data sets are chosen from International Corpus of Arabic (ICA) [26].Each word is tagged with features, namely, Lemma, Gloss, Pr1, Pr2, Pr3, Stem, Tag, Suf1, Suf2, Gender, Number, Definiteness, Root, Stem Pattern, Case Ending, Name Entity and finally Vocalization. It contains about 500,000 manually morphologically disambiguatedwords.

Good tagset design is particularly important for highly inflected languages. If all of the syntactic variations that are realized in the inflectional system were represented in the tag set, there would be a huge number of tags, and it would be practically impossible to implement or train a tagger. There are two criteria to distinguish the tag set design; external and internal criteria. The external criterion is that the tagset must becapable of making the linguistic (for example, syntactic or morphological) distinctions required in the output corpora. The internal criterion on tag sets is the design criterion of making the tagging as effective as possible [27].

Since the main target in this paper is to diacritize the Arabic text, the researcher made some normalization for the ICA tag set (which its target is morphological analysis)to be more normalized and effective in the diacritization system results. The normalization done for the prefixes tag set, the stem tags set and the suffixes tag set depending on the main tag sets of (ISO 12620) [28].The purpose of this tag set is providing the technical means for describing any linguistic behavior which should be done in a highly standardized manner, so that others could easily understand and exploit the data for their own benefit. The main intention is to create a harmonized system in order to make language resources as easily understandable and exchangeable as possible.

3 ARABIC DIACRITIZATION SYSTEM

This section aims to elaborate the methodology for building an Automatic Diacritizer for Modern standard Arabic texts beginning from the methodology for detecting the internal diacritics (morphological processing level), followed by the methodology for detecting the case-ending diacritics (syntactical processing level).

A. Morphological Level Processing

Morphological analysis techniques form the basis of most natural language processing systems. Such techniques are very useful for many applications, such as information retrieval, text categorization, dictionary automation, text compression, data encryption, vowelization and spelling aids, automatic translation, and computer-aided instruction.

Due to their non-concatenative nature, processing Semitic languages such as Arabic is not an easy task. For example, though Arabic words may be formed from concatenating morphemes, they are in fact normally formed using root pattern schemes. Morphologically, the Arabic language is a complicated and rich language. Tens or hundreds of words can be formed using one root, a few patterns, and a few affixes. Arabic also has a high degree of ambiguity for many reasons, such as the omission of vowels and the similarity of affixed letters to stem or root letters. Morphological analysis usually affects other higher levels of analysis such as syntactical and semantic analyses. [29]

It is important to distinguish between the problems of morphological analysis (what are the different readings of a word out-of-context) and morphological disambiguation (what is the correct reading of a word in a specific context). Once the morphological analysis is chosen in context the full POS tag, lemma and internal diacritics could be determined. So, the concern here is to select a model of morphological disambiguation to help in detecting the internal diacritics.

When trying to select a model of morphological analysis, there are two points that must be taken into consideration; firstly, the accuracy of morphological analysis systems where most morphological disambiguation systems consider the analyzers' output solutions as the input of their disambiguation systems. However, the accuracy of the morphological disambiguation process depends to a large extent on the ability of the analyzer to detect all possible solutions of the words. Secondly, what are the available morphological analyzers and disambiguation systems for research and evaluation?

There are many morphological analyzers for Arabic; some of them are available for research and evaluation while the rest are proprietary commercial applications. Among those known in the literature are Xerox Arabic Morphological Analysis and Generation [30], [31], [32], Buckwalter Arabic Morphological Analyzer [33],[30],[34],[35]Sakhr [33], [34], ArabMorpho (MORPHO3) [36], [32] and AlkhalilMorpho Sys[37]. The first two are the best known and most quoted in literature, and they are well documented and available for evaluation.

Among these systems there are two systems that are not commercial and can be used in morphological disambiguation process; Buckwalter Arabic Morphological Analyzer and AlkhalilMorpho sys. When trying to select between these two systems, some criteria have been taken into consideration. Firstly, which one of these systems is more helpful in producing solutions?

Secondly, when integrating one of these systems in the diacritization system, which one of these systems will be faster in retrieving the solutions of the input text?

Although BAMA has some disadvantage in its system but it has been selected as a model of analysis.The stem-based approach (concatenative approach) is adopted as a linguistic approach to analyze the input data. According to this linguistic approach, it was expected that a feature based on the right and left stems would lead to improvement in system accuracy. According to this adopted model in the morphological analysis, the word is viewed as composed of a basic unit that can be combined with morphemes governed by morphotactic rules. The three-part approach entails the use of three lexicons: Prefixes lexicon, Stem lexicon, and Suffixes lexicon. For a word to be analyzed, its parts must have an entry in each lexicon, assuming that a null prefixes or null suffixes are both possible.

There are some trials that used the database lexicons of BAMA in their disambiguation process [13],[38].The adopted morphological disambiguation algorithm here will be as of BAMAE [38] since in this algorithm the disambiguation process is done on two levels. The first level is the morphological disambiguating of the input words which detects the main tag of each word according its context. The second level is the semantic disambiguation level, resulted from missing the diacritics, which selects the suitable diacritic of each word according to its context.

The morphological level processing begins with disambiguating words that have one diacritized form; one morphological analysis without the case ending, and assigning this analysis to the word.For example, the word 'الاحتلال' has only one solution and hence one diacritized form 'الْإِحْتِلَال'. The second step of the system is extracting the relations among word forms or the analyzed words in the first step. The third step depends on extracting and implementing some Arabic linguistic rules to detect the analysis of some other words depending on the previous or the word if it is assigned with a certain tag. For example, if the word form to be analyzed is 'عمل' and the previous word's tag is preposition 'PREP' this word cannot be a verb or an adjective. So, this rule will eliminate all such solutions and the noun tag will be assigned in this case with the diacritized form 'عَمَل'.

In the previous example, the rule could detect one diacritized form, but this is not the case all the time. The Arabic extracted rules may eliminate the wrong solutions, but the remaining solutions would still be more than one. For example, if the word form is 'المدرسة' in the same previous condition, this case the rule will eliminate the adjective form of this word 'الْمُدَرَّسَة' but keep the noun forms of this word 'الْمُدَرِّسَة' and 'الْمَدْرَسَة'. In such case, when the rules fail to detect or choose one solution, the statistical model is applied. And the input solutions for the model will be the eliminated. This will reduce the solutions that the statistical model chooses among and hence reduce the mistake.

Moreover, this is not the only case for applying the statistical model. Another case is when the word to be analyzed has no rule to be applied over it. In this case, the input solution for the statistical model will be all of BAMA's solutions. If the word is assigned with the suitable analysis in this level, a rule that helps in disambiguate the next word could be applied, if it is not disambiguated yet.

Two important statistical features of any real text corpus have to be mentioned here:

1. Any finite-size text corpus, however large, is sparse. Sparseness means that; from all the possible m-gram combinations of the vocabulary words, a lot, in fact most, of these combinations occur rarely or do not occur at all within the text corpus.
2. Sparseness increases as m gets larger.

So, if a direct and naive, m-dimensional array is devoted to accommodate the occurrences of each of the possible *m*-grams in a training text corpus, the following two tough problems are encountered:

1. The needed storage for such an array is proportional toV^m; where V is the vocabulary size. If, for example, V=10000 (which is typically considered a small vocabulary size) and m=3, then the needed storage is prohibitively, and wastefully, large.
2. Due to sparseness, most of the elements of such an array, if ever implemented, will be either zeroes or very small. The minority of the elements which register considerable occurrences (neither zeroes nor very small numbers) can be regarded as reliable estimates of the actual frequency of the corresponding m-grams, whereas the majority zero or very small elements cannot be regarded as reliable. As computing m-gram probabilities directly relies on the frequency estimation, these estimates must be reliable enough.

The built language model in this level of the system depends on one of the effective techniques widely adopted today, namely "Bayes', Good-Turing Discount, Back-Off"Probability Estimation. It states that any entity in the language vocabulary must have usage in some context, though it seems endless to enlarge some corpus to cover all entities. The process of biasing the uncovered set on the expense of discounting the other regions is called smoothing. A disambiguation system that doesn't employ smoothing would refuse the correct solution if any of its input entities was unattested in the training corpus and consequently may miss the optimal (most likely) solution [32].This approach is applied depending on two phases; offline and runtime phases.

B. Syntactic Level Processing

Case ending diacritics play an important rule for understanding the meaning of Arabic statement where it gives the correct understanding of the statement.

The realization of nominal case in Arabic is complicated by its orthography, which uses optional diacritics to indicate short vowel case morphemes, and by its morphology, which does not always distinguish between all cases. Additionally, case realization in Arabic interacts heavily with the realization of definiteness, leading to different realizations depending on whether the nominal is indefinite, i.e., receiving nunation, definite through the determiner Al+ (ال+) or definite through being the governor of an EDAFAH possessive construction. [39]

In addition, case realization in Arabic interacts in some cases with other information; word pattern and feminine plural word forms. The diptote patterns in Arabic have special case where these words never receiving nunation. And, if these words are indefinite and genitive, the case ending will be fatha 'َ' not kasra 'ِ'. Concerning words that are feminine plural 'end with ات suffix', these words also have special case where if they are accusative, the case ending will be kasra 'ِ' not fatha 'َ'.

In order to set the case ending diacritics, a prior step is done where some Arabic linguistic rules have been extracted and implemented in the system to detect the definiteness of each word depending on its context or its selected morphological analysis. In addition, the stem pattern of each stem has been detected depending on the root, stem and lemma of each word.

After that, some Arabic linguistic rules that have been extracted from the training data set and implemented to detect the case ending depending on the context, the selected morphological analysis, definiteness feature and stem pattern feature, for each word. The difference between this stage and the previous stage of morphological processing level is that each extracted rule gives only one solution for the case ending. Consequently, those words do not have rules to assign them the suitable case ending and they will receive their case ending depending on the statistical approach.

The built language model in this level depends on the same adopted technique in the morphological processing level; "Bayes', Good-Turing Discount, Back-Off" Probability Estimation. The difference between the language models in both levels is the classifiers used in building the models. Figure 1 shows an example for the system's output:

Figure 1: Example for the system's output.

4 EVALUATION AND RESULTS

In this stage, the evaluation has been done using precision and recall measurements for 10% of the used corpus (50,000 words). A blind copy of the testing data set has been run using the diacritization system and then evaluated with its counterpart manually annotated data. It must be noted that the testing data set has never been used in extracting the Arabic linguistic rules or building the language models. Precision measurement was 89.1% while recall measurement was 93.4%. These results are expected to be enhanced by extracting more Arabic linguistic rules and implementing the improvements while working on larger amounts of data.

5 CONCLUSION

Most related works to diacritization depend in their systems on many of statistical approaches. This system is considered as a good trial to the interaction between rule-based approach and statistical approach, where the rules can help the statistics in detecting the right diacritization and vice versa. The evaluation has been done using precision and recall measurements for 10% of the used corpus. The results are expected to be enhanced by extracting more Arabic linguistic rules and implementing the improvements while working on larger amounts of data.

ACKNOWLEDGMENT

Thanking Bibliotheca Alexandrina for their permission to use a sample of their morphologically analyzed Arabic Corpus.

REFERENCES

[1] M. Diab, K. Hacioglu, & D.Jurafsky, *Automatic Processing of Modern Standard Arabic Text, In Arabic Computational Morphology* (pp. 159-179). Springer Netherlands.

[2] K. Shaalan, H. M. A. Bakr&I. A. Ziedan, *A Statistical Method for Adding Case Ending Diacritics for Arabic Text*. In Proceedings of Language Engineering Conference. (pp. 225-234). Cairo, Egypt,(2008, 17-18 December).

[3] M. Diab, M. Ghoneim& N. Habash. *Arabic diacritization in the context of statistical machine translation*. In Proceedings of MT-Summit. Copenhagen, Denmark,(2007, September)

[4] R. Nelken&S. M. Shieber, *Arabic diacritization using weighted finite-state transducers.* In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. (pp. 79-86). Association for Computational Linguistics. Michigan University, USA,(2005, June).

[5] M. Maamouri, A. Bies&S. Kulick,*Diacritization: A challenge to Arabic Treebank Annotation and Parsing.* In Proceedings of the Conference of the Machine Translation SIG of the British Computer Society. Linguistic Data Consortium, Pennsylvania University, USA, (2006).

[6] M. Rashwan, M. Al-Badrashiny, M. Attia&S. Abdou,*A Hybrid System for Automatic Arabic Diacritization*. In The 2nd International Conference on Arabic Language Resources and Tools. Cairo, Egypt, (2009).

[7] M. Alghamdi, Z. Muzaffar&H. Alhakami, *AUTOMATIC RESTORATION OF ARABIC DIACRITICS: A SIMPLE, PURELY STATISTICAL APPROACH.* The Arabian Journal for Science and Engineering, Volume 35, Number 2C. (pp. 125-135),(2010, Decemcer).

[8] A.Said, M. El-Sharqwi, A. Chalabi &E. Kamal. *A hybrid approach for Arabic diacritization. Natural Language Processing and Information Systems*. Lecture Notes in Computer Science, vol. 7934, (pp. 53–64). Springer, Berlin, (2013).

[9] O. Emam&V. Fisher, *A Hierarchical Approach for the Statistical Vowelization of Arabic text.* Technical report, IBM Corporation Intellectual Property Law, Austin, US, (2005).

[0] N. Habash&O. Rambow,*Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. (pp. 573-580). Association for Computational Linguistics. Ann Arbor, (2005, June).

[1] N. Habash&O. Rambow, *Arabic Diacritization through Full Morphological Tagging*. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (pp. 53-56). Association for Computational Linguistics. Rochester, NY, (2007, April).

[2] R. Roth, O. Rambow, N.Habash, M. Diab&C. Rudin.*Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking*. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (pp. 117-120). Association for Computational Linguistics. Columbus, Ohio, USA, (2008, June).

[3] N Habash, O. Rambow& R. Roth,*MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization*. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR). (pp. 102-109). Cairo, Egypt,(2009, April).

[4] Y. A. Gal, *An HMM Approach to Vowel Restoration in Arabic and Hebrew*. In Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages. (pp. 1-7). Association for Computational Linguistics,(2002, July).

[5] K. Shaalan, H. M. Abo Bakr &I. Ziedan,*A Hybrid Approach for Building Arabic Diacritizer*. In Proceedings of the 9th EACL Workshop on Computational Approaches to Semitic Languages. (pp. 27-35). Association for Computational Linguistics. Athens, Greece,(2009, March).

[6] M. Alghamdi&Z. Muzaffar, *KACST Arabic Diacritizer*. In Proceedings of the 1st International Symposium on Computers and Arabic Language (ISCAL). (pp. 73-79),(2007, 25-28 March).

[7] D. Vergyri&K. Kirchhoff,*Automatic diacritization of Arabic for Acoustic Modeling in Speech Recognition*. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING). (pp. 66-73). Association for Computational Linguistics. Geneva, (2004, August).

[8] ARAMEDIA Web Site: http://aramedia.com/nlp2.htm, [Accessed 24-11-2015].

[19]ARAMEDIA Web Site:http://aramedia.com/diacritizer.htm, [Accessed 24-11-2015].

[20] M. Al-Badrashiny, *AUTOMATIC DIACRITIZER FOR ARABIC TEXTS.* A Master's Thesis, Faculty of Engineering, Cairo University, (2009).

[2] M. A. Rashwan, M. Al-Badrashiny, M. Attia, S. M.Abdou&A. Rafea, *A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Un-factorized Textual Features*. Audio, Speech, and Language Processing, IEEE Transactions on, 19 (1), (pp. 166-175),(2011).

[22] RDI Web Site: http://www.rdi-eg.com/technologies/arabic_nlp.htm [Accessed 24-11-2015].

[23] Meshkal Arabic Diacritizer Web Site: http://tahadz.com/mishkal, [Accessed 24-1-2015].

[24]Harakat Arabic Diacritizer Web Site: http://harakat.ae/ [Accessed 24-11-2015].

[25] M. A. Rashwan, A. A. Al Sallab, H.M. Raafat& A. Rafea,*Automatic Arabic diacritics restoration based on deep nets*. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 65–72, Doha, Qatar, October 25, 2014.

[26] S. Alansary, M. Nagi&N. Adly, *Building an International Corpus of Arabic (ICA): progress of compilation stage*. In proceedings of the 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007.
[27] A. Feldman, *Tagset design, inflected languages, and n-gram tagging*. Editors: Paul Robertson and John Adamson, 3(1), 151, 2008.
[28 Wikipedia: https://en.wikipedia.org/wiki/ISO_12620 [Accessed 24-11-2015]
[29] I. A. Al-Sughaiyer&I. A. Al-Kharashi, *Arabic Morphological Analysis Techniques: A Comprehensive Survey*. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 55(3):189–213, February 1, 2004.
[30] Xerox Arabic Morphological AnalyzerWeb Site: https://open.xerox.com/Services/arabic-morphology, [Accessed 24-11-2015].
[3] S. Alansary, M. Nagi&N. Adly,*Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage*. In Proceedings of the 8th conference of The Egyptian Society Of Language Engineering (ESOLE). Cairo, Egypt, (2008, 17-18 December).
[32] M. Attia,*A large-scale computational processor of the Arabic morphology, and Applications*. A Master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt, (2000).
[33] BAMA Web Site: https://catalog.ldc.upenn.edu/LDC2004L02 [Accessed 24-11-2015].
[34] M. A. Attia,*An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks.* In proceedings of the Challenge of Arabic for NLP/MT Conference. The British Computer Society. (pp. 48-67). London, UK, (2006a).
[35] S. Alansary,*BAMAE: Buckwalter Arabic Morphological Analyzer Enhancer*, In proceedings of 4th international conference on Arabic language processing, Mohamed Vth University Souissi, Rebate, Morocco, May 2-3 2012.
[36]ArabMorpho Web Site: http://www.rdi-eg.com/technologies/Morpho.aspx, [Accessed 24-11-2015]
[37] M. S. S. Sawalha, *Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora*. University of Leeds, (2011).
[38] S. Alansary &M. Nagi, *The International Corpus of Arabic: Compilation, Analysis and Evaluation*. In the proceedings of EMNLP workshop. Doha, Qatar, (2014, August).
[39] N. Habash, R. Gabbard, O. Rambow, S. Kulick&M. P. Marcus, *Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features*. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. (pp. 1084–1092). Prague,(2007, June).

BIOGRAPHIES

Amany Fashwan: *Head of International Corpus of Arabic Unit, Arabic Computational Linguistic Center, Bibliotheca Alexandrina, Alexandria, Egypt.*

She graduated from Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University (2005). She participated with a team in building a tool for morphological analysis and generation of Arabic roots with excellent degree (field study). Her MSA thesis is in 'Automatic Diacritization of Modern Standard Arabic Texts: A corpus based approach'. Her main areas of interest are building Arabic corpora, corpus based studies, Arabic morphology, Arabic syntax and Arabic semantics. She has experienced in morphological analysis and extracting and implementing Arabic linguistic rules depending on morphologically analyzed Arabic words. She obtained a certificate for participation in 'The first annual forum for graduates at faculty of Arts, Alexandria University' in 2012.

Dr. Sameh Alansary:*Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina,Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

التشكيل الآلي للنصوص في العربية المعاصرة

Amany Fashwan[1], Sameh Alansary[2]

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

[1]amany.fashwan@bibalex.org

[2]sameh.alansary@bibalex.org

ملخص—**لقد جذبت العربية العديد من الأنظار لها في مجتمع المعالجة الآلية للغة العربية في الآونة الأخيرة. فاللغة العربية المعاصرة تكتب بدون علامات التشكيل مع وضع التشكيل في بعض الأحيان. فعلامات التشكيل في العربية مهمة جدا فبها يستقيم المعنى ويُفهم، ولذلك نجد العديد من الدراسات التي تتجه إلى التشكيل الآلي للغة العربية وذلك لوجود العديد من الأشكال لنطق الكلمة الواحدة الغير مشكّلة. تركز هذه الورقة على بناء نظام لتحليل النصوص في اللعربية المعاصرة بطريقة آلية. وسوف ينم معالجة مشكلة التشكيل الألي من خلال مرحلتين أساسيتن هما مرحلة المعالجة الصرفية للكلمات ومرحلة المعالجة النحوية للكلمات والتي تختص بوضع العلامات الإعرابية للكلمات التي تحتاج إلى تلك العلامات. وقد تحقيق ذلك من خلال عينة لغوية محللة تحليلا صرفيا وبجانب كل كلمة العديد من المعلومات اللغوية بجانب العلامة الإعرابية. وقد تم استخدام هذه العينة في استخراج القواعد اللغوية التي تساهم في فك اللبس الدلالي والصرفي للكلمات من خلال السياق، كما تم استخدامها أيضا في بناء نماذج لغوية لفك اللبس الصرفي والدلالي بطريقة إحصائية بالإضافة إلى استخدانها في عملية اختبار النظام المبني. وقد حقق هذا النظام نسبة صحة وصلت إلى 89.1%. ومن المتوقع أن تزيد هذه النسبة باستخراج المزيد من القواعد اللغوية وزيادة حجم العينة اللغوية المحللة.**

Text mining model using a hybrid of SOM and LSI Techniques

Abdelfattah ELsharkawi [*1], Ali Rashed [**2], Hosam Eldin Fawzan [*3]
[12]Department of Systems and Computer Engineering, Al-Azhar University, Egypt.
sharkawi_eg@yahoo.com
a_m_rashed@hotmail.com
[3]Department of Electrical and Computer Engineering, Faculty of Engineering Science, Sinai University, Egypt.
Hos.9876@yahoo.com

Abstract: Self-Organizing Maps (SOM) are good tools for clustering unseen data patterns, and hence allowing easy information retrieval based on the discovered clusters. This paper proposes a new technique for enhancing the learning capabilities of SOM as an aid for data mining. The idea is to combine the Latent Semantic Indexing (LSI) with SOM to speed up and improve the clustering process. LSI is used to reduce the dimension of data before training the SOM. The combination of LSI and SOM has enhanced the accuracy of clustering and information retrieval as well as the training speed. A comparative study with similar research work is also introduced.

1 INTRODUCTION

LSI is one of the dimension reduction methods used in text data mining [1]. It is designated to extract the meaning of words by using their co-occurrence with other words that appear in documents [2]. LSI uses continuous Vector Space Model (VSM) that maps words and documents into a low dimensional space [3]. With the use of proper matrix scale, LSI can effectively overcome the problems of synonymy and polysemy. It uses Term Document (TD) matrix to solve these two problems [2]. But TD matrix evaluates the rare terms with low weight which (in some cases) is considered as more informative than defining frequent terms. In practice, a weighting scheme that better captures the importance of a word in the document than VSM is TF-IDF (Term frequency-Inverse Document Frequency). TF-IDF is one of the feature factorization methods widely used in text mining that can reflect the importance of terms in documents, and hence it is used as the first process in text mining to extract the features of terms in a dataset. In this paper the TF-IDF matrix is used instead of TD matrix to increase the rarity of the term in the collection which means rare terms are up weighted to reflect their relative importance which is not available when using VSM.

On the other hand Kohonen's self- organizing map (SOM) represents one of the most machine learning techniques used in clustering and information retrieval. There are many challenges facing SOM parameters that govern the clustering process and hence achieve the expected results. Among these parameters are the initialization with random weights, the scheme of the neighborhood shrinking function, the map size, and the definition of the learning rate [4]. This paper is suggesting a solution that combines TF-IDF, LSI and SOM to present a new solution for a fast text search engine while overcoming the drawbacks of using each one of these techniques individually.

This paper used the Reuter-21578 "ApteMod" as a dataset for bench marking of information retrieval. The "ApteMod" is a collection of 10,788 documents partitioned into a training set of 7769 documents and a test set of 3019 documents. 250 documents of five categories were chosen for benchmarking in this paper (50 documents for each one; namely earn, acquisition, crude, trade, and interest).

A. *Latent Semantic Indexing (LSI)*

Latent Semantic Indexing (LSI) is a method for discovering hidden concepts in document data. In each document, the searched terms (words) are expressed as a vector with elements corresponding to these concepts. Each element in that vector gives a degree of participation of the document or term in the corresponding concept. The goal is not to describe the concepts verbally, but to be able to represent the documents and terms in a unified way for exposing document-to-term similarities or semantic relationships which are otherwise hidden[2]. LSI is a widely used continuous vector space model (VSM) that maps words and documents into a low dimensional space [5].

Singular Value Decomposition (SVD) is used to reduce the rank of a matrix without losing important content and to eliminate all noise (i.e. all data that obscure the content). It is combined with LSI to get the search results from corpus of documents. In LSI, the matrix TF-IDF is factored into the product of three matrices U, Σ and V using SVD function as in Fig. (1).

$$TF\text{-}IDF=U.\ \Sigma.VT \quad (1)$$

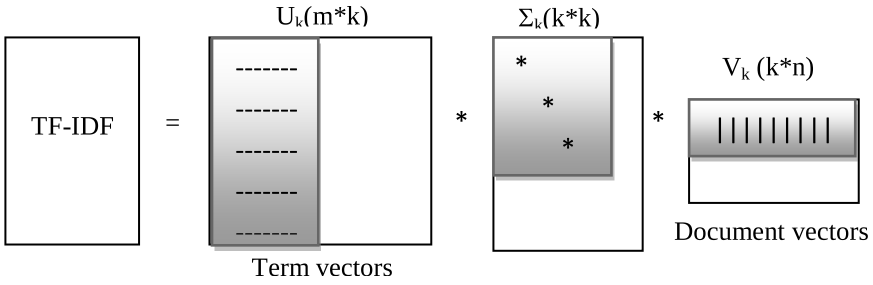


Figure 1: Analysis of TF-IDF to three matrix U, Σ and V

where U is an orthogonal (m×m) matrix whose columns are left singular vectors of TF-IDF, Σ is a diagonal matrix whose diagonal elements are singular values of matrix TF-IDF in descending order, V is an orthogonal (n×n) matrix whose columns are right singular vectors of TF-IDF [2]. The power of LSI comes from truncating the U, Σ and V matrices to K dimensions. Multiplying UkΣkVk produces the best rank-k approximation of the original term-document matrix [2]. So in Fig.(1), Uk is an (m×k) matrix whose columns are first k left singular vectors of TF-IDF, Σk is (k×k) diagonal matrix whose diagonal is formed by k leading singular values of TF-IDF and Vk is an (n×k) matrix whose columns are first k right singular vectors of TF-IDF. In LSI, the query vector has to be transformed into the same space as the document vectors before computing the cosine similarity. Each document vector is taken from a column of V', and the equation for transforming the query vector is [2]:

$$q=qT\ U_K\Sigma_{K-1} \tag{2}$$

This dimension reduction to k dimensions provided by SVD is the closest rank-k approximation available that allows eliminating noise and capturing the underlying latent structure [6]. Each document vector then has its cosine similarity taken with the query vector, and that result is recorded as the final relevance score for the document/query pair.

$$Sim(q,di)=di*q/|di||q| \tag{3}$$

where i ϵ [1, n] and then sort the results in descending order. By using these equations [1, 2 and 3] the user can get the relative document that he is searching for. The documents become relative if sim (q,di)>0.

2 TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY ALGORITHM

TF-IDF algorithm calculates an index for measuring the importance of a term to a document in a corpus. It is used for calculating the frequency of terms of a given word in a given collection of documents and calls it Term Frequency (TF) as shown in equation (4). It also calculates the Inverse Document Frequency (IDF) as in equation (5). The term count or the number of times the term appears in document indicates the importance of that term in this document [7]. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [7]. The TF-IDF is the product of term count (TF) and Inverse Document Frequency (IDF). The term frequency TF of term term ti in document dj is calculated by equation (1).

$$TFij = \frac{Nij}{\sum Nij} \tag{4}$$

Where Nij , is the number of occurrences of the considered term ti in document dj.

$$IDFi = \frac{\log|Dtotal|}{|d: ti \in d|} \tag{5}$$

Where |Dtotal | is total number of documents in the corpus and d : ti ε d is number of documents where the term ti appears. The TF-IDF for each term t can be defined as in Eq. (6) [7].

$$(TF\text{-}IDF)\ weight = TFij * IDFi \tag{6}$$

3 SELF-ORGANIZING MAPS (SOM)

Kohonen's self-organizing maps (SOM) are abstract mathematical models used for clustering of data [8]. The SOM algorithm is a competitive algorithm founded on the vector quantification principle: at each cycle of life in the network, the unit from SOM whose codebook is most similar to the input wins and called the best matching unit (BMU). The SOM consists of a topological grid of neurons typically arranged in one or two dimension lattice [9]. The SOM Learning algorithm steps are:

- Select an input vector X (t) = (x1 (t), x2 (t), ..., xn (t)
- Find winning node by calculating the Euclidian distance

  $ds = \min\|X(t) - W(t)\|$, where Wk (t)= (wk (t), wk (t), ..., wk (t)).
- Adjust weights as follows:

$$W(t+1) = W(t) + \eta(t)*(X(t) - W(t)), \quad \text{where } 0< \eta(t) <1. \tag{7}$$

where the learning rate function is:

$$\eta\,(t, k, s) = A1 * \frac{1}{e^{\left(\frac{t}{A2}\right)}} * e^{\left(\frac{t*d(k,s)}{2\sigma 2}\right)} \tag{8}$$

Where d (k, s) is the Euclidian distance between the node k and the winning node s in the two-dimensional grid, while A1 and A2 will be defined latter in Eq.(9) and Eq.(10). In the formula, the first Gaussian function A1 controls the weight update speed and the second Gaussian function A2 defines the neighborhood shrinkage function in SOM. The standard deviation σ decreases monotonically with time [10].

η_{start}:$0 < \eta_{start} < 1$, is the starting value (value at time t = 0) for η for the winning node s. Note that the time t goes from 0 to (C-1).

η_{end} : $0 < \eta_{end} < \eta_{start}$, is the final value (value at time t = (C-1)) for η for the winning nodes [11].

From Eq. (8) it is clear that at time t=0, η(0,k,s) = A1. Hence:

$$A1 = \eta_{start} \tag{9}$$

$$A2=(C-1)/\ln(\eta_{start}/\eta_{end}) \tag{10}$$

Dmax is the maximum distance in the map, i.e., the Euclidian distance between two opposite corners in the map

$$D_{max}=\mathrm{sqrt}[(Mr-1)(Mc-1)] \tag{11}$$

where Mr is the number of rows in the map and Mc is the number of columns in the map.

There are many challenges facing SOM parameters that govern the clustering process and hence achieve the expected results. Among these parameters are the initialization with random weights, the schedule of decreasing of the neighborhood shrinkage function, map size, and decreasing the learning rate [4]. Kohonen proposed that initialization should be based on random vectors in input space which lead to faster convergence between neurons [12]. Where in the initialization plays a critical role in convergence speed [13], neurons are initialized and organized in topologies that are preset by the designer of the network.

4 THE NEW MODEL

The idea of the new model is to combine the two techniques of LSI and SOM to enhance the accuracy of information retrieval and as well as the term clustering. The Vk and Uk extracted from LSI are used to train the SOM. The suggested model comprises a collection of processes; namely choosing the training dataset, preprocessing, feature extraction of the terms, LSI and hence training the SOM to get the relevant documents inside the SOM map. Fig (2) shows the stages of implementing that model. SOM uses TF-IDF in a feature extraction stage which is known as the best weighting scheme in information retrieval for training terms (which represent columns) and documents (which represent rows). For each query, the SOM maps all the training set documents in TF-IDF matrix. This consumes more time and reduces the accuracy of the relevant documents (precision and recall) in final SOM map by getting more documents that is not relevant to the query term. To overcome these two problems, the LSI with TF-IDF are used before training the SOM to reduce the dimension of documents by getting the maximum of relevant documents to the query term and pass them to train the SOM. This way improves the accuracy of finding relevant documents by SOM map and reduces the time needed for training the SOM.

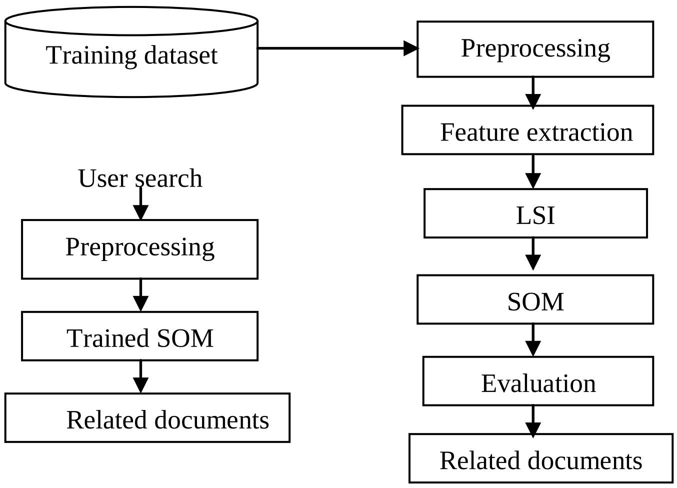


Figure 2: The main algorithm for clustering and searching of text documents

A. Preprocessing phase

Data mining techniques aims at having the data in a structured form and hence can easily obtain the knowledge. Aiming at having a web mining engine, the first step would be the removal of html tags as well as the removal of leading and tailing spaces from SGM (Standard Generalized Markup Language) files. The next step is the tokenization which acts for breaking up a sequence of strings into pieces such as keywords, functional phrases, symbols and other elements called tokens. The third step is cleaning the list of words from stop words (e.g. the, am, is, are etc.). The last step is de-stemming which means returning each word to its original form (root). This is done using porter de-stemming algorithm [14] that uses a set of 60 transformation rules which are applied in a succession of 6 steps. This process is used to make dimensional reduction of the total terms in the dataset. All these steps summarizes in Fig. (3).

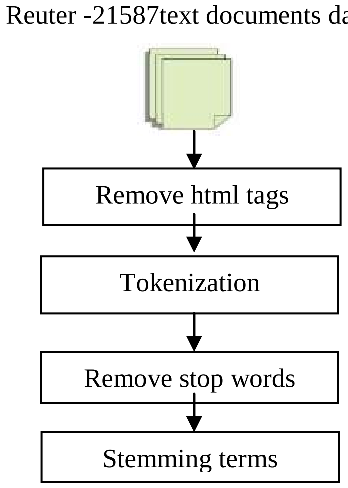


Figure 3: Steps of preprocessing

B. Feature extraction phase

Feature extraction process is concerned with transforming unstructured text data into numerical features usable for machine learning by SOM. Term frequency (i.e. the number of occurrences of one term in a document) and inverse term frequency (i.e. the measure of general importance of the term) composes these numerical features. This process then extracts necessary information required to describe a large set of data. The algorithm for Feature extraction algorithm will in the following steps:

1. Build the TF matrix for each term using Eq. (4).
2. Build the IDF matrix using Eq. (5).
3. Calculate the TF-IDF by multiplying TF and IDF.

C. LSI

LSI algorithm is running before training the dataset by SOM to determine the best value for K. This is done in three steps, the first starts by searching for a "company" term which is one of the 30 terms in table(1) and then measuring the accuracy of related documents returned for different K values (between 10 and 20) and then record the results in column F1 as in table(2). The second step is removing number tokens and then repeats the first step by searching for the same term and records the results of accuracy in column F2. The third step is removing date tokens and repeats the first step again and record results in column F3. The algorithm for LSI will declare in the following steps:

1. Enter a query term which is a “company”.
2. Perform SVD form TF-IDF
3. Choose best rank **k** approximation of the original term-document matrix(TF-IDF)
4. Execute the query as in Eq.(2)
5. Rank the documents by using cosine similarity as in Eq. (3).

D. SOM clustering algorithm adapted for document mining

1. Each node’s weights are initialized randomly
2. Select a random vector from a set of training documents(V_k) and presented in the lattice
3. Calculate the BMU.
4. Adjust the weights of the winning node and the weights of its neighboring node in the grid.
5. Adjust the learning rate L (t) as explained in Eq. (7).
6. Repeat step 2 N iterations.
7. Repeat step 1 to 6 but using training data (U_k) which represent terms.

N.B SOM is trained using Rows of U_k that represent terms and columns of V_k that represent documents

E. Searching phase

The Searching process starts by entering a search term. The preprocessing process then takes place as explained in section 5.1 to remove non- significant data. Matching the term using the SOM map then takes place. If the search term is found in the map, then related documents will be listed in browser by Matlab version 10, otherwise a message will appear telling the user that this term is not found. SOM role phase in the searching algorithm will be declared in these steps:

1. The user enters the searching query.
2. The preprocessing process removes non- significant terms in the query
3. If the term does exist in the SOM map, then all nodes that contain that term will be activated and colored
4. The relevant documents to the query term will be viewed in ascending order in another window.
5. The user selects a document that he/ she wish to open and read.
6. IF the term is not in the SOM map, then an error message appears.

F. Evaluation of information retrieval

The evaluation of information retrieval contexts is defined by three measures namely; Precision, Recall and F measure [15]. These measures are defined in terms of the retrieved documents as well as relevant elements. A set of retrieved documents is defined as containing the list of documents produced by a search engine for a query. A set of relevant documents is defined as containing the list of all documents on the dataset that are relevant for a certain topic.

The two measures, precision and recall are used together in calculating a single bench marking measure which is the F-measure [15].

$$\text{Precision} = |\text{Relevant} \cap \text{Retrieved}| \,/\, |\text{Retrieved}| \quad (12)$$

For a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

$$\text{Recall} = |\text{Relevant} \cap \text{Retrieved}| \,/\, |\text{Relevant}| \quad (13)$$

Recall is the number of correct results divided by the number of results that should have been returned. These two measures can be combined using F score as in the following equation:

$$\text{F-measure}=(2^* \text{Precision} *\text{Recall})/(\text{Precision} + \text{Recall}) \quad (14)$$

The F-measure can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and its worst score at 0.

5 IMPLEMENTATION AND DISCUSSION OF RESULTS

The numbers of tokens that has been extracted from the dataset after the preprocessing process executed are 4522 tokens. These numbers of tokens when trained by SOM will take too much time. So, selecting a few numbers of tokens to represent the dataset and reduce the time of training is required. This is done by selecting only the tokens that achieve a threshold 10% of DF (which means select tokens that occur in 10% of total documents or more). After applying this threshold there are 30 tokens remaining as in table (1) which lead to reduce the execution time of training by SOM.

Studying the use of TF-IDF for feature extraction and its effect of the SOM cluster size due to implementation, the TF and IDF for some 30 tokens extracted from the dataset are viewed in table (1). The last two columns refer to the size of clusters in a 10*10 SOM map for each token before and after the removal of non-significant tokens.

TABLE (1): TERMS FREQUENCY (TF) AND SIZE AREA IN KOHONEN MAP TRAINED BY SOM

No	Token	TF	DF	Size of clusters before removal of non-significant tokens (30 tokens)	Size of after removal of non-significant tokens (24 tokens)
1	Reuter	297	297	2	2
2	Mln	525	157	3	2
3	Dlrs	369	126	4	5
4	Cts	279	110	2	3
5	Vs	493	108	3	3
6	Net	193	104	1	2
7	Pct	252	85	0	6
8	Shr	147	84	6	7
9	Company	117	80	4	6
10	1986	113	69	0	0
11	Inc	80	53	0	5
12	Share	109	61	0	0
13	Lt	127	58	4	4
14	Corp	78	50	3	4
15	Note	51	50	5	5
16	Rev	92	53	5	5
17	Loss	166	44	5	4
18	Billion	186	57	4	0
19	Stock	93	45	4	4
20	April	66	50	5	0
21	Shares	86	42	2	7
22	March	53	34	4	0
23	Sale	51	32	5	5
24	1987	53	40	4	0
25	Record	46	41	4	5
26	Told	42	37	5	4
27	Nine	57	40	3	0
28	Five	42	34	2	0
29	Dlr	40	33	6	6
30	Mths	55	42	5	6

As an example and when training 30 tokens by SOM, table (1) shows that the term "Company" occupies 4 nodes out of 100 nodes in the SOM map while the accuracy percentage of reaching documents related to that term was 16%. The same term occupies 6 nodes and the accuracy percentage increases to 20% when removing non- significant tokens. This in turn reduces significant terms to 24 tokens. The logical conclusion here is that the increase in the number of SOM nodes sensitive to the searched term increases the accuracy of reaching the related documents as shown in Fig (4) and Fig (5).

A. LSI results:

This stage of implementation studied the best value approximation of K dimension to be used in calculation of the relative documents. Typical values for K between 10 and 20 were tested (To avoid having A_k very dense, the values of K less that 10 were neglected, and to avoid making it very sparse the values of K more than 20 were also neglected). The values of K were first calculated for one term chosen from the query terms. The corresponding F measure was then calculated. Table 2 shows three f-measure values (namely F1, F2, and F3) relative to different choices of token sets. (How do you know the best value of K?). The best value K is determined based on measuring the accuracies of related documents by F-measure in each K value when searching for "company" term. From table (2) it can be observed that, the highest accuracy values in column F2 and F3 is 72.3 and 79.3 when K is equal to 20 and 16 respectively. The best value of K is determined by choosing the mean value of K at heights two accuracy, So that the best value chosen for K is 18.

TABLE 2: K VALUES AND CORRESPONDING F-MEASURES FOR TERM "COMPANY"

K	F1	F2	F3
10	52.36	52.7	50.9
12	53.6	56.7	56.8
14	56.67	60	61.7
16	56.1	60.4	79.3
18	54.3	70	73.9
20	63.5	72.3	74.3

Table (2) shows the following results:

- F1 was calculated when using 30 tokens representing the original terms
- F2 was calculated when removing numeric tokens such as (1986 and 1987).
- F3 was calculated when removing dates tokens such as (March and April).

Numbers tokens or date tokens when removed affect the accuracy of related documents when searching for term "company". The accuracy has increased as appears in column F2 and F3 in table (2). Also theses tokens are considered as un-valuable because it can't be classified to any of the five categories that make up the dataset. Why date tokens

After the adoption of 18 for K value and removing the number and date tokens, run searching process for another 4 different entity names (dlrs, billion, stock and sales) and measure Precision, recall and F-measure for them. Table (3) shows results of these 4 terms when using K equal 18 as a best value.

TABLE 3: PRECISION, RECALL AND F MEASURE FOR 4 QUERY TERMS WHEN K=18.

Entity name	Precision	Recall	F-measure
Dlrs	77.9	84.2	80.9
Billion	48.6	96.5	64.7
Stock	55	97.7	70.4
Sales	39	100	45.24

The results in this table show that, the term that has high weight as token "Dlrs" (arranged early in Table 1) has higher accuracy than the term that has less weight as token "Sales"(arranged late in Table 1)

B. SOM clustering results

The training process by SOM runs twice to build two maps. The first time for training the 300 documents using the V_k matrix extracted from LSI process. The second time for training the 30 terms using U_k matrix in another SOM map. During each cycle of the first training, the program keeps the related documents that are connected to each node inside the map. After the first training finished, the second training of clustering terms starts and builds another map at end. These two maps are combined together to build one map hence each node at the final map contains terms and related documents. At this stage the accuracy percentage of related documents to each term is calculated and added to each node in the map.

Training of the 30 terms by SOM map as in Fig (4) shows that 29 of 30 terms has presented in the map which mean that 96.6% of total inputs are viewed in the map.

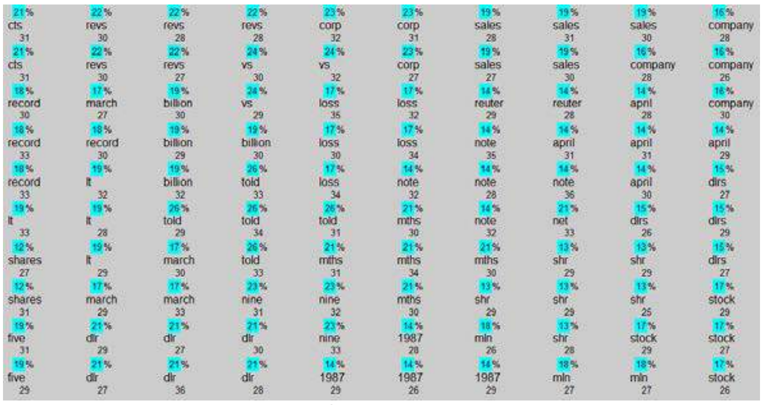


Figure 4: Clustering of 30 terms by 10*10 Kohonen map

When removing number and date terms, only 24 terms are found and all of them were presented in the SOM map. The process of removing these terms affected the precision, recall and F measure for different terms as shown in Fig(5).

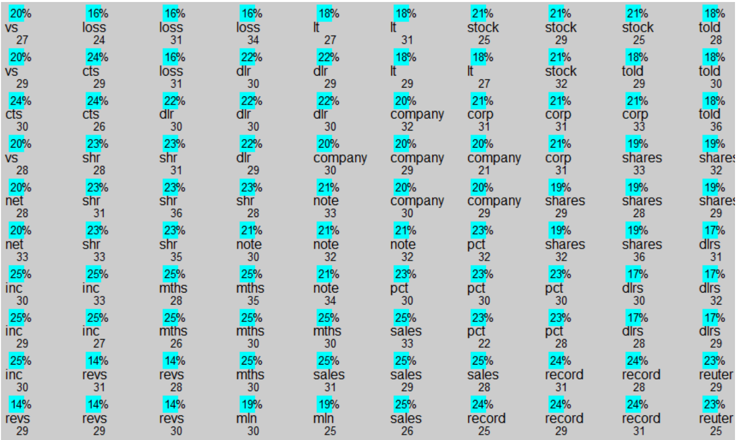
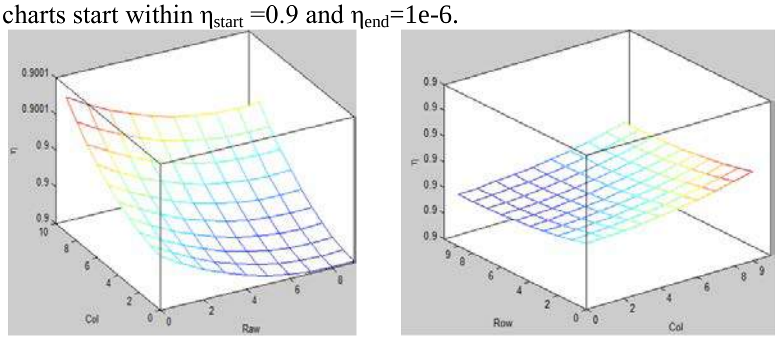


Figure 5: Clustering of 23 terms by 10*10 Kohonen map

After many experiments, the size 10*10 looks to be appropriate for arbitrary queries. Of course, increasing the size of the map will result in longer processing times, since many more weight vectors will need to be considered.

C. Speed of SOM clustering

The σ affect on the second Gaussian function in eq. (8) which affects the speed of neighborhood shrinking within period length C equal 2550 cycle. Fig. (6) Shows the speed of neighborhood shrinking for three different nodes (9,0) and (0,9) and values for σ=1e-3 on the left and σ=1e-6 on the right. Form the two figures notice that the speed of the neighborhood shrinking grows in the left figure and not in the right. This means that the speed of the neighborhood shrinking does not grow when σ becomes smaller. All the charts start within η_{start} =0.9 and η_{end}=1e-6.



[9,0] [0,9]

Figure 6: Influence of σ on the speed of the neighborhood shrinking for three nodes

D. Implementation of a new search engine

At this stage the algorithm of SOM training ended and the algorithm of searching start when the user enters the entity name in search engine form and press search as in Fig (7). After writing the query the preprocessing process is done for this query to remove all the non-significant terms like (stop words or punctuation marks).

Figure 7: GUI searching in SOM map.

The relatives documents for each search term view in GUI are split into two parts as in Fig. (8). A left part which consists of all relatives documents and a right part which views contents of the document when clicking on the document name in the left part. Five entity names representing five quires were used to evaluate SOM.

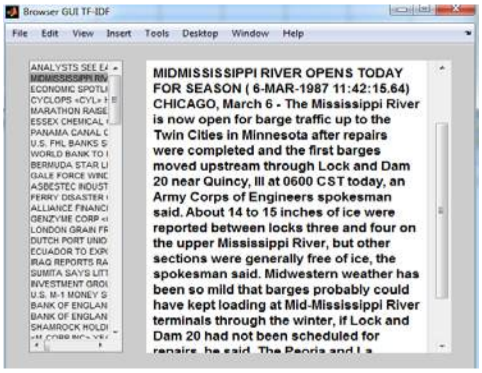


Figure 8: GUI to view the index of text document after searching.

E) Comparative study

The comparative study between the work in this study and Mohamed and Ahmed [11] summed up into five elements: map size, training process, view the accuracy percentage, Number "0" that appears beside some terms in SOM map and finally the number of clustering terms as shown in table (4).

TABLE 4. DIFFERENCES BETWEEN THIS PAPER AND MOHAMED AND AHMED [11].

	Factor	Mohamed and Ahmed. [11]	This study
1	Map size	10*14	10*10
2	Training process by SOM	Each user query require execute training process	Training process execute once for all user queries
3	Views the accuracy percentage of related documents in each node inside SOM map	SOM map doesn't views the accuracy	SOM map views the accuracy
4	Is some terms inside the SOM map contains zero related documents?	Yes, there are some terms has zero related documents	No, there is no term has zero related documents
5	Numbers of clusters in SOM map	(9-30)=30%	(26-30) =78% (22-24) =95.6%

Table (4) declares five differences between work in Mohamed and Ahmed [11] and the present work. From table (4) it is clear that the map size reduced to be 10*10 that lead to reducing the time of training. The training process (documents and terms) runs just one time for all queries instead of many times for many queries. The accuracy percentage of related documents that are connected to each term in each node inside SOM map views which consider as addition in this paper. The term inside the SOM map that contains zero related documents is replaced by a number more than zero. The final result is increasing the number of clustering 30 terms in SOM map from 9 clusters to be 26 clusters which mean increase the cluster percentage from 30% to be 78%. Also when removing the number and date tokens the cluster percentage has increased to be 95.6%.

6 CONCLUSIONS

The self-organizing map is a good technique for clustering and visual display of information retrieval of data mining such as text files. Using TF-IDF weight technique with LSI have increased the probability of representing rare terms in SOM map instead of using TF with LSI that evaluates the rare terms with low weight. LSI speeds the process of machine learning by reducing the dimensions of terms and documents vectors for training by SOM. Using the SOM in information retrieval system still needs more enhancements to achieve better results. The implementation of the new model showed differences of the SOM clustering phenomena compared to Mohamed and Ahmed [11]. Clustering of terms when using SOM has increased by 50% for the same number of terms. Also this paper has enhanced the SOM map by making each term in the map has related documents more than 0. The training process by SOM runs only one time for all user queries instead of it was run one time for each query. This paper has added an addition to the SOM map by views the accuracy percentage of related documents to each term. This addition is helping the user by knowing in advance the accuracy percentage of each term in the map before searching for his specific query and view the related documents to this query.

REFERENCES

[1] Tuomo Kakkonen, Niko Myller, Erkki Sutinen and Jari Timonen, "Comparison of Dimension Reduction Methods for Automated Essay Grading", ISSN 1436-4522 , Grading. Educational Technology & Society, 11(3), 275–288 (2008).

[2] Andy Garron , April Kontostathis, "Latent Semantic Indexing with Selective Query Expansion", Department of Mathematics and Computer Science, Ursinus College, Collegeville PA 19426 , 2011.

[3] Kai-Wei Chang , Wen-tau Yih , Christopher Meek ,"Multi-Relational Latent Semantic Analysis", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1602–1612, Seattle, Washington, USA, 18-21 October 2013.

[4] Baltic J. "Investigation on Learning Parameters of Self-Organizing Maps", Modern Computing, No.2, 45-55, Vol. 2 (2014).

[5] Tarek F. Gharib, Mohammed M. Fouad, Abdulfattah Mashat, Ibrahim Bidawi," Self Organizing Map -based Document Clustering Using WordNet Ontologies", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012.

[6] Tom Magerman, Bart Van Looy, Bart Baesens, Koenraad Debackere, "*Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents*", faculty of business and economics October 2011.

[7] D. Chandrakala, S. Sumathi, R. Bharath Raj,V. Kabilan, "Enhanced Emergent Trend Detection System Using PSO Based High Dimension Growing Self Organizing Map", European Journal of Scientific Research, India, 2011.

[8] Abdelfattah ELsharkawi , Ali Rashed , Hosam Eldin Fawzan, "Comparative Study of clustering phenomena of 2-D SOM Against 3-D SOM", journal of Al Azhar Universaity engineering sector, ISSN:1110-6409, Egypt, 2014.

[9] Dumidu Wijayasekara, "Visual, Linguistic Data Mining Using Self Organizing, Comput". Sci. Dept., Univ. of Idaho, Idaho Falls, ID, USA, June 2012.

[10] Z. Mohd Zin, M. Khalid, E. Mesbahi and R. Yusof , "Data clustering and topology Preservation using 3D visualization of self organization maps", Proceedings of the World Congress on Engineering, 2, 2012.

[11] Mohamed Salah Hamdi, Ahmed Bin Mohammed, "SOMSE: A semantic map based meta-search engine for the purpose of text information customization", Applied Soft Computing, 11:3870-3876, 2011.

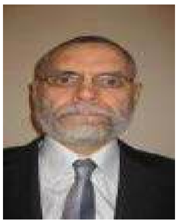
[12] Kohonen T ,"Self-organizing maps", 2nd edn. Springer, New York, (1995).

[13] Iren Valova, Derek Beaton, Alexandre Buer," Fractal initialization for high quality mapping with self-organizing maps", Springer-Verlag London Limitd 2010.

[14] Porter, M.F. An algorithm for suffix stripping. Program: electronic library and information systems, Vol. 40 Iss: 3, pp.211–218, (2006).

[15] Keneilwe Zuva , Tranos Zuva," evaluation of information retrieval systems", International Journal of Computer Science & Information Technology (IJCSIT) Vol. 4, No. 3, June 2012.

Bibliography

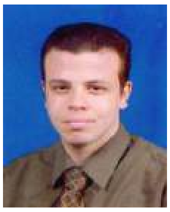


.[1]Dr. Abdelfattah El-Sharkawi

Associate professor , Software Engineering ,Al –Azhar University, Cairo, Egypt. Associate Professor, Software Engineering at Systems and Computer Engineering,Faculty of Engineering –Al-Azhar University, Cairo, Egypt Ph.D. degree inSystems and Computer Engineering (1993) Faculty of Engineering –Al-Azhar University. M. Sc. degree in inSystems and Computer Engineering (1986) Faculty of Engineering –Al-Azhar University. B.Sc. in inSystems and Computer Engineering (1981) Faculty of Engineering –Al-Azhar University

[2]Prof. Dr Ali Mahmoud Rashed

Ph.D. degree in electronics and communications engineering(1982) Faculty of Engineering - Ain Shams university. M. Sc. degree in electrical engineering (1977) -Faculty of Engineering - Al Azhar University. B.Sc. of Electronics and Communications Engineering (1968) MilitaryTechnicalCollege. Grade: Very Good. Supervise more than 50 Ph.D. and M.Sc. thesis and 100 B.Sc. projects. A member in the researches reviewers committee for Al Azhar Engineering research Journal. A member in the researches reviewers committee for Teacher Higher Staff Position – Al Azhar University. Holding a technical consultant position in ETCP (Egyptian Technical Colleges Project), Including Courses Revision, Quality Assurances and Human Resource managements since 2006 -2015.



[3]**Hosam Eldin Fawzan** received the B.Sc. from the faculty of Engineering, Al Azhar University, Egypt, in 2003. M. Sc. degree in system and Computer Engineering - Faculty of Engineering - Al Azhar University Egypt, in 2010. I'm joined the Electrical and Computer Engineering Department, Sinai University, Egypt in 2008.

نموذج لتعدين البيانات باستخدام مزيج من تقنيات خرائط التنظيم الذاتى

(SOM) و الفهرسة الدلالات الكامنة (LSI)

عبد الفتاح الشرقاوى[1] ، على راشد[2] ، حسام الدين فوزان[3]

1،2 قسم هندسة النظم والحاسبات،جامعة الأزهر - القاهرة- جمهورية مصر العربية

[1]sharkawi_eg@yahoo.com

a_m_rashed@hotmail.com[2]

3 قسم الهندسة الكهربائية والكمبيوتر،كليه العلوم الهندسيه ،جامعة سيناء-شمال سيناء – جمهورية مصر العربية

Hos.9876@yahoo.com[3]

ملخص

تعتبر خرائط التنظيم الذاتي (SOM) أدوات جيدة لتجميع أنماط البيانات غير المرئيه، ومن ثم السماح بسهولة استرجاع المعلومات استناداً إلى المجموعات التي تم اكتشافها. وتقترح هذه الورقة تقنية جديدة لتعزيز قدرات التعلم SOM كمساعدة لاستخراج البيانات. والفكرة هى الجمع بين فهرسة الدلالات الكامنة (LSI) مع ال SOM لتسريع وتحسين عملية التجميع. يستخدم LSI لتقليل البعد للبيانات قبل تدريب ال SOM. إن دمج ال SOM و LSI أدى لتحسين دقة تجميع واسترجاع المعلومات فضلا عن سرعة التدريب. كما قدم دراسة مقارنة مع عمل بحثي مماثل.

CMET: A Semantic Framework for Comparing and Merging Entities and Terms and its Application in Answer Selection

Mahmoud A. Wahdan[*1], Safia Abbas[*2], Mostafa Aref[*3]

[*]*Computer Science Department, Faculty ofComputers and Information Sciences, Ain Shams University*
Cairo, Egypt

[1]mahmoud.a.wahdan@gmail.com
[2]safia_abbas@yahoo.com
[3]mostafa.m.aref@gmail.com

***Abstract*—Named Entities are very important for many text-based applications. We present a general framework for detecting the semantics behind entities, Comparing and Merging Entities and Terms (CMET). The entities and terms should be of the same semantic type in the entity type hierarchy. The proposed framework is well-designed and flexible for future enhancements and can be extended to other languages than English. Many applications such as Question Answering Systems, Text Summarization and Co-Reference Resolution make use of entities similarity. We exploited knowing the semantic relations between entities in Question Answering system not only for boosting redundant answer candidate score, but also to support the scores of answer candidates based on its semantic similarity. We did an experiment to measure the impact of using our framework only in Question Answering system. We reported 6.1% increase in Answer Selection and 4% increase over the baseline in the end-to-end Question Answering system.**

1 INTRODUCTION

The usage of Named Entities and the relations between these entities are very important for many natural language applications. Information Retrieval (IR) Systems especially Question Answering (QA) Systems, Sentiment Analysis are examples of systems that depend heavily on Named Entities and relations between them [17].

String matching, lexical matching and string manipulation will fail to detect equality between two entities in many cases bedside failing to detect other relations between those entities [8]. The cause of its failing is that it depends heavily in comparing the words lexically and have no information about its meaning. For example: it has no information that "*Karl Malone*" is also known as "*The Mailman*". Unlike these approaches, knowing semantic relations between entities can be successfully used to enhance the accuracy of many natural language systems and applications. For example: the accuracy of Answer Selection phase in QA system beside the end-to-end QA system in [1, 2, 8, 9].

Question Answering system aims to automatically answer a natural language question by providing a precise answer. A common Architecture of QA system is shown in Figure 1 [17]. Question processing is the module which identifies the focus of the question, Named Entities, classifies the question, derives the expected answer type, and reformulates the question into semantically equivalent multiple questions. Reformulation of a question into similar meaning questions is also known as query expansion and it boosts up the recall of the information retrieval system [15]. IR system (Search) recall is very important for question answering, because if no correct answers are present in a document, no further processing could be carried out to find an answer [17]. Precision and ranking of candidate passages can also affect question answering performance in the IR phase. Answer extraction and selection is the final component in question answering system, which is a distinguishing feature between question answering systems and the usual sense of text retrieval systems.

Named Entity Recognition (NER) is a vital natural language component for many systems and applications including QA and so important in extracting candidate answers, scoring and selecting the right answer. Many approaches based on exploiting redundancy in candidate answers by doing some linguistic and lexical processing. This approach was early used by [4, 10, 11]. But these strategies usually consider candidate answers as independent entities. For example: for the question "*Where are the three pyramids located?*" the candidate answers *Giza* and *Egypt* are related since *Giza* is located in *Egypt*. In this example traditional manners will not report a relation between these two answer candidates and so the right answer - which is the answer with the highest score – will not be retrieved.

Many QA systems including OpenEphyra based on Ephyra [13] reward redundancy of the same answer candidate by removing one of them and boost the confidence score of the other one. Although the system counts redundant candidate answer, it may end up that none of the correct answer(s) will get the highest score and then the retrieved answer(s) will be wrong.

After doing error analysis, we found that the exact candidate answers might be considered different entities by previous QA systems because they are not lexically identical; however, it might be semantically the same. Another probable condition is there are some candidate answers that may be semantically related (i.e. Inclusion or Subsume relationship). Therefore, they must be used as an evidence to support each other rather than considering them as poles apart. So, for more reliable results, the proposed model will detect the similar entities and the degree of similarity between them not only based on lexical identity but also based on semantic structure in order to strengthen answers' candidate.

In order to decide that two entities are similar, they must be of the same entity type and exploiting the same semantic relation (i.e. similarity) based on the entity written nature. For example: when two dates are compared, some rules are considered that are different than those rules used for numbers or persons comparing process.

Moreover, other systems / tasks may use semantic relations such as: 1) Term Clustering Algorithms that can use a new similarity measure based on our framework; 2) Co-Reference Resolution task with enhanced nominal chaining; 3) Text Summarization tasks; 4) Enhancement of measuring the semantic similarity between two sentences and paraphrase detection.

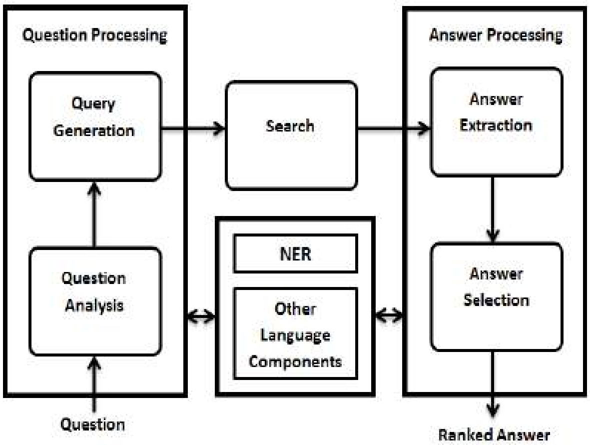


Figure 1: Common Architecture of QA systems

2 RELATED WORK

There are many trials to use redundancy to boost candidate answers' score in Answer Selection phase in QA systems [8, 9]. Some of these trials use just lexical comparisons, while a few works considered semantic relations between entities [9]. Unfortunately, lexical strategies will not detect all equal entities and also will not detect other semantic relations.

Normalization is the main step in some approaches. For instance, [1, 2] mainly compare and examine sets of answers to numerical (DATE and NUMBER) questions. This is done based on three dimensions (time, place or other restriction) between correct answers and they may decide to merge both some of these answers. Another approach [3] uses candidate answer normalization. Another approach [4] detects the relations between candidate answers. They made use of those relations in answer selection as final answer (the one that includes most of the others by tokens comparison). This is naïve matching strategy that will cause both false positives and false negatives.

Statistical approaches are also used to detect similarity. None of these statistical approaches such as [6] can explicitly detect the kind of similarity that exists between terms. Using taxonomies such as WordNet [7] may be good in measuring semantic similarity between words and try to get semantic sentence similarity. When we want to compare two terms or entities based in WordNet, it will not be the correct choice. The cause is entities and terms comparison based essentially on knowing the entity type of which these entities and terms is instantiated from. [16] is another probabilistic approach. It combines multiple evidences to rank answers and exploits the similarity between entities using traditional String-based similarity.

A framework in [8] made use of their claim in [5] to solve similarity issues. They designed a framework for entities comparisons and return semantic relationship. But they do not identify synonyms; which is an important strong point in our system. The approach has some weaknesses because they heavily depend on lexical relations and the nature of how people write entities of specific types. Another work [9] shortens the relations in four main relations originally introduced by [12]. This will be a limitation because some entities relations will not fit in these four relations based on the type of the compared entities. They focused on answers from specific categories NUMBER, DATE and ENTITY.

3 CMET FRAMEWORK

The basic usage of relations between candidate answers in Answer Selection is to use the redundancy of each candidate answer to support its score in ranking. Our hypothesis is that not only redundancy will support candidate answers score, but also using semantic relations between entities (i.e. candidate answer in this case) will support its score. The Semantic framework for Comparing and Merging Entities and Terms (CMET) also can be used in different applications. We chose to show the application of CMET framework in Answer Selection phase in QA system. Figure 2 shows CMET framework and its interaction with Answer Selection phase in OpenEphyra QA system.

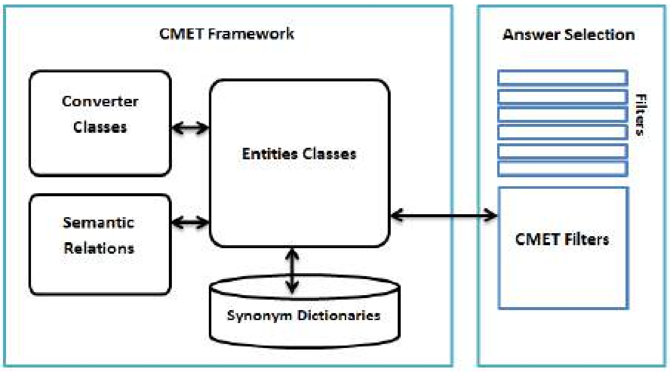


Figure 2: CMET framework and its interaction with Answer Selection phase

OpenEphyra's Answer Processing component consists of two main phases, Answer Extraction and Answer Selection. Answer Selection phase consists of many filters used to rank candidate answers – which are the results of Answer Extraction phase – and boost the score of every candidate answers based on some ranking criteria. The main approach is summing-up the scores resulted from running the filters on the previous score of candidate answer. There is no existence of any machine learning in Answer Selection phase, except in normalizing candidate answers scores which come from different sources or different Answer Extraction strategies. We will follow their strategy in summing-up scores. So, what we should do is to implement some filters that make use of discovering semantic relations between candidate answers (which is mainly named entities) to manage merging, duplicate removal and supporting candidate answers based on CMET framework. The new filters will mainly depend on discovering the semantic relations between two entities of the same type which is the functionality of CMET framework.

A. Synonym Dictionaries

Synonym identification is heavily used in CMET. We will define a strategy to generate a wide-range and precise enough Synonym Dictionaries. Synonym Dictionary will be used in synonym identification in semantically comparing the entities. We will show the strength of using such generated dictionaries in the following sections.

Free Encyclopedias like Wikipedia [18] and Freebase [19] are free sources of knowledge that many approaches and algorithms used to support, introduce evidence and reasoning for some semantic knowledge and facts. We exploited the coverage and variety of Wikipedia and Freebase to generate a synonym dictionary using two different approaches.

In Wikipedia, we used titles of "*redirect*" pages written in Wikipedia Dump file to generate the "*redirect list*" which consists of "*redirect-to*" term and "*redirected*" term and considered both terms to be synonyms. To fit the synonym dictionary in our task, we used the named entity lists provided in our Extended NER mentioned briefly in [17] to be a key in searching the resulted redirect list and make a synonym dictionary for each named entity type.

Freebase Dump is more structured and categorized by topic. We get lists of named entities that match our entity type hierarchy from Freebase as a search key in "*topic*" file which contains "*also known as*" relation (*aliases*).

The reason that using these synonym dictionaries will not hurt the system accuracy is that the Answer Extraction is based mainly on detecting named entities from the same type of expected answer type. This is done by our Extended NER [17] which uses precise named entity lists and the other extraction module uses patterns to extract answer candidates. So, the probability to get some garbage candidate answers and find it the same garbage in the synonym dictionary is very low and converges to zero.

B. Converter Classes

Converter classes are very important in comparing two named entities or terms. For example: the question "*How tall was Judy Garland?*" have many answer candidates of different length units, we will consider two of them; "*1.51 m*" and "*4 ft 11.5 in*" are exactly equal but how we can detect this equality while both are of different length units. Converter classes support wide-range of entities and wide-range of units per entity. Its main functionality is to convert a value from a specific unit to another unit that belongs to the same entity type of the first unit. Converter classes are very flexible, well-designed, reusable and language independent since it will not deal with any language interfaces. Frequently updated converter values (e.g. Currency Converter) can be updated also in CMET framework. For example: a web service can be used to get conversion rates daily.

C. Semantic Relations

Semantic Relations is an enumeration of different types of semantic relations that can exist between two entities of the same entity type. The relation can be extended and it is not mandatory for each entity to consider all semantic relations. For example: in RIVER entity EQUAL relation can be considered but it cannot consider SUBSUMES relation while we can consider both relations in DATE entity. Using relations enumeration makes the framework to be application independent. You will define relation score depending in your application either manually or by learning scores that leads to the maximum accuracy. Also the score of the same relation in different entity types can be different. The semantic relations in CMET framework are EQUAL for equality, EQUIVALENT for almost equal conversion rates, AROUND and CLOSE for approximate rates, OVERLAP if the two entities strings are just overlapping, SUBSUMES for hypernym, SUBSUMED for hyponym, LOCATED_IN if some place is located in another and NONE if no relation exists, and many others.

D. Entities Classes

Entities Classes is the most important part of CMET framework because it provides the main functionality and logic for how it can be used in applications. It supports 1) Entity Structure which shows how the entity can be structured from smaller peace of data. This involves parsing the entity string into entity structure fields; 2) Normalization routine that will be used to get the normalized version of the entity; for example: "*The Nile river*" of type RIVER will be normalized to "*Nile*" by removing both "*the*" and "*river*" words; 3) Compare routine that will take an entity of the same type as an input and return the semantic relation between the two entities as an output and this method contains the logic of how we consider the relations between entities; 4) Get Common Representation routine that will return the most common representation of the entity instance with its existing structure values. The common localized representation of entity can be retrieved. CMET is well designed and every class is an element in entities classes' hierarchy.

CMET parses and compares entities based on its natural type and exploiting natural strategies for writing some entities like Educational Institutions, Rivers, Mountains, Lakes, … etc. Previous works didn't use any type of synonym discovery routines except [9] used WordNet. WordNet is very good for detecting synonym for word senses but it will not fit in our solution because it provides very little knowledge about entity synonyms. Beside the source of our Synonym Dictionaries is internet users themselves while free encyclopedias allow contributions from internet users. So the resulted knowledge will be focused in their interests (i.e. how they name things). CMET has the knowledge of the type of the entity and also the hierarchical relations between entity types, so it can detect the hierarchical relations like what is done for word senses in WordNet. Unlike any previous works, CMET detects the majority of semantically redundant candidate answers beside other semantic relations. Consider the example: "*On what river is Strasbourg built?*" the candidate answers "*River Rhine*" and "*Rhein*" is EQUAL. This equality relation is detected by using Synonym Dictionary. This boosts the score of the first candidate answer and return it as the right answer which is true and can't be detected by normalization routine only which will remove "*the*" and "*river*" and do strict equality test as in [8]. Finding and discovering the natural ways of writing entity instances is useful in some cases but will not be extremely useful in all cases. As we should complete the solution by providing synonym usage; RIVER entity is an example to exploit this nature of writing rivers in sentences but it is also in need of using synonyms. Another example is the way EDUCATIONAL_INSTITUTION in USA is written; sometimes by writing U then state abbreviation, you can point to the state university; "*University of Virginia*" and "*UVA*" are equal, but if we only consider this natural feature we cannot detect the equality between "*MIT*" and "*Massachusetts Institute of Technology*".

Synonym Dictionaries beside entity type hierarchy are very important for CMET framework and the source of its strength that closes the gap of previous works in this field. Beside the approach used to get these Synonym Dictionaries is very easy, flexible, renewable and cover wide-range of knowledge. We detect semantic relations between entities not only EQUAL relation but also other semantic relations. We used Synonym Dictionaries in many other relations such as

LOCATED_IN relation. Synonym Dictionary also solved the problem reported by [8] in answering the questions that ask for "CauseOfDeath".

E. CMET Filters

CMET Filters are not part of CMET framework, but a direct implementation of Filter class in OpenEphyra. These new filters consult CMET framework to know the semantic relations between two entities to take a decision of merging; removing answer or mutual supporting for candidate answers score. The support strategy is simple as it mainly depends on the semantic relation returned from CMET. We manually added a weight for each semantic relation that will be used in "Support Routine" to define the degree of the boost in answer candidate score.

4 EVALUATION

We made experiments to evaluate CMET framework application in Answer Selection phase in QA system.

A. Experimental Setup

We will use a freely available QA test set from Text Retrieval Conference (TREC) QA track. We will use TREC11 QA track test set which consists of 500 FACTOID questions. It contains the questions as well as the answer patterns proposed by the competing systems participated in TREC11 competition. There are many wrong answers in answer pattern file. The wrong answers in the pattern files are due to many reasons. One of these reasons is that the answers in pattern files were right at the time of TREC11 competition but it is now wrong. For example: the question "*Who is the governor of Tennessee?*". The answer pattern file has "*Sundquist*" as an answer for this question while the right answer now is "*Bill Haslam*". There are a wide range of questions that change over the time and other questions that had no answer in pattern files. So, we should get the right answers to these questions. We follow [14] strategy with his collaboration with IBM Watson Research Center to extend the answer pattern files and this takes a lot of time and iterations.

B. Evaluation Metrics

We will follow TREC evaluation metrics at multiple levels. The first metric is the Accuracy (Acc) and calculated by equation (1). We will consider two types of Accuracy evaluation: 1) Accuracy of first result which will count "*correct answer*" if and only if the first answer is correct; 2) Accuracy of the first five results which will count "*correct answer*", if one of the first five answers is correct.

$$Acc = \frac{\sum correct\ answer}{n} \qquad (1)$$

The second metric is the Mean Reciprocal Rank (MRR) is the multiplicative inverse of the rank of the first correct answer and is computed as in equation (2). "$rank_i$" is the position of the first correct answer returned by the QA system for the question i.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \qquad (2)$$

We will do experiment to evaluate CMET framework impact on both Answer Selection and end-to-end QA system.

C. Experimental Results

The baseline system uses lexical similarity between candidate answers to merge and boost candidate answers score. After running OpenEphyra QA system with the baseline system, the results show accuracy of 46.4% for first answer, 51.6% for the first five answers and 58.6% for MRR in end-to-end QA system.

After running OpenEphyra QA system with CMET framework, the results show accuracy of 50.4% for first answer, 53.8% for the first five answers and 59.2% for MRR in end-to-end QA system. Experiments is done using web search. Table 1 shows the results of experiments.

TABLE 1
RESULTS OF EXPERIMENTS IN END-TO-END QA SYSTEM

	Acc_1	**Acc_5**	**MRR_5**
Baseline	46.4%	51.6%	58.6%
CMET	50.4%	53.8%	59.2%

We did error analysis and found that 35% of errors are from the early phases before Answer Selection phase. We did another experiment on Answer Selection phase only and get the results showed in Table 2.

TABLE 2
RESULTS OF EXPERIMENTS IN ANSWER SELECTION PHASE ONLY

	Acc_1	**Acc_5**	**MRR_5**
Baseline	71.4%	79.4%	90.2%
CMET	77.5%	82.8%	91.1%

5 CONCLUSION AND FUTURE WORK

We introduced a well-designed, flexible, renewable and extensible semantic framework. CMET framework compares and merges entities and terms based on the existing semantic relations between them. CMET can be used in variety of knowledge-based and text-based applications. We applied CMET on Answer Selection phase in OpenEphyra QA system. The results show 6.1% increase of first answer accuracy for Answer Selection phase and 4% increase of first answer for end-to-end QA system. We proved the advantage of using synonyms in semantic-based applications over the previous work strategies and proposed using CMET in other applications.

Since CMET mainly depend on named entities and its types, then there is a strong relation between CMET and NER. If we consider building a collaboration of NER and CMET and share the type hierarchy, it will be better especially in Answer Processing Component in QA system. Also, instead of manually adding weights to semantic relation when using CMET in Answer Selection phase, we can learn these weights and this will lead to best accuracy. Another idea is to use the semantic relation resulted from CMET as a feature in the feature vector for a statistical classifier for ranking and classifying candidate answers correctness instead of using ordinary support routine.

REFERENCES

[1]V. Moriceau, "Answer generation with temporal data integration," in *Proc. of 10th European Workshop Nat. Lang. Generation (ENLG-05)*, pages 197–202, 2005.
[2]V. Moriceau, "Numerical data integration for cooperative question-answering," in *Proc. of Workshop KRAQ'06 on Knowledge and Reasoning for Language Processing, KRAQ '06*, pages 42–49. ACL, 2006.
[3]J. Chu-Carroll, K. Czuba, J. Prager, and A. Ittycheriah, "In question answering, two heads are better than one," in *Proc. of NAACL'03*, pages 24–31. ACL, 2003.
[4] S. Buchholz and W. Daelemans, "Complex answers: a case study using a www question answering system," *Nat. Lang. Eng.*, 7:301–323, 2001.
[5] Prager, J.M., Duboue, P. and Chu-Carroll, J., "Improving QA Accuracy by Question Inversion, " *in Proc. of COLING-ACL 2006*, Sydney, Australia, 2006.
[6] Resnik, P, "Using information content to evaluate semantic similarity in a taxonomy," *in Proc. of the 14th IJCAI*, Montreal, August, 1995.
[7] Miller, G, "WordNet: A Lexical Database for English," *CACM 38(11) pp. 39-41*, 1995.
[8] John M. Prager, Sarah Luger, Jennifer Chu-Carroll, "Type Nano theories: A Framework for Term Comparison," *in Proc. of the sixteenth ACM conference on Conference on information and knowledge management,* Pages 701-710, 2007.
[9] Ana Cristina Mendes , Luısa Coheur, "An Approach to Answer Selection in Question-Answering Based on Semantic Relations," *in Proc. of the Twenty-Second International Joint Conference on Artificial Intelligence,* 2011.
[10] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng, "Data-intensive question answering," *in Proc. of 10th Text REtrieval Conference (TREC)*, pages 393–400, 2011.
[11] C. Clarke, G. Cormack, and T. Lynam, "Exploiting redundancy in question answering," *in Proc. of SIGIR*, pages 358–365. ACM Press, 2001.
[12] B.Webber, C. Gardent, and J. Bos., "Position statement: Inference in question answering," *in Proc. of LRECWorkshop on Question Answering: Strategy and Resources*, pages 19–25, 2003.
[13] Nico Schlaefer, Jeongwoo Ko, Justin Betteridge, Manas A. Pathak, Eric Nyberg, Guido Sautter, "Semantic Extensions of the Ephyra QA System for TREC 2007," *in Proc. of Text REtrieval Conference TREC 2007*, 2007.
[14] Nico Schlaefer, Jennifer Chu-Carroll, Eric Nyberg, James Fan, Wlodek Zadrozny, David A. Ferrucci, "Statistical source expansion for question answering," *CIKM 2011:* 345-354, 2011.
[15] Nico Schlaefer, Petra Gieselmann, Thomas Schaaf, and AlexWaibel, "A pattern learning approach to question answering within the ephyra framework," in *Proc. of TSD'06 the 9th international conference on Text, Speech and Dialogue* , Pages 687-694, 2006.
[16] Jeongwoo Ko, Luo Si, Eric Nyberg, "Combining Evidence with a Probabilistic Framework for Answer Ranking and Answer Merging in Question Answering,", 2010.

[17] Mahmoud A. Wahdan, Safia Abbas, Mostafa Aref, "Designing a Hybrid Approach for Answer Extraction for Factoid Questions," *in Proc. of the Fourteenth Conference on Language Engineering*, Cairo, Egypt, 2014.
[18] Wikipedia Web Site: https://www.wikipedia.org/ .
[19] Freebase Web Site: http://www.freebase.com/ .

BIOGRAPHY

Mahmoud A. Wahdan is a Software Engineer at Zalando Berlinand a Researcher in the field of NLP, Question Answering systems, Search Engines and Recommender Systems. He received B.Sc. degree in computer science from Ain Shams University in 2009 and currently a Master student in computer science at Ain Shams University. He worked for the leaders of Arabic NLP; Sakhr Software and RDI then for KngineR&D and then for Orange Labs.

Dr. Safia Abbas received her Ph.D. (2010) in Computer science from Nigata University, Japan, her M.Sc. (2003) and B.Sc.(1998) in computer science from Ain Shams University, Egypt. Her research interests include data mining argumentation, intelligent computing, and artificial intelligent. He has published around 15 papers in refereed journals and conference proceedings in these areas which DBLP and springer indexing. She was honoured for the international publication from the Ain Shams University president..

Mostafa Aref is a professor of Computer Science and Vice Dean for Graduate studies and Research, Ain Shams University, Cairo, Egypt. Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. of Electrical Engineering - Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.

ملخص

CMET : إطار دلالى لمقارنة ودمج الكيانات وتطبيقاتها فى "إختيار الإجابة"

مصطفى عارف [*3], صفية عباس [*2], محود عبدالرحمن وهدان [*1]

قسم علوم الحاسب ، كلية الحاسبات والمعلومات ، جامعة عين شمس [*]
القاهرة، مصر
[1]mahmoud.a.wahdan@gmail.com
[2]safia_abbas@yahoo.com
[3]mostafa.m.aref@gmail.com

تعتبرالكيانات المسماة (أسماء الكيانات) مهمة جدا للعديد من التطبيقات المستندة إلى النصوص. نقدم إطارا عاما للكشف عن المعانى الدلاليةللكيانات، مقارنة ودمج الكيانات (CMET). ينبغي أن تكون الكيانات من نفس النوع الدلالي في التسلسل الهرمي لأنواع الكيانات. الإطار المقترح مصمم بشكل جيد ومرن وقابل للتحسينات في المستقبل، ويمكن أن يمتد إلى لغات أخرى غير الإنجليزية. العديد من التطبيقات مثل أنظمة إجابات الأسئلة بطريقة آلية، تلخيص النص أليا من الممكن أن تستفيد من معرفة مقدار تشابه الكيانات. استغللنا معرفة العلاقات الدلالية بين الكيانات في نظام الإجابة على الأسئلة آليا، ليس فقط لزيادة دقة النتائج المقترحة، ولكن أيضا لدعم الإجابات المرشحه على أساس التشابه الدلالي لها. قمنا بعمل تجربة لقياس أثر استخدام الإطار فقط في نظام الإجابة على الأسئلة آليا. وصلنا إلى 6.1٪ زيادة في مهمة إختيار الإجابة و4٪ زيادة فى الإجابة على الأسئلة آليا ككل.

Graph Matching Based Technique for Words Segmentation in Arabic Sign Language

A. S. Elons*[1], M. F. Tolba*[2]

* Scientific Computing Department- Faculty of Computers and Information Sciences- Ain Shams University-Cairo-Egypt

[1]ahmed.new80@hotmail.com

[2]fahmytolba@gmail.com

Abstract - **Many previous systems were developed for recognizing sign languages in general and Arabic sign language specifically. They achieved good results for isolated gestures but none of them was exposed to connected sequence of gestures. This paper focuses on how to recognize connected sequence of gestures using graph-matching technique, and how the continuous input gestures are segmented and classified. Graphs are a general and powerful data structure useful for the representation of various objects and concepts. This work is a component of a real-time Arabic Sign Language Recognition system that applied Pulse Coupled Neural Network for static posture recognition in its first phase.**

Key Words: - Sign Language, Gesture, Posture, Arabic Sign Language Recognition (ASLR), Pulse Coupled Neural Network (PCNN), Graph Matching, Graph Isomorphism and Sub-Graph Isomorphism.

1 INTRODUCTION

Sign language as a kind of gestures is one of the most natural means of exchanging information for most deaf people. The aim of sign language recognition is to provide an efficient and accurate mechanism to transcribe sign language into a text or speech. Sign language is a visual and manual language made up of signs created with the hands, facial expressions, and body posture and movement. Sign language conveys ideas, information, and emotion with as much range, complexity, and versatility as spoken languages [1, 2]. Signs can be static (posture) or dynamic (gesture).

Arabic Sign Language (ASL) has more than 9000 gestures and uses 26 static hand postures and 5 dynamic gestures to represent the Arabic alphabet [3]. Attempts at machine sign language recognition have begun to appear in the literature over the past decade. However, these systems concentrated on isolated signs and small dataset. This paper focuses on how a real-time sequence of dynamic gestures (a whole sentence) can be represented and segmented into primitive gestures (words) to be translated. It presents a proposed model using the graph matching technique and a customized algorithm for connected gestures classification, which is a part of Arabic Sign Language Recognition (ASLR) System. Section 2 illustrates some previous sign language recognition system and some state in the art technology used. Section 3 discusses the pre- ASL recognition system that classifies static postures. Section 4 discusses graph matching problem and how it can be employed for connected dynamic signs representation. In section 5, the traditional traversal algorithm is discussed; section 6 shows how graphs can be constructed for representation of dynamic gestures. A proposed modification of tree traversal is explained in section 7; experimental results are illustrated in section 8.

2 SIGN LANGUAGE RECOGNITION

Previous work has been done in sign language recognition, Arabic and other languages. B. Bauer and H. Hienz [4] in 2000 developed a GSL (German Sign Language) recognition system that uses colored cloth gloves in both hands. The system is based on Hidden Markov Models (HMM) with one model of each sign. A lexicon of 52 signs was collected form one signer both for training and classification. A 94% recognition percentage was achieved. N. Tanibata et al. [5] -in 2001- proposed a method of extraction of hand features and recognition of JSL (Japanese Sign Language) words. For tracking the face and hand, they could recognize 64 out of 65 words successfully by 98.4%. Chen et al [6] introduced in 2003 a system for recognizing dynamic gestures (word signs) for TSL (Taiwanese Sign Language). They used frequency domain features (Fourier Transform) plus some information from motion analysis for recognizing 20 words. The data set was collected from 20 signers but the system is person dependent. HMMs were used as the classifier. An average of 92.5% recognition rate was achieved. In 2004 and 2005, J. Zieren et al. [7, 8] presented two systems for isolated recognition: the first is for recognizing GSL, on a vocabulary of 152 signs achieving a rate of 97%, using HMM. Compared to other sign languages, not much has been done in the automation of the Arabic sign language, except few individual attempts. M. Al-Rousan et al. [9] developed two systems for recognizing 30 static gestures of Arabic sign language, using a collection of Adaptive Neuro-Fuzzy Inference System (ANFIS) networks for training and classification depending on spatial domain features. In 2003 Assaleh et al. [10] used colored gloves for collecting varying size data samples for 30 manual alphabet of Arabic sign language. Polynomial classifiers were used as a new approach for classification. In a recent (2005) work in Arabic Sign Language, Mohandes et al. [12] developed a system that recognized 50 signs of words performed by one person having 10 samples per sign. They achieved a 92% recognition accuracy. In 2010 Tolba et al.[13-16] used PCNN for recognition of 30 alphabets postures and gained a 93% recognition accuracy. The system is signer-independent and achieved system invariance against rotation, scaling and color.

3 POSTURE RECOGNITION MODULE

This paper focuses on a single component in a whole ASL recognition system. The main system architecture is illustrated in Fig 1, first of all the input sequence of gestures will be divided into static postures (frames). These frames are 2D images containing meaningful postures and movement transitions. A modern image understanding technique (PCNN) was applied to convert the 2D image to 1D time series which is called "Image signature". This signature uniquely identifies the image and can be considered as image features. Meanwhile, the classification component applies Multi-Layers Perceptron (MLP) neural network to classify the features to a posture class.

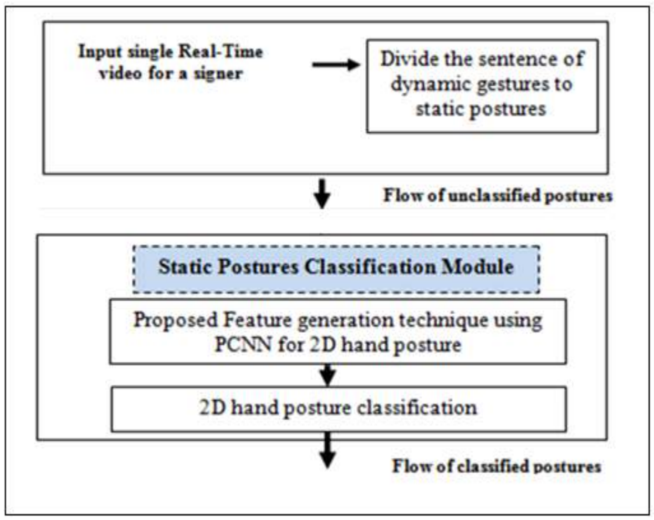


Figure 1: ASL static postures recognition system

A pulse-coupled neural network (PCNN) is a model of a biological network, specifically, a model of fragment of cat's sight network. It is a single-layer network [17, 18] composed of neurons. Each of them is linked to one pixel of the input image. Each neuron contains two input compartments: the feeding and the linking. The feeding receives an external stimulus as well as a local stimulus while the linking only receives a local stimulus [19, 20]. The local stimulus comes from the neurons within feeding radius. This local stimulus is hereafter called the firing information. The external stimulus is the intensity from the corresponding pixel in the picture. The feeding and linking are combined in a second order fashion to create the potential which then decides together with the output whether the neuron should fire or not [20]. There are several differences between the algorithms for the modified PCNN neuron and the exact physiological pulse coupled neuron. The differences are due to several simplifications made to the calculations, while still keeping the main features of the general theory. Each neuron in the modified PCNN could be described by the following set of equations [20]:

$$L(i) = L(i-1)\cdot e(-\alpha L) + VL\cdot (R*Ysur(i-1)) \quad (1)$$

$$F(i) = S+F(i-1)\cdot e(-\alpha F) + VF\cdot (R*Ysur(i-1)) \quad (2)$$

$$U(i) = F(i)\cdot[1 + \beta\cdot L(i)] \quad (3)$$

$$\theta(i) = \theta(i-1)e-(\alpha q)+V\theta\ Yout(i-1) \quad (4)$$

$$U > \theta(i) \Rightarrow Yout = 1 \quad \text{(Firing Condition)} \quad \text{otherwise} \Rightarrow Yout = 0 \quad (5)$$

Where L(i) is input linking potential, F(i) is input feeding potential and S represents the intensity of a given image element. U(i) is the activation potential of neuron, θ(i) is threshold potential of neuron and (i) is iteration step. Parameters αL, αF and αq decay coefficients, β is linking coefficient and parameters VL and VF are coefficients of the linking and threshold potential. Ysur is the firing information that indicates whether the surrounding neurons have fired or not and Yout indicates whether this neuron fires or not. R is the matrix of weight coefficients and * is convolution operator. An example of the modified PCNN neuron architecture is shown in Fig 2.

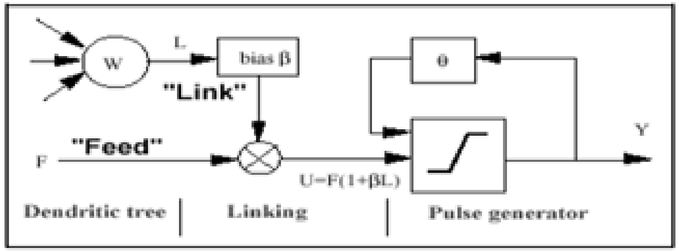


Figure 2. Model for Modified PCNN neuron

PCNN model with some characteristics, such as strong adaptive capturing ignition, has been widely applied to image de-noising, image smoothing, image processing, image segmentation and image fusion. It is also partly used in shortest path optimization, Structural layout optimization, etc.

Many feature generation methods have been developed using pulse-Coupled Neural Network (PCNN). The comparative study on them is deeply investigated by Tolba et al [13] which is out of this paper scope. Meanwhile, the equation proposed by Tolba [13] has been used to generate image signature:

$$g(n)=\frac{\sum_{i=1}^{n}(X(i)\times Y(i)\times CF(i))}{\sum_{ij} S_{ij}} \qquad (6)$$

Where Y(i) is output quantity based on step function, CF(i) is the continuity factor and X(i) is output quantity based on sigmoid function. Figure 3 shows the image signature using (8) of the same hand posture. One image has a uniform light distribution and another one is scaled and is distorted.

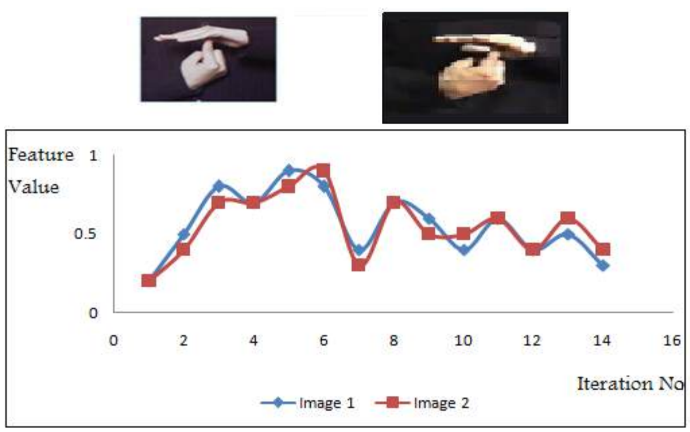


Figure 3- images and signatures for a uniform light distribution and another is a scaled distorted one

4 PROPOSED GRAPH MATCHING APPROACH

The main idea of this paper is to customize the graph matching problem and algorithms as a proposed solution for connected gestures classification. The gestures which represent alphabets or words are stored in database as models graphs; each graph consists of a group of vertices and edges and these graphs are attributed directed graphs. The connected flow of input gestures is represented by the input graph. Figure 4 illustrates the main steps for connected gestures recognition Let: $G = (V, E, \mu, \nu, L_v, L_e)$ be a model graph and M its corresponding $n \times n$ - adjacency matrix. Furthermore, let ***A(G)*** denotes the set of all permuted adjacency ,matrices of ***G*** , $A(G) = \{M_P | M_P = PMP^T \text{ where } P \text{ is a } n \times n \text{ permutation matrix}\}$. The total number of permuted adjacency matrices is $|A(G)| = n!$ [21] as there are $n!$ different permutation matrices of dimension n. For a model graph G with corresponding $n \times n$ adjacency matrix ***M*** and an input graph G_I with an $m \times m$ adjacency matrix M_I and $m \leq n$, determine whether there exists a matrix $M_P \in A(G)$ such that $M_I = S_{m,m}(M_P)$. If such a matrix M_P exists, the permutation matrix ***P*** corresponding to M_P describes a sub-graph isomorphism from G_I to ***G***, i.e. $M_I = S_{m,m}(M_P) = S_{m,m}(PMP^T)$. If G_I and ***G*** are of equal size, the permutation matrix ***P*** represents a graph isomorphism between G_I and ***G, i.e.*** $M_I = PMP^T$. One proposes to organize the set ***A(G)*** in a decision tree that each matrix in ***A(G)*** is classified by the tree. The features that will be used for the classification process are the individual elements in the adjacency matrices. One introduces a new notation for an $n \times n$ adjacency matrix $M = (m_{ij})$. One says that the matrix consists of an array of so-called ***row-column*** elements ***ai***, where each ***ai*** is a vector of the form [22-24]:

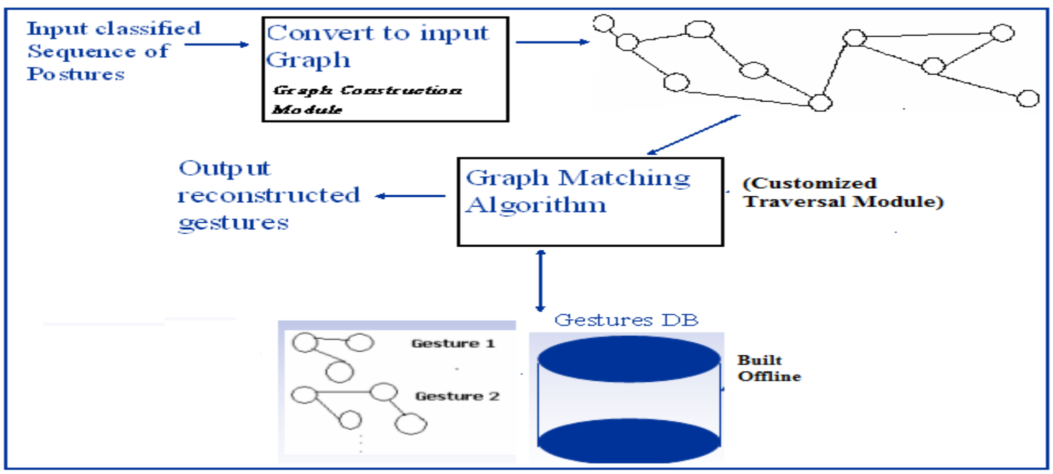


Figure 4: the main steps to recognize the connected gestures

$$a_i = (m_{1i}, m_{2i}, \ldots, m_{ii}, m_{i(i-1)}, \ldots, m_{i1})$$

The matrix can then be written as: $M = (a_1, a_2, \ldots, a_n); \; i = 1, \ldots, n$. Figure 5 illustrates the structure of an adjacency matrix ***M*** with regard to its row-column elements

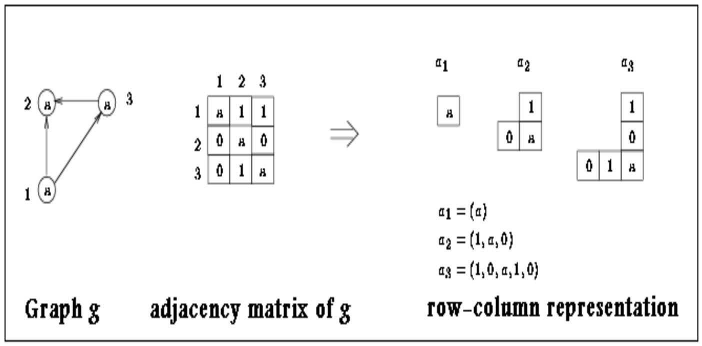


Figure 5: The row-column representation of the adjacency matrix

In Fig 6 a graph, g1, and its corresponding decision tree is shown. The nodes of the decision tree are represented by shaded circles. Each directed branch from one node to another has associated with it a row-column element. At the top of Fig. the set A(g1) of permuted adjacency matrices of g1 is listed.

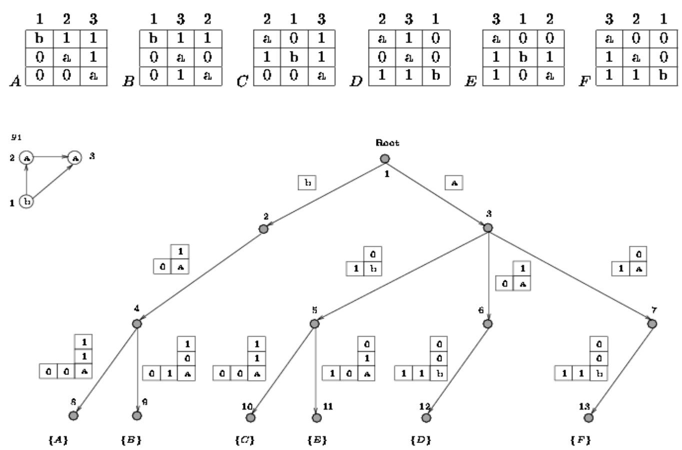


Figure 6: The decision tree

5 EXPERIMENTS

In this section, one illustrates the results of the gesture construction module; a new decision tree-based sub-graph isomorphism algorithm is customized and implemented. The experiment contains 30 connected sentences of total 100 words. First, a study is conducted to measure the effect of graph construction approaches on the decision tree size in terms of nodes count using the 3 approaches (graph construction section). Figure 7 illustrates the size of the offline built decision tree against the size of the graph models database size.

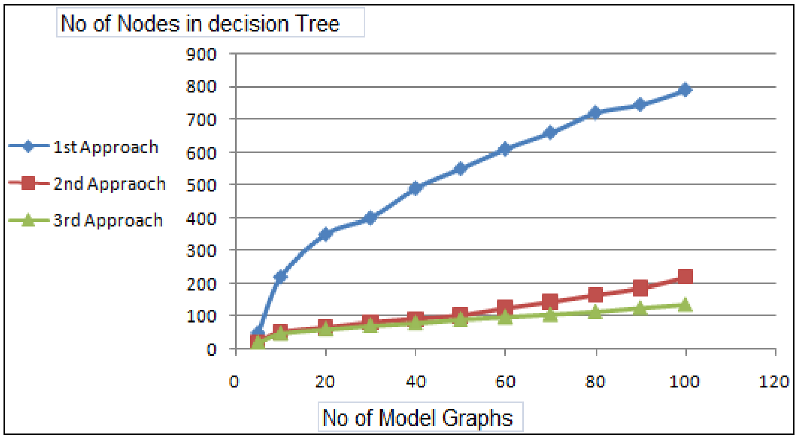


Figure 7: The tree size against model graphs database size

It is clear that applying the 3rd approach leads to a minimum tree size relative to the other approaches. The 1st approach gives the maximum tree size because it ignores the transitional movements; it causes that each frame represents a distinct vertex in the gesture graph. The equation used in the 2nd approach decreases the tree size because it omits transition frames. Meanwhile, it does not accomplish the minimum size because of its relative nature of the threshold values. The performance is then measured by: computing the execution time in seconds and counting the number of basic computation steps that are performed while searching for all graph and sub-graph isomorphism. A basic computation step is defined as the comparison of one model graph vertex and its incident edges to one input graph vertex and its incident edges. Figure 8 illustrates the system performance measures using execution time against model graphs database size. Moreover, the computational steps needed against the decision tree nodes count.

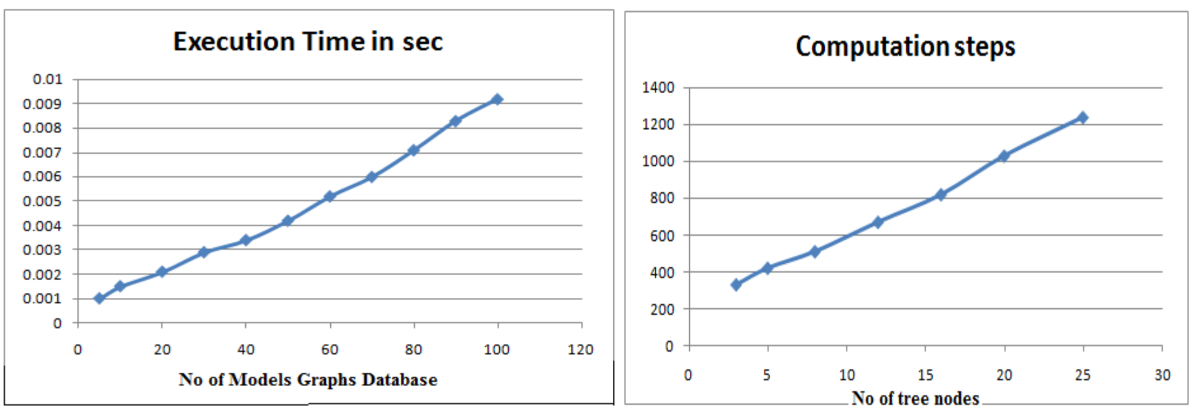


Figure 8- a) Execution time in sec against the database size. b) Computational steps against number of tree nodes.

The performance measurements emphasizes the polynomial nature of the proposed customized sub-graph isomorphism algorithm; it concludes that as much as the gestures graphs database size increases, the running time will be still bounded by a polynomial. As mentioned before, 30 connected sentences are tested using a graph database containing 100 dynamic gestures (words). The recognition accuracy of the proposed system is measured using the number of 3 words-sentences that have been successfully translated and the number of 4 words- sentences that have been successfully recognized.

6 CONCLUSION

This paper proposes and implements ASLR system of it focuses on recognizing the continuous gestures using graph-matching technique. Gestures have been divided into elementary elements, static postures. Gesture recognition is performed by "graph matching" algorithm. The gestures, which represent alphabets or words, are stored in database as models graphs. Each graph consists of a group of vertices and edges. The algorithm used for graph and sub-graph isomorphism detection is based on the decision tree paradigm. In the computational complexity analysis, it is shown that the new algorithm has a worst-case run time complexity that is only quadratic in the size of the graphs that are to be compared. The recognition rate does not lay down 70% for 100 gestures composing 30 continuous sentences.

REFERENCES

[1] B. Doner. "Hand Shape Identification and tracking for Sign Language Interpretation". Looking at People Workshop, Chambery, France, 1993.

[2] B. Bauer and H. Hienz, "Relevant Features for Video-Based Continuous Sign Language Recognition", Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, pp. 64– 75.

[3] Nashwa El-Bendary, Hossam M. Zawbaa , Mahmoud S. Daoud, Aboul Ella Hassanien, and Kazumi Nakamatsu. ArSLAT: Arabic Sign Language Alphabets Translator. International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 3 (2011) pp. 498-506.

[4] B. Bauer and H. Hienz, "Relevant Features for Video-Based Continuous Sign Language Recognition", Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, pp. 64– 75

[5] N. Tanibata, N. Shimada and Y. Shirai, Extraction of Hand Features for Recognition of Sign Language Words", Proceedings of the 15th International Conference on Vision Interface, Calgary, Canada, 2001.

[6] F-S. Chen, C-M. Fu and C-L. Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models, Image and Vision Computing, 2003; No. 21, pp. 745–758.

[7] J. Zieren and K-F. Kraiss, Non-Intrusive Sign Language Recognition For Human- Computer Interaction, in 9th IFAC/IFIP/IFORS/IEA Symposium Analysis, Design, and Evaluation of Human- Machine Systems, Atlanta, GA, 2004, pp. 221-228.

[8] J. Zieren and K-F. Kraiss, Robust Person- Independent Visual Sign Language Recognition, Proceedings of Pattern Recognition and Image Analysis, Second Iberian Conference, Estoril, Portugal, 2005.

[9] M. Al-Rousan, O. Aljarrah and M. Hussain, Automatic Recognition of Arabic Sign Language Finger Spelling, International Journal of Computers and Their Applications, Issue 1076-5204, Vol. 8, No.2. 2001, pp. 80-88.

[10] K. Assaleh and M. Al-Rousan, "Recognition of Arabic Sign Language Alphabet Using Polynomial Classifiers", EURASIP Journal on Applied Signal Processing society No. 13, 2005, pp. 2136-2145.

[11] M. Mohandes and M. Deriche, "Image based Arabic sign language recognition", Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium, Vol. 1, 2005, pp. 86- 89.

[12] M. Mohandes and M. Deriche, "Image based Arabic sign language recognition", Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium, Vol. 1, 2005, pp. 86- 89.

[13] Samir A. ,Abull-ela,Tolba, M.F., "Neutralizing lighting non-homogeneity and background size in PCNN image signature for Arabic Sign Language recognition," Neural Computing and Applications, Pages 1-7, 2012.

[14] M.E. Fathy, A.S. Hussein, and M.F.Tolba, "Fundamental matrix estimation: a study of error criteria," Pattern Recognition Letters, vol. 32, no. 2, pp. 383–391, 2011.

[15] S. Ali, A. S. Hussein, M. F. Tolba, and A. H. Yousef, "Large-Scale Vector Data Visualization on High Performance Computing," Journal of Software, vol. 6, issue 2, pp. 298-305, 2011.

[16] A.S Ali., A.S. Hussien, M.F. Tolba, and A.H. Youssef, "Visualization of large time-varying vector data," 3rd IEEE International Conference on Computer Science and Information Technology, art no. 5565176, pp. 210-215, 2010.

[17]Y. D. Ma, L. Li and Y. F. Wang, The principles of Pulse Coupled Neural Networks and its application. Beijing, Science Press, 2006.

[18]R. Eckhorn , H. J. Reitboeck , M. Arndt , P. Dicke, Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex, Neural Computation, v.2 n.3, p.293-307, Fall 1990.

[19] R.Forgá: Feature Generation Improving by Optimized PCNN, Applied Machine Intelligence and Informatics. SAMI 2008. 6th International Symposium on Volume ,Page(s):203 - 207, Issue , 21-22, Jan. 2008 ,2008.

[20]J. M. Kinser, andC. Nguyen, Pulse Image Processing Using Centripetal Autowaves. Proc SPIE, Vol. 4052, pp.278-284, 2000.

[21] Min-Wen Chao,Chao-Hung Lin, Chih-Chieh Chang and Tong-Yee Lee, " A graph-based shape matching scheme for 3D articulated objects" Computer Animation and Virtual Worlds, Volume 22, Issue 2-3, pages 295–305, April - May 2011.

[22] Manjari Gupta, " Design pattern mining using greedy algorithm for multi-labelled graphs" international Journal of Information and Communication Technology, Volume 3, Number 4/2011, pages 314-323.

[23] H. Qiu, E.R. Hancock: Graph partition for matching. In Proc. 4th Int. Workshop on Graph-Based Representations in Pattern Recognition, Springer LNCS2726, 2003, 178—189.

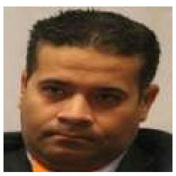
[24] M. Delalandre, E. Trupin, J.M. Ogier: Local Structural Analysis: A Primer. In Proc. Workshop on Graphics Recognition, Springer, LNCS3088, 2004, 220—231.

[25] H. Qiu, E.R. Hancock: 2003 " Graph partition for matching". In Proc. 4th Int. Workshop on Graph-Based Representations in Pattern Recognition, Springer LNCS2726, pp.178—189.

[26] B T Messmer, H Bunke," A decision tree approach to graph and subgraph isomorphism detection" Pattern Recognition (1999), Volume: 32, Issue: 12, Publisher: Elsevier, pp: 1979-1998.

BIOGRAPHY

Prof. Dr. Mohamed Fahmy Tolba is a Professor of Scientific Computing, FCSIS (1996-Present). Dr. Tolba has more than 150 publications in the fields of AI, Image Processing, Pattern Recognition, OCR, Scientific Computing, Simulation and Modeling. Also Dr. Tolba has supervised more than 50 M.Sc. and 25 Ph.D. degrees in Ain Shams University and other Egyptian Universities.



Dr. Ahmed Samir is a Lecturer at the Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt. His research interests: Image Processing and, pattern recognition and AI. He worked in Arabic Sign Language Recognition field from 2004 till now.

مقترح لتجزئة اشارات لغة الاشارة العربية باستخدام تقنية مطابقة الأشكال

أحمد سمير ، محمد فهمي طلبة

قسم الحسابات العلمية - كلية الحاسبات و المعلومات- جامعة عين شمس

خلاصة:

تقوم معظم محاولات الباحثين في مجال ترجمة لغة الاشارة العربية أو غيرها بالتركيز علي التعرف علي الاشارات المنفصلة. في هذا البحث تتم معالجة مشكلة تقسيم الاشارات المتواصلة الي اشارات منفصلة يتم التعرف عليها. يقوم الباحثون بتطبيق أداة رياضية مشهورة و عي مطابقة الاشكال و هي تتميز بميزات أهمها المحافظة علي استقرار الشكل اذا ما تم احداث متغيرات عليه و هو شئ لا يمكن افتراض عدم حدوثه.

Classification of Text Images on Social Network Using Linguistic and Behavioral Features

Ahmad M. Abd Al-Aziz*1,MervatGheith*2, Ahmed SharfEldien, Prof.*3

*The British University in Egypt (BUE)
Cairo, Egypt
1Ahmed.abdelaziz@bue.edu.eg

*Institute of Studies and Statistical Researches, Cairo University
Cairo, Egypt
2mgheith@cu.edu.eg

*Faculty of Computers and Information, Helwan University
Helwan, Egypt
3profase2000@yahoo.com

***Abstract*—The visual content including images provides important information beyond what actual text reveal on Social Network Sites. Studies analyze the images uploaded on Social Network Sites to infer emotionsand to analyze the sentiment by using image visual features e.g. low level features or linguistic features such as image metadata (e.g. captions, description…etc.) and comments. In this work, we propose a system to classify three different text images (Memes image, quote within image and condolences within image) into {*happy*, *emotionless*, *unhappy*}respectively by combining the behavioral features extracted from Facebook user behavioral actions on text images (like, comment) with linguistic features extracted from image comment text. Two machine learning classifiers are used; Support Vector Machine and Naive Bayes to classify 1127 text images extracted from Facebook. The experimental results show promising performance especially for Naive Bayes classifier.**

1 INTRODUCTION

Visual content including images and videos is becoming a medium for social interaction among users on the Internet, including the popular social network platforms such as: Facebook[1], Flickr[2], Twitter[3], etc... This visual content provides rich complementary information beyond what the actual text reveals. In such cases such as Facebook, extracting information from visual content is critical to understand the rich emotions, affect, or sentiment conveyed in the multimedia content [1]. To understand emotions in such online interactions, visual contentbased emotion detection has been studied recently and has been shown to achieve promising results in predicting sentiments expressed in multimedia tweets with photo content ([2], [1]). Among visual contents, studies revealed that images accounted for 75% of content posted by Facebook pages worldwide [3].

There are many types of posted images on Facebook: places, people, memes, animals, mobile screenshots…etc. Among these types, text images play an important role in Social Network Sites (SNS), there are many types of text image as shown in Fig. 1Quotes within image (a), Memes (b), Condolences text within image (c), Religion and spiritual text within image (d).

Figure 1: Sample of text images posted on Facebook

1 http://www.facebook.com
2 http://www.flickr.com
3 http://twitter.com

In literature, several studies focused on emotion analysis in social networks from text content e.g. tweets, comments or blogs [4]. Similar to the texts, images are also used to express individual emotions as people often prefer to use warm colors like such as pink to express *happiness* and cold colors such as blue to express *sadness*, accordingly, researchers used the visual features to infer emotions from image ([5], [6]).In addition, studies recently used a combination between the visual features and linguistic features extracted from image comments, description, caption…etc. to enhance the emotion detection task [7]. Most of these studies work on artistic photograph images or abstract paintings. In addition, the sentiment analysis is performed mostly in English texts.

The main idea of this work is to classify the text images (memes, condolences text within image, quote within image) posted on Facebook into {*happy*, *unhappy*, *emotionless*} respectively by using a combination between two features: linguistic features extracted from users' comments written in Arabic and English text, and behavioral features (e.g. like, comment) extracted from user's behavior of action toward the posted text image.

The paper is organized as follows. Section 2 introduces the related works. Behavioral and linguistic features are discussed in section 3 and section 4 respectively. The machine learning model is discussed in section 5, while the proposed system will be discussed in section 6. The experimental results are shown in section 7 before the concluding remarks in section 7.

2 RELATED WORKS

The role of visual content such as images or videos has become more important in increasingly popular social media such as Facebook. In such cases, extracting information from visual content is significant in understanding the rich emotions, affect, or sentiments in the multimedia content. Existing research of sentiment or affect analysis of social multimedia restricted to direct mapping of low-level visual features to affects using artistic photographs and abstract paintings ([5], [6]), these researches are conducted based on the justification for there being a link between visual content and evoked emotion/sentiment [8]. One of the first automatic emotional image analysis systems is proposed in [9], in their system, a novel technique to obtain a high-level representation of art images was proposed, allowing the extraction of emotional semantics such as *action*, *relaxation*, *joy* and *uneasiness*. Another study tries to attempt to map low level visual features to high-level affect classes to detect emotions from image ([10], [11]), despite the promising results, the visual features by directing mapping from low level features is quite limited due to the semantic gap and the emotional gap[12]. Recently, studies combine the visual features with linguistic features to detect emotions from images on social network application platforms. One of the most effective model was proposed in [7], they extracted visual features (e.g., five color theme) from the image and emotional words (e.g. "*amazing*", "*gorgeous*") appeared in comments, their main goal is to automatically extract emotions from the images by leveraging all the related information (e.g. visual features, comments, and friendships), the experiments on a Flickr dataset demonstrates that their model improves the performance on inferring users' emotions. Detecting emotion from images also studied by analyzing the image caption and description as in [13].

Based on previous research done on Facebook, sentiment analysis is performed mostly in English texts and few attempts in Arabic text, in addition the previous research did not focus on text images but the main concern is about artistic photographs or abstract painting. In order to provide another alternative, this work focuses on classifying text images posted in Facebook by combining the behavioral features and linguistic features in order to classify them into {*happy* , *unhappy* , *emotionless*} classes.

3 BEHAVIORAL FEATURES

In this work we define the behavioral features in order to use them in our emotion classification task. To define the behavioral features we test whether there is an existence explicit relationship between the text image type (meme, quote images, condolence text within image) and the user behavioral action (e.g. like or comment). In order to do that, we studied a sample of 477college undergraduate and graduate students (344 males, 133 females; mean age [*M*] = 20.80 years, standard deviation [*SD*] = 4.60). All participants had been using Facebook for at least one year. Participants were asked through a designed survey to select one of the possible Facebook behavioral actions (like or comment) to response on our three different types of text images and they asked to put any possible comment (s) if they choose comment action.

The results show as shown in Fig.2, there is a relationship between the text image type and the user behavioral actions reveal that some text image are super-likeable, but not so conversational (e.g. quotes within image), and some text image are super-conversational and receive more comments than other text image (e.g. meme images and condolences within image).

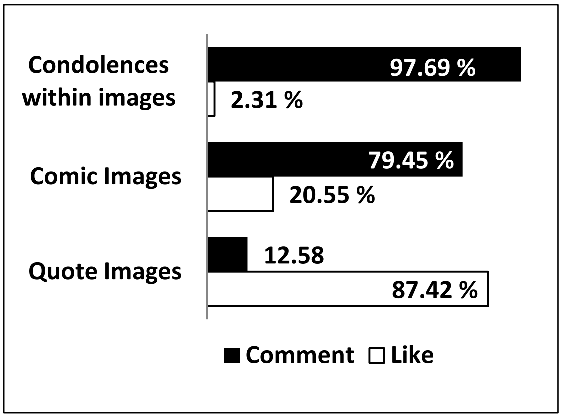


Figure 2: Percentage of (Like and Comment) toward three types of text images

Accordingly, we proposed two behavioral features: f_1 (likable) by quantifying the like ratio of text image defined as the total number of like hits/ total number of like hits and comment hits.

$$\text{Like (ratio)} = \frac{\sum_{i=1}^{n}(Like\_hits)}{\sum_{i=1}^{n}(like\_hists)+\sum_{j=1}^{n}(comment\_hits)}$$

f_2 (conversational) by quantifying the comment ratio of text image defined as the total number of comments divided by the total number of like hits and comments.

$$\text{Conversational (ratio)} = \frac{\sum_{i=1}^{n}(comments)}{\sum_{i=1}^{n}(like\_hists)+\sum_{j=1}^{n}(comment\_hits)}$$

The results also reveal that there is no difference between memes and image contains condolences text in their behavioral characteristics since both of them are more conversational than likable. So, we proposed another type of features in order to differentiate between them significantly as shown in the next section.

4 LINGUISTIC FEATURES

To define the linguistic features we test whether there is an existence explicit relationship between the text image type (meme, quote images, condolence text within image) and the pattern of user words and sentences used with each one? To investigate that we relied on the result of our pilot study and analyzing all participant comments on our three text image types. The results show the total number of comments on this type of text image was 400 all of them use same pattern of words and sentence with sympathy meaning (e.g. البقاء لله or "accept my condolences", R.I.P,…etc) and participants used mixed languages between Arabic and English as shown in Fig.3. Out of our sample, 46 participants suggest some comments to response to quotes image, the analysis of comments as shown in Figure 4 reveals that the response language was English and the most of words and sentences are with the same agreement meaning (e.g. true, agree…etc.). Finally, for the memes images the total number of participants who commented on this type of text image was 241, the analysis of comments reveal that their words and sentences in their comments have the same meaning showing their tendency of cognitive experiences to provoke laughter and provide amusement (e.g. hahaha, :D…etc.) and participants used mixed languages between Arabic and English.

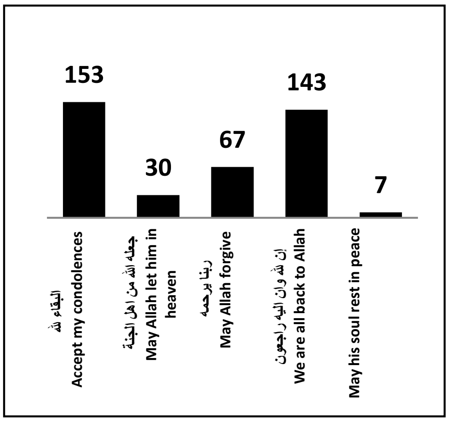


Figure 3: Frequencies of comments on image contains condolences text (*n=400*)

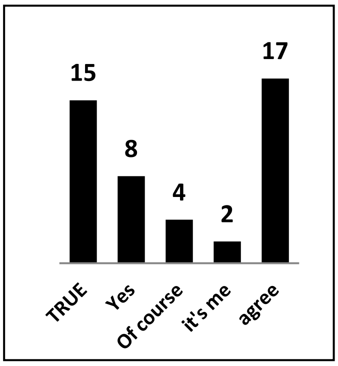


Figure 4: Frequencies of comments on image quote image (*n=46*)

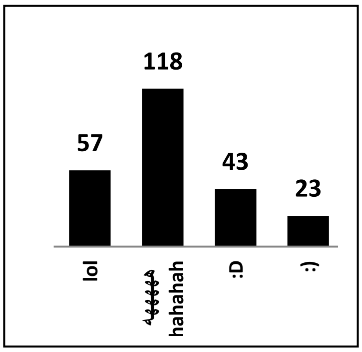


Figure 5: Frequencies of comments on memes image (*n=241*)

Based on the previous results we defined three different linguistic features.
f3 (agreement_word): Text contains agreement word(s).
f4 (sympathy_word): Text contains sympathy word(s).
f5 (happiness_word): Text contains humor, happiness or sarcastic word(s).

5 MACHINE LEARNING MODEL

Determining emotion of a text image can be defined as a classification problem. For that case, let ***B*** denote the user's behavior toward the text image, and ***a*** is one of embedded behavioral units {***like***,***comment***}, where ***a***∈***B***. Let **T** denote the text, and ***s*** an embedded linguistic unit, such as sentence, where ***s***∈***T***. Let ***E*** denote the emotions classes where ***E***= {em_1, em_2, em_3}, where $\boldsymbol{em_1}$ denotes {***happy***}, $\boldsymbol{em_2}$ denotes {***unhappy***} and $\boldsymbol{em_3}$ denotes {***emotionless***} or absence of emotion. The main goal is to determine a function $\boldsymbol{f :s,a \rightarrow em_i}$. The mapping is based on $F= \{f_1,\ldots,f_n\}$, where F contains the behavioral and linguistic features.

6 PROPOSED SYSTEM

Our proposed system as shown in Fig.6 starts by accepting the incoming text image, and then we extract the user responses toward this text image including the total number of like(s), the total number of comment(s) and finally, extracting the conversational comment text.

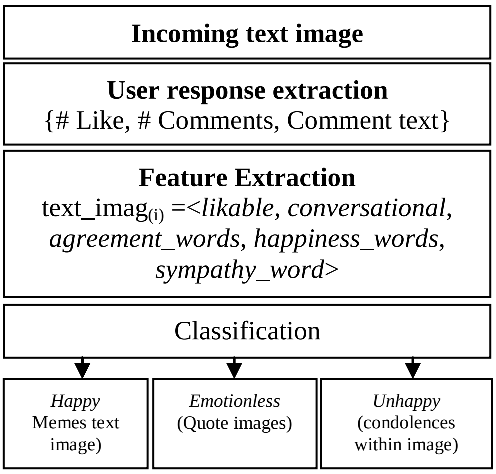


Figure 6: Proposed system

The system extracts features to generate a vector of five features (two behavioral features and 3 linguistic features) then by using a machine learning approach the system classifying the incoming text image into one of our three classes (*happy*, *unhappy*, *emotionless*).
For the linguistic features we preprocessed the comment text by usingAMIRA toolkit version 2.1, this toolkit is a group of tools for the processing of modern standard Arabic text, the output of this process is a number of segmented words included in the comment messages, from these chunks of word, agreement, happiness and sympathy words included are extracted.

7 EXPERIMENTAL RESULTS

A. Dataset Description and Experiment setup

Since the main aim of our work aims to classify the images posted on Facebook into our three emotion categories {*happy*, *emotionless*, *unhappy*}, we mainly focused on extracting three features for each image:

- Number of likes on image
- Number of comments on image
- Words in comments

In this work, we use Rfacebook- package version 0.5 (2014)[4] to extract images, number of like(s), number of comment(s) and their comment messages (if exists). The dataset consists of 1127 three types of images: Memes, Quotes within image and condolences within image. The number of extracted images for each type and examples are shown in Table 1.

TABLE I
EXAMPLE OF DATASET

Image Type	**Dataset Description**				
	Total Number	**Example**	**Number of Likes**	**Number of Comments**	**Comment Message(s)**
Memes	265		15	16	هههههه حلوة اوى يا لوله
Quotes within Image	790		59	1	that's a fact
Condolences within Image	72		10	23	البقاء لله
Total	1127				

This experiment starts by defining our feature set which consists of 5 features before running our machine learning classifier. Two machine leaning classifiers were run: Support Vector Machine (SVM) Naïve Bayes (NB) by using WEKA version 3.7. The results have been evaluated for each method and for both methods in order to analyze their performance of detecting our three classes from text image related.

B. Experiment Results

The results of ten-fold cross validation for classification experiments performed using NB for 1127 total number of text images in our dataset show correctly classified 89.26% and 10.74% incorrectly classification.. A detailed description of experimental results using NB classifier shows the classification precision, recall and F-measure for each category in our dataset as shown in Table 2.

4 This package provides a series of functions that allow R users to access Facebook’s API to get information about users and posts, and collect public status updates that mention specific keywords.

TABLE II
DETAILED RESULTS OF NB CLASSIFIER

Class	**Precision**	**Recall**	**F-Measure (100%)**
Happy	0.77	0.845	80.6
Emotionless	0.931	0.928	93
Unhappy	1	0.681	81

The experimental results reveal that the NB classifier achieved best results for {*emotionless*} class (93%). Whilethe worst results achieved for text images with {*happy*} class (80.6%).

The results of ten-fold cross validation for classification experiments performed using SVM for 1127 total number of text images in our dataset show correctly classified 88.56% and 11.53% incorrectly classification. A detailed description of experimental results using SVM method shows the classification precision, recall and F-measure for each category in our dataset as shown in Table 3.

TABLE III
DETAILED RESULTS OF SVM CLASSIFIER

Class	**Precision**	**Recall**	**F-Measure (100%)**
Happy	0.858	0.709	77.7
Emotionless	0.885	0.961	92.1
Unhappy	1	0.694	82

The experimental results reveal that SVM classifier achieved best results for {*emotionless*} class (92%.1). While the worst results achieved for text images with {*happy*} class (77.7%).

Experimental results reveal minor differences between two machine learning classifiers as shown in Figure 7.

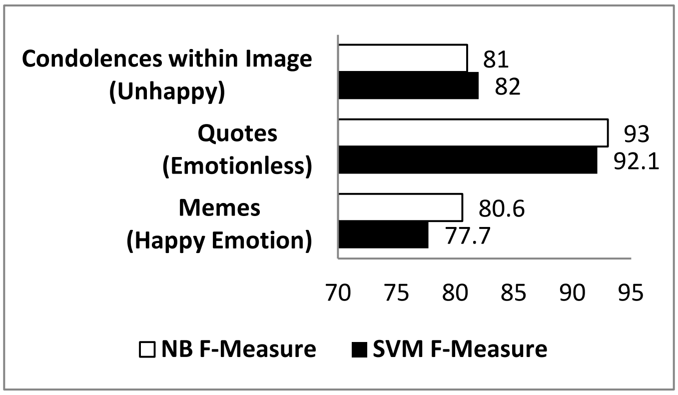


Figure 7: F-measure for NB and SVM classifiers for each text image category

The results reveal that both classifiers achieved best classification for {*emotionless*} class (93% and 92.1% for NB and SVM respectively) and the worst classification for {*happy*} (80.6% and 77.7% respectively).

We evaluate the results by using different dataset consists of 591 text images. We test the evaluated dataset by using NB classifier since it achieved better performance than SVM classifier. NB classifier achieved 85.96% correct classification for evaluated dataset not included in the training dataset. The detailed comparison between training and evaluated datasets is shown in Table 4.

TABLE IV
COMPARISON BETWEEN TRAINING AND EVALUATED DATASET USING NB CLASSIFIER

Class	**Training Dataset (1127 instances)**			**Training Dataset (591 instances)**		
	Precision	**Recall**	**F-Measure (100%)**	**Precision**	**Recall**	**F-Measure (100%)**
Happy	0.77	0.845	80.6	0.691	0.734	71.2
Emotionless	0.931	0.928	93	0.91	0.906	90.8
Unhappy	1	0.681	81	0.903	0.757	82.4

The results of evaluated dataset reveal that NB classifier achieved best classification for {*emotionless*} class (90.8%) and the worst classification for {*happy*} class (71.2%).

8 CONCLUSION

In this work, we use a novel behavioral feature combined with linguistic features to classify the text images on Facebook into {happy, emotionless, unhappy} categories by using two learning based approach classifiers; SVM and NB classifiers. The experiment has been tested on three types of text images posed on Facebook extracted (Memes, Quotes within images, Condolences within images). Selected Features contain two behavioral features (likable and conversational) and three linguistic features (agreement words, happiness words, sympathy words) collected from the results of our preliminary study on Facebook users who responses to our three type of text images. Results reveal minor difference between SVM (F-measure=88.56%) and NB (F-measure =89.26%) classifiers. We conclude that NB classifier can predict the emotion category from text image posted on Facebook based on the results achieved from our evaluated dataset. There is still much work that can be performed. First, considering high-quality and high-coverage of emotion word expressed in Arabic, English and transliteration forms, since users may add their comments in each one of these forms or even mixed between them; second; considering multi-words expressions, emotion word ambiguity; and finally, we need to expand the behavior features by adding "share" with "like" and "comment" in order to achieve better classification performance.

REFERENCES

[1]J., Yuan, Q., You, S. McDonough and, J., Luo, "Sentribute: Image sentiment analysis from a mid-level perspective", *inWorkshop on Sentiment Discovery and Opinion Mining*,2013.

[2]D., Borth, R., Ji,T., Chen, T., Breuel, and S.-F., Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs". *in Proc.ofACM Multimedia*. 2013.

[3]G., Wang, D., Hoiem and D., Forsyth, "Building text features for object image classification", *In Proc. of 19th international conference on pattern recognition*, 2009.

[4] J., Tang, Y., Zhang, J., Sun, J., Rao, W., Yu, Y. Chen, and A., Fong, "Quantitative study of individual emotional states in social networks", *IEEE Transactions on Affective Computing*, vol.3, No.2, pp.132–144, 2012.

[5]J. Machajdik and A., Hanbury, "Affective image classification using features inspired by psychology and art theory" *In Proc. of ACM Multimedia*, 2010.

[6]J., Jia, S. Wu, X., Wang, P., Hu, L., Cai, and J. Tang, "Can we understand van gogh's mood? learning to infer affects from images in social networks". *In Proc. of ACM Multimedia*, 2012.

[7] Y., Yang, J., Jia, S., Zhang, B., Wu, Q., Chen, J., Li, C., Xing and J ., Tang, "How Do Your Friends on Social Media Disclose Your Emotions?", *in proc. of Association for the Advancement of Artificial Intelligence (AAAI'14)*, 2014.

[8] S. Schmidt and W. G. Stock, "Collective indexing of emotions in images", *a study in emotional information retrieval. Journal of the American Society for Information Science and Technology*, Vol. 60, No. 5, pp. 863–876, 2009.

[9] C., Colombo, A., DelBimbo and P., Pala, "Semantics in visual information retrieval. *In proc. of IEEE Multimedia*, Vol. 6, No. 3, pp. 38–53, 1999.

[10] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory", *in Proc. of ACM Multimedia*, pp. 83–92, 2010.

[11] A., Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV", *IEEE Signal Processing Magazine*, 2006.

[12] W., Wang and Q., He, "A survey on emotional semantic image retrieval", *In Proc. of 15th IEEE International Conference on Image Processing (ICIP)*, 2008.

[13] M.,Ulinski, V., Soto, J., Hirschberg, "Finding emotion in image descriptions, *in Proc. of the 1stInternational Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'12)*, No. 8, 2012.

BIOGRAPHY

Prof. Ahmed SharafEldin, Prof. is a distinguished professor in computer science; he is among the first people in Egypt who specialized in computing since 60's. He has more than 180 published papers in national and international journals and conferences. Moreover, he authored 7 books in computing. He is currently the dean of faculty of information technology and computer science at Sinai University.

MervatGheith, Ph.D. is assistant professor at Institute of Statistical Studies and Researches (ISSR), Cairo University, her main research interests are artificial intelligence, pattern recognition, image processing and natural language processing. Mervat published several scientific papers in fields of natural language processing and artificial intelligence.

Ahmad Abd Al-Aziz, Ph.D., Ph.D. received his Ph.D. in computer science from Institute of Statistical Studies and Researches (ISSR), Cairo University. He is a lecturer at British University in Egypt (BUE). His main areas of research interest are natural language processing, sentiment analysis and opinion mining, social network analysis. He had multiple disciplines background: 1) social sciences, since he received his Ph.D. in psychology from Ain Shams University, 2) computer science which empowered him to scale out his research areas in several topics related to applying information theories on social sciences.

تصنيف الصور النصية على شبكات التواصل الإجتماعى بإستخدام الخصائص اللغوية والسلوكية

أحمد عبد العزيز- الجامعة البريطانية فى مصر
مرفت غيث –معهد الدراسات والبحوث الإحصائية – جامعة القاهرة
أحمد شرف الدين أحمد – كلية الحاسبات والمعلومات – جامعة حلون- عميد كلية تكنولوجيا المعلومات وعلوم الحاسب- جامعة سيناء

خلاصة:
يعد المحتوى المرئى الذى يحتوى على الصور مصدراً هاماً من المصادر التى تمدنا بالمعلومات الهامة أكثر مما تمده النصوص على مواقع شبكات التواصل الاجتماعى. هناك ابحاث اهتمت بدراسة وتحليل الصور الموضوعة على مواقع التواصل الإجتماعى لاكتشاف ما تحتويه من انفعالات و تحليل المشاعر عن طريق الخصائص المرئية لها (مثل الخصائص المرئية ذات المستوى المنخفض) و عن طريق الخصائص اللغوية الخاصة بالبيانات المتعلقة بالصورة (مثل التعليقات والتوصيف...الخ). فى الدراسة الحالية نقوم بعرض نظام يقوم بتصنيف ثلاثة أنواع من الصور المنتشرة على مواقع التواصل الاجتماعى (الصور الفكاهية التى تحتوى على شخصيات، الصور التى تحتوى على أقوال مأثورة، الصور التى تحتوى على عبارات العزاء) الى ثلاث أنفعالات (فرح ، معدومة المشاعر، حزن) عن طريق دمج الخصائص السلوكية لمستخدمى برنامج "فايس بوك" "Facebook" والتى تتضمن (الاعجاب والتعليق) مع الخصائص اللغوية المستخلصة من التعليقات الخاصة بالصور. تم استخدام نوعين من مصنفى تعلم الآلة وهما : Support Vector Machine و Naïve Bayes لتصنيف 1127 صورة تم استخلاصهم من برنامج "فايس بوك" . اظهرت نتائج التجربة أداء جيد فى التصنيف وخصوصاً النتائج الخاصة بمصنف Naïve Bayes .

NLP in Social Media: An Overview

Soha S. Ibrahim[*1], Mostafa M. Aref [*2]

[*] *Department of Computer Science, Faculty of Computer Science and information System*
Cairo, Egypt

[1] *sohaelshafey@yahoo.com*

[2] *mostafa.aref@cis.asu.edu.eg*

***Abstract*— Recently, it becomes very easy to express one's opinion or read others' opinions through the internet, and it shows that online reviews or opinions has a significant influence on many areas (e.g.: purchasing product, elections, tourist visit, movies, financial market, public events). Current research trend refocused on the analysis of social media (forums, reviews, blogs, etc) in order to get a feel for what people think about current topics of interest for individuals or organizations. An automated system is intended to be developed that can identify and classify opinion or sentiment represented in an electronic text. In this paper, a survey is presented to present a variety of issues related to opinion mining from social media, and the challenges they impose on a Natural Language Processing (NLP) system.**

Keywords: opinion mining, sentiment analysis, NLP in social media.

1 INTRODUCTION

In this new information age, thoughts and opinions are shared so prolifically through online social networks. In order to make best use of this information, we need to be able to distinguish what is important and interesting. There are obvious benefits to companies, governments and so on in understanding what the public think about their products and services, but it is also in the interests of large public knowledge institutions to be able to collect, retrieve and preserve all the information related to certain events and their development over time. The spread of information through social networks can also trigger a chain of reactions to such situations and events which ultimately lead to administrative, political and societal changes.

Finding opinion sources and monitoring them on the Web, however, can still be a formidable task because a large number of diverse sources exist on the Web and each source also contains a huge volume of information. In many cases, opinions are hidden in long forum posts and blogs. It is very difficult for a human reader to find relevant sources, extract pertinent sentences, read them, summarize them and organize them into usable forms. An automated opinion mining and summarization system is thus needed. Opinion mining, also known as sentiment analysis [7].

Opinion mining can be defined as a sub-discipline of computational linguistics that focuses on extracting people's opinion from the web. The recent expansion of the web encourages users to contribute and express themselves via blogs, videos, social networking sites, etc. All these platforms provide a huge amount of valuable information that we are interested to analyse. The basic components of an opinion [7, 11]:

- Opinion holder: The person or organization that holds a specific opinion on a particular object.
- Object: on which an opinion is expressed
- Opinion: a view, attitude, or appraisal on an object from an opinion holder.

Sentiment analysis, on the other hand, is about determining the subjectivity (subjective or objective), polarity (positive or negative) and polarity strength (weakly positive, mildly positive, strongly positive, etc.) of a piece of text – (in other words what is the opinion of the writer)[7,11].

2 AN OVERVIEW

Especially in the last few years, a lot of research work has been done in the area of opinion mining and sentiment analysis. Research on opinion mining started with identifying opinion (or sentiment) bearing words, e.g., great, amazing, wonderful, bad, and poor. Many researchers have worked on mining such words and identifying their semantic orientations (i.e., positive or negative). Firstly, the authors identified several linguistic rules that can be exploited to identify opinion words and their semantic orientations from a large corpus [18].After that, sentiment detection techniques can be roughly divided into lexicon-based methods [17, 20, and 24] and machine-learning methods [3]. The lexicon-based methods rely on a sentiment lexicon, a collection of known and precompiled sentiment terms, and it has been enhanced by using small set of given seed opinion words to find their synonyms and antonyms in WordNet[19, 21, 10]. On the other hand, the machine learning approaches make use of syntactic and/or linguistic features [12, 15, 23, 25, 26, 4], and the hybrid approaches are very common, with sentiment lexicons playing a key role in the majority of methods. Also, the model of feature-based opinion mining and summarization is proposed in [19, 22, and 26]. This model gives a more complete formulation of the opinion mining problem. It identifies the key pieces of information that should be mined and describes how a structured opinion summary can be produced from unstructured texts.

3 SENTIMENT ANALYSIS LEVELS

Much research exists on sentiment analysis of user opinion data, which mainly judges the polarities of user reviews. In these research, sentiment analysis is often conducted at one of the three levels [3]. Firstly, the document level which assume each document (or review) focuses on a single object (not true in many discussion posts) and contains opinion from a single opinion holder. The task is sentiment classification of document if it is positive, negative, or neutral [2, 14]. Secondly, the sentence level where subjective/opinionated sentences are identified. Then, sentence sentiment classification is positive, negative or neutral [2, 14]. Finally, the feature (attribute) level which object features have been commented on by an opinion holder are identified and extracted. Then, determine whether the opinions on the features are positive, negative or neutral. Finally, feature synonyms are grouped and summarized [2, 14].

4 OPINION MINING AND SENTIMENT ANALYSIS

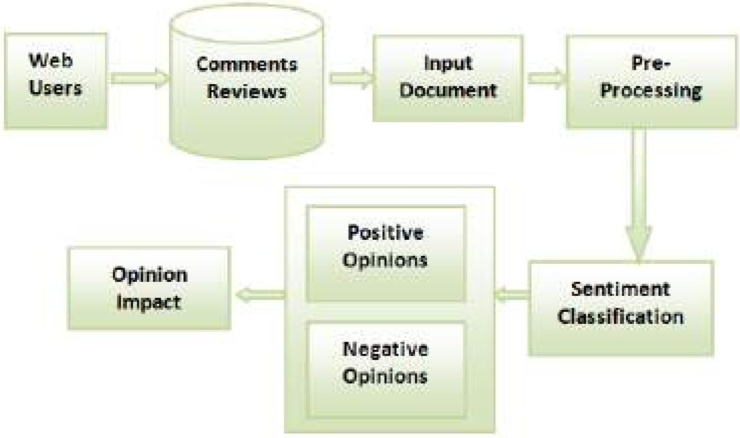


Figure1. Workflow of Opinion Mining [2]

Social network has become not only popular but also universally-acclaimed communication means that has thrived in making the world a global village, it has also given the users the privileges to give opinions with very little or no restrictions. Figure 1 has a workflow of opinion mining of how the opinions are being extracted from people review over their comment [2, 3, 7, 13, and 16]. ***Pre-processing*** In which raw data taken and is pre-processed for feature extraction. This phase includes sub-phases: tokenization, stop word removal, stemming, & case normalization [2]. ***Sentiment classification*** extracts features as a key step, and it must be too coarse by applying on both sentence level and document level, in order to determine precisely what users like or dislike. In order to address this problem, sentiment analysis aimed to extract opinion's product specific attributes from reviews [3]. ***Feature Extraction*** deals with feature types (which identifies the type of features used for opinion mining), feature selection (used to select good features for opinion classification), feature weighting mechanism (weights each feature for good recommendation) reduction mechanisms (features for optimizing the classification process) [2, 8]. Firstly, features types can be term frequency, term co-occurrence (features which occur together like unigram, bigram or n-gram), Part of speech tagger to separate POS tokens, opinion words (express positive (good) or negative (bad) emotions), negations which shifts sentiment orientation in a sentence, and syntactic dependency (a parse tree contains word dependency based features). Also, supervised and unsupervised pattern mining approaches can be applied to extract effective features types. Secondly, feature selection where good features are used for classification , popular selection techniques are information gain (presence and absence of a term in a document according to a threshold), odd ratio (binary class domain where it has one positive and one negative class for classification), and document frequency (number of appearances of a term in the available number of documents in the corpus and based on the threshold). Thirdly, features weighting mechanisms are term presence and term frequency (word which occurs occasionally contains more information than frequently occurring words), term frequency and inverse document frequency (TF-IDF), then documents are rated where highest rating is given for words that appear regularly in a few documents and lowest rating for words that appear regularly in every document [2]. Fourthly, feature reduction which reduces the feature vector size to optimize the performance of a classifier. Features vector reduction can be done in two different ways in which top n-features can be left in the vector and either removing low level or unwanted linguistic features [2]. In feature generation, adjectives are always important to impart inference from social media networks. They have been used most frequently as features amongst all parts of speech. A strong correlation between adjectives and subjectivity has been found. Although all the parts of speech are important people most commonly used adjectives to depict most of the sentiments and a high accuracy have been reported by all the works concentrating on only adjectives for feature generation. Adjective-Adverb Combination is another way to generate features, since most of the adverbs have no prior polarity. But when they occur with sentiment bearing adjectives, they can play a major role in determining the sentiment of a sentence. Adverbs alter the sentiment value of the adjective that they are used with. Adverbs of degree, on the basis of the extent to which they modify this sentiment value (e.g.: affirmation (certainly, totally), doubt (maybe, probably), etc.). Some of the positive adjectives are as follows dazzling, brilliant, phenomenal, excellent and fantastic. Negative adjectives: suck, terrible, awful, unwatchable, hideous [2, 9]. ***Opinion summarization*** finds what reviewers (opinion holders) liked and disliked (product features and opinions on the features), since the number of reviews on an object can be large. An opinion summary should be produced, a structured summary is desired to easy visualize and to compare [3,7].

5 KEY APPLICATIONS

The technology of opinion mining thus has a tremendous scope for practical applications. Opinions are so important that whenever one needs to make a decision [4,7, 16].

Individual consumers: If an individual wants to purchase a product, it is useful to see a summary of opinions of existing users so that he/she can make an informed decision. This is better than reading a large number of reviews to form a mental picture of the strengths and weaknesses of the product. He/she can also compare the summaries of opinions of competing products, which is even more useful.

Organizations and businesses: Opinion mining is equally, if not even more, important to businesses and organizations. For example, it is critical for a product manufacturer to know how consumers perceive its products and those of its competitors. This information is not only useful for marketing and product benchmarking but also useful for product design and product developments.

6 KEY CHALLENGES

Social media imposes a number of further challenges on an opinion mining system [1, 4, 5, 6, and 11]. They are:

A. *Relevance*

Not every comment on such pages will also be relevant to the topic or product discussed or displayed in article or review. This is a particular problem for social media, where discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand.

B. *Target Identification*

The topic of the retrieved document is not necessarily the object of the sentiment held therein. First the relevant entity must be identified and then look for opinions semantically related to this entity, rather than just trying to decide what the sentiment is without reference to a target.

C. *Contextual Information*

Social media, and in particular tweets, typically assume a much higher level of contextual and world knowledge by the reader than more formal texts. This information can be very difficult to acquire automatically. For example, in 2011, one tweet in the political dataset used likened a politician to Voldemort, a fictional character from the Harry Potter series of books. One advantage of tweets, in particular, is that they have a vast amount of metadata (e.g.: the date and time, the number of followers of the person tweeting, the person's location and even their profile) associated with them which can be useful, not just for opinion summarization and aggregation over a large number of tweets, but also for disambiguation and for training purposes.

D. *Volatility Over Time*

Social media exhibits a very strong temporal dynamic. Opinions can change radically over time, from positive to negative and vice versa.

E. *Opinion Aggregation and summarization*

In classical information extraction, aggregation can be applied to the extracted information in a straightforward way: data can be merged, if there are no inconsistencies, e.g. on the properties of an entity, opinions behave differently here.

F. *Spam and fake reviews*

The detection of spam, sarcastic and fake reviews, some different strategies are needed to deal with the linguistic issues imposed. For example, we incorporate detection of swear words, sarcasm, questions, conditional statements and so on, while our entity-centric approach focuses the opinions on specific topics and makes use of linguistic relations.

G. *Negation*

The simpler bag-of-words sentiment classifiers have the weakness that they do not handle negation well, the difference between the phrases “not good” and “good” is somewhat ignored, though they carry completely different meanings. A possible solution is to incorporate longer range features such as higher order dependency structures.

7 CONCLUSION

Monitoring social media to spot public opinion concerning different topics or events is presented. The opinions of individuals and groups are typically expressed as informal communications and are buried in the vast, and largely irrelevant, output of millions of bloggers and other online content producers. In this paper, an overview of related work presented in sentiment analysis and opinion mining in different forms of social media applications using different natural language processing techniques, opinion mining tasks, opinion mining workflow, and finally challenges in opinion to enhance sentiment analysis results in future research.

REFERENCES

[1]D. Maynard, K. Bontcheva, D. Rout., "Challenges in developing opinion mining tools for social media", In Proceedings of Workshop at LREC 2012, May 2012, Istanbul, Turkey.

[2]G.Angulakshmi, Dr.R.ManickaChezian, "An Analysis on Opinion Mining: Techniques and Tools" International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 3, Issue 7, July 2014.

[3] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 2, Issue 5, June 2012.

[4] Wilson, T., Wiebe, J. and Hwa, R., "Just How Mad Are You? Finding Strong and Weak Opinion Clauses", Proceedings of National Conference on Artificial Intelligence (AAAI'04), Montreal, Canada, August, 2004.

[5]Sudipta Roy, Sourish Dhar, Arnab Paul, Saprativa Bhattacharjee, Anirban Das, Deepjyoti Choudhury, "CURRENT TRENDS OF OPINION MINING AND SENTIMENT ANALYSIS IN SOCIAL NETWORKS ", International Journal of Research in Engineering and Technology (IJRET), Volume: 02, Special Issue: 02 , Dec-2013.

[6] David Osimo and Francesco Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis ",

[7] Bing Liu. "Opinion Mining", Invited contribution to Encyclopedia of Database Systems, 2008.

[8] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining ", 2012

[9] Nitish Bhardwaj, Anupam Shukla, Pradip Swarnakar, "Users' Sentiment Analysis in Social Media Context using NaturalLanguage Processing", The International Conference on Digital Information, Networking, and Wireless Communications (DINWC2014), 2014.

[10] WordNet 2.0, Available from: http://www.wordnet.princeton.edu/.

[11] Lars Kirchhoff, Katarina Stanoevska-Slabeva, Thomas Nicolai & Matthes Fleck,"Using social network analysis to enhance information retrieval systems",

[12] Ion SMEUREANU, Cristian BUCUR, "Applying Supervised Opinion Mining Techniques on Online User Reviews", Informatica Economică vol. 16, no. 2/2012.

[13] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications ", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009.

[14] Julia Kreutzer & Neele Witte, "Opinion Mining Using SentiWordNet", 2013

[15]Dave, D., Lawrence, A., and Pennock, D. Mining the Peanut Gallery, "Opinion Extraction and Semantic Classification of Product Reviews". Proceedings of International World Wide Web Conference (WWW'03), 2003.

[16] Bing Liu ,Lei Zhang, "A SURVEY OF OPINION MINING AND SENTIMENT ANALYSIS", chapter 1.

[17]Ding, X., Liu, B. and Yu, P, "A Holistic Lexicon-Based Approach to Opinion Mining", Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM'08), 2008.

[18] Hatzivassiloglou, V. and McKeown, K, "Predicting the Semantic Orientation of Adjectives" , ACLEACL'97, 1997.

[19]Hu, M and Liu, B, "Mining and Summarizing Customer Reviews", Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 2004.

[20] Kanayama, H. and Nasukawa, T. ,"Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis", Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06), 2006.

[21] Kim, S. and Hovy, E. ,"Determining the Sentiment of Opinions", Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), 2004.

[22] Liu, B., Hu, M. and Cheng, J. ,"Opinion Observer: Analyzing and Comparing Opinions on the Web", Proceedings of International World Wide Web Conference (WWW'05), 2005.

[23] Pang, B., Lee, L. and Vaithyanathan, S. ,"Thumbs up? Sentiment Classification Using Machine Learning Techniques", Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02), 2002.

[24] Popescu, A.-M. and Etzioni, O. ,"Extracting Product Features and Opinions from Reviews", Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP'05), 2005.

[25] Turney, P. ,"Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". ACL'02, 2002.

[26] Wiebe, J. and Riloff, E. ,"Creating Subjective and Objective Sentence Classifiers from Unannotated Texts", Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'05), 2005.

Bibliography

Soha S. Ibrahim is a researcher in department of computer science in faculty of computer science and information system. She received master's degree (M.Sc) of computer science in 2011 from faculty of computer science and information system in Ain Shams University. Her research focuses in Text knowledge representation.

Mostafa M. Aref is a professor of Computer Science and Vice Dean for Society and the Environment, Ain Shams University, Cairo, Egypt. Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. of Electrical Engineering - Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.

معالجة اللغة الطبيعية في وسائل الاعلام الاجتماعية: نظرة عامة

سها سعيد إبراهيم[1] ، مصطفى محمود عارف[2]
[1،2] *قسم علوم الحاسب, جامعة عين شمس*
القاهرة - جمهورية مصر العربية
[1] sohaelshafey@yahoo.com
[2] mostafa.aref@cis.asu.edu.eg

ملخص

في الآونة الأخيرة، يصبح من السهل جدا التعبير عن رأي شخص أو قراءة آراء الآخرين من خلال شبكة الإنترنت، وذلك يدل على أن الانتقادات على الانترنت أو الآراء لديها تأثير كبير على العديد من المجالات (على سبيل المثال: شراء منتج، والانتخابات، زيارة سياحية والأفلام و السوق المالية، والمناسبات العامة). اتجهت البحوث الحالية الى التركيز على تحليل وسائل الاعلام الاجتماعية (على سبيل المثال: المنتديات) من أجل التوصل الى ما يعتقده الناس حول المواضيع الراهنة ذات الاهتمام الاكبر بالنسبة للأفراد أو المنظمات. ويهدف النظام الآلي الى تحديد وتصنيف الاراء أو الشعور الممثل في النصوص الإلكترونية. في هذا البحث، يقدم نظرة عامة على الابحاث المتنوعة المتعلقة بالتنقيب عن الرأي من وسائل الاعلام الاجتماعية، والتحديات التي تفرضها على نظام المعالجة الآلية للغات الطبيعية.

كيفَ نبني مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًّا للُّغة العربيَّة بطريقة نصف آليَّة؟

المُعتزّ بالله السَّعيد
كُلِّيَّة دار العُلُوم، جامعة القاهرة، مصر
moataz@cu.edu.eg

المستخلص – تَسعى هذه الدِّراسةُ إلى تقديم منهجيَّةٍ لبناء مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًّا للُّغة العربيَّة بطريقة نصف آليَّة، وتهدفُ الدِّراسةُ إلى إيجاد رؤية واضحة لآليَّات بناء مُدوَّنة لُغويَّة مُوسَّمة تركيبيًّا، تُراعي طبيعة اللُّغة العربيَّة ونظامها الكتابيّ بحيثُ يُمكنُ توظيفُها مُستقبلًا في بناء أدوات التَّحليل التَّركيبيّ للُّغة العربيَّة والدِّراسات النَّحويَّة بصورةٍ عامَّة.وتقوم الفكرة الرئيسة للدِّراسة على توظيف تقنيات النَّحو العدديّ والكشَّاف السِّياقيّ في تحديد القرائن الدَّالَّة على أقسام الكلام العربيّ. وتأتي الدِّراسةُ في خمسة محاورَ أساسيَّة تتضمَّن مُقدِّمةً ثُمَّ عرضًا لإشكالات الدِّراسة. ويلي ذلك تقديمُ المنهجيَّة المُقترَحة لبناء المُدوَّنة اللُّغويَّة المُوسَّمة. وأخيرًا يستعرضُ الباحثُ نتائجَ الدِّراسة، ويعرضُ خُلاصةَ بحثه.

الكلمات المفاتيح
المُدوَّنات اللُّغويَّة – التَّوسيم – التَّحليل التَّركيبيّ – اللُّغة العربيَّة

1 المقدمة

1.1 في ماهيَّةالمُدَوَّنات اللُّغَويَّة.

تُعنى لسانيَّاتُ المُدَوَّنة *Corpus Linguistics* بالبحث في الظَّواهِر اللُّغوية وتفسيرِها من خلال مجموعةٍ مِن النُّصُوص التي تُمَثِّلُ الواقِعَ اللُّغَويّ. وأداةُ البحث في هذا المنهجِ هي "المُدَوَّنة اللُّغَويَّة *Linguistic Corpus*" باعتبارِها مجموعةً مِن نُصُوص اللُّغة المكتوبة أو المنطوقة الَّتي يمكن التَّعاملُ معها آليًّا والتَّحَكُّمُ في بياناتها ومُدخَلاتها بالإضافة أو الحذف أو التعديل من خلال قواعدِ بياناتٍ صُمِّمَت لتكونَ قادرةً على التَّعامُل مع هذه النُّصوص، حيثُ تُمَثِّلُ هذه القواعدُ مخزنًا كبيرًا للُّغة، يُرجَع إليه وقتَ الحاجة ويَتحمَّل أيَّ قدرٍ من النُّصوص الَّتي يُمكن أن تُضافَ إلى المادَّة الأساسيَّة للمُدَوَّنة اللُّغَويَّة مُستَقبلاً [1]. وتُعتبَرُ المُدوَّناتُ اللُّغَويَّةُ موردًا لُغويًّا رئيسًا يُستفادُ منهُ في مختلف ميادين حَوسبة اللُّغة *Computational Linguistics* ومُعالَجة اللُّغات الطَّبيعيَّة *Natural Language Processing (NLP)*، حيثُ يتمُّ تدريبُ نُصُوص المُدوَّنات وتحليلها آليًّا وإحصائيًّا بهدف إنتاج أدواتٍ لمُعالجة اللُّغة على كافَّة مُستوياتها.

1.2 توسيم المُدَوَّنات اللُّغَويَّة *Corpora Annotation*.

يُعنى التَّوسيمُ *Annotation* بتوصيف الوحدات اللُّغَويَّة للنُّصُوص، سواءٌ أكانت مقاطع أم كلمات أم غير ذلك. وهو بهذا إجراءٌ يُساعدُ على تحويل النُّصُوص من صُورتها الأوَّليَّة إلى صُورةٍ يسهُلُ التَّعامُلُ معها آليًّا. وبعبارةٍ أخرى، فإنَّ التَّوسيمَ هو العمليَّة الَّتي تتحوَّلُ من خلالِها المُدَوَّنات اللُّغَويَّة الخام *Raw Corpora* إلى مُدوَّناتٍ مُوَسَّمة *Annotated Corpora*. وتُعنى هذه الدِّراسةُ بالتَّوسيم التَّركيبيّ *Syntactic Annotation* الَّذي يُستفادُ من وُجوده في المُدَوَّنات اللُّغَويَّة العربيَّة في بناء المُحلِّلات التَّركيبيَّة وحَصر أنماط الجملة العربيَّة والإحصاء اللُّغَويّ، بالإضافة إلى ميادين التَّرجمة الآليَّة. ويتمُّ التَّوسيمُ التَّركيبيُّ باستخدام ما يُعرَفُ بـ "رُموز أقسام الكلام *Parts of Speech Tags*"، حيثُ يُرفَقُ رَمزٌ مُعَيَّنٌ بكلِّ قسمٍ من أقسام الكلام على حِدة، وفقًا لطبيعة النِّظام التَّركيبيّ للُّغة، وفي ضوء الهدف المنشُود من المُدَوَّنة اللُّغَويَّة المُوَسَّمة.

1.3 مناهج التَّحليل التَّركيبيّ للُّغة العربيَّة *Arabic Syntactic Analysis*.

يُمثِّلُ التَّحليلُ التَّركيبيُّ *Syntactic Analysis* مُستوى وسيطًا بينَ مُستويات التَّحليل اللُّغَويّ الَّتي تبدأ بمُستوى التَّحليل الصَّوتيّ وتنتهي بمُستوى التَّحليل الدِّلاليّ. وينبثقُ هذا المُستوى عن علم التَّركيب *Syntax* الَّذي يُعرَفُ كذلك بعلم نظام تركيب الجُمَل؛ وهو العلم الَّذي يدرس مُكوِّنات الجُملة والعلاقات بين عناصِرها، ويُعنى بدراسة أنواع الجُمَل وقواعدُ الإعراب وكيفيَّة التَّأليف بين أقسام الكلام لتكوين جُملةٍ مُفيدةٍ ومُنتظمةٍ وفقًا لقوانين النِّظام اللُّغَويّ، كما يُعنَى بتحليل الوحدات المُكَوِّنة للتَّركيب النَّحويّ. ولأنَّ وَحدة التَّحليل التَّركيبيّ هي الجُملة *Sentence*، فقد دَعَت الحاجةُ إلى توظيف الحاسُوب في تحليل عناصرِها من خلال أدوات التَّحليل التَّركيبيّ للنُّصُوص. وثمَّةَ ثلاثةُ مناهجَ أساسيَّة يعتمدُ عليها الباحثُونَ في بناء أدوات التَّحليل التَّركيبيّ المُوافقة لطبيعة اللُّغة العربيَّة، حيثُ يعتمدُ المنهجُ الأوَّلُ على المعطيات اللُّغَويَّة المستمدَّة من قواعد النَّحو العَرَبيّ، سواءٌ أكانت هذه القواعدُ في صُورةِ قوالبَ أم قواعدِ بيانات؛ ويعتمدُ المنهجُ الثَّاني على خوارزميَّة التَّحليل التَّركيبيّ الَّتي تُمثِّلُ صُورةً رياضيَّةً لقواعد النَّحو العَرَبيّ. أمَّا المنهجُ الثَّالثُ، فيقومُ على استخلاص قواعد النِّظام التَّركيبيّ من المدوَّنات اللُّغَويَّة العَربيَّة باعتبارها تمثيلًا لواقع اللُّغة المستخدَمة فعليًّا، حيثُ يتمُّ تدريبُ نُصُوص المُدَوَّنات بهدف استخلاص الأنماط التَّركيبيَّة، ثُمَّ تهيئة الآلة لاستقبال النَّتائج والتَّفاعل معها. وهذا المنهجُ الثَّالثُ هو الأكثرُ نجاعةً ومُناسَبةً للُّغة العربيَّة – مِن وُجهة نظر الباحث، لأسبابٍ أهمُّها أنَّهُ يُراعي تعدُّديَّة أنماط الجُملة العربيَّة، ويُراعي القواعدَ السَّماعيَّة التَّركيبيَّة الَّتي يصعُبُ التَّعبيرُ عنها بلُغة الآلة. ولأجل هذا، فإنَّ هذه الدِّراسةَ تسعى إلى الإبانة عن آليَّات بناء مُدَوَّنةٍ لُغَويَّةٍ مُوسَّمةٍ تركيبيًّا للُّغة العربيَّة، حيثُ يُشكِّلُ وُجُودُ مثلِ هذه المُدَوَّنة نقطةَ الانطلاق إلى تطوير أدواتٍ فعَّالةٍ للتَّحليل التَّركيبيّ العَربيّ.

2 إشكالات بناء شبكة للكلمات العربيَّة.

يستغرقُ بناءُ المُدَوَّنات اللُّغَويَّة المُوَسَّمة تركيبيًّا لُّغة العربيَّة وقتًا وجهدًا كبيرَين، الأمرُ الَّذي يُؤدِّي إلى زيادة تكلفة إنتاج هذا النَّوع من المُدَوَّنات. أضِف إلى ذلكَ أنَّ بناءَ المُدوَّنات المُوسَّمة يستدعي زيادةَ الموارد البشريَّة العاملة، لاسيَّما إذا تعلَّقَ الأمرُ بمُدَوَّناتٍ لُغَويَّةٍ كبيرةٍ نسبيًّا. وبالنَّظَر إلى طبيعة اللُّغة العربيَّة من ناحية، وواقع صناعة المُدوَّنات اللُّغَويَّة من ناحيةٍ أخرى، نستطيعُ أن نقفَ على ثلاثة إشكالاتٍ رئيسةٍ، نعرضُها فيما يلي.

1. المُرُونة في نظام بناء الجُملة العَرَبيَّة.

يتمتَّعُ نظامُ بناء الجُملة العَرَبيَّة بقدرٍ كبيرٍ من المُرُونة؛ حيثُ يسمحُ بالتَّقديم والتَّأخير بينَ عناصر الجُملة، كما يسمحُ بتعدُّد أنماط الجُملة وتمدُّد عناصرِها الَّتي قد تتجاوزُ أربعينَ عُنصرًا. ومن ناحيةٍ أخرى، يسمحُ نظامُ الجُملة العَرَبيَّة بتبادُل العناصرِ التَّالية لقسم الكلام المُحَدَّد. نُلاحظُ مثلًا أنَّ الضَّميرَ المُنفصِلَ الثَّابتَ في محلِّه الإعرابيّ يقبلُ أن يلحقَ به الاسم، نحو (أنتَ مُجتَهد)، ويقبلُ أن يلحقَ به الفِعل، نحو (أنتَ تجتهد)، ويقبلُ أن تلحقَ به الأداة، نحو (أنتَ لا تجتَهد) ... وهكذا. وتُمثِّلُ هذه المُرُونةُ إشكالًا عندَ توسيم المُدَوَّنات اللُّغَويَّة تركيبيًّا، لأنَّها تستدعي عملًا يدويًّا شاقًّا للبحث عن قِسم الكلام الَّذي يتبعُهُ كُلُّ عُنصُرٍ من عناصرِ الجُملةِ على حِدة. وحالَ التَّدخُّل الآليّ لتوسيم المُدَوَّنة، فإنَّ نسبةَ الخطأ لن تكونَ هيِّنة. وهذا يستدعي تدخُّلًا يدويًّا كبيرًا لمُعالجة الأخطاء النَّاجمة عن عمل الآلة.

2. طبيعة النِّظام الكتابيّ [الجرافيميّ] لُّغة العربيَّة.

اللُّغةُ العربيَّةُ لُغةٌ اشتقاقيَّة، يسمحُ نظامُها الكتابيُّ بأن تتشابَكَ فيها الوحداتُ الكتابيَّةُ [الجرافيمات *Graphemes*] بينَ جرافيمات الكلمة أو مجموعة الكلمات، على النَّحو الَّذي نجدُهُ مثلًا في المجموع الكتابيّ (فَسَيَكفيكَهُم) الَّتي تتكوَّنُ من خمس وحداتٍ صرفيَّة [مورفيمات *Morphemes*]، هي على التَّرتيب: (الفاء) و (السِّين) و (يكفي) و (الكاف) و (هم). ولكلِّ وحدةٍ من هذه الوحدات دلالةٌ تركيبيَّةٌ تجعلُها قسمًا مُستقلًّا من أقسام الكلام، حيثُ تدلُّ الفاءُ على الاستئناف، وتدلُّ السِّينُ على التَّسويف، ويدلُّ الفعلُ على المُضارعة والاستمراريَّة، ويدلُّ الضَّمير (الكافُ) على المُخاطَب المُفرَد المُذكَّر، ويدلُّ الضَّمير (هم) على الغائب الجَمع المُذكَّر. ومن ناحيةٍ أخرى، فإنَّ بعضَ أقسام الكلام تتماثلُ في رَسمها الكتابيّ مع اختلاف مبناها، على نحوِ ما نجدُ في الكلمات (من، بل، هل)؛ حيثُ تحتملُ كُلٌّ منها أن تكونَ اسمًا أو فعلًا أو أداةً، بحسب ضبطِها. ووفقًا لهذا النِّظام، فإنَّ توسيم المُدوَّنات اللُّغويَّة تركيبيًّا يفرضُ الجمعَ بينَ بعض أقسام الكلام المُتشابكة، كما يستدعي ضبطَ النُّصُوص بالشَّكل تحسُّبًا للالتباس المُحتَمَلِ وُقُوعُه عندَ توسيم الكلمات المُتماثلة في رَسمها.

3. الاختلاف حَولَ أقسام الكلام العَرَبيّ *Arabic PoS*.

تتكوَّنُ الجُملةُ العَربيَّة من مجموعةٍ من العناصر الَّتي تُعرَفُ بـ "أقسام الكلام *Parts of Speech (PoS)*". وقد صَنَّفَ النُّحاةُ القُدماءُ الكلامَ العَرَبيَّ إلى ثلاثة أقسامٍ، هي: الاسم *Noun* والفِعل *Verb* والأداة (الحَرف) *Particle*. ويَحيدُ بعضُ اللُّغويِّينَ المُعاصِرينَ عن هذا التَّصنيف، فيذهبُ فريقٌ إلى تقسيم الكلام العَرَبيّ إلى أربعة أقسامٍ [2]، هي: الاسم والفِعل والحَرف والضَّمير *Pronoun*؛ ويذهبُ فريقٌ آخَرُ إلى تقسيم الكلام العَرَبيّ إلى سبعة أقسامٍ [3]، هي: الاسم والصِّفة *Adjective* والفِعل والضَّمير والخالفة والظَّرف *Adverb* والأداة. وعلى جانبٍ آخَرَ يلجأ العاملُونَ في حَوسَبَة اللُّغة إلى ابتكار تصنيفاتٍ أخرى في محاولةٍ لتمكين الآلة من التَّعامُل مع قواعد النَّحو العَرَبيّ، على النَّحو الَّذي نجدُهُ في موارد "مُؤسَّسة البيانات اللُّغويَّة *Linguistic Data Consortium (LDC)*" [4]. ويُمثِّلُ هذا الاختلافُ إشكالًا عندَ توسيم المُدوَّنات اللُّغَويَّة تركيبيًّا، حيثُ يَستدعي تحديدَ الهدف من المُدوَّنة اللُّغَويَّة المُوَسَّمة، ثُمَّ بناء المُدَوَّنة وفقَ ما يُحقِّقُ هذا الهدف، كما يُقصِرُ الإفادةَ من المُدَوَّنات اللُّغَويَّة المُوَسَّمة على جوانبَ معلومةٍ سَلَفًا دونَ غَيرِها.

3 منهجيَّة بناء مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًّا لُّغة العربيَّة بطريقة نصف آليَّة.

قبلَ الشُّرُوع في بناء المُدَوَّنة اللُّغَويَّة المنشُودة، ينبغي أن نُحدِّدَ الهدفَ منها، لأنَّ حجمَ المُدَوَّنة وطبيعة النُّصُوص الَّتي تحويها يخضعان لذلك الهدف. والواقعُ أنَّ ما ننشُدُهُ في دراستِنا هو مُدَوَّنةٌ لُغويَّةٌ يُستفادُ منها في أغراض التَّحليل التَّركيبيّ لُّغة العربيَّة المُستخدَمة فعليًّا. ويعني هذا ضرورةَ اشتمال المُدَوَّنةِ المنشُودةِ على نُصُوص اللُّغة العربيَّة المُعاصرة من ناحية، وضرورةِ تنوُّع المادَّة المُتَضَمَّنة لتكونَ تمثيلًا حقيقيًّا لواقع اللُّغة العربيَّة من ناحيةٍ أخرى. وفي ضوء ذلك، ستعرضُ الدِّراسةُ فيما يلي لمنهجيَّة بناء المُدَوَّنة اللُّغَويَّة المنشودة عبرَ مجموعةٍ من المراحل المُتعاقبة.

3.1 بناء المُدوَّنة اللُّغَويَّة الخام *Raw Corpus*.

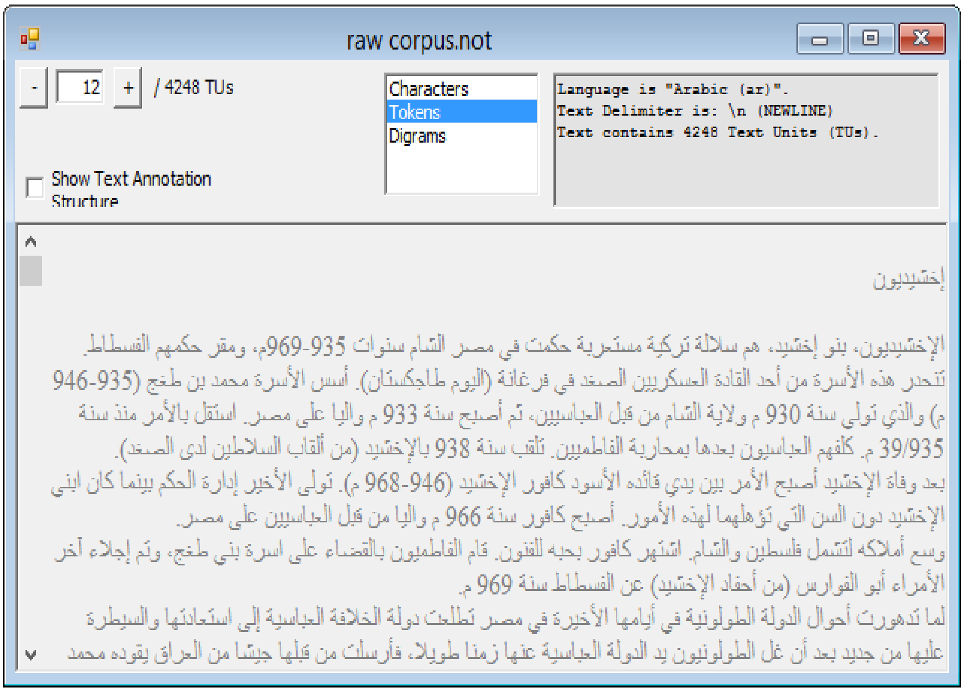
ثمَّةَ ثلاثُ وسائل لبناء المُدَوَّنات اللُّغَويَّة بصُورةٍ عامَّة، حيثُ تقومُ الوسيلةُ الأولى على "أسلُوب الحصر الشَّامل *Comprehensive Inventory Method*"، وتقومُ الوسيلةُ الثَّانية على "الاستبانة *Questionnaire*"، وتقومُ الوسيلةُ الثَّالثة على "نظريَّة العَيِّنات الإحصائيَّة *Statistical Sampling Theory*". وهذه الأخيرةُ هي الأكثرُ مُناسبةً لطبيعة دراستِنا، لأنَّها مَرِنةٌ بالقَدر الَّذي يُساعدُ على تمثيل لُغة المُجتمَع. وتحقيقًا للهدف من الدِّراسة، فقد صنَعَ الباحثُ مُدَوَّنةً لُغَويَّةً مُمثِّلةً للعربيَّة المُعاصِرة في صُورة عَيِّنةٍ قصديَّة [غَرَضيَّة] *Purposive Sampling* مُنتقاةٍ من الموسوعة الحُرَّة "ويكيبيديا *Wikipedia*" [5]. واختارَ الباحثُ هذه الموسوعةَ مصدرًا لمادَّة المُدَوَّنة لسبَبَين رَئِيسَين، هما:

- توافُر مِعيارَي المُعاصَرة والتَّنوُّع في مادَّتِها؛ حيثُ تتميَّزُ الموسوعةُ – في نُسختها العربيَّة – بحداثة النُّصُوص وتمثيلها لُّغة العربيَّة المُعاصرة؛ إذ بدأ العملُ في تحريرها رسميًّا في يوليو من عام 2003 على أيدي آلاف المُتطوِّعينَ من أبناء اللُّغة؛ وتتميَّزُ بتنوُّع مادَّتِها كونَها تُحَرَّرُ بطريقةٍ موسوعيَّةٍ تُراعي التَّفاوُتَ المعرفيَّ للقُرَّاء. وتتنوَّعُ المادَّةُ في الموسُوعة لتُغطِّي عشرةَ حُقُولٍ معرفيَّةٍ رئيسةٍ، هي: (الثَّقافة، والأعلام والتَّراجم، والجُغرافيا، والتَّاريخ، والرِّياضيَّات، والعُلُوم، والمُجتَمَع، والتِّقنيات، والفلسفة، والأديان). وتَتَّسعُ هذه المادَّةُ لتشملَ أكثرَ من 380 ألف مقالة، تشتملُ في مجموعها على عشرات الملايين من الكلمات [إحصاء 2015].
- أنَّها تخضعُ لرُخصة "جنو" للوثائق الحُرَّة *GNU Free Documentation License* [6]؛ الأمرُ الَّذي يُتيحُ استخدامَها لأغراض البحث العلميّ دونَ قُيُود؛ إذ إنَّها تُعامَلُ باعتبارها ملكيَّةً عامَّة، يُسمَحُ بالإفادة منها للجميع ما داموا يلتزمونَ بذكرها مصدرًا للمعلومات المُستَمَدَّة منها [7].

وقد مرَّت مرحلةُ بناء المُدَوَّنة اللُّغَويَّة الخام بأربع مراحلَ فرعيَّة، على النَّحو الآتي:

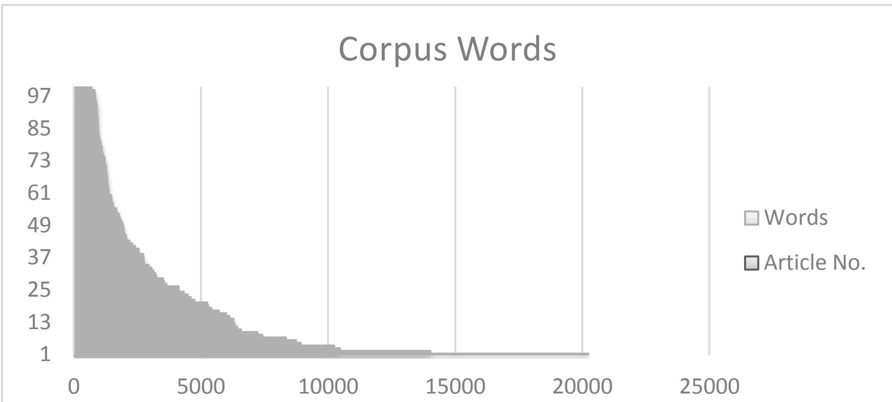
1. جُمِعَت المقالاتُ المُنتقاة، ووُزِّعَت في وثائقَ، بحيثُ تحتوي كُلُّ وثيقةٍ على مقالةٍ واحدة.
2. رُوجِعَت نُصُوصُ المُدوَّنةِ إملائيًّا بهدف تقليل نسبة أخطاء التَّحرير فيها. واستعانَ الباحثُ في ذلكَ بأدوات التَّدقيق الإملائيّ والفهرسة الآليَّة المُساعدة على كشف أكثر الأخطاء ترَدُّدًا في النُّصُوص.
3. قامَ الباحثُ بتحرير النُّصُوص المُتَضَمَّنة في وثائق المُدَوَّنة في صِيغةٍ قياسيَّةٍ مُوَحَّدةٍ تسمحُ بالتَّحكُّم فيها آليًّا، لتيسير عمليَّات المُعالجة الآليَّة والإحصائيَّة لاحقًا. وقامَ الباحثُ خلالَ ذلكَ بتحويل الوثائق من لُغَة تَوصِيف النُّصُوص التَّشعُّبيَّة [صيغة صفحات الويب] *Hyper Text Markup Language (HTML)* إلى لُغة التَّوصِيف القابلة للامتِداد *Extensible Markup Language (XML)*، ثُمَّ إلى الصِّيغة النَّصِّيَّة *TXT*. وقامَ الباحثُ بتشفير النُّصُوص بصيغة تشفير المحارف العربيَّة *CP-1256*.
4. قامَ الباحثُ بتنقية نُصُوص المُدَوَّنة من الرُّموز الَّتي قد تُعيقُ عمليَّةَ المُعالجة وتُؤثِّرُ على النَّتائج، كما قامَ بتنقية النُّصُوص من الكشائد [الزَّوائد] وعلامات الضَّبط لتوحيد رَسم المباني [الكلمات] المُتطابقة.

أمَّا عن توصيف المُدَوَّنة اللُّغَويَّة، فقد جُمِعَت من مِئة مقالةٍ مُتنوِّعة، وبلغَ مجموعُ عدد كلماتها (308550) كلمة. أما عددُ الكلمات الفريدة *Unique Words* فقد بلغَ (49048) كلمة [قبلَ التَّنقية]، و (48971) كلمة [بعدَ التَّنقية]. ويُوضِّحُ (الشَّكل 1) نموذجَ المُدَوَّنة اللُّغَويَّة في صُورتِها الخام، قبلَ توسيمِها.



الشكل رقم (1)
نموذج المُدوَّنة اللُّغَويَّة في صُورتها الخام – مِنَصَّة *Nooj*

وقد تبايَنَت أحجامُ الوثائق بحسب عدد الكلمات الَّتي تحويها؛ إذ اشتملت أكبر الوثائق على 20225 كلمة، واشتملَت أصغرُها على 696 كلمة. ويُوضِّحُ (الشَّكل 2) مُخطَّطًا بيانيًّا شريطيًّا بأعداد كلمات وثائق المُدَوَّنة اللُّغويَّة – مادَّة الدِّراسة – بعدَ ترتيبها تنازُليًّا.



الشكل رقم (2)

مُخَطَّط بيانيّ شريطيّ بأعداد الكلمات في وثائق المُدَوَّنة اللُّغويَّة

3.2 تَعيين رُموز أقسام الكلام *PoS Tags*.

يَمُرُّ التَّوسيمُ التَّركيبيّ للعربيَّة بمرحَلَتَين رَئيسَتَين، حيثُ تُعنى المرحلةُ الأولى بتعيين أقسام الكلام، وتُعنى المرحلةُ الأخرى بالإعراب *Parsing*. ونظرًا لطبيعة التَّركيب العَرَبيّ الَّذي يقومُ على بناءٍ شجريٍّ لا بناء خَطِّيّ، فإنَّ هذه الدِّراسةَ تُركِّزُ على المرحلة الأولى؛ إذ هي المرحلةُ الَّتي يُمكنُ إخضاعُ الآلة لفهمها. ذلكَ أنَّ المرحلةَ الأخرى [الإعراب] تستدعي توصيفًا دقيقًا للموقع الإعرابيِّ الَّذي تشغلُهُ كُلُّ كلمةٍ على حِدة؛ وهو أمرٌ يصعبُ إدراكُهُ عبرَ الآلة.

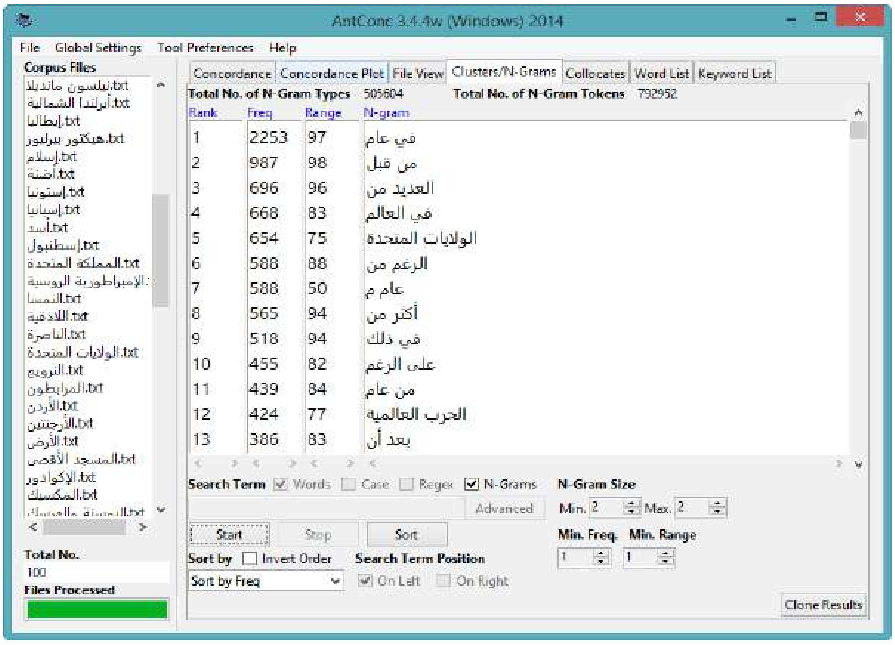
ولمَّا كانَ الهدفُ من الدِّراسة إيجادَ وسيلةٍ لإخضاع الآلة لفهم قواعد النَّحو العَرَبيّ، كانَ لزامًا أن نخرُجَ عن الأطُر التَّقليديَّة الَّتي وَضَعَها النُّحاةُ لأقسام الكلام إلى إطارٍ يُمَكِّنُ الآلةَ من استيعاب هذه الأقسام. وعليهِ فإنَّ الدِّراسةَ تقترحُ تقسيمَ الكلام إلى خمسة أقسامٍ رئيسة، يتفرَّعُ عنها خمسة عشر قسمًا فرعيًّا على النَّحو الوارد في (الجدول 1)، مع مُلاحظة أنَّ هذا التَّقسيمَ يَضُمُّ الصِّفات إلى الأسماء، ويضُمُّ الخوالِفَ إلى الأفعال، كما يُخالفُ ما جرى عليهِ النُّحاةُ بشأن الكلمات الدَّالَّة على الاستفهام، حيثُ يُوزِّعُونَها بينَ الأدوات (نحو: الهمزة، هل) والأسماء (نحو: أينَ، متى، كيفَ،...). ومنهجُ الدِّراسة أن تُوضَعَ هذه الكلماتُ ضمنَ الأدوات لدلالتها جميعًا على الاستفهام من ناحية، وجواز إحلالِ بعضِها مكانَ بعضٍ من ناحيةٍ أخرى.

الجدول رقم (1)
مُقتَرَح التَّقسيم الخُماسيّ للكلام العربيّ ورُمُوز الأقسام *Pos Tags* المُستخدَمة في التَّوسيم التَّركيبيّ

م	قسم الكلام	المُصطلح الإنجليزيّ	الرَّمز *PoS Tag*
		الاسم *Noun*	
1	الاسم الشّائع	*Common noun*	[*CN*]
2	اسم العَلَم	*Proper Noun*	[*PN*]
3	اسم الإشارة	*Determiner*	[*DE*]
4	الاسم الموصول	*Relative Pronoun*	[*RP*]
5	العدد/الرَّقم	*Cardinal Number*	[*CNU*]
		الفِعل *Verb*	
6	فِعل مُضارِع	*Imperfect Verb*	[*VI*]
7	فِعل ماضٍ	*Perfect Verb*	[*VP*]
8	فعل طلب (أمر)	*Request Verb*	[*VR*]
		الأداة *Particle*	
9	أداة استفهام	*Question*	[*QU*]
10	أداة استثناء	*Exception*	[*EX*]
11	أداة رَبط	*Conjunction*	[*CO*]
12	حَرف جرّ	*Preposition*	[*PRE*]
13	أداة أخرى	*Other Particle*	[*PO*]
		الضَّمير *Pronoun*	
14	الضَّمير	*Pronoun*	[*PRO*]
		الظَّرف *Adverb*	
15	الظَّرف	*Adverb*	[*AD*]

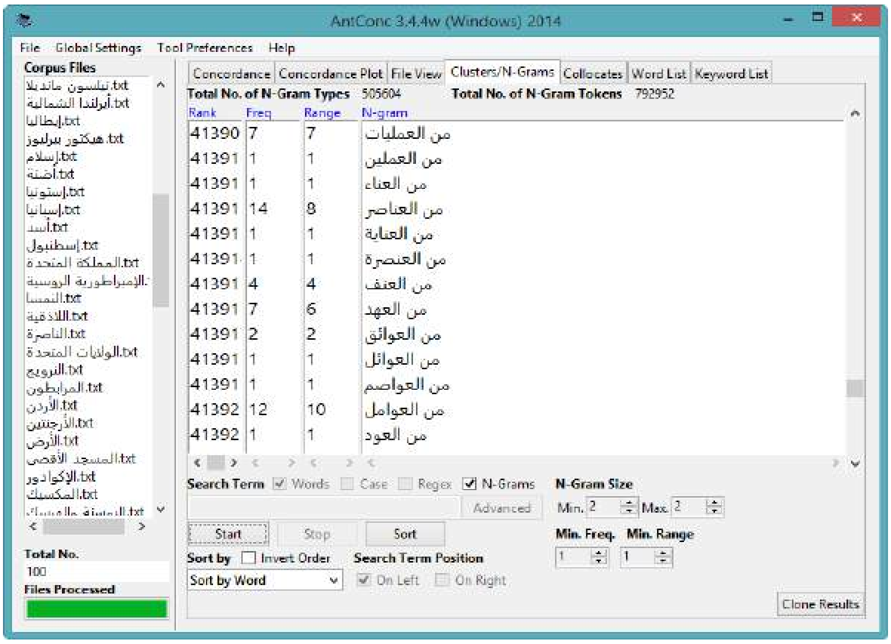
3.3 التَّوسيم التَّركيبيّ باستخدام تقنيات النَّحو العدديّ *N-Gram techniques*.

يُساعدُ النَّحوُ العدديُّ *N-Gram* في إحصاء ترَدُّدات الوحدات الكتابيَّة الكبرى [الكلمات ومُتسَلسِلات الكلمات]، الأمرُ الَّذي يُمكنُ مَعَهُ توسيمُ أعدادٍ هائلةٍ من الكلمات آليًّا [8، 9، 10]، دونَ الحاجة إلى الوُقُوفِ على كُلٍّ منها على حِدة. وتحقيقًا للهدف المنشُود، تقترحُ الدِّراسةُ ترتيبَ الوحدات الكتابيَّة الكُبرى بحسب ترَدُّداتِها أوَّلًا، على النَّحو الوارد في (الشَّكل 3) بهدف توسيم أكبر عددٍ مُمكن من الكلمات، ثُمَّ ترتيبَ هذه الوحدات ألفبائيًّا، على النَّحو الوارد في (الشَّكل 4) بهدف إزالة الالتباس الحادث في الكلمات الَّتي تتَّفقُ في رَسمِها وتختلفُ في قسم الكلام الَّذي تتبعُه.



الشكل رقم (3)

ترتيب الوحدات الكتابيّة الكبرى باستخدام تقنيات النّحو العدديّ بحسب تردُداتها – برمجيّة *AntConc 3.4*



الشكل رقم (4)

ترتيب الوحدات الكتابيّة الكبرى باستخدام تقنيات النّحو العدديّ ألفبائيًّا – برمجيّة *AntConc 3.4*

وتوضيحًا لآليَّة الإفادة من تقنيات النَّحو العدديّ في التَّوسيم التَّركيبيّ، تقترحُ الدِّراسةُ أن يكونَ التَّوسيمُ في بدايتِهِ على مُستوى النَّحو الأحاديّ *Uni-gram*، حيثُ يُمكنُ من خلالِهِ توسيمُ الكلمات المُتردِّدة بصُورةٍ كبيرة، لاسيَّما الكلمات الوظيفيَّة *Function words*. وبتطبيق ذلكَ على المُدَوَّنة اللُّغَويَّة – موضُوع الدِّراسة – نجدُ أنَّ أكثرَ الكلمات تتبعُ قسمًا واحدًا من أقسام الكلام؛ لكنَّنا سنجدُ بعضَ الكلمات تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميٍّ، مثل (من) الَّتي تحتملُ أن تكونَ حرفَ الجرِّ (مِن) أو الاسم الموصول (مَن)، وتحتملُ في حالاتٍ أقلَّ أن تكونَ الفعلَ الماضي (مَنَّ) أو الاسمَ (مَنّ).

ويُوَضِّحُ (الجدول 2) التَّوسيم التَّركيبيّ للكلمات الأكثر تردُّدًا في المُدوَّنة اللُّغَويَّة بعدَ استخلاصِها باستخدام النَّحو الأحاديّ.

الجدول رقم (2)
التَّوسيم التَّركيبيّ للكلمات الأكثر تردُّدًا في المُدوَّنة اللُّغَويَّة (النَّحو الأحاديّ)

م	الكلمة	التَّردُّد	التَّوسيم التَّركيبيّ
1	في	11849	في[PRE]
2	من	9069	من[PRE] من[RP] من[VP] من[CN]
3	على	4602	على[PRE]
4	إلى	3621	إلى[PRE]
5	أن	2606	أن[PO]
6	عام	1983	عام[AD] عام[VP] عام[CN]
7	التي	1781	التي[RP]
8	عن	1371	عن[PRE]
9	بعد	1141	بعد[AD] بعد[VP] بعد[CN]
10	مع	1127	مع[AD]
11	بين	1114	بين[AD] بين[VP] بين[CN]
12	كان	1106	كانت[VP]
13	أو	1094	أو[CO]
14	الذي	1085	الذي[RP]
15	هذه	1051	هذه[DE]
16	ما	1041	ما[RP] ما[PO]
17	و	922	و[CO]
18	ذلك	907	ذلك[DE]
19	هذا	859	هذا[DE]
20	وقد	849	وقد[PO-CO]
21	م	780	م[PO]
22	كانت	751	كانت[VP]
23	سنة	741	سنة[AD] سنة[CN]
24	كما	724	كما[PO]
25	حيث	715	حيث[AD]

وحتَّى نتمكَّنَ من توسيم الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميٍّ، تقترحُ الدِّراسةُ الانتقالَ إلى التَّوسيمِ على مُستوى النَّحو الثُّنائيّ *Bi-gram* ثُمَّ النَّحو الثُّلاثيّ *Tri-gram* ...، وهكذا، إلى أن تقلَّ احتمالاتُ تعدُّد الأقسام الكلاميَّة للكلمة الواحدة.

نُلاحظُ مثلًا عندَ توسيم المُدَوَّنة اللُّغويَّة على مُستوى النَّحو الثُّنائيّ أنَّ الكلمات المُلازمة لكلمة (من) تُقلِّلُ من احتمالات تعدُّد أقسام الكلام بصورةٍ كبيرة. ومع هذا تبقى احتماليَّةُ تعدُّد الأقسام في بعض السِّياقات، كما في الثُّنائيَّات (أكثر من، عدد من، كُلّ من، ...)، وهو أمرٌ يُمكنُ معالجتُهُ باستخدام النَّحو العدديّ الثُّلاثيّ.

ويُوَضِّحُ (الجدول 3) التَّوسيم التَّركيبيّ لثُنائيَّات الكلمات الأكثر تردُّدًا في المُدوَّنة اللُّغَويَّة بعدَ استخلاصِها باستخدام النَّحو الثُّنائيّ.

الجدول رقم (3)
التَّوسيم التَّركيبيّ لثُنائيَّات الكلمات الأكثر تردُّدًا في المُدوَّنة اللُّغويَّة (النَّحو الثُّنائيّ)

م	الكلمة	التَّردُّد	التَّوسيم التَّركيبيّ
1	في عام	506	[PRE]في [AD]عام
2	الولايات المتحدة	295	[CN]الولايات [CN]المتحدة
3	من قبل	269	[PRE]من [AD]قبل
4	العديد من	250	[CN]العديد [PRE]من
5	في العالم	207	[PRE]في [CN]العالم
6	أكثر من	196	[CN]أكثر [PRE]من ؛ [CN]أكثر [RP]من
7	إلا أن	183	[EX]إلا [PO]أن
8	بعد أن	178	[AD]بعد [PO]أن
9	عدد من	163	[CN]عدد [PRE]من ؛ [CN]عدد [RP]من
10	من خلال	161	[PRE]من [AD]خلال
11	بالإضافة إلى	150	[CN-PRE]بالإضافة [PRE]إلى
12	الدولة العثمانية	147	[CN]الدولة [CN]العثمانية
13	الحرب العالمية	141	[CN]الحرب [CN]العالمية
14	في ذلك	135	[PRE]في [DE]ذلك
15	في حين	127	[PRE]في [AD]حين
16	كل من	126	[CN]كل [PRE]من ؛ [CN]كل [RP]من
17	الرغم من	124	[CN]الرغم [PRE]من
18	إلى أن	118	[PRE]إلى [PO]أن
19	عن طريق	113	[PRE]عن [CN]طريق
20	وفي عام	108	[PRE-CO]وفي [AD]عام
21	في مصر	108	[PRE]في [PN]مصر
22	عمر المختار	105	[PN]عمر [PN]المختار
23	التي كانت	105	[RP]التي [VP]كانت
24	في القرن	102	[PRE]في [AD]القرن
25	من عام	100	[PRE]من [AD]عام

وبتطبيق تقنيات النَّحو العدديّ على كلمات المُدوَّنة، نلمسُ نتيجةً حقيقيَّة عندَ توسيم الكلمات المُتردِّدة الَّتي لا تحتملُ أكثرَ من قسمٍ كلاميٍّ، سواءٌ على مُستوى النَّحو الأحاديّ أم النَّحو الثُّنائيّ. ومع هذا، يبقى إشكالُ توسيم الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميٍّ قائمًا، إذ ينبغي أن نتحقَّقَ من قسم الكلامِ الصَّحيح لكلِّ سياقٍ على حِدة. وسعيًا إلى مُعالجة هذا الإشكال تقترحُ الدِّراسةُ إعادةَ ترتيب ثُنائيَّات الكلمات ألفبائيًّا، ثُمَّ بناء خوارزميَّة التَّوسيم آليًّا بالنَّظر إلى سوابقِ الكلمات المُلازمة للكلمة الَّتي ننشُدُ توسيمَها. وعلى سبيل المثال، سنُلاحظُ أنَّ كلمة (من) تنتمي إلى قسم الكلام [حرف الجرّ] حينَ تلحقُ بها سابقة التَّعريف (الـ)، وتنتمي إلى قسم الكلام (الاسم الموصول) حينَ تلحقُ بها سابقة المُضارَعة (يتـ)، وهكذا.

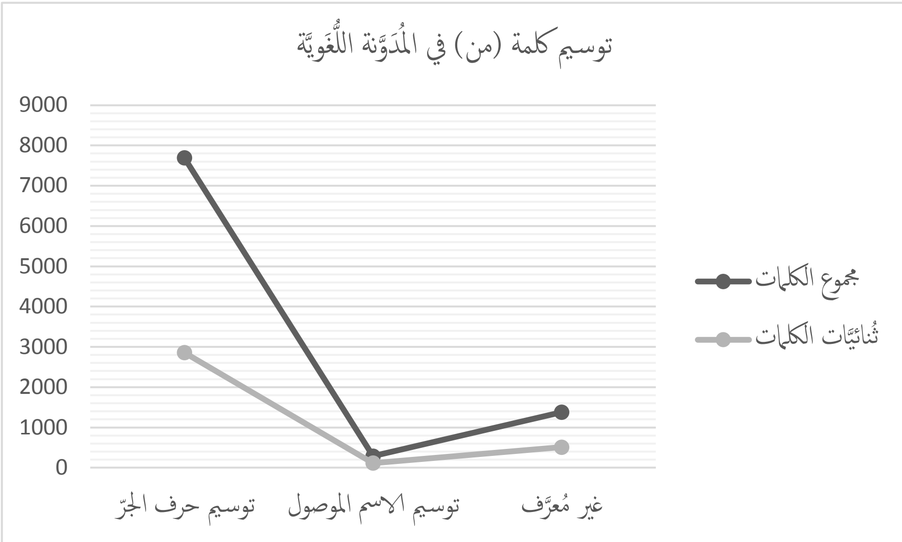
ويُوَضِّحُ (الجدول 4) نموذجًا لتوسيم كلمة (من) باعتبار سابقة الكلمة المُلازمة لها في المُدَوَّنة اللُّغويَّة موضوعِ الدِّراسة.

الجدول رقم (3)
نموذج التَّوسيم التَّركيبيّ لكلمة (من) باعتبار سابقة الكلمة المُلازمة في المُدَوَّنة اللُّغويَّة

م	الكلمة	التَّردُّد	التَّوسيم التَّركيبيّ
سابقة الكلمة المُلازمة (الـ)			
1	من الوصول	5	[PRE]من [CN]الوصول
2	من الوصي	1	[PRE]من [CN]الوصي
3	من الوضع	1	[PRE]من [CN]الوضع
4	من الوطن	1	[PRE]من [CN]الوطن
سابقة الكلمة المُلازمة (يتـ)			
5	من يتتبع	1	[RP]من [VI]يتتبع
6	من يتعاون	1	[RP]من [VI]يتعاون

7	من يتقلدها	1	من [*RP*][*VI*]يتقلدها
8	من يتلقى	1	من [*RP*][*VI*]يتلقى

لقد ورَدت كلمة (من) في المُدَوَّنة اللُّغويَّة في 9069 سياق [تتوزَّعُ على 3479 ثُنائيَّة]. وقامَ الباحثُ بالتَّوسيم التَّركيبيّ للكلمة آليًّا عبرَ تقنيات النَّحو العَدَديّ، فكانت النَّتيجةُ أن أمكنَ التَّعرُّفُ على قسم الكلام الَّذي تتبعُهُ الكلمة في 85٪ من السِّياقات، منها 82٪ تنتمي إلى قسم الكلام [حرف الجرّ]، بواقع 7692 سياق [تتوزَّعُ على 2857 ثُنائيَّة]، و 3٪ تنتمي إلى قسم الكلام (الاسم الموصول)، بواقع 285 سياق [تتوزَّعُ على 115 ثُنائيَّة]. وفي مُقابل ذلك لم تسمح تقنياتُ النَّحو العدديّ بالتَّعرُّف على 15٪ من السِّياقات، بواقع 1377 سياق [تتوزَّعُ على 507 ثُنائيَّة]، على النَّحو الوارد في (الشَّكل 5)؛ وهُوَ ما يعني إمكانيَّةَ إخضاعِ الآلة لتوسيم 85٪ من السِّياقات الَّتي وَرَدت فيها كلمة (من)، على أن يتمَّ توسيمُ النِّسبةِ المُتبقِّية يدويًّا؛ وقِس على ذلكَ مجموعةَ الكلماتِ الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميّ.

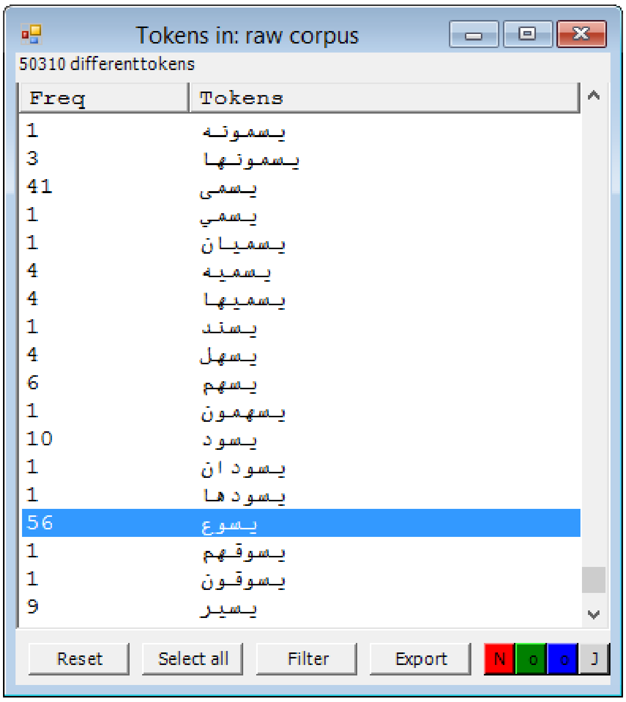


الشكل رقم (5)
مُخَطَّط بيانيّ خطِّيّ بنتائج التَّوسيم التَّركيبيّ لكلمة (من) في المُدَوَّنة اللُّغويَّة باستخدام تقنيات النَّحو العدديّ

3.4 التَّوسيم التَّركيبيّ باستخدام الكشَّاف السِّياقيّ *Concordancer*.

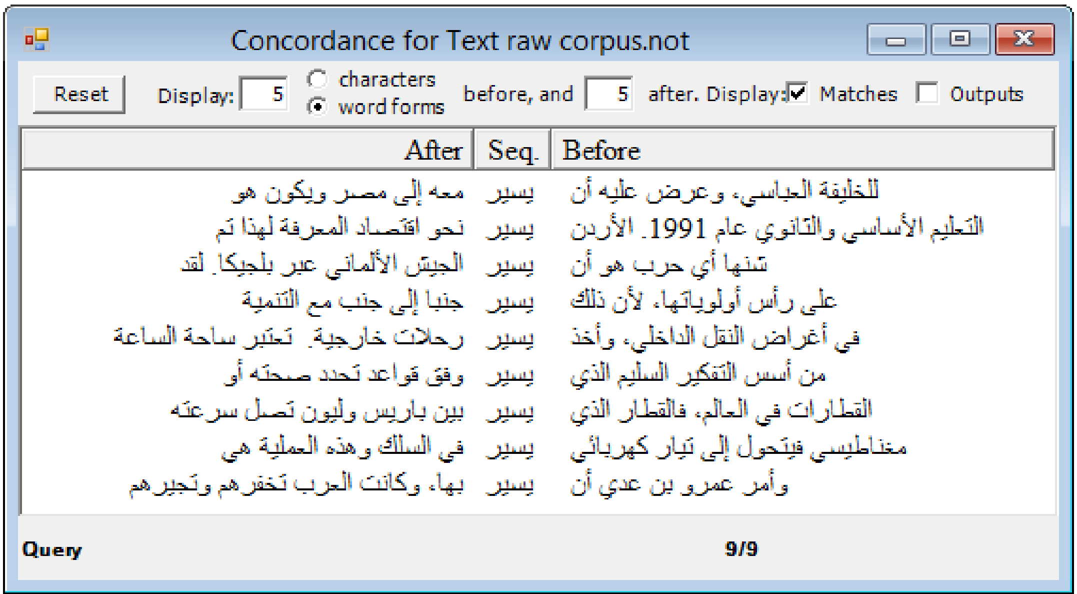
تظهرُ الفائدةُ الحقيقيَّةُ لتقنيات النَّحو العَدَديّ في التَّوسيم التَّركيبيّ في الكلمات الأكثر تردُّدًا في المُدَوَّنة اللُّغويَّة، لكنَّها قد لا تكونُ مُجديةً بصورةٍ كبيرة في الكلمات الأقلّ تردُّدًا. ولهذا، تقترحُ الدِّراسةُ إكمالَ عمل تقنيات النَّحو العَدديّ باستخدام الكشَّاف السِّياقيّ *Concordancer*؛ حيثُ يُتيحُ تعقُّبَ الوحداتِ الكتابيَّة [الجرافيمات *Graphemes*] الَّتي تُلزمُ الكلمةَ قسمًا كلاميًّا مُعيَّنًا باستخدام المُفهرس الآليّ للنُّصُوص *Text Indexer* [القائم أساسًا على النَّحو العدديّ]، كما تُساعدُ في مُراجعة المُطابقة بينَ كلمات المُدوَّنة اللُّغويَّة وأقسام الكلام الَّتي تُوسَمُ بها إذا احتمَلَت الكلمةُ التَّوسيمَ بأكثر من قسمٍ كلاميّ، من خلال الكشف عن سياقات كُلِّ كلمةٍ على حِدة.

وعلى سبيل المثال، يُلزمُ الجرافيمان المُتتاليان (ي، س) في بداية الكلمة قسمًا كلاميًّا هُو (الفعل المُضارع). وباستخدام الكشَّاف السِّياقيّ، نستطيعُ الكشفَ عن كلمة (يسوع) الَّتي خالَفَت القاعدة لتتبع القسمَ الكلاميّ (اسم العَلَم) على النَّحو المُوضَّح في (الشَّكل 6)، ونستطيعُ الكشفَ عن سياقات كلمة (يسير) الَّتي تحتملُ أن تكونَ اسمًا أو فعلًا، على النَّحو المُوضَّح في (الشَّكل 7).



الشكل رقم (6)

نموذج كلمات المدوّنة اللّغويّة مفهرسة آليًّا – برمجيّة *Nooj Concordance 3.2*



رقم **الشكل (7)**

الكشّاف السّياقيّ لكلمات المدوّنة اللّغويّة– برمجيّة *Nooj Concordance 3.2*

4 نتائج الدِّراسة.

1. أبانَت الدِّراسةُ عن ماهيَّة المُدوَّنات اللُّغَويَّة ومفهوم التَّوسيم التَّركيبيّ، كما أبانَت عن ثلاثة مناهج للتَّحليل التَّركيبيّ للُّغة العربيَّة؛ حيثُ يعتمدُ المنهجُ الأوَّلُ على المعطيات اللُّغَويَّة المستمدَّة من قواعد النَّحو العَرَبيّ؛ ويعتمدُ المنهجُ الثَّاني على خوارزميَّة التَّحليل التَّركيبيّ الَّتي تُمثِّلُ صُورةً رياضيَّةً لقواعد النَّحو العَرَبيّ؛ ويقومُ المنهجُ الثَّالثُ على استخلاص قواعد النِّظام التَّركيبيّ من المدوَّنات اللُّغَويَّة العَربيَّة باعتبارها تمثيلًا لواقع اللُّغة.
2. أبانَت الدِّراسةُ عن إشكالات بناء مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًا للُّغة العربيَّة؛ وتمثَّلَت هذه الإشكالاتُ في: المُرُونة في نظام بناء الجُملة العَرَبيَّة، وطبيعة النِّظام الكتابيّ للُّغة العربيَّة، والاختلاف حَولَ أقسام الكلام العَرَبيّ.
3. اقترَحَت الدِّراسةُ منهجيَّةً لبناء مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًّا للُّغة العربيَّة بطريقة نصف آليَّة عبرَ أربع خُطواتٍ رئيسة، تبدأ ببناء المُدوَّنة اللُّغويَّة الخام بحيثُ يتوافرُ فيها معيارا المُعاصَرة والتَّنوُّع، ومرُورًا بتعيين أقسام الكلام بما يتوافقُ مع الهدف المنشود، ثُمَّ التَّوسيم التَّركيبيّ باستخدام تقنيات النَّحو العدديّ [الأحاديّ، والثُّنائيّ، والثُّلاثيّ، ...]، وانتهاءً بالتَّوسيم التَّركيبيّ باستخدام الكشَّاف السِّياقيّ.
4. اقترَحَت الدِّراسةُتقسيمَ الكلام العربيّ إلى خمسة أقسامٍ رئيسة، هي: الاسم (ويتفرَّعُ عنه: الاسم الشَّائع، واسم العلم، واسم الإشارة، والاسم الموصول، والعدد)، والفعل (ويتفرَّعُ عنه: الفعل المُضارع، والفعل الماضي، وفعل الطَّلَب)، والأداة (ويتفرَّعُ عنها: أداة الاستفهام، وأداة الاستثناء، وأداة الرَّبط، وحرف الجرّ، وأداةٌ أخرى)، والضَّمير، والظَّرف.
5. أبانَت الدِّراسةُ عن إمكانيَّة توظيف النَّحو العدديّ في توسيم الكلمات الأكثر تردُّدًا في المُدوَّنات اللُّغَويَّة، واقترَحت حلولًا لتوسيم الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميّ، كما أبانَت الدِّراسةُ عن جدوى توظيف الكشَّافات السِّياقيَّة في التَّوسيم التَّركيبيّ للكلمات الأقلّ تردُّدًا، وأبانَت كذلك عن إمكانيَّة الإفادة من الكشَّافات السِّياقيَّة في مُراجعة المُطابقة بينَ كلمات المُدوَّنة وأقسام الكلام الَّتي تُوسَمُ بها إذا احتمَلَت الكلمةُ التَّوسيمَ بأكثر من قسمٍ كلاميّ.
6. قامَ الباحثُ بتطبيق منهجيَّتِهِ على مُدوَّنةٍ لُغويَّةٍ مُستمدَّة من الموسوعة الحُرَّة (ويكيبيديا)، تشتملُ على 48971 كلمة فريدة. وانتهى إلى إمكانيَّة توسيم 92٪ من جُملة كلمات المُدوَّنة آليًّا، في حين تستدعي النِّسبةُ المُتبقِّية التَّدخُّلَ اليدويّ.

5 الخُلاصة.

يستغرقُ بناءُ المُدَوَّنات اللُّغَويَّة المصنوعة لأغراض التَّحليل التَّركيبيّ في العربيَّة وقتًا وجهدًا كبيرَين، الأمرُ الَّذي يُؤدِّي إلى زيادة تكلفة بناء هذا النَّوع من المُدَوَّنات. ووُقوفًا على طبيعة اللُّغة العربيَّة وحاجتها إلى هذا النَّوع من المُدَوَّنات اللُّغويَّة بهدف توظيفها في تطوير أدوات التَّحليل التَّركيبيّ للنُّصوص، فإنَّ هذه الدِّراسةَ تَسعى إلى تقديم منهجيَّة لبناء مُدَوَّنة لُغَويَّة مُوَسَّمة تركيبيًّا للُّغة العربيَّة بطريقة نصف آليَّة. وينطلقُ الباحثُ في دراستِهِ من تقنيات النَّحو العَدديّ *N-Gram*الَّتي تسمحُ بإحصاء وترتيب الكلمات ومُتسلسلات الكلمات وفقَ نَسَقٍ يُساعدُ على إيجاد القرائن الدَّالَّة على البنيَة التَّركيبيَّة، كما يستخدمُ الكشَّافات السِّياقيَّة الَّتي تُساعدُ على تعقُّب الوحداتِ الكتابيَّة الَّتي تُلزمُ الكلمةَ قسمًا كلاميًّا مُعيَّنًا، كما تُساعدُ في مُراجعة المُطابقة بينَ كلمات المُدوَّنة وأقسام الكلام، من خلال الكشف عن سياقات كُلِّ كلمةٍ على حِدة. ومن ناحيةٍ أخرى، يسعى الباحثُ إلى ضبط منهجيَّتِه باستخدام القواعد القياسيَّة للنَّحو العَرَبيّ على مُستوى أقسام الكلام بما يضمنُ بناءَ المُدَوَّنة في صُورةٍ تُحقِّقُ الإفادةَ القُصوى من مادَّتِها.

قائمة المراجع

[1] السَّعيد (المُعتزّ بالله): المدوَّنات اللُّغَويَّة، ضِمنَ كتاب (مُقدِّمة في حوسَبة اللُّغة العربيَّة)، مجموعة من المؤلِّفين، تحرير: مُحسن رَشوان، والمعتزّ بالله السَّعيد، قَيد النَّشر بمدينة الملك عبد العزيز للعُلُوم والتِّقنية، الرِّياض، 2015م.

[2] أنيس (إبراهِيم): مِن أسرار اللُّغة، مكتبة الأنجلو المِصْريَّة، القاهِرة، ط4، 1972م، ص 279.

[3] حَسَّان (تَمَّام): اللُّغة العَرَبيَّة "مَعناها وَمَبناها"، الهَيئة المِصريَّة العامَّة للكِتاب، القاهِرة، ط2، 1979م، ص 86.

[4] الموقع الإلكترونيّ للمُؤسَّسة:

https://www.ldc.upenn.edu/

[5] الموقع الإلكترونيّ لموسوعة "ويكيبيديا" في نُسختها العربيَّة:

http://ar.wikipedia.org.

[6] صيغة الرُّخصة في إصدارها الأخير (نوفمبر 2008) عبرَ الموقع الإلكترونيّ:

http://www.gnu.org

[7] السَّعيد (المُعتزّ بالله): نحوَ مُعجمٍ للُّغة العربيَّة للنَّاطِقينَ بغيرِها "مُعالجة حاسُوبيَّة إحصائيَّة"، مجلَّة "التَّواصُل اللِّسانيّ" – المجلَّة الدَّوليَّة لهندسة اللُّغة العربيَّة واللِّسانيَّات العامَّة، *LINGUISTICA COMMUNICATIO (International journal of Arabic Language Engineering & General Linguistics)*، فاس، المغرب، المُجلَّد 18، 2015.
عُمَر (أحمد مُختار): مُعجم اللُّغة العربيَّة المُعاصرة، عالَم الكُتُب، القاهِرة، 2008.

[8] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman and Slav Petrov. (2012). Syntactic annotations for the Google Books Ngram Corpus. ACL '12 Proceedings of the ACL 2012 System Demonstrations Pages 169-174.

[9] Xiaofei Lu. (2014). Computational Methods for Corpus Annotation and Analysis. Springer. P. 151.

[10] S. Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, David Yarowsky. (2013). Natural Language Processing Using Very Large Corpora. Springer Science & Business Media.P. 102.

السيرة الذاتية

المُعتزّ بالله السَّعيد

يَعمل مُدَرِّسًا بقسم علم اللُّغة والدِّراسات السَّامِيَّة والشَّرقِيَّة في كُلِّيَّة دار العُلُوم بجامعةالقاهرة، وخبيرًا لُغويًّا حاسُوبيًّا بالمركز العربيّ للأبحاث في الدَّوحة. حصلَ من جامعة القاهِرة على درجة الدُّكتوراه في علم اللُّغة والدِّراسات السَّاميَّة والشَّرقِيَّة. نَشَرَ أكثرَ من عشرين ورقة بحثِيَّة في دوريَّاتٍ علميَّةٍ ومؤتمراتٍ دوليَّة، بالإضافة إلى ستَّة كُتُبٍ في المُعجَمِيَّة العَرَبِيَّة والدِّراسات اللُّغَوِيَّة المُعاصِرة؛ منها: علم اللُّغة والتِّقنيات المُعاصرة، علم الدِّلالة ونظريَّة المعنى، مُقَدِّمة في حَوسَبة اللُّغة العربيَّة "بالاشتراك"؛ كما ساهمَ في أكثر من عشرة مشرُوعاتٍ بحثيَّةٍ دوليَّةٍ تُعنى بحوسَبة اللُّغة العربيَّة وتقنياتها، منها: مشرُوع مُعجم الدَّوحة التَّاريخيّ للُّغة العربيَّة، ومشرُوع بناء شبكة دلاليَّة للُّغة العربيَّة *SEWAC*، ومشرُوع بناء نظام الخليل للتَّحليل الصَّرفيّ للُّغة العربيَّة. شاركَ في الإشراف على عددٍ من الأطروحات العلميَّة في مصرَ وإسبانيا؛ وهو عُضوٌ بهيئة تحرير المجلَّة الدَّوليَّة للتَّقدُّم والبحث العلميّ*IJSPR*والمجلَّة الدَّوليَّة لهندسة اللُّغة العربيَّة واللِّسانيَّات العامَّة *LINGUISTICA COMMUNICATIO*والمجلَّة الدَّوليَّة لعُلُوم وهندسة الحاسب *IJCSEA* وعُضوٌ باللِّجان العلميَّة لعدد من المُؤتمرات الدَّوليَّة، منها: ورشة عمل المعالجة الآليَّة للُّغة العربيَّة *WANLP 2015*، 2015، بكين – الصِّين، والمُؤتمر الدَّوليّالثَّالث للتَّطبيقات الإسلاميَّة في عُلُوم الحاسُوب وتقنياته*IMAN 2015*، 2015، قونية – تركيا، والنَّدوة الدَّوليَّة الخامسة حول المعالَجة الآليَّة للُّغة العربيَّة *CITALA'14*، 2014م، وجدة – المغرب، وورشة عمل المعالجة الآليَّة للُّغة العربيَّة *EMNLP 2014*، 2014م، الدَّوحة، قَطَر. حصلَ في عام 2012 على جائزة المُنَظَّمة العربيَّة للتَّربية والثَّقافة والعُلُوم (ألكسو *ALECSO*) للإبداع والابتكار التِّقَنِيّ [المركز الأوَّل] في مَيدان "المَعلُوماتيَّة والمُعالَجة الآليَّة للُّغة العربيَّة".

Discourse Tagging of Political Speeches: A Corpus-based Study

Marwa Adel Abu El Wafa[*1], Sameh Alansary[**2], Shadia El Soussi[***3]

[*]*Language and Translation Department, College of Language and Communication, Institute for Language Studies*
Arab Academy for Science, Technology and Maritime Transport
Miami, Alexandria, Egypt
[1]marwa.adel.abuelwafa@aast.edu

[**]Phonetic and Linguistics Department, Faculty of Arts, University of Alexandria
ElShatby, Alexandria, Egypt
Bibliotheca Alexandrina, Alexandria, Egypt
[2]sameh.alansary@bibalex.org

[***]*Institute of Applied Linguistics, Faculty of Arts, University of Alexandria*
ElShatby, Alexandria, Egypt
[3]shadia.elsoussi@bibalex.org

Abstract— **This paper discusses the creation of a tag set on the discourse level through tagging various rhetorical devices employed by both the American President Barack Obama in seven of his speeches and the African American leader Martin Luther King in seven of his speeches. This is done on the path of discourse tagging as a means of creating a discourse-based tag set of the devices and annotated corpus of political speeches. This tag set is meant to be fed into a concordance program namely MonoConc Pro 2.2. Once the speeches are manually annotated by the researcher, the tagged speeches are then analyzed by the concordance program searching for and counting the frequencies of the devices. The results help draw conclusions about the style of each character as well as the similarities and differences between each. This study might open the way for creating a discourse based corpus that can be used by other researchers experimenting in the same field.**

1 INTRODUCTION

The study is not a critical discourse analysis study, but rather a corpus-based study. This corpus-based study aims at building an annotated corpus of political speeches. The annotations are tags created by the researcher. This is from where the originality of the present study stems. To the researcher’s knowledge, there are no studies on annotating a corpus of political speeches on the discourse level neither in English nor in Arabic. The created tag set is totally original as it is created by the researcher for the purpose of analyzing the present corpus. The study aims at creating an annotated corpus on the discourse level through analyzing speeches searching for the rhetorical devices used by the politicians to bond with their audience and evoke their emotions. Two speakers are selected for this study; the African American leader Martin Luther King and the American President Barack Obama. Through the rhetorical devices, politicians convince their audience with a certain frame work or view points or certain perspectives. In the present study the term rhetorical devices is the one used to refer to the devices which in other studies may be referred to as stylistic devices. The political speeches chosen serve as the corpus of the study.

2 STUDY SIGNIFICANCE

The originality of this study stems from the endeavor of creating an output of annotated corpus on the discourse level. A corpus of this nature has never been available before to help researchers portray the style of different writers or speakers. As a matter of fact, the shortage of annotated data for linguistic and language engineering research was a motive behind conducting this study. An annotated corpus is rich with linguistic data which can open the door to multiple linguistic and language engineering researches whose results open gates for language users in general and reveals secrets about language. This research could be one of the fewest resources in discourse tagging as there are few endeavors to tag on the discourse level. Moreover, the study aims at creating a discourse- based tag set. This tag set stands for the selected rhetorical devices for the study. This can enable other researchers to analyze texts in terms of the language used with a press of a button bringing out numbers of used rhetorical devices. Through annotating the texts using the assigned tags, the researcher can arrive at clear numbers of the occurrences of the devices. Consequently, conclusions can be drawn to identify and describe the style of the different writers or speakers whether belonging to the political field or any other field.

3 DEFINING RHETORIC

Rhetoric has been defined as the art of speaking or writing effectively as a means of communication or persuasion. It is also a skill in the effective and creative use of speech and the use of language. Rhetoric is a tool that is used to enrich language in order to persuade, inform, express ideas and entertain. It is no surprise that the skill of persuasion is often in evidence with great politicians or religious leaders throughout history. Using rhetoric and its devices, a writer or speaker is capable of invading audiences' minds and changing or guiding their perspective. Rhetoric gives the power to communicate diverse messages through the use of powerful imagery or referring to reputable figures thus evoking emotions and creating the bond needed with the audience. Persuasion, although is present as an aim of any use of language, is viewed as one of four aims of using rhetoric. Informing is the second aim. Using rhetoric to inform may not appear as powerful as when it is used to persuade. Informing is clear in cases as teaching. A teacher uses the tools of rhetoric to bring ideas closer to the learners.

In rhetoric, a rhetorical device is any of the techniques that an author or speaker uses to convey to the audience a meaning with the goal of persuading him or her towards considering a topic or a number of topics or an ideology different from or similar to his or her from a different perspective. Not only do rhetorical devices evoke an emotional response in the audience and consequently bond them with their politicians, but also the main goal behind using them is to persuade the audience towards a particular frame of view, view point or a particular course of action. In this sense, appropriate rhetorical devices are used to shape the language that is designed both to make the audience receptive through emotional changes and to provide a rational argument for the frame of view, view point or course of action.

4 THE RHETORICAL DEVICES AND THEIR CATEGORIES

The selection of the devices was done in a very cautious manner. They are grouped according to their function into four classes. Each group or class encompasses devices that are employed for a certain purpose and a certain effect. The devices that belong to the first category are known to be used to present a strategy or point of view. The second group includes devices that give depth to the argument through stressing the ideas in a certain manner. The third group embraces devices that are used to organize the ideas. The fourth group includes devices that give a distinctive style to the writing. The four categories are presented as they conventionally appear in the literature. Devices that share the same or similar effect or purpose are grouped together under the same category. Neither the devices nor the categories are presented in priority order. Therefore, they could be alternating.

5 TAGGING THE DEVICES

Once the devices are selected, the phase of designing the tag set starts. These tags play the role of codes that stand for each device of the thirty five devices selected. These tags are used in the analysis stage and are annotated into the corpus selected. The tags assigned to the parts of speech are either one capital letter or three capital letters. For instance, verbs take the tag V, nouns N, prepositions P, adjectives ADJ and determiners DET. The second pattern is used by the researcher for the rhetorical devices chosen in the study. Three capital letters that resemble the device's pronunciation are given to each device as a tag (Bird & Liberman, 1999[1]). Table I displays all the devices included in the study, the meaning of each and their tags.

A. Annotation

In the present study the term 'annotation' is used to refer to the process of adding interpretative linguistic information to the corpus (Bird &Liberman, 1999). Any act of corpus annotation is, by definition also an act of interpretation, either of the structure of the text or of its content. An unannotated corpus is simply a raw text where linguistic information and linguistic phenomena are hidden. On the other hand, an annotated corpus transforms texts into banks of linguistic information available for investigation and analysis. Annotating a corpus helps make the retrieval and extraction of linguistic information and the study of linguistic phenomena easier and faster thus enabling researchers to arrive at findings that would not have been feasible without the presence of an annotated corpus. Annotated corpora make up reusable resources for many researchers with multiple purposes. Hence, a linguistic database is available for analyses and studies can be compared and contrasted adding richness to the field. There are many levels of corpus annotations starting with the phonological moving to the morphological, then the lexical and finally the highest level which is the discourse level.

B. Leech's Annotation Maxims

The linguistic information that is added to a corpus is governed by the seven maxims of Leech. According to Leech, there should be flexibility in dealing with the annotated corpora. In other words, after annotation there should still be the possibility of recovering the corpus to its raw state. If the first maxim is the head of the coin, then Leech's second maxim is actually its tail. The first and the second maxims accentuate that on one hand the corpus can be regressed to its raw state without the annotations and on the other hand the annotations themselves can be solely extracted from the corpus. The first two maxims are put in such manner so as to ensure maximum flexibility for the manipulation of the corpus by the user. This totally applies to the corpus in the present study. In other words, the tags can be removed from the corpus and it can appear in its raw state once more. This is because the tags are not inserted into words and so removing them would destroy the words, but rather surround extracts. The third maxim is concerned with the end user and so stresses on the availability of clear guidelines for the annotation scheme adopted by the researcher. For this reason, a clear description of all the chosen rhetorical devices and their corresponding tags is given to ensure that other users can benefit from the present study in future research. The fourth maxim confirms that it should be made obvious how and by whom the annotation was performed. In the present study the corpus is manually annotated by the researcher. Manual annotation is one of the types of annotation which is highly valued for its accuracy.

Table I
The Rhetorical Devices and the Tag Set

Device	**Description/Function**	**Tags**
Allusion	A short reference to a famous person, event, history, Greek mythology, literature or reference to religion.	ALU
Understatement	A statement consciously weakened or expressed as less important than it actually is, either to soften the message for politeness and tact or to sound ironical.	UNS
Litotes	A figure of speech generated by denying the opposite or contrary of the word which otherwise would be used. It is a form of understatement. Litotes intensify the sentiment intended by the writer.	LTO
Antithesis	Opposition or contrast of ideas or words expressed often in parallel construction. It emphasizes the contrast between two ideas to draw the readers' attention directly to the contrast.	ANT
Hypophora	Question raised and then answered by the author / speaker.	HYP
Rhetorical question	Question without a direct answer. It is used for effect, emphasis, or provocation, or for drawing a conclusionary statement from the facts at hand.	RHQ
Procatalepsis	Allowing an argument to continue through anticipating an objection and answering it, putting into consideration points or reasons opposite to the train of thought.	PRO
Distinctio	Offering the meaning or meanings of a word in order to remove ambiguity.	DST
Simile	A direct comparison between two different things that resemble each other at least in one way, often by using the words' *like'* or *'as'*.	SIM
Analogy	Overlaps with similes Comparing two things with similarities in several aspects without adding *'like'* or *'as'*.	ANG
Metaphor	Comparing two totally different things by asserting that one thing *is* another thing.	MET
Eponym	A particular attribute of a famous person famous of such attribute.	EPM
Exemplum	Citing an example through offering an illustrative story.	EXM
Sententia	A means of quoting a wise saying or a statement of wisdom.	SNT
Anaphora	The same word or phrase is used to *begin* successive clauses or sentences. This draws the readers'/listeners' attention to the message of the sentence.	ANA
Epistrophe	The counterpart of anaphora where the repeated part comes at the end of successive phrases, clauses or sentences.	EPS

Symploce	Combining anaphora and epistrophe. This is displayed by repeating one word or phrase at the beginning and another is repeated at the end of successive phrases, clauses or sentences.	SYM
Personification	Metaphorically representing inanimate objects or animals or abstract terms as having human qualities.	PER
Amplification	Repeating a word or expression while offering more details as a means of emphasizing its importance.	AMP
Aporia	Expresses doubt about an idea or conclusion. It is a way to raise a number of choices without being obliged to any of them.	APR
Climax	Climax consists of arranging words, clauses, or sentences in an ascending order or the order of increasing importance for continuity and emphasis.	CLX
Parallelism	Similarly structuring successive clauses or sentences as a means to concentrate on the message to show that the ideas in the parallel structures are equal in importance as well as to create a musical effect.	PAR
Chiasmus	It is usually called 'reversed parallelism', because the second part of a grammatical construction is paralleled with the former but in reverse order.	CIA
Metabasis	A brief statement of what has been said before and what will follow. It acts as a sort of transitional summary to keep the discussion ordered and keep the audience focused.	MTA
Anadiplosis	The last word of one phrase, clause or sentence is being repeated at the beginning or very near to the beginning of the next.	AND
Conduplicatio	A key word is being repeated from a preceding phrase, clause or sentence to the beginning of the next.	CND
Apostrophe	Interrupting the discussion and directly addressing a person or personified entity either present or absent.	APS
Polysyndeton	The use of a conjunction between each word, phrase, or clause as an attempt to encompass something complex.	POL
Asyndeton	Omitting conjunctions between words, phrases, or clauses as an attempt to give the effect of multiplicity and spontaneity. It is the opposite of polysyndeton.	ASN
Zeugma	Zeugma includes grammatically linked parts of speech by another part of speech. This is done with two or more parts of speech.	ZGM
Synedoche	Any portion, section, or main feature stands for the whole itself or vice versa.	SYN
Metonymy	Another form of metaphor The thing chosen for the metaphorical image is closely associated with the subject with which it is compared.	MTN
Alliteration	Repetition of the *initial consonant sound* in neighboring words. Alliteration draws attention to the phrase and is often used for emphasis.	ALT
Expletive	A single word or short phrase, usually interrupting normal syntax, used to lend emphasis to the words immediately proximate to the expletive. The expletive can be placed at the beginning, middle or at the end. The words on each side are emphasized in order to maintain continuity of the thought.	EXP
Tricolon	A rhetorical term for a series of three parallel words, phrases, or clauses.	TRI

The fifth maxim sounds as an advice for the end user. This advice is concerned with clarifying that the annotation done in the corpus should not be viewed as a perfect and flawless production, but is a tool that can aid in future research. The sixth maxim stresses that any scheme used in the annotation process should be based on theory-neutral principles. That is, principles that are widely agreed upon by linguists and not controversial ones. The seventh maxim is both an advice for the annotator and the end user (Leech, 1993[2]). The maxim emphasizes that no annotation scheme is to be considered as a standard. Standards are considered as such after general accord and this can happen only after the annotation scheme is practically applied. These maxims are taken very closely into consideration in the analysis of the corpus of the present study. The researcher focuses on meeting all the maxims of annotation so as to create an annotated corpus that not only would be of help to other researchers but also helps provide findings.

C. Different Types of Annotation

There are three types of corpus annotation; fully manually, fully automatically and semi- automatically (Bird &Liberman, 1999). All the three types have pros and cons. The fully manually annotated corpus has the virtue of being of highest quality, yet it is tremendously time consuming and still a human researcher's annotation is prone to error. Humans are of course more accurate than machines since they embrace the value of reasoning. This is the one used in the analysis of the corpus in the study. Annotation in this study plays the pivotal role in the analysis of the political speeches, the corpus of the study. The researcher analyses the speeches searching for the different instances or occurrences of the rhetorical devices to which the tags are assigned. The second type of annotation is the one automatically carried out. Although this automatic type of annotation is quick, yet it is consistently full of errors. A computer program, no matter how suitable for the task, commits a high number of errors.

The third type of annotation is a mix between the first two types. This type entails automatic annotation with manual post annotation editing. Accordingly, the tags are annotated into the speeches to indicate the occurrence of the devices they stand for. Once an instance is spotted, the tag is placed at the beginning and at the end of the instance. An illustration of this is the following example taken from King's speech "I Have a Dream":
"<TRI> <PAR> <ALT> Life, Liberty <ALT/> and the pursuit of Happiness <TRI/> <PAR/>."

The above example shows several occurrences of several devices at the same time. Tricolon, alliteration and parallelism are assigned the tags <TRI>, <ALT> and <PAR> respectively. The tag is placed twice to surround the instance thus simplifying and clarifying the tracing of all the various instances. The brackets < TAG> surround the tag that opens at the beginning of the instance. The end of the instance is surrounded by the same brackets but includes an oblique <TAG/> to indicate that the instance has ended.

The study encompasses a corpus of approximately 40,000 words included in the fourteen speeches.

D. The Concordance Program: MonoConc Pro 2.2

The program chosen for the study is MonoConc Pro 2.2 for concordancing and corpus analysis (http://www.athel.com/mono.html.). The program's user interface makes the software easy to deal with. The program helps researchers upload a corpus and search. The search results appear in just a few seconds and are displayed in a very clear manner. It also offers expression searches and tag searches. This of course requires that the tag set is uploaded to the concordance program along with the annotated corpus. The program searches for word lists and frequency lists, for words and phrases, and also for collocates and collocations.

According to Barlow, MP 2.2 has newly added features such as highlighting the frequent collocates in a different color and they appear in the concordance result window. The results or the retrieved examples appear in a form of keywords that are highlighted and are shown in context. By clicking on the highlighted example appearing in the results window, the whole sentence where the word or occurrence lies appears in the context window. This helps identifying the data visually with utmost ease (Barlow, 2008[3]). The originality of the present study stems partly from its distinctive tag set. Such tags are created by the researcher and they need a software program for the analysis. Most concordance programs analyze the part of speech tags, but for the present study the tags are discourse based. After the whole corpus is annotated, it is uploaded to the MonoConc. The tags are of course also added to the software to be able to spot them as needed for the researcher's purposes. One device is searched at a time, and the program displays all instances of the required tag search. Numbers of the occurrence of every device are displayed to the researcher, who then starts collecting the results to arrive at conclusions. The conclusions are related to the type of devices used by every speaker and the amount of usage. Once the search is done, the program spots the specified device and brings it forth to the researcher in the results window and other information is also displayed.

6 CHOOSING THE TWO POLITICAL FIGURES

The study has at its heart a corpus of political speeches. These speeches were given by two political figures described by many writers as two very eloquent orators. The first figure is Martin Luther King, Jr., the clergy man and the son of the African American Baptist church, who managed to change history through his eloquent speeches. He was a man driven by his dream of achieving equality for all of "God's children" as he always describes mankind in his speeches. The 4th of April, 2015 served as the 47th anniversary of King's assassination. Although Martin Luther King died at the age of 39, he had several contributions in various areas springing from his connections to the peace and social justice, humanist and civil rights movements of his time. He acted as a source of inspiration and a muse for a variety of the intellectual, cultural and political developments belonging to the twentieth century. King spent years of his life fighting to gain the dignity of the oppressed people all around the world and not only the blacks.

His oratory, infused with the experience he gathered from his readings in theology as well as his own insights, had a glowing effect on so many as was evident in his preaching activities. He joined and created so many associations and movements calling for the rights of the blacks. The Montgomery Improvement Association (MIA) which was formed by a number of notable Montgomery black leaders including Ralph Abernathy, his lifelong companion, is only one illustration of the many leaders who fought by his side. King took the role of the primary spokesperson of the year-long Montgomery bus boycott which he actually spoke about in his speeches. His oratory, deep beliefs in the equality of all human beings, theological background and Mohandas Gandhi's teachings of nonviolence of which King was an advocate, transformed him into a leader capable of expressing himself in memorable words thus mobilizing forces to fight by his side.

King's speeches are rich with the variety of rhetoric employed throughout. As a political and religious leader, King's aim is definitely to move and persuade thus leading to the major end desired from the listeners; to act. His speeches influenced masses of people belonging to his similar school of thought and others from different walks of life. In chapter three, the chapter responsible for the analysis, the language of the speeches is analyzed and the rhetoric of both speakers is put on display. Not only do King's speeches have linguistic richness between its lines that can help researchers arrive at theories and investigate language, but also King's speeches have many contributions and legacies in many areas of life. The year 2008 unfolded on the 45th anniversary of King's most famous speech "I Have A Dream" and in the same year Barack Obama became the first African American to accept the presidential nomination of a major political party at the 2008 Democratic National Convention. This is definitely the realization of one of King's dreams that all human beings are equal and that they should be assessed by "the content of their character and not the color of their skin". Consequently, King can be described as Obama's god father.

Assessed by the content of his character rather than the color of his skin or his African roots, Barack Obama is now the 44th Unites States President succeeding George W. Bush. Obama is the second political figure in this study whose political speeches serve as the other half of the study's corpus. Barack Obama became President at noon on January 20th, 2009 which is a date specified by the Twentieth Amendment of the Constitution. The Amendment requires that the president starts officially holding the office at noon on January 20 following the year of the presidential election. This day is known in America as the Inauguration Day thus marking the four-year term of both President and Vice President. Obama and King are highly connected for several reasons. Both speakers have African origins and both are always referred to as great eloquent orators who can stir and enchant audiences.

Obama's capability of stirring an audience is many linguists' area of research. He is described by many writers as having the ability to use simple words in his speeches, yet manage to elevate and inspire the audience through techniques that he uses (Assumndson, 2008[4]). The speeches that Obama gave during his campaign running for presidential election are widely praised as master pieces which have inspired many writers to work on analyzing Obama's style. Many writers search for Obama's secret behind his ability to stir the crowd. Obama's election itself has its historical value and the way he uses his simple words to awaken, stir, inspire and stimulate the audience to revive their hopes that a better America in particular and a much better world in general is possible. Analyzing his words and looking deep into the stylistic or rhetorical devices used is the concern of the researcher of the present study.

7 ANALYSIS

Uploading the annotated corpus to the program, MonoConc Pro, was followed by the entering of the tag set designed by the researcher. The search takes place one tag at a time. Figure 1 displays an example of the occurrences of one of the devices in the uploaded corpus. The tag of the device shows in the middle of the window in blue surrounded by the fragments in which they occur.

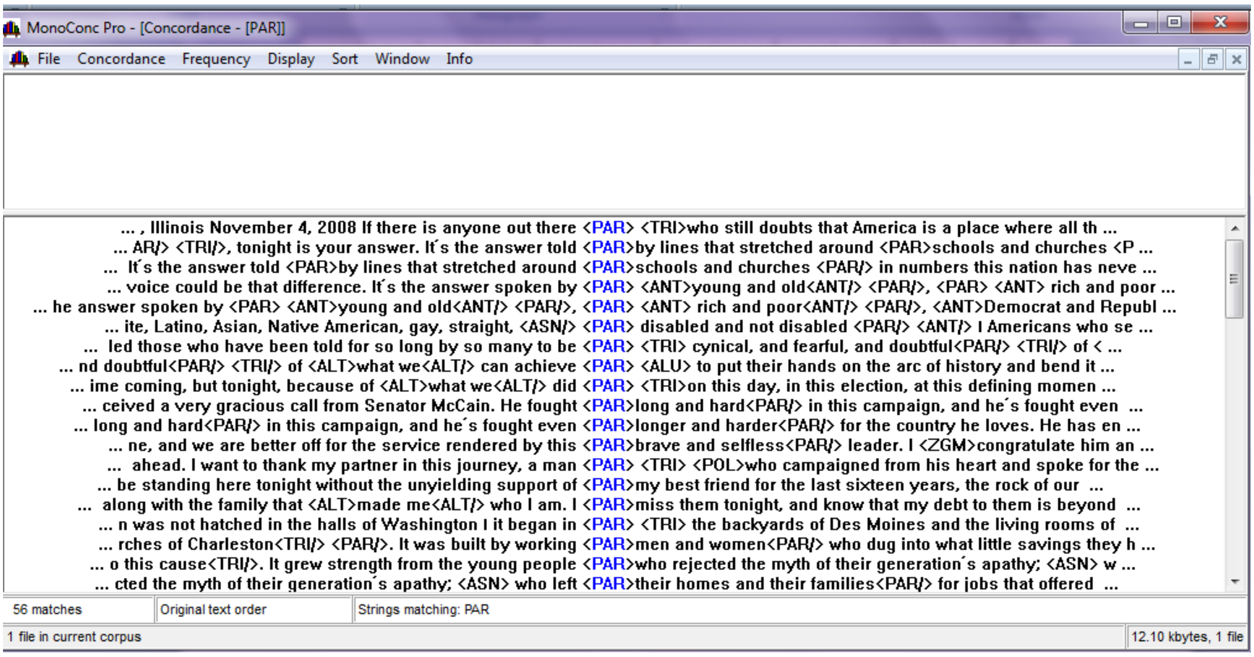


Figure 1: Occurrence of parallelism

By clicking on any of the highlighted occurrences, the whole instance of the device shows in an upper window. The results then appear in a double window where the researcher can clearly read the whole instance as a better way to understand the device in context. This is shown in Figure 2.

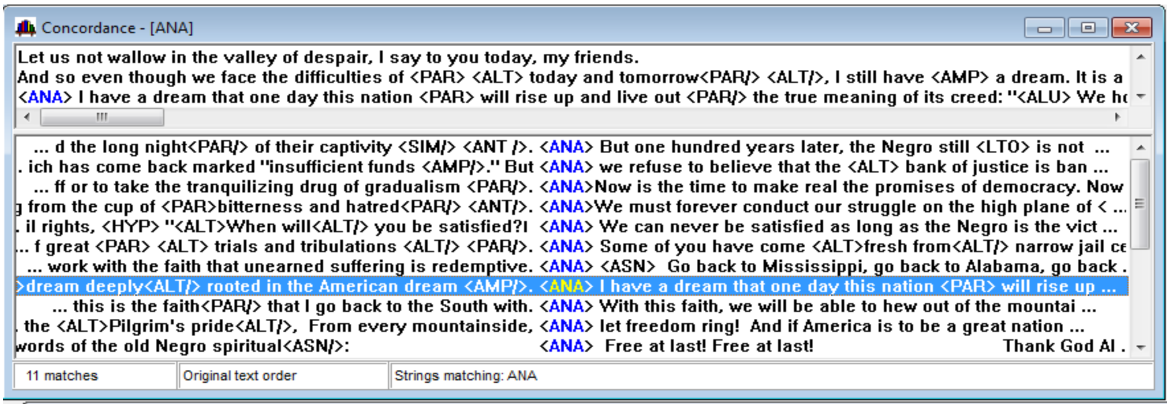


Figure 2: Search results of anaphora in two windows

A. *The results of the analysis of King's speeches*

"Give Us the Ballot", the first speech in this selection, does not show high numbers in the use of rhetorical devices. Parallelism, alliteration and allusion are the top three devices used occurring 55 times, 29 times and fourteen times respectively. This group is followed by antithesis and tricolon both used nine times. Anaphora is used six times, expletives five times with two similes in the speech. Understatement, sentential, amplification, conduplicatio, asyndeton and zeugma appear only once. The other devices have no occurrences at all.

The second speech in King's collection is The Great March on Detroit. Parallelism and alliteration occur 49 and 43 times respectively. These two numbers are followed by ones that are smaller, for example antithesis appears sixteen times. Anaphora occurs 12 times, allusion nine, tricolon eight and asyndeton seven times. Metaphor, personification and polysyndeton all occur four times. Symploce occur twice whereas rhetorical question and amplification appear only once.

The third speech is King's speech "I Have a Dream". The highest number of occurrences of a device goes for parallelism which occurs 38 times in the speech. This is followed by alliteration which is used for 35 times, antithesis occurring 21 times and anaphora 11 times. Some other devices are used in the speech but in little numbers, such as allusion which is used six times and climax which occurred only once. On the other hand, some devices did not occur at all in the speech. These devices are rhetorical question, procatalepsis, distinctio, understatement, eponym, exemplum, sentential, epistrophe, personification, aporia, chiasmus, metabasis, anadiplosis, conduplicatio, apostrophe, zeugma, syndoche and expletive.

The fourth speech is the one that King gave in Oslo when he was receiving the Noble Prize. The numbers of occurrences of all devices in general in this speech are not as huge as its counterparts. The highest occurrence is of parallelism which occurs thirty seven times in the speech. This is followed by the second highest number of device occurrence which is alliteration. Tricolon occurs seven times in the speech and antithesis and anaphora occur five and four times respectively. Allusion appears three times and personification and conduplicatio both occur twice. Distinctio, simile, metaphor, symploce, amplification, zeugma and expletive each have a single occurrence only.

The following speech in this section is Our God Is Marching On. 70 occurrences of parallelism are found followed by 65 occurrences of alliteration. Following these two devices are allusion and anaphora appearing 14 and 13 times respectively. Antithesis occurred 11 times whereas tricolon and hypophora appeared nine and seven times respectively. Both personification and polysyndeton occur five times, while metaphor and amplification occur only twice. Each of exemplum, metonymy and expletive has only one occurrence.

The sixth speech in this collection is Beyond Vietnam. Numbers of occurrences of devices in this speech outnumber the first two speeches. Largest numbers of occurrences are scored by parallelism scoring 131 occurrences, followed by 79 occurrences of alliteration, thirty three occurrences of allusion and lastly thirty occurrences of tricolon. These huge numbers are followed by rhetorical question and antithesis appearing 25 and 24 times respectively. Litotes is used sixteen times, expletive used 12 and anaphora is used 10 the same as metonymy in the speech and personification is used nine times. There are six occurrences of asyndeton and five for metaphor. Procatalepsis, simile, symploce and zeugma all occur three times, while polysyndeton appears twice. Hypophora, epistrophe, amplification, metabasis and conduplicatio are all used only once. The other devices are not used in this speech at all.

I See the Promised Land is the last speech chosen for King. As usual, parallelism occupies the highest number of occurrences scoring fifty as shown in figure 13. This is followed by 38 occurrences of alliteration and 20 of allusion. Anaphora is used 12 times succeeded by rhetorical question appearing eight times and asyndeton six times. Both hypophora and symploce are employed five times in the speech while both antithesis and metonymy four times. Expletive and amplification appear twice and each of tricolon, distinctio, simile, personification, conduplicatio and apostrophe show only a single occurrence.

B. The results of the analysis of Obama’s speeches

The first speech is the one he gave in South Carolina. The search results of the number of occurrences in the South Carolina speech show forty eight occurrences of parallelism, followed by fourteen occurrences of alliteration. Obama used antithesis eleven times in the current speech, and used tricolon ten times. These scores are followed by eight uses of polysyndeton. The three of anaphora, apostrophe and asyndeton are employed for five times. Metonymy is used for three times whereas amplification, conduplicatio and expletive are used only twice.

The second speech in Obama’s selection is “Super Tuesday”. The highest number of occurrences is scored by parallelism which occurs fifty times throughout the speech. The second highest number is shown through the occurrence of alliteration. Tricolon occurs 14 times and both expletive and asyndeton occur eight times. This is followed by polysyndeton and anaphora occurring seven times. Antithesis occurs six times whereas allusion four times. Each of rhetorical question, exemplum, amplification, conduplicatio and apostrophe occurs only once.

‘Night Before the Election’ is the speech that Obama gave one day before he was announced President of the United States of America. This speech is the fourth in the selection. Parallelism occurs 46 times in this speech followed by anaphora and asyndeton which occur 13 times and 11 times respectively. This is followed by tricolon and alliteration that occur nine times and eight times respectively. Metonymy occurs seven times whereas antithesis occurs six times. Polysyndeton, apostrophe and expletive occur in five, three and two times in that order. Finally, exemplum, amplification and climax each occurs only once.

The Election Night Victory Speech is the fourth speech in this selection. In this speech the highest number of occurrences of rhetorical devices goes to parallelism which occurs fifty six times. Alliteration follows parallelism occurring thirty six times. Tricolon makes the third highest number. These three high numbers are followed by nine occurrences of antithesis,

seven for apostrophe, six for allusion, four for anaphora and three for asyndeton. Rhetorical question, amplification and polysyndeton all occur twice. Distinctio, exemplum, zeugma and expletive each occurs only once.
The fifth speech is the Inaugural Speech which makes the first speech for Obama as President. Ninety four occurrences of parallelism are found in the Inaugural Speech, followed by 31 alliterations. Tricolon makes up eighteen occurrences, while anaphora occurs twelve times throughout the speech. One occurrence less than anaphora, antithesis occurs eleven times. This is followed by asyndeton, apostrophe and allusion occurring nine and eight and six times respectively. There are four occurrences of metonymy, three of amplification, and one for zeugma. Both symploce and climax occur twice.

The sixth speech in Obama's selection is the speech he gave in University of Cairo. This speech has a huge number of occurrences of both parallelism and alliteration occurring one hundred fifty four times and one hundred and five times respectively. These very huge numbers are followed by numbers that are close to each other. Tricolon and metonymy occurred thirty and twenty seven times respectively. Allusion occurs twenty times and antithesis, only two occurrences less, occurs eighteen times. Fifteen occurrences of expletives are found in this speech. Anaphora has eight occurrences and amplification occurs four times. Symploce and litotes and zeugma have three occurrences. Finally metaphor and epistrophe occur only once.

The last speech is the one that Obama gave after the Egyptian President Hosny Mubarak stepped down on the eleventh of February 2011. The speech does not contain a big number of devices used. Parallelism is used seventeen times followed by alliteration thirteen times. Tricolon is used for four times and allusion is used three times. Antithesis, simile and metonymy are used twice, while anaphora, amplification, polysyndeton, asyndeton and expletive are used one time.

Figure 3 is a bar chart showing the results of all the occurrences of all devices in the fourteen speeches of both King and Obama. The bars show the differences between the two speakers' usage of the devices belonging to the four categories. There is a similarity in the use of the devices that belong to the last two categories. These categories include devices that are used to organize the writing and the other includes devices that create certain structural pattern to add a distinctive style to the writing. In these two categories Obama shows a wider range of use of devices in both categories scoring 498 and 480 occurrences in the third and fourth categories respectively whereas King scored 438 and 459 occurrences respectively. King then shows higher numbers of occurrences than his counterpart with a difference of 182 occurrences in the first category and a difference of 51 in the second category.

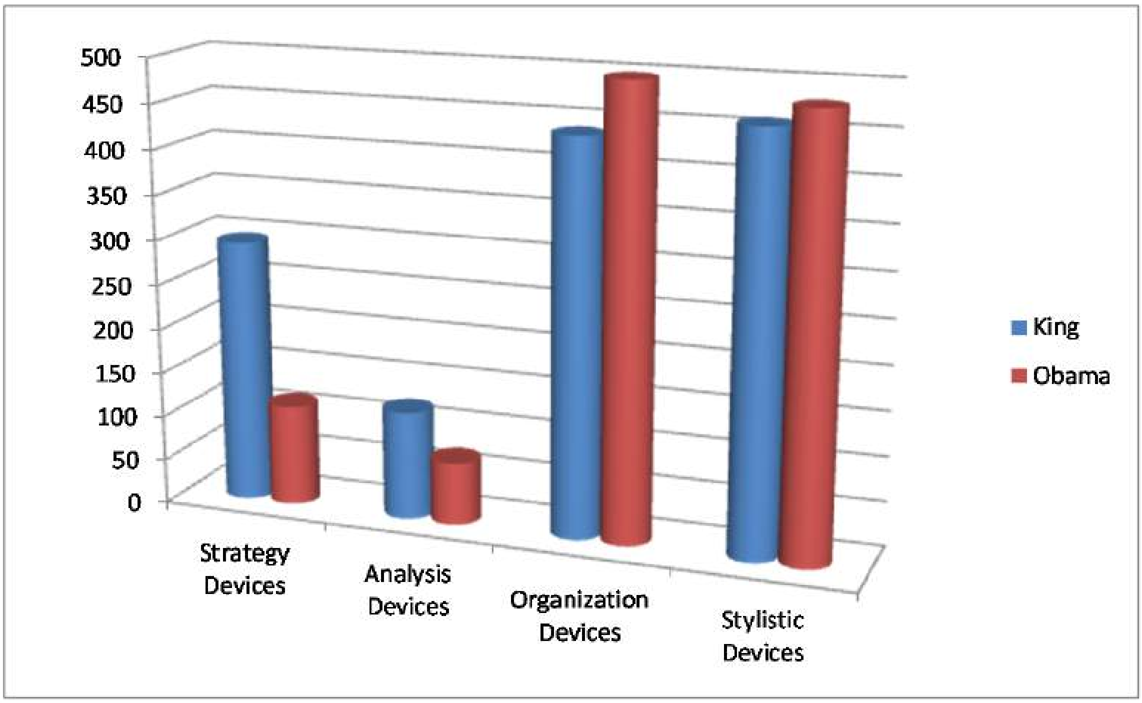


Figure 3: A Bar Chart Showing the Differences Between King and Obama's Use of the Devices in the Four Categories

8 CONCLUSION AND RECOMMENDATIONS

The present study is a corpus based study where the aim is to create an annotated corpus on the discourse level. Creating such output will facilitate the job of many other researchers trying to identify the style of speakers through analyzing the language they use. Having such output available will enable them to come up with the findings with the press of a button. In the present study, part of the methodology is to create a tag set of the selected rhetorical devices. The tags were designed on the same lines of part of speech tagging. Three capital letters resembling the word being tagged were created for every device. The annotated corpus was then uploaded to the concordance program and the search began. The tags proved success as well as the choice of the computer program, the MonoConc Pro. Although many tags are sometimes placed in the same paragraph representing their occurrences, yet the search for the tags was not problematic at all. Once the search of a certain tag starts, the software separates the searched for tag from the others and the occurrences appear clearly. The program not only produces the occurrences of the tags, but also certain phrases.

The tagged corpus transformed the speeches from being an undiscovered creation into a living body of data that has tags within its lines. These annotations or tags help unfold the secrets behind stirring an audience and behind making them laugh or cry, behind the rise of a leader as a world leader and the fall of another. This annotated output enables researchers to investigate its language looking for the used rhetorical devices which will help know the style of the authors and speakers. They can search for any phenomenon they might be working on investigating. This annotated corpus enables researchers to find the power behind the language of political speeches. The investigations, analyses and results finally arrived at could not have been feasible except through a corpus as such.

The research questions that the study started with are answered through the analysis and investigation. A main aim of the study is to build a discourse- based corpus. Such output is the annotated corpus which embraces the designed tags corresponding to the chosen devices. The study proved that building an annotated corpus on the discourse level is possible. A second research question is concerned with the creation of the tags. The tags were created to represent the rhetorical devices and they follow the same pattern of the part of speech tags. The tags are placed in the corpus and uploaded to the computer program. The results of the search showed that the tag set worked successfully embracing the occurrences of the various devices. The chosen rhetorical devices are organized in categories based on the purpose of using them. Consequently, after the analysis of the tagged speeches using the concordance program, the search results clarified which devices are used and to which categories do these devices belong. This can enable researchers to both clarify the effect of the devices and also to identify the style of the politicians.

Leech's annotation maxims were a very good guide in the annotation process. The annotated corpus can be reverted to its original state through the removal of the tags. Equally, the annotations can be extracted by themselves from the corpus. Since the study is of good use to other researchers, a clear description of the annotation scheme is provided. This description also includes that the annotation was carried out by the researcher fully manually and of course such annotation scheme might be prone to error and is not presented as a standard but as the primary endeavor. The originality of the study stems from its discourse based corpus. Such output was never available before.

This output can definitely be enlarged in future research. If more speeches for the same speakers are annotated, this will enable researchers to arrive at more reliable conclusions about the speakers' styles. Annotating speeches that belong to different stages in the speakers' lives can also help trace the changes or spot the similarities in their styles as a means of arriving at a better understanding of their way of thinking. That is, this output can also be enlarged through using parallel corpora. Speeches in both Arabic and English can be annotated and the differences or the similarities be pointed out for further analysis and investigation. Enlarging the set of rhetorical devices will also add to the annotated output. Through a wide range of rhetorical devices, which will be assigned new tags of course, more reliable conclusions can be drawn about the speakers' styles. Researches on this path won't be possible before the presence of an annotated corpus as the one present in this study. With the press of a button the researcher can come up with numbers of the occurrences of the different devices, no matter how large these numbers might be. Such conclusions will allow the researcher to create theories about language.

REFERENCES

[1] Bird, Steven & Liberman, Mark. (March,1999). *Formal Framework for Linguistic Annotation*. Tech.Rep. MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 1999. http://xxx.lanl.gov/abs/cs.CL/9903003.

[2] Leech, Geoffrey. (1993). Corpus Annotation Schemes. In Literary and Linguistic Computing, 8(4): 275-281.

[3] Barlow, Michael. (2008). Parallel texts and corpus-based contrastive analysis, In: GÃ³mez GonzÃ¡lez, M., Mackenzie, L. and GonzÃ¡lez Alvarez, E. (eds.), *Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives.,* Benjamins, 101-121.

[4] Assmundson, Mikael (2008). Persuading the Public: A Linguistic Analysis of Barack Obama's Speech on "Super Tuesday" Dalarna University, School of Languages and Media Studies, English.

[5] A. Harris, Robert. (2003). *Writing with Clarity and Style: A Guide to Rhetorical Devices for Contemporary Writers.* Los Angeles: Pyrczak Publishing House.

[6] Aaarts, J. & Meijs, W. (Eds). (1984). *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Research.* Amsterdam: Rodopi.

[7] Aaarts, J. & Meijs, W. (Eds). (1990). *Theory and Practice in Corpus Linguistics.* Amsterdam: Rodopi.

[8] Aarts, J., Haan, P. & Oostidik, N. (Eds). (1993). E*nglish Language Corpora: Design, Analysis and Exploitation.* Amesterdam: Rodopi.

[9] Abu El Wafa, Marwa. (2013). *Persuasion in Political Speeches: Discourse Tagging of Political Speeches.(*Unpublished Master's Thesis). University of Alexandria, Alexandria, Egypt.

[10] Aijmer, K. & Altenberg, B. (Eds). (1991). *English Corpus Linguistics.* London: Longman.

[11] Allen, J., and Core, M. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/.

[12] American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th Edition, Second Printing). Washington, DC: Author.

[13] Anonymous. (2008). Text Annotation for Political Science Research. *Journal of Information Technology & Politics*, 5(1).

[14] Araki, Masahiro, Kimura, Yukihiko, Nishimoto, Takuya & Niimi, Yasuhisa. (n.d.). *Development of a Machine Learnable Discourse Tagging Tool.* Department of Electronics and Information Science. Kyoto Institute of Technology.

[15] Barlow, Michael, and Bowker, L. (2008). A comparative evaluation of bilingual concordancers and translation memory systems, In: Yuste Rodrigo, E. (ed.), *Topics in Language Resources for Translation and Localisation.*, Benjamins, 1-22.

[16] Barlow, Michael. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics.*

[17] Beaugrande, Robert de. (January, 2006). Critical Discourse Analysis: History, ideology, methodology. *Journal of Studies in Language & Capitalism,* 29 – 56.

[18] Biber, Douglas. (2006). University Language: *A Corpus-based Study of Spoken and Written Registers*. Amesterdam: John Benjamins.

[19] Biber, Douglas, Connor, Ulla & Upton, Thomas A.(2007). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure.* Amsterdam: John Benjamin.

[20] Biber, Douglas, Conrad, Susan & Reppen, Randi. (1998). *Corpus Linguistics: Investigating Language Structure, and Language Use*. London:Cambridge University Press.

[21] Carlson, Lynn, Marcu, Daniel & Okurowski, Mary Ellen.(n.d.) *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*.

[22] Carlson, Lynn & Marcu, Daniel. (2001). Discourse Tagging Reference Manual. Technical Report, Information Science Institute.

[23] Carlson, Lynn & Daniel Marcu. (2012). Ecological Evaluation of Persuasive Messages Using Google Ad Words. Journal: CoRR, Vol.abs/1204.5369.

[24] Danieli, Morena. (1999). Discourse Tagging. Towards Tools and Standards for Discourse Tagging. (ACL-99 Workshop) June 22, 1999 University of Maryland College Park, MD, USA. URL: http://www.mri.mq.edu.au/conf/acl99/html.

[25] Finegan, Edward. (1996). [Review of Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture, by Michael Stubbs].

[26] Garside,R Leech, G. & McEnery, Tony. (1997). *Corpus Annotation. Linguistic Information from Computer Text Corpora*. London: Longman.

[27] Halmari, Helena & Virtanen, Tuija (Eds). (2005). *Persuasion Across Genres*. Amsterdam:John Benjamin's Publishing Company.

[28] Hunston, Susan. (2002). *Corpora in Applied Linguistics.* United Kingdom: Cambridge University Press.

[29] J., Thomas & M., Short. (Eds.) (1996). *Using Corpora for Language Research*. London: Longman.

[30] Johansson, Stig & Oksefjell, Signe. (Eds). (1994). *Corpora and Cross-Linguistic Research: Theory, Method, and Case Studies.* USA: John Benjamins Publishing Company.

[31] Kennedy, Graeme. (1998). *An Introduction to Corpus Linguistics.* New York: Longman Pearson Education Inc.

[32] Lapadat, Judith C. (2007). Discourse Devices used to Establish Community, Increase Coherence, and Negotiate Agreement in an Online University Course. *Journal of Distance Education*, Vol. 21(3), 59-92.

[33] Leech, Geoffrey. (1992). Corpora and Theories of Linguistic Performance. In Svartvik (1992)

[34] Nomoto, Tadshi & Matsumoto, Y. (1999). Learning Discourse Relations with Active Data Selection. In P. Fung and J. Zhou (eds.), Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 158-167, University of Maryland, MD.

[35] M., Stubs. (1996). *Text and Corpus Analysis*. Oxford:Blackwell.

[36] M., Oakes. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

[37] Maurice De Schryver, Gilles. (2001). Corpus-based Activities versus Intuition-Based Compilations by Lexicographers, the Sepedi Lemma-Sign List as a Case in Point. Nordic Journal of African Studies 10(3): 374-398.

[38] McEnery, Tony. (2003).Corpus Linguistics. In Ruslan Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp.448-456). New York: Oxford University Press.

[39] McEnery, Tony & Wilson, Andrew. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

[40] McEnery, Tony & Wilson, Andrew. (2003). Corpus Linguistics. Retrieved from Essex.ac.UK: http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/types/annotated.html.

[41] Meyer, Charles F. (2002). *English Corpus Linguistics: An Introduction*. United Kingdom: Cambridge University Press.

[42] Mitchell P., Marcus, Mary Ann, Marcinkiewicz & Santorini, Beatrice. (1993). Building a large annotated corpus of English: The Penn Treebank. Journal,19(2). 313–330. URL: http://portal.acm.org/citation.cfm?id=972475.

[43] Mladová, Lucie, Zikánová, Šárka & Hajičová, Eva. (2008). From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In Proceedings of the 6th International Conference on Language Resources and Evaluation.

[44] Moliken, Paul, Grudzina, Douglas & McGuigan, Brendan. (2007). *Rhetorical Devices: A Handbook and Activities for Student Writers.*USA:Prestwick House, Inc.

[45] Nesselhauf, Nadja. (2005). Corpus Linguistics: A Practical Introduction. Amsterdam: Benjamins.

[46] Newman, John. (2008). Aiming low in linguistics: Low-level generalizations in corpus-based research. Proceedings of the 11th International Symposium on Chinese Languages and Linguistics (IsCLL-11), May 23-25 2008, National Chiao Tung University, Hsinchu, Taiwan.

[47] Rapillo, Miriam. (n.d.). Corpus Linguistics: A General Introduction. Retrieved from:http:// www.givemeawhisper.tk.html.

[48] Schaffner, Christina. (1996). Editorial: Political Speeches and Discourse Analysis, Current Issues in Language and Society, 3: 3, 201 — 204.

[49] Schiffrin, Deborah, Tannen, Deborah & Hamilton, Heidi E. (Eds). (2001). *The Handbook of Discourse Analysis.* United Kingdom: Blackwell.

[50] Sherris, Ari. (n.d.). Corpus Linguistics as Agency for Quantitative and Qualitative Research. Retrieved from Find That PDF: http://www.findthatpdf.com/search-9594609-hPDF/download-documents-essayoncorpuslinguistics.pdf.html.

[51] Sinclair, John & Carter, Ronald (Eds).(1991). *Corpus Concordance Collocation.* New York: Oxford University Press.

[52] Sinclair, John. (Eds). (1992). The Automatic Analysis of Corpora. New York: Routledge.

[53] Sinclair, John & Carter, Ronald (Eds). (2004). *Trust the Text: Language, Corpus and Discourse*. New York: Routledge.

[54] Teun A. van Dijk. (2001).Critical Discourse Analysis. In Schiffrin, Deborah, Tannen, Deborah & Hamilton, Heidi E. (Eds). *The Handbook of Discourse Analysis* (pp.103-136). United Kingdom: Blackwell.

[55] Teun A. van Dijk. (2001). Political Discourse and Political Cognition. In Schiffrin, Deborah, Tannen, Deborah & Hamilton, Heidi E. (Eds). *The Handbook of Discourse Analysis* (pp.203-233). United Kingdom: Blackwell.

[56] Urbanavičienė, Irena. (2004). Political Speeches: Exertion of Power through Linguistic Means Retrieved from Ceeol: http:// www.ceeol.com.html.

[57] Walker, Marilyn. (1999). Towards Standards and Tools for Discourse Tagging. University of Maryland, USA: Association for Computational Linguistics.

[58] Wasowa, Thomas & Arnold, Jennifer. (2004). Intuitions in Linguistic Argumentation. Retrieved from www. Science direct.com: DOI:10.1016/j.lingua.2004.07.001.

[59] Y. Kawaguchi, M. Minegishi & J. Durand (Eds.) (n.d.). *Corpus Analysis and Variation in Linguistics.* Amsterdam: John Benjamins.

[60] Yeates, Stuart& H. Witten, Ian. (n.d.). On Tag Insertion and its Complexity. Department of Computer Science, University of Waikato, Hamilton, New Zealand. Retrieved from S. Yeates, I. Witten @cs.waikato.ac.nz.

Biography

Marwa Adel is an Assistant Lecturer at the Arab Academy for Science, Technology & Maritime Transport. She is originally a graduate of Faculty of Arts, University of Alexandria, English Department, Literature Section. She teaches Pragmatics, Syntax and Linguistics in College of Language and Communication, Language and Translation Department and also teaches a variety of ESP courses in different colleges including College of Engineering, College of Maritime Transport, College of Logistics and College of Business Administration. She obtained her Applied Linguistics Diploma in 2008. She then registered for the Masters Degree in 2009 in University of Alexandria in the field of Applied Linguistics working on a pioneering topic in the field of text tagging. She obtained the Degree in 2013 with a general grade Excellent. In the same year, in September 2013, she got enrolled in the PhD programme and registered the PhD thesis in September 2014 in the field of abstract meaning representation.

Her academic interests are in the fields of Corpus- based Studies, Discourse Tagging, Discourse Analysis, Political Discourse Analysis and Meaning Representation.

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Shadia el-Soussi is a Lecturer at the Institute of Applied Linguistics and Translation, Faculty of Arts, University of Alexandria. She has extensive teaching experience in teaching and course management of English and Arabic as Foreign languages (TEFL &TAFL), English for Specific Purposes, Testing and Evaluation. She has supervised and co-supervised a large number of MA and PhD degree students.

Her academic interests are largely in Sociolinguistics, Discourse Analysis, Critical Discourse Analysis and Cultural Studies.
She has also been working a consultant for the Arabic Bibliographic Department, Library Sector, Bibliotheca Alexandrina, since 2009.

عنونة الخطب السياسية: دراسة تحليلية

مروة عادل أبو الوفا
قسم اللغة و الترجمة، كلية اللغة و الاعلام، معهد دراسات اللغات، الأكاديمية العربية للعلوم و التكنولوجيا و النقل البحرى
ميامى، الأسكندرية ، مصر
سامح الأنصارى
قسم الصوتيات و اللغويات، كلية الأداب، جامعة الأسكندرية
الشاطبى، الأسكندرية، مصر
مكتبة الأسكندرية، الأسكندرية، مصر
شادية السوسى
معهد اللغويات التطبيقية،، كلية الأداب، جامعة الأسكندرية
الشاطبى، الأسكندرية، مصر

ملخص
يناقش هذا البحث امكانية تصميم مجموعة من الأكواد على مستوى النص و ذلك من خلال تكويد أو عنونة أساليب بلاغية متنوعة تستخدم فى الخطب السياسية و غيرها من أنواع النصوص بغرض الاقناع و توصيل الأفكار بطريقة بلاغية و فى هذة الدراسة تم تكويد الأساليب البلاغية التى أستخدمها كلا من الرئيس الأمريكى باراك أوباما و القائد الأمريكى مارتن لوثر كنج فى سبع خطب لكل منهم. و يتم هذا على طريق عنونة النصوص كوسيلة لتصميم مجموعة أكواد خاصة بالباحث و استخدامها لتكويد أو عنونة الخطب الأربعة عشر التى تم اختيارها للبحث .كما تم اختيار برنامج **MonoConc Pro 2.2.** لرصد اعداد الأساليب البلاغية المختارة عن طريق تتبع الأكواد التى تم ادخالها على البرنامج من قبل الباحث. و قد تؤدى الأ رقام الناتجة عن الأعداد للوصول لأسلوب كل شخصية سياسية كما تبين ما يتشابهون فية و مايختلفون فية و قد تتيح هذة الدراسة الفرصة لباحثين اخرين من الاستفادة من الدراسة فى نفس المجال.

Building a POS-Annotated Corpus for Egyptian Children

Heba Salama, Sameh Alansary

Phonetics and linguistics Department, Faculty of Arts Alexandria University
Salamaheba00@gmail.com
Phonetics and linguistics Department, Faculty of Arts Alexandria University
Sameh.Alansary@bibalex.org

***Abstract*—In this paper, we present an attempt at developing a POS annotated corpus for Egyptian children. Linguistic annotation of the corpora provides researchers with better means for exploring the development of grammatical constructions and their usage. This is an initial annotated corpus for Egyptian children. It implements part of speech tag (POS) especially a morphologically annotated corpus of spoken Arabic child language. POS are made in "%mor" 'morphology' tiers manually. Coding language transcripts for computer analysis is a daunting task. It approximately took 170 hours, and thus manual annotation focused on a particular child. The POS coding process started with a purely manually annotation of 2701words. 1380 words annotated for an adultand 1321 annotated words for the child was handled. Annotated child language proved to be challenging, and time consuming task. The MOR grammar exists in many languages, such as English, French, German, Japanese, Cantonese, Hebrew, and they are generated automatically, the CLAN has the automatic coding system "MOR program". In Egyptian Arabic, this is not applied for two reasons. First, there is no previous Egyptian Arabic work done on a constructing system for such a representation. Second, morphology of Egyptian Arabic is very rich and different from other languages. Thus, their rules cannot be applied to Arabic. In the two Arabic studies of Qatari and Emirati languages, semi-automatic and mini automatic MOR is used. Finally, certain applications of linguistic analysis commands are provided by using CLAN software. The analyses include frequency counts, word searches, co-occurrence analyses; MLU (mean length of utterance) counts and analyzes specified pairs of utterances. Transcript data provide some morphological analysis, such as mean length of utterance (MLU) counts, lexical analysis, such as frequency (FREQ) count, syntactic analysis, such as searching the data for specified combinations of words or complex string patterns (COMBO) count, as well as the discourse and interactional analysis, such as analyzes specified pairs of utterances (CHIP) count.**

Key words:POS annotated corpus, CHILDES database.

1 INTRODUCTION

A part-of-speech tagging is usually called (POS) tagging, or simply tagging, but is also known as grammatical tagging or morphosyntactic annotation [1] takes place at word level and adds morphosyntactic information next to each word in the corpus. The information added makes the grammatical category to which each word belongs explicit, by adding codes such as: adjective, comparative; noun, countable, singular; verb, simple present, third person. It increases specificity of data retrieval from a corpus, and helps in syntactic parsing, and semantic field annotation. It allows us to distinguish between the homographs. The aim of a Part of speech annotation is to assign each lexical unit in the text a code indicating its part-of-speech. Different tagsets may distinguish a different number of categories, and consequently include a different number of tags, and they may use very different codes for the same categories. POS-tagged corpora allow corpus linguists to perform advanced searches in the corpus.

Corpus annotation has become a major effort in recent years, both for linguistic research and for natural language processing applications. Linguistic annotation of the corpora provides researchers with better means for exploring the development of grammatical constructions and their usage. The main advantage of the use of a standard representation of morphosyntactic coding enable is to test the impact of universality in the development of grammatical marking and syntax in corpora from different languages. Conventions and procedures described in the present research are based on the CHAT conventions of CHILDES system. The CHAT conventions have been modified to achieve a targeted coding scheme for the Egyptian Arabic, based on the classification of [2]. The coding scheme focuses on the development of grammatical marking and syntax. This required the use of a standard representation of morphosyntactic coding.

2 PART OF SPEECH CODES

The codes for grammatical categories were from the CHAT, but with some adaptation to suit the Arabic language. More subcategories were added in Arabic that were not found in English. The morphological codes on the "% mor" line begin with a part-of-speech code. The basic scheme for the part-of-speech code is a category: subcategory: subcategory. The colon character is used as the field separator. The subcategory fields contain information about syntactic features of the word that is not marked. For example, /ʔækɪl/ "ate" is a past verb and there is no single morpheme signaling past, so the part-of-speech code is **v: past**. Information that is marked by a prefix or suffix is not incorporated into the part-of-speech code. The information is found in the right of the **|** delimiter.

A. Stems

The codes for the stem are found on the right hand side of the **|** delimiter, following any pre-clitics or prefixes. Every word on the "% mor" tier must include a "lemma" or stem as a part of the morpheme analysis. A single form is selected for each stem. Thus, the Arabic definite article is coded as **det|ʔel** with the lemma /ʔel-/ whether the actual form of the article is /ʔel-/ or /ʔe-/ if /l/ is omitted from the moon letter.

B. Affixes

The codes for affixes and clitics are in the position in which they occur in relation to the stem. CHAT conventions are used to encode the morphological structure of word forms. For example, the delimiters (-) are used for a suffix, e.g., n|qeṣaṣṣ-BROK&PL, the symbol (&) is employed to indicate inherent features (like the gender of nouns), and morphemes that are not separable. The (&) is used to mark affixes that are not realized in a clearly isolable phonological shape. For example, the form /tuffæ:ħ/ "apples" cannot be broken down into a part corresponding to the stem /tuffæ:ħ/ "apples" and a part corresponding to the plural marker. For this reason, the word is coded as n|tuffæ:ħ&PL. Several codes indicated with the & after the stem e.g., the form /ʔækıl/ "ate" is coded v|ʔækıl&PAST&1s.

3 EGYPTIAN ARABIC PARTS OF SPEECH

Languages vary considerably in morphological complexity. English, for example, has a simple morphology compared with languages, such as Arabic and Hebrew [3]. Arabic is a language of rich morphology compared to other languages especially European languages. It is based on both derivational and inflectional morphology. The richness of Arabic morphology makes the analysis process difficult to deal with. On the one hand, the morphological analysis process is used in the most of the NLP (natural language processing) applications, such as information retrieval, spell checking, and machine translation. In general, morphological analysis of any given word consists of determining the values of a large number of features, such as basic part of speech (i.e., noun, verb), gender, person, number, voice information about the clitics[1][4].

The grammar of Arabic is standardized for centuries. An initial tagset was derived from this grammatical tradition rather than from an Indo-European based tagset. Morphological tag cannot do successfully using methods developed for English because of data sparseness. Indeed, Egyptian Arabic is a very different language from Indo-European languages and should have its own tagset. In addition, Arabic linguists are basically focusing their studies on a traditional Arabic grammar rather than on Indo-European grammar. Arabic grammarians traditionally analyze all Arabic words into three main parts-of-speech. However, according to the present study parts-of-speech are categorized into more detailed ones, which collectively cover the whole of the Egyptian Arabic language [5]. The three main parts of-speech are:

A. Noun

A noun in Arabic is a name or a word that describes a person, thing, or idea. Traditionally the Noun class in Arabic is sub-divided into Derivatives (that is, nouns derived from verbs, nouns derived from other nouns, and nouns derived from particles) and Primitives (nouns not so derived). These nouns are sub-categorized by number, gender, and case. This class also includes what, in traditional European grammatical theory, is classified as participles, pronouns, relatives, demonstratives, and interrogatives.

B. Verb

The verb classification in Arabic is similar to that in English, although the tenses and aspects are different. The tag for the verb is sub-categorized into perfect, imperfect, and imperative. Further, sub-categorization of the verb class is possible using number, person, and gender.

C. Particle

The Particle class includes Prepositions, adverbs, conjunctions, interrogative particles, negative particle, quantifiers, communicators, determiners, and fillers.

Sometimes, it is difficult to decide to which part of speech a word belongs. Parts of speech should be clearly clarified, and the possible description of Egyptian Arabic is reviewed, as there is no previous work for part of speech in Egyptian

[1]A clitic: is a morpheme that has syntactic characteristics of a word, but shows evidence of being phonologically bound to another word. For example, in Arabic the definite article, equivalent to “the” in English, appears as a two-letter proclitic at the beginning of the noun.

Arabic. Thus, this is applied to the possible literature dealing with more examples of Egyptian Arabic word classes to enable us tag words. The researcher reviewed a lot of description for Egyptian Arabic words in [6], [7], [8], [9], [10], [11], [12], and [13] as well as the whole description of Egyptian Arabic and the classification and examples of words.

4 INSIGHTS INTO EGYPTION ARABIC MORPHOLOGICAL PARADIGMS

Arabic is the most widespread member of the Semitic group of languages. The Arabic language is the most complicated and richest language. This section presents an overview of the Egyptian colloquial Arabic morphological paradigms used in POS annotated data.The following sections present the morphological paradigms of Egyptian Arabic.

A. Noun

Arabic nouns are classified according to gender and number. Arabic nouns have two genders (masculine-feminine). Gender in Arabic is animatenouns, such as those referring to people, usually have the grammatical gender corresponding to their natural gender, but for inanimate nouns the grammatical gender is largely arbitrary. Most feminine nouns end in / -a/, such as cities, countries and certain body parts. Nouns that do not fit in any of these categories are masculine.

[11] classifies noun in Arabic into three categories: singular, dual, and plural. Singular noun is a base form, which dual or plural affixes are added to it. A dual noun is created by adding the suffix /-en/ to the stem or by adding number two before a noun. Plural nouns are sub-categorized into regular and irregular forms. Regular plurals are suffixes, /-in/ for masculine, such as /mudærris**i:n**/ 'teachers'/ and /-at/ for feminine, such as / ħæjæwæn**æ:t**/ 'animals'. Some nouns have both counted plural, such as /bed̲a:t̲/ 'eggs' and collective plural such as /be:d̲/ 'eggs'. Irregular plural "broken plural" is predicted in some nouns, such as /ko:ra/ 'ball' , /kowwar/ 'balls', and in other nouns is unpredicted, such as /ra:gel/ 'man', /riggæ:læ/ 'men'. When the noun is counted except for the dual form, the cardinal number precedes the noun in the noun phrase. Numerals 3 to 10 have two forms, long and short. The long form ends in /-a/ such as /**tælætæ** kilo/ 'three kilos'. The short forms end without /-a/ such as, /**tælæt**tuffæħæ:t/ 'three apples'. Numerals 11and above consist of a base which is an allomorph of numerals 1 and 2 and the suffix /-aʃar/ such as /ʔetn**aʃar**/ 'twelve'. Ordinal numbers tell the order of things in a set: first, second, third, such as /ʔettæ:ni / 'the second'.

Another type of nouns is a noun possessive. It is expressed by the word /bitæ:ʕ/ masculine 'belong', /bitæ:ʕæ/ feminine 'her', and /bitu:ʕ/ plural 'their'. It is the most common alternative to construct a phrase and indicate possession between two nouns such as /ʔekkitæ:bbitæ:ʕʔelbent/ 'the girl's book'. It is also used next to the suffix pronouns such as /bitæ:ʕu/ /ʔelʔælæmbitæ:ʕu/ 'the boy's pen', /bitæ:ʕhæ/ /ʔelʔælæmbitæ:ʕhæ/ 'the girl's pen'.

A proper noun is the special word or name that we use for a person, place, or country. A proper noun has two distinctive features: 1) it names a specific item, and 2) it begins with a capital letter. Nouns are tagged with n for common nouns, and **n:prop** for proper nouns (names of people, places, fictional characters, brand-name products).

1) Occupational Nouns:

The feminine of the most occupations is formed by adding /–a/ such as /mudærres/ 'male teacher', /mudærres**æ**/ 'female teacher'. Occupational nouns are tagged **n:occu|mudærres**

2) Place and Time Nouns:

Place and time nouns express the place or time of a verbal action or state. They are formed by prefixing /ma-/. For example, /**ma**t̲bax/ 'kitchen' (from /t̲abaxa/ 'to cook'), /mustæʃfæ/ 'hospital' (from the verb /istæʃfæ/ 'to cure'). Place and time nouns are tagged **n:plac|mustæʃfæ**.

3) Instrumental Nouns:

Instrumental nouns express the instrument by which the action is performed. They are prefixed with /mi-/ and formed only by verb form I, according to the following pattern. For example, /muftæ:ħ/ 'key' from /fætæħ/ 'to open'. Instrumental nouns are tagged **n:inst|muftæ:ħ**. Example of noun paradigm is shown in table 1.

TABLE I
PARADIGM OF NOUNS

<table>
<tr><td rowspan="2">Gender</td><td>Masc</td><td colspan="5">mudærris'teacher'</td></tr>
<tr><td>Fem Adding /-æ/</td><td colspan="5">mudærresæ'teacher'</td></tr>
<tr><td rowspan="4">Number</td><td>Singular</td><td>Dual /-in/</td><td colspan="4">Plural</td></tr>
<tr><td rowspan="3">ʔi:d 'hand'</td><td rowspan="3">ʔidi:n 'hands'</td><td colspan="2">Regular</td><td>Irregular</td><td>Collective</td></tr>
<tr><td>Masc/-in /</td><td>fællæ:ħin 'farmers'</td><td rowspan="2">ko:wwar 'balls'</td><td rowspan="2">so:kkar 'sugar'</td></tr>
<tr><td>Fem /–æt /</td><td>ʕarabı:jj-a:t̲ 'cars'</td></tr>
<tr><td rowspan="2">Numerals</td><td>11and above</td><td>ʔetna:ʃar 'twelve'</td><td>Possessive noun</td><td colspan="3">bitæ:ʕ 'belong'</td></tr>
<tr><td>Ordinal numbers</td><td>ʔettæ:ni 'The second'</td><td>1st possessive noun</td><td colspan="3">xæ:li 'my uncle'</td></tr>
<tr><td>Proper noun</td><td colspan="2">Farah, Sindebæ:d</td><td>Occupational nouns</td><td colspan="3">mudærresæ 'teacher'</td></tr>
<tr><td>Instrumental noun</td><td colspan="6">muftæ:ħ 'key'</td></tr>
</table>

B. Adjective

An adjective is a word that describes a noun. Adjectives are inflected for gender (masculine-feminine) and number (singular-plural). The masculine singular form of the adjective is the base form and is the stem to which feminine and plural affixes are added as mentioned in [11]. The suffix /-æ/ is added to the stem to form a feminine adjective. Adjectives are also inflected for plural by adding /-in/ /s̲uɣajjar**i:n**/ 'small'. The adjective is inflected for comparative by adding /ʔæ-/ such as /**ʔa**kbar/ 'older', and inflected for superlative as well by adding /ʔel-/ such as /**ʔil**akbar/ 'the oldest'. Adjectives follow the noun they modify and agree with singular nouns in gender and number. An adjective is tagged with **Adj.** An example of adjectiveparadigm is shown in table 2.

TABLE 2
PARADIGM OF ADJECTIVE

Gender	**Singular**	**Plural**
Masc	kibi:r'old'	kuba:r'old'
Fem	kibi:r-**æ**'old'	kuba:r'old'
Comparative	ʔakbar 'older'	
Superlative	ʔilakbar'the oldest'	

C. Determiner

Determiners include definite and indefinite articles. The definite article in Egyptian Arabic is /ʔel-/. It expresses the definite state of a noun of any gender and number. Definite article /ʔel-/ assimilated to a number of consonants, so the article in pronunciation is expressed only by geminating the initial consonant of the noun [8]. The gemination is expressed by putting /ʃæddæ/ on the following letters /t/, /ɵ/, /d/, /ð/, /r/, /z/, /s/, /ʃ/, /s̲/, /d̲/, /t̲/, /z̲/, /l/, /n/. The 14 letters are called "sun letters" while the remaining 14 are called "moon letters". Determiners are tagged **def:art:moonL|ʔel**. Example of definite and indefinite article paradigm is shown in the following table 3.

TABLE 3
PARADIGM OF DEFINITE ARTICLE

Definite article	**Example**
ʔel +Moon letters	ʔelħæflæ 'the party'
ʔe + gemination Sun letters	ʔeʃʃæ:rʕ 'the street'

D. pronouns

1) Personal subject-independent pronoun:

Personal pronouns in Egyptian Arabic have singular and plural, the second and third persons differentiate gender, while the first person does not. Personal pronouns are not needed with verbs, as it is clear from the verb, but it is common to use them, especially for emphasis. They are often used with participlesas stated in [7]. Personal pronouns are tagged **pron:subj:sg|ʔænæ**. Examples of paradigm of subject pronouns are shown in table 4.

TABLE 4
PARADIGM OF SUBJECT PRONOUN

Person		Singular	Plural
1st		ʔænæ 'I'	ʔiħna'we'
2nd	Masc	ʔæntæ'you'	ʔintu'you'
	Fem	ʔenti'you'	
3rd	Masc	huwwæ 'he'	humma'they'
	Fem	hijjæ'she'	

2) Possessive Objective Dependent Pronoun

Dependent personal pronouns in Egyptian Arabic are affixed to various parts of speech, with varying meanings. Egyptian Arabic object pronouns are clitics. They attach to the end of a noun, verb, or preposition, with the result forming a single phonological word rather than separate words. Personal pronouns are affixed to various parts of speech, with various meanings: Dependent personal pronouns are affixed to nouns, where they have the meaning of possessive demonstratives, e.g. /be:t**i**/ 'my house' , /be:tı**k**/ 'your house' , /be:t**u**/ 'his house'. They are affixed to verbs, where they have the meaning of direct object pronouns, e.g. /–ni/ 'me' / ʃu:fte**ni**/ ' saw me', /–k/ 'you' /ʃu:fte**k** 'saw you', /-hum/ 'them' /ʃu:ftu**hum**/ 'saw them'. With verbs, indirect object clitic pronouns are formed using the preposition /li-/ plus a clitic. Both direct and indirect object clitic pronouns can be attached to a single verb: /ʔægi:b/ 'I bring', /ʔægíbu/ 'I bring it', /ʔægibhu:lik/ 'I bring it to you', /mægibhulki:ʃ/ 'he did not bring it to you'. They are also affixed to prepositions, where they have the meaning of objects of the prepositions, e.g. /ʕændi/ 'to me', /ʕændek /'to you', /ʕændu/ 'to him'. Dependent personal pronouns are tagged **pron:dep|hæ**. Example of possessive/objective-dependent pronoun paradigm is shown in the following table 5.

TABLE 5
PARADIGM OF POSSESSIVE/OBJECTIVE –DEPENDENT PRONOUN

Direct object/Possessive				Indirect object	
Person		Pronoun	Example	Pronoun	Example
Singular					
1st		-i , ni	be:**ti** 'my house'	-li	gæ:bli 'brought me'
2nd	masc	-k -	be:t**æk** 'your house'	-læk	gæ:blæk 'brought you'
	fem	-ik-	be:t**ik** 'your house'	-lik	gæ:blik 'brought you'
3rd	masc	-u -	be:t**u** 'his house'	- lu	gæ:blu 'brought him'
	Fem	-hæ	be:t**hæ** 'her house'	-lhæ	gæ:blæhæ 'broughther'
Plural					
1st		-næ	be:t**næ** 'our house'	-lnæ	gæ:blenæ 'brought us'
2nd		-ku	be:t**ku** 'your house'	-lku	gæ:bleku 'brought you'
3rd		-hum	be:t**hum** 'their house'	-lhum	gæ:blhum 'brought them'

3) Pronouns with Suffixed Prepositions

A suffix pronoun is attached to prepositions, such as /fi/ 'in', /li-/ 'to', min/ 'from', /mæʕæ/ 'with', /ʕælæ/ 'on'. Pronouns with suffixed preposition are tagged **Prep|fi~Pro|hæ**. Examples of pronouns with suffixed prepositions paradigm are shown in table 6.

TABLE 6
PARADIGM OF PRONOUN WITH SUFFIXED PREPOSITIONS

Person		Pronoun	Pronouns with prepositions
1st		-jæ	lijæ'for me'
2nd	Masc	-k	li:k'for you'
	Fem	-ki	li:ki'for you'
3rd	Masc	-h	li:h 'for him'
	Fem	-hæ	li:hæ'for her'
Pl		-næ	li:næ'for us'
		-ku	li:ku'for you'
		-hum	li:hum'for them'

4) *Demonstrative Pronouns*

Demonstrative pronouns point to and identify a noun or a pronoun. Demonstrative pronouns are /dæ/ 'this, that', /di/ 'this, that ', and /do:l/ these, those'. They occur after the noun as demonstrative adjectives or before the noun as demonstrative pronouns. Other words also classified with demonstratives are /ʔæhu/ 'here is, there is', /ʔæhe:h/ 'here is, there is' , and /ʔæhum/ 'here are, there are' for dual and/or plural. They follow or precede the noun or occur in isolation. Demonstrative pronoun is tagged **dem|ʔæhu**. Examples of demonstrative pronoun paradigm are shown in the following table 7.

TABLE 7
PARADIGM OF DEMONSTRATIVE PRONOUN

Gender	Singular		Plural	
Masc	dæ	ʔerra:geldæ '**this** man' ʔelwælæddæħelw'**that** boy is handsome'	ʔæhum	ʔæhuʔaṣħa:bi '**there are** my friends'
	ʔæhu	ʔæhuʔelbe:t'**Here is** the house' ʔæhuʔelwælæd '**there is** the boy'	do:l	ʃo:ftʔelleʕæbdo:l 'I saw **these** toys' do:lʕarabijja:ṯ '**those** are cars'
Fem	di	ʔelbent di '**this** girl' ʔelbent di weħʃæ '**that** girl is bad'		
	ʔæhe:h	ʔæhe:hʔelħæjæwænæ:t '**here** are the animals'		

5) *Indefinite Pronouns*

In Egyptian Arabic indefinite pronouns are words like /ʔæjħædd/ 'anybody', /ħæ:gæ/ 'something'. In Egyptian, these made up of two words, but they used in exactly the same way as in English. Indefinite pronouns are tagged **Pron:indep|ħæ:gæ.** Examples of indefinite pronoun paradigm are shown in table 8.

TABLE 8
PARADIGM OF INDEFINITE PRONOUN

Indefinite pronouns	Example
Somebody	ħædd
Anybody	ʔæjħædd
Nobody	wælæħædd
Something	ħæ:gæ
Anything	wælæħæ:gæ
Nothing	wælæ

6) Relative Pronoun

The Egyptian Arabic has only one relative pronoun /ʔilli/ to represent 'that, who, and which'. There is only one relative pronoun used in reference to all nouns, regardless of gender/number. The relative pronoun is tagged **pron:rel|ʔilli**.

7) Interrogative pronouns

Egyptian Arabic pronouns indicate questions are /ʔe:hdæʔ/ 'What is this?', /mi:n/ ' who', /ʔezzæj/ 'how'. Interrogative pronouns are tagged **pro:wh|ʔe:h?.**

8) Reflexive Pronouns

The noun "næfs" is used as a reflexive pronoun followed by a suffix pronoun to mean that a person does an action by "himself". Egyptian Arabic reflexive pronouns are /næfsi/ 'myself ', /næfsæk/ 'yourself ', /næfsu/ 'himself'. Reflexive pronouns are used after a noun or a verb. Reflexive pronouns are tagged **Pron:ref|benefsu.**

E. Verb Tenses

[5] Classifies Egyptian Arabic into two basic tenses in Arabic. The "perfect "refers to a finished action, corresponds to the English past tense. The "imperfect" refers to an incomplete action (on going or future) and corresponds to our present, progressive, and future tenses. The imperfect is usually preceded by /bi-/ to denote present continuous and by /ħæ-/to denote the future tense. The imperative is used to give instructions or orders. There are three forms: masculine, feminine and plural. Examples of tenses paradigm are shown in the following table 9.

TABLE 9
PARADIGM OF VERB TENSE

Person		Past	Present imperfect	Present continuous	Future	Imperative
Singular						
1st		kætæbt 'I wrote'	ʔakætæb 'I write'	bæktib 'I 'm writing'	ħækteb 'I will write'	
2nd	masc	kætæbt 'you wrote'	tekætæb 'you write'	bitektib 'you are writing'	ħætíkteb 'you will write'	ʔiktib 'write'
	fem	kætæbti 'you wrote'	tekætæbi 'you write'	bitektíbi 'you are writing'	ħætiktebi 'you will write'	ʔiktibi 'write'
3rd	masc	kætæb 'he wrote'	jekætæb 'he writes'	bijektib 'he is writing'	ħæjíkteb 'he will write'	
	fem	kætæb-it 'she wrote'	tekætæb 'she writes'	bitektib 'she is writing'	ħætíkteb 'she will write'	
Plural						
1st		kætæbnæ 'we wrote'	nekætæb 'we write'	binektib 'we write'	ħæníkteb 'we will write'	
2nd		kætæbtu 'you wrote'	tekætæb 'you write'	bitektíbu 'you write'	ħætiktebu 'you will write'	ʔiktebu 'write'
3rd		kætæbu 'they wrote'	jekætæb 'they write'	bijektíbu 'they write'	ħæjiktebu 'they wilwrite'	

1) Voice participle

An Egyptian Arabic participle is derived from a verb, but is used like an adjective with the verbal meaning [8]. There are two types of participles: active and passive. Active voice is the "normal" way of using a verb; it has the form of an adjective or noun. Active participles act as adjectives, and so they must agree with their subject. There are three forms: masculine, feminine, and plural. Active participles are tagged **v:activ:partic|ʕæ:rf**. Passive participles, like active participles, act as adjectives or nouns, and so they must agree with the noun they're describing. Passive participles are tagged **v:pass:partic|mækto:b**. Examples of passive and active participles are shown in the following table 10.

TABLE II0
PARADIGM OF VOICE PARTICIPLE

Person	Active participle	Passive participle
m.sg	kæ:teb'writer'	ʔit-kætæb'was written'
f.sg	kæ:tbæ'writer'	ʔit-kætabt 'was written'
Pl	kæ:tbi:n'writer'	ʔit-kætabu'was written'

F. Negation

Negation in Egyptian Arabic appears in the free particles, such as /meʃ, læʔʔ, læ/ or negation bound prefix /mæ-/ and the suffix /–iʃ/. Negation is used with a verb, pronoun, adjective, and participles [11]. Negation is tagged **neg|læʔʔ**. Examples of negation paradigm are shown in table 11.

TABLE III1
PARADIGM OF NEGATION

Negation verb		Past	Present		Future	Imperative
Singular						
1st		mækætæbte∫ 'I did not write'	mækteb∫ 'I don't write'	mæbækteb∫ 'I am not writing'	me∫ħækteb 'I will not write'	
2nd	Mas	mækætæbte∫ 'he did not write'	mætekteb∫ 'you don't write'	mæbitekteb∫ 'you are not writing'	me∫ħætíkteb 'you will not write'	mæíktib∫ 'don't write'
	Fem	mækætæbi∫ 'she didn't write'	mætektebi:∫ 'you don't write'	mæbitektebi∫ 'you are not writing'	me∫ħætektebi 'you will not write'	mætiktíbi:∫ 'don't write'
3rd	Mas	mækætæb∫ 'he did not write'	mæjekteb∫ 'he doesn't write'	mæbijektib∫ 'he is not writing'	me∫ħæjíkteb 'he will not write'	
	Fem	mækætæbite∫ 'she didn't write'	mætekteb∫ 'she doesn't write'	mæbitekteb∫ 'she is not writing'	me∫ħætekteb 'she will not write'	
Plural						
1st		mækætæbnæ:∫ 'we didn't write'	mænekteb∫ 'we don't write'	mæbinekteb∫ 'we are not writing'	me∫ħænekteb 'we will not write'	
2nd		mækætæbtu:∫ 'they didn't write'	mætektebu:∫ 'you don't write'	mæbetektebu:∫ 'they are not writing'	me∫ħætektebu 'you will not write'	mætektíbu:∫ 'don't write'
3rd		mækætæbu:∫ 'they didn't write'	mæjektebu;∫ 'they don't write'	mæbejektebu;∫ 'they are not writing'	me∫ħæjektebu 'they will not write'	
neg, pron		me∫ 'not' [ʔænæ 'I' - ʔentæ 'you (mas)' - ʔenti 'you (fem)'- huwwæ 'he'- hijjæ 'she']				
neg, prep		mæfi:∫ 'there isn't' , mæʕhæ:∫ 'he hasn't got', mæʕændu:∫ 'he doesn't have'				
neg, adj		me∫ħelw 'not good'				
neg, parti		læ 'no' - læʔʔ 'no' - me∫ 'not'				
neg, bound		mæ- -∫				

G. Communicators

Communicators are used for interactive and communicative forms, which fulfill a variety of functions in speech and conversation. Many of these are formulaic expressions, such as ba:j 'bye', bravo, ∫okran 'thank you', ʔæhlæn 'welcome', sæ:læmoʕæleko 'hello'. Words used to express emotion, as well as imitative and onomatopoeic forms, such as "ʔah, boom, mhm, wow" are included in this category [13]. Communicators are tagged **co|ʔuh**.

H. Conjunctions

Conjunctions in Egyptian Arabic are the useful little words that join clauses together to make sentences that are more complex. Conjunctions conjoin two or more words, phrases, or sentences. A coordinating conjunction is a particle, which connects two words, phrases, or clauses together [5]. The most common conjunction is the prefixed particle /wæ/ 'and ', /fæ/ 'and so'. A coordinating conjunctions are tagged **conj:coo|wæ**. Subordinating conjunctions introduce a subordinate clause. Most subordinating conjunctions are single words, such as //bæss/ 'that's it', but, /zæj/ 'like', /bæʕd/ 'after', /ʔizæ/ 'if', /ʕæ∫æ:n/ 'because', /læmmæ/'when', /jebʔæ/ 'well then' , /wælæ/ 'or' , /bardu/ 'as well'. Subordinating conjunctions are tagged **conj:sub|ʕæ∫æ:n.**

I. Fillers

Fillers in Egyptian Arabic are a sound or word that is spoken in conversation by one participant to signal to others that he/she has paused to think, but has not yet finished speaking /jeʕni/ 'that means' and /wallahi/ 'A word used for swearing' are common fillers[7]. Fillers are tagged **fil|jeʕni.**

J. Quantifier

Quantifier in Egyptian Arabic is a word or phrase, which is used before a noun to indicate the amount or quantity. Quantifier is used with both countable and uncountable nouns, such as /kul/ 'all'[9]. Quantifier is **tagged qn|kul**.

K. Vocative Particle

The vocative particle /jæ/ is followed by a noun or proper noun for both genders [9].The vocative particle is tagged **Part:voc|jæ**.

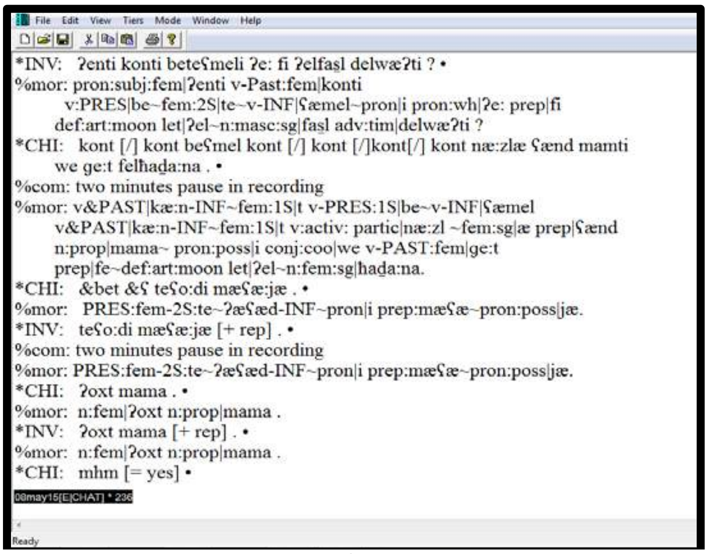


Figure 1: Transcribed File after Annotation Process.

5 METHOD

The second stage in building child corpus is POS coding process, which is the direct result of our previous transcription process[2]. POS are made in "%mor" 'morphology' tiers manually. We hand annotated one file, it approximately took 170 hours, and thus manual annotation focused on a particular child. Hand coding of a "%mor" tier for many children would require perhaps many years of work. The POS coding process started with a purely manually annotation of 2701words. 1380 annotated words for adult and 1321 annotated words for the child was handled.This initialEgyptian Arabic annotated corpus was used to run CLAN program for morphological analysis. The total number of the tagsets used in the data is 92 tags an example of the tagset was shown in table 12. CHAT codes were used with some adapting to fit the classification of Egyptian Arabic language. The morphological features applied to classify the words of the data were 92tagsets. The POS annotated corpus and the project are available at [14]. Following, ananalysis of the transcript as the application of CLAN program is overviewed. The commands applied in the data and analysis results are presented as well in the next section.

[2]Salama. H., Alansary, S (2014). Building a spoken Arabic corpus for Egyptian children: data collection and transcription.In *Proceedings of the Conference of language engineering*,*3*(4).Egyptian Society of Language Engineering.

TABLE IV2

EXAMPLE OF MORPHOLOGICAL TAGGING OF ARABIC

Class	**Examples**	**Coding of Examples**
adjective masculine	kibi:r 'old'	adj|kibi:r- MAS
adjective feminine	kibi:ræ 'old'	adj|kibi:r~fem|a
adjective regular plural	soɣajjarin'small'	adj| soghajjar~PL|in
adjective irregular plural	kuba:r'old'	adj| kubar~ir:PL
adjective, color (fem)	ħamra'red'	adj:col:fem|ħamra
adjective, color (mas)	ʔaħmar'red'	adj:col:mas|ʔaħmar
adjective, comparative	ʔakbar'older'	adj|ʔakbar- CP
adjective, superlative	ʔilʔakbar'the oldest'	adj|ʔilʔakbar –SP
adverb, locative	henæ 'here'	adv:loc|henæ

6 SOME FINDINGS from ANALYZING CHILD LANGUAGE TRANSCRPIT with CLAN PROGRAM

Analyzing child transcript is the final stage in building child corpus. Once a file is transcribed and annotated, the analytic work of CLAN is performed by a series of commands. These commands run from the Commands window, search for strings, and compute a variety of indices. CLAN allows the performance of a large number of analyses on transcript data; there are 29 programs inside the CLAN. The analyses include frequency counts, word searches, co-occurrence analyses, MLU counts, interaction analyses, and text changes. The CLAN programs are designed to support linguistic analysis [15]: morphological analysis, lexical analysis, syntactic analysis, discourse, and interactional analysis. The following lines review how these linguistic analyses perform in CLAN programs.

A. Morphological Analysis

Once a complete %mor tier is available, a vast range of morphological and syntactic analyses become possible. Many of the most important questions in child language require the detailed study of specific morphosyntactic features and constructions.

1) MLU

The MLU (mean length of utterance) is a command used primarily to determine the mean length of utterance of a specified speaker. It also provides the total number of utterances and of morphemes in a file. The ratio of morphemes over utterances (MLU) is derived from those two totals. [16]Manifests the value of thinking of MLU in terms of morphemes, rather than words. Brown is interested in the ways in which the acquisition of grammatical morphemes reflects syntactic growth and he believes that MLU in morphemes would reflect this growth more accurately than MLU. The output of the command**mlu +t*CHI farah.cha** perform MLU analysis on the child's tier (+t*CHI) is shown in Fig.1.The MLU for investigator output is:The total number of utterances is 308 and morphemes are 2459 in a file. The ratio of morphemes over utterances (MLU) is 7.984.Where the MLU for child is:The total number of utterances is 58 and morphemes are 2374 in a file. The ratio of morphemes over utterances (MLU) is 40.931 as shown in Fig.1.

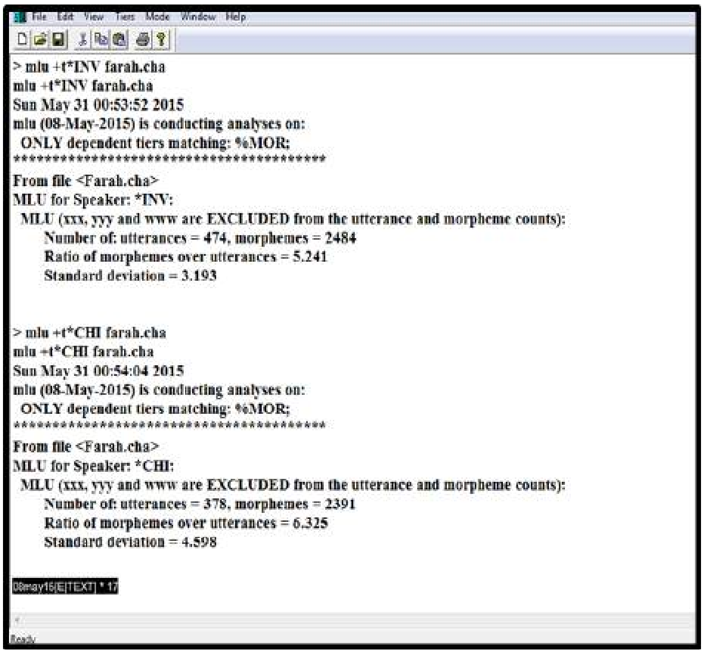


Figure 1: MLU analysis

B. Lexical Analysis

This is the easiest type of CLAN analyses, which looks at the frequencies and distributions of particular word forms. The programs for lexical analysis like FREQ (frequency) and KWAL (Key Word And Line) focus on the ways of searching for particular strings. The strings to be located can be entered in a command. Many studies used these techniques to track the development of lexical fields, such as morality, kinship, gender terminology, mental states, causative verbs, and modal auxiliaries. It is also possible to track words of a given length or a given lexical frequency. An example for FREQ and KWAL is clear in the following sections.

1) FREQ:

The FREQ (frequency) command is powerful and quite flexible, permitting frequency analysis. FREQ counts the frequencies of words used in selected files. It also calculates the type–token ratio typically used as a measure of lexical diversity. It generates an alphabetical list of all the words used by all speakers in a transcript indicating frequency of each word form (morpheme) and frequency of grammatical categories.A frequency word count is the calculation of the number of times a word occurs in a file or a set of files. FREQ produces a list of all the words used in the file, along with their frequency counts, and calculates a type–token ratio. The type–token ratio found by calculating the total number of unique words used by a selected speaker (or speakers) and dividing that number by the total number of words used by the same speaker(s). It is generally used as a rough measure of lexical diversity.The output of the command **freq +t*CHI farah.cha** shows how many times a child used the word. In the last output, it is a total of 1321 words or tokens used with only five different word types. The type–token ratio is found by dividing the total of unique words by the total of words spoken. For example, the type–token ratio would be 544 divided by 1321 or a ratio of 0.412as shown in Fig. 2.

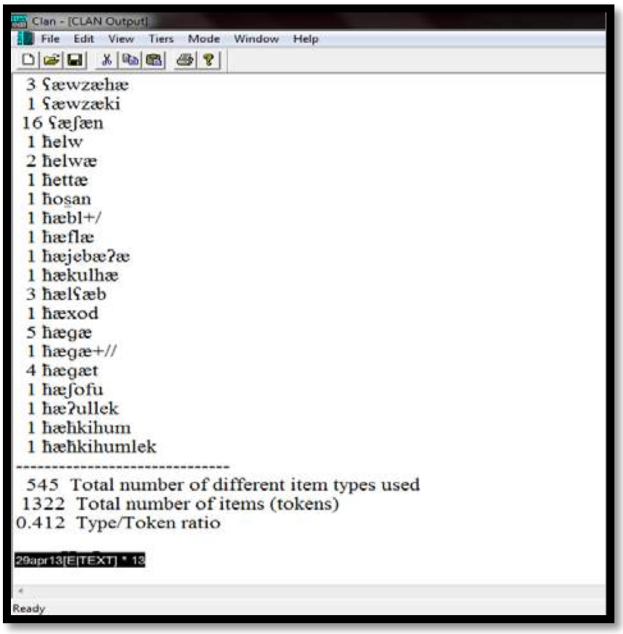


Figure 2: Frequency analysis

2) FREQPOS:

The FREQPOS (frequency position) program is a minor variant of freq. Freqpos is different in the fact that it allows us to track the frequencies of words in initial, final, and second position in an utterance. This is useful in studies of early child syntax. For example, using freqpos on the main line enables users to track the use of initial pronouns or auxiliaries. For an open class, an item such as verbs, freqpos is useful in analyzing codes on the %mor line. For example, freqpos allows studying the appearance of verbs in second position; initial position, final position, and other positions. The frequency position command **freqpos +d farah.cha** is shown in Fig.3.

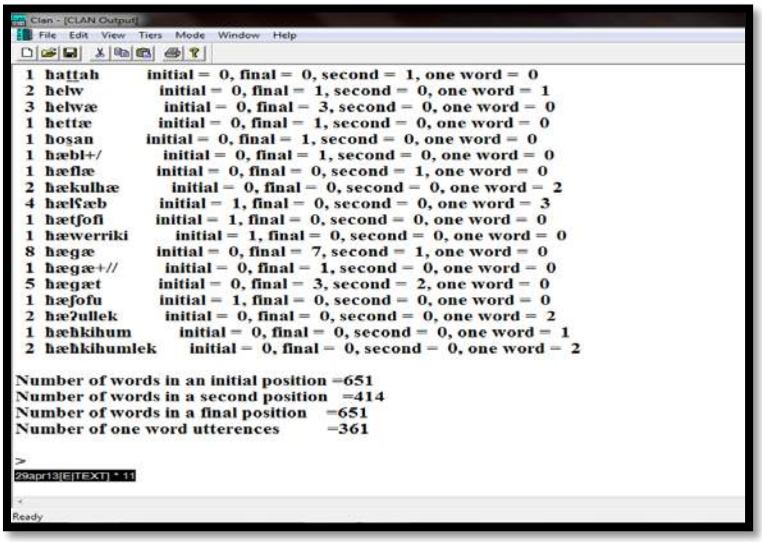


Figure 3: Freqpos analysis

3) KWAL:

KWAL is short for (Key Word And Line). It is the second major tool for conducting lexical analyses is the KWAL program. The analysis takes a word and finds the lines on which that word occurs in each transcript. This analysis is necessary to find out which lines the targets are on and in what position in the utterance each target is located. The outputs are not merely the frequencies of matching items, but also all the full context of the item. The KWAL command for the mother used the word /ʕæʃæ:n/ 'because'**kwal +sʕæʃæ:n -w2 +w2 farah.cha** is shown in Fig. 4. In this analysis, a mother used the word /ʕæʃæ:n/ 'because' nineteen times.

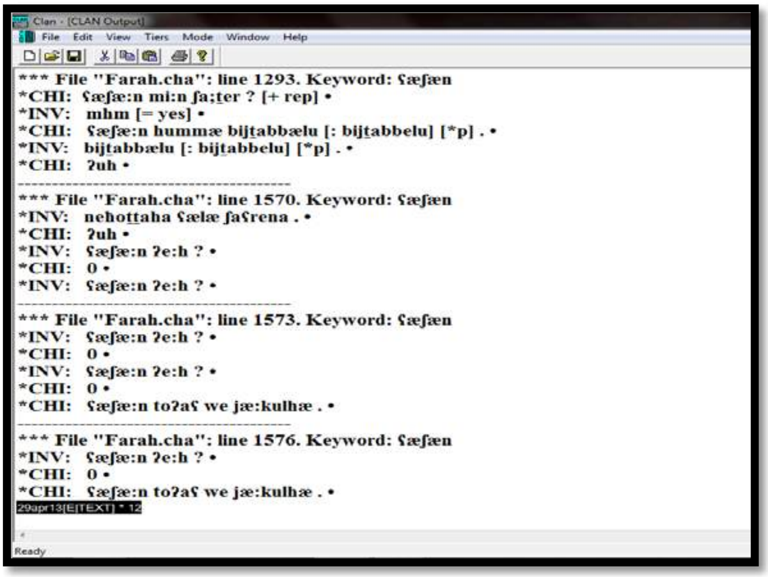


Figure 4: KWAL analysis

C. Syntactic Analysis

1) COMBO:

COMBO (combination) is a powerful program that searches the data for specified combinations of words or complex string patterns. For example, COMBO finds instances where a speaker says "beʕmelsels̠a:l" 'I am making dough' twice in a row within a single utterance. The command **combo +tCHI +s"beʕmel ^s̠els̠a:l" farah.cha** searches a child’s tiers (+t*CHI) of the specified file 0042.cha as in Fig.5. The output shows that the combination "beʕmels̠els̠a:l" 'I am making dough' is found once in the speaker’s speech as in shown in Fig.5.

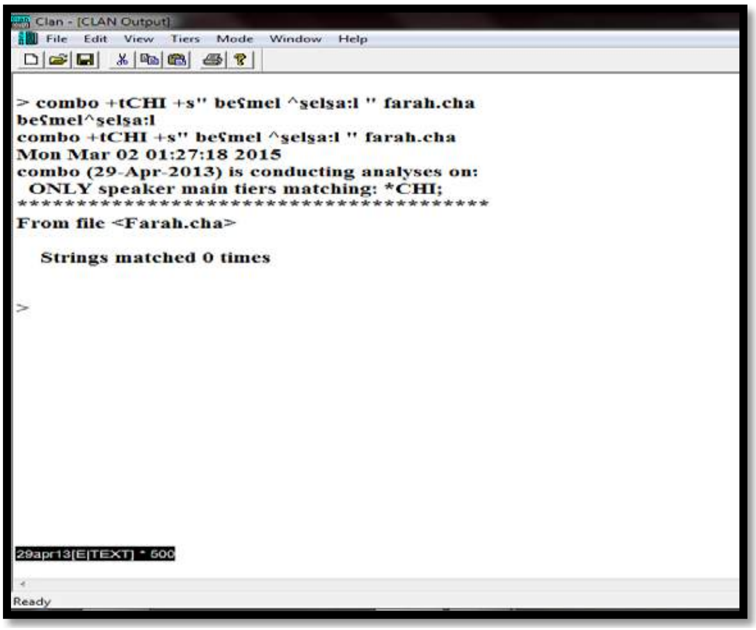


Figure 6: COMBO analysis

D. Discourse and Interactional Analysis

2) CHIP:

CHIP is useful for tracking the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. [17]Have used it successfully to demonstrate the availability of useful instructional feedback to a language-learning child. The program analyzes specified pairs of utterances. CHIP is used to explore parental input, the relation between speech acts and imitation, and individual differences in imitativeness in both normal and language-impaired children. CHIP compares two specified utterances and produces an analysis that then is inserted onto a new coding tier. The first utterance in the designated utterance pair is the "source" utterance and the second is the "response" utterance. The response compared to the source. An example of a minimal CHIP command **chip +bMOT +cCHIfarah.cha** is shown in Fig.6. The output of the first ten lines shows that CHIP introduces % csr tier. This tier is an analysis of the child's self-repetitions expressed by the code $REP. Here the child is both the source and the response as shown in Fig.6.

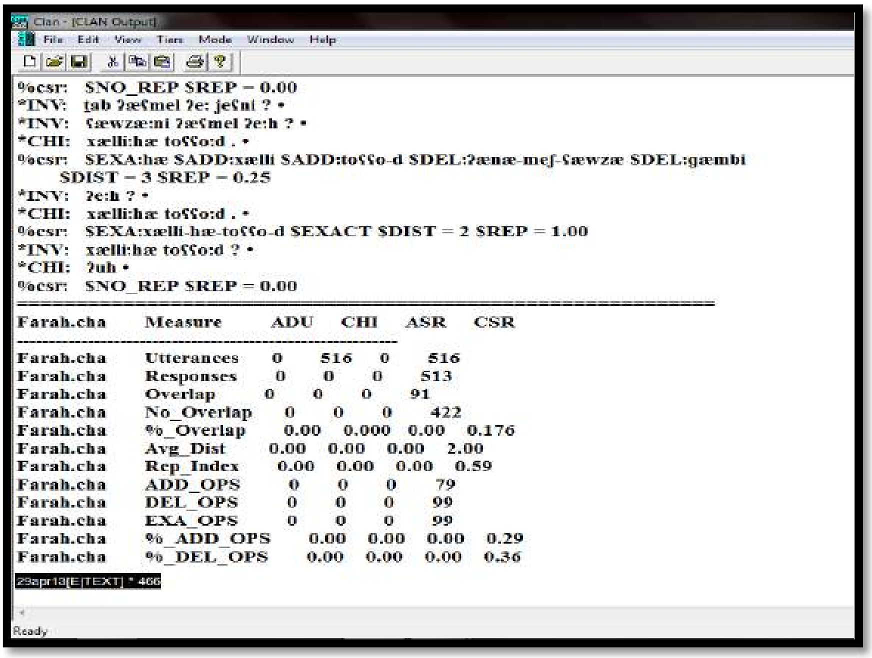
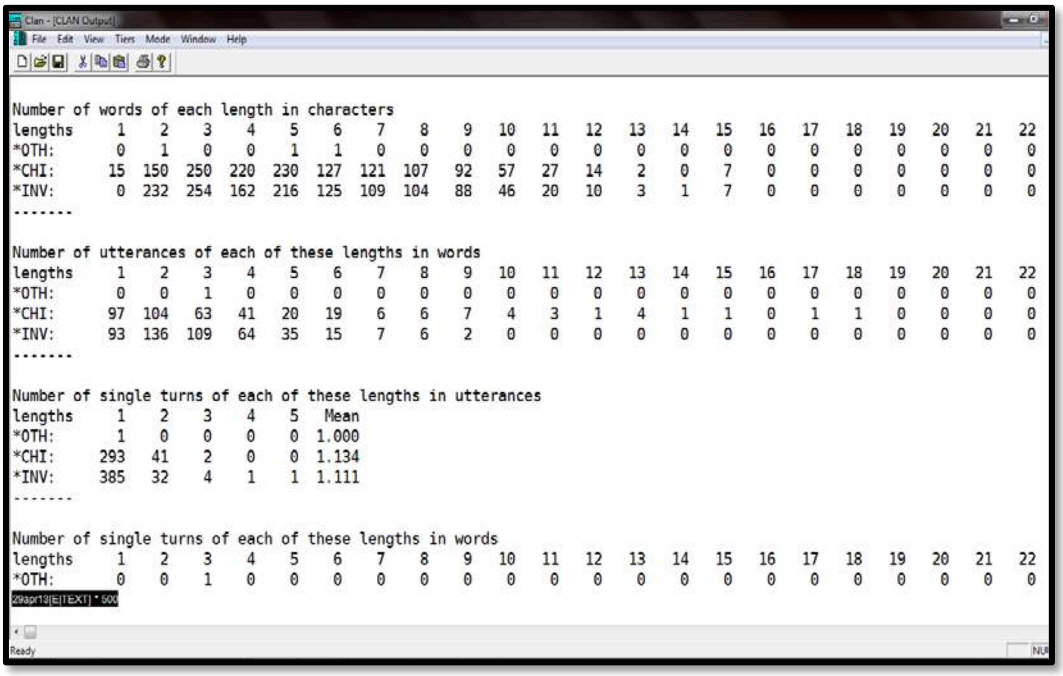


Figure 6: CHIP analysis

3) WDLEN:

The WDLEN (word length) program tabulates the lengths of words, utterances, and turns. The WDLEN program generates a histogram of maternal utterance lengths. It highlights the very high frequency of very short utterances that present language-learning children with either no or very few segmentation decisions in their efforts to locate words in the input. The command **wdlenfarah.cha** tabulates the lengths of words in child's tiers. The output shows that the investigator utterances consisted of zero single word as shown in Fig.7. An additional 230 are two words long, and an additional 255 are three words long. Thus, 485 words of child directed utterances in this analysis consist of investigator turns.



Number of words of each length in characters

lengths	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
*OTH:	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*CHI:	15	150	250	220	230	127	121	107	92	57	27	14	2	0	7	0	0	0	0	0	0	0
*INV:	0	232	254	162	216	125	109	104	88	46	20	10	3	1	7	0	0	0	0	0	0	0

Number of utterances of each of these lengths in words

lengths	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
*OTH:	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*CHI:	97	104	63	41	20	19	6	6	7	4	3	1	4	1	1	0	1	1	0	0	0	0
*INV:	93	136	109	64	35	15	7	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0

Number of single turns of each of these lengths in utterances

lengths	1	2	3	4	5	Mean
*OTH:	1	0	0	0	0	1.000
*CHI:	293	41	2	0	0	1.134
*INV:	385	32	4	1	1	1.111

Number of single turns of each of these lengths in words

lengths	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
*OTH:	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

29apr13[E|TEXT] * 600

Ready

Figure 7: WDLEN analysis

7 CONCLUSIONS

We introduced POS coding and analysis by using CLAN program and CHAT format 18]. Linguistic analysis performed by using CLAN commands. Seven types of linguistic analysis were applied as an application for CLAN program. The outputs of lexical analysis, such as FREQ and KWAL help to look at the frequencies and distributions of particular word forms. The output of MLU in morphological analysis helps the researchers to investigate the grammatical development of children. The syntactic analysis, such as COMBO searches the data for specified combinations of words or character strings. Moreover, the discourse and interactional analysis, such as CHIP track the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. This corpus is a research tool for future investigations of Egyptian Arabic child and child-directed language, language development, language disorder, and psycholinguistics in general.

In recent years, Corpora are considered basic resources for language analysis and research. There was a major shift towards the empirical study of language rather than intuitive study. The technological advance of computers changes the area of language research. This change of trend is because of the introduction of computer and corpora in linguistic research, which, subsequently, illuminated numerous new applications of language and linguistics in the field of information exchange. Moreover, the empirical approach to language study is distinguished to be more dependable and authentic than rationalistic approach, which is based on intuition. These corpora can be useful for producing many advanced automatic tools and systems, besides being good resources for language description and theory making. When child language is transcribed and compiled in a computerized database, it forms linguistic corpora. Corpora play an important role in child language research. The researchers of all theoretical persuasions make use of corpus data to investigate the development of children's linguistic knowledge. This is a high time to turn our attention towards using corpora for linguistic research. There are a lot of areas where corpora can lead to new perspectives in child language research, such as first language acquisition, second language learning, phonetic and prosodic analysis, and speech disorders.

REFERENCES

[1]McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*, Edinburgh: Edinburgh University Press.
[2] Mitchell, T. F. (1956). *An Introduction to Egyptian colloquial Arabic*. London: Oxford university press.
[3] Savoy, J. (1999). A stemming procedure and stop word list for general French corpora. *Journal of the American Society for Information Science*, 50, 944–952.
[4] Habash, N., &Rambow, O. (2005). Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In *Proceedings of the Conference of American Association for Computational Linguistics* (pp. 578-580).
[5]Haywood, J.A. and Nahmad, H.M. (1962). *A new Arabic grammar of the written language*. London: Lund Humphries.
[6]Hopkins, S. (1984). *Studies in the Grammar of Early Arabic: Based Upon Papyri Datable to Before 300 AH/912 AD*. Oxford University Press.
[7]Pipes, D. (1983). *An Arabist's Guide to Egyptian Colloquial*. Daniel Pipes.
[8] McGuirk, Russell H. (1986). *Colloquial Arabic of Egypt*.Psychology Press.
[9]Abu-Chacra, F. (2007). *Arabic: an essential grammar*. Routledge.
[10] McLoughlin, L. (2009). *Colloquial Arabic (Levantine)*.Routledge London and New York.
[11]Omar, M. (1970). The Acquisition of Egyptian Arabic as a Native Language, JanuaLinguarum, Series Practica, 160, The Hague : Mouton, 1973. (Formerly a Ph. D. dissertation at Georgetown University).
[12] Habash, N. Y. (2010). Introduction to Arabic natural language processing.*Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.
[13]Abdel-Massih, E. T. (2011). An Introduction to Egyptian Arabic.University of Michigan.
[14]project site http://eacc.ga
[15] Crystal, D., Fletcher, P., & Garman, M. (1989). *The grammatical analysis of languagedisability*. Second Edition, London: Cole and Whurr.
[16] Brown, R. (1973). *A first language: the early stages*.Cambridge, Mass.: Harvard University Press.
[17]Sokolov, J. L., &Moreton, J. (1994). Individual differences in linguistic imitativeness.In J. Sokolov& C. Snow (Eds.), *Handbook of research in language development using CHILDES*. Hillsdale, NJ: Lawrence Erlbaum Associates.
[18] MacWhinney, B. (2012). The CHILDS project.Tool for analyzing talk Electronic Edition.part 2 : the CLAN programs. Carnegie Mellon university available on line at http://childs.psy.cmu.edu/manuals/clan.

BIOGRAPHY

Heba Salama has a master's degree in corpus linguistics from the faculty of Arts phonetics and linguistics department Alexandria University 2015. She is interested in child language research. Her main interest is to collect corpus data to study child language development. She is searching for standard criteria to collect and transcribe data. She likes corpus linguistic because it is more methodology that is powerful, scientific and open objective verification of results. Electronic corpora have advantages, which is unavailable to their paper based equivalents. The availability of data exchange allows the researcher to answer questions by looking for the transcript of spontaneous speech of many data, rather than single study. Sharing data make a revolution in the study of child language. She found that the most obvious advantage of using computer for language study is the speed of processing and the ease of data manipulation.E.g., searching, sorting, and formatting. Advances in computer technology enable to share child language data more readily.The database is very important in helping the researcher to manage the problem they faced and wishes to test a detailed theoretical prediction on naturalistic samples.

TRANSLATED ABSTRACT

بناء مدونة لغوية محللة علي مستوي أقسام الكلام للاطفال المصريين

هبه سلامة[1],سامح الانصاري[2]
[1]ماجستيرفي المدونات اللغوية كلية الاداب- قسم الصوتيات واللغويات- جامعة اسكندرية
[2]أستاذ اللغويات الحاسوبية كلية الاداب- قسم الصوتيات واللغويات- جامعة اسكندرية

ملخص

تهدف الدراسة إلى عرض طريقة عنونه الكلمات وعرض تحليل للغة الطفل عن طريق برنامج CLAN . يعمل البحث على أقسام الكلام فيما يتعلق ببنية اللغة العربية المنطوقة لدى الأطفال. و إن الشرح اللغوى للمجموعات توفر للباحث وسائل أفضل للبحث فى التركيبات النحوية و استخدامها و تطويرها. يقوم البحث بعمل بعض التحليلات المرفولوجية مثل طول الجملة المنطوقة (MLU) و كذلك التحليلات اللفظية مثل عدد مرات التكرار (FREQ) والبحث عن كلمة معينة داخل السياق KWAL. هذا و إن بناء مدونة للأطفال قد ظهر مع وجود الثورة التكنولوجية و ثورة الحاسبات، و لقد قامت إثنتا و ثلاثون دولة حول العالم بعمل مدونة لغويه للأطفال تعتمد على قاعدة البيانات (CHILDES). أما بالنسبة للدول العربية، فلقد قامت كل من قطر و الإمارات بعمل مدونة خاصة بهما، و

قامتا بعرض المدونتين على مواقع الإنترنت، إلا أن المدونة الخاصة بالعربية المصرية لم تكن متوفرة بعد. وقد قام هذا البحث بعمل اول مدونة لغوية عربية منطوقة للاطفال المصريين وعرضها علي الانترنت من اجل الاسهام في تبادل المعلومات بين الباحثين. كما يفيد ايضا فى مجال علم اللغة النفسى و البحث فى التركيبات النحوية و كذلك فى التحليل اللغوى. كما أن البحث التجريبي يمكن أن يعرفنا الكثير عن الاضطرابات اللغوية التي تحدث للاطفال ومن ثم سرعة اكتشافها وعلاجها مبكرا، كما أننا بحاجة إلى البحث فى كيفية تفاعل الطفل واستخدامه للغة فى المواقف العادية. فنحن بحاجة إلى ملاحظة و تسجيل و تحليل النماذج اللغوية التلقائية، إلا أن دراسة تلك النماذج التلقائية يتطلب وقت كبير فى جمع البيانات و الكتابة الصوتية و التحليل ، و من ثم فعمل مدونة مصرية للأطفال يسهل عملية تحليل كلام الاطفال و يساعد فى دراسة لغة الأطفال. و لقد أحدث مشروع نظام تبادل البيانات اللغوية للأطفال (CHILDES) تغيرات جذرية فى طرق البحث على العديد من المستويات (الصوتية و التركيبية و اللفظية)، فهذا المشروع مبادرة لجمع البيانات للكتابة الصوتية من مختلف الدراسات التى أجريت على لغة الأطفال وفقا لصيغة CHAT و باستخدام برنامج CLAN.

Automatic Part-of-Speech Tagging of Arabic-English Dictionary Senses through WordNet

Diaa M. Fayed[*1], Aly A. Fahmy[*2], Mohsen A. Rashwan[**3,] Wafaa K. Fayed[***4]

[*]*Computer Science, Faculty of Computers and Information, Giza, Egypt*
[1]diaa.fayed@outlook.com
[2]a.fahmy@fci-cu.edu.eg

[**]*EECE, Faculty of Engineering, Giza, Egypt*
[3]mrashwan@ieee.org

[***]*Arabic Language and Literatures, Faculty of Arts, Giza, Egypt*
[4]wafkamel@link.net

***Abstract*—This paper proposed an algorithm for part-of-speech (POS) tagging senses of a bilingual dictionary. The algorithm is applied on the Al-Mawrid Arabic-English dictionary. The tagging task is accomplished by transferring the POS tags of the English translation equivalences (TEs) to the dictionary senses after dis-ambiguities process. The English POS tags of senses are acquired from the Princeton WordNet. POS tagging of bilingual dictionary senses is prerequisite to link a bilingual dictionary to WordNet and/or standardizing that dictionary into WordNet-LMF format where the synset (set of synonyms), not word, is the basic brick. The registered accuracy is high though the cost is little. Building NLP/HLT tools needs linguistic experts, large investments, and long time. For statistical approach, we need large annotated corpora and for rule-based approach, we need large lexicon that contains rich linguistic and world knowledge. That motivates the appearance of what are called resource-light approaches to develop natural language processing (NLP) tools for poor-resource languages.**

1 INTRODUCTION

Vast researches and investments were made for Latin languages specially English in the field of natural language processing (NLP) and human language technology (HLT). The results of these are much amount of resources and tools such are lexicons, thesauruses, annotated corpora, morphological analyzers, syntactic parsers, etc. On the other hand, other languages, such as Arabic, are poor of those resources and tools. Some researches such as parallel text processing try to benefit from resources and tools that are built for Latin languages to build resources and tools for other poor-resources languages[1-11]. Yarowsky and Ngai [12]stated that we can overcome on resource shortage problem of some languages by leveraging the annotated data and tools for resource-rich languages (such English, French and Japanese).

Feldman [7]summarized resource-light approaches to NLP tasks as unsupervised or minimally supervised approaches and cross-language knowledge induction. Instances of the former approach are unsupervised POS tagging and minimally supervised morphology learning. Instances of the latter approach are cross-language knowledge transfer using parallel texts, bilingual lexicon acquisition, and cross-language knowledge transfer without parallel corpora.

Annotated language sources such as corpora and dictionaries are required in both HLT and NLP. The annotations are any information that augmented to text so as to make computer to either understand the text or used in training. Annotating includes syntactic and semantic annotations. Manning [13] defined the part-of-speech (POS) tagging as “the task of labeling (or tagging) each word in a sentence with its appropriate part of speech; we decide whether each word is a noun, verb, adjective, or whatever”. .Part-of-speech (POS) tagging is to assign one or more POS tags such as noun, verb, adjective, etc. to a lemma or synset (set of synonyms).

In this study, we consider the Arabic-English Al-Mawrid [14] dictionary as a parallel corpus of Arabic and English. The Al-Mawrid is not nearly annotated by any part-of-speech tags as stated by Fayed et al. [15, 16]. This study will exploit the translation equivalences (TEs) on the English side of the dictionary to assign part-of-speech tags to the Al-Mawrid senses. Assigning POS tags to senses of a bilingual lexicon is required in both HLT and NLP application. Furthermore, this step is required before translation of the English WordNet, linking a bilingual lexicon to the WordNet, or standardization of bilingual dictionary into WordNet-LMF.

The POS tagging task is composed of two steps. First, use the translation equivalences (TEs) of a sense and get the POS tags of them from the WordNet. Then, intersect the sets to acquire the most probable POS tags. The idea of this disambiguation process is that the most common POS tags among POS tags of TE are the most probable ones that represent the POS tags of a sense.

The contributions of this paper are:
- Proposing an independent-language algorithm that can be used to POS tag senses of any bilingual dictionary. This requires a repository of POS tags of the target language.
- Implementing and applying the algorithm on the Arabic-English Al-Mawrid lexicon.

2 STRUCTURES OF DATA SOURCES

A. Al-Mawrid

The Arabic-English Al-Mawrid dictionary is a general-purpose dictionary. The headwords of the Al-Mawrid are arranged alphabetically according to the first letters. An entry of the Al-Mawrid starts by a bold headword. When a headword has more than one meaning or sense, each meaning occupies a subentry that is cited in separated lines. Subentries that contain collocations, idioms, terms, or examples are cited lately. A subentry consists of a mandatory Arabic section and an optional English section. The Arabic section consists of three fields that we name header, explanation, and cross-reference. The header is optional but the explanation and cross-reference are optional. A colon separates the header from its explanation. A dash may precede the cross-reference field. A subentry may express declaration, question, or exclamation [14, 15].

The three fields of an Arabic section have the same structure: each field consists of one or more words or phrases that are separated by either an Arabic comma or the conjunction word "أو". Comma-separated phrases are almost synonymous or near synonyms. The header has the headword of an entry if its subentry represents a sense of the headword. If a subentry does not represent a sense of the headword, it contains collocations, idioms, terms, or examples. The morphologic, syntactic, or semantic information is scattered in the header or explanation fields. The English section has one or more translation equivalence groups (TEGs) that are separated by semicolons. Each TEG has one or more translation equivalence (TE) phrases that are separated by comma. The phrases of a TEG are synonyms. The Al-Mawrid is not annotated by part-of-speech (POS) tags. Only very low number of two part-of-speech tags (approximately fifteen tags of nouns and adjectives) exists [14, 15]. Fig. 1 illustrates the microstructure of the Al-Mawrid.

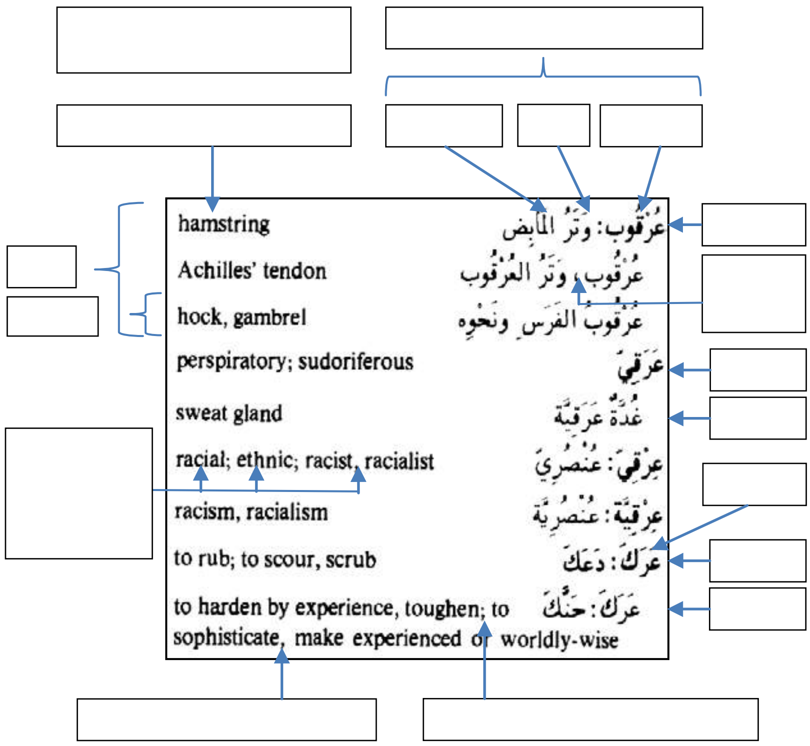


Figure 1 The micro structure of the Al-Mawrid

B. WordNet

WordNet [17] is a lexical system composed of an on-line English lexical database and software utilities. English concepts are organized into set of synonyms (synsets). Each synset is composed of words or phrases and is associated with glosses and illustrative examples. The database is divided into four categories: nouns, verbs, adjectives, and adverbs. Synsets are linked to other synsets by lexical and semantic relations. Fig. 2 and Fig. 3 show format of the WordNet sysnset[18] and examples respectively.Examples are excerpted from data.noun file and data.adj file of the WordNet database.

synset_offset lex_filenum ss_type w_cnt word lex_id [word lex_id...] p_cnt [ptr...] [frames...] | gloss

synset_offset
Current byte offset in the file represented as an 8 digit decimal integer.

ss_type
One character code indicating the synset type:
n NOUN
v VERB
a ADJECTIVE
s ADJECTIVE SATELLITE
r ADVERB

word
ASCII form of a word as entered in the synset by the lexicographer, with spaces replaced by underscore characters (_).

gloss
Each synset contains a gloss. A *gloss* is represented as a vertical bar (|), followed by a text string that continues until the end of the line. The gloss may contain a definition, one or more example sentences, or both

Figure 2 The format of the WordNet synset

10502950 18**n** 02 **racist** 0 **racialist** 0 004 @ 09853645 n 0000 + 01155044 n 0202 + 06203758 n 0101 + 01155044 n 0101 | **a person with a prejudiced belief that one race is superior to others**

01928283 00 **s** 01 **racist** 0 001 & 01927654 a 0000 | **based on racial intolerance; "racist remarks"**

00285905 00 **s** 03 **racist** 0 **antiblack** 0 **anti-Semite**(a) 0 002 & 00285148 a 0000 + 09797742 n 0301 | **discriminatory especially on the basis of race or religion**

Figure 3 Examples of synsets

The WordNet API search uses morphy function that preprocesses the searched string before looking up the database files. The preprocessing includes exceptional lists, morphological rules, collocations, hyphenations, etc. Table 1 contains suffixes that the morphy function uses to process the input string. For more explanation on the WordNet search function, see[18].

TABLE-1
RULES OF DETACHMENT

POS	Suffix	Ending	POS	Suffix	Ending
NOUN	"s"	""	VERB	"es"	"e"
NOUN	"ses"	"s"	VERB	"es"	""
NOUN	"xes"	"x"	VERB	"ed"	"e"
NOUN	"zes"	"z"	VERB	"ed"	""
NOUN	"ches”	"ch"	VERB	"ing"	"e"
NOUN	"shes”	"sh"	VERB	"ing"	""
NOUN	"men”	"man"	ADJ	"er"	""
NOUN	"ies"	"y"	ADJ	"est"	""
VERB	"s"	""	ADJ	"er"	"e"
VERB	"ies"	"y"	ADJ	"est"	"e"

3 TAGGING ALGORITHM

Fig. 4 illustrates the general components of the proposed POS tagger. The main components are bilingual dictionary, monolingual dictionary or morphological analyzer of the second or target language, and POS disambiguator.

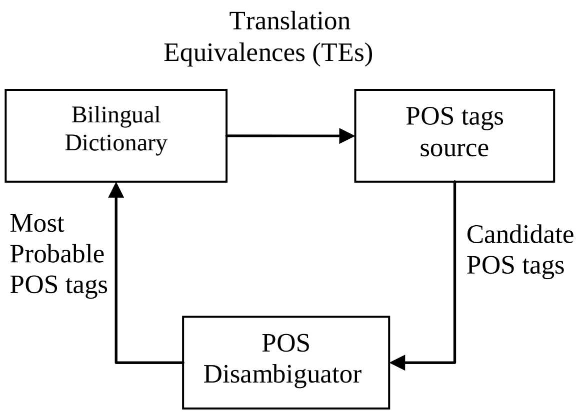


Figure 4 Diagram of POS Tagging algorithm

Fig. 5 contains the pseudo code for the proposed POS tagging algorithm.Table 2 contains illustrative examples for the algorithm.

```
For each headword (HW) of a bilingual dictionary
    For each subentry sense (S) of HW
        Get the second language (SL) translation equivalences (TEs)
            For each TE word or phrase
                Get the part-of-speech (POS) tags set by consulting
                the SL dictionary or a SL morphological analyzer.
        Intersect the POS tag sets of all TEs
        If the result is not empty
            Put tags set equal to the result of intersection
        Else
            Put tags set equal to the most frequent POS tags
```

Figure 5 Pseudo code Algorithm of Part-of-speech Tagging

We POS annotate senses of the Arabic-English Al-Mawrid dictionary by projecting the Tags from the English section to the Arabic section. The task accomplished by locking up WordNet database via the translation equivalence phrases, and then using a disambiguating method in the case of existing ambiguity in the POS tags. The disambiguation is simply accomplished by intersecting the sets of POS tags to get the common tag of sets. If the results of intersection are empty, the most frequent tag/tags will be the candidate POS tag of a sense. Table 2 contains illustrative examples for the algorithm.

TABLE-2
ILLUSTRATIVE EXAMPLE FOR ALGORITHM

Mawrid	Intersection	WordNet
عَرَقِيّ perspiratory sudoriferous	$\{\emptyset\} \cap \{\emptyset\} = \emptyset$	… …
غُدَّةٌ عَرِقَيَّة sweat gland	$\{n\}$	(n) sweat gland, sudoriferous gland (any of the glands in the skin that secrete perspiration)
عِرْقِيّ: عُنْصُرِيّ racial	$\{a, a\} \rightarrow \{2a\}$ $\{a\}$	(a) racial (of or related to genetically distinguished groups of people) (a) racial (of or characteristic of race or races or arising from differences among groups)
عِرْقِيّ: عُنْصُرِيّ ethnic	$\{n, a, a\} \rightarrow \{n, 2a\}$ $\{a\}$	(n) ethnic (a person who is a member of an ethnic group) (a) cultural, ethnic, ethnical (denoting or deriving from or distinctive of the ways of living built up by a group of people) (a) heathen, heathenish, pagan, ethnic (not acknowledging the God of Christianity and Judaism and Islam)
عِرْقِيّ: عُنْصُرِيّ racist	$\{n, a, a\} \cap \{n\} = \{n\}$	(n) racist, racialist (a person with a prejudiced belief that one racial group is superior to others) (a) racist (based on racial intolerance) (a) racist, antiblack, anti-Semite (discriminatory especially on the basis of race or religion)
عِرْقِيّ: عُنْصُرِيّ racialist		(n) racist, racialist (a person with a prejudiced belief that one racial group is superior to others)
عِرْقِيَّة: عُنْصُرِيَّة racism	$\{n, n\} \cap \{n\} = \{n\}$	(n) racism (the prejudice that members of one race are intrinsically superior to members of other races) (n) racism, racialism, racial discrimination (discriminatory or abusive behavior towards members of another race)
عِرْقِيَّة: عُنْصُرِيَّة racialism		(n) racism, racialism, racial discrimination (discriminatory or abusive behavior towards members of another race)

4 Experimental Setup and Evaluation

A. Dataset and Tools

We used the chapter Ayn "ع" of the Arabic-English Al-Mawrid dictionary [14] to evaluate the proposed algorithm that disambiguates part-of-speech tagging. The definitions of the Al-Mawrid are structured following the method of Diaa et al. [16]. In addition, we used the Princeton WordNet 3.0 [19, 20] as a source of part-of-speech tags of senses. We implemented the proposed algorithm in python and used the WordNet database that is implemented in Natural Language Toolkit (NLTK)[21].
In addition to preprocessing of the Al-Mawrid data in Diaa et al. [15, 16], we made additional preprocessing to the translation equivalences before used them in querying the WordNet API. Table 3 shows some of those modifications.

Table -3
Examples of preprocessing

Sense	TE set
عَجَّلَ: حَثَّ على العَجَلَة to hurry, rush, urge, impel, press	{ hurry, rush, urge, impel, press }
عَجَّجَ: أثَارَ (الغُبَارَ) to raise, swirl up (the dust)	{raise, swirl up}
العَجَلَةُ والجُزْع wheel and axle	{ wheel and axle, wheel-and-axle, wheel_and_axle, wheelandaxle }

B. Experiments

The WordNet API search functions make some morphological processing on the query word or phrase. We make some modifications on the WordNet interface and utilities codes. In each subsequent experiment, we augmented the steps by further modifications or adaptations on the previous experiment. Table 4 shows some modifications that can improve the accuracy of the algorithm.

Table-4
Modifications to improve the accuracy of the algorithm

Experiment	Modification
1	• Suppress all the exception lookup. • Suppress all the functions of morphology. • Make the following preprocessing for the translation equivalences (TEs) phrases: – remove "to " from beginning phrases that defining verbs, – remove all parentheses, – replace inner space in the collocations by score, underscore, space, and nothing.
2	Allow the morphology for plurals and apply rules in the morphology function. See **Table 1** for "Noun".
3	Set up the "Verb" tag as for any TE phrase starting by "to ".
4	Set up "Noun" as the default POS tag when the result of the algorithm is empty. The reason of that is the most frequent word class in any dictionary is the noun, making default POS tag will prevent empty results and increase coverage.

C. Evaluation Metrics and Evaluation Procedure

We used the precision, recall, and F-measure metrics [22, 23] to evaluate the proposed POS tagging algorithm. The definitions of the metrics are in equations (1), (2), and (3).

$$Precision = \frac{number of correct POS tags in tagged data}{number of correct POS tags in golden data} (1)$$

$$Recall = \frac{number of correct POS tags in tagged data}{number of total POS tags in tagged data} (2)$$

$$F = \frac{Precision \times Recall}{Precision + Recall} (3)$$

The evaluation procedure as following:

- Define a set of part-of-speech tags {noun, verb, adj, adv, phi} to define the senses of the dictionary. The first four senses are the POS tags of the WordNet. Phi is a POS tag for the undefined POS tags of senses. Examples of entries that take the Phi are sentences, verbal phrases, etc.
- first the senses of the Ayn chapter of the Al-Mawrid is tagged manually -as golden standards for evaluation,
- then the same chapter is tagged automatically following the proposed algorithm,
- finally, the *person* and *recall* are computed according to formulas.

5 RESULTS AND DISCUSSION

Table 5 contains the results of the four experiments. The first experiment is considered the base-line of the proposed algorithm. The values of precision and recall are moderate for the base-line experiment. The preci-sion and recall increased slightly by the experiment 2. However, in the experiment 3 and experiment 4, the values of precision and recall are increased dramatically.

TABLE -5
EVALUATION RESULTS

Experiment	Precision	Recall	F
1	71.58	74.67	36.55
2	72.30	75.18	36.86
3	89.12	87.84	44.24
4	93.10	87.36	45.07

The lesson of the previous results is that we can exploit the characteristics of the bilingual dictionary to increase the accuracy and coverage of the baseline algorithm.

6 RELATED WORKS

Our work is inspired by Pianta et al.[24] who developed an aligned multilingual database for the Italian language. They designed an assigning procedure that takes an input sense of an Italian word of the Italian-to-English section of the Collins dictionary and outputs a set of English candidate senses arranged by confidence scores. The confidence scores are computed based on a group of linking rule and each rule participates in the final score by weighted quantity. The Synset intersection is the linking rule that inspires our work. The synset intersection rule is based on the fact that TGRs may have multiple TEs which are synonymous. We can use other TEs to disambiguate an ambiguity of a TE. The rule takes different sets of candidates of TEs and intersects them. The candidates that are in the intersection get a partial confidence score.

Cucerzan and Yarowsky [25] bootstrapped a multilingual POS tagger using (1) an online or hard-copy pock-et-sized bilingual dictionary, (2) a basic library reference grammar, and (3) access to an existing monolingual text corpus in the language. As one step in the bootstrapping the POS tagger, they extracted a preliminary POS distributions from an untagged monolingual translation lists. For a given English translation word ei in the translation list (TL), the prior POS distribution probabilities are estimated from a large and balanced corpus. The combination of the Brown and WSJ corpora are used.

7 Conclusions and Future Work

In this work, we proposed an independent-language algorithm that POS tags senses of a bilingual dictionary by using the translation equivalences of the target language and monolingual repository of senses. The algorithm is composed of two main steps: acquiring the sets of part-of-speech tags and then intersect them to get the POS tags that are common. We applied the algorithm on the Al-Mawrid Arabic-English dictionary and WordNet.

In future work, we will use more than one resource for part-of-speech tags will increase the accuracy and coverage. We plane also to link the senses of the A-Mawrid to the WordNet.

REFERENCES

[1] D. Yarowsky, G. Ngai, and R. Wicentowski, "Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora," in *Proceedings of the first international conference on Human language technology research*, 2001, pp. 1-8.

[2] B. Cavestro and N. Cancedda, "Literality Based Sample Sorting for Syntax Projection," 2005.

[3] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak, "Bootstrapping Parsers via Syntactic Projection Across Parallel Texts," *Natural language engineering*, vol. 11, pp. 311-326, 2005.

[4] V. B. Mititelu and R. Ion, "Cross-Language Transfer of Syntactic Relations using Parallel Corpora," in *Cross-Language Knowledge Induction Workshop, Romania*, 2005.

[5] V. B. Mititelu and R. Ion, "Automatic Import of Verbal Syntactic Relations using Parallel Corpora," in *Cross-Language Knowledge Induction Workshop*, 2005.

[6] S. Padó and M. Lapata, "Cross-Linguistic Projection of Role-Semantic Information," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 859-866.

[7] A. Feldman, "Portable Language Technology: A Resource-light Approach to Morpho-syntactic Tagging," Citeseer, 2006.

[8] A. Feldman, J. Hana, and C. Brew, "Experiments in Cross-Language Morphological Annotation Transfer," in *Computational Linguistics and Intelligent Text Processing*, ed: Springer, 2006, pp. 41-50.

[9] A. Feldman, J. Hana, and C. Brew, "A Cross-Language Approach to Rapid Creation of New Morpho-Syntactically Annotated Resources," *Gen*, vol. 115, pp. 6-6, 2006.

[10] A. Feldman and J. Hana, *A Resource-Light Approach to Morpho-Syntactic Tagging*: Rodopi, 2010.

[11] T. D. Szymanski, "Morphological Inference from Bitext for Resource-Poor Languages," The University of Michigan, 2011.

[12] D. Yarowsky and G. Ngai, "Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora," 2001.

[13] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*: MIT press, 1999.

[14] R. Baalbaki, "Al-Mawrid: A Modern Arabic-English Dictionary," 18 ed. Beirut, Lebanon: Dar El-Elm Lilmalayin, 2004.

[15] D. M. Fayed, A. A. Fahmy, M. A. Rashwan, and W. K. Fayed, "Extracting Knowledge from an Arabic-English Machine-Readable Dictionary Using Information Extraction," presented at the 5th International Conference on Arabic Language Processing (CITALA 2014), Oujda, Morocco, 2014.

[16] D. M. Fayed, A. A. Fahmy, M. A. Rashwan, and W. K. Fayed, "Towards Structuring an Arabic-English Machine-Readable Dictionary Using Parsing Expression Grammars," *International Journal of Computational Linguistics Research*, vol. 5, pp. 1-13, 2014.

[17] G. A. Miller, "Five papers on WordNet," *Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University*, 1993.

[18] *WordNet 3.0 Reference Manua*:http://wordnet.princeton.edu/wordnet/documentation/, (accessed 23 October 2015)

[19] *WordNet*: http://wordnet.princeton.edu/, (accessed 23 October 2015)

[20] *WordNet 3.0 Database*: https://wordnet.princeton.edu/wordnet/download/current-version/, (accessed 23 October 2015)

[21] *Natural Language Toolkit (NLTK)*: http://www.nltk.org/, (accessed 23 October 2015)

[22] C. Hagerman and ク. ヘガマン, "Evaluating the Performance of Automated Part-of-Speech Taggers on an L2 Corpus," 2012.

[23] S. Thouësny, "Modeling second language learners’ interlanguage and its variability," Dublin City University, 2011.

[24] E. Pianta, L. Bentivogli, and C. Girardi, "MultiWordNet: Developing an Aligned Multilingual Database," in *Proc. 1st Int’l Conference on Global WordNet*, 2002.

[25] S. Cucerzan and D. Yarowsky, "Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day," in *proceedings of the 6th conference on Natural language learning-Volume 20*, 2002, pp. 1-7.

Diaa El-Din Mohamed Abo-Fayed received the B.E. degree in the Electronics, Faculty of Engineering, Mansoura University, 1995. He received M.Sc. degree in the Automatic Control from the Faculty of Engineering, Mansoura University, 2001, in the field of “Data Mining”. Now Diaa is a PhD Candidate in the computer science in the Faculty of Computers and Information, Cairo University; the PhD is in the field of Arabic NLP.Diaa worked as a software research engineer in the COLTEC ME company in building HLT for Arabic. Diaa also worked in the News Group Co. as software research engineer. The News Group Co. specialize in the sourcing, distribution, creation, monitoring and analysis of news content in the emerging markets of the Middle East, Africa and the Indian sub-continent. Currently,Diaa is acommunications engineer in the National Water Research Center.

Aly Aly Fahmy is the former Dean of the Faculty of Computers and Information, Cairo University and a Professor of AI and ML, in the Department of Computer Science. He graduated from the Department of Computer Engineering, Technical College with honor degree. He specialized in Mathematical Logic. He received a master’s degree from the ENSAE, Toulouse, France, 1976 in the field of Logical DB systems and then obtained his PhD from the Center for Studies and Research – CERT-DERI, 1979, Toulouse – France in the field of Artificial Intelligence. He participated in several national projects, built expert systems, and published many papers and books. Prof. Fahmy’s main research areas are: Data and Text Mining, Mathematical Logic, Computational Linguistics, Text Understanding and Automatic Essay Scoring and Technologies of Man–Machine Interface in Arabic. He Directed the first Center of Excellence in Egypt in the field of Data Mining and Computer Modeling (DMCM). Prof. Aly Fahmy is currently involved in the implementation of the exploration project of Master’s and Doctorate theses of Cairo University.

Mohsen Abdelrazek Rashwan received the B.Sc. and M.Sc. degrees in electronics and electrical communications from the Faculty of Engineering, Cairo University, Cairo, Egypt. The M.Sc. degree in systems and computer engineering from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree in electronics and electrical communications from Queen’s University, Kingston, ON, Canada. He currently serves as a Professor in the Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, Cairo Egypt, and as the Managing Director of the Engineering Company for the development of Computer Systems that cofounded in 1993. Over the past research as well as building commercial products of Arabic NLP, digital speech processing, image processing, OCR, and e-learning. Among other several national and international mega projects, he has seved/been serving as a Senior Scientist in the EC’s FP7 projects on Arabic HLT NEMLAR and MEDAR , and as a Co-PI in the Egyptian Data Mining and Computer Modeling Center of Excellence (DMCM-CoE).

WafaaKamel Fayed B.A., M.A., Ph.D., Arabic Department, Faculty of Arts, Cairo University. Professor, Arabic Department - Faculty of Arts, Cairo University. Correspondent member, Arabic Language Academy, Damascus. Expert , Arabic Language Academy , Cairo.Supervisor of 49 Ph.D. and M.A. Degrees. Award of Ideal Professor at Cairo University. Honorary Certificate of distinction, 2004. Cairo University's Incentive award of human and social sciences, 2004. The 1st thesis using computer in applications of Arabic studies at 1974.Cairo University's appreciation award of humanities and pedagogic sciences 2013. Published 58 academic researchers at universal journals. The author of 7 academic books and translator of 2 books.

توسيم معاني معجم عربي-إنجليزي بأنواع الكلمات بطريقة آلية

ضياء الدين محمد محمد أبوفايد*[1]، علي علي فهمي*[1]، محسن عبد الرازق رشوان**[2]، وفاء كامل فايد***[3]

كلية الحاسبات والمعلومات، قسم علوم الحاسب، جامعة القاهرة، الجيزة، مصر*
[1]diaa.fayed@outlook.com
[2]a.fahmy@fci-cu.edu.eg
كلية الهندسة، قسم الإلكترونيات والإتصالات الكهربية، جامعة القاهرة، الجيزة، مصر**
[3]mrashwan@ieee.org
كلية الآداب، قسم اللغة العربية وآدابها، جامعة القاهرة، الجيزة، مصر***
[4]wafkamel@link.net

ملخص

يقترح البحث خوارزمية لتوسيم معاني معجم عربي-انجليزي بأقسام الكلام بطريقة آلية. طبقت الخوارزمية على معجم المورد. تتم عملية التوسيم بنقل أقسام الكلام لمكافئات الترجمة الإنجليزية إلى معاني المعجم بعد عملية إزالة اللبس. يأخذ الخوارزم التوسيمات لمكافئات الترجمة من الوردنت. نحتاج توسيم معاني معجم عند ربط المعجم بالوردنت أو تحويل المعجم للصيغة القياسية WordNet-LMF حيث تكون مجموعة المعاني هي وحدة بناء المعجم وليس الكلمة.

بناء أدوات لعمل تطبيقات معالجة اللغات الطبيعية يتطلب خبراء واستثمارات ضخمة وفترات زمنية طويلة. فاذا كانت المقاربات المستخدمة احصائية فان ذلك يحتاج لذخائر لغوية ضخمة وموسمة؛ واذا كان المقاربات المستخدمة قاعدية، فان ذلك يحتاج لمعاجم ضخمة غنية بالمعلمات اللغوية. هذه المتطلبات المكلفة أدت لبزوغ مايسمى بالمقاربات المخفضة المصادر للغات الفقيرة في هذه المصادر لبناء أدوات تستخدم في تطبيقات معاجة اللغات الطبيعية.

Developing an Approach for Solving Ambiguity in Requirements Specification to UML Conversion

Somaia Osama[*1], Safia Abbas[*2], MostafaAref[*3]

[*]*Computer Science Department, Faculty of Computer and Information Science,Ain Shams University*
Cairo, Egypt

[1]somaia.osama.r@gmail.com

[2]safia_abbas@yahoo.com

[3]aref_99@yahoo.com

Abstract-- **Requirements Engineering is one of the most essentialactivities in the Software Development Life Cycle. The success of the software is mostly dependent on how well the users' requirements have been understood and converted into suitable functionalities in the software. Usually, the users express their requirements in natural language statements that initially appear easy to represent. However, being represented in natural language, the statement of requirements often tends to suffer from ambiguities. Ambiguity is a critical issue in the software requirement specifications. Ambiguity occurs when different readers can interpret a sentence differently. The proposed work is aimed to detect and resolve the ambiguity and find the UML components and the relationship between them to generate the UML Diagram. The tool helps analysts by providing an efficient and fast way to produce the accurate the UML diagram from their requirements. A case study has been solved to show that the use of tool in automated ambiguity detection.**

Key words: Natural language processing (NLP), ambiguity, Requirement engineering, Software Requirement Specification, Unified Modeling Language.

1 INTRODUCTION

Requirements engineeringis the activity that involves the functions associated with the extraction, modeling, analysis, verification and specification of the user's requirements [1]. The RE activity often starts with the vaguely defined requirements [2] and results finally in to a Software Requirements Specification (SRS) document. The SRS is a part of the contract and it must define the user and the system requirements obviously, accurately and unambiguously. An SRS that has inconspicuous, incomplete, unmanaged, unspecified, inaccurate or ambiguous requirement definition may eventually lead to cost and time overruns [3, 4, 5]. An important research problem in Requirements Engineering is resolving ambiguity. An ambiguity is "a statement having more than one meaning". An ambiguity can be lexical, *syntactic, semantic,pragmatic, vagueness, generality and language error* ambiguity [6]. Although the fact that the requirements specified in natural language tend to inappropriate interpretations, the requirements are most often specified in natural language. So, it is necessary to develop the approaches that deal with resolving the ambiguities from the user requirement specifications.Manually resolving ambiguity from software requirements is a tedious, time-consuming, error-prone, and therefore expensive process [6]. Therefore, an automated and semi-automated approach to resolve ambiguities from the requirements statement is needed.There exist various approaches, starting from manual glossaries approach to automatic ontology based approach to reduce ambiguity from the Software Requirement Specification. In addition, there are a number of diverse tools such as, QuaARS[7], RESI [8], WSD [9], SREE [10,11], ARM [12], NAI [13, 14], and NL2OCL [15], SR-Elicitor [16]developed to detect and resolve ambiguities.

2 AMBIGUITY

"An important term, phrase, or sentence essential to an understanding of system behavior has either been left undefined or defined in a way that can cause confusion and misunderstanding." [17].Ambiguous requirements lead to confusion, wasted effort and time and rework. Ambiguity is the possibility to interpret a phrase/word in several ways. It is one of the problems that occur in natural language texts. An empirical study by Kamsties et al [6] depicts that "Ambiguities are misinterpreted more often than other types of defects". An ambiguity has two sources: incomplete information and communication mistakes. Some errors can be resolved without domain knowledge like grammatical error though some error needs domain knowledge like the lack of detail that wants user. The Ambiguity Handbook [6] presents different types of ambiguities, categorized as Lexical, Syntactic, Semantic, Pragmatic, Vagueness, Generality and Language Error.

TABLE I. TYPES OF AMBIGUITY [6]

Type of Ambiguity	Subtype	Description with example
Lexical Ambiguity	Homonymy Ambiguity	Two different words have the same written and phonetic representation, but unrelated meanings and different etymologies. E.g.: The airport shall be a major hub for Departures from Australia to Asia. "major" (important/an army officer of high rank/ specialize in a particular subject at a college)
	Polysemy Ambiguity	A word has several related meanings but one etymology.
Syntactic Ambiguity	Analytical Ambiguity	The role of the constituents within a phrase or sentence is ambiguous. E.g.:The software will follow the applicable regulatory and utility technical requirements in its speculated calculations and selection process.(can refer regulatory technical requirements and utility technical requirements or regulatory requirements and utility technical requirements)
	Attachment Ambiguity	A particular syntactic constituent of a sentence, such as a prepositional phrase or a relative clause, can be legally attached to two parts of a sentence. Or a phrase can be placed in different positions in the parse tree.
	Elliptical Ambiguity	When it is not certain whether or not a sentence contains an ellipsis.
	Coordination Ambiguity	More constituents joined by coordinative conjunctions (and, or). E.g.: The system shall print a login session report to every Manager and Database Administrator. (can refer The system shall print a login session report to very Manager and every Database Administrator or The system shall print a login session report to every person who is both a Manager and a Database Administrator.
Semantic Ambiguity	Scope Ambiguity	A sentence has more than one way of reading it within its context although it contains no lexical or structural ambiguity.
Pragmatic Ambiguity	Referential Ambiguity	An anaphor can take its reference from more than one element, each playing the role of the antecedent. E.g.: If the ATM accepts the card, the user enters the PIN. If not, the card is rejected.
	Deictic Ambiguity	Pronouns, time and place adverbs, such as now and here, and other grammatical features, such as tense, have more than one reference point in the context. The context includes a person in a conversation, a particular location, a particular instance of time, or an expression in a previous or following sentence.
Vagueness		If it is not clear how to measure whether the requirement is fulfilled or not. E.g.: The System shall be easy as possible.

3 THE PROPOSED ARCHITECTURE OF OUR TOOL

A proposed automated tool aims to generate accurate and complete UML diagrams from the Natural Language specification (NLS). So we will develop an automated system detect and remove ambiguities from full text documents. Figure 1 shows the system design architecture. The initial input is a complete requirements document. The output is UML diagrams. Our tool consists of four modules viz. Text Preprocessing module, Ambiguity detection and removal module, UML generation module.

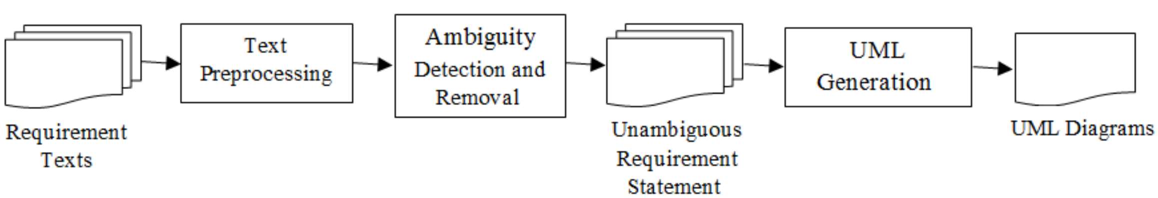


Fig. 1 System Architecture

A. Text Preprocessing Module

Given a text document as input, our tool first executes several text preprocessing steps, including sentence splitting, part-of speech(POS) tagging, and produce parse tree. At first, the text is split into a set of sentence. Then, for each sentence, the Stanford parser is used to obtain POS tags (e.g., noun, verb, adjective, adverb, etc.) of individual words.

E.g.:**"The system provides maximum output."**

After POS Tagging

"**"The/DT system/NN provides/VBZ maximum/JJ output/NN ./."** "

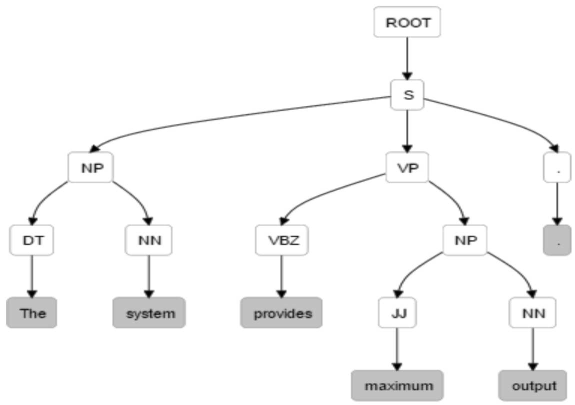


Fig. 2 Parse Tree

The text is syntactically analyzed and a parse tree is producedfor further semantic analysis. Figure 2 shows the generated parse tree of the above example.

B. Ambiguity Detection and removal Module.

A tool could apply a several ambiguity measures to a requirement specification to recognize possibly ambiguous sentences in the requirement specification. The core goals for this tool for detecting and measuring ambiguities in natural language requirement specification are: to detect which sentences in a natural language requirement specification are ambiguous and, for each ambiguous sentence, resolve the ambiguity from the sentence, and consequently improve the natural language requirement specification.

a) Detect the Ambiguity.

Corpus is the main element of ambiguity detection. Ambiguous words that result in misinterpreted requirements are analyzed and stored into the corpus. The major aim of this process is to check and validate whether the data which is a part of Software Requirements Specification document is ambiguous or not.

1. Identify Referential Ambiguity

The Referential corpus contains the possible ambiguity indicators: I, it, its, itself,he,she, her, hers, herself, him, himself, his,me, mine, most, my, myself, that, their, theirs, them, themselves, these,they, you, your, yours, yourself, and yourselves,anyone, anybody, anything, everyone, everybody, everything, nobody, none, no one, nothing, our, ours, ourselves, someone, somebody, something, this, those, us, we, what, whatever, which, whichever, who, whoever, whom, whomever, whose, and whosever.

2. Identify Coordination Ambiguity

The Coordination corpus contains the possible ambiguity indicators: and, and/or, or, but, unless, if then, if and only if, and also.

3. Identify Scope Ambiguity

The Scope corpus contains the possible ambiguity indicators: a, all, any, few, little, several, many, much, each, not, and some.

4. Identify Vague

The *Vague* corpus contains the possible ambiguity indicators: /, <>, ( ), [ ], { }, ;, ?, !, adaptability, additionally, adequate, aggregate, also, ancillary, arbitrary, appropriate, as appropriate, available, as far as, at last, as few as possible, as little as possible, as many as possible, as much as possible, as required, as well as, bad, both, but, but also, but not limited to, capable of, capable to, capability of, capability, common, correctly, consistent, contemporary, convenient, credible, custom, customary, default, definable, easily, easy, effective, efficient, episodic, equitable, equitably, eventually, exist, exists, expeditiously, fast, fair, fairly, finally, frequently, full, general, generic, good, high-level, impartially, infrequently, insignificant, intermediate, interactive, in terms of, less, lightweight, logical, low-level, maximum, minimum, more, mutually-agreed, mutually-exclusive, mutually-inclusive, near, necessary, neutral, not only, only, on the fly, particular, physical, powerful, practical, prompt, provided, quickly, random, recent, regardless of, relevant, respective, robust, routine, sufficiently, sequential, significant, simple, specific, strong, there, there is, transient, transparent, timely, undefinable, understandable, unless, unnecessary, useful, various, and varying[10].

C. *Extraction using heuristics module:* Finally, This section focuses in heuristics and their application to develop the generation of object oriented concepts from natural language texts. Usually, candidate classes can be detected by determining the noun phrases in the text of the requirements. Candidate relationships can be found in the same way by determining verb phrases, with the UML diagrams being presented to the user as the final step.

4 CONCLUSIONS

One of the most essential stages of software development is requirement gathering. Rest of the project depends on this step i.e. how requirements are understood, collected and described. If requirements are not correctly understood, or software requirements specification is not correctly designed, then the result will be ambiguous software requirements specification document. Ambiguities in software requirements specification presents conflicts in the software project, as different interpretations can be stated by team members while understanding requirements, which finally affect the quality of system to be develop. One way to resolve this problem is to detect and resolve ambiguities early, in the requirement analysis stage. So our tool is designed that finds ambiguities in software requirements specification document and resolve it.The future work, our tool will extract the objectoriented information from softwarespecification requirements such as classes, instances and their respective attributes, operations, associations, aggregations, and generalizations to enhancethe text analysis process to generate UML diagrams like use-case,activity diagram, collaboration diagram and sequence diagram.

REFERENCES

[1] Sommerville, I. and Sawyer, P. 1997."Requirements Engineering, A good practice guide". Chichester: John Wiley & Sons Ltd.

[2] Nuseibeh, B., & Easterbrook, S. 2000, May. "Requirements engineering: a roadmap". *In Proceedings of the Conference on the Future of Software Engineering* (pp. 35-46).

[3] Belev, G. C. 1989, January. "Guidelines for specification development". In Reliability and Maintainability Symposium, 1989.Proceedings., Annual (pp. 15-21). IEEE.

[4] Christel, M. G., & Kang, K. C. 1992. Issues in requirements elicitation, (No.CMU/SEI-92-TR-12). CARNEGIE-MELLON UNIV., PITTSBURGH, PA, SOFTWARE ENGINEERING INST.

[5] Donald G. Firesmith. 2007. Common Requirements Problems, Their Negative Consequences, and Industry Best Practices to Help Solve Them. In Journal of Object Technology, vol. 6, no. 1, January-February 2007, pp. 17-33

[6] Berry, D.M., Kamsties, E., Krieger, M.M.: "From contract drafting to software specification: Linguistic sources of ambiguity," http://se.uwaterloo.ca/~dberry/handbook/ambiguityHandbook.pdf, 2003.

[7] Fabbrini, F., M. Fusani, S. Gnesi, and G. Lami. 2001. The Linguistic Approach to the Natural Language Requirements Quality: Benefit of the use of an Automatic Tool. SEW'01, Proceeding of the 26th annual NASA Goddard Software Engineering Workshop, IEEE Computer Society Washington, DC, USA, 97.

[8] Sven Körner and Torben Brumm, 2009, RESI-A natural language specification improver. IEEE International Conference on Semantic Computing (ICSC).

[9] Nancy Ide and Jean Véronis, 1998. Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics - Special issue on word sense disambiguation, Volume 24, Issue 1, PP. 2-40.

[10] Sri Fatimah Tjong, 2008. Avoiding ambiguity in requirements specifications. Thesis submitted to the University of Nottinghamfor the degree of Doctor of Philosophy.

[11] Tjong, Sri Fatimah, and Daniel M. Berry. 2013. The Design of SREE—A Prototype Potential Ambiguity Finder for RequirementsSpecifications and Lessons Learned. Requirements Engineering:Foundation for Software Quality. Springer Berlin Heidelberg, 2013. PP.80-95.
[12] Willis, Alistair, Francis Chantree, and Anne De Roeck. 2008. Automatic Identification of Nocuous Ambiguity. Research on Language & Computation, 6 (3-4), 1-23.
[13] Hui Yang, Alistair Willis, Anne De Roeck, Bashar Nuseibeh. 2010. Automatic Detection of Nocuous Coordination Ambiguities in Natural Language Requirements. Proceedings of the IEEE/ACMinternational conference on Automated software engineering, 53-62. ISBN: 978-1-4503-0116-9. DOI=10.1145/1858996.1859007.
[14] Hui Yang, Anne de Roeck , Vincenzo Gervasi, Alistair Willis Bashar Nuseibeh. 2011. Analyzing anaphoric ambiguity in natural language requirements. Requirements Engineering - Special Issueon Best Papers of RE'10: Requirements Engineering in a MultifacetedWorld, Volume 16 Issue 3, 163-189. DOI=10.1007/s00766-011-0119 y.
[15] Imran Sarwar Bajwa. 2012. Resolving Syntactic Ambiguities in Natural Language Specification of Constraints. CICLing'12 Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing, Volume 1, 178-187.
[16] Basili, Victor R., Scott Green, Oliver Laitenberger, Filippo Lanubile, Forrest Shull, Sivert Sorumgard. 1995. The Empirical Investigation of Perspective-Based Reading. Technical report theempirical investigation of perspective based reading.
[17] Gracia, Jorge; Lopez, Vanessa; d"Aquin, Mathieu; Sabou, Marta; Motta, Enrico and Mena, Eduardo, "Solving semantic ambiguity to improve semantic web based ontology matching," in The 2nd International Workshop on Ontology Matching, Busan, South Korea, 2007.

BIOGRAPHY

Somaia Osama: She graduated from faculty of Computer Science in 2009 at Akhabr El Yom Academy, Cairo, Egypt. She started working as a teaching assistant in the Computer Science department at Akhabr El Yom Academy since Sept 2009 till now. Then she got a diploma in software architect from Information Technology Institute, Smart Village, Egypt in 2011.

Dr. Safia Abbas: She received his Ph.D. (2010) in Computer science from Nigata University, Japan, her M.Sc. (2003) and B.Sc.(1998) in computer science from Ain Shams University, Egypt. Her research interests include data mining argumentation, intelligent computing, and artificial intelligent. She has published around 15 papers in refereed journals and conference proceedings in these areas which DBLP and springer indexing. She was honored for the international publication from the Ain Shams University president.

Mostafa Aref is a professor of Computer Science and Vice Dean for Graduate studies and Research, Ain Shams University, Cairo, Egypt. Ph. D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio.M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. of Electrical Engineering - Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.

وضع نهج لحل الغموض في متطلبات المواصفات لتحويلها لرسم تخطيطي

سمية اسامة، صفية عباس، مصطفى عارف
قسم علوم الحاسب، كلية الحاسبات والمعلومات، جامعة عين شمس

ملخص
هندسةالمتطلبات هي واحدة من أكثرالأنشطة الحيوية في دورة تطوير البرمجيات. نجاح النظام يعتمدإلى حد كبيرعلى مدى تمفهم متطلبات المستخدمين وتحويلهاإلى وظائف مناسبة في البرنامج. عادة،للمستخدمين التعبيرعن احتياجاتهم في تصريحات باللغة الطبيعية التي تظهرسهلة التعامل في البداية. ومع ذلك، مع استخدام اللغة الطبيعية، ينتج بيان متطلبات غالبا مايميل للمعاناة من الغموض. والغموض هومشكلة خطيرة في تحديد مواصفات متطلبات البرمجيات. يحدث التباس عند مختلف القراء فكل قارئ له تفسيرمختلف. لذلك يهدف هذا العمل المقترح لكشف وحل الغموض والعثورعلى مكونات النظام والعلاقة بينهما لتوليد الرسم التخطيطي. فهو أداة تساعد المحللين من خلال توفير وسيلة فعالة وسريعة لإنتاج رسم تخطيطي دقيق لاحتياجاتهم. و قد تم عمل دراسة حالةلاستخدام أداة الكشف الآلي من الغموض.

Case Based Reasoning of Semantic Knowledge on Medical System

Passent ElKafrawy[*1], Rania A. Mohamed[**2]

[*]*Mathematics and CS Department, Faculty of Science, Menofia University*
ShebinElkom Menofia, Egypt
[1]basant.elkafrawi@science.menofia.edu.eg
[**]*Faculty Computer Science, Modern University for Technology & Information*
Cairo, Egypt
[2]rania.a.mohamed@gmail.com

***Abstract*— This paper presents a new approach to Case-Based Reasoning (CBR) using Semantic knowledge (SCBR) for the representation of cases, case structure, and case based ontologies in biology and medicine. The approach could be extended to other application domains of CBR. The major advantage of such approach is that Semantic data systems are designed to understand the content of the real world as accurately as possible within the data set. This paper also makes a comparison between traditional CBR and SCBR where there are some problems in traditional CBR such as adaptation may be difficult; cases may need to be created by hand; large processing time to find similar cases; and CBR systems generally give good or reasonable solutions, this is because the retrieved case often requires adaptation. SCBR framework can handle these problems.**

Keywords: —Case-Based Reasoning, Ontology, Semantic Knowledge.

1 INTRODUCTION

The importance of the medical field in today's life cannot be underestimated simply because there seems to be a continuing advancement in the complexity and severity of many diagnosed medical maladies. The medical field is the scientific discipline that deals with finding cure for every conceivable type of illness and disease so this paper use case based reasoning in medical field to help doctors diagnose diseases, to find the appropriate treatment for the patient and to analyze causes and/or treatments.

Case-based Reasoning is an emerging field in Artificial intelligence. The common application areas of CBR includes help-desk and customer service, recommender systems in electronic commerce, knowledge and experience management, medical applications and applications in image processing, applications in law, technical diagnosis, design, planning and applications in the computer games and music domain. Case-based reasoning is an approach, which utilizes the experience gained from past solved problems [1]. This approach maintains all information of past problems solved (i.e. experience) that is called the case. The collection of all these past experiences is stored in a form of case based. There are various factors which define the efficiency of this approach [2]. The major factor is that a solution to a new problem is projected from the number of past experiences stored in the case based. A new problem should be matched to the closest problem of past experiences faced. The new upcoming problem is considered as a new case. The strategies of finding a similar case for the new case regarding past cases stored in the case based is another major factor of defining the efficiency of the case-based reasoning approach.

Case-based reasoning systems have some drawbacks such as: occupies a large storage space for all the cases, take large processing time to find similar cases in case-based, cases may need to be created by hand, adaptation may be difficult, requires a case-based, case selection algorithm, and possibly case-adaptation algorithm. When required best solution or optimum solution, then CBR may not be able to handle such solutions. Hence, we propose a case based reasoning mechanism with semantic knowledge to handle these problems. Semantic data is the information that allows machines to understand the meaning of information. It describes the technologies and methods that convey the meaning of information.

Using semantics, data can be accessed more intelligently as it contains automated agents [3] to understand information. Basically, it breaks down the information into its simplest form so that it is quickly and deeply understood by the machine. Semantic data is not formally defined and it incorporates the following: Resource Description Framework, data interchange formats and notations, and the Web Ontology Language which all give a defined answer for concepts, terms, and relationships in a specific domain. The concept of semantic data as a whole has remained unclear and is generally speculated upon as not being a workable service.

The purpose of semantic data would be to allow computers to understand and figure out information without the help of a human user. In order to handle knowledge each piece of information must be programmed in details, however, using semantic technology, knowledge is self-defined and easily handled. The computer would be able to find information on its own, combine it with other information as needed, and act upon the information it received in an appropriate way. Semantic data plays an important part in this because the way that the data is lined up allows for the rest of the data in a sequence to be automatically figured out. The data is interpreted according to its relationship to each other and the result knows what the next data would be in sequence based on these relationships. All of the data and parts of a sequence are in an ordered hierarchy so that there is only one choice for the next part in a sequence. It's sort of like solving a math equation, there is only one correct answer almost all of the time. This relationship of data is what allows machines to work alone without human intervention.

In many domains Case-based Reasoning (CBR) has become a successful technique for knowledge-based systems and especially the medical domain. In medical domains, attempts to apply the complete CBR cycle are rather exceptional. Some systems have recently been developed [4], which on the one hand use only parts of the CBR method, mainly the retrieval, and on the other hand enrich the method by a generalization step to fill the knowledge gap between the specificity of single cases and general rules [5]. So we discuss the appropriateness of CBR for medical knowledge-based systems, point out problems, limitations and possibilities how they can partly be overcome.

This paper is organized in the following after this introduction; a theoretical background is illustrated in section 2.Section 3 gives an outline on related work. Section 4 introduces a brief description about semantic ontology. The description of the proposed architecture is given in section 5. Section 6 presents a comparative study after testing and comparing CBR applications. The conclusion and future work is presented in section 7.

2 THEORETICAL BACKGROUND

A. Case Based Reasoning

Case-Based Reasoning (CBR) is a problem solving paradigm that solves a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation [6]. More specifically, CBR uses a database of problems to resolve new problems. The database can be built through the knowledge engineering (KE) process or it can be collected from previous cases.

In a problem-solving system, each case would describe a problem and a solution to that problem. The reasoning engine solves new problems by adapting relevant cases from the library [7]. Moreover, CBR can learn from previous experiences. When a problem is solved, case-based reasoning can add the problem description and the solution to the case library. A new case generally represented as a pair <problem, solution> is immediately available and can be considered as a new piece of knowledge.

According to Doyle et al. [8], Case-Based Reasoning is different from other Artificial Intelligence approaches in the following ways:

- Traditional AI approaches rely on general knowledge of a problem domain and tend to solve problems on a first-principle while CBR systems solve new problems by utilizing specific knowledge of past experiences.
- CBR supports incremental, sustained learning. CBR solves a problem then it will make the problem available for future problems.

The CBR Cycle can be represented by a schematic cycle, as shown in Figure 1. The first phase is the retrieve phases, which identify features via noticing the feature values of a case, initially match a list of possible candidates and select the best match from the cases.
Second phase is the reuse phase where the difference between the new and the old case is determined by copying from the old case and adapting by transforming or reusing the old solution.
The third phase is the revise phase, in this phase, if the solution from the last phase is incorrect, then this solution must be evaluated in a real environment setting and finds the errors/flaws of the solution if the solution was evaluated badly.
Finally, the retain phase which is the fourth phase, in this phase incorporate the lesson learned from the problem-solving experience into the existing knowledge by extracting or indexing. By extracting we mean if the problem was solved using an old case, the system can build a new case or generalize an old case. By indexing we mean via deciding what types of indexes to use for future retrieval and integrate by modifying the indexing of existing cases after the experience

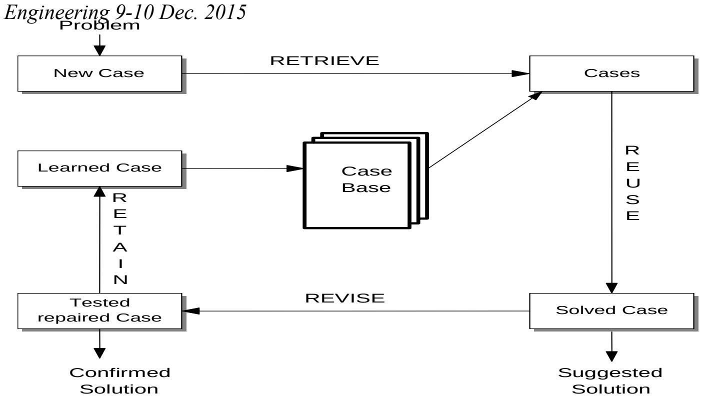


Figure 1: Case-based reasoning

There are three main types of CBR that differ significantly from one another concerning case representation and reasoning. The first one is Structural [9] in which cases are stored using a common structured vocabulary, i.e. ontology. The second is Textual [10] in such way cases are represented as free text, i.e. strings. The third is the Conversational CBR [11] in which a case is represented through a list of questions that varies from one case to another; knowledge is contained in customer / agent conversations.

During the past twenty years, many CBR applications have been developed, ranging from prototypical applications build in research labs to large-scale fielded applications [12]developed by commercial companies.

There are some disadvantages of CBR such as: can take large storage space for all the cases, can take large processing time to find similar cases in case-based, cases may need to be created by hand, adaptation may be difficult, needs case-based, case selection algorithm, and possibly case-adaptation algorithm. Optimum solution or best solution cannot be achieved using CBR.

CBR systems generally give good or reasonable solutions this is because the retrieved case often requires adaptation.

B. Semantic ontology

Semantic web is actually an extension of the current web in that it represents information more meaningfully for humans and computers alike. It enables the description of contents and services in machine-readable form, and enables annotating, discovering, publishing, advertising and composing services to be automated. It was developed based on Ontology, which is considered as the backbone of the Semantic Web. In other words, the current Web is transformed from being machine-readable to machine-understandable. In fact, Ontology is a key technique with which to annotate semantics and provide a common, comprehensible foundation for resources on the Semantic Web. Moreover, Ontology can provide a common vocabulary, a grammar for publishing data, and can supply a semantic description of data, which can be used to preserve the Ontologies and keep them ready for inference [13, 14].

Ontologies [15], which are used in order to support interoperability and common understanding between the different parties, are a key component in solving the problem of semantic heterogeneity, thus enabling semantic interoperability between different web applications and services.

Recently, ontologies have become a popular research topic in many communities, including knowledge engineering, electronic commerce, knowledge management and natural language processing. Ontologies provide a common understanding of a domain that can be communicated between people, and of heterogeneous and widely spread application systems. In fact, they have been developed in Artificial Intelligence (AI) research communities to facilitate knowledge sharing and reuse.

The goal of ontology is to achieve a common and shared knowledge that can be transmitted between people and between application systems. Thus, ontologies [16] play an important role in achieving interoperability across organizations and on the Semantic Web [17], because they aim to capture domain knowledge and their role is to create semantics explicitly in a generic way, providing the basis for agreement within a domain. Ontology is used to enable interoperation between Web applications from different areas or from different views on one area. For that reason, it is necessary to establish mappings among concepts of different ontologies to capture the semantic correspondence between them. However, establishing such a correspondence is not an easy task.

The primary use of the word “ontology” is in the discipline of philosophy, where it means “the study or theory of the explanation of being” [18]; it thus defines an entity or being and its relationship with an activity in its environment. In other disciplines, such as software engineering and AI, it is defined as “a formal explicit specification of a shared conceptualization” [18]. The foundations of this definition are:

- All knowledge (e.g. the type of concepts used and the constraints on their use) in ontology must have an explicit specification.
- An ontology is a conceptualization, which means it has a universally comprehensible concept

3 RELATED WORK

Case based reasoning (CBR) is a known problem solving technique based on reutilizing specific knowledge of previously experienced problems stored as cases. The CBR cycle consists of four major stages: Retrieve, Reuse, Revise and Retain(as shown in figure 1). In the Retrieve stage, the system selects a subset of cases from the case based that are relevant to the current problem. The Reuse stage adapts the solution of the cases selected in the retrieve stage to the current problem. In the Revise stage, the obtained solution is verified (either by testing it in the real world or by examination by an expert), which provides feedback about the correctness of the predicted solution. Finally, in the Retain stage, the system decides whether or not to store the new solved case into the case based.

Fuchs and Mille [19] have proposed a modeling of the CBR at the knowledge level. They have distinguished four knowledge models: the conceptual model of the domain describing the concepts use to describe the domain ontology independently of the reasoning; the case model which separates the case in 'problem, solution', and track of reasoning; the tasks reasoning models which include a model of specification and other one of tasks decomposition and; reasoning supports model.

D'Aquin [20] worked on the integration of the CBR in semantic Web. For that purpose, they have proposed an extension of OWL (Ontology Web Language) allowing representing the adaptation knowledge of the CBR. The expression of domain and cases knowledge in OWL allowed them to add to the CBR system the appropriate reasoning capacities of OWL by exploiting, for example, the subsumption and the instantiation.

Bichindaritz has demonstrated the use of ontologies for facilitating case structuring and acquisition [21]. Diaz-Agudo and Gonzalez Calero [22] proposed architecture independent from the domain which helps to integrate ontologies in CBR applications. Their approach consists in building integrated systems which combine cases specific knowledge with generic models of the domain knowledge. They presented CBROnto [23], as task / method ontology which supplies the necessary vocabulary to describe implied elements in the CBR processes

Case-based reasoning generally takes large storage space for all the cases, and also take large processing time to find similar cases in case-based. Moreover, cases may need to be created by hand which is another overhead when using case based reasoning and adaptation may be difficult. This paper introduces Case based Reasoning using semantic (SCBR) knowledge where it defines the cases semantically. The cases are semantically defined before being stored in the CBR system. New cases are semantically represented before being matched to the stored experiences so the case won't take large storage space, consequently, it can handle the drawbacks of CBR as shown in the next section.

4 PROPOSED SCBR SYSTEM

The new system mainly depends on defining the cases semantically. The cases are semantically defined before being stored in the CBR system; hence cases are better understood and thus represented with higher level of understanding. Consequently, new cases are initially also semantically represented before being matched to the stored experiences, for easier matching and verification. This has been achieved by dividing the system into three main layers. Layer 1 is the GUI interface, Layer 2 is responsible for semantic knowledge representation, and finally layer 3 is concerned about the CBR process (semantic case storage and retrieve cases).

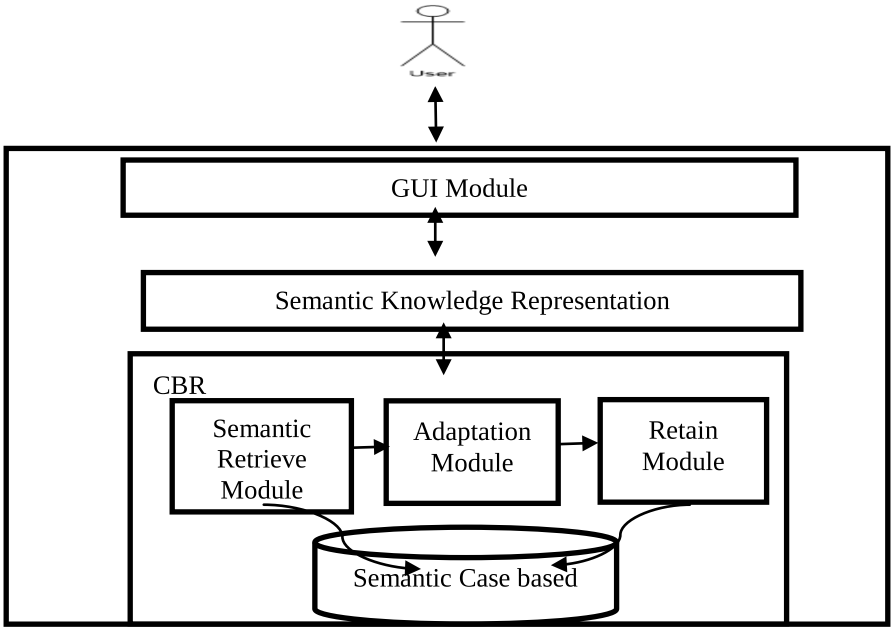


Figure 2: SCBR System modules

1) Layer 1: GUI Module

This module is the first layer in the system and is responsible for the graphical user interface, in other words, the interaction between the system and the user of the system. The user (patient) will insert Symptoms of the disease and all his precautions then apply the SCBR system to diagnose the patient's condition and find the right medicine.

2) Layer 2: Semantic Knowledge Representation

This module is responsible for understanding the input case and provides a semantic definition to the description of the case. We extract Metadata information from the input case to store our semantic understanding of the case.

Metadata is textual data, which contains a description of the concepts of the case. The module extracts a list of feature and their values from the input that will be used as extra attribute values in the retrieval phase (not included as attributes in the case). This extra information provides a deep understanding of the content of the case that helps later in the CBR processes. It guaranties a higher accuracy in matching or searching past experiences.

The paragraph is divided into a set of sentences and each sentence contains a set of tokens as shown in Figure 3.

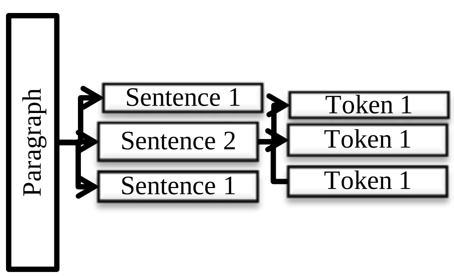


Figure 3: Basic Paragraph Structure

Each word in the text is represented as a token stored as an object. These objects store information like [24, 25]:

- Stop word (word without sense) does not contain important significance to be used in Search Queries. Usually these words are filtered out from search queries because they return vast amount of unnecessary information(i.e. a, about, before, above, after, again, the, that …)

- Main name inside the sentence: it is the direct word, which is related to the medical field such as Kidney disease, pregnant, headache, etc.
- The stemmed word: is to reduce the word to its origin. The term doesn't have to reduce the word to its root, as some times it gives a completely different meaning. A stem may consist of just a root. However, it may also be analyzed into a root plus derivational morphemes for example, the words "argue", "argued", "argues", "arguing", and " argus" reduce to the stem "argu" (illustrating the case where the stem is not itself a word or root) but "argument" and "arguments" reduce to the stem "argument".
- The Part-Of-Speech tag of the token: also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context. It is simple form to the identification of words as nouns, verbs, adjectives, adverbs, etc.
- A list of relations with other similar tokens: it means the relation between the tokens (i. e. worlds) extracted from the input text.

The organization in paragraphs, sentences and tokens is performed by NLP methods depending on the chosen implementation. The information extracted from the text is stored in the IEtext object. There are several types of information that will be obtained:

- Phrases identified in the text.
- Features: identifier-value pairs extracted from the text.
- Topics: combining phrases and features of a topic that can be associated to a text. A topic is a classification of the text.

Phrases and Features are stored using the objects implemented in the jcolibri.extensions. Textual.IE.representation.info sub-package. That package store three objects that aid in the representation of the extracted information:

- Phrase Info: stores extracted phrases.
- Feature Info: stores extracted features.
- Weighted Relation: represents a weighted relation between two tokens. These relations are found by the glossary and thesaurus methods.

Figure 4 illustrates the complete organization:

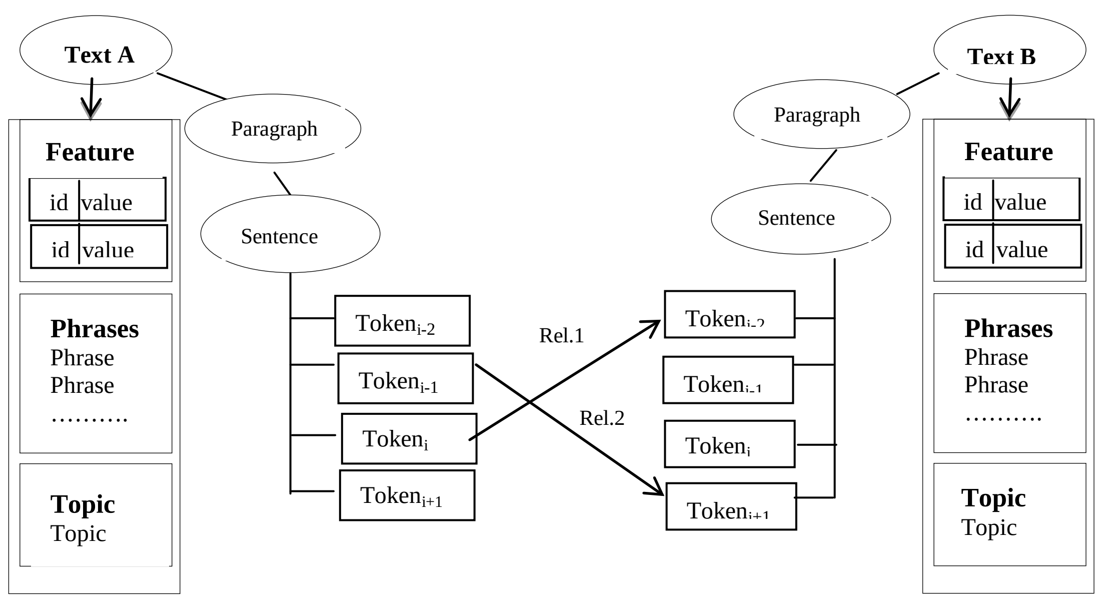


Figure 4: Global view of the representation of texts for IE.

Finally, the case will be stored in the case based with both the description of the case and the solution of the case as shown in figure [5].

The description part includes stored attributes with their values and an extra metadata that describes the case and will be used to extract an extra attribute that are not stored directly.

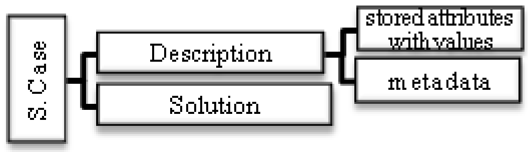


Figure 5: Semantic Case Structure

3) Layer 3: Semantic CBR Module

This module is divided into three stages that represent the main CBR processes. These are the retrieval module, adaptation module and the retain module.

1. Semantic Retrieve Module

Measure the similarity of the cases and retrieve most N similar cases [26].

Computing Similarity

The OpenNLP[X] library is a machine learning based toolkit for the processing of natural language text [27]. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and reference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy and perceptron based machine learning [28]. The methods of the IE extension extract information from texts and store it into the other attributes of the case. These attributes can be compared using normal similarity functions as shown in Figure 6.

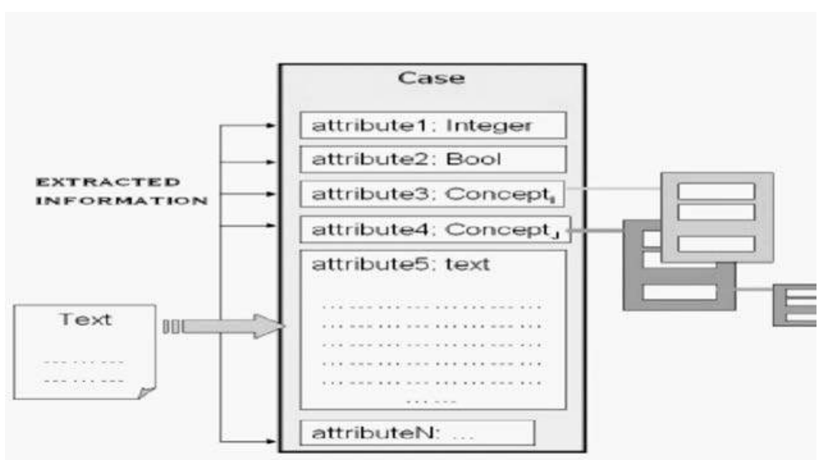


Figure 6: Common organization of textual cases

The generated entities (words) from OPenNLP[X] will be sent in the background request to the DBpediaSpotlight[X] to annotate text and get the most important features [29].

Due to defining the cases semantically by using ontology, it solves some of the main problems in the traditional case based reasoning. Large processing time to find similar cases is solved in semantic retrieval while providing a target solution in little time. This is because it depends on complex semantic case structures.

2. Adaptation Modules

Adaptation, as one of the most difficult tasks (especially in a complex problem domain) in traditional CBR, relies on both the retrieval of proper cases that need less adaptations and the utilization of appropriate domain knowledge. In this step the Adapting module transform or reuse the old solution because in many situations the case returned is not the exact solution needed.

Semantic representation of cases reduced the need for adaptation. Traditional CBR requires case adaptation when given solution is not as required or far from the real solution. This is reduced as case understanding is enhanced and deeper understanding of case knowledge is achieved. The new case can be adapted easily through the semantic relations of its knowledge.

3. Retain Module

In the retain step useful new cases are stored in the Semantic case for future reuse. This way the SCBR system has learned a new experience (knowledge based learning). SCBR is intuitive - it's how we work, no knowledge elicitation to create rules or methods this makes development easier and systems learn by acquiring new cases through use.

5 IMPLEMENTATION OF SCBR FRAMEWORK

A. Semantic Ontology

The ontologies are useful for designing SCBR applications because they allow the knowledge engineer to use knowledge already acquired, conceptualized and implemented in a formal language, like DLs based languages, reducing considerably the knowledge acquisition bottleneck. Ontologies used to build models of general domain knowledge. Although in a SCBR system the main source of knowledge is the set of previous experiences, the approach is to CBR is towards integrated applications that combine case specific knowledge with models of general domain knowledge. The more knowledge is embedded into the system, the more effective is expected to be. Semantic CBR processes can take advantage of this domain knowledge and obtain more accurate results.

As an example appeared in jCOLIBRI figure7 shows an example shows how to map a case into ontology.

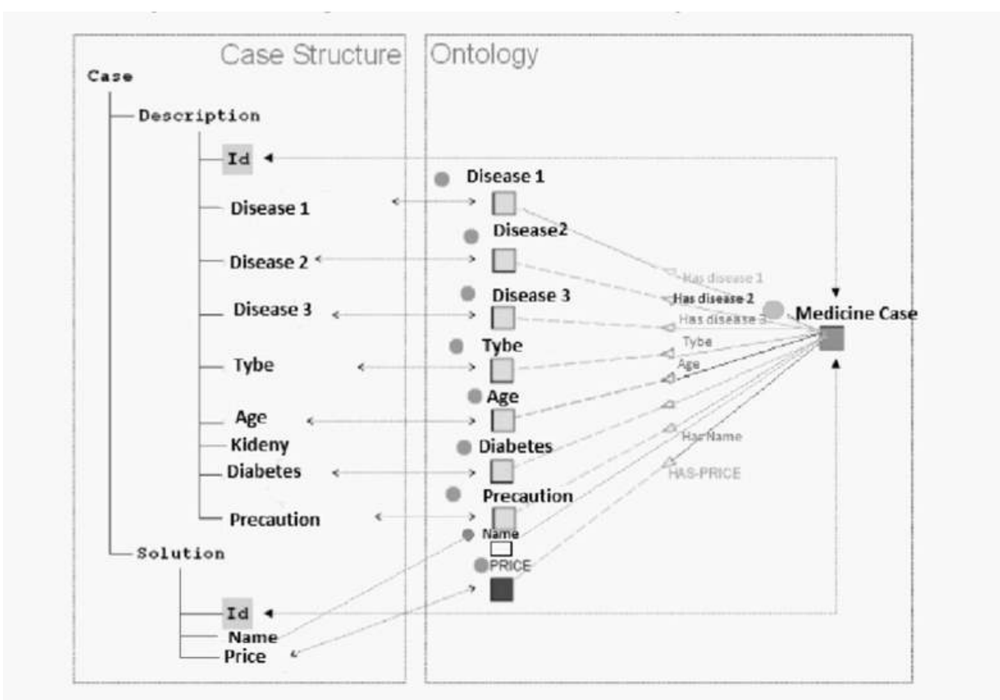


Figure 7: Case representation ontology

a) The algorithm to use case based reasoning with semantic knowledge is as follows:

1. Acquire input text (Symptoms of the disease and precautions)
2. Use open NLP to extract the precautions from the input text
3. Use dbpedia-spotlight (semantic representation) to understand and pick the important precaution and symptoms and represent them as a query
4. Apply the 4 steps of Case based reasoning to the extracted knowledge (as evaluated and represented semantically), cases are stored to determine the best medicine

   i. Retrieve:
      - Identify features: noticing the feature values of a case
      - Initially match a list of possible candidates by using ontology
      - Select the best match from the cases

   ii. Reuse:

      Here, we try to find the difference between the new and the old case by the following:
      - Copying: the solution is simply copied from the old case.
      - Adapting: transforming or reusing the old solution.

   iii. Revise:

      If the solution from the last phase is incorrect, then we must:
      - Evaluate this solution in a real environment setting.
      - Find the errors/flaws of the solution if the solution was evaluated badly.

   iv. Retain:

      Incorporate the lesson learned from the problem-solving experience into the existing knowledge by:
      - Extracting: the problem was solved using an old case; the system can build a new case or generalize an old case.

The following figures show the result when using Semantic case-Based Reasoning (SCBR) where the user insert the text (This an test of pregnant to and have asthma and stomach gastric and intestinal ulcer and kidney problem be University by sugar child and high pressure data will be adults only Gordon and maybe he is a children heart) then the open NLP extract the precautions from the input text (asthma, stomach, intestinal, pepticulcer, kidney, blood pressure, heart) then used bpedia-spotlight to understand and pick the important precaution and symptoms and represent them as a query finally apply four steps of CBR to determine the best three medicine as shown in figure 8

a

b

Figure 8: a,b Show the result when using SCBR

7 RESULTS

The case based used for testing the previously mentioned software obtained from the UC Irvine Machine Learning Repository, which contains details for 1000 cases for used patients (Diagnosis) [30].

This paper introduces a comparative study after testing and comparing the CBR applications mentioned previously in table 1 using the same case based

There are a number of major concerns when studying case-based reasoning approach. These major concerns are listed below:

- What is the structure of the cases?
- What are the numbers of correct match cases?
- What are (Recall, F-Measure and Accuracy)?
- What is the time taken to find target solution?

TABLE 1 CBR SHELL COMPARISON

CBR Shell	Case Structure	Correct Match Cases	Precision	Recall	F-Measure	Accuracy	Total Time
CBR Shell	Textual	70	0.666	0.7	0.682	0.703	3 minutes
Free CBR	Textual	81	0.784	0.81	0.786	0.813	2.3 minutes
jCOLIBRI	Xml /text	93	0.902	0.93	0.961	0.934	1.58 minutes
myCBR	Object	90	0.881	0.90	0.906	0.891	2.1minute
eXiTCBR	Custom CSV	61	0.603	0.61	0.606	0.613	4 minutes
SCBR System	Textual	99	0.933	0.99	0.961	0.994	57 seconds

The results show that the highest accuracy reached and the number of cases retrieved and matched also take the least time of cases retrieved and matched through the SCBR followed by jCOLIBRI, myCBR, FreeCBR,CBR Shell and eXiTCBR respectively.

8 COMPARATIVE STUDY BETWEEN TRADITIONAL CBR AND SEMANTIC CBR

This section introduces a comparative study after testing and comparing the CBR applications mentioned previously in table 1 using the same case based.

There are a number of major concerns when studying case-based reasoning approach. These major concerns are listed below:

- What are the selection strategies for finding similar cases?
- How is the case being retrieved?
- How is the selected case being revised?
- How is the suggested case being stored in case based?
- How is the suggested case being indexed for faster access?
- How to deal with noisy data or missing values?

According to the previous mentioned points, a comparative study between the traditional CBR software and case based reasoning using semantic (SCBR) mentioned previously in table 1. Next paragraphs describe the effect of each factor to each CBR software respectively.

No interfaces to external systems and DB are available in myCBR. It is valid regarding the interfaces to real-time or diagnostic systems. On Retain phase, myCBR allows saving the Query as a new case, also to use an old case as a basis for new Query. MyCBR is entirely based on GUI, providing a ready-windows templates and forms for defining classes, attributes, SFs, queries to the case-based DB, visualization of found results and more.

myCBR platform can be used for non-complex CBR applications development with partial CBR R4 cycle and with small number of cases in text file. For CBR application development, no time for programming is needed but it is needed only for case configuration. MyCBR is not suitable to be applied with large number of attributes with text solution, especially when they must be visually presented in one window. Table 1 summarizes the comparisons between the selected CBR software.

SCBR system has a very simple and powerful GUI; it represents cases semantically in a very simple way so the cases don't need to be created by hand. There are a number of case retrieval algorithms applicable in case based reasoning. These algorithms are based on the similarity metric that allows resemblance between cases stored in case based. The nearest neighbor retrieval algorithm & induction retrieval algorithms are two chief algorithms used in this process. Nearest-neighbor retrieval is a straightforward approach that computes the similarity between relevant cases found through indexing.

According to using ontology in retrieval stage theses algorithm don't take large processing time to find similar cases in case-based and CBR systems generally give good or reasonable solutions and possibly case don't need adaptation algorithm and if it needs SCBR can make maintenance easy and justification through precedent

TABLE 1CBR SHELL COMPARISON.

CBR Shell	**Case Structure**	**Selection strategies**	**Case retrieval**	**Case revised**	**Case storage**	**Case indexed**	**Graphical User Interface(GUI)**	**Dealing with uncertain data**
CBR Shell	Textual	distance method	Two methods KNN Threshold	Manual	Text	No	Very simple GUI	Can't handle
FreeCBR	Textual	weighted Euclid distance	Simple matching	Manual	Text	No	Simple and easy but limited	Can't handle
jCOLIBRI	Xml /text	Similarityfunctions	method k-NN, Threshold, Ontology, Textual, OpenNLP and GATE Recommender	Automatic	CSV XML	Yes	Simple and powerful Use wizard to simplify	Handle as null
myCBR	Object	similarity functions	Query model	Manual	CSV XML	No	user can customize the GUI and handle most of things	Handle as _unknown_ or _undefined
eXiTCBR	Custom CSV	Distance method or similarity measure	Simple Querying	Manual	Text	No	Very simple , no options	Can't handle

Table 2 can be summarize the difference between traditional case based reasoning and case based reasoning using semantic knowledge as proven in proposed SCBR system.

TABLE 2: DIFF. BETWEEN TRADITIONAL CBR &SCBR

Traditional Case based Reasoning(CBR)	**Semantic Case based reasoning(SCBR)**
Can take large storage space for all the cases Can take large processing time to find similar cases in case-based Cases may need to be created by hand Adaptation may be difficult Needs case-based, case selection algorithm, and possibly case-adaptation algorithm if you require the best solution or the optimum solution CBR may not be for you CBR systems generally give good or reasonable solutions this is because the retrieved case often requires adaptation	CBR is intuitive - it's how we work no knowledge elicitation to create rules or methods this makes development easier systems learn by acquiring new cases through use this makes maintenance easy justification through precedent Adaptation may be easy SCBR system give the target solution this is because it depend on semantic Complex case structures Knowledge-based learning

9 CONCLUSION AND FUTURE WORK

Case-based reasoning systems (CBR) have some drawbacks such as: occupies a large storage space for all the cases, take large processing time to find similar cases in case-based and cases may need to be created by hand. This paper proposed a case based reasoning mechanism with semantic knowledge to handle these problems where the new system mainly depends on defining the cases semantically. New cases are semantically represented before being matched to the stored experiences.

This paper also introduces a comparison among most common used traditional CBR software and the proposed SCBR system. It also mentions the advantages and disadvantages of each software. Moreover, this paper applies the same case based to the six CBR software to compare and evaluate the results using the predetermined factors and calculating Precision, Recall, F-Measure and Accuracy for each one. As a conclusion CBR, Free CBR and eXit CBR are very simple software including simple GUI and only include the selection and retrieval of similar cases using traditional techniques. On the other hand, Proposed SCBR system, myCBR and jCOLIBRI are more complex and can be used for complex CBR.

The proposed system can prove that combining Ontology technology and CBR has a positive impact on search results and the more cases are stored to increase system performs. We recommend applying this approach to cases on Wikipedia in other fields. Also, for future work we will investigate the methodology for building ontology from unstructured data such web pages and documents. Moreover, more investigation can be done for reducing the case storage size and time.

REFERENCES

[1] A. Aamodt and E. Plaza, “Case-based reasoning: foundational issues, methodological variations, and system approach” AI Communications 7(1), 39–59, 1994.

[2] Janet L. Kolodner, “An Introduction to Case-Based Reasoning” Artificial Intelligence Review 6, 3-34, 1992

[3] Abadi, D., A. Marcus, S. Madden, and K. Hollenbach (2009). SW-Store: a vertically partitioned DBMS for Semantic Web data management.

[4] Intelligence and Applications (AIA 2009), IASTED, Innsbruck, Austria, Editor(s):V. Devedžic, February, 2009.

[5] Bichindaritz I, Marling C. Case-based reasoning in the health sciences: What’s next? In Artificial Intelligence in Medicine.36(2), 2006.

[6] Simon C.K. Shiu, “Case-Based Reasoning: Concepts, Features and Soft Computing” Applied Intelligence 21, 2004.

[7] Zhi-We Ni, Shan-Lin “Integrated Case-based Reasoning” Proceedings of the Second International Conference on Machine Learning and Cybernetics, XI; 2-5 November 2003.

[8] Leake, David, "CBR in Context: The Present and Future (http:/ / www. cs.indiana.edu/~leake/papers/p-96-01_dir.html/paper. html)", In Leake, D., editor, Case-Based Reasoning: Experiences, Lessons, and Future Directions. AAAI Press/MIT Press, 1-30, 1996.

[9] R. Bergmann, K.-D.Altho®, S. Breen, M. GÄoker, M. Manago, R. TraphÄoner, and S. Wess. Developing industrial case-based reasoning applications: The INRECA methodology. LNAI 1612. Springer, 2nd Edition, 2003.

[10] M. Lenz and K. Ashley, editors. Proceedings of the AAAI98 Workshop on Textual Case-Based Reasoning. AAAI Press, 1998.

[11] D. Aha, L. A. Breslow, and H. Munoz-Avila. Conversational case-based reasoning. Applied Intelligence, 2001.

[12] I. Watson. Applying case-based reasoning: techniques for enter-prise systems. Morgan Kaufmann Publishers Inc., 1998.

[13] Antoniou, Grigoris, & Frank van Harmelen, “ A Semantic Web Primer. Cambridge”, MA: MIT Press.Bridgman, & Percy, W. (1922). Dimensional Analysis. New Haven, CT: Yale University Press, 2004.

[14] S. Bechhofer, I. Horrocks, C.A. Goble, and R. Stevens, "OilEd: a Reason-able Ontology Editor for the Semantic Web", In Proceedings of Description Logics, 2001.

[15] D. Dou, D. McDermott, and P. Qi, "Ontology Translation on the Semantic Web", Presented at on Data Semantics Journal, 2005.

[16] D. Fensel, "Ontologies: Silver Bullet for Knowledge Management and Electronic
Commerce", Springer, 2001.

[17] D. Tidwell, "Web Services-The Web's Next Revolution", IBM Web Service Tutorial, 29 Nov. 2000, http://www-106.ibm.com/developerworks/edu/ws-dwwsbasics-i.html.

[18] H. Mihoubi, A. Simonet, and M. Simonet, "An Ontology Driven Approach to Ontology Translation", In Proceedings of DEXA, 2000.

[19] B. Fuchs and A. Mille, “Une modélisation au niveau connaissance du raisonnement à partir de cas”, in L'HARMATTAN, Ed., Ingénierie des connaissances, 2005.

[20] M. D'aquin, J. Lieber and A. Napoli, “Artificial Intelligence: Methodology, Systems, and Applications”, Lecture Notes in Computer Science, chapter Case-Based Reasoning within Semantic Web Technologies. Springer Berlin / Heidelberg, vol. 4183, 2006.

[21] I. Bichindaritz, Thesis: “Case Based Reasoning Meets the Semantic Web in Biology and Medicine”, Advances in Case Based Reasoning, 7th European Conference, LNAI, Spain, 2004.

[22] B. D-Agudo and P. G-Calero, “ An architecture for knowledge intensive CBR systems”, Advances in Case-Based Reasoning”. (EWCBR'00). Springer-Verlag, Berlin Heidelberg New York, 2000.

[23] B. D-Agudo and P. G-Calero, “CBROnto: a task/method ontology for CBR”. In S. Haller and G. Simmons, editors, Procs. of the15th International FLAIRS'02 Conference. AAAI Press, 2002.

[24] M. Brown, C. Förtsch, and D. Wissmann. Feature extraction - the bridge from case-based reasoning to information retrieval. In Proceedings of 6th German Workshop on Case-Based Reasoning, 1998.

[25] S. Brüninghaus and K. D. Ashley. The role of information extraction for textual CBR. In Proceedings of the 4th International Conference on Case-Based Reasoning, Springer-Verlag, 2001.

[26] Croft et al., Bruce Croft, Donald Metzler, and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison Wesley, 2009.

[27] K. M. Gupta and D. W. Aha. Towards acquiring case indexing taxonomies from text. In Proceedings of the 17th Int. FLAIRS Conference, Miami Beach, FL, 2004.

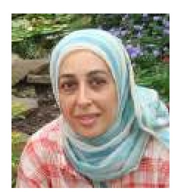
[28] E. Hatcher and O. Gospodnetic. Lucene in Action (In Action series). Manning Publications Co., Greenwich, CT, USA, 2004.

[29] Bird, Steven, Ewan Klein, and Edward Loper. Natural Language Processing with Python, Sebastopol, CA: O'Reilly, 2010.

[30] http://archive.ics.uci.edu/ml/

BIOGRAPHY

PassentElKafrawy, Associate Professor, Faculty Science, Menofia University



Dr. Passent M ElKafrawy is an Associate Professor since 2013, she got her PhD from the University of Connecticut in United states on 2006 in Computer Science and Engineering in the field of computational geometry as a branch of Artificial Intelligence. Then she taught in Eastern State University of Connecticut for one year. In 2007 she worked as a Teacher in Faculty of Science, Menoufia University, Mathematics and computer science department since that time till now.

She has over 20 publications and member of ACEE, ECOLE and TIMA research organizations. One of the organizing members of the following conferences: SPIT and ESOLEC. Supervising over 10 research studies between PhD and MSc. Member of the faculty projects for education development as CIQAP, DSAP, and Question Bank.

Rania Ahmed , Assistant Lecturer in Modern University for Technology and Information,

was born in Giza, Egypt, in 1985.She received the Bachelor in Computer Science degree from the Helwan University in 2006 and the Master in Computer Science degree from the Helwan University in 2010. She is currently pursuing the PH.D degree with the Department of Computer Science.

دلالات الالفاظ لحالة البرمجيات القائمة على منطق المعرفة الطبية

بسنت محمد الكفراوى* , رانيا احمد محمد**

*أستاذ مساعد , كلية العلوم جامعة المنوفية

**مدرس مساعد بالجامعة الحديثة للتكنولوجيا والمعلومات*

الملخص

يقدم هذا البحث اسلوبا جديدا في المنطق القائم على الحالة (CBR) باستخدام دلالات المعرفة(SCBR) لتمثيل الحالات، بنيةحالة، وحالةالمفاهيم القائمة في علم الأحياء والطب. ويمكن توسيع نطاق النهج إلى مجالات التطبيقات الأخرى منCBR. والميزة الرئيسية لهذا الاسلوب هو أن نظم الدلالات اللفظية في البيانات تم تصميمها لفهم المحتوى الحقيقي من الكلمة بأكبر قدر ممكن من ضمن مجموعة البيانات. هذه الورقة أيضا تقدم المقارنة بين طرقCBR التقليدية و SCBR حيث هناك بعض المشاكل في طرقCBR التقليدية مثل تعديل الحالات قد يكون من الصعب؛ قد تحتاج الحالات إلى إنشاء باليد؛ الوقت المستغرق في المعالجةلإيجاد حالات مماثلة طويل. ونظم CBR تعطي عادة حلول جيدة أومعقولة وذلك لأن حالة استردادها غالبا ما يتطلب التكيف. يمكن إستخدام SCBR لمعالجة هذه المشاكل.

BASMA: BibAlex Standard Arabic Morphological Analyzer

Sameh Alansary

Director of Arabic Computational Linguistics Center Bibliotheca Alexandrina
sameh.alansary@bibalex.org
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University

***Abstract*—Arabic morphology poses special challenges to computational natural language processing systems. Its rich morphology and the highly complex word formation process of roots and patterns make computational approaches to Arabic very challenging.Morphological analyzers are preprocessors for text analysis. This paper sheds the light on BASMA-Tool (BibAlex Standard Arabic Morphological Analyzer) that has been initiated at Bibliotheca Alexandrina (BA). The BASMA tool is based on Buckwalter Arabic Morphological Analyzer (BAMA). It focuses on fixing its problems, adding a set of useful morphological features that BAMA does not provide, and disambiguating its multiple solutions. This is done depending on a well training data and a hybrid system (Rule based and memory based). Precision and Recall are the evaluation measures used to evaluate BASMA tool. At this point, precision measurement was 93.37%while recall measurement was 96.9%.The percentages are expected to rise by implementing the improvements while working on larger amounts of data.**

1 INTRODUCTION

Arabic is a language of rich morphologycompared to other language especially Europeanlanguages. It is based on both derivational andinflectional morphology. The richness of Arabicmorphology makes the analysis process difficultto deal with. On the one hand, morphological analysisprocess is used in most of the NLPapplications such as information retrieval, spellcheckingand machine translation. On the otherhand, morphological analysis is the first stepbefore syntactic analysis. Furthermore, it is anessential step in semantic analysis.[1]

Arabic has a high degree of ambiguity resultingfrom its diacritic-optional writing system and common deviation from spelling standards (e.g., Alif and Ya variants).[2]

Morphological analysis for text corpora is a prerequisite for many text analytics applications, which has attracted many researchers from different disciplines such as linguistics (computational and corpus linguistics), artificial intelligence, and natural language processing, to morphosyntactically analyze text of different languages including Arabic. Recently, several researchers have investigated different approaches to morphological and syntactic analysis for Arabic text. Many systems have been developed which vary in complexity from light stemmers, root extraction systems, lemmatizers, complex morphological analyzers, part-of-speech taggers and parsers.[3]

In 2007, Bibliotheca Alexandrina (BA) has started an important project of building the “International Corpus of Arabic (ICA)”.It is a serious attempt to build a representative Arabic corpus as being used all over the Arab world that is able to support research on Arabic. It is planned to contain 100 million words morphologically, syntactically and semantically analyzed. The first stage of linguistic analysis of the International corpus of Arabic is to analyze the 100 million words of the ICA corpus morphologically.[4][5][6]

The stem-based approach “concatenative approach” has been adopted as a linguistic approach to analyze the ICA morphologically. There are many morphological analyzers for Arabic;some of them are available for research and evaluation while the rest are proprietary commercial applications. Buckwalter Morphological analyzer (BAMA) is one of the well-known analyzers in the literature and has even been considered the “most respected lexical resource of its kind” [6].It is designed as a main database of word forms interacting with other concatenation databases. In Buckwalter, every word is entered separately, and the stem is used as the base form of a word. Words are viewed as being composed of basic units that can combine with morphemes governed by morphotactic rules; thus, Buckwalter Morphological Analyzer entails the use of three lexicons: a Prefixes Lexicon, a Stem Lexicon, and a Suffixes Lexicon.

Section 2 of this paper will discuss the trials that use BAMA in the morphological disambiguation process. Section 3 will review the BibAlex Standard Morphological Analyzer system and why there was a need to enhance BAMA, through explaining and discussing some of the main problems noticed in its output. This section will also introduce to what extent it is different from BAMA (2004). Moreover, section 4will show the current state of the development and BASMA`s results and section 5includes a comparison between BASMA and MADA. Finally, section 6 will state the conclusion.

2 RELATED WORK

MSAmorphological analysis, disambiguation, part-of-speech(POS) tagging, tokenization, lemmatization and diacritization have received a lot of focus; for an overview, see[7]. And more recently, there has been growing body of work on Dialectical Arabic (DA)[8], [9], [10] and [11] among others. In this paper, the discussion will be focused on two systems that are commonly used by researchers in Arabic NLP: MADA [12], [13], [14] and [11] and AMIRA [15].

The primary purpose of Morphological Analysis and Disambiguation for Arabic(MADA3.2) is to extract as much linguistic information as possible about each word in the text, from given raw Arabic text, in order to reduce or eliminate any ambiguity concerning the word. MADA uses ALMORGEANA (an Arabic lexeme-based morphology analyzer) to generate every possible interpretation of each input word. It then applies a number of language models to determine which analysis is the most probable for each word, given the word's context.

MADA uses up to 19 orthogonal features in order to choose, for each word, a proper analysis from a list of potential analyses derived from the Buckwalter Arabic Morphological Analyzer (BAMA) [16]. The BAMA analysis that most closely matches the collection of weighted, predicted features is chosen. The 19 features include 14 morphological features that MADA predicts using 14 distinct Support Vector Machines (SVMs) trained on the PATB. The other five features that MADA capture information such as spelling variations and n-gram statistics.

Since MADA selects a complete analysis from BAMA, all decisions regarding morphological ambiguity, lexical ambiguity, tokenization, diacritization and POS tagging in any possible POS tag set are made in single action [11], [17], and [18]. The choices are ranked in terms of their score. MADA has over 96% accuracy on basic morphological choice (including tokenization, but excluding case, mood, and nunation) and on lemmatization. MADA has over 86% accuracy in predicting full diacritization (including case and mood). More detailed comparative evaluations can be found in [12], [17] and [13].

The AMIRA toolkit includes a tokenizer, a part of speech tagger (POS), and a base phrase chunker (BPC), also known as a shallow syntactic parser. The technology of used in AMIRA iscompletely different from that of MADA, since it is based on supervised learning with no explicit dependenceon knowledge of deep morphology, it relies on surface data to learn generalizations.

AMIRA was enhanced, in later versions, with a morphological analyzer and a named-entity recognition (NER) component. Moreover, both tools are similar in using a unified framework that postpones each of the component problems as a classification problem to be solved sequentially. AMIRA adopts a multi-step approach to tokenization, part-of-speech tagging and lemmatization, in contrast to MADA that handles all of these and more in a single action. The analysis that MADA provides is deeper than that of AMIRA, namely by identifying syntactic case, mood and construct state in the morphological tag, however, it is slower in processing. In addition, AMIRA provides additional utilities - BPC and NER - that are not supported by MADA. Both tools are somewhat brittle, academic prototypes implemented in Perl; they rely on third-party software utilities which the end-user must install andconfigure separately. [2]

3 BIBALEX STANDARD ARABIC MORPHOLOGICAL ANALYZER (BASMA)

Initially, Buckwalter Arabic Morphological Analyzer (BAMA) has been selected, since it was the most suitable lexical resource to our approach[4]. Although it has many advantages including its ability to provide a sufficient amount of information such asLemma, Vocalization, Part of Speech (POS), Gloss, Prefix(s), Stem, Word class, Suffix(s), Number, Gender, Definiteness and Case or Mood, it does not always provide all the information the ICA requires, and in some cases, the provided analyses would need some modification. The obtained results may vary between giving the right solution for the Arabic input word, provide more than one result that needs to be disambiguated to reach the best solution, provide many solutions, but none of them is right, segment the input words wrongly without taking the segmentation rules in consideration or provide no solutions. Consequently, solutions enhancement would be needed in these situations.

Number, gender and definiteness need to be modified according to their morphosyntactic properties. Some tags had been added to the ICA lexicon, some lemmas and glossaries had been modified and others had been added. In addition, new analysis and qualifiers had been added as root, stem pattern and name entities [5].

The process of developing a morphological analyzer tool for ICA began in 2007 which is known as BibAex Arabic Morphological Analyzer Enhancer (BAMAE). It is a system that has been built to morphologically analyze and disambiguate the Arabic texts depending on BAMA's output. It was preferred to use BAMA's enhanced output of ICA, since it contains more information than any other BAMA's enhanced systems. And this is the reason why the members of ICA team aimed to build their own morphological analyzer tool.

In order to reach the best solution for the input word, BAMAE preforms automatic disambiguation process carried on three levels, depends primarily on the basic POS information (Prefix(s), Stem, Tag and Suffixes) obtained from enhanced BAMA's output. [5], [6]:

- Word level which avoids or eliminates the impossible solutions that Buckwalter provides due to the wrong concatenations of prefix(s), stem and suffix(s).
- Context level where some linguistic rules have been extracted from the training data to help in disambiguating words depending on their context.
- Memory based level which is not applicable in all cases; it is only applicable when all the previous levels failed to decide the best solution for the Arabic input word.

After selecting the best POS solution for each word, BAMAE detects the rest of information accordingly. It detects the lemmas, roots (depending primarily on the lemmas), stem patterns (depending on stems, roots and lemmas), number (depending on basic POS and stem patterns), gender (depending also on basic POS, stem patterns and sometimes depending on number), definiteness (depending on POS or their sequences), case (depending on definiteness and sequences of POS) and finally it detects the vocalization of each word.

Figure 1 shows BAMAE architecture starting from the input text and the numerous solutions for each word in order to predict the best POS solution for each word and then detect the rest of information accordingly.

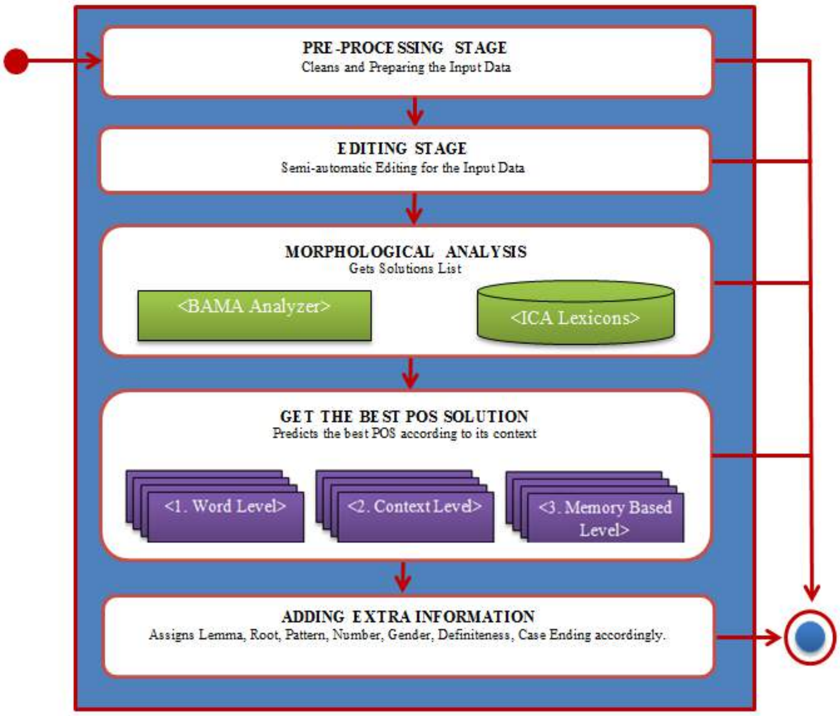


Figure 1: BAMAE Architecture.

The precision measurement of BAMAE was 92% while recall measurement was 89%. These percentages were expected to be raised by implementing the improvements while working on larger amounts of data [6].

After discovering BAMA's output problems, and handling these problems in the BAMAE, the decision was to handle these problems in BAMA.But, not all of BAMA's outputproblems have been handled in BAMA. Others have been handled by implementing Arabic linguistic rules, depending on the kind of the problem. Handling these problems required some modifications in the Perl code of BAMA (AraMorph). Moreover, more development was needed such as a new feature that Buckwalter does not provide, was added to BAMA's lexicons namely stem pattern as well as another feature that is found in lexicons, but does not appear in BAMA's output solutionsnamely root. By handling these problems and revamping some functions in BAMAE another update has been released known as BASMA. The following sub-sections review how these problems have been handled and implemented in BASMA:

A) Problems handled in BAMA's lexicons:

As mentioned before, not all problems are necessarily handled in this stage, it only handles problems that are related to the lackin grammar-lexis specifications, uncovered concatenations of some words, uncovered prefixes or suffixes in

Arabic, wrong segmentations, wrong lemmas, wrong roots and wrong tags. These problems have been fixed in BAMAs' lexicons and/or their compatibility tables[1] according to the problem type.

The problems that are related to the lack in grammar-lexis specifications, uncovered prefixes or suffixesin Arabic and wrong tags have been fixed in both BAMAs' lexicons and their compatibility tables, because if a new prefix, tag or suffix is added, some constrains must be added to rule which combinations of these prefixes, tags and suffixes are linguistically acceptable and which are not, depending on the nature of Arabic language. In addition, the lack in grammar-lexis requires adding more constrains to avoid the wrong combinations that BAMA does not constrain. Figure 2 shows an example for the problem of detecting wrong tags andlack ingrammar-lexis specifications for some words and how it has been handled in this stage.

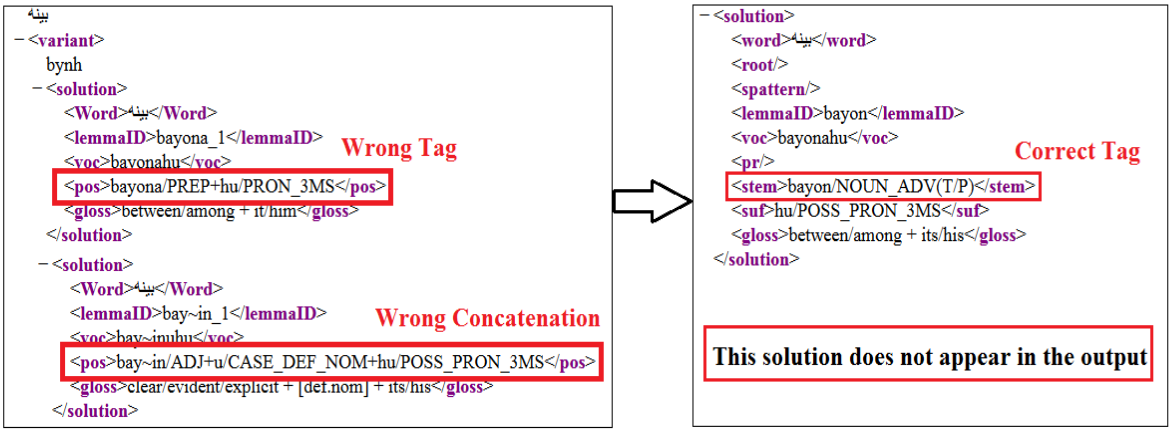


Figure 2: Example for wrong tags and concatenations.

The problems that are related to wrong lemmas or roots and wrong or new glosses have been handled in BAMA's lexicons and specifically in dicStems lexicon without being handled in the compatibility tables. As mentioned before the root feature does not appear in BAMA's output, although it is found in the dicStems lexicon. Moreover, unfortunately not all of the roots that are available founded in this lexicon are Arabic root, so there has to be some modifications in these roots. After reviewing all roots in the dicStems lexicon, they are displayed in the output.

Although the stem pattern is not used in BAMA's lexicon at all, it is found that the stem pattern feature is very useful in enriching the lexicons, we have depended onit in the disambiguation process of ICA texts. The stem patterns have been detected automatically, depending on root and stem of some words and depending on root, lemma and stem in other words. Then, these stem patterns have been added and mapped in the dicStems lexicon. Figure 3 shows an example for the problem of wrong lemmas and rootsand how the roots and stem patterns appear now in the output:

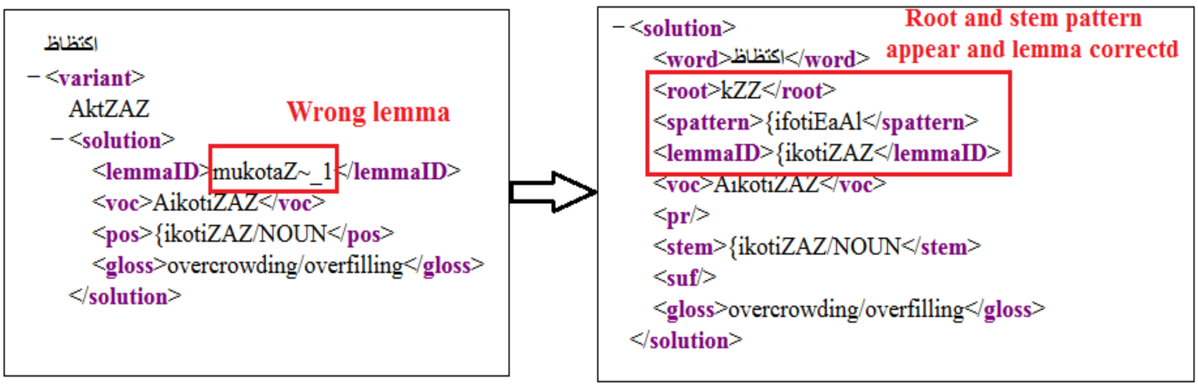


Figure 3: Example for wrong lemma and displayed root and stem pattern in the output.

Some words did not have any solutions for one of three reasons. First, some words are not analyzed altogether by BAMA; second, some words are analyzed, but none of the provided solutions is suitable to their contexts in the text; third, some words are wrongly segmented by BAMA [5] and [6]. Such words have been inserted in BAMA's dicStems lexicon with it suitable constrains to generate it correctly. An example of the second category of unanalyzed words is the passive form

1 For more information about BAMA lexicons visit: https://catalog.ldc.upenn.edu/docs/LDC2004L02/readme.txt(Last Access 19-11-2015).

of the word 'حرموا''be forbidden/be deprived'. After inserting the suitable transliteration, stem, tag (with suitable constraints) and gloss for this word, it is analyzed correctly as figure 4 shows:

```
- <solution>
    <word>حرموا</word>
    <root>Hrm</root>
    <spattern>fuEill</spattern>
    <lemmaID>Haram-iu</lemmaID>
    <voc>HurimuwA</voc>
    <pr/>
    <Stem>Hurim/PV_PASS</Stem>
    <Suf>uwA/PVSUFF_SUBJ:3MP</Suf>
    <gloss>be forbiden/be deprived + they [verb]</gloss>
  </solution>
```

Figure 4: Example of recently inserted word.

It must be noted that after handling the problem of wrong concatenations and the lack in grammar-lexis specifications, there will be no need to handle this part in BASMA. Furthermore, these modifications are still in progress to enhance the input solution source for BASMA as much as possible, hence enhancing the morphological analysis results.

B) Problems handled by Arabic Linguistic rules:

The problems that have been handled through the linguistic rulesare the problems that are related to morphosyntactic properties such as number, gender, definiteness and case ending. There are some linguistic rules that have been extracted from the previously analyzed data to help in assigning the right solutionin the next time the data is analyzed. The assigning process may depend on basic POSand stem pattern of each best selected solution such as in number and gender,or it may depend either on the POS ofeach best selected solution or the POS of the surrounding words in addition toits POS such as definiteness and case (context based). These features are no longer displayed in BAMA's output, since the correctness of detecting the number and gender by the linguistic rules in BASMA is more adequate. In addition, both definiteness and case ending features are context based features, so there is no need to display such information in the solutions list while selecting the best morphological analysis for each word.

Figure 5 shows some words that BAMA has assigned the wrong number and gender to them, and how these words have been handled in BASMA.

Figure 5: Example for the corrected gender and number features.

It must be noted that in order to prevent such features from appearing in BAMA's output some handling have been done in dicSuffixes BAMA's lexicon. All information that refer to any of these features have been deleted.The accuracy of rules in detecting gender and number are acceptable and can be enhanced, while the accuracy of rules in detecting definiteness and case ending still needs more modifications, since these features need more syntactic information.

C) *Needed modifications in BAMA's AraMorph Perl file:*

There are some modifications that are needed in BAMA's *AraMorph* Perl file. These modifications needto be compatible with the new added features in BAMA's output; root and pattern. In addition, there are some needed modifications to make the parsing process of BAMAs' solutions in BASMA easier. These modifications are 1) separating the prefixes and suffixes from the stem, 2) displaying the input word of every word, and 3) showing the x_solution of BAMA with only the words that have no solutions at all. Figure 6 shows BAMA's output solutions after these modifications.


```
- <solution>
    <word>وإنسانيته</word>
    <root>'ns/nws</root>
    <spattern>fiEolaAniy~</spattern>
    <lemmaID><inosAniy~ap</lemmaID>
    <voc>wa<inosAniy~athu</voc>
    <pr>wa/CONJ</pr>
    <Stem><inosAniy~/NOUN</Stem>
    <Suf>at/NSUFF+hu/POSS_PRON_3MS</Suf>
    <gloss>and + humanity + his/its</gloss>
  </solution>
- <solution>
    <word>وإنسانيته</word>
    <root>'ns/nws</root>
    <spattern>fiEolaAniy~</spattern>
    <lemmaID><inosAniy~ap</lemmaID>
    <voc>wa<inosAniy~athi</voc>
    <pr>wa/CONJ</pr>
    <Stem><inosAniy~/NOUN</Stem>
    <Suf>at/NSUFF+hi/POSS_PRON_3MS</Suf>
    <gloss>and + humanity + his/its</gloss>
  </solution>
```


Figure 6: BAMA's output after modifications.

4 RESULTSANDEVALUATION

To evaluate BASMA, a blind test data set (1,000,000 representative words) was run using BASMA, and results were compared to a manually annotated version. Precision, Recall and accuracy are the evaluation measures used to evaluate the BASMA system. Precision is a measure of the ability of a system to present only relevant results. Recall is a measure of the ability of a system to present all relevant results. The evaluation has been conducted on two levels; the first level includes the precision, recall and accuracy for each qualifier separately as table 1 shows. The second level includes the basic POS in addition to adding a new qualifier each time to investigative how it would affect the accuracy as table 2 shows.

TABLE 1
PRECISION, RECALL AND ACCURACY FOR QUALIFIERS SEPARATELY

Qualifier	Precision	Recall	Accuracy
Lemma	97.16	99.95	97.07
Pr1	98.50	99.90	97.00
Pr2	99.90	99.96	99.80
Pr3	100	100	100
Stems	96.83	99.95	93.67
Tags	96.39	99.96	92.78
Suf1	96.27	99.25	95.82

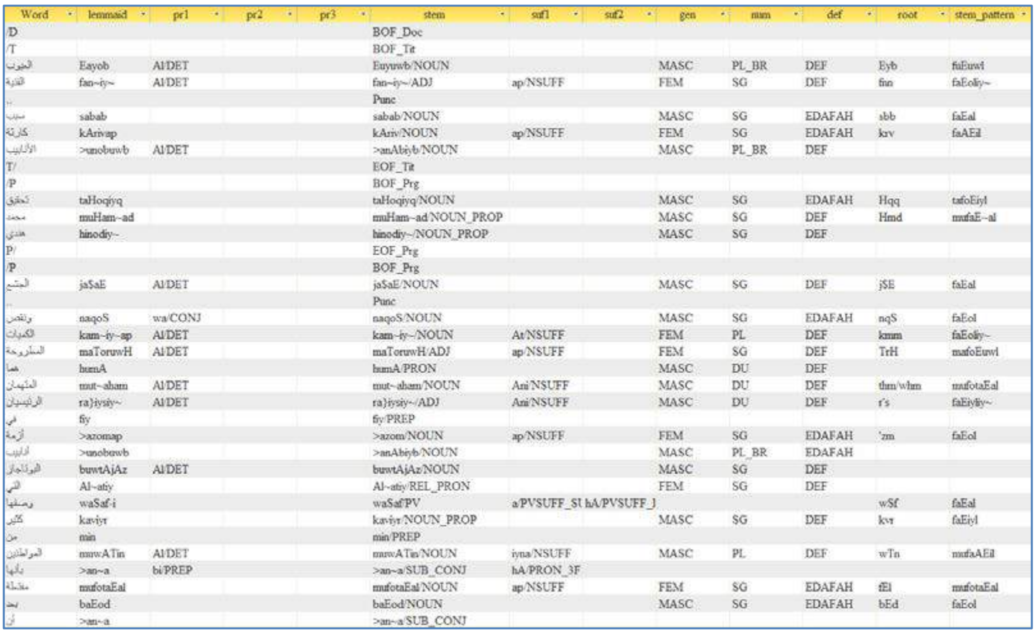
Suf2	99.86	99.97	99.72
Gender	98.46	99.87	97.74
Number	98.84	99.78	97.67
Definiteness	93.94	98.51	87.89
Root	99.30	99.80	98.60
Stem Pattern	97.80	99.80	95.60

TABLE 2

ACCURACY DECREASING AS A RESULT OF ADDING NEW QUALIFIER EACH TIME TO THE MAIN POS TAG

POS + Qualifiers	**Accuracy**
Prefix(s) + Stem + Tag + Suffix(s)	93.37
Prefix(s) + Stem + Tag + Suffix(s) + *Lemma*	93.11
Prefix(s) + Stem + Tag + Suffix(s) + *Lemma + Root*	92.95
Prefix(s) + Stem + Tag + Suffix(s) + *Lemma + Root + Pattern*	92.95
Prefix(s) + Stem + Tag + Suffix(s) + *Lemma + Root + Pattern + Number*	92.41
Prefix(s) + Stem + Tag + Suffix(s) + *Lemma + Root + Pattern + Number+ Gender*	92.03
Prefix(s) + Stem + Tag + Suffix(s) + *Lemma + Root + Pattern + Gender + Number + Definiteness*	88.10

Finally, precision measurement was 93.37% while recall measurement was 96.9%. The percentages are expected to increase by implementing the improvements while working on larger amounts of data. Figure 7 shows an example of some features of BASMA's results.



Word	lemmaid	pr1	pr2	pr3	stem	suf1	suf2	gen	num	def	root	stem_pattern
/D					BOF_Doc							
/T					BOF_Tit							
العيوب	Eayob	Al/DET			Euyuwb/NOUN			MASC	PL_BR	DEF	Eyb	fuEuwl
الفنية	fan~iy~	Al/DET			fan~iy~/ADJ	ap/NSUFF		FEM	SG	DEF	fnn	faEoliy~
..					Punc							
سبب	sabab				sabab/NOUN			MASC	SG	EDAFAH	sbb	faEal
كارثة	kArivap				kAriv/NOUN	ap/NSUFF		FEM	SG	EDAFAH	krv	faAEil
الأنابيب	>unobuwb	Al/DET			>anAbiyb/NOUN			MASC	PL_BR	DEF		
T/					EOF_Tit							
/P					BOF_Prg							
تحقيق	taHoqiyq				taHoqiyq/NOUN			MASC	SG	EDAFAH	Hqq	tafoEiyl
محمد	muHam~ad				muHam~ad/NOUN_PROP			MASC	SG	DEF	Hmd	mufaE~al
هندي	hinodiy~				hinodiy~/NOUN_PROP			MASC	SG	DEF		
P/					EOF_Prg							
/P					BOF_Prg							
المجتمع	ja$aE	Al/DET			ja$aE/NOUN			MASC	SG	DEF	j$E	faEal
..					Punc							
ونقص	naqoS	wa/CONJ			naqoS/NOUN			MASC	SG	EDAFAH	nqS	faEol
الكميات	kam~iy~ap	Al/DET			kam~iy~/NOUN	At/NSUFF		FEM	PL	DEF	kmm	faEoliy~
المطروحة	maToruwH	Al/DET			maToruwH/ADJ	ap/NSUFF		FEM	SG	DEF	TrH	mafoEuwl
هما	humA				humA/PRON			MASC	DU	DEF		
المتهمان	mut~aham	Al/DET			mut~aham/NOUN	Ani/NSUFF		MASC	DU	DEF	thm/whm	mufotaEal
الرئيسيان	ra}iysiy~	Al/DET			ra}iysiy~/ADJ	Ani/NSUFF		MASC	DU	DEF	r's	faEiyliy~
في	fiy				fiy/PREP							
أزمة	>azomap				>azom/NOUN	ap/NSUFF		FEM	SG	EDAFAH	'zm	faEol
أنابيب	>unobuwb				>anAbiyb/NOUN			MASC	PL_BR	EDAFAH		
البوتاجاز	buwtAjAz	Al/DET			buwtAjAz/NOUN			MASC	SG	DEF		
التي	Al~atiy				Al~atiy/REL_PRON			FEM	SG	DEF		
وصفها	waSaf-i				waSaf/PV	a/PVSUFF_SU	hA/PVSUFF_I				wSf	faEal
كثير	kaviyr				kaviyr/NOUN_PROP			MASC	SG	DEF	kvr	faEiyl
من	min				min/PREP							
المواطنين	muwATin	Al/DET			muwATin/NOUN	iyna/NSUFF		MASC	PL	DEF	wTn	mufaAEil
بأنها	>an~a	bi/PREP			>an~a/SUB_CONJ	hA/PRON_3F						
مفتعلة	mufotaEal				mufotaEal/NOUN	ap/NSUFF		FEM	SG	EDAFAH	fEl	mufotaEal
بعد	baEod				baEod/NOUN			MASC	SG	EDAFAH	bEd	faEol
أن	>an~a				>an~a/SUB_CONJ							

Figure 7: BASMA output results.

5 COMPARING BASMA WITH MADA

MADA (Morphological Analysis and Disambiguation for Arabic) is selected to be compared with BAMAE since both of them use Buckwalter's output analyses to help in disambiguating the Arabic texts. The primary purpose of MADA 3.2 is to extract as much linguistic information as possible about each word in the text, from given raw Arabic text, in order to reduce or eliminate any ambiguity concerning the word. MADA does this by using ALMORGEANA (an Arabic lexeme-based morphology analyzer) to generate every possible interpretation of each input word. MADA then applies a number of language models to determine which analysis is the most probable for each word, given the word's context.

In order to compare between BASMA and MADA, a text; to be used to evaluate both systems, was selected from ICA training data to facilitate the comparing process. To make the comparing process more accurate some modifications have been done in MADA's format to be compatible with BASMA's format. For example, in the number qualifier the feature of singular (s) was modified to be (SG), in the case qualifier the feature of nominative (u) was modified to be (NOM), in the tags qualifier the verbs were handled with relation to aspect and stem category. The comparing process will be done among some qualifiers; diacritization, tags, stems, number, gender and definiteness including Arabic words only as Table 2 shows:

TABLE 3
COMPARING RESULTS BETWEEN BASMA AND MADA

Qualifier	BASMA	MADA
Diacritization	91.11	78.78
Tags	95.94	85.28
Stems	97.08	91.34
Number	99.10	64.93
Gender	99.12	66.67
Definiteness	97.53	60.61

There are some notes that must be taken into consideration:

- The problems of detecting the diacritization in BAMAE are related to either the wrong prediction of the case ending or wrong prediction of the whole solution.
- The problems of detecting the diacritization in MADA are related to the wrong prediction of the case ending, wrong prediction of the whole solution, missing some diacritics in some words, or missing all diacritics in some words.
- The problems of detecting the tags in MADA are related to either the wrong prediction of the tags or the differences in some tags from BASMA. For example the adverbs of time or place in BASMA are assigned with 'NOUN_ADV(T)' or 'NOUN_ADV(P)', while they are assigned with 'NOUN', sub conjunction 'SUB_CONJ', and preposition 'PREP' in MADA. This happens as a result of using BAMA's output without enhancing these tags. In addition, the wrong concatenations of BAMA's output causes problems in detecting some tags.
- The problems of detecting stems in both BASMA and MADA are related to the wrong prediction of the solution.
- The problem of detecting number, gender and definiteness in MADA are related to using BAMA's output without regarding the morphosyntactic properties.
- The cases in BASMA and MADA can't be compared, since MADA assigns case without regarding the diacritics of the case. For example, it assigns the accusative case 'ACC' for both 'a/ACC' and 'i/ACC' which are differentiated in BASMA.
- There are some qualifiers in BASMA which are not used in MADA; Root and Stem Pattern. The root qualifier has been assigned with accuracy 99.45% while the stem pattern qualifier has been assigned with accuracy 96.34%.
- The lemma qualifier has been assigned in BASMA with accuracy 97.64%, while it is not used in MADA.

6 CONCLUSIONS

About 20 million words have been disambiguated using (BASMA). The evaluation has been done using precision and recall measurements for 1,000,000 words. Precision measurement was 93.37% while recall measurement was 96.9%. The percentages are expected to increase by implementing the improvements while working on larger amounts of data. If the analysis tools reach a deadlock and cannot improve any more enhancements, the data will be corrected manually.

REFERENCES

[1]M. Gridach& N. Chenfour, *Developing a new system for Arabic morphological analysis and generation*. In Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP'11) (pp. 52-57), November 2011.

[2]A. Pasha, A. Mohamed, M. Diab, A. El Kholy, R. Eskander, N. Habash , M. Pooleery, O. Rambow& R. Rohth, *Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic*. In Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2004.

[3] M.Sawalha, E. Atwell &M. Abushariah, *SALMA: Standard Arabic Language Morphological Analysis*. In Communications, Signal Processing, and their Applications (ICCSPA) 1st International Conference on (pp. 1-6). IEEE, February 2013.

[4]S. Alansary, M.Nagi& N. Adly, *Towards Analysing the International Corpus of Arabic (ICA): Progress of Morphological Stage*. In Proceedings of 8th International Conference on Language Engineering, Egypt, 2008 .

[5] S.Alansary. 2012, BAMAE: *Buckwalter Arabic Morphological Analyser Enhancer*. In Proceedings of 4th international conference on Arabic language processing, Mohamed Vth University Souissi, Rebate, Morocco, May 2-3, 2012.

[6]S.Alansary&M. Nagi,*The International Corpus of Arabic: Compilation, Analysis and Evaluation*. ANLP August 2014.

[7] N.Habash, *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers, 2010.

[8] R. Al-Sabbagh&R.Girju, A supervised POS tagger for written Arabic social networking corpora. InJancsary, J., editor, Proceedings of KONVENS, pages 39–52. ÖGAI, September. Main track: oral presentations, 2012.

[9]E. Mohamed, B. Mohit& K. Oflazer, *Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic*.InProceedings of the LanguageResources and Evaluation Conference (LREC), Istanbul, 2012.

[1 0]N. Habash, R. Eskander, & A. Hawwari.*A Morphological Analyzer for Egyptian Arabic*. InNAACL-HLT Workshop on Computational Morphology and Phonology (SIGMORPHON), pages1–9, 2012.
[11] NHabash, R. Roth, O.Rambow, R.Eskander, & N. Tomeh, *Morphological Analysis and Disambiguation for Dialectal Arabic*. InProceedings of Conference of the North American Chapter of theAssociation for Computational Linguistics: *Human Language Technologies*(NAACL-HLT), Atlanta, GA, 2013.
[12]N. Habash& O. Rambow, *Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. InProceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 573–580, Ann Arbor, Michigan, 2005
[13] R. Roth, O. Rambow, N. Habash, M. Diab& C. Rudin, *Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking*. In proceedings of the Conference of the Association for Computational Linguistics (ACL); Companion Volume, Short Papers, Columbus, Ohio, June. Associationfor Computational Linguistics, 2008 .
[14]N. Habash, O. Rambow& R. Roth, *MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization*. In Proceedings of the 2nd International Conference on Arabic Language Resources andTools. The MEDAR Consortium, April, 2009.
[15] M.Diab, K. Hacioglu, & D.Jurafsky, *Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking*. In proceedings of Arabic ComputationalMorphology: Knowledge-based and Empirical Methods. Kluwer/Springer, 2007.
[16] T. Buckwalter, *Buckwalter Arabic Morphological Analyzer Version 2.0*. Linguistic Data Consortium (LDC) catalogue number LDC2004L02, ISBN 1-58563-324-0, 2004.
[17]N. Habash& O. Rambow,Arabic diacritization through full morphological tagging, In the Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (pp. 53-56). Association for Computational Linguistics, 2007.
[18] R. Roth, O.Rambow, N.Habash, M.Diab&C.Rudin, *Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking*. In Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages (117–120), Columbus, Ohio, USA, June 2008.

BIOGRAPHY

Dr. SamehAlansary

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

محلل مكتبة الأسكندرية الصرفي للعربية المعاصرة (BASMA)

سامح الأنصاري

مدير مركز اللغويات الحاسوبية العربية – مكتبة الأسكندرية
sameh.alansary@bibalex.org

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الأسكندرية

ملخص—يعد الصرف العربي أحد التحديات الأساسية في الأنظمة المستخدمة في المعالجة الآلية للغة العربية. فالعربية غنية بالكثير من التنوعات والتعقيدات الصرفية حيث نجد أنه من الجذر الواحد يمكن توليد العديد من الكلمات المختلفة في الوزن الصرفي. تركز هذه الورقة الضوء على أحد المحللات الصرفية الآلية الذي تم بناؤه في مكتبة الأسكندرية (المحلل الصرفي للغة العربية المعاصرة لمكتبة الأسكندرية). وهذا المحلل يقوم بتحليل الكلمات تبعا لتواردها في سياقات مختلفة بالاعتماد على التحليلات الصرفية الواردة من المحلل الصرفي الشهير تيم باك ولتر. فيقوم هذا المحلل بمعالجة المشاكل الواردة من باك وولتر، كما يعتمد في عملية فك اللبس الصرفي على نظام هجين يعتمد على بعض القواعد اللغوية وبعض النماذج اللغوية الإحصائية المستخلصة من عينة لغوية، وهذهالعينة اللغوية عبارة عن مجموعة نصوص محللة تحليلا صرفيا، وقد وصلت نسبة الصحة في هذا المحلل الصرفي إلى 93.37% حيث استطاع المحلل التعرف على 96.9% من التحليلات الصرفية للكلمات. ومن المتوقع أن تزيد هذه النسبة بتطبيق مزيد من التحسينات على ذلك المحلل.

Part-of-Speech Tagging and Disambiguation for Arabic Language Understanding

Sameh Alansary

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
Bibliotheca Alexandrina, Alexandria, Egypt
sameh.alansary@bibalex.org

***Abstract*—There are different approaches to the problem of assigning each word in a text with a parts-of-speech tag, which is known as Part-Of-Speech (POS) tagging as well as many approaches to the problem of disambiguation in languages. The paper introduces general definitions about the POS tagging and disambiguation. The topic has a great significance in the Natural language processing (NLP). After general definitions about the topics, a more detailed explanation is provided for rule-based (constraint-based) part-of-speech tagging and morphological disambiguation system.The introduced system has been incorporated in many NLP applications such as Language-to-Interlanguage-to-Language System Based on UNL (LILY) and the knowledge extraction system (KEYS).The percentage of accuracy is 95% while the percentage of errors is 5%.**

1 INTRODUCTION

It has recently become clear that automatically extracting linguistic information from a sample text corpus can be an extremely powerful method of overcoming the linguistic knowledge acquisition bottleneck inhibiting the creation of robust and accurate natural language processing systems. A number of part-of-speech taggers are readily available and widely used, all trained and retrainable on text corpora [1]–[4].

There are two methodologies concerning part of speech tagging and disambiguation; supervised taggers, which typically rely on pre-tagged corpora to serve as the basis for creating any tool to be used throughout the tagging process. Pre-tagged models are used to acquire information about the tag-set, word-tag frequencies, rule sets etc. Therefore, any increase in the size of corpora will generally lead to a better performance of the models.

Unsupervised taggers do not require a pre-tagged corpus, but instead they use sophisticated computational methods such as the Baum-Welch algorithm to automatically detect word groupings (i.e. tag sets) and those automatic groupings could be used either to calculate the probabilistic information needed by stochastic taggers or to induce the context rules needed by rule-based systems.

In terms of those two methodologies, there are different approaches that have been used for Part-of-Speech (POS) tagging; rule-based approach, stochastic approach and the transformation-based approach. In this section, each of them will be introduced in details.

Rule-based approach uses contextual information to assign tags to unknown or ambiguous words. These rules are often known as context frame rules. As an example, a context frame rule might say something like: “if an ambiguous/unknown word X is preceded by a determiner and followed by a noun, tag it as an adjective”[5].

Two stage architecture was applied for automatically assigning part-of-speech. Firstly, in the initial stage a dictionary is used in order to assign each and every word a list of potential parts of speech. In the second stage, large lists of hand-written disambiguation rules are used with the purpose ofreducing this list to just a single part-of-speech for each word. Supervised training is required usually in the rule based tagging models that is pre-annotated corpora. The main disadvantages of the rule based systems are the necessity of a linguistic background and the need to manually construct the rules.In addition to contextual information, many taggers use morphological information to aid in the disambiguation process. One such rule might be: “if an ambiguous/unknown word ends in an -ing and is preceded by a verb, label it a verb”[5].

Some systems go beyond using contextual and morphological information by including rules pertaining to such factors as capitalization and punctuationfor English language. Information of this type is of greater or lesser value depending on the language being tagged. In German for example, information about capitalization proves to be extremely useful in the tagging of unknown nouns.

Rule based taggers most commonly require supervised training; however, very recently there has been a great deal of interest in automatic induction of rules. One approach to automatic rule induction is to run an untagged text through a tagger and see how it performs. Then, the output of this first phase is manually revised and corrected if there are any erroneously tagged words. The properly tagged text is then submitted to the tagger, which learns correction rules by comparing the two sets of data.

The stochastic approach uses a large training corpora to get statistical information in order to choose the most probable tag for a word. A part of the corpus is used inthe training phase in order to get a statistical model, which will be used to tag untagged texts and the remaining of the corpus is used to test the statistical model.
The simplest stochastic taggers disambiguate words based solely on the probability that a word occurs with a particular tag; the tag which is most frequent in the training set is the one assigned to an ambiguous instance of that word[5]. The problem with this approach is that while it may yield a valid tag for a given word, it can also yield inadmissible sequences of tags. Most of the probabilistic methods are based on Hidden Markov Model (HMM), Maximum Likelihood Estimation, Decision Trees, Maximum Entropy, Support Vector Machines and Conditional Random Fields, but the most common techniques are HMM and N-grams.

The n-gram technique which calculates the probability of a given sequence of tags can be used as an alternative to the word frequency approach. Using this technique, the best tag for a word can be determined by the probability that it occurs with the n previous tags, where the value of n is set to 1, 2 or 3 for practical purposes. These models are termed as unigram, bigram and trigram.
Before a N-gram tagger could be used in tagging data, it must be trained on a training corpus. It uses the corpus to determine which tags are most common for each word. The N-grams taggers will assign the default tag "None" to any token that was not encountered in the training data.While, the intuition behind HMM and all stochastic taggers is a simple generalization of the "pick the most likely tag for this word". The unigram tagger only considers the probability of a word for a given tag; the surrounding context of that word is not considered.
On the other hand, for a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes the formula: P (word | tag) * P (tag | previous n tags)

The transformation-based approach combines the rule-based approach and statistical approach. It picks the most likely tag based on a training corpus and then applies a certain set of rules to see whether the tag should be changed to anything else. It saves any new rules that it has learnt in the process, for future use. Taken together, the transformation with rewrite rule and triggering environment when applied to the word can correctly change the mis-tagged[6]. One example of an effective tagger of this category is the Brill tagger technique[7].
In 1990's, Brill introduced a method to induce the constraints from tagged corpora, which is called transformation based error-driven learning. Nowadays, all of the approaches are used together to get better results.

In this paper, we present a POS tagger and disambiguation system for Arabic language understanding that performs with high efficiency for Arabic language. The system is based on the rule based approach that uses contextual information to assign tags to unknown or ambiguous words; however, it may also use the unigram in order to choose the most frequent tag for a specific word. The system works within the framework of the Universal Networking Language (UNL) which is composed of Universal Words (UWs), Relations and Attributes. UWs constitute the vocabulary of the UNL language; they are labels that stand for abstract language-independent units of knowledge (concepts) belonging to any of the open lexical categories (nouns, verbs, adjectives or adverbs). Relations and Attributes, on the other hand, represent the syntax of this language. Relations stand for the links between the UWs in a given sentence [8]. The UNL system is robust enough by enriching the dictionary with all the levels of linguistic information (morphological, semantic, syntactic information) which in turn have a great effect in disambiguating the words. Moreover, the notion of concepts in defining the words has a vital role in disambiguating the words in our disambiguation system, which is our main focus in this paper.
In this paper, a training and test corpus will be described in section 3, then the system algorithm will be presented as well as our POS tagset and the used tool IAN, in section 4; next how these methodologies perform for Arabic will be presented in section 5 and 6; finally section 7 will include the evaluation of our results and section 8 will conclude the paper.

2 THE STATE OF THE ART

The first trials for building a rule-based POS tagger was by Klein and Simmons. Their main purpose was to avoid the labor of constructing a very large dictionary. Their algorithm uses a set of 30 POS categories. First, itlooks each word in up dictionaries, then checks for suffixes and special characters as clues. Then, the context frame tests are applied. These work on scopes bounded by unambiguous words. However, Klein and Simmons have specified an explicit limit of three ambiguous words in a row. The pair of unambiguous categories bounding such scope of ambiguous words, is mapped into a list. The list includes all known sequences of tags occurring between the particular bounding tags; anysequences that have the correct length become a candidate. Then, the program then matches the candidate sequences are matched against the ambiguities remaining from previous steps of the algorithm. When there is only one sequence that is possible, the disambiguation is considered successful. This algorithm correctly and unambiguously tags about 90% of the words in several pages of the Golden Book Encyclopedia [5].
Moreover, one of the most important taggers,is TAGGIT.It was developed by Greene and Rubin in 1971. The tag set used is very similar to that of Klein and Simmons, but somewhat larger, at about 86 tags. The dictionary used is derived

from the tagged Brown Corpus, rather than from the untagged version. In TAGGIT, the task of category assignment is divided into two phases; initial (potentially ambiguous) tagging, and disambiguation. The tagging process is performed as follows; first, the program consults an exception dictionary of about 3,000 words. Among other items, this contains all known closed-class words. It is able to handle various special cases, such as words with initial "$", contractions, special symbols, and capitalized words. Subsequently, a word's ending is checked against a list of suffixes of about 450 strings, that was derived from the Brown Corpus. In case after going through all these steps TAGGIT has not assigned some tag(s) , the word is tagged as a noun, a verb and an adjective, in order to provide the disambiguation routine with something to work with. This tagger correctly tags approximately 77% of the million words in the Brown Corpus (the rest is completed by human post-editors)[5].

The Constraint Grammar is a very successful constraint-based approach for morphological disambiguation. It was developed in Finland, from 1989 to 1992, by four researchers: Fred Karlsson, ArtoAnttila, JuhaHeikkila and AtroVoutilainen. In this framework, the parsing process is divided into seven modules; four of them are related to morphological disambiguation, the other three are used for parsing the running text. The context-dependent morphological disambiguation is one of the most important steps of Constraint Grammar, where ambiguity is resolved using some context-dependent constraints. For this purpose they wrote a grammar, which is composed of a set of constraints based on descriptive grammars and studies of various corpora. Each constraint is a quadruple consisting of domain, operator, target and context condition(s).

Reference [9] has implemented a rule based POS tagger. However, this tagger requires laborious work, it requires writing hand crafted rules by human experts and continuous efforts from many linguists for many years. Moreover, the feasibility of their proposed rule based method for Bangla is questionable, since they have not reported a performance analysis of their work.

The Lancaster-Oslo-Bergen (LOB) Corpus tagging algorithm, later named as CLAWS is similar to TAGGIT program. The tag set used is very similar to that of the TAGGIT program, but rather larger, at about 130 tags. Moreover, the dictionary used is derived from the tagged Brown Corpus, rather than from the untagged version. CLAWS main contribution is the use of a matrix of collocation probabilities, indicating the relative likelihood of co-occurrence of all ordered pairs of tags and this matrix can be mechanically derived from any pre-tagged corpus. CLAWS had made extensive use of the Brown Corpus, with 200,000 words. CLAWS has been applied to the entire LOB Corpus with an accuracy of between 96% and 97%.

This general approach has several advantages over the rule-based approach. First, it can handle scopes of unlimited length. Second, it is possible to give a precise mathematical definition for the fundamental idea of CLAWS. However, CLAWS main drawback is being time- and storage-inefficient in the extreme.

Later in 1988, DeRose have tried to handle the inefficiency problem of the CLAWS, so he proposed a new algorithm called VOLSUNGA. The algorithm depends on a similar empirically-derived transitional probability matrix to that of CLAWS, and has a similar definition of optimal path. The tag set consists of 97 tags. The optimal path is defined to be the one whose component collocations multiply out to the highest probability. However, the more complex definition applied by CLAWS, using the sum of all the paths at each node of the network, is not used. By applying this change VOLSUNGA has overcame the complexity problem. Application of the algorithm to Brown Corpus resulted with the 96% accuracy.

A form of Markov model has also been widely used in statistical approaches. This model is based on the assumption that words depend probabilistically on just their part-of-speech category, which in turn depend solely on the categories of the preceding two words for each word. Two types of training have been used with this model. The first uses a tagged training corpus. The second method does not require a tagged training corpus. The Baum-Welch algorithm could be used in this situation. In this case, the model is called a Hidden Markov Model (HMM), as state transitions (i.e., part-of-speech categories) are assumed to be unobservable. Hidden Markov Model taggers and visible Markov Model taggers are among the most efficient of the tagging methods and they could be implemented using the Viterbi algorithm.

Tree Tagger: it is a language-independent POS tagger, free for academic use, easily downloaded, comes with free language models for approximately 10 languages. However, in order to be downloaded, it requires a signed license agreement, comes with language models for German and English. SVM Tool: It is open source tagger with models for Catalan, English, and Spanish. However, it must be trained by using the non open-source SVM light software which can be used for freely for academic purposes only. It is based on Support Vector Machines.

Stanford Log-linear Part-Of-Speech Tagger is also an open source tagger, providing models for English, Arabic, Chinese, and German, It is based on the Maximum Entropy framework. It can be trained on any language on a POS-annotated training text for the language.

Apache UIMA Tagger: is an open source taggerthat comes with models for English and German. It is HMM tagger as part of the Apache Unstructured Information Management Architecture (UIMA) framework.

Chris Biemann'sunsupos: it is unsupervised open source POS tagging. It provides models for a number of languages including Danish. It is not clear what type of material the Danish model is based on, it is unsupervised POS tagging that does not require an annotated training corpus. Instead, word categories are determined by analyzing a large sample of monolingual, sentence-separated plain text. The tag set probably cannot be determined by the user/linguist.

Eric Brill's simple rule-based part of speech tagger: the source code is accessible at Plymouth Tech, it is based on rules derived from a training corpus. It is implemented in C language. It is also implemented in Python as part of NLTK.

Sujit Pal's HMM-based tagger: itssource code is available inSujit Pal's blog. It comes with a model for English derived from the Brown Corpus, it is a HMM tagger based on [10].

For Arabic language, there are some trials and the most common are: Abuleil, S. Alsamara, Kh. and Evens, M [11], have described a learning system that can analyze Arabic nouns to produce their paradigms with respect to both gender and number using a rule-base that uses suffix analysis as well as pattern. Reference [12] has described a system for automatically building an Arabic lexicon by tagging Arabic newspaper text. References [12] and [13] have described some initial findings in the development of an Arabic part-of-speech tagger. ShreenKhoja, Roger Garside and Garry Knowles have proposed a tag-set for the morpho-syntatic tagging of Arabic that described morpho-syntactic tag-set that is derived from the ancient Arabic grammar. Reference [14] documents some of the hurdles that were encountered during a long semester project to implement Brill's POS tagger for Arabic. Reference [15] describes the design and implementation of a question answering (QA) system called QARAB. John Maloney and Michael described a fast, high-performance name recognizer for Arabic texts. It combines a pattern-matching engine and supporting data with a morphological analysis component. Reference [16], in his thesis has implemented an industry-quality computational processor of the Arabic morphology − called Morpho3− along with a host of dependent applications as well as complementary utilities [17].

3 CORPUS COMPILATION

In order to build an adequate corpus to be representative of the different issues of the disambiguation, 105,878 words of Arabic text were compiled from different resources which are parallel data, texts compiled from the Arabic Wikipedia and texts compiled from the Arabic book"مصر أصل الحضارة"EGYPT, where the civilization began'. The corpus includes documents from various genres and domains which means that the coverage rate is high and the corpus is considered robust. It is segmented automatically into sentences. The total number of sentences is 21,021 sentences. The maximum length of the sentences is 17 words. This corpus has been divided into a training corpus which contains 79,408 words and a test corpus which contains 26,196 words. By carefully studying the training corpus, different issues and cues have been detected. These issues and cues will be discussed in section 6. An example of the tagger output is shown in figure 1, the untagged corpus:

و_COO قد_PTC ترك_VER ال_ART إنسان_NOU ال_ART بدائي_ADJ رسوم_NOU ا_SUF لا_NEG يمكن
_AUX أن_PTC تفسر_VER ب_PER ال_ART غاية_NOU ال_ART نفعية_ADJ. إذ_PTC أن-PTC ال_ART
روح_NOU ال_ART فني_ADJ واضح_ADJ في_PER ها_SPR. و_COO لكن_COO يجب_AUX أن_PTC
نعترف_VER ب_PER أن_PTC ال_ART قبر_NOU ال_ART مصري_ADJ كان_AUX أحد_NOU ال_ART
أصول_NOU. أو_COO على الأقل_AAV ال_ART وسائل_NOU ل_PER ل_PER نحت_NOU و_COO ال
_ARTرسم_NOU. ف_COO كان_AUX ال_ART تمثال_NOU يصنع_VER من_PER ال_ART خشب_NOU
أو_COO ال_ARTحجر_NOU. و_COOوجه_NOU ال_ARTمومياء_NOU يرسم_VER
ب_PERال_ARTألوان_NOU.و_COOقبل_PER أسابيع_NOU ذكرت_VER ال_ARTصحف_NOU
خبر_NOUا_SFX غريب_ADJا_SFX.
هو_PPR أن_PTC بعض_QUA ال_ARTلصوص_NOU سرقوا_VER جثة_NOU وجيه_NOU من_PER
وجهاء_NOU المنيا_PPN. و_COOلابد_AUX أن_PTC هؤلاء_DEM ال_ARTلصوص_NOU هم_PPR من_PER
سلالة_NOU أولئك_DEM ال_ARTلصوص_NOU الذين_RPR كانوا_AUX يسرقون_VER قبور_NOU

Figure 1: Morphological analyzer output of the example sentence.

4 SYSTEM ALGORITHM

This section discusses the linguistic and technical resources used to build an efficient rule based system for POS tagging and disambiguation. This section will describe the developed dictionary; its format, the different linguistic information provided to each word and the environment in which it was developed. Furthermore, the used tool will be presented and its algorithm, its grammar formalism and the different types of rules and its format will be explained.

A. Dictionary

The Arabic dictionary is a bilingual dictionary, where Arabic natural language words are matched with their corresponding abstract Universal Words (UWs) (concepts), along with the corresponding linguistic features. This dictionary is developed through the UNLarium[1] which is an integrated development environment for producing language resources for natural language processing (NLP). It is mainly a web-based database management system, where registered users are able to create, to edit and to export dictionary entries according to the UNDL foundation[2] standards for language engineering. However, the UNLarium environment and the data it contains could be used in several NLP systems, other than UNL-based applications. Furthermore, the system is meant to be used as a research workplace for exchanging information and testing several linguistic constants that have been proposed for describing and predicting natural language phenomena.

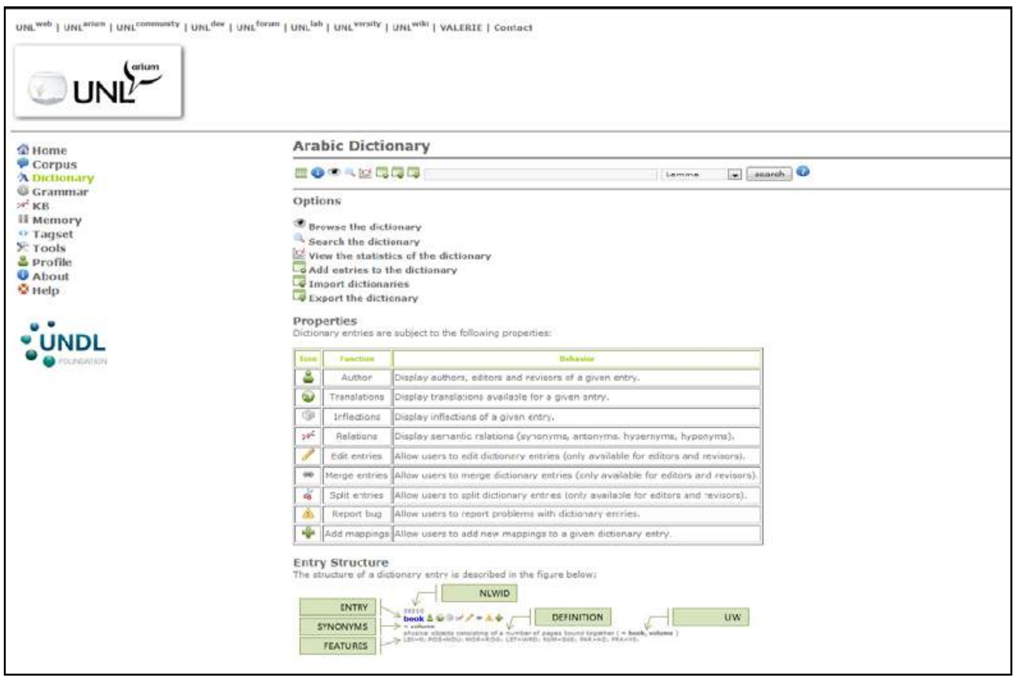


Figure 2: The UNLarium environment

The dictionary follows the format:

```
■ UNL-NL Dictionaries

[NLW]  {ID}  "UW"  (ATTR , ... )  < FLG , FRE , PRI >; COMMENTS
```

Figure 3: The Dictionary format

Where: NLW is the Arabic word. It can be a multiword expression, a compound, or a simple word. UW is the abstract concept representing the natural language word; they are "universal" in the sense that they are uniform identifiers to the entities defined in the UNL Knowledge Base, which is expected to map everything that we know about the world, and that is used to assign translatability to any concept. ATTR is the list of linguistic features of the NLW, the linguistic features of the dictionary entries have been assigned to all words through the UNLarium encompassing different linguistic levels: morphological information, syntactic information and semantic information see the entry in figure 4.

---

[1] http://www.unlweb.net/unlarium

2 The UNDL Foundation is a non-profit organization based in Geneva, Switzerland, which has received, from the United Nations, the mandate for implementing the UNL

```
[باحث]{116422}"110523076"(LEMMA=باحث, BF=باحث, LEX=N, POS=NOU,
LST=WRD, GEN=MCL, NUM=SNG,
PAR=M532, FRA=Y0, ABN=CCT,
ANI=ANM, SEM=HUM)<ara,11,1>;
```

Figure 4: Arabic dictionary entry

UNL uses a standard and universal list of features (Tagset) to describe all types of the linguistic information concerning every Arabic word. This tagset is a set of features in a UNL dictionary depending on the structure of the natural language. Several of those linguistic constants have been already proposed in the Data Category Registry (ISO 12620), and represent widely accepted linguistic concepts. The purpose of this tagset is providing the technical means for describing any linguistic behavior which should be done in a highly standardized manner, so that others could easily understand and exploit the data for their own benefit. The main intention is to create a harmonized system in order to make language resources as easily understandable and exchangeable as possible, see the list of tags in figure 5.

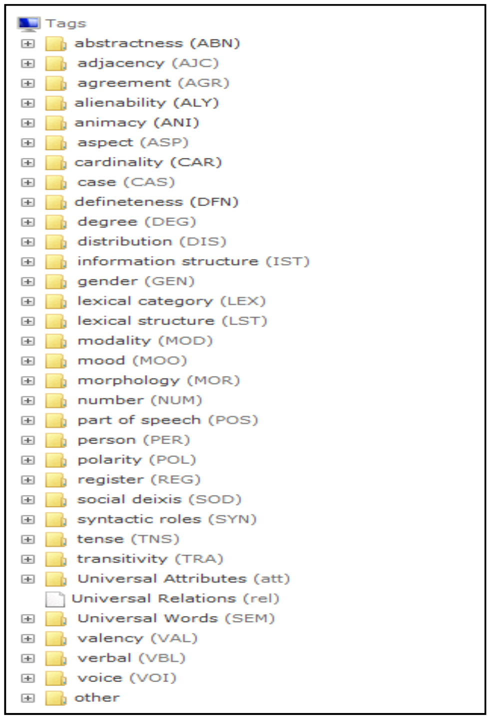


Figure 5: List of tags in alphabetical order

The linguistic information field inside the dictionary are four types: entry`s lemma, entry`s base form, list of simple features and a list of inflection rules. First, lemma is the canonical form of a lexeme, the word as it would appear in the dictionary. Lexemes, as a set of different word forms with different inflectional affixes, but with the same stem, are normally referred to by a citation (default) word form called lemma. The lemma, more generally referred to as headword, is essentially an abstract representation, subsuming all the formal lexical variations which may apply within the same lexeme. For instance, the lexeme comprising the word forms "قال", "يقول", "نقول", is normally referred to by the lemma "قال". Second, base form, or simply BF, is the form used to generate all variants of a given lexeme. The lemma is not always the most adequate form used to generate the inflections of a given lexeme.Third, a list of simple features describing the lexical structure of words; their part of speech (POS); gender and number for nouns; types of verbs with their transitivity, valancy and aspect; and much other information about adjectives and adverbs. Fourth, a list of inflection rules to describe the morphological behavior of Arabic words and to generate different word forms of each base form. For example the noun "باحث" 'researcher', has 12 different word forms that will be generated including the forms "باحث"'male researcher' - "باحثة"'female researcher'- "باحثان"'two male researchers' - "باحثتان"'two female researchers' - "باحثون"'male researchers'.

FRE is the frequency of NLW in natural texts.The same Arabic word may occur with different senses as in the word 'قصيدة' it might be the feminine form of the adjective of 'قصيد' 'broken' as in 'نافذة قصيدة' 'broken window' or it may be the noun 'قصيدة' 'poem'. Frequency specifies which senseof these two is the most frequent of this word and orders different senses from the most frequent to the least frequent. Frequency was detected through counting the occurrences of Arabic words and their possible senses in the ICA corpus. Using frequency helps in choosing the most frequent sense and reduces lexical ambiguity.

[قصيدة]{255278}"300289082"(LEMMA=قصيد,BF=قصيد,LEX=J,POS=ADJ,LST=WRD,GEN=FEM,NUM=SNG,DEG=PST,PAR=M716,FRA=Y0) < ar,**16**,0>;

[قصيدة]{276232}"106377442"(LEMMA=قصيدة,BF=قصيدة,LEX=N,POS=NOU,LST=WRD,GEN=FEM,NUM=SNG,PAR=M581,FRA=Y0,ANI=NANM,ABN=ABT,ALY=ALI,ANI=NANM,CAR=CTB,SEM=CMN,SFR=K0) < ar,**20**,1>;

Figure 6: Frequency in the Arabic dictionary.

PRI is the priority of NLW in natural texts. The same sense may have many word synonyms. As in the case of the concept 'begin (icl>start)', it is represented by two Arabic words 'بدأ' and 'شرع'. Priority specifies which wordof these two is the most common of this sense and orders different words from the most common to the less common. Priority was detected through counting the occurrences of Arabic words with the same sense in the ICA corpus.

[بدأ]{209302}"202608347"(LEMMA=بدأ,BF=بدأ,LEX=V,POS=VER,LST=WRD,TRA=TSTD,PAR=M706,FRA=Y0,SEM=STT) < ar,74,**20**>;

[شرع]{225121}"202608347"(LEMMA=شرع,BF=شرع,LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M107,FRA=Y524,SEM=STT) < ar,2,**15**,>;

Figure 7: priority in the Arabic dictionary

B. Used Tool (IAN)

This sub-section discusses the tool used in the disambiguation system. The UNDL foundation has developed a tool called Interactive Analyzer (IAN). IAN is a natural language analysis system. In its current release, it is a web application developed in Java and available at the UNLdev[3] [18]. It is a universal engine, IAN must be parameterized to the source languages with the dictionary and grammar files that are provided through IAN's interface.

IAN performs different procedures over the input file: Segmentation, i.e., the division of the input document into a series of processing units (sentences), which are processed one at a time. Tokenization, i.e., the identification of the tokens (lexical items) of each sentence of the input document. Disambiguation, i.e., the identification of the right sense of each token of the input document. Transformation, i.e., the application of the transformation rules of the grammar over each tokenized sentence in order to analyze the document syntactically and semantically [18].Tokenization and disambiguation phases will be the focus of the paper.

IAN follows general guidelines which illustrate how the tokenization algorithm works and this will be useful in building tokenization grammar later. The following is the general principles:

I. The tokenization algorithm is strictly dictionary-based:

The system tries to match the strings of the natural language input against the entries existing in the dictionary. In case it does not succeed, the string is treated as a temporary entry. There are no predefined tokens: spaces and punctuation signs have to be inserted in the dictionary in order to be treated as non-temporary entries. For instance, if the dictionary is empty, the string "Barking dogs seldom bite" will be considered as a single token. If the dictionary contains only the entry [d], the input will be tokenized as [Barking][d][ogssel][d][om bite].

II. The tokenization algorithm tries to match first the longest entries in the dictionary:

The system tries to match first the longest entries. If the dictionary contains only two entries: [d] and [do], the string "Barking dogs seldom bite" will be tokenized as [Barking][do][gssel][do][m bite], instead of [Barking][d][ogssel][d][om bite], because the length of [do] is larger than the length of [d].

III. The tokenization algorithm takes into consideration the frequency of the entries included in the dictionary (the most frequent entries come first):

The system observes the frequency defined in the dictionary. If the dictionary contains only two entries: [do] and [og], but the frequency of [og] is higher than the frequency of [do], the string "Barking dogs seldom bite" will be tokenized as [Barking d][og][s sel][do][m bite], instead of [Barking][do][gssel][do][m bite].

IV. The tokenization algorithm observes the order of the entries in the dictionary (the system selects the first to appear in case of same frequency):

3 http://dev.undlfoundation.org/index.jsp

The system observes the order defined in the dictionary. If the dictionary contains only two entries: [do] and [og], with the same frequency, but [og] appears first in the dictionary, the string "Barking dogs seldom bite" will be tokenized as [Barking d][og][s sel][do][m bite], instead of [Barking][do][gssel][do][m bite].

V. The tokenization algorithm goes from left to right:
The system tokenizes the leftmost entries first. If the dictionary contains only two entries: [do] and [og], with the same length and with the same frequency, the string "Barking dogs seldom bite" will be tokenized as [Barking][do][gssel][do][m bite], instead of [Barking d][og][s sel][do][m bite], because [do] appears before [og].

VI. The tokenization algorithm is case-insensitive, except in case of regular expressions:
The string "a" is matched to both [a] and [A], but the entry [/a/] will match only the string "a".

VII. The tokenization algorithm assigns the feature TEMP (temporary) to the strings that were not found in the dictionary:
If the dictionary contains only the entry [d], the input will be tokenized as [Barking][d][ogssel][d][om bite], and the tokens [Barking],[ogssel] and [om bite] will receive the feature TEMP.

VIII. The tokenization algorithm blocks tokens or sequences of tokens prohibited by D-rules:
If the disambiguation grammar contains the rule ("do")("gssel")=0, and the dictionary contains only two entries: [do] and [og], the string "Barking dogs seldom bite" will be tokenized as [Barking d][og][s sel][do][m bite], regardless the frequency and the order of [do] and [og], because the possibility of "do" being followed by "gssel" is prohibited by the grammar.

IX. In case of several possible candidates, the tokenization algorithm picks the ones induced by D-rules, if any:
If the disambiguation grammar contains the rule ("og")("s sel")=1, and the dictionary contains only two entries: [do] and [og], the string "Barking dogs seldom bite" will be tokenized as [Barking d][og][s sel][do][m bite], regardless the frequency and the order of [do] and [og], because the possibility of "og" being followed by "s sel" is induced by the grammar. Retokenization can be done only in the case of entries having the feature TEMP.

There are two different types of rules that are used in IAN; disambiguation (D-rules) and transformation rules (T-rules). [18].

1) *Disambiguation Rules (*D-rules*):* D-rules or disambiguation rules are used to prevent wrong lexical choices, to provoke best matches and to check the consistency of graphs, trees and lists. D-rules follow the general syntax:

STATEMENT=P;

where STATEMENT is the left side (condition) and P, which can range from 0 (impossible) to 255 (necessary), is the probability of occurrence of the STATEMENT. There are two types of disambiguation rules:

1- Linear disambiguation rules, when the rule applies over lists of nodes.
2- Non-linear disambiguation rules, when the rule applies over non-linear relations between words.

Linear disambiguation rules apply over the natural language list structure to constrain word selection (dictionary retrieval). They have the following format:

(word 1)(word 2)(...)(word n)=P;

where (word 1), (word 2) and (word n) are word, and P is an integer (from 0 to 255).
Non-linear disambiguation rules apply over the syntactic structure. They have the following format:

REL1(arg1;arg2;...)REL2(arg3;arg4;...)...RELN(argx;argy;...)=P;

where REL1, REL2 and REL2 are syntactic or semantic relations, with their corresponding arguments (arg1, arg2, ...), and P is an integer (from 0 to 255).

2) *Transformation Rules (*T-rules*):* T-rules are rules that alter the state of words. The transformation rules follow the very general formalism :

α:=β;

where the left side α is a condition statement, and the right side β is an action to be performed over α.

There are special types of transformation rules. A-rule is a specific type of T-rule used for affixation (prefixation, infixation and suffixation). C-rule is a specific type of T-rule used for composition (word formation in case of compounds and multiword expressions). L-rule is a specific type of T-rule used for handling word order. N-rule is a specific type of T-rule used for segmenting sentences and normalizing the input text. S-rule is a specific type of T-rule used for handling syntactic structures.

A lot of ambiguity problems could be solved through the two phases of T-rules:

- LL - List Processing (List-to-List)
- LT - Surface-Structure Formation (List-to-Tree)

The List to List (LL) rules are responsible for preprocessing the natural language input by analyzing it morphologically in order to match the input words with the dictionary entries and assign each stem to the concept it conveys.

Then, List-to-Tree Rules (LT) parse the resulting list structure into a surface tree structure. This type of rules is only employed in the analysis process. They specify the syntactic relations between the words of the input sentence to form a surface tree structures.

The following sections will discuss the usage of the different types of rules and their role in solving the disambiguation issues.

5 TOKENIZATION AND PART OF SPEECH TAGGING

In this section, tokenization process and the problem of ambiguity will be covered. However, before the tokenization process began,a preprocessing phase should take place, if needed. The pre-processing phase is called normalization process; pre-processing rules (N-rules) apply over the string stream to fix the most common spelling mistakes. For example, a word like "موسيقى" 'music' it is common to be written wrongly as "موسقى". So, normalization rules substitute the wrong form by the right one. Then, the tokenization process begins. Tokenization is the process of splitting the natural language input into lexical items. The tokenization follows mainly the general guidelines stated previously in section 4. The tokenization process depends on D-rules.There are two types of disambiguation rules; negative and positive rules. Negative rules follow the same format mentioned in section 4 and they are used to prevent the sequence specified in the left side (condition). Positive disambiguation rules also follow the same format of D-rules, but the probability mentioned in the right side should be higher than 1. Generally, in the following, the usage of the D-rules clarified with different examples.

Tokenization starts with preventing joined lexical items; in Arabic, lexical items are separated with blank spaces. Then, it identifies the different suffixes and prefixes that could be attached to each lexical category. The tokenization process makes use of the engine`s algorithm, for example, the engine will automatically segment a word like 'الولد' "the boy" correctly, although the dictionary contains 'الو' "twist" and 'لد' "lod", as well as 'ولد' "boy" and, 'ال' "the" and both of 'الو' "twist" and 'ولد' "boy" have the same length.However, the frequency of both 'ال'"the" and 'ولد'"boy" will be the determining factor, since they are higher in the dictionary than the other two items. So, the engine is able to tokenize automatically some of words correctly based on the dictionary and assign the correct POS to words. On the other hand, the larger the number of entries in the dictionary, the more the ambiguity during tokenization increases. For example, the word "القلب" 'heart' would be automatically segmented as [الق]+[لب], given the fact that the dictionary includes [ART ال] 'the',[الق] 'throw', [لب] 'answer' and [N قلب] 'heart'. But, D-rule prevents two verbs to be joined without a blank space. So, it selects the [قلب] + [ال] as the appropriate combination. Also, the words that are not included in the dictionaryare considered by the tokenizer as a temporary entry (TEMP). As in a word like 'الابريسم' "Alibrism",it would be automatically segmented as [ال] "the'+[ابري] +[سم] "name". But, a D-rule prevents this sequence as the determiner 'ال' "the" is not an allowed as a prefix for verbs. Then, the lexical item would be retokenized as [الابري] "TEMP" +[سم] "name" which will also be refused by D-rules, because TEMP should be followed by blank space. Finally, the D-rules will select [TEMP الابرسيم] as the appropriate tokenization.

If words have spelling mistakes or morpho-syntactic changes they would be considered as a TEMP from the tokenization process, while they already exist in the dictionary. For example, the most common mistake in the Arabic writings is /Hamza/ in the initial position as in "اقتنع" 'convince'. Rules will try to investigate the morphological pattern of the wrongly spelled word by the regular expression techniques. For example, if a five-letters word begins with the sequence "/(ا|إ|أ).ت../" as in the pattern "افتعل" /?iftaʕala/, the wrong written /Hamza/ ("أ", "إ" or "آ") will be modified to the correct "ا" according to the Arabic grammar, then the correct concept will be retrieved from the dictionary. Many challenges arise in the spelling correction process; morphological patterns could be misleading sometimes, as in "أوهام" 'delusions', if the first Hamza was wrongly written as ("ا", or "آ"), it will be wrongly categorized under the morphological pattern "إفعال" /?ifʕal/ and the wrong Hamza will be modified to "إ", thereforethe new modified word "إوهام" should be retrieved, however no result will appear. So, rules will change the modified Hamza to the default Hamzain Arabic "أ"; rules consider "أ" as the default and the most common orthographical form for Hamza. Then, the right concept would be retrieved from the dictionary.

6 VALIDATION ANDDISAMBIGUATION

This module is concerned with preventing the wrong automatic lexical choices from the dictionary. Some linguistic indicators can help in solving the lexical ambiguity which are morphological, adjacency and structural indicators. The following sub-sections will discuss those three indicators.

A. Morphological indicators.

Affixation has an important role as the first level of part of speech disambiguation, as prefixes and suffixes are the smallest processing units rules can begin with. The rules used in this level of disambiguation are the D-rules. Prefixes can help as indicators in determining correct lexical choices. For example, in the word "الكتب", the noun "كتب" 'books' is chosen instead of the verb "كتب" 'write', since it is preceded by the definite article prefix "ال" 'the'. The rule in (1a) rejects this combination; (1a) states that if the definite article 'ال' 'the' which is an (ART) is followed by a verb (VER), then this combination should be rejected which is expressed in the rule as (= 0;). Moreover, suffixes can solve the lexical ambiguity, as in the word "أمه", if the conjunction "أم" 'or' is chosen instead of the noun 'mother', then this means that the conjunction is followed by the masculine third person pronoun suffix "ه" 'his'; however rule in (1b) rejects this structure. The rule (1b) states that if disjunction (COO) is followed by suffix (SFX), this structure should be rejected.

1- (a) (ART)(VER)=0;
(b) (COO, [أم])(SFX)=0;

Sometimes, suffixes can help in disambiguating verbs that underwent morpho-phonological changes such as "ضربني" 'he hit me', the automatic segmentation may choose the past feminine plural verb "ضربن" 'they (feminine) hit' + the 1st person pronoun "ي" 'me' instead of "ضرب"'hit' + "ن" that is added for morpho-phonological necessity + the 1st person pronoun "ي" 'me'. In the Arabic morpho-phonological system, the protection noon "نون الوقاية" is attached to verbs predicated to the object first person pronoun "ي" 'me'. The rule in (2) rejects the structure of a verb (V) followed by the suffix "ي" (1PS). In rule (2), the operator "|" is used to mean "or" to make the rule more comprehensive; to prevent the structure of a verb followed by the first person singular (1PS) *or* first person plural (1PP) pronouns.

2- (VER)(SFX,{1PS|1PP}) =0;

As protection noon is meaningless, therefore it will be deleted in a subsequent phase; this phase is responsible for retrieving the surface morphological form to the underlying form.

B. Adjacency indicators.

After disambiguating the POS on the word level, the role of the adjacent word will take its effect as the second level of disambiguation. D-rules will also be used in this level. In this level, the meaning and part of speech choice could be controlled.

1) *Number and Gender qualifiers:* There are two different meanings for the quantifier "كل" 'each' and 'all' as in "كل كتاب" 'each book' and "كل الكتب" 'all books'. It is determined by the number of the following noun, as stated in the rule in (3a); if the quantifier "كل" is tokenized as to mean 'each' and not all (^@all), followed by a blank (BLK), definite article (ART) and plural noun (PLR), then it is disambiguated as 'all' not 'each' by (3a). But, if it is followed by a singular noun, then it means 'each'.
Moreover, agreement in number and gender of the nearest modifiers playsa vital indicator. For example, the plural (PLR), non-animate (NANM) noun should be modified by singular (SNG) feminine (FEM) adjective in case of nouns and their adjectival modifiers. For example, "قطع رائعة" 'wonderful parts', given the fact that the dictionary includes [N,SNG,MCL,NANMقطع] 'cutting', [N,PLR,FEM,NANM قطع] 'parts' and [ADJ,SNG,FEM رائعة] 'wonderful'. The sequence of [N, SNG, MCL, NANM قطع] + [ADJ, SNG, FEMرائعة] should be blocked as there is no agreement between them. rule in (3b) states that, if the masculine (MCL), non-animate (NANM), singular (SNG) noun is followed by a singular feminine adjective, then the sequence should be rejected and the singular noun 'cutting' should be changed to the plural one 'parts'.

3- (a) ([كل],^@all)(BLK)(ART)(NOU,PLR)=0;
(b)(N, SNG, MCL, NANM)(BLK)(ADJ, SNG, FEM)=0;

2) *Functional word qualifier:* Particles could be used as indicators for disambiguating the part of speech, as there are particles for verbs and others for nouns. For example, the particles of "لم" ,"لن","قد" and "لقد" are a verb particles. In "لم نمل", if the word "نمل" is chosen as a noun 'ants'and preceded by "لم" particle (PTC).The rule in (4a) should reject this sequence and backtrack it to the verb form "نمل" 'get bored'.

Another example is that prepositions and adverbial nouns should be followed by a noun. In the context "في تقبل", the word "تقبل" is chosen automatically as a verb 'accept', the rule in (4b) rejects that a preposition (P) could be followed by a verb (V), and it changes it to the noun "تقبل" 'acceptance'.

4- (a) ({[لم]|[لن]|[قد]|[لقد]|[إذ]},PTC)(BLK)(^VER)=0;
(b) (PREP)(BLK)({VER|ADJ})=0;

3) *Lemma qualifier:* Lemma can be used as a cue in disambiguating the correct meaning of specific words, some nouns always co-occur with certain nouns such as the word "فرقة" and its forms; plural and dual, they mostly occur with nouns such as "إنقاذ" 'rescue' and "جيش" 'arm'. Therefore, if the automatic choice is "فرق" 'difference' and it is followed by the noun "إنقاذ" in "فرق إنقاذ"' 'rescue teams', the rule in (5) will reject this sequence and chose the plural form of the lemma "فرقة" 'team'.

5- (a) (NOU, LEMMA=فرق)(BLK)(NOU,{[إنقاذ]|[جيش]}) = 0;

4) *Semantic feature qualifier:*In order to build more productive and constraint grammar, the co-occurrence of specific words and words with specific semantic features should be considered. For example, "أنحاء العالم"'parts of the world', if the word "أنحاء" 'parts' is modified by another noun, this noun should be a location. In the context "أنحاء العالم" 'parts of the world', if the automatic choice of the word "عالم" is 'scientist'and it modifies the word "أنحاء" 'parts', this sequence should be rejected by rule (6). Rule in (6) states that, if the noun "أنحاء" 'parts' is not followed by a location noun (^LCT), then change it to the locative noun"'عالم"'world'.

6- (NOU,[أنحاء])(BLK)(ART)(NOU,^LCT)=0;

C. Structural indicators.

To a great extent, the integration of words lexical and syntactic information provides a good solution for disambiguation. Structural indicator depends on both the lexical and syntactic structure tags. In order to start the structural disambiguation, some small syntactic trees should be built for the input sentences. Constituents should be established to form block of words or small tree such as (NPs, PPs, APs, CPs …., etc.).
Constituent boundaries are very helpful cue in the structural disambiguation phase. The structural level is the third and final level in the disambiguation process [19]. T- Rules will be used for this indicator. In this level, the mis-tagged words due to distant modifiers judgment will be disambiguated as possible.

In (7a), the sentence "هلا أريتني في الكتاب صفحتك", there are several items between the verb "أريت" and "صفحتك". The two nodes are separated by the protection noon, pronoun, and the prepositional phrase which exceeded the length specified in the rules stated in the above sections (4 nodes max). The sequence "صفحتك" is automatically analyzed as the verb "صفحت" 'forgave' + the pronoun "ك" 'your' as in (7b). Past verbs predicated to 3rd person feminine pronoun end with "ت", as in "صفحت" 'she forgave' has the similar surface form as that for noun "صفحة" 'page' when attached to the pronoun "page" 'your' (Because of similarity in the orthographic shape). The rule in (8a) explained in (8b) is applied over the syntactically tagged sentence in (7c) to change the verb "صفحت" 'forgave' to the noun "صفحة" 'page' as in (7d):

7- (a) هلا أريتني في الكتاب صفحتك
(b) POS tagging: هلا_IPR أريت _VER ن_SFX ي_SPR في_PER ال _ART كتاب_NOU صفحت _VER ك_SPR
(c) Syntactic tagging: هلا_IPR أريت_VER ن_SFX ي_SPR [في_PER [ال_ART كتاب_NOU] $_{NP}$] $_{PP}$صفحت_VER ك_SPR
(d) After structural disambiguation: هلا_IPR أريت_VER ن_SFX ي_SPR [في_PER [ال_ART كتاب_NOU]$_{NP}$]$_{PP}$ [صفحة_NOU ك_SPR]$_{NP}$

8- (a)
(VER,PAS,modified,%x)(SPR,%p)(PP,%t)(VER,PAS,{3PS,FEM},^changed,%w)(SPR,%b):=(%c)(%x)(%p)(%t)(%l)("ة"<"ت",changed, NOUN, modified, %w)(%b);

(b)**Condition:** if a modified past verb (PAS) followed by a connected pronoun (SPR), prepositional phrase(PP), another past verb predicated to the mentioned pronouns, and another connected pronoun.
Action: the final "ت" should be changed to "ة" and the new form "صفحة" should be retrieved as a noun from the dictionary.

The sentence in (9a), which is tagged in the lexical level as in (9b),the word "درسنا" is tagged as a verb 'we studied' which make the sentence syntactically ill-formed. As the verb requires its arguments to be expressed in the sentence as in "درسنا للتاريخ المصري مظاهر عديدة" 'we studied for Egyptian history several aspects'. The NP " مظاهر عديدة" 'several aspects' here is the object or the complement of the verb. In the sentence in hand, the pronoun "هو" RPR, cannot act as the complement of the verb, as it should act as a subject in any context. Free word order in Arabic permits the occurrence of the subject in a distant place from the verb, but not when the pronoun is the prominent pronoun "الضمير الظاهر; it only can appear before the verb.

9- (a) درسنا للتاريخ المصري هو دراسة للشخصية المصرية

(b) **POS tagging:**
درسنا_VER ل_PREP ال_ART تاريخ_NOU مصري_ADJ هو_PPR دراسة_NOU ل_PREP ال_ART شخصية_NOU مصرية_ADJ.

(c) **Syntactic tagging:**
درسنا_VER [ل_PREP [ال_ART تاريخ_NOU مصري_ADJ]$_{NP}$]$_{PP}$ هو_PPR [[دراسة_NOU [ل_PREP ال_ART شخصية _NOU مصرية_ADJ]$_{PP}$]$_{NP}$.

(d) After structural disambiguation:
[درس_NOU نا_SPR]$_{NP}$ [ل_PREP [ال_ART تاريخ_NOU مصري_ADJ]$_{NP}$]$_{PP}$ هو_PPR [[دراسة_NOU [ل_PREP ال_ART شخصية _NOU مصرية_ADJ]$_{PP}$]$_{NP}$.

Considering the pronoun as constituent boundary, the rule in (10) is applied over the syntactically tagged sentence in (9c) and the verb can be changed to the noun "درس" 'studying' and the connected pronoun as 'نا' 'our'.

10- (V,1PP,%x)(PP,%y)(PPR,%r):=(%s)(%c)(%x,-att,NOUN,modified,2>"",-POS,-LEX,-@past,-NUM, -PER, -ATE,[[]])(?[نا])(%y)(%r);

In addition to the constituent boundaries, another lexical-syntactic cue can help in the disambiguation which is coordination structure. Sentence in (11a) contains 4 coordination elements which should share the same POS. In the POS automatic tagged sentence in (11b), the word "خفت" is chosen as a verb 'fade' which is not suitable for the coordination syntactic structure so it should be disambiguated as noun

11- (a) فاستحسنوا لونه وخفته ومرونته ونصاعته
(b) POS tagging:ف_COO استحسنوا_VER لون_NOU ه_SPR و_COO خفت_VER ه_SPR و_COO مرونة_NOU ه_SPR و_COO نصاعة_NOU ه_SPR

(c) **Syntactic tagging:**
ف_COO استحسنوا_VER [لون _NOU ه_SPR]$_{NP}$ و_COO خفت_VER ه_SPR و_COO [مرونة_NOU ه_SPR]$_{NP}$ و_COO [نصاعة_NOU ه_SPR]$_{NP}$

(d) **After structural disambiguation:**
ف_COO استحسنوا_VER [لون _NOU ه_SPR]$_{NP}$ و_COO [خفة_NOU ه_SPR]$_{NP}$ و_COO [مرونة_NOU ه_SPR]$_{NP}$ و_COO [نصاعة_NOU ه_SPR]$_{NP}$

7 EVALUATION AND LIMITATION

Evaluation has been performed in order to investigate the accuracy and robustness of the rules. The used data consists of 105,878 words. The set of data is divided into a training set which includes 79,408 words and a testing set contains 26,496 words. The overall performance of our WSD system was very positive. The percentage of accuracy is 95% while the percentage of errors is 5%. The errors are divided into 4% due to problems in the disambiguation process and 1% due to wrong tokenization which consequently leads to wrong POS tagging. Our developed POS tagging and disambiguating system is capable of disambiguating many Arabic language problems as stated in the different sections of the paper. However, the system has some limitations, for example, the system was unable to correctly disambiguate the sequence "وهم" as conjunction "و" 'and' and pronoun "هم"'they', in the context " وهم في بداية نهضتهم"'and they in the beginning of their progress', because of the algorithm. The algorithm automatically assign to the sequence "وهم" the tag noun 'illusion', since it is the longest match and the context does not contain anything that can be used as a cue to correctly disambiguate that sequence. However, the same sequence can be correctly disambiguated in other context such as "وهم مستديرو الرؤوس" 'and they have rounded heads', because the system was able to overwrite the automatic tagging because of the plural adjective. Another example for the limitation of the system is that it was unable to disambiguate the sequence "لسعة" as "ل" 'because of' and "سعة" 'extent', in the context "وهو لسعة ثقافته يدأب في المقابلات والمقارنات" 'and because of the extent of his education he devotes himself to collations and comparisons', but it was automatically tagged as a noun "لسعة" 'sting'. The wrong disambiguation was due to the longest match rule. Evaluation results show that our system achieves significantly better accuracies.

8 CONCLUSION

In this paper, we have presented and evaluated a POS tagging and disambiguating system based on the UNL algorithm for obtaining language models oriented for POS tagging and disambiguation. The system is based on the rule based approach that uses contextual information to assign tags to unknown or ambiguous words. The system acts with high efficiency for Arabic language. We have directly applied the acquired models with other required models in different NLP applications such as information retrieval, summarization, machine translation and etc, and we have obtained fairly good results. Our developed models for POS tagging and disambiguation learning and testing have been performed on the corpus of 106.878 words. In this article, the infrastructure of the system is discussed. The linguistic resources and the tools involved are presented; they are all open-source resources. The accuracy of the output has been evaluated on the level of the tokenization and disambiguation. The percentage of accuracy is 95% while the percentage of errors is 5%.

REFERENCES

[1] E. Brill, "*Asimple rule-based part of speech tagger*", In Proceedings, Third Conferenceon Applied Natural Language Processing, ACL,Trento Italy, 1992.

[2] K. Church, "*A stochastic parts program and noun phrase parser for unrestricted text*" In Proceedings, Second Conference on Applied Natural Language Processing, ACL, Austin, TX,1988.

[3] D. Cutting,J. Kupiec, J. Pedersen and S. Penelope, "*A practical part-of-speech tagger*" In Proceedings, Third Conference on Applied Natural Language Processing, ACL, Trento, Italy, 1992.

[4] R.Weischedel,M. Meteer, R.Schwartz, L.Ramshaw and J. Palmucci, "*Coping with ambiguity and unknown words through probabilistic models*" Computational Linguistics,1993.

[5] A. Gülen, E. Saka, "*Part of Speech Tagging*". A Term Paper Submitted To Ceng463 Introduction To Natural Language Processing Course Of The Department Of Computer Engineering Of Middle East Technical University, December, 2001.

[6] F. Hasan, N. UzZman, M. Khan, "*Comparison of Different POS Tagging Techniques (n-gram, HMM and Brill's tagger) for Bangla*", Center for Research on Bangla Language Processing, BRAC University, Bangladesh,2006.

[7] E. Brill, "*Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging*" Computational Linguistics, vol. 21, pp. 543 565, 1995.

[8] S. Alansary, M. Nagi, N. Adly, "*UNL+3: The Gateway to a Fully Operational UNL System*", in Proc. of 10th Conference on Language Engineering,Cairo, Egypt, 2010.

[9] M.S.A. Chowdhury, N.M. MinhazUddin, M.Imran, M.M. Hassanand M.E.Haque"*Parts of Speech Tagging of Bangla Sentence*", In Proc. of the 7th International Conference on Computer and Information Technology (ICCIT), Bangladesh, 2004.

[10] M. Konchady, "*Text Mining Application Programming*". Programming Series. Charles River Media, 1 edition, 2006.

[11] S. Abuleil, kh. Alsamara, M. Evens, "*Discovering Lexical Information by Tagging Arabic Newspaper Text*", Workshop on Semitic Language Processing. COLING-ACL.98, University of Montreal, Montreal, PQ, Canada, Aug 16 1998, pp 1-7.

[12] S. Khoja, "*APT: Arabic Part-of-speech Tagger*". Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania. June 2001.

[13] S. Khoja, P. Garside, G. Knowles, "*A tagset for the morphosynactic tagging of Arabic*". Paper presented at Corpus Linguistics 2001, Lancaster University, Lancaster, UK, 2001.

[14] A. Freeman, "*Brill's POS tagger and a Morphology parser for Arabic*". NAACL Student Research Workshop, Lancaster University, 2001.

[15] B. Hammo, H. Abu-Salem, S. Lytinen, "*QARAB: A Question Answering System to Support the Arabic Language*", Proceedings of the Computational Approaches to Semitic Languages Workshop, University of Pennsylvania, 11th July 2002.

[16] M. Attia, "*A Large-Scale Computational Processor of the Arabic Morphology, and Applications*",2000.

[17] G. Kanna, R. Al-Shalab, M. Sawalha, "*Full automatic Arabic text tagging system*". The proceedings of the International Conference on Information Technology and Natural Sciences, Amman/Jordan, 2003.

[18] S. Alansary, M. Nagi, N. Adly, "*IAN: An Automatic tool for Natural Language Analysis*", in Proceeding of 12th Conference on Language Engineering, Cairo, Egypt, 2012.

[19] S. Alansary, "*Towards a Large Scale Deep Semantically Analyzed Corpus for Arabic: Annotation and Evaluation*", Empirical Methods in Natural Language Processing. 2014.

BIOGRAPHY

Dr. Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

TRANSLATED ABSTRACT

وسم أقسام الكلام وفك اللبس من أجل فهم اللغة العربية آليا

سامح الأنصاري

مكتبة الإسكندرية، الشاطبي، الإسكندرية، مصر
قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر
sameh.alansary@bibalex.org

ملخص
هناك أساليب مختلفة لوسم أقسام الكلام في النص(POS) وأيضا العديد من النهج لفك اللبس الدلالي في اللغات. هذه الورقة البحثية تقدم تعريفات عامة عن الوسم المعجمي وفك اللبس. هذان الموضوعان لهما أهمية كبيرة في المعالجة الألية للغات الطبيعية(NLP) . وبعد وضع الملامح العامة عن هذه الموضوعاتفإن هذا البحث سوف يقدم شرحا مفصلا عن نظام الوسم المعجمي وفك اللبس الصرفي والدلالي باستخدام القواعد. النظام المقدم قد شارك في العديد من تطبيقات المعالجة الألية مثل نظام ترجمة من اللغة الطبيعية للغة وسيطة للغة طبيعية قائم علىUNL (LILY) ونظام استخراج المعلومات (KEYs). وبلغت نسبة الدقة 95٪ في حين بلغت نسبة الأخطاء 5٪.

معالجة الالتباس الدلالي في نتائج تحليل المحلل الصرفي العربي تيم باكولتر

أحمد عبد الغني[1]*، سامح الأنصاري[2]*

*قسم اللسانيات والصوتيات/كلية الآداب/جامعة الإسكندرية

[1]hmd_abdelghany@yahoo.com

[2]sameh.alansary@bibalex.org

ملخص___يُعَدّالمحلل الصرفي العربي تيم باكولتر (Tim Buckwalter) من أشهر المحللات الصرفية في أدبيات معالجة اللغة العربية آليًا ، وذلك قد يرجع إلى أسباب مرتبطة بسهولة استخدامه ، وإتاحته، وإمكانية تطويره والتعديل فيه مما شجّع الكثير من الباحثين على تناوله بالتطوير والتعديل من أجل خدمة أهدافهم البحثية ، أو من أجل مجرد التحسين لخدمة مجال معالجة اللغة آليًا والبحث اللغوي بشكل عام. ويتناول هذا البحث معالجة الجانب الدلالي للتحليلات الصرفية المقررة التي تم اختيارها آليًا في مرحلة معالجة الالتباس الصرفي مستفيدًا في ذلك من الخصائص الصرفية المقررة، ويقتصر هذا البحث على معالجة شكل واحد من أشكال الالتباس الدلالي وهو الناتج عن الاشتراك اللفظي للكلمات ، ولا يشمل الالتباسات الدلالية الناتجة عن غياب علامات التشكيل مثل الالتباس في كلمة "رجل" التي قد تحتمل "رَجُل" أو "رِجْل". ويهدف هذا البحث إلى استنباط مجموعة العوامل اللغوية وغير اللغوية التي تساهم في معالجة التباس الكلمات المشتركة في اللفظ في اللغة العربية من خلال استقراء مدونة ضخمة ممثلة للغة العربية المعاصرة (corpus-based study) ، ثم صياغة تلك العوامل في شكل قاعدة معلومات يمكن تضمينها في نظام آلي يعالج الالتباس ويحدد المعنى الدلالي المقصود ، ثم وضع تصور لخطوات عمل النظام المعالج مستفيدًا من قاعدة معلومات عوامل معالجة الالتباس. ويُعدّ هذا البحث إضافة جديدة في طريق تطوير المحلل الصرفي العربي Tim Buckwalter ، وذلك بالاستفادة من المعلومة الدلالية المهملة في نتائج التحليل (gloss) وتفعيلها من خلال إجراء (Procedure) يعالج الالتباس الدلالي الناتج عن تعدد معاني الكلمة المحللة ، مما يؤدي تدريجيًا إلى بناء مدونة عربية محللة دلاليًا يمكن استخدامها (بشكل مباشر) في بناء وتطوير أدوات وتطبيقات معالجة اللغة العربية آليًا ، أو إعادة استخدامها (بشكل غير مباشر) في تطوير خوارزميات(algorithms) أكثر ذكاءً ، وفاعلية ، ودقة ، واحترافية في معالجة الدلالة.

الكلمات المفتاحية: اللبس الدلالي – المعالجة الآلية للغة العربية- المحلل الصرفي – تيم باكولتر- المشترك اللفظي –علم الدلالة.

أولًا: مقدمة

تنقسم الألفاظ العربية من حيث دلالاتها إلى ثلاثة أقسام :

1. المتباين: وهو أكثر اللغة ، وهو أن يدل اللفظ الواحد على معنى واحد.
2. المشترك: وهو أن يدل اللفظ الواحد على أكثر من معنى.
3. المترادف: وهو أن يدل أكثر من لفظ على معنى واحد[1][6]

يقول سيبويه "واعلم أن من كلامهم ، اختلاف اللفظين لاختلاف المعنيين ، واختلاف اللفظين والمعنى واحد ، واتفاق اللفظين واختلاف المعنيين"[2][12].

والأصل في اللغة أن يستخدم اللفظ الواحد في الدلالة على معنى واحد ، وأن يكون للمعنى الواحد لفظ واحد ، لكن يتولد من المعاني المفردة عدة معانٍ بشكل تدريجي وبطيء ، وهذا ما نسميه تطور المعنى ، فيستخدم نفس اللفظ للدلالة على معنى آخر قريب ، ومنه إلى ثالث متصل به ، وهكذا حتى تصل الكلمة أحيانًا إلى معنى بعيد كل البعد عن معناها الأول[3][1].

واختلف العلماء في إثبات هذه الظاهرة في اللغة العربية ، فمنهم من ينكر هذه الظاهرة بالكلية محتجًا بأن الأصل في اللغة الإبانة ، والإبانة تقتضي امتناع الالتباس[4][11] ، ومنهم من يثبت وقوعها في اللغة محتجًا بأن المعاني غير

[1]أحمد مختار عمر : علم الدلالة، ص:145.

[2]سيبويه : الكتاب، تحقيق عبد السلام هارون ، ص: 7/1.

[3]علي عبد الواحد وافي : علم اللغة ،ص : 314 (نقلًا عن المشترك اللفظي في الحقل القرآني).

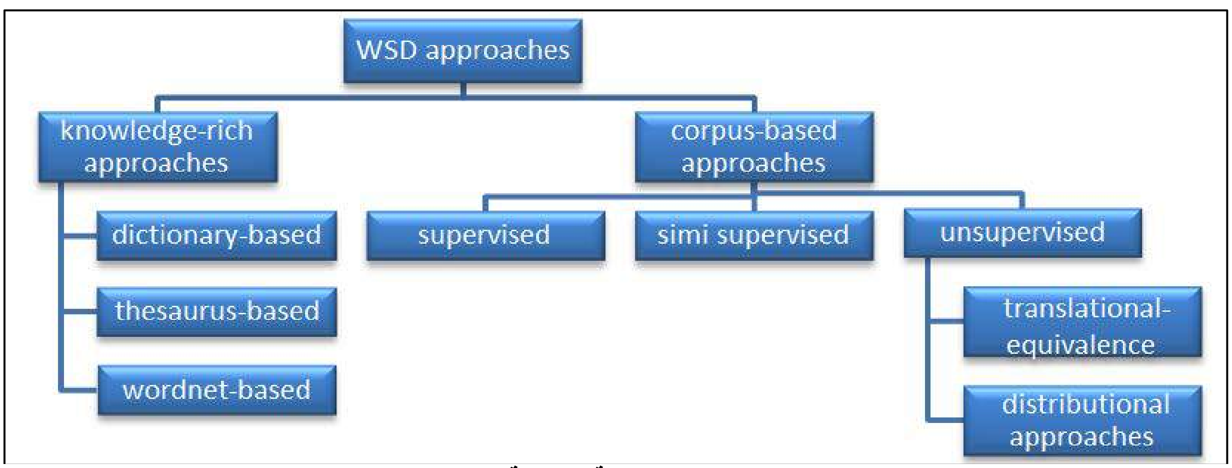
[4]عبد العال سالم مكرم :المشترك اللفظي في ضوء غريب القرآن الكريم ، ص12.

متناهية والألفاظ متناهية ، فإذا وُزّع لزم الاشتراك[5][7] ، واختلف المثبتون في تحديد إطار ومجال الظاهرة ، فمنهم من أطلقها ، ومنهم من قيّدها ، ومنهم من غالى في تقييدها إلى الحد الذي قصره على اللفظة التي تؤدي إلى معنيين مختلفين كل الاختلاف ، ليس بينهما أدنى ملابسة ، أو أية علاقة ، أو أي نوع من أنواع الارتباط.
وتنقسم أسباب حدوث الاشتراك اللفظي إلى أسباب خارجية متعلقة بالبيئة (مثل اختلاف اللهجات واقتراض الألفاظ من لغات أخرى) وأخرى داخلية متعلقة بتغير اللفظ أو تغير المعنى ، أما تغير اللفظ فيكون نتيجة تغير النطق بسبب العمليات الصوتية كالإبدال والقلب المكاني ، وأما تغير المعنى فيكون إما مقصود (كما في المصطلحات العلمية) أو تلقائي كما في التطور الدلالي بسبب ظواهر المجاورة والمشابهة والتقادم وغيرها.
وتُعد ظاهرة الاشتراك اللفظي أحد أهم أسباب وأبرز أشكال الالتباس الدلالي في اللغة العربية. وتُعد مشكلة الالتباس الدلالي في اللغة العربية الأكثر حدوثًا عنها في أي لغة أخرى ، وذلك لأن الالتباس ينشأ في أي لغة من اشتراك اللفظ في أكثر من معنى ، وهذا ينطبق على اللغة العربية ، ولكن يضاف إلى ذلك شكل آخر من أشكال الالتباس الدلالي وهو الناشئ عن غياب علامات التشكيل (Miss of diacritics) في اللغة العربية على وجه الخصوص ، فهذا السبب يضيف كمية كبيرة من الالتباسات الدلالية التي لا توجد في غيرها من اللغات ، وهذا ما يجعل ظاهرة الالتباس الدلالي في اللغة العربية الأكثر انتشارًا والأولى بالاهتمام والمعالجة.
ولم تحظ اللغة العربية بمحاولات كثيرة لمعالجة الالتباس الدلالي ، فمعظم الاتجاهات والخوارزميات المبتكرة لمعالجة الدلالة تم تطبيقها على اللغة الإنجليزية ولغات أخرى ، وقد حققت معدلات صحة في معالجة الالتباس تصل إلى 90%[6][14]. ويرجع سبب تأخر اللغة العربية في تطوير أنظمة معالجة الدلالة إلى الافتقار إلى المدونات العربية المحللة لغويًا التي تعتبر أساس عمل الأنظمة الموجّهة (supervised approach) في المعالجة الآلية.

ثانيًا: الاتجاهات المختلفة في معالجة الاشتراك اللفظي آليًا

يوجد اتجاهان رئيسان لمعالجة التباس المعنى من حيث مصدر معلومات اللغة المستخدم في المعالجة، هما :

1. الاتجاه المعياري المعتمد على مصادر اللغة التقليدية (Knowledge-Based Approaches)
2. الاتجاه الوصفي المعتمد على المدونات وأنظمة تعليم الآلية (Corpus-Based Approaches).



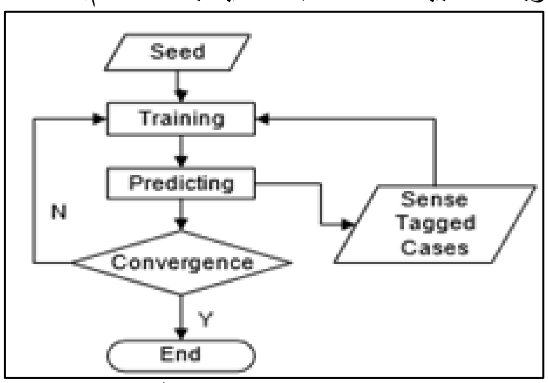
شكل (1) الاتجاهات المختلفة لمعالجة الالتباس الدلالي

وقد تنوعت الخوارزميات التي تندرج تحت كل من الاتجاهين السابقين ، فظهرت طريقة المتعلقات النحوية والقيود الدلالي (selection preferences and arguments) ، وخوارزمية تداخل التعريفات المعجمية (Overlap Based Approaches) ، وخوارزمية السير العشوائي (Random walk algorithm) ، وخوارزمية ووكر

[5] المزهر : 369/1

[6] Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation.

(WALKER) ، وخوارزمية يورفيسكي (YAROWSKY) لتستفيد من مصادر اللغة المعيارية (كالمعاجم والقواميس والموسوعات ومصنفات المفردات وغيرها) من أجل معالجة الالتباس الدلالي.
أما الاتجاه المعتمد على مصادر اللغة الوصفية فيندرج تحته ثلاثة اتجاهات فرعية ، هي :
(1) الاتجاه الموجّه (supervised corpus-based disambiguation)
(2) الاتجاه الشبه موجّه (Minimally or Semi-supervised Disambiguation)
(3) الاتجاه الغير موجّه (Unsupervised corpus-based disambiguation)
أما الاتجاه الموجه فتتصف باعتماده على مدونات محللة مسبقًا على المستوى الدلالي من أجل استخدامها كوسيلة للتدريب (training) وبناء الحسابات الإحصائية التي تستخدم بعد ذلك في اختبار وتحليل نصوص جديدة غير محللة (testing). وقد حقق هذا الاتجاه نتائج أفضل من نظيريه الشبه موجّه والغير موجّه في معالجة الالتباس الدلالي[7] [25]. ومن أشهر الخوارزميات التي تندرج تحت هذا المسمّى خوارزمية مصنف Bayes البسيط ، وخوارزمية قوائم القرار (decision lists) ، وطريقة آلات الدعم الموجّهة SVM ، وطريقة نموذج Markov ، وخوارزمية الأمثلة المدرّبة (memory-based) ، وطريقة شجر القرار (decision trees) ، وطريقة الشبكات العصبية (Neural Network) ، وغيرها.
أما الاتجاه الشبه موجّه فيتصف بالاستفادة فقط من أقل قدر من النصوص المحللة وبالتالي أقل قدر من التدخل البشري (knowledge acquisition bottlenecks) ، ثم التحول تدريجيًا إلى الميكنة الآلية الكاملة ، وهو ما يسمى بأسلوب Bootstraping أو التحسين المتكرر (recursive optimization) ، وهذا الأسلوب يستخدم في معالجة الالتباس الدلالي بشكل خاص ، وبناء تطبيقات المعالجة الآلية بشكل عام.

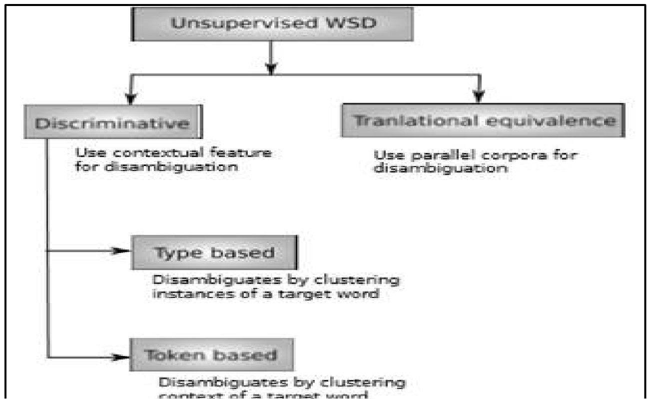


شكل (2) مخطط انسيابي لعمل خوارزمية التحسين المتكرر

أما الاتجاه الغير موجّه فيتصف باعتماده على مدونات صمّاء (Raw corpora) خالية من أي نوع من التحليل اللغوي ، لذلك فهي توصف بأنها طرق هزيلة المعلومات (knowledge-lean methods) ، وبسبب ذلك فإنها تفتقر إلى التحديد الدقيق لمعنى الكلمة الملتبسة (assigning sense tags) ، فهي فقط تستطيع تمييز (discrimination) المعاني المختلفة في فصائل من الكلمات أو السياقات يُطلق عليها اسم عناقيد (clusters) (كما في الطرق العنقودية). وتنقسم الطرق التابعة لهذا الاتجاه إلى طرق تمييزية[8] (discriminative approaches) معتمدة على المدونات الصمّاء أحادية اللغة (monolingual corpora) ، وطرق الترجمة المقابلة (translational equivalence) المعتمدة على المدونات الصمّاء المترجمة (parallel corpora) ، وذلك من خلال استخدام تقنيات المحازاة الكلامية (alignment techniques) التي تطابق مفردات المدونات المتوازية ، وبالتالي تحديد معاني الكلمات الملتبسة.

[7] Word sense disambiguation: A survey.

[8] تسمى أيضًا طرق عنقودية (clustering approaches) ، وطرق توزيعية (distributional approaches)



شكل (3) الطرق المختلفة التي تنتمي إلى الاتجاه الغير موجّه[9] [24]

ثالثًا: المحلل الصرفي العربي تيم باكوالتير

المحلل الصرفي العربي تيم باكوالتير هو أشهر المحللات العربية الصرفية العربية في أدبيات حوسبة اللغة العربية ، وقد تم تطويره بواسطة Linguistic Data Consortium (LDC) بلغة برمجة PERL ، ويتبع المحلل الصرفي العربي تيم باكوالتير الاتجاه التلاصقي المسوق بالمعجم (Concatenative lexicon-driven approach) في التحليل الصرفيبحيث يتم تمثيل قواعد الكتابة (Orthographic rules) ، وتوارد المورفيمات (Morphotactics) في المعجم مباشرةً ، وبذلك يكون القدر الأكبر في بناء المحلل هو بناء المعاجم الملحقة به. وبسبب استخدام المحلل الاتجاه التلاصقي في التحليل فإنه يكون من المتناسب مع ذلك استخدام التجذيع (stemming) في التحليل والتعرف على الكلمة بدلًا من مطابقة الجذر (root) والوزن الصرفي (pattern) الذي يتناسب مع اتجاه المستويين (Tow level approach) في التحليل الصرفي ، لذلك يعتبر جذع الكلمة(stem) هو أبسط شكل (base form) للكلمة في هذه الطريقة بخلاف الجذر (root) الذي يعتبر أبسط شكل للكلمة في اتجاه المستويين في التحليل الصرفي. ويتكون النظام من ثلاثة مكونات رئيسة هي المعاجم (lexicons)، وجداول التوافق (compatibility tables)، وخوارزمية التحليل (analysis algorithm). فأما المعاجم فتشمل معجم الجذوع الذي يحتوي على الأشكال التصريفية المختلفة للمداخل المعجمية العربية ، ومعجم السوابق الذي يحتوي على أشكال تتابع السوابق في اللغة العربية المعاصرة ، ومعجم اللواحق الذي يحتوي على أشكال تتابع اللواحق في اللغة العربية المعاصرة ، واحتواء معجم الجذوع على أشكال جذوع المداخل المعجمية يجعل المعجم أكبر حجمًا ، وخوارزمية التحليل أكثر بساطة. أما جداول التوافق فهي التي تحدد العلاقة التوافقية بين السوابق والجذع واللواحق ، فمجرد التعرف على السابقة من خلال معجم السوابق ، والتعرف على اللاحقة بواسطة معجم السوابق ، والتعرف على الجذع من معجم الجذوع ليس دليلًا على صحة التحليل ، ولكن لابد من التأكد من صحة توافق المكونات الثلاث. وأما خوارزمية التحليل فتتسم ببساطتها ، فهي فقط مجرد تنظيم وترتيب لخطوات التحليل ، أما منطق التحليل فقد تم تمثيله وصياغته في طريقة بناء المعاجم وجداول التوافق.

رابعًا: أنواع المعلومات اللغوية في نتائج التحليل

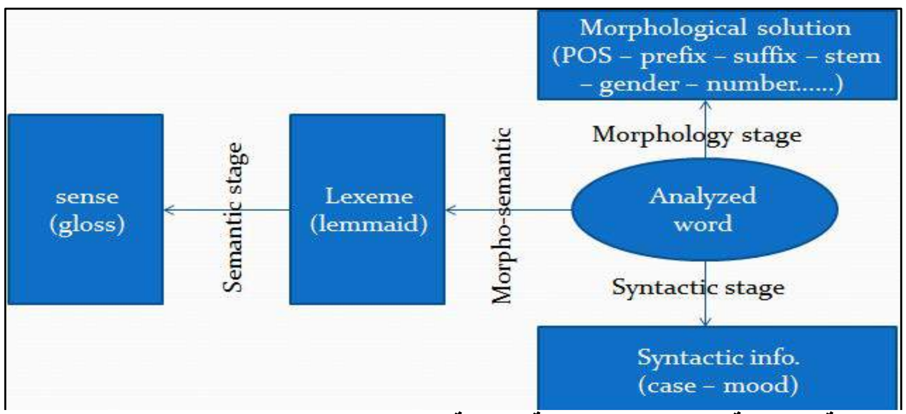
يمكن تقسيم المعلومات التي يَعرِضْها المحلل في نتائج التحليل إلى ثلاثة أقسام :

1. معلومات صرفية : متمثلة في تحليل الكلمة صرفيًا إلى سوابق وجذع ولواحق ، ثم وسم كل جزء من الكلمة بأقسام الكلام ، إلى جانب بعض الخصائص الصرفية من نوع (gender) ، وعدد (number) ، وشخص (person) ، وحالة التعريف (state) ، والزمن (aspect).
2. معلومات نحوية : متمثلة في عرض الحالات الإعرابية المختلفة المحتملة للأسماء المعربة (case) ، وكذلك الحالات الإعرابية للفعل المضارع المعرب (mood).

[9]Unsupervised Corpus-Based Methods for WSD.

3. معلومات دلالية : متمثلة في تحديد المداخل المعجمية (lemma) المحتملة للكلمة المحللة ، والمعاني المرتبطة بتلك المداخل (gloss).

ومن ثَمّ فإن عملية اختيار التحليل المناسب للسياق ليست مجرد عملية بسيطة لفك التباس صرفي ، لكنها عملية مركّبة من ثلاث عمليات رئيسة متدرجة ، تبدأ بفك الالتباس الصرفي الذي يحدد وسم الكلمة ، ثم فك الالتباس النحوي الذي يحدد علامة الإعراب ، ثم فك الالتباس الدلالي الذي يتم في مرحلتين ، أولاهما الالتباس الدلالي الناتج عن غياب علامات التشكيل (missing of diacritics) ، ويتم في هذه المرحلة تحديد المدخل المعجمي (lemma) ، وثانيهما الالتباس الدلالي الناتج عن الاشتراك اللفظي ، وهو يختص بتحديد المعنى المقصود لهذا المدخل في حالة احتماله لأكثر من معنى.
وبعد معالجة الالتباس الصرفي والنحوي والدلالي يسهل تحديد النسخ النطقي للكلمة ، ولا يمكن تحديده قبل ذلك لأنه يتضمن المعلومات الصرفية (تحديد أجزاء الكلمة) ، والنحوية (علامة الإعراب في نهاية الكلمات المعربة) ، والدلالية (المدخل المعجمي).
والشكل التالي يوضح المراحل التحليلية المختلفة التي تمر بها الكلمة العربية من أجل اختيار التحليل المناسب.



شكل (4) المراحل التحليلية المختلفة التي تمر بها الكلمة العربية

خامسًا: محاولات معالجة الالتباس الصرفي في نتائج المحلل تيم باكوالتير

1. نظام MADA+Token لمعالجة الالتباس الصرفي

من أبرز تلك المحاولات نظام MADA+Token[10] الذي يُعدّ امتداد لبرنامج ALMORGEANA (Arabic Lexeme-base MORphologicalGEnerator /ANAlyzer) الذي أُستخدِمت أحد إصداراته (ALMOR) في إجراء الجزء التحليلي في نظام MADA+Token[11][19]. أما معالجة الالتباس الصرفي وتحديد التحليل الأنسب للسياق فكان بشكل إحصائي تمامًا باستخدام تقنيات تعليم الآلة وبرامج نمذجة اللغة ، حيث يقوم النظام (في مرحلة التدريب training) بالتدرب على عشرة خصائص صرفية بشكل منفرد ، وتشمل تلك الخصائص الوسم صرفي (15 POS) ، وجود ضمير مرتبط بالكلمة (presence of a pronoun) ، وجود رابط بالكلمة (presence of a conjunction) ، وجود أداة مرتبطة بالكلمة (presence of a particle) ، وجود معرّف (presence of a determiner) ، النوع (gender) ، العدد (number) ، الشخص (person) ، الزمن (aspect) ، البناء للمعلوم والمجهول (voice). ثم يتم التنبؤ بخصائص الكلمة المراد تحليلها من حيث الخصائص العشرة السابقة (في مرحلة الاختبار testing). ثم يتم اختيار أقرب التحليلات المحتملة التي يعرضها المحلل ، وهي الأكثر اتفاقًا من حيث الخصائص العشرة المحددة مع الخصائص التي تم التنبؤ بها في مرحلة الاختبار ، وتتم تلك الخطوة باستخدام خوارزمية decision tree. وقد حققت عملية معالجة الالتباس نسبة صحة وصلت إلى 95.6%[12][21].

2. نظام مشروع المدونة العربية العالمية لمعالجة الالتباس الصرفي

وهو نظام هجين يوظّف كل ما يمكن الاستفادة به من قواعد اللغة النحوية والصرفية في معالجة الالتباس كليًا أو جزئيًا ثم يلجأ (في حالة استنفاذ القواعد اللغوية المعيارية) إلى الحلول الإحصائية متمثلةً في مراجعة مجموعة من النصوص المحللة (memory-based approach) ، ثم تزويد النصوص المحللة بأمثلة جديدة (بعد مراجعتها) من أجل تحسينها ، ثم إعادة استخدامها في تحليل نصوص جديدة ، وهكذا بشكل تكراري (Bootstrapping). ويقوم هذا النظام على اعتبار اختيار التحليل المناسب عملية مركبة من معالجة الالتباس الصرفي ، ثم النحوي ، ثم الدلالي.

ولم يتطرق أحد من الذين حاولوا معالجة الالتباس في نتائج تحليل المحلل Tim Buckwalter إلى قضية الالتباس الدلالي، ونظرًا لكون المحلل الصرفي العربي Tim Buckwalter من أشهر المحللات في مجال معالجة الصرف العربي ، وانتهازًا لفرصة عرض المحلل لكل الاحتمالات الدلالية للكلمات المحللة (حتى لو كان المعنى هو المتغير الوحيد في الكلمة فالمحلل يُفرد لها احتمال تحليل إضافي) ، ونظرًا للمحاولات المتعددة لمعالجة الالتباس الصرفي واختيار التحليل المناسب للسياق صرفيًا ، فإن كل ذلك كان حافزًا وراء محاولة تفعيل تلك المعلومات الدلالية وتحسين أداء المحلل ليشمل المستوى الدلالي إلى جانب الصرفي ليكون نواة بناء مدونة عربية محللة دلاليًا يمكن استغلالها مباشرةً في تطبيقات معالجة اللغة ، أو بشكل غير مباشر في بناء أنظمة موجّهة (supervised) لمعالجة الدلالة والتي لا يمكن لها أن تقوم إلا في ظل قاعدة واسعة من النصوص المحللة.

سادسًا: المداخل المعجمية المحتملة لأكثر من معنى دلالي

إن طريقة بناء معجم جذوع (Dicstem) المحلل الصرفي Tim Buckwalter تساعد في الاستدلال على المداخل المعجمية التي تحتمل لأكثر من معنى ، فعند التباس المدخل المعجمي الواحد يتم تمييز معانيه باللواحق (_1 ، _2 ، _3 ، ....) ، ويتم اعتبار المعاني المختلفة مداخل جديدة داخل المعجم ، وبالتالي تظهر في نتائج التحليل كاحتمال جديد لتحليل الكلمة.

والشكل التالي يعرض المعاني المختلفة لكلمة "لواء" من داخل معجم الجذوع (Dicstem) ، وكيفية تمييزها داخل المعجم.

[10] Morphological Analysis and Disambiguation for Arabic.
[11] Introduction to Arabic Natural Language Processing, p : 86
[12] Arabic computational morphology, p : 163 (Automatic Processing of Modern Standard Arabic Text)

```
;; liwA'_1
lwA'     liwA'    N_L banner;flag
lwA'     liwA'    NF  banner;flag
lwA&     liwA&    Nuh_L   banner;flag
lwA}     liwA}    Nihy_L  banner;flag
;; liwA'_2
lwA'     liwA'    N_L major general
lwA'     liwA'    NF  major general
lwA&     liwA&    Nuh_L   major general
lwA}     liwA}    Nihy_L  major general
;; liwA'_3
lwA'     liwA'    N_L brigade
lwA'     liwA'    NF  brigade
lwA&     liwA&    Nuh_L   brigade
lwA}     liwA}    Nihy_L  brigade
;; liwA'_4
lwA'     liwA'    N_L district;province
lwA'     liwA'    NF  district;province
lwA&     liwA&    Nuh_L   district;province
lwA}     liwA}    Nihy_L  district;province
>lwy     >alowiy  Nap districts;provinces
;; liwA'_5
```

شكل (5) المعاني المختلفة لكلمة "لواء" من داخل معجم الجذوع وكيفية تمييزها داخل المعجم

وبتصنيف المداخل المعجمية الموسومة رقميًّا (_1 _2 _3.....) في معجم جذوع المحلل الصرفي العربي تيم باكوالتير ، وبحصر المداخل التي سبب وسمها رقميًّا اختلاف المعنى الدلالي ، ولا يمكن الاستدلال على معناها باختلاف الوسم الصرفي ، نحصل على قائمة من المداخل المعجمية الملتبسة مكوّنة من 232 فعلًا و1453 اسمًا.

سابعًا: عوامل معالجة الالتباس الدلالي

يتضح لدينا من خلال استقراء مجموعة من السياقات لكلمات ملتبسة وجود أربعة أقسام من عوامل معالجة الالتباس ، هي :

- عوامل لغوية معيارية مرتبطة بقواعد اللغة النحوية الصرفية
- عوامل لغوية وصفية مرتبطة بالسلوك النحوي والصرفي العام للمعاني
- عوامل لغوية مرتبطة بالسياق اللغوي المحيط بالكلمات الملتبسة
- عوامل غير لغوية تعتمد على ملحوظات حول استخدام معاني الكلمة الملتبسة مثل الاعتماد على مجال النص ، ومدى تكرار حدوث المعاني المختلفة للكلمة.

1.<u>العوامل اللغوية المعيارية</u>

فأما العوامل اللغوية المعيارية فتشمل الخصائص الصرفية والتركيبية المعيارية (prescriptive rules) التي تنظم سلوك استخدام بعض المعاني ، وبالتالي تمكّن من كشف المعنى المقصود.

<u>ومن مظاهر تأثير الخصائص النحوية والصرفية في معالجة الالتباس</u> :

- عندما يتلازم اختلاف المعنى مع اختلاف الإطار التركيبي للكلمة (subcategorization frame) ، ويظهر ذلك بوضوح في حالة الأفعال الملتبسة بين معنى متعدي بحرف جر) ، وآخر متعدي لمفعول مباشر، مثل الفعل "صمّم" الذي يحتمل معنى "الإصرار" ، وفي هذه الحالة يكون فعلًا متعديًا بحرف جر "صمّم على" ، ويحتمل معنى "التخطيط والابتكار" صمم بيتًا ، وفي هذه الحالة يتعدى لمفعول مباشر، فظهور الفعل متبوع بحرف الجر "على" يحسم الالتباس. كذلك كلمة "محافظة" التي تحتمل كونها مصدرًا للفعل "حافظ"، وفي هذه الحالة تتعدى بحرف الجر "على" ، وتحتمل معنى "وحدة إدارية تمثل جزءًا من الدولة"، وفي هذه الحالة لا ترتبط بحرف جر بعدها. كذلك

كلمة "تعليق" كما (التعليق على الموضوع – تعليق عضويتها في الامم المتحدة)، وكلمة "معقود" كما في (الاتفاق المعقود "بين" - الأمل معقود "على" (معلّق) – سكر معقود (مذاب)) ، وغيرها.

- عندما يتلازم اختلاف المعنى مع اختلاف قبول بعض الأسماء للتعريف بأل ، مثل كلمة "شطر" التي قد تعني "جزءًا من" (شطر ماله) ، أو قد تعني "تجاه" (شطر المسجد الحرام) ، فالمعنى الأول يقبل التعريف بأل ، والآخر يلزم الإضافة لاسم ظاهر أو ضمير،وبالتالي يمتنع تعريفه بأل للزومه الإضافة ، مما يسهّل من تمييز المعنيين في حالة التعريف بأل. أيضًا كلمة "نحو" التي قد تعني "قدْر" أو "تجاه" (towards - approximately) ، فإنه يمتنع تعريفُها بأل للزوم الاضافة، بينما في حالة "علم النحو" أو "الطريقة" كما في (على النحو التالي) فيصح تعريفها بأل.
- عندما يتبع اختلاف معنى الفعل اختلاف في صفة الزوم التعدي وبالتالي قبول الضمائر المتصله كما في الفعل "اعتمد" الذي قد يعني "وافق وأنفذ" ، كما في "اعتمد القرار" ، وفي هذه الحالة يقبل الاتصال بالضمائر مباشرًة ، وقد يعني "اتّكل" ، كما في "اعتمد على نفسه" ، وفي هذه الحالة لا يقبل الاتصال بالضمائر بشكل مباشر. وكذلك الفعل "أدّى" الذي قد يعنى "أتم وأنجز وقضى" (أدي الفريق مرانه) ، ويتصل في هذه الحالة بالضمائر لتعديه بمفعول مباشر ، وقد يعنى "نتج عنه" كما في (أدى إلى) ، ولا يتصل في هذه الحالة بالضمائر لتعديه بحرف الجر "إلى".كذلك الفعل "دَقّ" الذي قد يحتمل المعنى اللازم ("صَغُر وخَفِيَ وقَلَّ" أو "نبض وخفق" كما في "دق جسمه" و"دقت الساعة") ، وفي هذه الحالة لا يتصل بضمير، وقد يحتمل المعنى المتعدي ("قرع وضرب ونقر" كما في "دقّ البابَ") ، وفي هذه الحالة يمكن اتصاله بضمير (دقّه).
- عندما يتلازم اختلاف المعنى مع اختلاف التصريف كما في بيت (بيوت (المسكن) – أبيات (الشعر)) وكذلك كلمة "ترجمة" التي تُجمع على "تراجم" في حالة قصد "السيرة الذاتية" ، أو "ترجمات" في حالة قصد "النقل من لغة إلى لغة" ، وكلمة ضابط (ضباط - ضوابط) حسب معناها ، وكلمة قرينة (قرائن - قرينات) ، وكلمة سائل (سائلون - سوائل) ، وكلمة "عامل" (عمال - عوامل) ، إلى غير ذلك من الأمثلة.
- عندما يتبع اختلاف المعنى اختلاف المصدر من حيث دلالته على معنى المصدر الجنسي المطلق (ولا يمكن جمعه في هذه الحالة) أو المصدر المقيد بنوع أو عدد (بالتالي يمكن جمعُه)، كما في كلمة "إجراء" التي قد تكون مصدرًا بمعنى الجنس المطلق ، كما في "إجراء عملية جراحية" ، أو تكون محددة بنوع أو عدد فتجمع كما في "إجراءات مشدّدة" وكذلك كلمة "قضاء" التي قد تعني (العدالة – الإبادة – ما يقدره الله – أداء – بذل الوقت) ، ولا تجمع في هذه الأحوال لأنها معاني مصدرية مطلقة ، وقد تأتي بمعنى "حي أو منطقة" كما في (قضاء صلاح الدين) فتجمع في هذه الحالة على "أقضية". فورود الكلمة بصيغة الجمع يحسم اللبس، وكذلك كلمة "فصْل" إذا قصد بها المصدر لا تجمع ، أما فصل الشتاء (أو الكتاب أو المدرسة أو المسرحية) فتجمع ، وكلمة "قلب" التي تجمع على "قلوب" لغير المصدر بخلاف المعنى المصدري "تحويل الشيء عن وجهه" (inversion) ، وغير ذلك من الأمثلة الكثير.

2.العوامل اللغوية الوصفية

أما العوامل اللغوية الوصفية فترجع أهميتُها الى حقيقة أن الكثير من الكلمات العربية تكون قابلة للعديد من التصريفات والاستخدامات الدلالية نظريًا (من واقع اللغة والمعجم) ، أما عمليًا (من واقع الاستخدام) فنجد تلك الكلمات منحصرة في استخدامات وتركيبات وتصريفات محدّدة ، والذي يهمّنا هو معالجة اللغة المستخدمة في الواقع وليس اللغة النظرية الموصوفة في المعاجم ، لأن اللغة بصفتها المعيارية (prescriptive) نظام معقّد يصعب تتبعُه وتطويعه للمعالجة بشكل كامل ، كما أنه لن يعود علينا بفائدة أن نعالج استخدامات ووظائف نظرية غير موجودة على أرض الواقع.

ومن مظاهر تأثير السلوك الصرفي في معالجة الالتباس :

من حيث الإفراد والجمع

التزام كلمة مواصلات صيغة الجمع إذا قصد بها وسائل المواصلات، وصيغة المفرد إذا قصد بها المعنى المصدري (مواصلة المسير)، كذلك كلمة "مصير" تلتزم صيغة الجمع (مصارين) إذا قصد بها الأمعاء وصيغة المفرد إذا قصد بها

المآل (والى الله المصير) ، كذلك كلمة "طقس" التي تحتمل معنى "حالة الجو" وتأتي بصيغة المفرد في هذه الحالة أو معنى "نظام العبادة والشعائر الدينية" وتأتي جمعًا بهذا المعنى (طقوس)،

ومن حيث التعريف والتنكير

حيث تتسم بعض المعاني بعدم أو ندرة ظهورها معرّفة بأل مما يساعد في إيجاد ضابط فاصل بين المعنيين. ومن أمثلة ذلك كلمة "شارع" (street - legislator) ، إذ يتميز السلوك الصرفي لمعناها "المشرِّع" بعدم ظهوره نكرة في السياقات المختلفة التي تم فحصُها ودراستُها ، وبذلك يكون الالتباس محسوم في حالة ورود الكلمة مفردة نكرة. أيضًا كلمة "قضاء" لا تعرّف بأل إذا قُصد بها بذل الوقت (spending) (قضاء وقت ممتع).

ومن حيث الاتصال بالضمائر

أما من حيث الاتصال بالضمائرفنظريًا أغلب الأسماء تقبل الاتصال بالضمائر ، لكن الاستخدام الفعلي للمفردات قد يفرض واقعًا مختلفًا. ومن أمثلة متابعة اختلاف المعنى لاختلاف قبول الضمائر المتصلة كلمة "حامل" لا تتصل بالضمائر إذا قصد بها "pregnant" بخلاف المعنى "carrier"، وكلمة "نحو" لا تتصل بالضمائر إذا قصد بها "grammar" بخلاف المعاني " approximately - towards -manner" كما في "على النحو التالي – نحو الهدف – عددهم نحو 20 رجلًا"،وكلمة "براءة" إذا قصد بها براءة الاختراع " license" لا تتصل بالضمائربخلاف المعنى الآخر "innocence"، وكلمة "مَيْسَرة" إذا قصد بها نحو (حين ميسرة) لا تتصل بالضمائر "Comfort" بخلاف المعنى " left wing" (ميسرة الجيش)، وكلمة "قرش" يندر اتصالها بالضمائر إذا قصد بها سمك القرش "shark"، ويكثر اتصالُها بالضمائر إذا قصد بها "piaster"، وكلمة "عين" يندر اتصالها بالضمائر إذا قصد بها "عين الماء" بخلاف المعاني " arabic letter - eye"،وكلمة "سائل" يندر اتصالها بالضمائر إذا قصد بها المائع "liquid" بخلاف معنى اسم الفاعل من سأل، وكلمة "شارع" يندر اتصالها بالضمائر إذا قصد بها "المشرِّع" بخلاف المعنى "street"، إلى غير ذلك من الأمثلة.

ومن مظاهر تأثير السلوك التركيبي في تمييز المعاني :

اختصاص كل معنى من معاني كلمة تحقيق (achievement – investigation) بمجموعة محددة من النماذج التركيبية عند ظهورها بصيغة المفرد النكرة في السياقات المختلفة كما هو موضح :

جدول (1) السلوك التركيبي لمعاني كلمة "تحقيق" عند ظهورها بصيغة المفرد النكرة

حالة الظهور	الضابط	تحقيق	
		investigation/verification/interrogation	achievement/realization
sin_indef	نماذج تركيبية	تحقيق + صفة تحقيق + علامة ترقيم (: -) + علم على شخص تحقيق + علم على شخص (مسبوق بلقب) تحقيق + فعل تحقيق + اسم معطوف بحرف عطف	تحقيق + مضاف إليه (اسم مجرد من أل) تحقيق + مضاف إليه (معرف بأل) تحقيق + مضاف إليه (اسم إشارة) تحقيق + مضاف إليه (اسم موصول) تحقيق + مضاف إليه (أي - كل)

3.العوامل السياقية

في أحيان كثيرة يُحكَم الالتباس بين معاني الكلمة الواحدة، فتتحد التصريفات والوصف النحوي، ويتحد السلوك الصرفي والتركيبي العام للمعاني المختلفة، ولا يبقى أي سبيل للفصل بين المعاني إلا السياق للدلالة على المعنى. وتتمثل العوامل السياقية في : المتصاحبات اللفظية (collocation) والكلمات السياقية البارزة (salient words).

فأما المتصاحبات اللفظية فتكثر مع الكلمات التي يرتبط ظهورها بتركيب معين لا تنفك عنه ("مسقط رأسه" بخلاف "مسقط مائي أو مسقط الخريطة" – "تل ابيب" بخلاف "شهر أبيب" – "ميسرة الجيش" بخلاف "حين ميسرة")، ويُعد

ظهور الاسم بصيغة المفرد أو الجمع المجرد من "أل"ومن الإضافة للضمائر المتصلة من أكثر أوضاع الظهور التي يبرز فيها دور المتصاحبات اللفظية وتكون عاملًا مؤثرًا في كشف المعنى، إذ تكثر فرصة إضافتِهِ إلى أسماء ظاهرة والتي كثيرًا ما تكون محددًة للمعنى المقصود.

أما الكلمات السياقية البارزة (salient words) : فهي هي الكلمات التي تظهر بشكل ملحوظ في السياقات للدلالة على معنى معين ، وبالتالي تكون مؤشرًا في الاستدلال على هذا المعنى، والاستدلال بالكلمات السياقية على المعنى المقصود يكون في حالات أهمُها :

- ظهور الكلمات الملتبسة غير مقيدة بقيد يكشف معناها مثل الإضافة أو الوصف المميِّز للمعنى.
- ظهور الكلمة الملتبسة مضافة لضمير متصل لأن الضمير المتصل يقلل من فرصة عمل المتصاحبات اللفظية.
- في حالة الكلمات التي يرتبط اختلاف معناها باختلاف مجال الكلام ("صرف" (علم اللغة - الاقتصاد) – "جذر" (علم اللغة –علم النبات) – "عجلة" (لها معنى اصطلاحي في الفيزياء) – "علَم وظرْف" (لهما معني اصطلاحي في علم النحو) – تسديد (اقتصاد - رياضة) – أَجْر (اقتصاد – دين) – نقدي (الأدب - الاقتصاد) – سهم (لها معني اصطلاحي في الاقتصاد)).

1. العوامل الغير لغوية

وتتمثل العوامل الغير لغوية في الاعتماد على تكرار المعنى ومجال النص، فقد يحقق الاعتماد على مجال النص نتائج جيدة وسريعة في حالة المعاني المرتبطة بمجالات معينة، والمجالات المتَّسمة بمفردات مميّزة مثل النصوص الدينية، والاقتصادية، لكن يظل الاعتماد على الكلمات السياقية أقوى في الحُجّة على إثبات المعنى المقصود لاعتماده على عوامل مباشرة ودقيقة في تحديد المعنى. كذلك الاعتماد على التكرار ونسبة حدوث كل معنى قد يحقق نتائج جيدة وسريعة في حالة المعاني الملتبسة التي بينها فروقات ملحوظة في نسبة التكرار داخل النصوص.

ثامنًا: وسائل الترجيح

في بعض الحالات لا يتحقق أي من العوامل المرجّحة لأي من المعاني الملتبسة، وفي أحيان أخرى قد تتحقق الضوابط المرجّحة لِكلا المعنيين المحتملين ، فكان لابد من وجود وسائل لترجيح المعنى المقصود. وقد تم تحديد أربع وسائل لترجيح المعنى المقصود في الحالات السابقة ، وتتلخص في الآتي :

1. توسيع نطاق السياق بما يسمح بظهور ضوابط ودلائل تحسم الالتباس.
2. الترجيح بالتكرار : وشرط عمل تلك الوسيلة أن يكون أحد المعنيين غالب الحدوث في النصوص ، والآخر نادر الحدوث ، ويظهر ذلك بوضوح عندما يكون العامل المميّز للمعاني هو الكلمات السياقية أو المتصاحبات اللفظية حين لا يظهر أي من الكلمات المميّزة لأي من المعاني المحتملة في سياق الكلمة المراد تحديد معناها، عندئذٍ نلجأ للتغليب بالتكرار.

مثال

كلمة "بيت" حال ورودها مثنى نكرة في حالتي النصب أو الجر يكون تمييز معنييها بالكلمات السياقية كما في الجدول التالي :

جدول (2)

الكلمات السياقية المميزة لمعاني كلمة "بيت" عند ورودها نكرة بصيغة المثنى المنصوب أو المجرور

بيت		العامل	حالة الظهور
Verse	House		
نص – مقطع – فكرة – صورة – شعر – الجاحظ – قصيدة – مطران – أنشد – لفظة – كلام – يصوغ – صيغة – لفظ – ذم – شاعر – نظَم – نظْم – ديوان – فني – أبيات – عبيد بن الأربص – قوافي – شعري – ابن المعتز – الأخطل – معنى – عمرو بن كلثوم – النابغة – التنبي – ينشد – شكيب أرسلان – رثاء حسي – قافية – مجانسة – زهير بن أبي سلمى – شعراء – أبو نواس – استعارة – أسلوب – الحصر – المسعودي – موسيقا – وزن – عروض – عروضي – مؤلفين – جناس – قوله – صدر – شطر – معنى – ابن هشام – ضرار بن الخطاب – ارتجل – امرئ القيس – بياني – أبي القاسم الشابي	دكان – فناء – بيوت – ثمن – اشترى – زوجة – جدار – منور – أهل - الجيزة	كلمات سياقية	Du_indef_gen_acc

لكن قد نجد أنّ بعض السياقات لا يظهر فيها أي من الكلمات المميزة السابقة مثل :

"وينشأ عن ذلك أن الجملة المسرحية التي تكون أطول من أن يستوعبها بيت واحد تنشطر في بيتين تفصل بينهما فصلا واضحا ، ليس من السهل على المستمع أن يغفل عنه"

فيكون التغليب هنا بالمعنى الأكثر تكرارًا في النصوص العربية بهذه المواصفات الصرفية ، وهو البيت الشعري ، وذلك بسبب ملاحظة تفوق هذا المعنى في التكرار - بشكل ملحوظ - في نصوص المدونة عند ظهوره بصيغة المثنى النكرة المجرورة أو المنصوبة.

3. التر<u>جيح بالتمايز</u> : وشرط هذا النوع أن يكون أحد المعنيين أكثر تمايزًا وأسهل في التحديد، على الرغم من أنه قد يكون الأقل تكرارًا ، (كأن يكون المعنى مرتبطًا بظهور مجموعة محددة من المفردات المميزة له ، أو يكون مرتبطًا بمجال نصوص معين لا يخرج عنه ، والمعنى الآخر غير مقيد بمفردات ، ولا قواعد استخدام مميزة ، ولا مجال محدد).

<u>مثال</u>

كلمة "قرش" حال ظهورها معرّفة بأل "القرش" ، فالمعنى الأول ("shark") يكون أكثر تمايزًا بمجموعة من المفردات السياقية مثل "حوت – أسماك – بحار – تصيد ......"، والمركبات المميزة مثل "القرش المفترس – أسماك القرش الضارية – أنواع القرش – حيوان القرش ......" (نادرًا ما يظهر في سياقات محايدة) ، أما المعنى الآخر ("Piaster") فهو أقل في درجة التمايز (كثيرًا ما تحيط به كلمات محايدة) ، ويمكن أن يرد في أمثلة وليس بجواره أو حوله كلمات مميّزة لمعناه كما في المثال التالي :

"الطفل سعيد يضع قدمه على القرش وينضم إليه بعض الأطفال منهم حسن ابن زكية"

وفي هذه الحالة ترد كلمة "القرش" ولا يتحقق أي من المركبات أو الكلمات السياقية المرجّحة لأي من المعنيين ، فيترجّح المعنى الأقل تمايزًا لضعف احتمال ورود المعنى الأول بدون كلمات مميّزة ، وقوة احتمال ورود المعنى الثاني بلا قرائن مرجّحة.

4.<u>الترجيح بالمفاضلة بين عوامل معالجة الالتباس</u> : وشرط ذلك أن يتحقق العامل المميّز لكل من المعنيين ، وأن يكون العاملان المستخدمان في تحديد المعنيين مختلفين (بحيث يكون أحدهما أقوى من الآخر)، فتتم المفاضلة بين العوامل ، ويترجّح المعنى المرتبط بالعامل الأقوى. ويتضح ذلك كثيرًا في حالة وجود كلمات سياقية ترجّح أحد المعنيين، ومتصاحبات لفظية ترجّح المعنى الآخر ، فيترجّح المعنى المرتبط بالمتصاحبات اللفظية لأنه العامل الأقوى. وترتيب

عوامل معالجة الالتباس من حيث القوة يبدأ بالعوامل المعيارية، ثم الوصفية، ثم المتصاحبات اللفظية، ثم الكلمات البارزة، ثم العوامل الغير لغوية.

مثال

"الدكتور أنيس يعتبر الهمزة مزمارية وليست حلقية ، لتشكل صوتها عند فتحة المزمار"

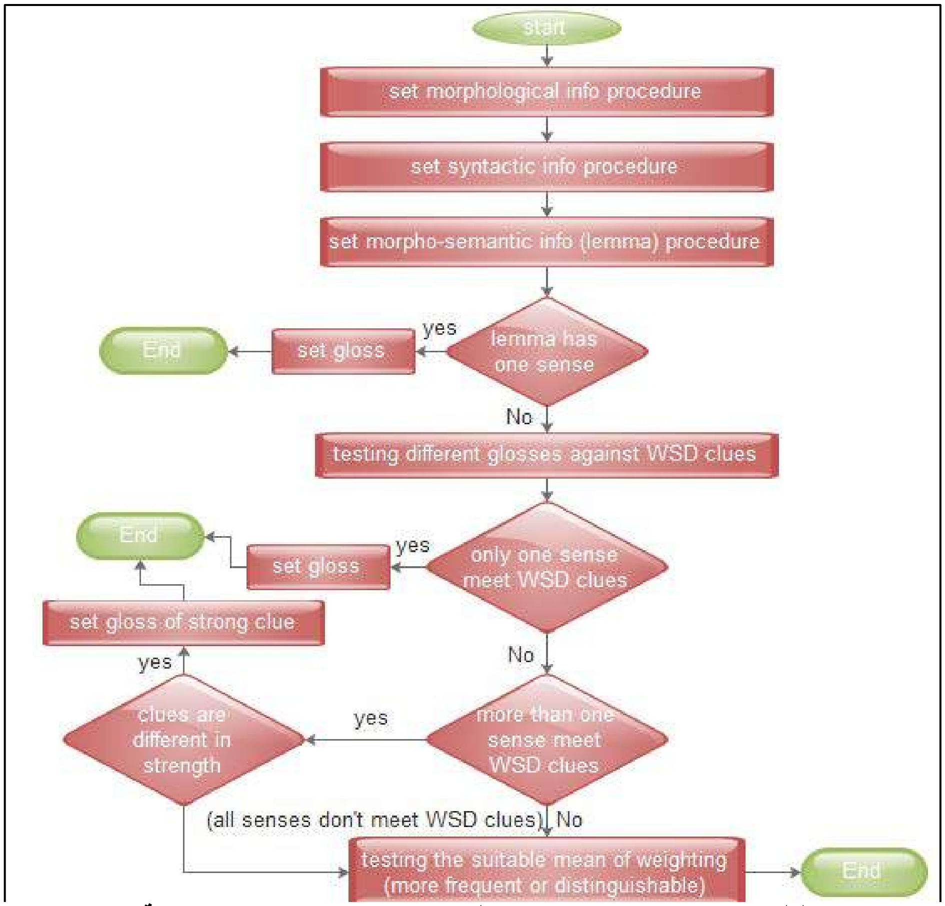
كلمة "فتحة" بصيغة النكرة المفردة يتحدد معناها "Arabic short vowel 'a'" بمجموعة محددة من الكلمات السياقية (salient words)، والمتصاحبات اللفظية (collocation) مثل " سكون – تشديد – كسرة – الهمزة – الصرف – ألف – ضمة – واو – كسرة ....." ، و"فتحة اللام – فتحة الهمزة – فتحة أو ضمة – نضع فتحة – عوض عن فتحة......" ، أما المعنى الآخر "opening/porthole" فيتحدد بمجموعة أوسع من المركّبات (collocation) مثل "فتحة الباب - فتحة الخروج – فتحة المزمار – فتحة الأنف – فتحة العينين – فتحة التهوية – فتحة ضيقة – فتحة في الجدار – عبر فتحة – فتحة تتسع.. "، وفي المثال السابق يوجد قرينة ترجّح المعنى الأول ، وهي ظهور كلمة "الهمزة" في السياق ، وقرينة ترجّح المعنى الآخر ، وهي ظهور الكلمة الملتبسة في المركب الإضافي "فتحة المزمار" ، ومعلوم أن عامل المتصاحبات اللفظية أقوى من الكلمات السياقية في الدلالة على المعنى ، وبذلك يكون المعنى المرجَّح هو المرتبط بالضابط الأقوى ("opening/porthole").

تاسعًا: تصوّر لعمل الإجراء الخاص بمعالجة الدلالة

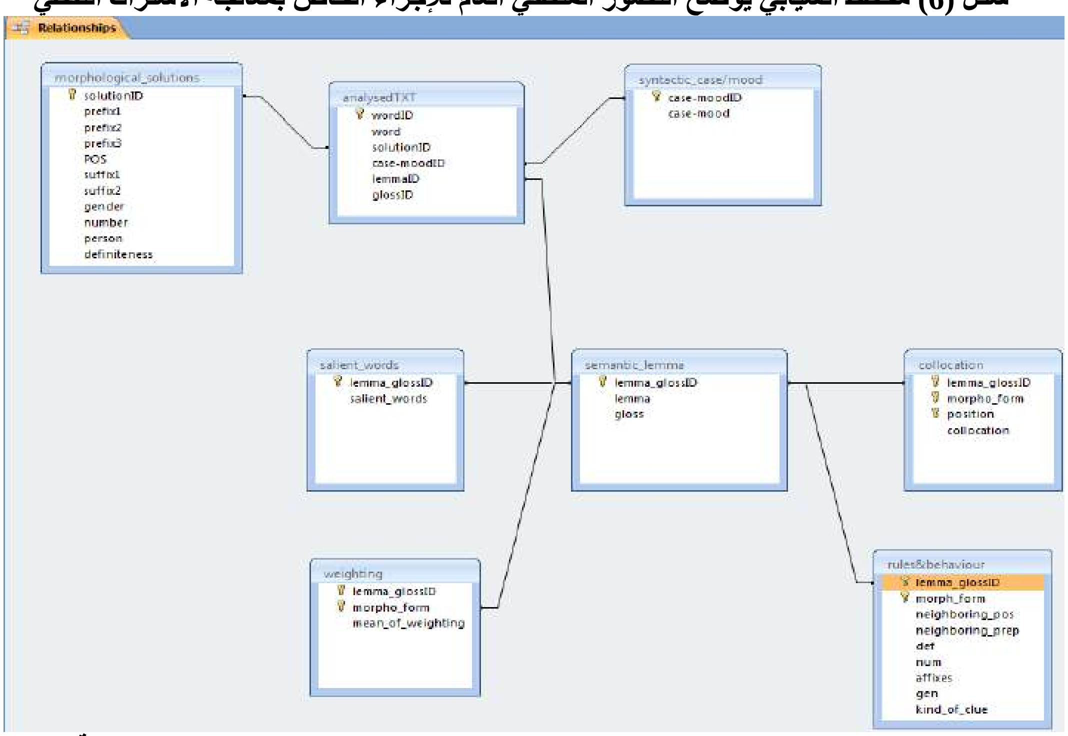
تنقسم عملية اختيار التحليل المناسب للسياق من بين التحليلات المحتملة التي يعرضها المحلل الصرفي Tim buckwalter إلى أربعة إجراءات (procedures) فرعية ، هي :

- تحديد الخصائص الصرفية للكلمة، وتشمل الوسم الصرفي لأجزائها، ونوع الكلمة من حيث النوع والعدد والشخص (في حالة الضمائر والأفعال).
- تحديد الخصائص النحوية للكلمة، وتشمل الحالة الإعرابية في حالة الاسم (case) ، وفي حالة الفعل المضارع المعرب (mood).
- تحديد المدخل المعجمي للكلمة (lemma)، ومعالجة الالتباس الناتج عن غياب علامات التشكيل، وتعدّد المداخل المعجمية (صدور).
- تحديد معنى المدخل المعجمي في حالة اشتراك لفظه واحتماله لأكثر من معنى.

والذي يخصنا هنا هو عملية تحديد معنى المدخل المعجمي في حال اشتراك لفظه. ونعرض فيما يلي المخطط الانسيابي الذي يوضّح التصوّر المنطقي العام لعمل هذا الإجراء (procedure) ، ويليه تصوّر لبناء قاعدة المعلومات التي تربط بين الجداول الخاصة بالنوعيات المختلفة للمعلومات اللغوية التي تساهم في معالجة الالتباس ، ثم تصوّر لشكل واجهة النظام الذي يستفيد من قاعدة المعلومات، ويطبّق المخطط الانسيابي المقترح.



شكل (6) مخطط انسيابي يوضح التصور المنطقي العام للإجراء الخاص بمعالجة الاشتراك اللفظي



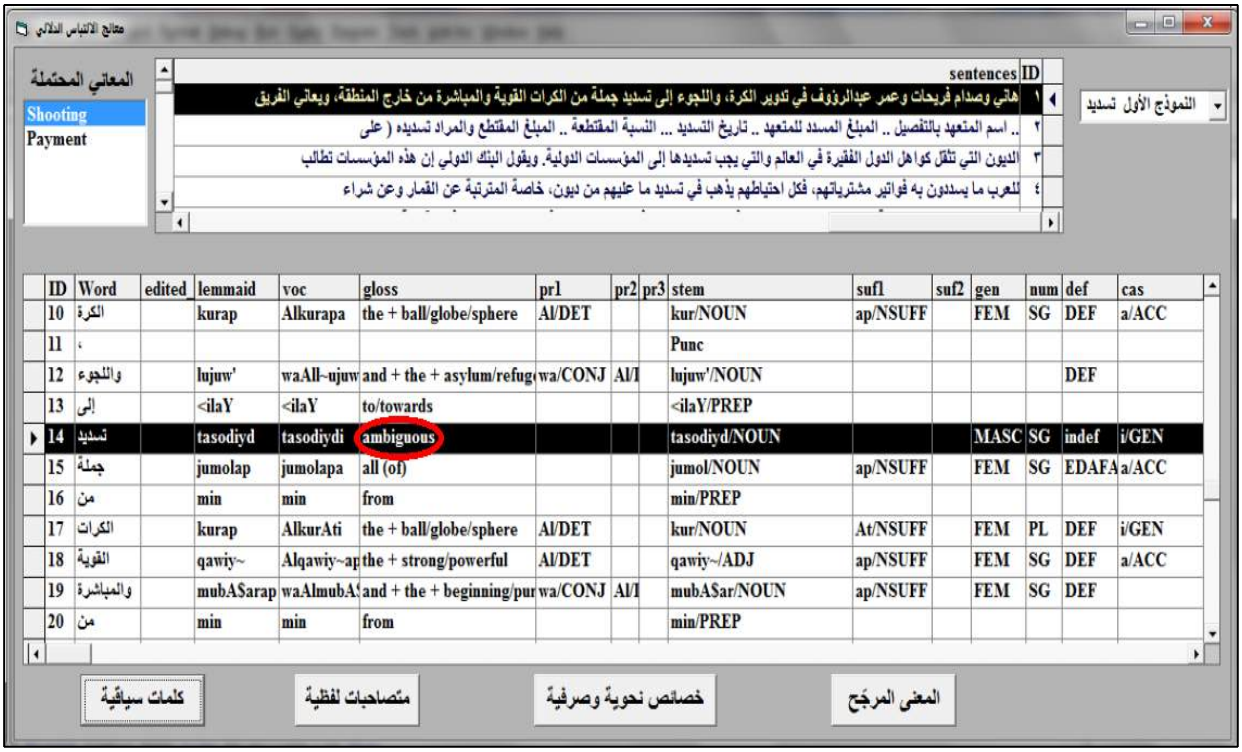
شكل (7) تصور هيكلي لقاعدة بيانات اختيار التحليل المناسب يوضح التكامل بين النوعيات المختلفة للمعلومات اللغوية وكيفية تمثيل عوامل تحديد المعنى المقصود فيها

شكل (8) واجهة متصفح التحليلات المختارة (توضح الخصائص الصرفية والنحوية والدلالية لكلمة "محافظة"

شكل (9) واجهة معالج الالتباس الدلالي (توضح المعاني المحتملة لكلمة "محافظة" والمعنى المرجّح وضابط الترجيح

عاشرًا: معالج الالتباس الدلالي

هو نموذج مصغّر يوضّح تطبيق عوامل معالجة الالتباس على الوجه الذي تم عرضه مسبقًا من أجل اختبار كفاءة تلك العوامل على مجموعة من الأمثلة الجديدة. ويعرض الشكل التالي واجهة البرنامج حيث يظهر فيها جدول يضم مجموعة من سياقات كلمة "تسديد" المحتملة للمعنى "دفع" (Payment) ، والمعنى "توجيه وتصويب" (shooting) ، ويظهر في الجدول الآخر تفاصيل التحليل الصرفي الذي تم اختياره آليًا في مرحلة معالجة الالتباس الصرفي للمثال المظلّل في جدول الأمثلة، وعند تظليل الكلمة محل الالتباس في الجدول الثاني تظهر قائمة المعاني المحتملة (في أعلى يسار واجهة البرنامج) من واقع جدول المداخل المعجمية ومعانيها (semantic_lemma table) في قاعدة بيانات معالجة الالتباس الدلالي. أمّا اختبار عوامل معالجة الالتباس في حالة كل معنى محتمل وترجيح المعنى المقصود فيكون باختيار المعنى من قائمة المعاني المحتملة والضغط على الأزرار الموضحة. ويتم تسجيل جميع نتائج الاختبارات بشكل مفصّل في جدول النتائج (Archive) في قاعدة البيانات. وقد ساهم هذا البرنامج في اختبار العديد من النماذج.



شكل (10) واجهة معالج الالتباس الدلالي

أخيرًا: الخاتمة

سَعَت هذه الدراسة إلى توسيع نطاق الاستفادة من المحلل الصرفي العربي Tim Buckwalter من خلال تفعيل المعلومة الدلالية التي يعرضها ضمن نتائج التحليل والاستفادة منها في تطوير مدونة عربية محللة على المستوى الدلالي إلى جانب المستوى الصرفي لتكون بمثابة النواة أو نقطة الانطلاق التي تمكّن اللغة العربية من الخوض والمنافسة بقوة في تطوير أنظمة معالجة الدلالة الحديثة والمعروفة على مستوى باقي اللغات التي سبقتنا بسبب امتلاكها مدونة ممثلة للغة محللة دلاليًا. كما سعت هذه الدراسة إلى التعرّف على الاتجاهات المختلفة لمعالجة الالتباس الدلالي ومحاولة الاستفادة منها في معالجة الالتباسات الدلالية لنتائج تحليل المحلل الصرفيTim Buckwalter ، إلا أن اعتماد أغلب الخوارزميات على مدونة محللة مسبقًا على مستوى الدلالة أدى بالباحث إلى التفكير أولًا في بناء القدر المطلوب من النصوص المحللة بشكل أقل احترافية مستفيدًا من الخصائص الصرفية للكلمة الملتبسة دلاليًا بعد تحليلها صرفيًا ، لذلك اتجه الباحث إلى حصر مجموعة العوامل التي تمكّن من تمييز المعاني المختلفة للكلمات الملتبسة بالرجوع إلى الأبحاث الحديثة والقديمة في معالجة الدلالية، بالإضافة إلى استكشاف سياقات المعاني المختلفة لبعض الكلمات الملتبسة في المدونة العربية العالمية (ICA) من أجل الوقوف على تلك العوامل، ثم صياغة تلك العوامل في نظام منطقي متكامل يعالج الالتباس الدلالي لمجموعة المداخل المعجمية المحتملة لأكثر من معنى في معجم جذوع المحلل الصرفي العربي Tim Buckwalter.

ملحق (1) : نموذج لمدخل معجمي ملتبس وتطبيق عوامل معالجة الالتباس من واقع استقراء أمثلة من المدونة العربية العالمية (ICA)

<u>تسديد (payment - shooting)</u>

جدول (3) عوامل معالجة التباس معاني المدخل المعجمي "تسديد" (PAYMENT - SHOOTING)

المعنى	حالة الظهور	الضابط	تسديد
تصويب - توجيه aiming/shooting	Sin_indef	مركبات مميِّزة	تسديد (النظرات – خطاه – لكمات – الكرة – الرماية – اللاعبين – الخطى – (أي) ضربة - أجوبته)
	(**) Sin_def	مركبات مميِّزة	التسديد (من قِبل اللاعبين - على المرمى – من بعيد – إلى الهدف – الفردي)
			(محكم - تدرب على - موقع - دقة - أحسنَ – خط) التسديد
	Sin_pro		كلمات سياقية
	Pl		تسديدات
	<u>كلمات سياقية</u> : اللاعبين – المرمى – الهدف – ركل – تمرير – الكرة – مرمى – مراوغة – هجوم – منطقة الجزاء – المنتخب – هجمات – اللاعب – مهارة – الفريق – زناد – إصبع – بولنج – تدريب – مناورة – قنص – رمي – صاروخ - المقص – حارس المرمى – الحارس – منطقة الجزاء - قوي		
أداء – دفع payment/paying /settle/pay off (debt)	(**) Sin_indef	مركبات مميِّزة	تسديد (جميع – كل – هذه – كافة - أي) (المستحقات – الغرامة – قسط – ديونه – الاشتراكات – رسوم – رسم – الفواتير – ثمن – الديون – مليار - المطالبات – مسبق – مبلغ – أتعاب – الحركات – قيمة – المخالفات – الحسابات – الضرائب – نفقات – المستحقات – الدَين – القرض – الفوائد – مبالغ – مديونيتها – أصل الدين – العجز – أقساط – جميع الديون – قيود بضائع – التزاماته المالية – الأعواز – جزء – المهر – التزاماته)
			دفعات تسديد
	sin_def	مركبات مميِّزة	التسديد (الآلي - لاحقًا - على أقساط)
			(لحين - تاريخ – جدولة – في حالة عدم – تأخر عن – امتنع عن – ممتنع عن - برنامج) التسديد
	(*)Sin_pro		كلمات سياقية
	Pl		غير وارد
	<u>كلمات سياقية</u> : أيام – دَين – سنوات – أموال – أنفق – مدة طويلة – غرامة – نقدي – سيولة – سداد – أشهر – أقساط – فواتير – خزينة – الدعم – مبالغ – مشروع – ديون – أعوام - الضريبة – النقود - فواتير – متخلفين – مصرف – أقساط – شهرية – عقوبة – إغلاق – محل – عميل – جدولة – ديون – بنك - خدمات – نقدي – خصم – إشعار – شهر – المبلغ – مدة – تأخر – تسهيلات – صندوق – ممتنع – سَنَة - أموال - مصرف – أجرة – عميل - حاجات – الدَين – مليار – تأمين – تأميني - قسط – ديونه – الاشتراكات – رسوم – رسم – (هذه) الفواتير – ثمن – الديون – مليار - المطالبات – مسبق – مبلغ – أتعاب – الحركات – قيمة – المخالفات – الحسابات – الضرائب – نفقات – المستحقات – الدين – القرض – الفوائد – مبالغ – مديونيتها – أصل الدين – العجز – أقساط – جميع الديون – قيود بضائع – المالية – الأعواز – جزء – المهر – التزاماته		

ملحق (2) : عينة من قائمة المداخل المعجمية الملتبسة في معجم جذوع المحلل الصرفيTim buckwalter
أولًا : الأسماء

جدول (4) عينة من قائمة المداخل المعجمية الملتبسة في معجم جذوع المحلل الصرفيTim buckwalter (الأسماء)

المدخل المعجمي	النطق	الوسم الصرفي	المعنى	التكرار
تذكير	ta*okiyr	NOUN	reminding	2
تذكير	ta*okiyr	NOUN	reminder/memento	2
تايمز	tAyomz	NOUN_PROP	Thames	2
تايمز	tAyomz	NOUN_PROP	Times	2
تبشير	tabo$iyr	NOUN	evangelization	2
تبشير	tabo$iyr	NOUN	announcement	2
تثليث	tavoliyv	NOUN	making three-fold/triangulating	2
تثليث	tavoliyv	NOUN	trinity	2
تثمين	tavomiyn	NOUN	appraisal/rating	2
تثمين	tavomiyn	NOUN	octagonal/eightfold	2
تجاوز	tajAwuz	NOUN	exceeding/overstepping	2
تجاوز	tajAwuz	NOUN	surmounting/overcoming	2
تحجير	taHojiyr	NOUN	petrification	2
تحجير	taHojiyr	NOUN	ban/interdiction	2
تحرير	taHoriyr	NOUN	liberation/liberating	2
تحرير	taHoriyr	NOUN	editorship/editing	2
تحسين	taHosiyn	NOUN	improving/making better	2
تحسين	taHosiyn	NOUN	improvement/beautification	2
تحصين	taHoSiyn	NOUN	immunization	2
تحصين	taHoSiyn	NOUN	fortification	2
تحقيق	taHoqiyq	NOUN	investigation/verification/interrogation	2
تحقيق	taHoqiyq	NOUN	achievement/realization	2
تحكّم	taHak~um	NOUN	control/controlling	2
تحكّم	taHak~um	NOUN	arbitrariness/despotism	2
تحلية	taHoliyap	NOUN	decoration/sweetening	2
تحلية	taHoliyap	NOUN	softening (water)/desalination	2
تخريج	taxoriyj	NOUN	graduation ceremony	2
تخريج	taxoriyj	NOUN	upbringing/extraction/derivation	2
تخشيبة	taxo$iybap	NOUN	wooden shed	2
تخشيبة	taxo$iybap	NOUN	jail cell	2
تخطيط	taxoTiyT	NOUN	planning/projecting	2
تخطيط	taxoTiyT	NOUN	graphing/imaging	2
تخلّف	taxal~uf	NOUN	tardiness/being late	2
تخلّف	taxal~uf	NOUN	backwardness/underdevelopment	2
تداخل	tadAxul	NOUN	reaction (against)/conflict (with)	2
تداخل	tadAxul	NOUN	interference/intervention	2
تربة	turobap	NOUN	grave/graveyard	2
تربة	turobap	NOUN	dust/ground	2
ترجمة	tarojamap	NOUN	translation/interpretation	2
ترجمة	tarojamap	NOUN	biography	2
تردّد	tarad~ud	NOUN	frequency	2
تردّد	tarad~ud	NOUN	frequentation/reluctance	2
ترويض	tarowiyD	NOUN	sports	2
ترويض	tarowiyD	NOUN	domesticating/pacifying/regulating	2
تسديد	tasodiyd	NOUN	payment/paying/settle/pay off (debt)	2
تسديد	tasodiyd	NOUN	aiming/shooting	2
تشخيص	ta$oxiyS	NOUN	personification/characterization	2
تشخيص	ta$oxiyS	NOUN	diagnosis/analysis	2

ثانيًا : الأفعال

عينة من قائمة المداخل المعجمية الملتبسة في معجم جذوع المحلل الصرفيTim buckwalter (الأفعال)

جدول (5) عينة من قائمة المداخل المعجمية الملتبسة في معجم جذوع المحلل الصرفيTim buckwalter (الأفعال)

المدخل المعجمي	النطق	المعنى	التكرار
أبد	>abada	persist/remain/stay + he/it [verb] be untamed/escape + he/it [verb]	2
أحقّ	>aHaq~a	enforce/make right + he/it [verb] be right/be allowed + he/it [verb]	2
أدّى	>ad~aY	perform (function)/carry out (duty) + he/it [verb] direct/guide/lead + he/it [verb]	2
أدرك	>adoraka	reach/attain + he/it [verb] comprehend/realize + he/it [verb]	2
أراب	>arAba	disquiet/fill with misgivings + he/it [verb] make curdle + he/it [verb]	2
أرّخ	>ar~axa	date + he/it [verb] report/chronicle + he/it [verb]	2
أساء	>asA'a	do badly/mismanage + he/it [verb] harm/offend + he/it [verb]	2
أسمى	>asomaY	elevate/exalt + he/it [verb] name/designate + he/it [verb]	2
آسى	|saY	grieve/afflict + he/it [verb] console/comfort + he/it [verb]	2
أشر	>a$ara	cut with a saw + he/it [verb] sharpen/file + he/it [verb]	2
أصفر	>aSofara	empty + he/it [verb] be empty-handed + he/it [verb]	2
أغار	>agAra	make jealous + he/it [verb] attack/invade/raid + he/it [verb]	2
أغلى	>agolaY	boil/make boil + he/it [verb] raise (price)/make expensive + he/it [verb]	2
أقام	>aqAma	reside/live + he/it [verb] install/establish/erect + he/it [verb]	2
أقرّ	>aqar~a	ratify/accept + he/it [verb] console/pacify + he/it [verb]	2
أكسد	>akosada	oxidize/rust + he/it [verb] be stagnant/be paralyzed + he/it [verb]	2
آلف	|lafa	adapt/familiarize + he/it [verb] befriend/adapt to + he/it [verb]	2
آنس	|nasa	entertain/perceive + he/it [verb] be friendly/entertain + he/it [verb]	2
ائتلف	{i}otalafa	form a coalition + he/it [verb] be accustomed/be harmonious + he/it [verb]	2
ابتعث	{ibotaEava	send/dispatch + he/it [verb] exhume/revive + he/it [verb]	2
اتّصل	{it~aSala	be connected or related (to) + he/it [verb] contact/get in touch (with) + he/it [verb]	2
اختطّ	{ixotaT~a	trace/mark + he/it [verb] plan/devise + he/it [verb]	2
اختطف	{ixotaTafa	abduct/kidnap + he/it [verb] hijack + he/it [verb]	2
ارتطم	{irotaTama	be involved/be implicated + he/it [verb] crash/impact + he/it [verb]	2
استباح	{isotabAHa	allow/seize + he/it [verb] behave licentiously + he/it [verb]	2
استحيى	{isotaHoyaY	let live/keep alive + he/it [verb] be embarrassed/be shy + he/it [verb]	2

أهم المراجع العربية

[1] علي عبد الواحد وافي ، علم اللغة ، نهضة مصر للطباعة والنشر ، الطبعة التاسعة ، إبريل (2004م).
[2] صبحي صالح ، دراسات في فقه اللغة ، دار العلم للملايين ، بيروت ، لبنان ، الطبعة الثالثة ، 1388هـ.
[3] إبراهيم أنيس : دلالة الألفاظ، مكتبة الأنجلو المصرية ، الطبعة الخامسة ، (1984م).
[4] إبراهيم أنيس :في اللهجات العربية ، مكتبة الأنجلو المصرية ، القاهرة ، الطبعة الثالثة ، (1965م).
[5] إدريس ميموني : قضايا الدلالة في اللغة العربية بين الأصوليين واللغويين: المشترك اللفظي نموذجا ، مجلة محكمية علوم إنسانية ، السنة السابعة: العدد 42: (2009م).
[6] أحمد مختار عمر ، علم الدلالة، عالم الكتب مصر، الطبعة الثانية، (1988م).
[7] السيوطي: المزهر في علوم اللغة وأنواعها، شرح وتعليق، محمد أحمد جاد المولى بك، محمد أبو الفضل إبراهيم، علي محمد البجاوي، الطبعة الثالثة ، مكتبة دار التراث ، القاهرة.
[8] عبد الواحد وافي: فقه اللغة، دار نهضة مصر للطباعة والنشر، الفجالة، القاهرة.
[9] إبراهيم أنيس: في اللهجات العربية، المطبعة الفنية الحديثة.
[10] صبحي الصالح: دراسات في فقه اللغة، دار العلم للملايين، بيروت ، لبنان.
[11] عبد العال سالم مكرم : المشترك اللفظي في ضوء غريب القرآن الكريم ، عالم الكتب ، القاهرة ، (2009م).
[12] سيبويه : الكتاب، تحقيق عبد السلام هارون، دار الكتب، بيروت، لبنان(1403هـ/1983م).

أهم المراجع الأجنبية

[13] A. Zouaghi, L. Merhben, and M.Zrigui (2010). "*Ambiguous Arabic Words Disambiguation: The results*". Published in: Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), 11th ACIS International Conference on June 2010, pages 157 – 164
[14] A. Zouaghi, L.Merhbene and M.Zrigui (2012) "*Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation*". Published in: Journal Artificial Intelligence Review, Volume 38, Issue 4, December 2012 , Pages 257-269.
[15] T. Buckwalter, (2003) "*Buckwalter Arabic Corpus*" homepage. http://www.qamus.org/wordlist.htm. (accessed June 14th, 2005)
[16] David Yarowsky (1995) "*Unsupervised word sense disambiguation rivaling supervised methods*". Proceeding ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Pages 189-196
[17] D.yarowsky. (1992). "*Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora*". Proceedings of the 14th conference on Computational linguistics:COLING -92, Volume 2.
[18] N. Habash&O. Rambow (2005). "*Arabic Tokenization, Part-of-Speech tagging and morphological disambiguation in one fell swoop*". In Proceedings of the Association for Computational Linguistics (ACL, pp. 573–580).
[19] N. Habash. (2010)"*Introduction to Arabic Natural Language*" Processing. Morgan & Claypool Publishers.
[20] N. Habash, O. Rambow and R. Roth (2009). "*MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization*". In

Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

[21] S.Abdelhadi et al. (2007) "*Arabic Computational Morphology: Knowledge-based and Empirical Methods*", Springer Publishing Company.

[22] M. Lesk, (1986), "*Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*", in 'SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation' , ACM, New York, NY, USA , pp. 24--26 .

[23] M.Diab (2003) "*Word sense disambiguation within a multilingual framework*". PhD dissertation, University of Maryland, USA.

[24] T. Pedersen (2006) "*Unsupervised Corpus-Based Methods for WSD*". Book "*Word Sense Disambiguation : algorithms and applications*", Agirre, Eneko, Edmonds, Philip (Eds.), Text, Speech and Language Technology, Vol. 33 chapter 6 , pp 133-166.

[25] R. Navigli. "*Word Sense Disambiguation: a Survey*". ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date: February 2009 [A complete State of the Art in Word Sense Disambiguation]

[26] R.Mihalcea and EhsanulFaruque (2004) "*Minimally supervised word sense disambiguation for all words in open text*". In Proceedings of ACL/SIGLEX Senseval-3.

[27] S.Elmougy, T. Hamza, and H.Noaman. (2008) "*Naïve Bayes Classifier for Arabic Word Sense Disambiguation*", The 6th International Conference on Informatics and Systems, INFOS2008, March 27-29, 2008.

[28] T. Pedersen (1998) "*Knowledge lean word sense disambiguation*". In Proceedings of the Fifteenth National Conference on Artificial Intelligence.

Word Sense Disambiguation for Buckwalter Arabic Morphological Analyzer results

Ahmed Abdelghany[1] ،Sameh Alansay [2]

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

[1] hmd_abdelghany@yahoo.com

[2]sameh.alansary@bibalex.org

***Abstract*-Tim Buckwalter Arabic morphological analyzer is considered one of the most popular Arabic morphological analyzers in literatures of Arabic language processing automatically, this may be due to reasons related to ease of use, availability, and possibility of modifying lexicons and analysis algorithm freely. These reasons and others encourage researchers to enhance results of analysis through expanding lexicons and modifying algorithm to disambiguate solutions automatically. This research aims to handling the semantic aspect of the morphological solutions disambiguated automatically in morphological disambiguation stage making use of the morphological properties of the defined solutions. The scope of the research includes semantic disambiguation results in polysemy (and doesn't include that results in missing of diacritics). In this paper, the researcher will deduce a list of linguistic and nonlinguistic cues for disambiguating word senses through exploring a representative corpus of Arabic (ICE). Then the researcher will propose a model for implementing these cues logically.**

Ahmed Abdelghany

Ahmed abdelghany is an assistant lecturer of computational linguistics in the Department of Linguistics and Phonetics Department, Faculty of Arts, Alexandria University. He obtained his MA (Building a morphologically analyzed corpus for Modern Standard Arabic) in 2010, His main areas of interest are concerned with corpus work, morphological analysis and generation, and traditional Arabic morphology and syntax.

He is also a member of the Arabic center team of UNL project in bibliotheca Alexandrina, his work was since 2005 till now. He studied MCIT scholarship in New Horizons in 2006 for 9 months studying Visual Basic.NET programming language, SQL server developing database, English language courses and soft skills (preparing presentation, working in teams, problem solving, understanding leadership, project management and customer service and sales skills).

Dr. Sameh Alansary

Director of Arabic Computational Linguistics Center Bibliotheca Alexandrina

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Speaker Identification Based on Temporal Parameters

Eman M. Yousri[*1], Mervat Fashal[**2]

[*] *Post-graduate student of Phonetics & Linguistics Dep., Faculty of Arts, University of Alexandria. Alexandria, Egypt.*

[1]emanyousri88@yahoo.com

[**]*Professor of phonetic sciences, Phonetics & Linguistics Dep., Faculty of Arts, University of Alexandria. Alexandria, Egypt.*

[2]mervat.fashal@alexu.edu.eg

Abstract — **The subject of this study is to identify unknown speakers particularly from their speaking tempi represented in Speech Rate SR and Articulation Rate AR as temporal parameters. The fundamental goal of this study, on the acoustical level, is to prove acoustically that every speaker has a significant speech rate SR and articulation rate AR through which the unknown speaker can be discriminated and to investigate which of them (SR or AR) could be more benefit for identifying unknown speakers and to what extent. Also, the present study is essentially concerned, on the perceptual level, with listeners' perceptual abilities in perceiving and differentiating different speaking tempi for identifying unknown speakers in order to utilize this exceptional ability in forensic speaker identification FSI; aiming to provide some useful acoustical and perceptual data to be used in forensic phonetic filed. The most important characteristic of the temporal aspects of speech, that they are not easily disguised or imitated by accent or fundamental frequency leveling; so they could be useful for identifying unknown speakers particularly in forensic phonetic field.**

The speech rate SR and articulation rate AR of ten unknown speakers / informants of colloquial Arabic are calculated. The speakers were recorded while talking spontaneously for a radio program. Only 30 seconds of speech are cut for each speaker from the entire episode. After that 60 naïve listeners are asked to listen carefully to the 10 unknown informants in order to mark the fastest speaker and the slowest speaker depending only on their ears.

1 INTRODUCTION

Speaker Identification is the task of deciding and determining a given sample of speech (uttered by unknown speaker), who among many candidate speakers said it. The unknown speaker is defined as the speaker hose model best matches the given utterance (Furui 2008). *Forensic Speaker Identification FSI* is considered as one of the most significant practical applications of speaker identification. FSI is defined as the most central aspect of forensic phonetics and acoustics which mainly concerned with solving problems related to identification of the unknown speaker in criminal investigation to identify suspects who were heard but not seen committing a crime including; murder, blackmail threats, ransom calls, kidnapping, political corruption, bomb threats, terrorist activities, etc. (Singh, Khan& Shree, 2012; Jessen, 2008; Nolan, McDougall, DeJong & Hudson, 2006; Lindh, 2004; Eriksson, 2005; Rose, 2002; and Nolan, 2005).

There will be always differences (which are always audible, measurable and quantifiable) between speech samples, even if they come from the same speaker. This is due to two kinds of variability: *1) organic vs. phonetic variability*, and *2) between speaker vs. within speaker variability.* Consequently, the main task of Forensic Speaker Identification FSI is to find all the sources of variability in order to make a clear distinction for the correct evaluation.

For speaker identification in forensic situation as evidence in the court, there are four main phonetic/acoustic parameters depending on the speaker through them he / she can be discriminated and identified:

1. The Fundamental Frequency F_0.
2. The formants frequencies of the vowels.
3. The resonance of the nasal consonants.
4. Tempo of speaking.

Tempo of speaking; the fourth parameter is our concern here; it is a multidimensional phenomenon and revealing the temporal aspects of the speech. It is also one of the prosodic cues which considered as non-linguistic factor that signaling *paralinguistic* information (about the situation and the inner state of the speaker's attitudinal or emotional state) and also *extra-linguistic* information (about the speaker's identity, personality and individuality) (Trouvain, 2003; and Rose, 2002). *Tempo of speaking* can be exhibited by two methods, one is *Speech Rate (SR),* and another is *Articulation Rate (AR).* Both of SR and AR can be defined as "the number of syllables per second". The biggest difference between SR and AR is that the SR includes pause intervals but the AR does not (Gold, 2012; and Koreman, 2006).

Tempo of speaking has significant importance in Forensic Speaker Identification FSI Demenko (2000) because it is:

1. Carrying the individual-identifying information about the speaker.
2. Affected by the individuals variations in speaking.
3. Not affected by the frequency characteristics of the transmission systems and at the level at which the speaker talks.

4. Not easy to imitate or disguise.
5. Not controlled by the speaker.

2 METHODOLOGY

A. Data Collection

The experiment includes 10 unknown speakers (5 females and 5 males) of colloquial Arabic language, with no recorded speech disorders. Speaker's ages estimated between 19 to 40 years old. Natural spontaneous speaking style is elicited for 30 seconds for each speaker trying to avoid the effect of any stress or the domination of any specific emotion. All the data are collected through a radio program called "the press in their eyes *الصحافة فى عيونهم*" which is a daily program that announced every day at Alexandria Radio (Bakous Alex, frequency 101.1). The announcer of the program goes down to the street every day and asks one of the public. This one of the public could be a male or a female who was reading one of the daily newspaper and his or her identity is unknown for the announcer and for the listeners. The announcer asks a simple question which is: what's your comment about one of the news that you have been read at that daily journal? Then, the unknown speaker starts to talk spontaneously, without any recommended preparation, about any topic that he or she chooses. Accordingly, that unknown speaker is one of the Alexandrian populations who may get intermediate education (which enables that unknown speaker to read the daily journals) or may be well educated.

B. Recordings

The data are collected and elicited through the announcer who asks the unknown speaker about his/ her comments or opinions about any piece of news of the daily journal headlines. The whole duration of each episode is (about 5 minutes for every speaker) directly recorded from the radio channel using **Samsung mobile phone recorder as wav. files**; to avoid any transmission distortions. Then, all the episodes (10 episodes of 10 unknown speakers, each of which is 5 minutes) are transmitted into a laptop device for editing. Therefore, the researcher used cutter software for cutting only 30 seconds of continuous and spontaneous speech of each speaker from the whole speaking time (from the whole episode which is 5 minutes).this cutter software is called **"Easy audio ogg wma wav cutter software** (www.koyotesoft.com). At last all the edited data (only 30 seconds of spontaneous speech for 10 unknown speakers) are exposed to **Praat software** (www.praat.org) for the analysis (next step).

C. Analyses

All the data are analyzed manually with the aid of Praat software for all speakers. The analysis procedure is composed of three sequential steps which are:
The first step is the transcription process in which every 30 seconds of recording spontaneous speech for each unknown speaker are phonetically transcribed by using IPA symbols. The researcher transcribed all the data manually through the careful listening depending on the ears of the researcher with the aid of Praat software as a listening tool. Broad transcription type is used for this research because the main concern of that transcription process is counting the number of the pronounced syllables in a particular time (which is 30 seconds of spontaneous speech for each informant). So, no matter of how an informant is pronouncing a particular phoneme as long as does not affect the number of the pronounced syllables.
The second step is the segmentation process which means dividing the transcribed speech into syllables; this process is done manually by the researcher.
The third step is the calculation process in which speech rate SR and the articulation rate AR are calculated with their durations. Also the number of pauses and the duration of each pause are counted too.

D. Measurements

All the acoustical measurements illustrated with their mean of calculation for all the ten unknown speakers:

- *Fundamental frequency f_0* is measured for all speakers using praat voice report.
- *Intensity* is measured for all the speakers with praat software through getting the mean intensity.
- *The number of the pronounced syllables* for each speaker, how many numbers of syllables the speaker has pronounced in only 30 seconds. The number of the pronounced syllables calculated manually by the researcher through counting all the produced syllables after segmentation process.
- *Speech rate SR* is measured according to the following definition "the number of syllables per second including the whole speaking time (with all pauses and hesitations)"; which is 30 seconds for each speaker.
- *Articulation rate AR* is measured according to the following definition "the number of syllables per second excluding the pause time and all the hesitation duration". Note that the excluded pause time and hesitation duration will vary from one speaker to another.
- *All pauses durations* are measured by combining the duration of each pause in each speaker's utterance and the duration of each pause between utterances.

- *The number of pauses* for each speaker; is counted manually by the researcher, through counting the number of all pauses (filled and silent) occurred in the whole speaking utterance (occurred in 30 seconds for each speaker).
- *The duration of each pause* occurred in the whole speech sample (in 30 seconds) for each speaker with the aid of praat software. And also, determining the type of each pause.
- *Percentage of pause time* is measured manually by the researcher, through calculating the proportion of all the pauses time (the duration of all pauses) to the whole time of the speech sample (which is 30 seconds).
- *The degree of hesitancy* is measured manually by the researcher for each speaker through calculating the proportion of filled pauses to all pauses for the overall speech sample.

E. Perceptual Test

Sixty listeners of university students aged between 17 and 25 years old, with no recorded history of hearing impairments. Each listener was sitting directly in front of a laptop computer device with approximately three feet distance. The listeners were listening to the voice line-up (mp3 playlist, with 2 seconds interval between each informant and the following) through a loud speaker (attached to the laptop computer device) which was set up on medium volume.
The listeners were received some instructions from the researcher for doing the perceptual test perfectly:

1. Each listener received a listening sheet (see Figure 2) which contained the ten unknown speakers (5 females and 5 males listed one by one) titled as informant 1, informant 2, …………., informant 10.
2. The listeners are asked to listen carefully to the voice line-up of the ten unknown informants three times at most in order to enabling them to select the fastest speaker and the slowest speaker.
3. Then, each listener selected the fastest speaker and the slowest one by marking (√) in front of his or her title at the listening sheet (see Figure 1).

Observe that, the ten informants' voices intended to be listed one by one (male followed by female) in the voice-line up; in order to distract the listeners' attentions from the gender of the speaker. Because, almost all acoustic measurements and perceptual expert descriptions show experimentally that there are no significant differences in speech tempo between men and women. In other words, tempo of speech has no relation to the gender of the speaker.

Speakers المتكلمون	**The FASTEST الأسرع**	**The SLOWEST الأبطأ**
Informant 1		
Informant 2		
Informant 3		
Informant 4		
Informant 5		
Informant 6		
Informant 7		
Informant 8		
Informant 9		
Informant 10		

Figure 1: the listening sheet; where the involved listeners are marking (√) in front the fastest informant and the slowest one too.

3 RESULTS

A. Perceptual Test Results

The following figure (Figure 2) showing the distribution of all the listeners' selections percentages for both the fastest speaker and the slowest speaker as well. Through glancing over Figure 2, it's noticed that, the percentages of listeners' selections are highly distributed across all the ten informants with varied degrees which reveal that there is no absolute agreement about a particular speaker whether the slowest or the fastest.

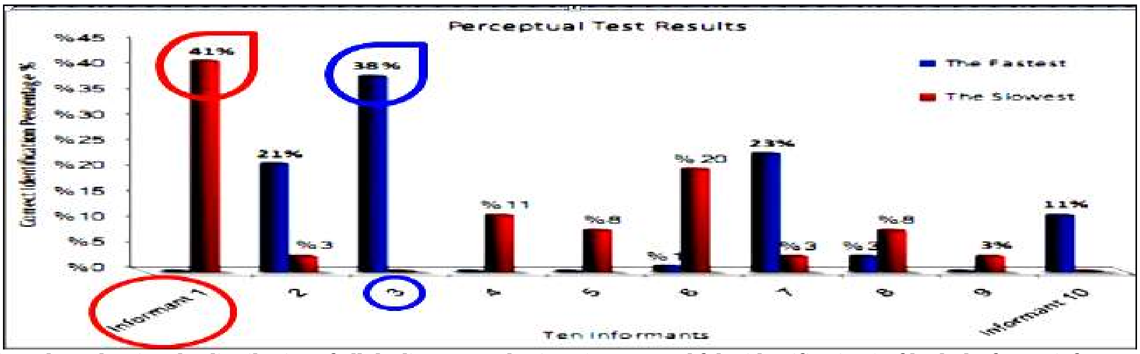


Figure 2: line- chart showing the distribution of all the listeners' selections (correct and false identifications) of both the fastest informant and the slowest informant depending only on their ears.

B. Acoustical Test Results

Figure 3 represented speech rate SR and articulation rate AR values for all the ten informants. With respect to *the fastest speaker*; informant 3 (male) is the fastest speaker with the highest SR= 7.2 S.S. and AR= 8.520 S.S. He pronounced the largest number of syllables in 30 seconds (216 syllables); he also has the highest F_0 between male speakers (187 Hz). Regarding *the slowest speaker speech rate SR*; informant 9 (male) is the slowest speaker with SR= 5.2 S.S. and he produced the minimum number of syllables in 30 seconds (155 syllables).

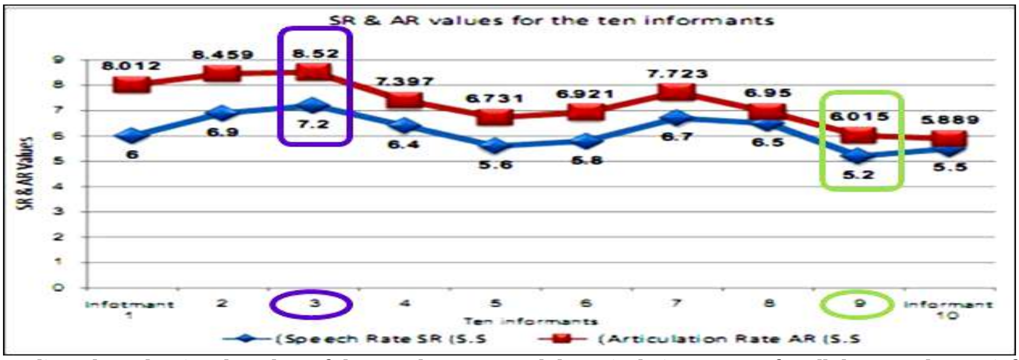


Figure 3: line- chart showing the values of the speech rate SR and the articulation rate AR for all the ten unknown informants.

According figure 4 that showing us the mean intensity of all the ten unknown speakers, regarding *high intensity degrees*, informant 2 (female) recorded the highest degree of intensity = 82 dB. Whereas, informant 1 (male) recorded *the lowest intensity degree* = 60 dB.

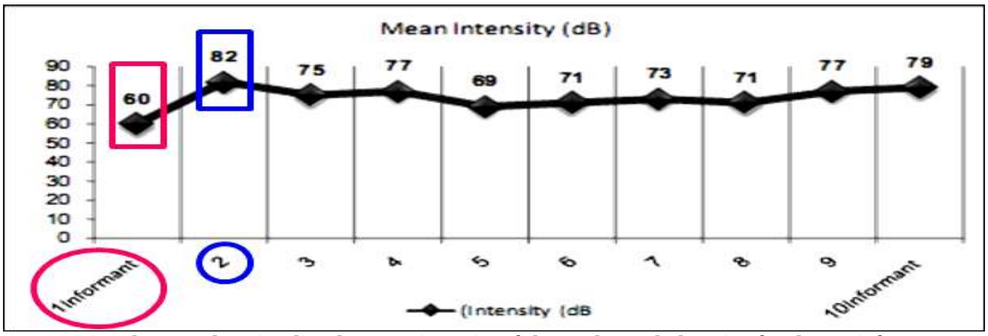


Figure 4 : Line chart used to visualize the mean intensity of the total speech duration for the ten informants.

Primarily the percentage of all pauses to the overall speech sample is depending on both; the number of pauses (pauses' frequencies of occurrences) as well as their durations. According to the following figure (Figure 5), respecting *the highest percentage of all pauses to overall speech sample*, informant 1 has the highest percentage of pauses (28 % of his speech sample is consisting of pauses). Whereas, informant 8 has the lowest percentage of pauses (only 10.33 % of her speech sample is consisting of pauses).

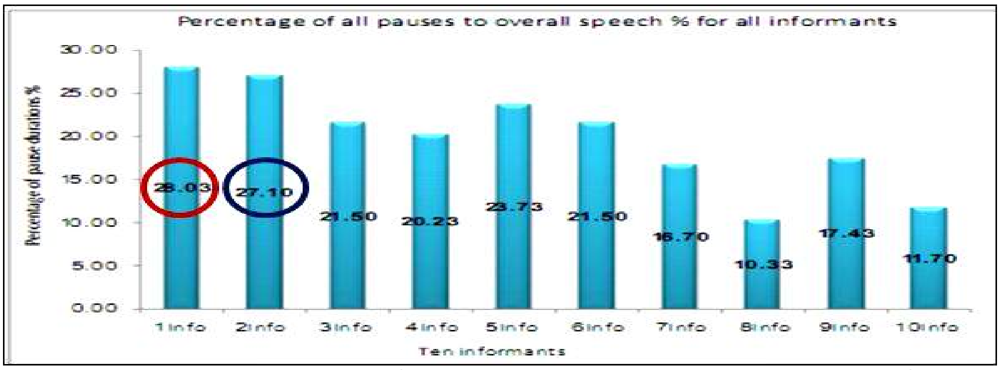


Figure 5: column - chart showing the percentage of all pauses and hesitations to the overall speech sample for each speaker.

The degree of hesitancy for each informant shows the proportion of filled pauses to all pauses for the overall speech sample to indicate large differences between speakers (intra speaker variation) and relatively small differences within speaker (inter speaker variation). Figure 6 indicates that informant 5 (who is arranged as the third slow speaker according his speaking rate and he has the lowest F_0 between male speakers) has *the highest degree of hesitancy* (66.7 %), which may indicate that the high degree of hesitancy may negatively affect the perceived speaking rate. In other words; high degree of hesitancy may be considered as a sign of slow speaking rate. To confirm this, we need more experimental research. Figure 6 also indicates that informant 3 (who is the fastest speaker according to his speaking rate and he has the highest F_0 between male speakers) has *the lowest degree of hesitancy* (23.5 %). Regarding the results of the present experiment; the degree of hesitancy seems to have an inverse relation with speaking tempo particularly at fast speaking tempo. In other words; the fastest speaker (according to speech rate SR and articulation rate AR) has the least degree of hesitancy. And this relation is compatible with only the fast speaking rate.

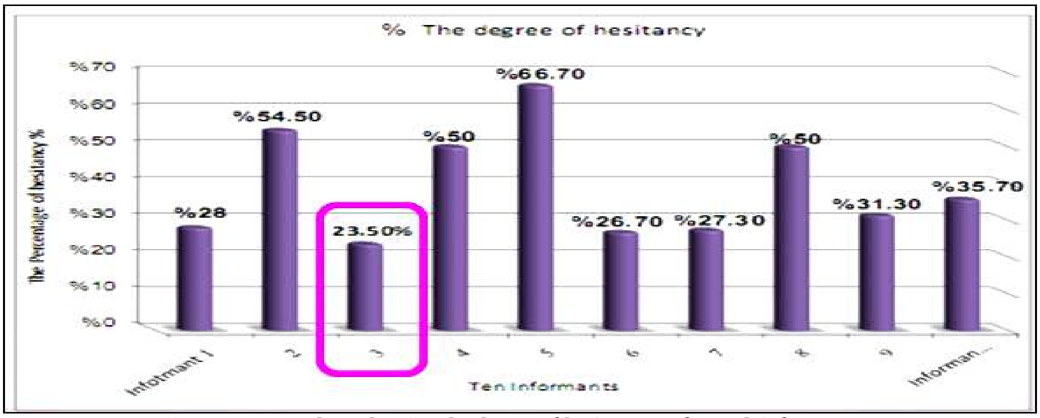


Figure 6: cone - chart showing the degree of hesitancy % for each informant.

4 CONCLUSIONS

There are many phenomena observable in speeded up and slowed down speech. And there are lots of parameters that influenced the rate of speech as well as its perception. So, from the preceding experiment and its results; acoustically as well as perceptually; we can deduce the following conclusions:

1. Acoustically and perceptually, SR is most powerful in identifying unknown speakers than AR. But this does not mean to exclude the articulation rate AR.
2. The percentage of all pauses plays a double-edged role. On the perceptual level, large percentage of pauses durations considered one of the most important factor that influencing the listeners' perception of the rate of speech. Whereas, on the acoustical level, they don't have any obvious effectiveness on modifying the rate of speaking.
3. The degree of hesitancy, acoustically, it is considered as a remarkable factor for the fastest speaking tempo. But not in identifying the slowest speaking tempo.
4. F0 is an important acoustic cue in identifying the speaker's speaking rate acoustically and perceptually as well. High F0 (for male or female speaker) indicated fast speaking tempo.
5. Mean intensity, perceptually, is a remarkable cue for listeners' perception in identifying the rate of speaking of the speaker (whether the slowest or the fastest). High intensity indicated fast speaking tempo; and low intensity indicated slow speaking tempo.

ACKNOWLEDGMENT

I want to give the great thank to my supervisor Dr. Mervat Fashal, the professor of phonetic sciences at the department of phonetics and linguistics, for her guidance, supporting and for giving me this great opportunity. I want also to express my sincere appreciation to Dr. Sameh Al-Ansary, the headmaster of the phonetics and linguistics department for his remarkable effort for our department.

REFERENCES

[1] Demenko G. (2000) Analysis of supra-segmental features for speaker verification. Institute of linguistics, Adam Mickiewicz University, Poznan, Poland.

[2] Eriksson, A. (2005) Tutorial on forensic speech science. Part I: Forensic phonetics. In Interspeech, Eurospeech 2005. Proceedings of the 9th *European conference on speech communication and technology*. Department of Linguistics, Gothenburg University, Gothenburg,Sweden2005.Website:http://www.york.ac.uk/media/languageandlinguistics/documents/currentstudents/Eriksson_tutorial_paper.pdf

[3] Furui, S. (2008) Speaker Recognition. Tokyo Institute of Technology. Scholarpedia, 3(4):3715. doi:10.4249/scholarpedia.3715. Website: http://www.scholarpedia.org/article/Speaker_recognition

[4] Gold, E. (2012) Articulation Rate as a Discriminant In Forensic Speaker Comparisons. *UNSW Forensic Speech Science Conference* Sydney, Australia. Website: http://sydney2012.forensic-voice-comparison.net/

[5] Jessen M. and Bundeskriminalamt BKA (2008) Forensic Phonetics. *Language and Linguistics* Compass 2/4 (2008): 671–711.

[6] Koreman, J. (2006) The role of articulation rate in distinguishing fast and slow speaker. Institute of Phonetics Saarland University, Germany. jkoreman@coli.uni-saarland.de.

[7] Lindh, J. (2004) Handling the "Voiceprint" Issue. In Proceedings, FONETIK 2004, Dept. of Linguistics, Stockholm University.

[8] Nolan F., McDougall K., Jong D. G. & Hudson T. (2006) A Forensic Phonetic Study of 'Dynamic' Sources of Variability in Speech: The DyViS Project. Department of Linguistics, University of Cambridge, United Kingdom, fjn1@cam.ac.uk , kem37@cam.ac.uk , gd288@cam.ac.uk , toh22@cam.ac.uk.

[9] Nolan, F. and Catalin G. (2005) A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12 (2): 143-173.

[10] Rose, P. (2002) *Forensic Speaker Identification.* London, UK: Taylor and Francis.

[11] Singh, N., Khan, R. A. and Shree, R. (2012) Applications of Speaker Recognition. *International conference of modeling, optimization and computing (ICMOC 2012),* Procedia Engineering Conference 2012.

[12] Trouvain, J. (2003) *Tempo variation in speech production, implications for speech synthesis*. PH.D dissertation.

BIOGRAPHY

Eman M. Yousri I am graduated from the department of phonetics and linguistics 2009. My graduation project was about voice print analysis for forensic speaker identification in courts. I got my master degree in phonetic science in June 2015. My thesis also was about forensic speaker identification depending on temporal parameters.

Prof. Mervat Fashal is a professor of phonetic sciences, and in particular in psycho-acoustic field, Dept. of Phonetics & Linguistics, Faculty of Arts, Alexandria University.

She had studied Ph.D. in Germany. She has articles and experiments on various areas of phonetics. The significant contributions in the field are mainly in the following topics:

- Speech production and perception
- Prosodic and discourse analysis
- Acoustic analysis of normal and abnormal speech
- Speech recognition and speaker identification in the field of forensic phonetics

Mervat Mohamed Ahmed Fashal is a Full Professor since 2008, a Head of Phonetics Department from 2003 - 2012.

التعرف على المتكلم اعتماداً على معايير السرعة الزمنية

إيمان يسري [1]*,ميرفت فشل[2]**

* طالبة دراسات عليا بقسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الإسكندرية، مصر

[1]emanyousri88@yahoo.com

**أستاذ قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الإسكندرية، مصر

[2]mervat.fashal@alexu.edu.eg

هدف هذه الدراسة هو التعرف على هوية المتكلمين غير المعروفين من سرعة كلامهم، وقد تم فى هذا البحث -**على المستوى الإدراكي** - عمل تقييم للقدرة الإدراكية للمستمعين غير المدربين فى التعرف المتكلم اعتمادا على سرعة كلامه، وإدراك ما إذا كان الأسرع أم الأبطأ بين المتكلمين العشرة الذين تم اختيارهم للتجربة. أما **على المستوى الأكوستيكي،** فقد تم رصد المعايير الفيزيائية الأساسية للتعرف على صوت المتكلم وهى كالآتى:

1. التردد الأساسى F_0
2. الترددات المكونة للصوائت(F1, F2, F3).
3. الرنين الأنفي للصوامت (الغنة).
4. معدل سرعة الكلام (SR) ومعدل سرعةالمنطوقات (AR)

وقد اختير العنصر الأخير وهو سرعة الكلام (Speech Tempo) كموضوع لهذه الدراسة. وقد تم عمل التحليل الفيزيائى لكلام المتحدثين وقياس معدل سرعة الكلام ومعدل سرعة المنطوقات والوقفات فى كلام كل متحدث (أطوالهم وأعدادهم). هذا فضلا على قياس التردد الأساسى لكل متكلم F_0 وشدة الصوت (I).

هناك العديد من الأسباب الأساسية التي توضح مدى أهمية المعايير الزمنية ومعدل سرعة الكلام فى التعرف على المتكلم للأغراض القضائية وهى كالآتى:

- لا يمكن محاكاة سمات السرعة الزمنية للكلام.
- لا يمكن للمتكلم السيطرة على السرعة الزمنية لكلامه بشكل واع.
- الفروق الفردية بين المتكلمين تُعد من أهم مصادر التغير التي تُؤثر على معدل سرعة الكلام.

تشمل هذة التجربة عشرة أشخاص (خمس نساء وخمسة رجال) غير معروفين الهوية ومتحدثين أصليين للهجة العامية العربية وتقدر أعمارهم بين 19 و 40 عام. تتكون المادة من كلام تلقائي لمدة نصف دقيقة (30 ثانية) لكل متكلم مع تجنب تأثير أو سيطرة أي نوع من أنواع المشاعر السلبية للمتكلمين.

تم تسجيل المادة من خلال برنامج "الصحافة فى عيونهم" الذي يذاع يومياً على راديو إذاعة الإسكندرية. وتم تحليل المادة المسجلة لكل متكلم يدويا وكتابتها بالرموز الصوتية Transcription وذلك عن طريق الإستماع الجيد لهذه المادة المسجلة مراراً وتكراراً بواسطة Praat Software. ثم تمت عملية فصل المقاطع Segmentation Process وذلك لحساب SR & AR. كما تم أيضا قياس التردد الأساسي و شدة الصوت لكل متكلم ودرجة التلعثم و عدد الوقفات وزمن كل وقفة ونوعها ونسبة كل الوقفات إلى مدة الكلام الكاملة .

ستون مستمع من طلبة الجامعات ومتحدثين أصليين أيضاً للعامية العربية المصرية وتتراوح أعمارهم بين 17 و25 عام ، جميعهم تطوعوا للإشتراك فى هذا الإختبار. المهمة الأساسية للمستمعين هى الإستماع بحرص شديد إلى المتكلمين العشرة وتحديد المتكلم الأسرع وأيضا المتكلم الأبطأ من حيث سرعة الكلام عن طريق وضع علامة (√) أمام الرمز الدال عليه.

تشير النتائج إلي أن:

1. أكوستيكياً و إدراكياً: سرعة الكلام موضحة فى معدل سرعة الكلام (SR)هى المعيار الأقوى فى التعرف على المتكلمين غير المعروفين; بينما معدل سرعة نطق الأصوات (الصوامت والصوائت) (AR) كان أقل تأثيرا على تحديد سرعة المتكلم.
2. النسبة المئوية للوقفات تلعب دوراً مهما جداً على المستويين الإدراكي والأكوستيكي؛ على المستوى الإدراكي فإن زيادة النسبة المئوية للوقفات تُعدّ من أهم العناصر التي تؤثر على إدراك المستمعين للسرعة الزمنية للكلام، حيث تشير إلى سرعة الكلام البطيئة. أما على المستوى الأكوستيكي: فليس لها أي تأثير واضح على زيادة أو نقصان سرعة الكلام للمتكلم.
3. درجة التلعثم فى الكلام (الوقفات المملوءةfilled pauses)، تُعد من العناصر المُميزة فى التعرف على المتكلم الأسرع من حيث سرعة الكلام للمتكلم. ومع ذلك فليس لها أي دور فعّال فى التعرف على المتكلم الأبطأ .
4. التردد الأساسي للمتكلم يُعد من العناصر الأكوستيكية المُميزة لتحديد سرعة الكلام للمتكلم ،بحيث زيادة التردد الأساسي للمتكلم تشير إلى زيادة معدل سرعة كلامه إدراكيا وأكوستيكيا.
5. متوسط شدة الصوت لدى المتكلم يُعد من الناحية الإدراكية من العناصر المُميزة بالنسبة إلى آذان المستمعين، بحيث زيادة شدة الصوت تشير إلى زيادة معدل سرعة الكلام للمتكلم ,وأيضا نقصان شدة الصوت تدل على نقصان معدل سرعة الكلام للمتكلم. ولكن هذه النتائج لا تنطبق على المستوى الأكوستيكي.