# Panel Discussion:
# Towards A Road Map of
# Human Language Technology Development
# For the Arabic Language

## Abstract

Arabic language speakers are over 300 millions world wide, especially that Arabic is one of the few official languages used in the UN. Accordingly, the interest in Arabic language exceeds the Arabic region. In this paper a road map for the research and development in the area of human language technology (HLT) is introduced. An organization is suggested that has the following mandates:

1. Conduct Situational analysis
2. Draw a strategic plan for Arabic HLT for 10 years ahead
3. Conduct Training and specialized diploma
4. Motivating the community through: Language Resources for training and benchmarking and periodical competitions to discover the real potential of the community.
5. Cooperation with other complementary associations

## 1. Introduction

- **HLT means** all the natural language processing (NLP) and speech processing technologies (see appendix A for examples of such technologies) [9].

- **Arabic Language spread:**
  Arabic Speakers are over 300millions all over the glob. Arabic is one of the few language officially used in the UN. The interest in Arabic exceeds the Arabic region.

- **HLT need in the digital era:** The growing size of the digital content on the internet with different languages is the reason for the pressing demand for Natural Language Processing for text and speech applications.

  The great spread of the mobile sets lunches the mobile era in almost all applications. M-learning, M-health,… etc are starting to set new standards in all disciplines; that adds more spread and in the same time more challenges to the NLP applications and technologies.

## 2. Arabic Language challenges in HLT

Arabic language has more challenges than a language like English; and among these reasons [11],[12]:

- **Morphologically-rich language:** Arabic is a very morphologically–rich language that produces very high number of word forms for a given root. That leads to a lot of challenges for NLP technologies and applications and for speech processing.

- **Flexible Syntax structure:** Arabic has more flexible word sequence. The Arabic sentence can be with or without verb.
- **No enough punctuation marks:** In Arabic it is quite easy to connect two sentences by a single character, so the Arabic writers used to write relatively long sentences without effective use of punctuation marks. This makes the Arabic language automatic syntax analysis (besides other processing) a very challenging task.
- **Dialects**: As many other languages there are quite a few dialects for the Arabic language.
- **Connected orthography:** Not only is the handwriting in Arabic but also the typewritten always connected (most of the characters). That means a more challenged OCR systems.

## 3. Arabic Language Challenges in digital age

Although there is some interest in the Arabic language from the NLP community for different reasons, but still this language has not been served enough.

- **No wide view and no clear roadmap:** Although many other languages with even much less native speakers has enjoyed a clear roadmap with support of many specialized associations and organizations, Arabic language until now has no clear and agreed upon roadmap. This means that the fragmented work here and there will leave serious gabs in the NLP applications that are needed along the road.
- **Insufficient Language Resources:** believe it or not, until this moment there is no Arabic machine readable dictionary. Bilingual dictionaries are still not mature or not available for the research community. There is dramatic shortage in the annotated text or speech corpora.
- **Inconvenient technology:** For the features of the Arabic language mentioned above, we believe that special tools and engines are needed to enhance the situation for the NLP research and the applications based on.

## 4. The need for an Association for the HLT for the Arabic Language

For the above reasons there is a need for a dedicated association to serve the area of Arabic HLT (Arabic NLP and Arabic speech processing). This association should belong to the Arab league to

be able to serve the entire Arab region and to be able to get support and fund from all the relevant associations and researchers.
The mandate of the suggested association:

A. **To make a situational analysis:** By probing the community needs from the area of the NLP**,** and continuously monitoring the advances and the gabs in applications and hence the technologies and language resources that might change along the time.

B. **To draw a roadmap for the Arabic language, for 10 years ahead [1-10]:** We mean by roadmap, a complete vision for the needed applications with its supporting technologies and language resources infrastructure along a timeframe. It is too late but it has to be done and revisited every three years to redirect the path if needed with the fast moving advances of technology and needs of the society.
    - **i.** To do that we need to cooperate with all the associations, workgroups, individuals who are working in the Arabic NLP. We also need to build on any work that has been done along this direction like the Arabic BLARK [1] as an excellent effort done by the NEMLAR project [1-5] as well as the MEDAR project as a current project under fp7. We also have to build on the experience of the others who led the work in NLP for other languages.

C. **Training and specialized diploma:** A serious training programs are needed in this area for both the linguists and the computer science researchers and developers. This should be done in cooperation between academia and industry with full cooperation with all players in the Arabic region. A diploma in NLP will be studied to elevate this kind of work needed in this field. We target to create a critical mass of well trained and prepared researchers, linguists and developers in this field. The study should be in both Arabic and English. Most of the linguists need to study in Arabic; while the engineers and the computer science graduate will prefer to study in English. My suggestion for the study is as follows:
    - i. **Introduction for NLP and HLT**: That will include:
        1. A brief of the applications to form a clear site for the area (see appendix A).
        2. A good course in probability and statistics.
        3. A review for the general algorithms used in this area namely:
            - a. The rule based algorithms

b. The statistical/corpus based and machine learning algorithms

4. A computer programming language.

ii. **Specialized track of study:** that could include:

1. A track for the linguistics for handling language resources; that could include:
   a. Speech data
   b. Text data
   c. OCR data, .. etc.
2. A track for the engineers/computer science; that should include techniques for computational linguistics. That has too many techniques to be studied in a single track. So after studying a general course, a multi-track could be followed up. This could be done be setting some courses and let the students (with the guidance of the staff) study selection of these courses according to their needs.

iii. **A project:** like the graduation projects of the engineers. This is extremely important to put the students on the track.

D. **Motivating the community:** Like in any serious research in many disciplines we need:

i. **Language Resources:** for training and **benchmarking**

ii. **Periodical competitions:** To discover the real potential of the community to solve the given problems.

E. **Cooperation with other relevant associations:** The continuous task of building the infrastructure for the NLP for any language is not a simple task that one entity can perform; so it is a must to cooperate with other similar entities and to help collectively in serving the community.

## Some notes on the suggested Organization:

It is not enough to create a center of excellence COE and secure some money for its initial activities. We need to be sure that there is an effective entity (NGO) that can really support the development of the Arabic HLT in the Arab region. This is why I suggest an association (NGO) not just a virtual COE, to have an easy mechanism to attract and secure donations from different places. So I suggest:

- An association that belongs to the Arab league.
  - To be able to collect donations from the entire region without any legal or logistic problems.

    - To convince the Arabic and the international donors that the activities and the results will serve the entire region.
- Formed from members from the entire Arabic region.
- With international key persons in this field (HLT) to form a real mass (scientifically) for any donors.
- Cooperate and integrate with the other existing associations and entities that serve the same purpose.

**Egypt and Arabic HLT:** Egypt has the critical mass in the human resources that is badly needed to establish any serious contribution to the NLP area.

- **Linguists:** They are many linguists in Egypt graduated from many colleges along the country. They are in 10,000's with relatively high standard linguistic information. But most of them are poor regarding the computer experience. They need some training and practice to be useful to the area of HLT. They will be of great help for developing **language resources** and to contribute in problems where the statistical solution is not enough.
- **Computer science graduates and Engineers:** They are also in 10,000's we need only increase those who are working in the HLT field. That needs some computational linguistic courses.
- **Industry:** Most of the HLT applications for the Arabic language are originated from Egypt. Name it; MT, search engine, OCR, TTS**,** ASR, spellchecker, computer aided language leaning,…etc are done by Egyptian industry. Sakhr, RDI, Caltech, IBM Egypt,…etc are well known companies among others for Arabic NLP based in Egypt. This means that there are high calibers from all disciplines that are needed in the area of NLP and speech processing.

**So we prefer to make Egypt the central hub for such an organization.**

## 5. Conclusions

In conclusion I suggest cooperating to lunch an association that is formed by many well known people in the area of HLT and organizations within the Arabic region. It will be better if this organization belongs to the Arab League. Egypt with its human resources will be an excellent central office for this organization. A clear role for this association is mentioned in some details in this paper.

# 6. References:

1. Report on Basic Language Resource Kit (BLARK) for Arabic, updated September 2006 In case of new information updates will be made September, December, March, June

2. Report on Survey on Arabic Language Resources and Tools in Mediterranean Countries In case of new information updates will be made September, December, March, June, www.NEMLAR.org

3. Yaseen, M., M. Atiyya, C. Bendahman, B. Maegaard, K. Choukri, N. Paulsson, S. Haamid , H.FersØe,S. Krauwer, M. Rashwan, B. Haddad, C. Mukbel, A. Mouradi, A. Al- Kufaishi, M. Shahin, A. Ragheb, Chenfour (2006): Building Annotated Written and Spoken Arabic LRs in NEMLAR project.In:LREC 2006 Proceedings, Genova.

4. Maegaard, B., L. Damsgaard JØrgensen, S.Krauwer, K. Choukri (2004): NEMLER: Arabic Language Resources and tools, In:K. Choukri and B. Maegaard (ed.): Proceedings of Arabic Language Resources and Tools Conference, P.42-54, Cairo.

5. Monachini, M., F. Bertagna, N. Calzolari, N. Underwood,C. Navarretta (2003): Towards a standard for the creation of Lexica , ELRA, Paris.
Nikkhou, M., K. Choukri (2004): Survey on the existing institutions and Language Resource using or developing Arabic, NEMLAR report,
www.nemlar.org

6. Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, H. Strik, C. Cucchinari (2002) A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, In : Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation), Las Plamas de Gran Canaria Spain.

7. Cieri, C., M. Maxwell, S. Strassel (2003): Core Linguistic Resources for the World's Languages. In: International Roadmap for Language Resources, Workshop Paris 2003, http://www.enabler-network.org/documents/workshop/Cieri-Maxwell-Strassel.Zip

8. Krauwer, Steven (1998): ELSNET and ELRA: A common past and common future. In: The ElRA Newsletter,Vol.3, n.2, Paris.

9. Andreas Eisele, Dorothea Ziegler-Eisele, "Towards a Road Map on Human Language Technology: Natural Language Processing", version 2 (March 2002), utrecht.elsnet.org/roadmap/docs/rm-**eisele**-v2.pdf

10. The ELSNET Roadmap for Human Language Technologies, Version 1.0, Nov. 30, 2002, DFKI GMBH, www.elsnet.org

11. M.Sc.; Mohamed Attia Mohamed Elaraby Ahmed; A large-Scale Computational Processor of The Arabic Morphology, and Applications, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

12. Ph.D. thesis; Mohamed Attia Mohamed Elaraby Ahmed; "Theory and Implementation of A Large-Scale Arabic Phonetic Transcriptor, And Applications", Cairo Univ., Faculty of Eng., Electronics & Comm. Department, 2005.

# Appendix A

## Some applications for the HLT

- Spell checkers
- Grammar checkers
- Document classification (content based classification)
- Document retrieval
- Automatic switchboards
- Text to speech (Talking pen)
- Natural language interface to information sources (e.g. traffic information)
- Dialogue systems via telephone (e.g. automatic reservation)
- Automatic dictation (e.g. in hospitals)
- Machine translation (restricted areas, generally)
- New enabling tools for disabled (speech technology, word prediction, etc)
- Summarization
- Stylometery
- TTS
- ASR
- Audio Indexer
- Information Retrieval
- OCR (offline, online)
- Many of the above technologies on mobile platform

# Table of Contents

page

## I. Large Corpora

## II. Machine Translation

6. **Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage**
Sameh Alansary[1,2] Magdy Nagi[1,3] Noha Adly[1,3]
*[1] Bibliotheca Alexandrina, Alexandria, Egypt.*
*[2] Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Alexandria, Egypt.*
*[3] Computers and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria, Egypt.*

7. **Machine Translation Using the Universal Networking Language (UNL)**
Sameh Alansary[1,2] , Magdy Nagi[1,3] , Noha Adly[1,3]
*[1]Bibliotheca Alexandrina, Alexandria, Egypt.*
*[2]Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University Alexandria, Egypt.*
*[3]Computer and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria, Egypt.*

8. **A Corpus Based Linguistic Tool for Machine Translation: A Case Study of '~INGs' in English and Its Equivalents in Tamil Through Grammatical Association Approach**
Dr. S. Kamakshi Devi
*ICFAI University, Regional Office, TN, East, Chennai*

9. **English to Arabic Hybrid Machine Translation System for Languages with Scarce Resources**
Eng. Ahmed Hatem, Prof. Dr. Amin Nassar
*Elect. & Comm. Eng. Dept., Faculty of Engineering, Cairo University, Giza, Egypt*

10. **Word Sense Disambiguation in Machine Translation Using Monolingual Corpus**
Ola M. Ali [1] , Mahmoud GadAlla[2] , Mohammad S. Abdelwahab[1]
*[1] Faculty of Computer and Information Sciences, Department of Scientific Computing, Ain Shams University*
*[2] Military Technical Collage, Computer science Department*

## III. Spoken Language Understanding

11. **دور المؤثرات السياقية في تقدير المدى الزمني للفونيم**
أحمد راغب أحمد
*مركز الإنسان للدراسات والتطوير –الرياض*

## IV. Speech Processing

## V. Natural Language Processing for Information Retrieval

## VI. Language Analysis and comprehension

24. **A Bilingual Approach for Arabic Paraphrases Acquisition: Preliminary Experiments**
Rania Al-Sabbagh[1] , Khaled Elghamry[2]
*[1] Faculty of Al-Alsun (Languages), Ain Shams University, EGYPT*
*[2] College of Liberal Arts and Science, University of Florida, USA*

25. **من مشكلات التحليل النصي للمحتوى العربي على شبكة الإنترنت**
*د/ سلوى السيد حماده ,عمرو جمعه*
*معهد بحوث الإلكترونيات*

## VII. Automatic Character Recognition

26. **Egyptian License Plate Recognition System Using DWT and Template Matching**
Ahmed R. EL-Barkouky[1] , Salwa H. El-Ramly[2] , Mohamed I. Hassan[3]
*[1]Engineering Physics and Mathematics Department, Faculty of Engineering, Ain Shams University*
*[2]Electronics and Communication Engineering Department, Faculty of Engineering, Ain Shams University*
*[3]Engineering Physics and Mathematics Department, Faculty of Engineering, Ain Shams University*

27. **Fractal Image Compression Applied on Document Images**
Salwa H. El-Ramly[1] , Ramy F. Taki El-Din[2]
*[1]Electronics and Communication Engineering Department, Faculty of Engineering, Ain-Shams University*
*[2]Engineering Physics and Mathematics Department, Faculty of Engineering, Ain-Shams University*

## *The Eighth Conference on Language Engineering*
## *Final Program*

### Wednesday 17 December 2008

9.00 - 10.00 Registration
10.00 - 10.30 Opening Session
10.30 - 11.00 Coffee Break
11.00 - 11.45 **Session 1** : **Invited Paper 1**:
Chairman : Prof. Dr. Ibrahim Farag

**معالجة المحتوى المعجمي الدلالي في المعجم العربي الحاسوبي - مقاربة لغوية حاسوبية**
أ.د. وفاء كامل - أ.د. محسن رشوان- د. عبد العاطي هواري
*كلية الآداب – جامعة القاهرة ,كلية الهندسة – جامعة القاهرة*

11.45 - 12.30 **Session 2** : **Invited Paper 2:**
Chairman : Prof. Dr. Adeeb Riad Ghonaimy

**Formal Semantics: Luxury or Necessity**
Allan Ramsay
*Manchester University, United Kingdom*

12.30 - 13.00 Break
13.00 - 13.45 **Session 3** : **Invited Paper 3 :**
Chairman : Prof. Dr. Adeeb Riad Ghonaimy

**Information Extraction and Arabic Named Entities**
Taghrid Anbar
*Faculty of Alson, Ain Shams University*

13.45 - 15.15 **Session 4** : Machine Translation
Chairman : Prof. Dr. Mohamed Fahmy Tolba

**1. Designing and Implementing Arabic Sign Language (ArSL) for Deaf Peoples**
Hassanin M. Al-Barhamtoshy, Sami M. Halawani and Sakher F. Ghanem
*King Abdel Aziz University, Jeddah, Saudi Arabia*

**2. Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage**
Sameh Alansary[1,2] Magdy Nagi[1,3] Noha Adly[1,3]
*[1] Bibliotheca Alexandrina, Alexandria, Egypt.*
*[2] Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Alexandria, Egypt.*
*[3] Computer and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria, Egypt.*

**3. Machine Translation Using the Universal Networking Language (UNL)**
Sameh Alansary[1,2] , Magdy Nagi[1,3] , Noha Adly[1,3]
*[1]Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.*
*[2]Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, El Shatby, Alexandria, Egypt.*
*3Computer and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria, Egypt.*

15.15 16.00 Lunch

16.00 - 17.00 **<u>Session 5</u>** : **Room A** : Machine Translation
Chairman Prof. Dr. Taghrid Anbar

**1.A Corpus Based Linguistic Tool for Machine Translation: A Case Study of '~INGs' in English and Its Equivalents in Tamil Through Grammatical Association Approach**
Dr. S. Kamakshi Devi
*ICFAI University, Regional Office, TN, East, Chennai*

**2.English to Arabic Hybrid Machine Translation System for Languages with Scarce Resources**
Eng. Ahmed Hatem, Prof. Dr. Amin Nassar
*Elect. & Comm. Eng. Dept., Faculty of Engineering, Cairo University, Giza, Egypt*

**3.Word Sense Disambiguation in Machine Translation Using Monolingual Corpus**
Ola M. Ali [1] , Mahmoud GadAlla[2] , Mohammad S. Abdelwahab[1]
*[1] Faculty of Computer and Information Sciences, Department of Scientific Computing, Ain Shams University*
*[2] Military Technical Collage, Computer science Department*

16.00 17.00 **<u>Session 6:</u> Room B** : Large Corpora
Chairman: Prof. Dr. Mohamed Younis Elhamalwy

**1.Bilingual Dictionaries: From Theory to Computerization**
Sherif A. Sattar Okasha
*Advanced Multilingual Systems Co., Cairo, Egypt*

2. **المعالجة الحاسوبية للتطور الدلالي للمشتقات في العربية**
د. مدحت يوسف السبع
*كلية دار العلوم / جامعة القاهرة*
*والأستاذ المساعد بجامعة الإمام محمد بن سعود الإسلامية الرياض*

## <u>Thursday 18 December 2008</u>

9.00 - 11.00 **<u>Session 9</u>** : **Room A** : Tutorial
Title: Text Mining: Methodologies and Approaches
Speaker: Prof. Dr. Ibrahim F. Imam

10.00 - 11.00 **<u>Session 8</u>** : **Room B** : Text Processing
Chairman : Prof. Dr. Hany Kamal Mahdy

**1. Egyptian License Plate Recognition System Using DWT and Template Matching**
Ahmed R. EL-Barkouky[1] , Salwa H. El-Ramly[2] , Mohamed I. Hassan[3]
*[1]Engineering Physics and Mathematics Department, Faculty of Engineering, Ain Shams University*
*[2]Electronics and Communication Engineering Department, Faculty of Engineering, Ain Shams University*
*[3]Engineering Physics and Mathematics Department, Faculty of Engineering, Ain Shams University*

**2. Fractal Image Compression Applied on Document Images**
Salwa H. El-Ramly[1] , Ramy F. Taki El-Din[2]
*[1]Electronics and Communication Engineering Department, Faculty of Engineering, Ain-*

*Shams University*
[2]*Engineering Physics and Mathematics Department, Faculty of Engineering, Ain-Shams University*

11.00 - 11.30 Coffee Break

11.30 - 12.15 **<u>Session 10</u> : Room A : Invited Paper 4**
Chairman : Prof. Dr. Ali Ali Fahmy
**Statistical Machine Translation: Current Trends**
Ahmed Rafea
*Computer Science and Engineering Department*
*American University in Cairo*

12.15 - 13.45 **<u>Session 12</u> : Room A : Language Analysis and Comprehension**
Chairman : Prof. Dr. Ahmed Rafea

**1. Learning Generator: Neuro-Linguistic Programming and Learning Styles in English Text Books**
Dr. Mrs. Eva Zanuy Pascual
*PhD on Linguistics*
*Universidad Nacional de Educación a Distancia,Spain*

**2. A Bilingual Approach for Arabic Paraphrases Acquisition: Preliminary Experiments**
Rania Al-Sabbagh[1] , Khaled Elghamry[2]
[1] *Faculty of Al-Alsun (Languages), Ain Shams University, EGYPT*
[2] *College of Liberal Arts and Science, University of Florida, USA*

**3. من مشكلات التحليل النصي للمحتوى العربي على شبكة الإنترنت**
د/ سلوى السيد حماده ,عمرو جمعه
*معهد بحوث الإلكترونيات*

12.15 - 13.45 **<u>Session 13:</u> Room B: Spoken Language Understanding**
Chairman : Prof. Dr. Walid Fakhr

**1. دور المؤثرات السياقية في تقدير المدى الزمني للفونيم**
أحمد راغب أحمد
*مركز الإنسان للدراسات والتطوير –الرياض*

**2. Towards a Prototype Intonational Transcription System for Egyptian Arabic: Testing the Local f0 Contour Properties of Intonational Pitch Accents in Spontaneous Speech.**
Dr Sam Hellmuth
*Department of Language & Linguistic Science*
*University of York, Heslington, York, United Kingdom*

**3.Spoken Term Detection For Arabic Educational Media**
M. Hesham[1] and M. F. Abu-EL-Yazeed[2]
[1] *Professor at Engineering Math. & Physics Dept., Cairo University*
[2] *Professor at Electronics Engineering Dept., Cairo University*

13.45 - 14.30 Lunch

14.30 - 16.00 **<u>Session14:</u> Room A : Natural Language Processing for Information Retrieval**
Chairman : Prof. Dr. Mohamed Zaki Abdel Mageed

**1. A Statistical Method for Adding Case Ending Diacritics for Arabic Text**

Hitham M. Abo Bakr [1] ,    Khaled Shaalan[2]   ,   Ibrahim Ziedan[1]
*[1] Computer & System Dept., Zagazig University, Zagazig, Egypt*
*[2] The Institute of Informatics, The British University in Dubai, United Arab Emirates*

**2. Experiments for Automatic Arabic Diacritization**
Mohsen A. A. Rashwan[1, 2], Mohammad Al-Badrashiny[1], Mohamed Attia[1]
*[1] The Engineering Company for the Development of Computer Systems; RDI, Egypt*
*[2] Dept. of Electronics and Electrical communications, Faculty of Engineering, Cairo University, Egypt*

**3. Diacritization and Transliteration of Proper Nouns from Arabic to English**
Hamdy S. Mubarak,  Mohamed Al Sharqawy, Esraa Al Masry
*Arabic NLP Dept, Sakhr Software, Cairo, Egypt*

**4. An Empirical Analysis of Clustering Techniques for Arabic Documents**
Doaa Farag  ,  Ibrahim F. Imam
*Department of Computer Science, College of Computing and Information Technology,*
*Arab Academy for Science, Technology and Maritime Transport*
*Masaken Sheraton, Cairo, Egypt*

14.30 - 16.00 **Session 15 : Room B : Speech Processing**
Chairman : Prof. Dr. Afaf Abelfattah

**1. Wavelet Packets Best Tree 4 Points Encoded (BTE) Features**
Amr M. Gody
*Fayoum University, Fayoum, Egypt*
*Department of Electrical Engineering*

**2. Voiced/Unvoiced and Silent Classification Using HMM Classifier based on Wavelet Packets BTE features**
Amr M. Gody
*Fayoum University, Fayoum, Egypt*
*Department of Electrical Engineering*

**3. Proposed Biometric Key for DES Scheme Applying Formant's Arc-tangents**
Zaki. T. Fayed
*Faculty of Computers & Information Sciences, Ain Shams University*
*Department of Computer Science, Cairo, Egypt*

16.00 - 17.00 **Panel discussion**
Chairman: Mohsen A. A. Rashwan
Title:  Towards A Road Map of Human Language Technology Development For the Arabic Language

17:00 - 17:30 Closing session

# Wavelet Packets Best Tree 4 Points Encoded (BTE) Features

Amr M. Gody[1]
Fayoum University

## Abstract

The presented research aimed to introduce newly designed features for speech signal. The newly developed features are designed to normalize the dynamic structure of best tree decomposition of wavelet packets. The 4 points encoded vector is full of information just like the original best tree's structure. It is a loss less encoding system that grantees 100% reconstruction of the original best tree. The encoding process for BTE features vector is developed such as to minimize the distance based on frequency adjacency. The implied scoring system makes BTE suitable for recognition problems that's because the scoring system consider the adjacency in both frequency bands and frequency level at wavelet packet tree.

## 1. Introduction

It is known that human speech is decomposed of short time duration's units called phonemes. Each phoneme contributes with specific piece of information. We can assume it as the characters that construct the whole word in any written language. Information in each phoneme is encoded into the frequency domain. Simply the information is a pattern of frequency components [1]. Features are extracted from the speech signal to best represent such information.

It is believed that human hearing system is the best recognition system. By trying to simulate human hearing system, good practical results may be achieved. Speech signal is processed in this research in such a manner that low frequency components have more weights than high frequency components [2]. The human ear responds to speech in a manner such as that as indicated by Mel scale in figure 1. This curve explains a very important fact. Human ears cannot differentiate between different sounds in high frequency scale while it can do this in low frequency scale. Mel scale is a scale that reflects what human can hear. As shown by figure 1, a change in frequency from 4000(HZ) to 8000 (HZ) makes only 1000 (Mel) change in Mel scale. This is not the case in the low frequency range which starts at 0(HZ) and ends by 1000 (HZ). In this low frequency range it appears that 1000(Hz)'s change is equivalent to 1000(Mel) change in Mel scale. This explains that the human hearing is very sensitive for frequency variation in low range while it is not the case in high range.

Wavelets are short duration waveforms that can express any function by scaling and shifting of certain mother signal that is called mother wavelet [5]. Wavelet algorithm is acting as filter banks on the input signal. The output of the filter banks are the wavelet signal's amplitudes.

[1] Department of Electrical Engineering, Email: amg00@fayoum.edu.eg

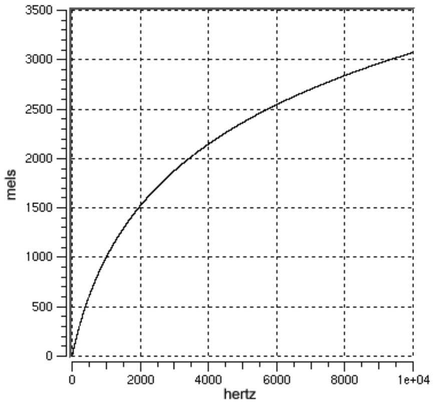


**Figure 1: Mel scale curve that models the human hearing response to different frequencies [3].**

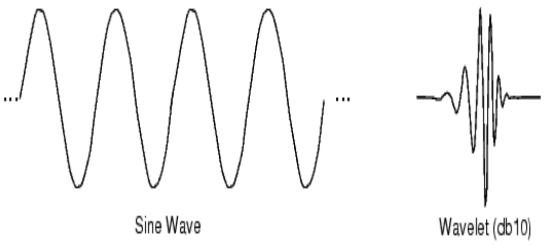


**Figure 2: Sine wave is used for Fourier representation of the signal while wavelet function is used in wavelet representation for Daubechies 10 pointes filter. Sine wav is infinite in time but finite in frequency domain while wavelet is finite in both time and frequency domains [5].**

Figure 2 indicates a very important property of wavelet function. Wavelet function is finite in time. It is also finite in frequency [4]. This is not the case of "Sine" basis functions (harmonic functions) used for Fourier analysis. All derived wavelets are orthogonal. This makes each wavelet acts as an identifier of the signal in a certain band. Figure 3 gives a brief comparison between different possible spaces to express certain function [5].

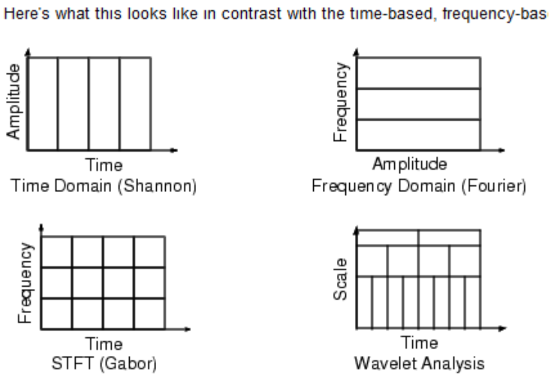


**Figure 3: Comparison between different signal spaces [5].**

Wavelet packets are extension to wavelet transform. It includes the high frequency parts in the analysis for more signal resolution of the frequency spectrum as shown in figure 4.

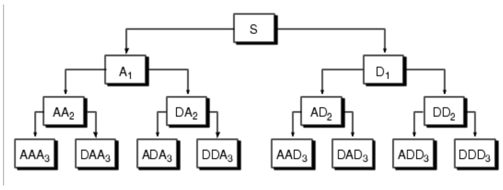


**Figure 4: Signal decomposition using wavelet packets [5].**

To simplify the subject, let us discuss Fourier series as a signal representation tool.

$$\boldsymbol{f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right)} \qquad (1)$$

Equation 1 indicates the Fourier series representation of function f(x). By the same approach, f(x) may be expressed using wavelet packets as in equation 2.

$$\boldsymbol{f(x) = \sum_j \sum_{n=0}^{2^{j-1}} b_{jn} W_{j,n}(x)} \qquad (2)$$

"b" is wavelet coefficients and "W" is wavelet packet. Let us start with the two filters of length 2N, where h(n) and g(n), corresponding to the wavelet filters.

$$W_{2n}(x) = \sqrt{2} \sum_{k=0}^{2N-1} h(k) W_n\,(2x-k) \qquad (3)$$
$$W_{2n+1}(x) = \sqrt{2} \sum_{k=0}^{2N-1} g(k) W_n\,(2x-k) \qquad (4)$$

g(k) and h(k) are filter banks. Where:

$W_o(x) = \emptyset(x)$ is called the scaling function.

$W_1(x) = \psi(x)$ is called wavelet function.

Where:

$$\boldsymbol{W_{jnk}(x) = w_n(2^{-j}x - k)} \qquad (5)$$

K is not a dynamic parameter after the decomposition of the signal, rather it is a constant value for each wavelet packet W. This makes it much better to abstract (5) as :

$$W_{j,n} = w_n(2^{-j}x - k) \quad k \epsilon Z \qquad (6)$$

Hence:

$$W_{0,0}(x) = \Phi(x-k) \qquad (7)$$
$$W_{1,1}(x) = \Psi(\frac{x}{2} - k) \qquad (8)$$

The idea is explained by figure 5. Scaling "ф" and wavelet "Ψ" functions are used to generate W functions that cover all the frequency-scale space. The parameter k is used to indicate the time location of certain W function. K is chosen to best fit the original function to be expressed by wavelet packets while the scaling and wavelet functions are designed such that all W functions be orthogonal.

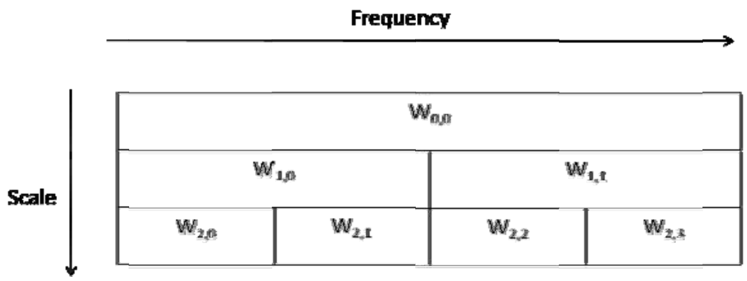


**Figure 5: Frequency-Scale space for wavelet packets.**

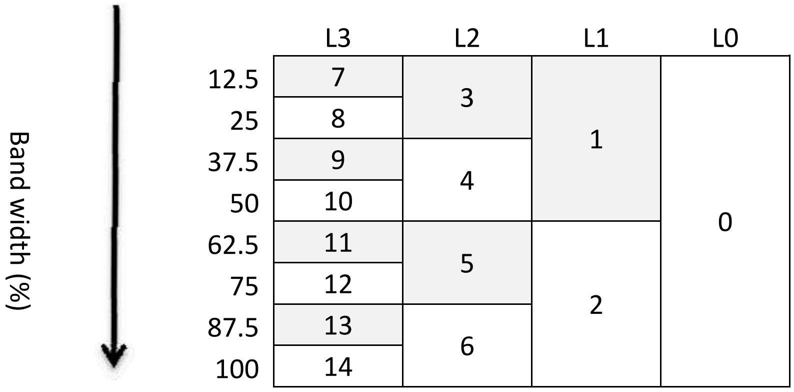
Many researchers deal with the best way to optimize the full binary tree in a way to best describe the contained information [6]. Different entropy functions may be used in such optimization [7,8].

The objective of this paper is to introduce new features for speech signal. Features are developed from the wavelet packets best tree decomposition of speech signal. This research aims to explain the proposed features in details. Also it targets to introduce the benefits of using the proposed features in speech recognition problems.

## 2. Feature extraction

In this section the process of feature extraction will be explained. Best Tree 4 point Encoded features (BTE) will be explained now. Wavelet packets process is very similar to filter banks. Both of them are filter banks in nature. The wavelet packets method is a generalization of wavelet decomposition that offers a richer signal analysis. Wavelet packet atoms are waveforms indexed by three naturally interpreted parameters: position, scale (as in wavelet decomposition), and frequency. For a given orthogonal wavelet function, we generate a library of bases called wavelet packet bases. Each of these bases offers a particular way of coding signals, preserving global energy, and reconstructing exact features. The wavelet packets can be used for numerous expansions of a given signal. We then select the most suitable decomposition of a given signal with respect to an entropy-based criterion [9].

The first step in BTE is to align the neighboring bands. This is very important for a good scoring process. Scoring process tries to score adjacent bands in such a way that is minimizing the distance. For our case of best tree by Matlab, adjacent bands are indexed not in sequence.



| | L3 | L2 | L1 | L0 |
|---|---|---|---|---|
| 12.5 | 7 | 3 | 1 | 0 |
| 25 | 8 | | | |
| 37.5 | 9 | 4 | | |
| 50 | 10 | | | |
| 62.5 | 11 | 5 | 2 | |
| 75 | 12 | | | |
| 87.5 | 13 | 6 | | |
| 100 | 14 | | | |

**Figure 6 : Wavelet packet tree analysis chart to figure out adjacent bands.**

The objective is to remap node indices such that adjacent node indices lay in adjacent frequency bands. To explain this subject consider the following table that represents the indices in a typical wavelet packet tree for 4-levels decomposition. Figure 6 represents band indexes in Matlab wavelet packets for 3 levels decomposition. Node indices are written inside the boxes that represent the nodes in the wavelet tree decomposition. As shown in figure 6 that node 7 and node 6 are too far in frequency while they are subsequent nodes as wavelet packets indexing system. This problem needs to be altered such that adjacent frequency bands are listed as contiguous numbers. This way we will ensure that indexing system reflects frequency scale. This property may be used in the scoring system. Information in figure 6 is tabulated in table 1 to make it simple to figure out adjacent bands. Traversing tree as Left → Right → Center will be very logical to make good criteria for adjacency. Figure 7 explains the new indexing system.

Now we are ready to apply the best tree algorithm to optimize the full binary tree shown in figure 7. The optimization minimizes the number of tree nodes such that it best fit the information included in the speech signal. The entropy is used in the optimization algorithm.

Now we can apply the encoding by considering clusters of 7 bands. Each cluster will be encoded in 7 bits such that each bit is associated to a certain band. Figure 8 explains the clusters.

**Table 1 : Bandwidth distribution over wavelet packet decomposition bands.**

| Filter bank's Upper Limit with respect to total bandwidth (%) | Filter Bank's Node- index according to wavelet packet indexing system |
|---:|---:|
| 100 | 0 |
| 50 | 1 |
| 100 | 2 |
| 25 | 3 |
| 50 | 4 |
| 75 | 5 |
| 100 | 6 |
| 12.5 | 7 |
| 25 | 8 |
| 37.5 | 9 |
| 50 | 10 |
| 62.5 | 11 |
| 75 | 12 |
| 87.5 | 13 |
| 100 | 14 |

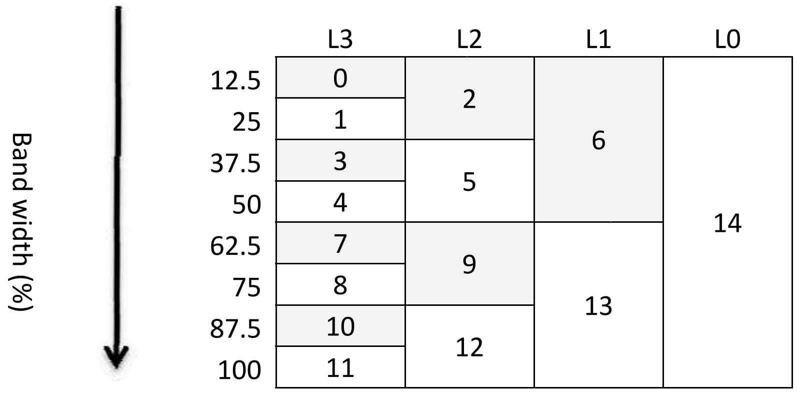


**Figure 7 : Proposed indexing to solve the adjacency problem due to wavelet packet's indexing system.**

In figure 8, clusters are surrounded by bold black boxes. Bits are ordered as in figure 8.The least Significant Bit (LSB) is assigned to band number 0 and the Most Significant Bit (MSB) is assigned to band number 6.

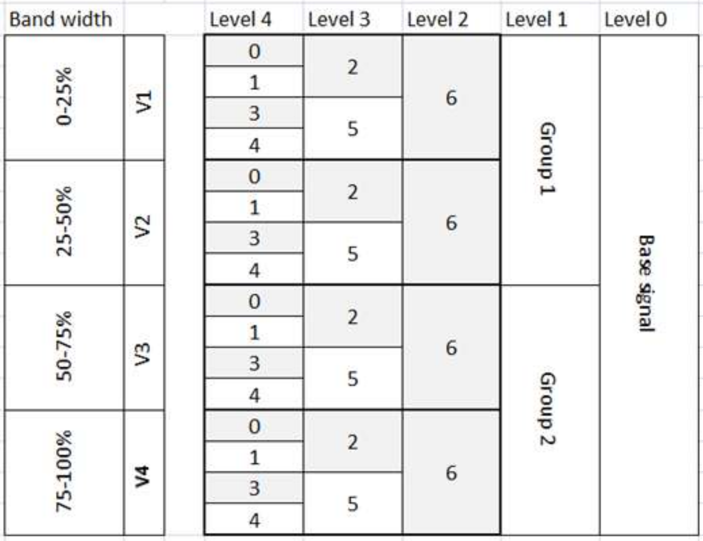


**Figure 8 : Clustering chart to explain the 4 points encoding algorithm.**

As shown in figure 8, each cluster will be encoded by 7 bit valued number. The number is formed such that it reflects the tree structure within the cluster. Trees that cover the same bands will be almost adjacent trees. This property will be utilized in the scoring system. By considering all clusters, a vector of 4 components will be formed. Each vector's component represents a certain cluster. And each cluster covers a certain area in the total bandwidth. This is the 4 point encoded method that construct BTE features vector.

Figure 9 introduces a simple example to explain features encoding for a frame of speech signal. Circles in figure 9 represent leave nodes in the best tree decomposition.

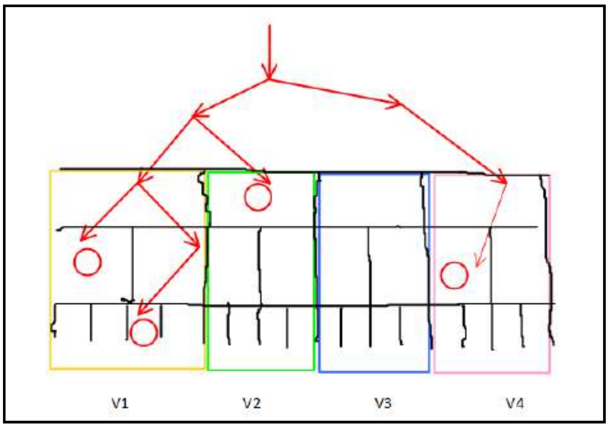


**Figure 9: Best tree 4 point encoding example.**

The indicated tree structure in figure 9 will be encoded into features vector of 4 elements as shown in table 2.

**Table 2 : Best tree 4 point encoding evaluation.**

| Element | Binary Value | Decimal value | Frequency Band |
|---|---|---|---|
| V1 | 0001100 | 12 | 0 - 25 % |
| V2 | 1000000 | 64 | 25% - 50% |
| V3 | 0000000 | 0 | 50%-75% |
| V4 | 0000100 | 4 | 75%- 100% |

Features vector for this example speech frame will be:

$$\boldsymbol{F} = \begin{pmatrix} 12 \\ 64 \\ 0 \\ 4 \end{pmatrix} \qquad (9)$$

Matlab is used to implement BTE features extraction. The following code snippet is the core part of Matlab function to implement BTE features extraction.

```
function [res] = BTE (frame, depth)
    nbIn = nargin;
    nbout = nargout;
    if nbIn < 1 ,    error('Not enough input arguments.');
    elseif nbIn == 1,     level = 4;
    elseif nbIn == 2,     level = depth;
    end;
    if nbout < 1 , error('Not enough output arguments.'); end;
    t = wpdec(frame,level,'db4','shannon');
    u = leaves (t);
    bt =  besttree(t);
    v = leaves (bt);
   % res = score(v,0,4)/1000;
```

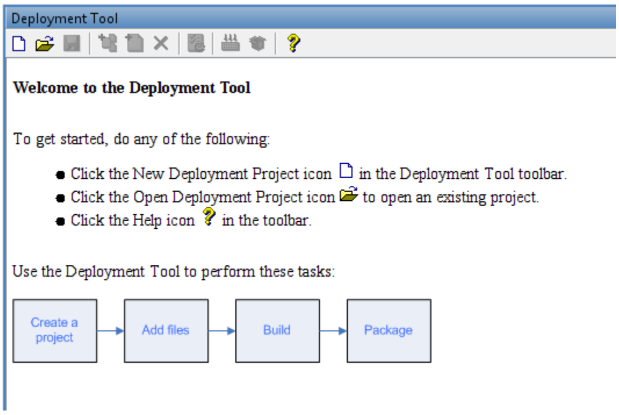
```
    res = box4encoder(v);

end
```

The function "box4encoder" in the above code snippet is responsible for encoding Best tree as indicated in table 2. Matlab functions needed for this research are all packaged into a Class Library[2]. This step makes it easy to call Matlab functions from within the C# development environment[3] that is being used as Business and Cue Logic[4] "BCL". The following Matlab command is used to invoke the packaging tool in Matlab:

```
                                 Deploytool
```

Figure 10 explains the deploy tool utility that is available in Matlab 7.5. This is a very useful tool that enables calling for all Matlab functionalities from other more advanced software development environments.



Deployment Tool

**Welcome to the Deployment Tool**

To get started, do any of the following:

- Click the New Deployment Project icon 🗋 in the Deployment Tool toolbar.
- Click the Open Deployment Project icon 📂 to open an existing project.
- Click the Help icon ? in the toolbar.

Use the Deployment Tool to perform these tasks:

**Figure 10: Deployment tool for packaging Matlab functions into Class Library suitable for calling from C# development environment.[5]**

The Matlab function called "wav2BTE" is developed in Matlab. Part of the code of "wav2BTE" is indicated in the following cod snippet.

```
    [y fs] = wavread(file);
    S = 20e-3*SamplingRate;
    F = framing(y,S,0,0);
    A = BTE (F(:,1));
    for i = 2:n
      A = [A  BTE (F(:,i))];
    end;
    version = uint32([3 1]);
    Frame = uint32(20);
    wpdepth =uint32( 4);
    fid = fopen(outfile, 'wb');
    fwrite(fid,version,'int32');
    fwrite(fid,Frame,'int32');
    fwrite(fid,wpdepth,'int32');
    fwrite(fid,uint32(fs),'int32');
    fwrite(fid,size(A),'int32');
    fwrite(fid,2,'int32');
```

[2] Class library is the name of the entity used by Microsoft in the dot net framework to package functions and procedure. By packaging all needed functions int class library, we can reuse the functions from any dot net programming language for further use.

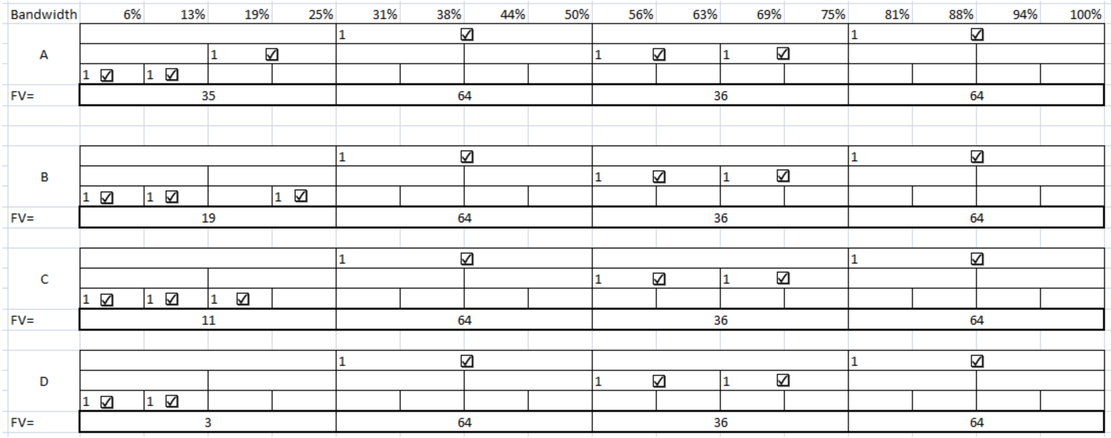
[3] Dot net programming language by Microsoft Corporation.

[4] Business and Cue Logic "BCL" is a name for all program snippets that is being written to control program sequencing. This includes loops, conditions, input and outputs.

```
    fwrite(fid, uint16(A),'int16');
    status = fclose(fid);
```

## 3. Testing BTE scoring system

This section is dealing with testing the scoring system of BTE features. As indicated before the scoring system is designed as to minimize the distance based on frequency coverage. Signals that have similar frequency spectrum are close and signals that have different frequency component are far.  Figure 11 introduces the score of 4 BTE feature vectors. Check marks mark the frequency bands being covered by leaf's nodes. FV is the abbreviation for Feature Vector. As it is shown in the figure, Vectors A, B, C and D are almost identical vectors. They just differed in 19% and 25% Bandwidth components of wavelet packets. The scoring makes B and C are too close while A and D are too far. This is logical as vector A has no resolution in level 4 while B and C have adjacent components in level 4. Also D has no component at all in 19% and 25%.  Also C is equally distant from D and B. This is also logical as the 19% component at level 4 for vector C is in the middle between the 13% component at level 4 for vector D and the 25% component at level 4 for vector B.



| Bandwidth | 6% | 13% | 19% | 25% | 31% | 38% | 44% | 50% | 56% | 63% | 69% | 75% | 81% | 88% | 94% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | | ☑ | | | | | | 1 | | ☑ | |
| A | | | 1 | ☑ | | | | | 1 | ☑ | 1 | ☑ | | | | |
| | 1 ☑ | 1 ☑ | | | | | | | | | | | | | | |
| FV= | 35 | | | | 64 | | | | 36 | | | | 64 | | | |
| | | | | | 1 | | ☑ | | | | | | 1 | | ☑ | |
| B | | | | | | | | | 1 | ☑ | 1 | ☑ | | | | |
| | 1 ☑ | 1 ☑ | | 1 ☑ | | | | | | | | | | | | |
| FV= | 19 | | | | 64 | | | | 36 | | | | 64 | | | |
| | | | | | 1 | | ☑ | | | | | | 1 | | ☑ | |
| C | | | | | | | | | 1 | ☑ | 1 | ☑ | | | | |
| | 1 ☑ | 1 ☑ | 1 ☑ | | | | | | | | | | | | | |
| FV= | 11 | | | | 64 | | | | 36 | | | | 64 | | | |
| | | | | | 1 | | ☑ | | | | | | 1 | | ☑ | |
| D | | | | | | | | | 1 | ☑ | 1 | ☑ | | | | |
| | 1 ☑ | 1 ☑ | | | | | | | | | | | | | | |
| FV= | 3 | | | | 64 | | | | 36 | | | | 64 | | | |

**Figure 11: Scoring sheet that explains the scoring of 4 different feature vectors.**

The above discussion explains that the scoring hold information that we can rely upon in the recognition system.  The above discussion is summarized in figure 12. As it is indicated in figure 12, vector C are in the middle path between B and D. Vector A and D are at the far limits.

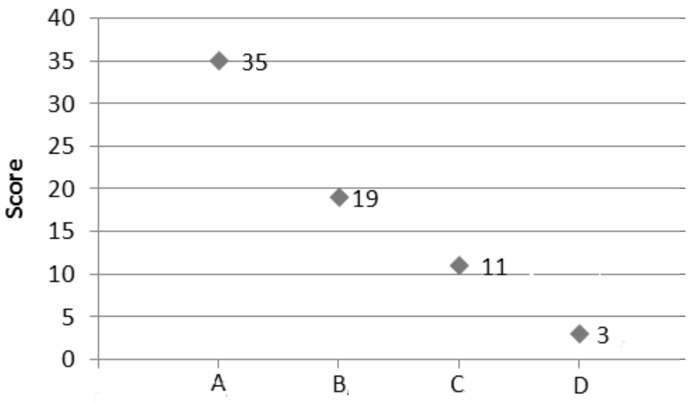


**Figure 12: Summary results of scoring system**

## 4. Conclusions

Wavelet packets make a similar processing on speech signal as the Filter banks method. It is much smarter than filter banks in that the number of filters is adapted by considering signal entropy to find the best tree. The problem of having dynamic size feature vectors is solved by considering the 4 points encoding algorithm. The proposed encoding system grantees minimizing the distance between feature vectors based on adjacency in frequency domain. This adjacency based on frequency domain of feature vectors distance calculation makes (BTE) features to be highly promising in speech recognition systems.

## 5. References

[1] Amr M. Gody, "Natural Hearing Model Based On Dyadic Wavelet", The Third Conference on Language Engineering CLE'2002, Page(s): 37-43,October 2002

[2] Alessia Paglialonga, "Speech Processing for Cochlear Implants with the DiscreteWavelet Transform: Feasibility Study and Performance Evaluation", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006

[3] Mel scale, http://en.wikipedia.org/wiki/Mel_scale

[4] Gilbert Strang, "Wavelets and filter banks", Wellesley-Cambridge Press, ISBN: 0-9614088-7-1, pp. 37-86, ©1996.

[5] MatLab,http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/ch06_a11.html.

[6] Coifman, R.R.; M.V. Wickerhauser (1992), "Entropy-based algorithms for best basis selection," IEEE Trans. on Inf. Theory, vol. 38, 2, pp. 713-718.

[7] Hai Jiang, Meng Joo Er and Yang Gao ," Feature Extraction Using Wavelet Packets Strategy", Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii USA, December 2003

[8] http://en.wikipedia.org/wiki/Information_entropy.

# Voiced/Unvoiced and Silent Classification Using HMM Classifier based on Wavelet Packets BTE features

Amr M. Gody[1]
Fayoum University

## Abstract

Wavelet Packets Best Tree Encoded (BTE) features is used here as base features for HMM classifier. The research aimed to introduce the newly designed features that are discussed in [1]. The considered problem is Voiced, Unvoiced and Silent classification. Comparison to the 19 filter banks features is provided. Although it is simple and straight forward, BTE makes comparable results to the 19 elements features vector based on filters bank. A very accurate hand labeled database called SCRIBE is used. Voiced sounds are recognized in 81% success rate. Silent periods are detected in 84.5% success rate. The unvoiced sounds are not recognized using the proposed features. It gives a 5.5% success rate. This low rate of unvoiced detection affects the overall performance. The overall performance of 64.5% is achieved. This overall performance is expected to be dramatically changed in case of adding some unvoiced attributes to BTE.

## 1. Introduction

Using good features is the key of accurate speech recognizer. Recognizer's success depends on three main factors. The first factor is the database used in the training phase. The second factor is the features used to train the model. The third factor is the mathematical model used to recognize the different classes in the speech signal.

BTE features are discussed in [1]. BTE inherits some human attributes by considering the human hearing mechanism in processing the received speech. Received speech's stream is classified into logarithmic bands before it is being processed by human brain [1]. This human nature is described by Mel scale as shown in figure 1.

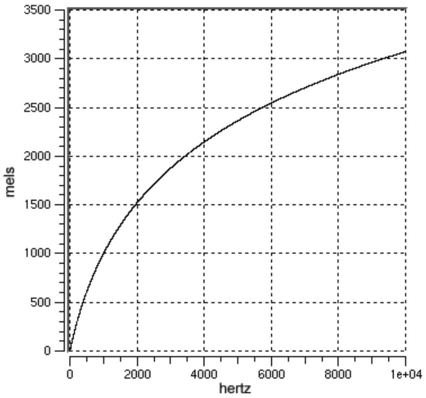


**Figure 1: Mel scale curve that models the human hearing response to different frequencies [2].**

[1] Department of Electrical Engineering, Email: amg00@fayoum.edu.eg

Mel frequency reflects what human can discriminate. It is a scale that reflects what human can hear. As shown in figure 1, from 4000(HZ) to 8000 (HZ) only 1000 (Mel) change while from 0(HZ) to 1000 (HZ) a 1000(Mel) change is appeared. The curve in low frequency till 1000(HZ) indicates that the human ear can be highly discriminative. This property starts to be degraded toward the higher frequencies. As shown in figure 1, change after 4000(Hz) in frequency tends to make almost very low change in Mel scale. This phenomenon indicates that human ear is much sensitive to low frequencies than to high frequencies. It is expected that most of the information contained into speech is located in the low frequency area of the total bandwidth of speech signal. Recognizers based on Filter banks tries to satisfy this logarithmic relation that was explained by figure 1. It is not wise to handle the frequency band in a linear manner while it is not like that in human hearing mechanism.

The objective of this paper is to test BTE through a comparative study. This research is a preliminary work to introduce Wavelet Packets Best Tree 4 point Encoded (BTE) features. The database is selected to be very accurate hand labeled database. SCRIB[2] database is selected. SCRIBE consists of a mixture of read speech and spontaneous speech. The read speech material consists of sentences selected from a set of 200 "phonetically rich" sentences and 460 "phonetically compact" sentences and a two-minute continuous passage. Hidden Markov Model (HMM) is chosen as the mathematical model of the speech recognizer. This model is chosen for its good reputation in speech recognition domain. The model is implemented using The Hidden Markov Model Toolkit HTK[3]. Signal processing, system evaluation and results preparation are calculated using Speech Filling System SFS[4]. All programming and logic are made using Microsoft C Sharp[5] (C#).

Voiced, Unvoiced and silent are three main classes in any spoken language. Almost all phonetics is either Voiced or unvoiced. Silent periods are those periods where no speech exists. The detection of unvoiced speech in the presence of additive background noise is complicated by the fact that unvoiced speech is very similar to white noise [3]. The problem of detecting unvoiced appears in this preliminary research. It is almost confused in equal proportion between Voiced and silent.

## 2. Framework

In this section the framework of this research will be fully explained. The system has many actors/ resources.

1. SFS platform.
2. HTK platform.
3. Batch and Cue Logic platform (BCL). Microsoft C# is used to implement BCL platform.
4. Hand labeled Database (SCRIBE).
5. Matlab platform.

[2] SCRIBE database: www.phon.ucl.ac.uk/resource/scribe

[3] Hidden Markov Model Toolkit HTK: http://htk.eng.cam.ac.uk/

[4] Speech Filling System SFS: http://www.phon.ucl.ac.uk/resource/sfs/

[5] Microsoft C# is one of the programming languages by Microsoft Corporation. C# implements Object Oriented Programming (OOP). It bears both simplicity and advanced programming technique. It is considered the number one language nowadays. For more information go to http://en.wikipedia.org/wiki/C_Sharp_(programming_language)

**Step 1 [Creating SFS files]**

BCL is used to import all sound files provided by the SCRIBE into SFS formatted files. Also BCL is used to import all corresponding label files into the same SFS files. After this step each SFS file contains waveform item and annotation item.

**Step 2 [Mapping phonetic labels into Voiced, Unvoiced and Silent labels]**

A new map file is constructed. The map file contains the map for each phonetic symbol into one of the three classes {VOI, UNV and SIL}. Symbols are VOI for Voiced sound, UNV for Unvoiced sound and SIL for Silent or pauses. SFS is used to apply the map to SFS file. Then BCL is used to apply the SFS map to all SFS files.

**Step 3 [Preparing two different groups for two parallel experiments]**

SFS files are cloned into two sets. This is to use each SET into different experiment. SET_A will be used in VOC19 feature experiment and SET_B will be used in BTE features experiment.

**Step 4 [Apply feature extraction function on all SFS files]**

SET_A: SFS is used to apply VOC19 to the available samples.

SET_B: Matlab is used to extract BTE features. BCL is used to apply the Matlab function to all available samples. Then finally SFS is used to import all feature vectors into the SFS files.

**Step 5 [HMM preparation]**

SET_A: Three HMM models are prepared. Each model is 3 states. HTK is used to initialize each model based on training samples, label files and feature vector files.

SET_B: The same process as in SET_A is followed.

**Step 6 [Training HMM models]**

In both sets, HTK is used to train the available HMM models. The training depends on the feature vector files, label files and selected training files list. Training continues till a convergence in log probability happened for each model.

**Step 7 [Testing HMM models against test files]**

Test files are some SFS files that were never being used in the training phase. HTK is used to test HMM models against the selected test files in both groups. HTK generates label file for each test file. Each label file is imported using BCL to the corresponding SFS file. This will cause that each SFS test file contain two annotations. One is the reference annotation generated in step 2 and the other one is the test annotation.

**Step 8 [Evaluation]**

Each SFS test file will be analyzed using SFS. A confusion matrix is generated for each SFS test file. Results are registered for both groups. Then results are tabulated and graphs are obtained to view and compare the results.

## 3. Database

SCRIBE database is used in this research. It is multi-speakers database. Each file is phonetically transcribed and segmented.

The following commands invoke SFS to add Sound file and the corresponding annotation file into SFS formatted file. This SFS file will act as a container for all items {Speech file, features data and annotation data}

```
Slink -i1.01 -f SamplingRate  Sound_file  sfsfile
Anload -S Annotation_file sfsfile
```

To apply the same command to all the available samples, BCL is used. The following is the program written to do the function. All speech files in SCRIB are listed into a string array called "files". Speech files in SCRIBE have an extension called "PES". Speech is sampled at 20000 (HZ). All annotation files in SCRIBE have an extension called "PEA". The Annotation file is located in the same folder as the corresponding speech file in SCRIBE database. All functionalities for SFS are packaged into a class library called "SPLib[6]". A certain class for processing SFS commands is implemented. It is called "SFSFile". It is located into "SPLib".

```
foreach (string file in files)
        {
          SPLib.SFSFile f = new SPLib.SFSFile();
          string sfsFile;
          string anFile;
          int start = file.LastIndexOf('\\');
          int end = file.LastIndexOf('.');
          sfsFile = targetdir + file.Substring(start, end - start) + ".sfs";
          anFile = file.Substring(0, file.Length - 1) + "a";
          f.open(sfsFile);
          f.AddSPItem(file, 20000);
          f.AddANItem(anFile);
         }
```

"AddSPItem" and "AddANItem" are subroutines that contain the SFS commands which are previously listed.

Now we have SFS file for each of the database files. This SFS file will act as a container for speech signal, associated annotation symbols and associated features.

The next operation to be applied on the speech database files is to map the phonetic annotations into Voiced, Unvoiced and Silent annotations. This is important for HTK to understand the classes to be trained. Let us denote the new annotation set with a suitable name for the upcoming references. The new set is called speech type set (STS). So, STS contains three speech type symbols

1. Voiced speech symbol (VOI).
2. Unvoiced speech symbol (UNV).
3. Silent periods or no speech symbol (SIL).

The first step is to make a MAP file that links each phonetic symbol into a suitable type symbol in STS. This file is manually created (by human not by program). Part of the map file is shown below:

```
#      SIL
##     SIL
%tc    SIL
+      SIL
/      SIL
3:     VOI
3:?    VOI
3:a    UNV
3:af   UNV
3:f    UNV
3:~    VOI
=l     VOI
=lx    VOI
=lx?   VOI
=lxf   UNV
```

The first column is the phonetic symbol and the second column is the map to STS symbol. A complete version of the map file may be downloaded from [4]. To apply

[6] SPLIB: Speech lib class library. It is a C# class library by Amr M. Gody to work with SFS and Matlab. It encapsulates all needed logic and business to work with speech signal using SFS or Matlab.

the map operation to annotation item inside an SFS file, the following command line is invoked:

```
Anmap -m  mapfile sfsfile
```

To apply the map operation on all the available SFS files, the following program is written as BCL:

```
foreach (string file in files)
            {
                SPLib.SFSFile f = new SPLib.SFSFile();
                f.open(file);
                f.MapAnnotation(mapfile);
            }
```

All SFS files are listed into string array called "files". Then for each "file" in "files" the map annotation is applied. " `MapAnnotation`"  is a subroutine contains the SFS command that was indicated above.

## 4. Features extraction

In this section the process of feature extraction will be explained.  We have two different groups as indicated in section 2. SET_A will be designated for VOC19 while SET_B will be chosen for BTE features.  SFS will be used to apply VOC19 on all available samples.  The following is the SFS command to do the function:

```
voc19 sfsfile
```

To apply the above command to all available samples in SET_A, the following program is written as BCL:

```
foreach (string file in files)
            {
                SPLib.SFSFile f = new SPLib.SFSFile(file);
                f.VOC19();
            }
```

All SFS files in SET_A are listed into string array called "files". Then the loop is applied on each file into files. VOC19 subroutine contains the SFS command that was indicated above.  After this step, each SFS file in SET_A contains a new item that express VOC19 features. It is called coefficients item.

Now the coefficients item, contained into the SFS file that represents VOC19 features, needs to be exported into an HTK formatted file to be used in further step during the training of HMM model using HTK. The following SFS command do this function:

```
Colist -H sfsfile
```

To apply the above SFS command on all the available SFS files into SET_A, the following code snippet in BCL is used:

```
foreach (string file in files)
            {
                SPLib.SFSFile f = new SPLib.SFSFile();
                f.open(file);
                f.Co2HTK();

            }
```

In the above code snippet, all SFS files in SET_A are listed into a string array called "files". The function "Co2HTK" contains the SFS command needed to export the features from the SFS file to HTK formatted file.

It is also needed to extract the annotation item from the SFS file to an HTK formatted annotation file. This is achieved by calling the following SFS command:

```
anlist -h -O sfsfile
```

Then BCL is used to apply the function on all the available SFS files in SET_A and SET_B. The following code snippet in BCL is used to achieve this objective:

```
foreach (string file in files)
            {
                SPLib.SFSFile f = new SPLib.SFSFile();
                f.open(file);
                f.An2HTK();
            }
```

In the above code snippet, all SFS files in SET_A and in SET_B are listed into string array called "files" Then the function "An2HTK" is called for each file in the string array. "An2HTK" contains the SFS command mentioned above.

A parallel process is implemented on SFS files in SET_B. This time the proposed features (BTE) will be extracted. Matlab instead of SFS is used to implement BTE features extraction process. The following code snippet is the core part of Matlab function to implement BTE features extraction.

```
function [res] = BTE (frame, depth)
    nbIn = nargin;
    nbout = nargout;
    if nbIn < 1 ,    error('Not enough input arguments.');
    elseif nbIn == 1,     level = 4;
    elseif nbIn == 2,     level = depth;
    end;
    if nbout < 1 , error('Not enough output arguments.'); end;
    t = wpdec(frame,level,'db4','shannon');
    u = leaves (t);
    bt =  besttree(t);
    v = leaves (bt);
    res = box4encoder(v);
end
```

The function "box4encoder" in the above code snippet is responsible for encoding Best tree as indicated in [1].

To apply BTE algorithm on all available samples in SET_B, BCL is used. The following code snippet is used to apply BTE to all available speech samples assigned for SET_B experiment.

```
foreach (string file in files)
            {
                SPLib.SFSFile f = new SPLib.SFSFile();
                f.open(file);
                f.ExportWAV();
                string wfile = file.Substring(0, file.Length - 3) + "WAV";
                mat.wav2bte (1, wfile);

            }
```

All SFS files in SET_B are listed into a string array called files. The speech waveform is exported from the SFS file to a known format called WAV file format. This is important to pass the sound file to the Matlab function. The Matlab function called "wav2bte" is called for each file in the string array "files". For more information on the function "wav2bte" you may referee to [1].

Now it is needed to prepare the generated BTE files for being used by HTK. BCL is used to write such a converter. The following cod snippet is written for the converter function:

```
public static void BTEtoHTK(string BTEfile, string htkfile)
        {

BTEfile f1 = new BTEfile(BTEfile);
HTKFile f2 = new HTKFile();
int samplein100ns = Convert.ToInt32 (  f1.SampleLength * 1e-3 / 100e-9);
short bytesperhtksample =Convert.ToInt16 (  f1.BytesPerSample * 4 /
f1.BytesPerElemnt); // HTK is 4 bytes/element
f2.create(f1.NumberOfSamples, samplein100ns, bytesperhtksample,
SPLib.HTKParamKind.USER, htkfile);
            int n;
            n = f1.NumberOfSamples;
            int m = f1.ElementsPerSample;
            for (int i = 0; i < n; i++)
            {


                for (int j = 0; j < m; j++)
                {
                    int elm =(int) f1.ReadInt16();
                    f2.write( elm);
                }
            }
            f1.close();
            f2.close();
        }
```

By the end of the above step, we should have all the training files needed by HMM for both groups as shown in table3.

**Table 1: Snapshot of HMM training files generated so far by the end of feature extraction step.**

| Group | Feature type | HTK feature files | HTK annotation files |
|---|---|---|---|
| SET_A<br>(24 files) | VOC19 | AAPA0001.dat<br>⋮ | AAPA0001.lab<br>⋮ |
| SET_B<br>(24 files) | BTE | AAPA0001.htk<br>⋮ | AAPA0001.lab<br>⋮ |

## 5. Training HMM

After features extraction step, it is the time for testing the features into a pattern recognition process. As indicated in Table 1, two sets of files are prepared. They are both ready for training HMM models using HTK.

First step in this phase is to design HMM model that best fit the information needed to be recognized. The model here is 3 states left to right model. This assumes that the recognized pattern is assumed to have three different parts. The parts are consequent parts. The first part is the left one and the last part is the right one. This assumption is very close to the reality as the consecutive sounds is supposed to have a transition periods at the boundaries and a stable period at the middle.  Figure 14 explains the relation between the proposed design model and speech sound.  In this experiment there are there sounds to be recognized {Voiced (VOI), Unvoiced (UNV) and silent (SIL)}.

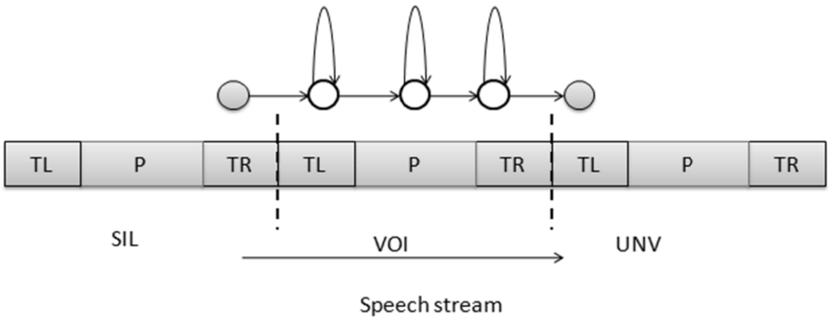


**Figure 2: Relation between HMM model and speech sounds to be recognized. TL is the leading transition period and TR is the trailing transition period while P is the stable phone period.**

The model contains two non emitting states which appear in gray color in figure 2. The non emitting states are important in HTK to indicate the entry and the exit points to the model. Gaussian Probability Distribution Function (PDF) is used in each state to fit the variability of the sound. To define an HMM model for HTK the following script is written into a separate text file.

```
~o
<STREAMINFO> 1 19
<VECSIZE> 19<NULLD><FBANK><DIAGC>
~h "SIL"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<MEAN> 19
 1.404372e+001 1.146923e+001 1.089870e+001 7.190041e+000 2.423316e+000 1.263892e+000
1.634800e+000 4.447996e-002 1.654768e+000 3.511177e+000 4.378909e+000 3.915090e+000
8.674143e-002 1.574287e-001 -6.058807e-001 -1.295122e+000 -2.223198e+000 -
2.120687e+000 -9.183334e-001
<VARIANCE> 19
 3.086359e+002 2.563524e+002 1.796036e+002 1.263492e+002 8.962412e+001 7.586296e+001
7.740655e+001 6.912480e+001 8.007733e+001 9.697153e+001 1.057107e+002 1.151898e+002
7.569676e+001 7.488948e+001 6.804313e+001 6.226783e+001 5.014687e+001 5.277541e+001
6.593863e+001
<GCONST> 1.210665e+002
<STATE> 3
<MEAN> 19
 -1.550593e+000 -3.269745e+000 -4.287612e+000 -5.243945e+000 -5.796925e+000 -
5.822775e+000 -5.803507e+000 -5.910261e+000 -5.803027e+000 -5.761147e+000 -
5.671880e+000 -5.702754e+000 -5.948490e+000 -5.959568e+000 -5.960902e+000 -
5.970729e+000 -5.993655e+000 -5.991790e+000 -5.961835e+000
<VARIANCE> 19
 8.319610e+001 4.678292e+001 2.388155e+001 8.794478e+000 2.526003e+000 1.837515e+000
2.220891e+000 1.482371e+000 2.225448e+000 2.601058e+000 3.406267e+000 3.234430e+000
8.082389e-001 8.287914e-001 8.249093e-001 5.735081e-001 2.694195e-001 3.057849e-001
6.422817e-001
<GCONST> 5.132733e+001
<STATE> 4
<MEAN> 19
 4.351817e+000 2.717550e+000 2.140117e+000 1.390050e+000 -1.326494e+000 -1.220001e+000
4.477349e-001 -1.336653e+000 -1.657292e-002 9.020020e-001 2.071208e+000 3.250645e+000
-2.780049e-001 -8.483851e-001 -9.663507e-001 -9.924042e-001 -1.579547e+000 -
1.328159e+000 -4.995179e-001
```

```
<VARIANCE> 19
 1.096481e+002 9.111847e+001 8.397858e+001 7.596156e+001 5.026793e+001 4.964570e+001
7.590692e+001 6.081587e+001 6.890288e+001 7.740077e+001 8.851939e+001 1.061187e+002
7.655155e+001 7.167074e+001 7.540655e+001 8.126243e+001 7.148788e+001 7.071327e+001
8.899091e+001
<GCONST> 1.172263e+002
<TRANSP> 5
 0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
 0.000000e+000 7.436733e-001 2.563267e-001 0.000000e+000 0.000000e+000
 0.000000e+000 0.000000e+000 9.159696e-001 8.403045e-002 0.000000e+000
 0.000000e+000 0.000000e+000 0.000000e+000 7.777685e-001 2.222315e-001
 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
<ENDHMM>
```

The above script defines all model parameters. It defines the following items

1- Number of states.
2- Feature type.
3- Transition Matrix.
4- Gaussian PDF parameters in each state (Means and Variance).
5- The name of the class. In the above example it is "SIL".

For more details you may referee to [5].

The above is an initial definition. This definition will be adapted on the vision of the training data available for the training phase. Two similar HMM models will be defined to model VOI and UNV sounds.

Now it is important to split the available speech database files into two groups. One group will be used for training and the other group will be used for test. The following list is the training set files. It is saved into a text file called "train.lst".

```
AAPA0002.dat
AAPA0003.dat
AAPA0004.dat
ACPA0002.dat
ACPA0003.dat
ACPA0004.dat
AEPA0002.dat
AEPA0003.dat
AEPA0004.dat
AFPA0002.dat
AFPA0003.dat
AFPA0004.dat
AHPA0002.dat
AHPA0003.dat
AHPA0004.dat
AMPA0002.dat
AMPA0003.dat
AMPA0004.dat
```

Another list will be prepared for test. The following is the test list. It is saved into a text file called "TEST.LST".

```
AAPA0001.dat
ACPA0001.dat
AEPA0001.dat
AFPA0001.dat
AHPA0001.dat
AMPA0001.dat
```

We have two sets of files for testing the model as indicated in section 4. SET_A for the filter banks features and SET_B for BTE features. It is important to configure HTK such that it can understand the type of features under test. The following text is saved into a text file called "config.txt".

```
# config.txt - HTK basic parameters
SOURCEFORMAT = HTK
TARGETKIND = FBANK
NATURALREADORDER = T
```

The configuration file will be provided for any HTK command to configure it to correctly understand the provided features during the test or the training phases. The above is the configuration file for Filter banks features. The following is the configuration file for BTE features type.

```
# config.txt - HTK basic parameters
SOURCEFORMAT = HTK
TARGETKIND = USER
NATURALREADORDER = T
```

BTE is a new feature type so that it should be provided to HTK as user defined type as indicated in the above script "TARGETKIND = USER".

After defining the three HMM models, it is better to initialize them using the avalible database. This is good before starting the training phase. The following HTK command is used to intitialize each HMM model:

```
hinit -T 1 -c config.txt -s train.lst SIL
hinit -T 1 -c config.txt -s train.lst VOI
hinit -T 1 -c config.txt -s train.lst UNV
```

The above three commands should initialize the available three HMM models on the vision of the available database for each class. The above step will be applied on the initial models in {SET_A and SET_B}.

Now the models are ready to be trained using the HTK. The following commands are used to train the models in both groups:

```
HRest -T 1 -C config.txt -S train.lst -l VOI VOI
HRest -T 1 -C config.txt -S train.lst -l UNV UNV
HRest -T 1 -C config.txt -S train.lst -l SIL SIL
```

The above step may be repeated till log probability approaches to 0 or approaches to stable value. As soon as the model is well trained, we can start the testing phase. By the end of this step we should have 6 HMM models as indicated in table 2.

**Table 2: HMM files after the training phase.**

| Group | Feature type | HMM files |
|---|---|---|
| SET_A<br>(3 files) | VOC19<br>(Filter Banks) | VOI<br>UNV<br>SIL |
| SET_B<br>(3 files) | BTE | VOI<br>UNV<br>SIL |

## 6. Testing HMM

As shown in table 2, we have three HMM models for each group of files. Now it is needed to test the trained models using the available testing files in each group. First it is needed to prepare dictionary and word net. Dictionary contains all recognized words while the Word net contains the grammar. Both of them are important for HTK to get it correctly functioning. The problem we address in this research is a simple classification problem that may not need grammars. So the word net will be prepared in such way that matches with our needs. Figure 3 explains the way HTK alters grammars. The top in the hierarchy is the word. Each word may be expressed in a network of phones as each phone network represents certain pronunciation for the associated word. And finally each phone is expressed in HMM model.

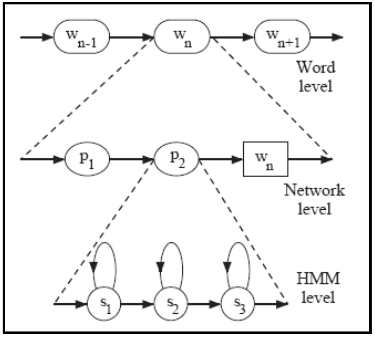


**Figure 3: HTK recognizer in depth. The abbreviations are explained as W for Word, P for Phone and S for state. The dotted boundary explains the decomposition of the root element[5].**

In our case the network is very simple. It will be constructed statistically from the available samples. Each sample has a label file that explains sample contents in term of VOI, UNV and SIL symbols. The following is a part of certain label file:

```
24917000 26727500 UNV
26727500 30867000 VOI
30867000 32193500 UNV
32193500 32375000 SIL
32375000 32414500 VOI
```

The first column indicates the beginning of the segment and the second column indicates the end of the segment. Numbers are in term of 100(ns). For example 24917000 means segment (UNV) will start at$24917000 \times 100 \times 10^{-9} = 2.4917(sec)$. The following HTK command is invoked to build the word net from the available label files.

```
HBuild voices.dic voices.net
```

The above command uses the symbols in the file "voices.dic" to construct the file "voices.net" using all label files exist in the same directory. The file "voices.dic" may be like the following script:

```
SIL [SIL] SIL
VOI [VOI] VOI
UNV [UNV] UNV
```

The dictionary file maps the word to its possible pronunciation phone streams. Here in our example the word is the same as the phone stream. Word VOI is constructed of phone VOI. In our experiment the word is the same as the phone. It is a one level recognition. We do not have further resolutions for each word. The word between the square brackets is the output symbol. It is used by HTK to provide suitable output when word is recognized. It is an optional parameter. The network file generated by "HBuild" command is like the following:

```
 VERSION=1.0
N=7    L=9
I=0    W=!NULL
I=1    W=!NULL
I=2    W=UNV
I=3    W=SIL
I=4    W=VOI
I=5    W=!NULL
I=6    W=!NULL
J=0     S=0    E=1    l=0.00
J=1     S=5    E=1    l=0.00
J=2     S=1    E=2    l=-1.10
J=3     S=1    E=3    l=-1.10
J=4     S=1    E=4    l=-1.10
J=5     S=2    E=5    l=0.00
J=6     S=3    E=5    l=0.00
J=7     S=4    E=5    l=0.00
J=8     S=5    E=6    l=0.00
```

J means joint, S means start, E means end, W means word symbol, l means log probability and I is node identifier. Figure 4 explains the structure of the word net generated by "HBuild".

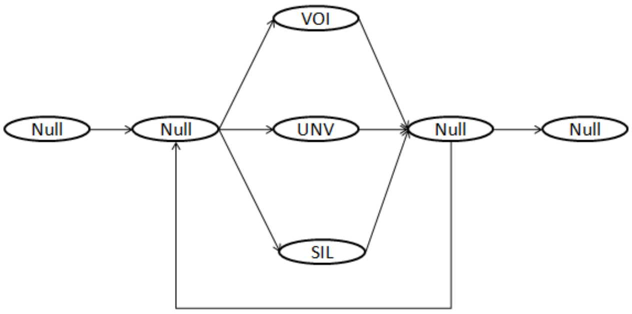


**Figure 4: Word net structure.**

After constructing the dictionary and word net files, it is possible to start testing the models. All test files will be fed to HTK for the recognition process. The following HTK command is used to start testing the models against the available testing files:

```
HVite -T 1 -C config.txt -w voices.net -o S -S test.lst voices.dic words.lst
```

The above HTK command should be applied to both sets {SET_A and SET_B}. The file "words.lst" contains the name of HMM model files. In this experiment it is like the following script:

```
SIL
VOI
UNV
```

After executing the above command, all recognition results will be exported into files in the same name as the test files with a new file extension ".rec". The generated ".rec" file is just similar to the standard label file. The following is a part of such generated files:

```
11800000 13600000 VOI
13600000 30000000 SIL
30000000 31200000 VOI
```

```
31200000 32800000 SIL
32800000 36000000 VOI
36000000 37400000 SIL
```

To start results mining process, it is required to import the recognition files into the associated SFS files. BCL will be used to apply the import process on all files in both sets {SET_A and SET_B}. The following code snippet is written to make the function:

```
foreach (string file in files)
            {
                int index = file.LastIndexOf('.');
                string sfsfile = file.Substring(0, index) + ".sfs";
                Process a = new Process();
                a.StartInfo.FileName = "anload";
                a.StartInfo.Arguments = "-h "+file +" " + sfsfile;
                a.StartInfo.RedirectStandardOutput = true;
                a.StartInfo.UseShellExecute = false;
                a.Start();
                a.WaitForExit();
            }
```

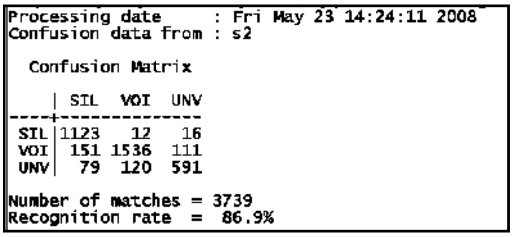
The above code snippet call the following SFS command for each file in the group to be imported into the SFS file:

```
anload -h recfile  sfsfile
```

After importing all recognition files into the associated SFS files, each SFS file will contain two annotation items. The reference annotation item and the recognized annotation item will be used by SFS to estimate the recognition rate. The results will be obtained by comparing the reference annotation item to the recognized annotation item for each 10(ms) of the spoken period. The following two SFS commands perform the annotation comparison. The output is a confusion matrix. The following lists the commands:

```
ancomp -r an.02 -t an.03 -f -m - AcPA0001.sfs > s1
Conmat s1
```

The above two commands are applied on all testing SFS files. The generated confusion matrix is like the one in figure 5.



```
Processing date      : Fri May 23 14:24:11 2008
Confusion data from : s2

  Confusion Matrix

    | SIL  VOI  UNV
----+---------------
 SIL|1123   12   16
 VOI| 151 1536  111
 UNV|  79  120  591

Number of matches = 3739
Recognition rate  =  86.9%
```

**Figure 5: Confusion matrix for a certain recognition process.**

The confusion matrix gives allot of information for the recognition process. Figure 5 gives such an example of the matrix. In this matrix we can notice that SIL sounds is recognized as SIL in 1123 matches out of (1123+12+16 = 1151). SIL is recognized as VOI for 12 times and as UNV for 16 times.

## 7. Results

The results of this research will be analyzed in this section. We have two sets of files. SET_A deals with Filter banks (VOC19) and SET_B deals with Best Tree 4 point Encoded features (BTE).

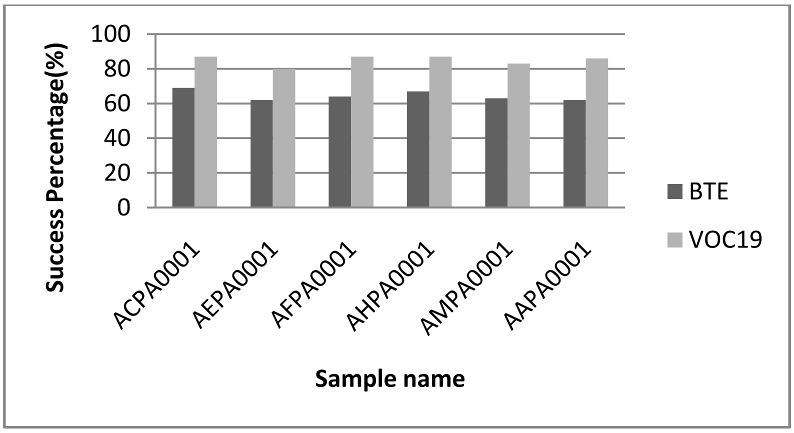


**Figure 6: Comparison chart of overall system performance.**

Figure 6 indicates the overall performance. As it is the first round in using BTE, Filter Banks (VOC19) features indicate a significant better performance than BTE. Many enhancements still may be added in the future to go around the drawback in the currently proposed BTE features. The detailed analysis of the results is introduced below to figure out the obtained results. As it will be shown below, the features failed in recognizing the UNV sounds while it makes comparable results in recognizing SIL and VOC. Figures 7, 8 and 9 provide the comparison results for the three classes under test on both features {BTE and VOC19}.

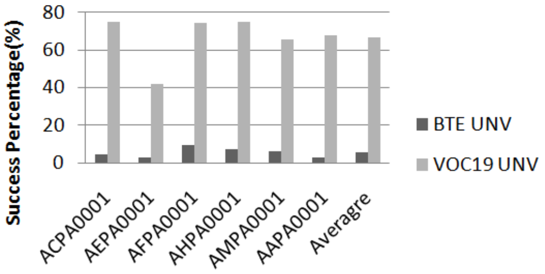


**Figure 7: Comparison between BTE and Filter banks in recognizing (UNV) sound.**

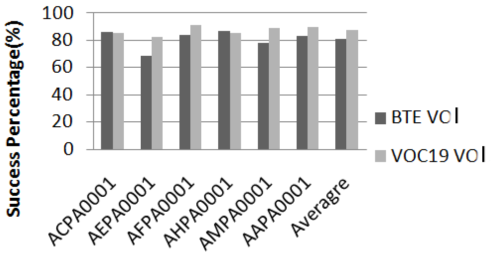


**Figure 8: Comparison between BTE and Filter banks in recognizing (VOI) sound.**

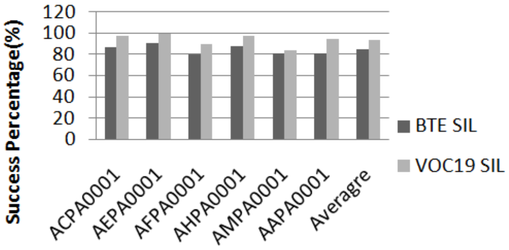


**Figure 9: Comparison between BTE and Filter banks in recognizing (SIL).**

## 8. Conclusions

This is a preliminary study to introduce BTE features. Many enhancements may be included in the future to minimize the confusion results being discussed in section 7.

## 9. References

[1] Amr M. Gody,"Wavelet Packets Best Tree 4-Points Encoded (BTE) Features", The 8th Conference on Language Engineering.2008, Cairo, Egypt.

[2] Mel scale, http://en.wikipedia.org/wiki/Mel_scale

[3] Giridharan, K. Smolenski, B.Y. Yantorno, R.E, "Statistical and model based approach to unvoiced speech detection", Intelligent Signal Processing and Communication Systems, 2004. ISPACS 2004. Proceedings of 2004 International Symposium, On page(s): 816 – 821

[4] University College London: http://www.phon.ucl.ac.uk/resource/sfs/howto/htk.htm.

[5] http://htk.eng.cam.ac.uk/

# دور المؤثرات السياقية في تقدير المدى الزمني للفونيم

**أحمد راغب أحمد**

مركز الإنسان للدراسات والتطوير –الرياض

**Ragheb31@yahoo.com**

## ملخص الدراسة:

تعتبر أحكام المد والقصر الأساس الأول لدراسة الأحكام الكمية في علم التجويد، وهو من أبرز الحلى المزينة لتلاوة القرآن الكريم، مع التسليم بدور هذه الحلى اللفظية في خدمة السياق الدلالي العام لأغراض الخطاب القرآني، تحاول الدراسة معرفة التغيرات السياقية التي توثر في زمن نطق الفونيمات العربية عن طريق رصد وحساب متوسطات الحركات القصيرة والطويلة للغة العربية "vowels" وذلك في جميع صور الأداء القرآني، وتتضمن الورقة مبحثين متتابعين على النحو التالي:

المبحث الأول: الإطار النظري لأحكام للحركات "vowels"، ويتناول مفهوم المد والقصر، وأقسام المد ولواحقه وأحكامه ومسمياته.

المبحث الثاني: القسم التطبيقي، ويتناول تحليل الأداء الصوتي لآيات القرآن الكريم مقروءة بأصوات قراء إذاعة القرآن الكريم بمصر على طريقة الترتيل، وهؤلاء القراء، وكانت المادة الصوتية مسجلة على أسطوانات أوديو، وكلها خالية من المؤثرات الصوتية الموجهة –(Echo)– وغير الموجهة –(Noise)–، وكان اختيار مادة البحث على اعتبار أنها المادة التي تمت مراجعتها واعتمادها لدى الشركة الهندسية لتطوير نظم الحاسبات RDI ضمن فعاليات مشروعات المعلم الآلي للتجويد ((حفص)).

ويمكن إجمال الأهداف المرجوة من هذه الدراسة في النقاط التالية:

1. رصد وتقييم المدة الزمنية للمد في الأداء القرآني ومعرفة العوامل والمؤثرات التي تؤدي إلى زيادة أو نقص هذه المدة الزمنية.
2. نقد وتقييم نظرية الحركة والحركتين التي اعتمدها علماء التجويد والقراءات في تحديد مقدار زمن الحركات القصيرة والطويلة في تلاوة القرآن الكريم.
3. رصد علاقة زمن المد بنوع الصامت السابق له.
4. رصد علاقة زمن المد بظاهرة النبر.
5. ترتيب أنواع المد من ناحية المدة الزمنية.

# المبحث الأول

# الإطار النظري لأحكام المد والقصر

## مفهوم المد والقصر

المَدُّ في اللُّغةِ : المط والتطويل والإكثار والزيادة (1).

واصطلاحًا: هو إطالة زمن جريان الصوت بأصوات المد واللين عند ملاقاة أحد سببي المد –الهمز والسكون–.

وقد تقاربت تعريفات المد "في اللفظ، وتطابقت في الدلالة، فهي تجمع على أن المد إطالة صوت المد زيادة على ما فيه من مد طبيعي لا تقوم ذات الحرف إلا به، ولتلك الزيادة أسباب، ولها مقدار"(2).

**حروف المد:**

حروف المد ثلاثة: الأَلِفُ السَّاكِنَةُ المَفْتُوحُ مَا قَبْلَهَا دائمًا، وَالوَاوُ السَّاكِنَةُ المَضْمُومُ ما قَبْلَهَا، وَالياءُ السَّاكِنَةُ المَكْسُورُ ما قَبْلَهَا.

وتسمى هذه الأصوات بأصوات المد واللين لخروجها بامتداد ولين من غير كلفة على اللسان لاتساع مخرجها، وقد دأب علماء التجويد على تضمين هذه الحروف الثلاثة في كلمة واحدة هي: **نوحيها.**

"وتسمى هذه الحروف حروف مد ولين؛ لامتدادها في لين وعدم كلفة، كما تسمى جوفية لخروجها من الجوف، وهوائية لقيامها بهواء الفم، وخفية لخفاء النطق بها، فهي أخفى الحروف، وأخفاهن الألف، ثم الياء، ثم الواو"(3).

**حرفا اللين(4) :**

إذا كانت الواو والياء ساكنتين، وانفتح ما قبلهما كانتا حرفي لين، ومثاله: **{خَوْف}**، **{قُرَيْشٍ}**، علمًا بأنه إذا أطلق حرف المد فيراد به المد واللين، وإذا قيد باللين، فيختص به حينئذ.

"فإذا كان ما قبل الواو مفتوحًا نحو: **{خوف}**، **{نوم}**، **{الموت}**، وكان ما قبل الياء مفتوحًا نحو: **{بيع}**، **{غير}**، **{والصيف}** كانا حرفي لين فقط، ولا يمدان أصلًا إلا إذا تلاهما ساكن عارض عند الوقف، أو ساكن لازم، والأول يقع في القرآن كثيرًا بخلاف الثاني الذي لم يقع في القرآن إلا بعد الياء وذلك في فاتحتي: مريم والشورى"(5).

وعليه فإن كل حرف مد حرف لين، بينما ليس كل حرف لين حرف مد، فالمد أخص واللين أعم.

"وقصارى القول أن الألف لا يكون إلا حرف مد ولين لسكونها وانفتاح ما قبلها دائمًا، وأن الواو والياء تارة يكونان حرفي مد ولين إذا جانسهما ما قبلهما بأن سكنت الواو بعد ضم، وسكنت الياء بعد كسر، وتارة يكونان حرفي لين فقط إذا سكنتا وانفتح ما قبلهما"(6).

[1] نصر، عطية قابل. غاية المريد في علم التجويد، ط3، س1413 هـ، دار الحرمين للطباعة، القاهرة، ص:91.

[2] د. الحمد، غانم قدوري، الدراسات الصوتية عند علماء التجويد، ص: 523.

[3] الحصري، محمود خليل، أحكام قراءة القرآن الكريم، ص: 175.

[4] اللين: يقال: نزلوا بلين الأرض. وحروف اللين: الألف والواو والياء. واللين: كل نوع من أنواع النخل سوى العجوة، الواحدة: لينة، وفي التنزيل العزيز (ما قطعتم من لينة). [المعجم الوسيط، ص: 850، مادة (لين)].

[5] الحصري، محمود خليل، أحكام قراءة القرآن الكريم، ص 176.

[6] الحصري، محمود خليل، أحكام قراءة القرآن الكريم، ص: 176، 177.

**أقسام المد:**

ينقسم المد إلى قسمين: طبيعي، وفرعي. فالمد الطبيعي هو المد الذي لا يرتبط بسبب ولا يقوم الحرف إلا به، بينما المد الفرعي هو الذي يعتمد على سبب كهمز أو سكون، وإليك تفصيل ذلك.

**المد الطبيعي أو الأصلي:** هو ما لا تقوم ذات الحرف إلا به، ولا يتوقف على سبب كهمز بعده أو سكون، مثاله: {قال} {قيل} {يقول} {طه}. ومقدار مده حركتان(7)، ولا يجوز الزيادة أو النقصان عن الحركتين.

**وينقسم المد الطبيعي إلى قسمين فرعيين:**

- المد الطبيعي الحرفي: وهو ما كان في فواتح السور، وذلك في خمسة حروف مجموعة في كلمة **حي طهر** ، ومقدار مده حركتان.
- المد الطبيعي الكلمي: وهو ما كان في كلمة واحدة.

**المد بسبب الهمز:**

إذا اجتمع حرف المد مع الهمز نتج عنه أحد صور المد الآتية:

❖ مد البدل.

❖ المد المتصل.

❖ المد المنفصل.

**أ- إن كان الهمز قبل حرف المد فهو مد البدل:**

**ب- إن كان الهمز بعد حرف المد: فهو نوعان:**

**المد المتصل:** هو أن يأتي حرف المد والهمز بعده في كلمة واحدة، ويسمى المد الواجب المتصل. ويمد أربع أو خمس حركات.

**المد المنفصل:** هو أن يأتي حرف المد في آخر كلمة، والهمز بعده في بداية الكلمة التي تليها، ويسمى المد الجائز، ويمد أربع أو خمس حركات، ونستطيع أن نقصره إلى حركتين.

**المد بسبب السكون: وهو نوعان:**

**أ- سكون عارض:** وهو أن يكون الحرف قبل الأخير من الكلمة حرف مد، والحرف الأخير متحرك، فإن درجنا الكلام ووصلنا الكلمة بما بعدها كان المد طبيعيًا، وإن وقفنا على الحرف الأخير بالسكون صار المد الذي قبل الحرف الأخير مدًا بسبب السكون العارض، ويسمى: مدًا عارضًا للسكون، يمد ست حركات، أو أربع، أو حركتان.

**ب- سكون لازم:** وهو أن يأتي بعد حرف المد سكون لازم وصلًا ووقفًا في كلمة واحدة، ومقدار مده ست حركات، وهو نوعان:

1. **الكلمي:** وهو أن يأتي بعد حرف المد حرف ساكن في كلمة، فإن أدغم -أي كان الحرف الذي بعد المد مشددًا- فيسمى مثقلًا.

2. **الحرفي:** ويوجد في فواتح بعض السور، في الحرف الذي هجاؤه ثلاثة أحرف أوسطها حرف مد والثالث ساكن. وحروفه مجموعة في: **[[نَقص عسلكم]]** فإن أدغم سمي مثقلًا. مثاله: **(الم)** [البقرة/١]، **ﭽ** المر **ﭼ** [الرعد/١]، **ﭽ**طسم**ﭼ**

[7] سوف أتناول قضية الحركة والحركتين في نهاية هذا الباب إن شاء الله، انظر ص:25 .

[الشعراء/١].

وإن لم يدغم سمي مخففًا. مثاله: **ﭽ**ن والقم وما يسطرون**ﭼ** [القلم/١]، **ﭽ** ق والقرآن المجيد**ﭼ** [ق/١]، **ﭽ**المص**ﭼ** [الأعراف/١].

أما حرف العين من فاتحتي مريم والشورى فقد "اختلف أهل الأداء في إشباعها وتوسطها وقصرها، فمنهم من أجراها مجرى حرف المد، فأشبع مدها لالتقاء الساكنين، ومنهم من أخذ بالتوسط نظرًا لفتح ما قبل الياء ورعاية للجمع بين الساكنين"(8).

**لواحق المد:**

والمقصود بلواحق المد: مجموعة المدود الفرعية الطارئة على الحرف، والتي يمكن إجمالها مثل مد العِوَض ومد التمكين ومد اللين ومد الصلة.

(8) الحصري، محمود خليل، أحكام قراءة القرآن الكريم، ص: 189.

# المبحث الثاني

# المعالجة التطبيقية لأصوات المد

## قاعدة البيانات

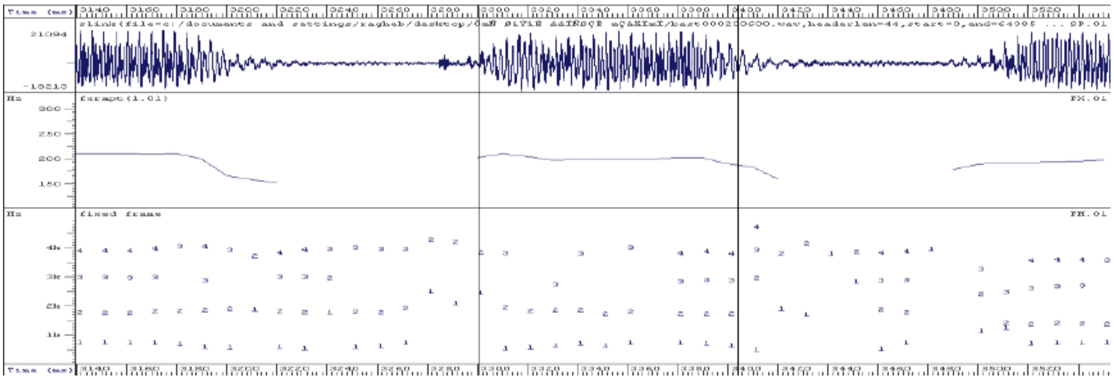
يقوم هذا البحث على تحليل آيات القرآن الكريم المقروءة على طريقة الترتيل بأصوات خمسة من أشهر قراء القرآن الكريم بإذاعة القرآن الكريم بمصر، وهم: الشيخ/ **محمود خليل الحصري**، والشيخ/ **مصطفى إسماعيل**، والشيخ/ **محمد علي البنا**، والشيخ/ **عبد الباسط عبد الصمد**، والشيخ/ **محمد صديق المنشاوي**، وكانت المادة الصوتية مسجلة على أسطوانات أوديو، وكلها من تسجيلات شركة صوت القاهرة للصوتيات والمرئيات، وكلها خالية من المؤثرات الصوتية الموجهة –**(Echo)**– وغير الموجهة –**(Noise)**–، وكان اختيار مادة البحث على اعتبار أنها المادة التي تمت مراجعتها واعتمادها لدى الشركة الهندسية لتطوير نظم الحاسبات **RDI** ضمن فعاليات مشروعات المعلم الآلي للتجويد **((حفص))**، وقد تم إعدادها بواسطة فريق الدعم اللغوي بقسم أبحاث ومعالجة الصوتيات بالشركة، وذلك خلال الفترة من يناير 2001م، وحتى مارس 2004م.

### الأهداف

يمكن إجمال الأهداف المرجوة من دراسة المدود في هذا الفصل في النقاط التالية:

1. رصد وتقييم المدة الزمنية للمد في الأداء القرآني ومعرفة العوامل والمؤثرات التي تؤدي إلى زيادة أو نقص هذه المدة الزمنية.
2. نقد وتقييم نظرية الحركة والحركتين التي اعتمدها علماء التجويد والقراءات في تحديد مقدار زمن الحركات القصيرة والطويلة في تلاوة القرآن الكريم.
3. رصد علاقة زمن المد بنوع الصامت السابق له.
4. رصد علاقة زمن المد بظاهرة النبر.
5. ترتيب أنواع المد من ناحية المدة الزمنية.

### التحليل: الفتحة القصيرة المرققة /a/ :



شكل 1 يعرض صورة طيفية للفتحة القصيرة /a/ من خلال كلمة **كفروا** التي وردت في قوله تعالى: ﭽ إن الذين كفروا سواء عليهم أأنذرتهم أم لم تنذرهم لا يؤمنونﭼ [البقرة/٦]،

ويظهر في الشكل صوت الفتحة القصيرة، والذي بدأ من الثانية 03:090، وانتهى عند الثانية 03:241، مستغرقًا زمنًا مقداره 0.151 ميللي ثانية، وتعرض هذه الصورة ثلاثة مستويات للتحليل الصوتي:

1. المستوى الأول الأعلى يعرض الشكل الموجي –(wave form)– ويبدو جليًا أنها لصوت مجهور؛ حيث وجود إشارة الذبذبات التي تقترن دائمًا بالأصوات المجهورة، بخلاف الصوت التالي أو السابق لها.
2. المستوى الثاني يعرض النغمة الأساسية أو منحنى التنغيم الأساسي –(Fundamental Frequency)– ونلاحظ اتصال الخط القاعدي لها، وهو أمر ملازم للأصوات المجهورة فقط.
3. المستوى الثالث formants يعرض المعالم الأولى والثانية والثالثة –(f1,f2,f3)– ونجدها موزعة توزيعًا منتظمًا متتابعًا مما يدل على انتماء هذا الصوت إلى مجموعة الأصوات المجهورة وعليه فإن صوت الفتحة القصيرة /a/ صوت مجهور لا تظهر فيه أية معالم من معالم الهمس.

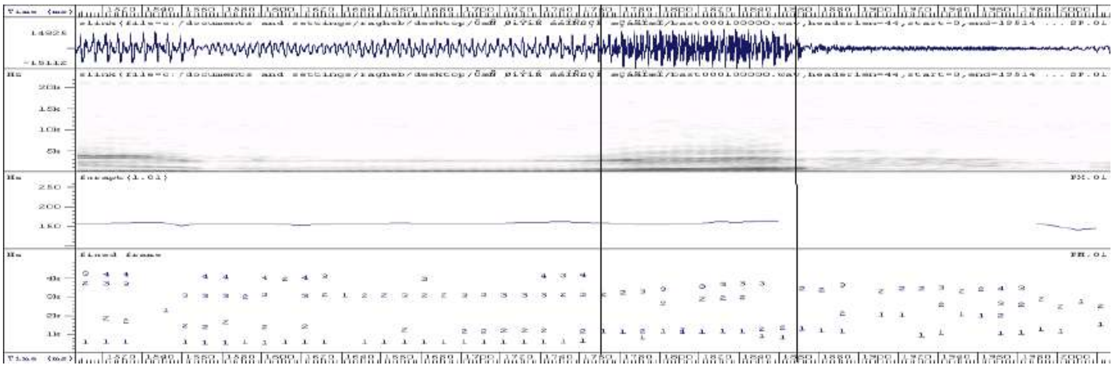
وهذه الحزم الصوتية والتي يطلق عليها formants أو المعالم هي "الترددات أو مجموعة الترددات –( groups of frequencies)– التي تشكل نوع الصوت –(Timpre)– وتميزه عن الأصوات الأخرى ذات الأنواع المختلفة"(9).

ومن استقراء القيم التي حواها ملف التحليل الصوتي نجد الآتي:

بلغ متوسط قيمة المعلم الأول 684 ذبذبة والمعلم الثاني 1602 ذبذبة والمعلم الثالث 2839 ذبذبة،

ونلاحظ أثناء تحليل قيم المعالم الثلاثة –(formants)– ارتفاع قيم المعلم الأول مع انخفاض قيم المعلم الثاني بشكل ملحوظ، وهو أمر مرده عملية الترقيق، وقد أدى هذا الانخفاض في قيم المعلم الثاني إلى التأثير في قيم المعلمين الأول والثاني لصوت "الفاء /f/ " التالي للفتحة؛ حيث أثرت هذه الفتحة القصيرة علي الصامت المجاورة فانخفضت بداية المعلم الأول للصوت اللاحق من 1072 د/ث إلي 975 د/ث، كما أثرت الفتحة القصيرة علي بداية المعلم الثاني للصوت اللاحق فانخفض من 2187 د/ث إلى 2083 د/ث(10).

**الفتحة القصيرة المفخمة /A/ :**

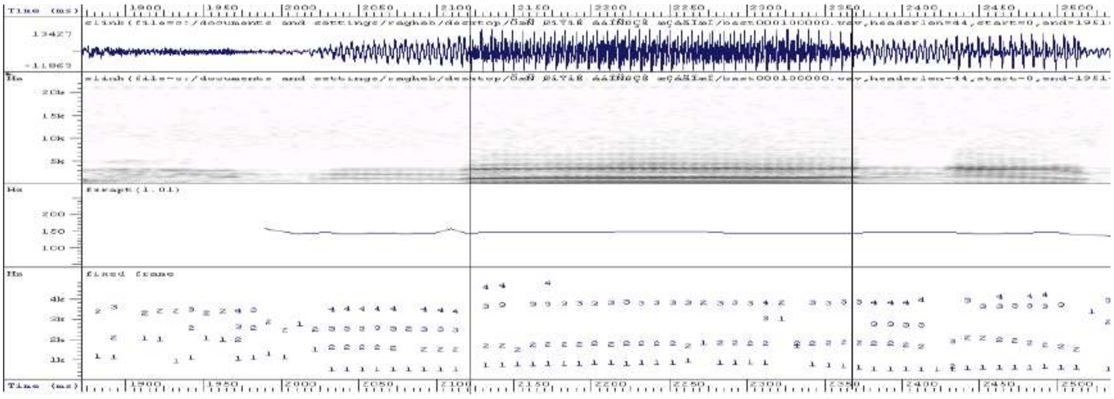


شكل 2 يعرض صورة طيفية للفتحة القصيرة المفخمة /A/ من خلال كلمة **الرحمن** التي وردت في قوله تعالى: ﴿بسم الله الرحمن الرحيم﴾ [الفاتحة/١].

(9) د. عمر، أحمد مختار، دراسة الصوت اللغوي، ص: 34.

(10)هذا نموذج للتحليل الطيفي لصوت الفتحة القصيرة المرققة، وسيتم تنفيذه على باقي الحركات في حال اعتماده من قِبل الأستاذ الدكتور/ هويدي شعبان إن شاء الله.

### الفتحة الطويلة المرققة /a2/ :



شكل 3 يعرض صورة طيفية للفتحة القصيرة المفخمة /A/ من خلال كلمة الرحمن التي وردت في قوله تعالى: ﭽبسم الله الرحمن الرحيمﭼ [الفاتحة/١].

ومن خلال رصد المدى الزمني للمدود في الصور الطيفية السابقة وفي الجداول الملحقة يمكنني مناقشة قضية الحركة والحركتين وعلاقتها بزمن المد التي تناولها علماء التجويد على النحو التالي:

### زمن المد:

غلب على علماء التجويد تحديد زمن المد بالحركات، فإذا استوفى حرف المد نصيبه من المد انتقل بذلك من الحركة إلى الحرف، وهذه الخاصية ثابتة لحروف المد دون غيرها من الأصوات الجامدة "لا سيما الشديدة –الانفجارية– فإنها آنية الحدوث، وكذلك الرخوة –الاحتكاكية– فإنها وإن كانت زمانية يمتد بها الصوت مدة، لكن ذلك الامتداد لا يبلغ مقدار ألف، أي مقدار نطق حرف المد"(11).

وقد عقد الأستاذ الدكتور **غانم قدوري الحمد** مبحثا خاصًا بالمدود في كتابه القيم **الدراسات الصوتية عند علماء التجويد** ذكر فيه أقوال علماء التجويد التي تباينت كثيرًا في تقدير زمن المد فذكر أن مقادير المد تكاد "تنحصر بين المد مقدار ألفين، أي ضعف المد الطبيعي، وبين المد مقدار خمس ألفات، وبين ذلك مراتب من المد بحسب مذهب القراء، وبحسب نوع المد ومكانه، وبحسب أسلوب القراءة من الحدر والتحقيق"(12).

ثم ذكر أن علماء التجويد قد حاولوا ابتكار وسائل لقياس مقادير المد وضبطها "فالقول أن مقدار المد ألف أو ألفان مثلًا لا يكفي لبيان الزمن الذي يحتاجه نطق المد، فلابد من إيجاد وسيلة تساعد في ضبط زمن نطق الوحدة المستعملة في قياس طول المد وهي الألف، أي زمن طق صوت الألف"(13).

ثم جمع نتائج دراسته لأقوال علماء التجويد في مسألة قياس وضبط زمن المد في وجود خمسة طرق "لقياس زمن نطق الألف الذي اتخذه علماء التجويد أساسًا لقياس مقادير المدود، وتلك الطرق هي:

1. أن تقول آ مرة أو مرتين أو أكثر، كل مرة تساوي نطق ألف.
2. العقد بالأصابع، ولعل معناه الطرق بأي من الأصابع على الإبهام، كل طرقة تقابل نطق ألف.

(11) د. الحمد، غانم قدوري، الدراسات الصوتية عند علماء التجويد، ص: 536.

(12) السابق، ص: 539.

(13) السابق، ص: 540.

3. أن تعد عددًا، فتقول: واحد، اثنان، ثلاثة..إلخ. وقد انفرد بذكر هذه الطريقة طاش كبرى زاده، وهو موضع نظر، لأن كل واحد من الأعداد المذكورة يتضمن صوت الألف إلى جانب أصوات أخرى، فكل كلمة تعادل في النطق أكثر من ألف.

4. أن تمد صوتك بقدر قولك: ألف ألف.

5. أو كتابتها، أي كتابة **ا** وليس كتابة **ألف** فيما نرجح، وانفرد علي القاري بذكر هاتين الطريقتين"(14).

والحق أن كل هذه الطرق المذكورة لا تصمد ولو للحظات أمام البحث الموضوعي، بل هي في أغلب الأحيان حجة من لا يملك تعليلا، أو تعليل من لا يملك حجة. وقد شعر بذلك الأستاذ الدكتور **غانم قدوري الحمد** نفسه، فختم حديثه عن هذه المسألة – مسألة مقادير المدود– بما يشير من طرف خفي إلي عدم قناعته بكل تلك الطرق التي تبدو غير موضوعية، لكنه وجد لنفسه عذرًا في اعتمادها في بحثه لتعذر حصوله علي أجهزة دقيقة لرصد مقادير المدود "وإذا كان استخدام أجهزة القياس الدقيقة في ضبط مقادير المدود غير متيسر الآن، فإن الطرق السابقة التي ذكرها علماء التجويد تظل صالحة للاستخدام حتى يتيسر استخدام طرق أكثر دقة وتحديدًا لقياس مقادير المدود"(15).

والحق أنني قصدت عمدًا الإطناب في نقل جهود الأستاذ الدكتور غانم قدوري الحمد في هذه المسألة لأؤكد أنني هنا لن أحاول مجاراته في جمع أقوال من هنا وهناك تبين القيمة الزمنية للمد، ولن أحلل مقولات لبعض علماء التجويد أو علماء الأصوات، لكنني سأبدأ من حيث انتهي؛ لأن العذر الذي وجده لنفسه والذي نقله نصًا من كتاب أستاذنا الدكتور إبراهيم أنيس(16) لم يعد بإمكاني أن أحتمي خلفه أو أستتر من ورائه، وعليه فقد شرعت في تحليل زمن المد في الآيات عينة الدراسة، حيث قمت بعرض الملفات الصوتية المذكورة علي تقنية برنامج HTK والذي قام بدوره بتحديد الأزمان التي استغرقها كل فونيم ورد في قاعدة البيانات بالميلي ثانية، ثم قمت بمراجعتها واعتمدها ضمن المواصفات التقنية لشركة RDI ، **وعن طريق رصد نتائج تلك الجداول يمكن ترتيب الحركات تصاعديًا على النحو التالي:**

| م | رمز الفونيم | م الزمن | م | رمز الفونيم | م الزمن |
|---|---|---|---|---|---|
| 1 | i | 0.148 | 9 | i4 | 0.842 |
| 2 | a | 0.151 | 10 | u4 | 0.877 |
| 3 | A | 0.152 | 11 | A4 | 1.238 |
| 4 | u | 0.157 | 12 | a4 | 1.451 |
| 5 | i2 | 0.392 | 13 | i6 | 1.764 |
| 6 | a2 | 0.41 | 14 | a6 | 2.302 |
| 7 | u2 | 0.416 | 15 | A6 | 2.43 |
| 8 | A2 | 0.436 | | | |

جدول 1 يبين متوسط أزمان المدود المختلفة التي جمعت في قاعدة بيانات الدراسة

ولا تقتصر نتيجة هذا الجدول على إدراك زمن المد بأشكاله المختلفة بل تتعداه إلى إبراز نتيجة هامة دار حولها الخلاف قديمًا على النحو التالي:

(14) السابق، ص: 541.

(15) السابق، ص: 541.

(16) انظر: د. أنيس، إبراهيم، الأصوات اللغوية، ص: 159.

**تفاوت مقادير المدود:**

**جدول متوسط مقادير الحركات والمدود حسب الجهر والهمس:**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a_Hams | 0.144 | a2_Hams | 0.379 | a4_Hams | 0.672 | a6_Hams | |
| a_Gahr | 0.162 | a2_Gahr | 0.404 | a4_Gahr | 1.14 | a6_Gahr | 2.371 |
| | | | | | | | |
| i_Hams | 0.138 | i2_Hams | 0.352 | i4_Hams | 0.453 | i6_Hams | |
| i_Gahr | 0.156 | i2_Gahr | 0.389 | i4_Gahr | 0.827 | i6_Gahr | 1.764 |
| | | | | | | | |
| u_Hams | 0.142 | u2_Hams | 0.367 | u4_Hams | 0.689 | u6_Hams | |
| u_Gahr | 0.166 | u2_Gahr | 0.384 | u4_Gahr | 0.877 | u6_Gahr | |
| | | | | | مد<br>عارض | | |

جدول 2 يوضح متوسط مقادير الحركات والمدود حسب الجهر والهمس حيث:

1. **a_Hams** تعني الفتحة القصيرة المسبوقة بصوت مهموس
2. **a_ Gahr** تعني الفتحة القصيرة المسبوقة بصوت مجهور
3. **i_Hams** تعني الكسرة القصيرة المسبوقة بصوت مهموس
4. **i_ Gahr** تعني الكسرة القصيرة المسبوقة بصوت مجهور
5. **u_Hams** تعني الضمة القصيرة المسبوقة بصوت مهموس
6. **u_ Gahr** تعني الضمة القصيرة المسبوقة بصوت مجهور
7. الأرقام الملازمة لرموز الحركات تعني قيمة هذه الحركات ((حركتان، أربع حركات، ست حركات)).

ويظهر الجدول بصفة عامة زيادة قيم الحركات والمدود الزمنية إذا أتبعت بصامت مجهور عن مثيلاتها المتبوعة بصامت مهموس ويمكن توضيح ذلك من خلال الثنائيات التالية:

- بلغ متوسط الفتحة القصيرة المسبوقة بصوت مهموس 144 ميللي ثانية، بينما بلغ متوسط الفتحة القصيرة المسبوقة بصوت مجهور 162 ميللي ثانية.
- بلغ متوسط الكسرة القصيرة المسبوقة بصوت مهموس 138 ميللي ثانية، بينما بلغ متوسط الكسرة القصيرة المسبوقة بصوت مجهور 156 ميللي ثانية.
- بلغ متوسط الضمة القصيرة المسبوقة بصوت مهموس 142 ميللي ثانية، بينما بلغ متوسط الضمة القصيرة المسبوقة بصوت مجهور 166 ميللي ثانية.
- بلغ متوسط الفتحة الطويلة –حركتان– المسبوقة بصوت مهموس 379 ميللي ثانية، بينما بلغ متوسط الفتحة الطويلة –حركتان– المسبوقة بصوت مجهور 404 ميللي ثانية.
- بلغ متوسط الفتحة الطويلة –أربع حركات– المسبوقة بصوت مهموس 672 ميللي ثانية، بينما بلغ متوسط الفتحة الطويلة – أربع حركات– المسبوقة بصوت مجهور 1014 ميللي ثانية.
- خلت أصوات العينة من صوت الفتحة الطويلة –ست حركات– المسبوقة بصوت مهموس، بينما بلغ متوسط الفتحة الطويلة – ست حركات– المسبوقة بصوت مجهور 2371 ميللي ثانية.

- بلغ متوسط الكسرة الطويلة –حركتان– المسبوقة بصوت مهموس 353 ميللي ثانية، بينما بلغ متوسط الفتحة الطويلة –حركتان– المسبوقة بصوت مجهور 389 ميللي ثانية.
- بلغ متوسط الكسرة الطويلة –أربع حركات– المسبوقة بصوت مهموس 453 ميللي ثانية، بينما بلغ متوسط الفتحة الطويلة – أربع حركات– المسبوقة بصوت مجهور 827 ميللي ثانية.
- خلت أصوات العينة من صوت الكسرة الطويلة –ست حركات– المسبوقة بصوت مهموس، بينما بلغ متوسط الفتحة الطويلة – ست حركات– المسبوقة بصوت مجهور 1764 ميللي ثانية.
- بلغ متوسط الضمة الطويلة –حركتان– المسبوقة بصوت مهموس 367 ميللي ثانية، بينما بلغ متوسط الفتحة الطويلة –حركتان– المسبوقة بصوت مجهور 384 ميللي ثانية.
- بلغ متوسط الضمة الطويلة –أربع حركات– المسبوقة بصوت مهموس 689 ميللي ثانية، بينما بلغ متوسط الفتحة الطويلة – أربع حركات– المسبوقة بصوت مجهور 877 ميللي ثانية.
- خلت أصوات العينة من صوت الضمة الطويلة –ست حركات– المسبوقة بصوت مهموس أو مجهور.

**المد المتصل:**

| نوع المد | زمن المد |
|---|---|
| **A4-@** | **1.475** |
| **a4-@** | **1.538** |
| **u4-@** | **1.480** |

جدول 3 يوضح متوسط مقادير المد المتصل حيث:

A4-@ تعني مد متصل مفتوح مفخم، و a4-@ تعني مد متصل مفتوح مرقق، وu4-@ مد متصل مضموم.

**ترتيب مدى الحركات زمنيًا:**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 0.152 | A2 | 0.436 | A4 | 1.238 | A6 | 2.43 |
| a | 0.151 | a2 | 0.41 | a4 | 1.451 | a6 | 2.302 |
| i | 0.148 | i2 | 0.392 | i4 | 0.842 | i6 | 1.764 |
| U | 0.157 | u2 | 0.416 | u4 | 0.877 | | |

جدول 4 يوضح ترتيب مدى الحركات زمنيًا

**مد البدل:**

| | |
|---|---|
| **@-a2** | **0.400** |
| **@-i2** | **0.400** |
| **@-u2** | **0.380** |

جدول 5 يوضح متوسط قيم مد البدل في القرآن الكريم ويظهر الجدول ما يلي:

1. ترتيب أنواع مد البدل تصاعديًا –المتبوع بضم فالمتبوع بفتح ثم المتبوع بكسر–.
2. تقارب متوسطات مد البدل مع متوسطات المد الطبيعي (410، 416، 392).

### نتائج الجداول:

1. زيادة المدة الزمنية للحركات المتبوعة بصوت مجهور عن مثيلتها المتبوعة بصوت مهموس، سواء أكانت الحركة طويلة أم قصيرة، وسواء أكانت مفتوحة أم مكسورة أم مضمومة.
2. اختلاف المدى الزمني للمد العارض للسكون عن المد الطبيعي وعن المد المتصل والمنفصل، وذلك بسبب عامل النبر الذي أدى إلى زيادة المدة الزمنية للمد العارض للسكون حتي كادت أن تصل إلى ضعف المد الطبيعي مع الأخذ في الاعتبار أن المشايخ عينة الدراسة التزموا جميعًا بوجه قصره على حركتين.
3. الجدول الثاني يوضح متوسط قيم الحركات القصيرة المتبوعة بهمزة والتي ظهرت بقيم أكبر من من قيم الحركات المتبوعة بصوت مهموس وأقل من قيم الحركات المتبوعة بصوت مجهور، وهذا دليل على عدم انتماء صوت الهمزة إلى أي من القسمين.
4. خلت العينة من المد الطبيعي المتبوع بالهمزة وذلك لأن الهمزة سبب من أسباب المد الفرعي.
5. الجدول الثالث يوضح متوسط قيم الحركات القصيرة المفتوحة المرققة والتي ظهرت بقيم أصغر من قيم الحركات المفتوحة المفخمة.
6. كما أظهر الجدول الثالث أيضًا ترتيب قيم الحركات المفتوحة والمكسورة والمضمومة.
7. أظهر الجدول الأخير أن متوسطات مد البدل جاءت مقاربة لمتوسطات المد الطبيعي.

### نتائج الدراسة

وختامًا فقد حاولت هذه الدراسة معالجة دور المؤثرات السياقية في تقدير المدى الزمني للفونيم حال النطق بأصوات القرآن الكريم معالجة حاسوبية، وقد خلصت إلى مجموعة من النتائج يمكن إبرازها كالتالي:

1. اعتمدت اللغة العربية بشكل أساسي على الصوائت، حيث وردت 210596 مرة، وشكلت نسبة 42.13% من قيمة الأصوات العربية
2. تقارب متوسطات مد البدل مع متوسطات المد الطبيعي (410، 416، 392).
3. زيادة المدة الزمنية للحركات المتبوعة بصوت مجهور عن مثيلتها المتبوعة بصوت مهموس، سواء أكانت الحركة طويلة أم قصيرة، وسواء أكانت مفتوحة أم مكسورة أم مضمومة.
4. اختلاف المدى الزمني للمد العارض للسكون عن المد الطبيعي وعن المد المتصل والمنفصل، وذلك بسبب عامل النبر الذي أدى إلى زيادة المدة الزمنية للمد العارض للسكون حتي كادت أن تصل إلى ضعف المد الطبيعي مع الأخذ في الاعتبار أن المشايخ عينة الدراسة التزموا جميعًا بوجه قصره على حركتين.
5. الجدول الثاني يوضح متوسط قيم الحركات القصيرة المتبوعة بهمزة والتي ظهرت بقيم أكبر من من قيم الحركات المتبوعة بصوت مهموس وأقل من قيم الحركات المتبوعة بصوت مجهور، وهذا دليل على عدم انتماء صوت الهمزة إلى أي من القسمين.
6. متوسط قيم الحركات القصيرة المفتوحة المرققة ظهرت بقيم أصغر من قيم الحركات القصيرة المفتوحة المفخمة.

# ثبت المراجع

1. **أبو الفضل محمد بن مكرم ابن منظور:** لسان العرب، مطبعة بولاق، ط1.
2. **أبو بشر عمرو بن عثمان بن قنبر سيبويه:** الكتاب، تحقيق: عبد السلام محمد هارون، دار الجيل، بيروت، ط1.
3. **أحمد راغب أحمد:** فونولوجيا القرآن "دراسة لأحكام التجويد في ضوء علم الأصوات الحديث"، رسالة ماجستير، كلية الآداب، جامعة عين شمس، 2004م.
4. **أحمد محمد قدور:** أصالة علم الأصوات عند الخليل من خلال مقدمة كتاب العين، دار الفكر المعاصر، بيروت– لبنان، ط1، 1419هـ/1998م.
5. **أحمد مختار عمر:** دراسة الصوت اللغوي، عالم الكتب، 2000م.
6. **إبراهيم أنيس:** الأصوات اللغوية، مكتبة الأنجلو، ط 5 / 1975م.
7. **إبراهيم ضوة:** محاضرات في اللغة العربية والحاسب، دار الثقافة العربية، ط1، 2000م.
8. **أرنست بولجرام:** في علم الأصوات الفيزيقي، ترجمة د. سعد مصلوح، ط1، 1977م.
9. **تمام حسان:**
   - مناهج البحث في اللغة، دار الثقافة، الدار البيضاء، 1394م، 1974م، ط2.
   - اللغة العربية معناها ومبناها، الهيئة المصرية العامة للكتاب، ط2 1979م.
10. **جاسم علي جاسم**، وزيدان علي جاسم،: نظرية علم اللغة التقابلي في التراث العربي، مجلة الآداب الأجنبية، اتحاد الكتاب العرب، دمشق، النسخة الإلكترونية.
11. **جون ليونز:** اللغة وعلم اللغة، ترجمة مصطفى زكي التوني، دار النهضة العربية، 1988م.
12. **ديفيد كريستال:** التعريف بعلم اللغة، ترجمة د. حلمي خليل، الهيئة المصرية العامة للكتاب، 1979م.
13. **ريمون الطحان:** الألسنية العربية، ط2، بيروت 1981.
14. **سعد مصلوح**، دراسة السمع والكلام، القاهرة، 1980م.
15. **سلمان حسن العاني:** فونولوجيا العربية، ترجمة ياسر الملاح، مطبوعات النادي الأدبي الثقافي بجدة، ط1، 1403هـ – 1983م.
16. **صبحي عبد الحميد عبد الكريم:** النون وأحوالها في لغة العرب، مطبعة الأمانة، 1986م.
17. **عبد الحميد محمد أبو سكين:** دراسات في التجويد والأصوات اللغوية، مطبعة الأمانة، القاهرة، 1404هـ/ 1983م.
18. **عبد الرحمن أيوب:** الكلام إنتاجه وتحليله، ط جامعة الكويت، 1984 م.
19. **عبد الرحمن بن إسماعيل أبو شامة المقدسي:** إبراز المعاني من حرز الأماني، تحقيق غانم قدوري الحمد.
20. **عبد الغفار حامد هلال:**
   - أصوات اللغة العربية، مكتبة الأنجلو المصرية، ط2، 1988م.
   - تجويد القرآن الكريم من منظور علم الأصوات الحديث، مكتبة الآداب، ط1، 2007م.
   - أبنية العربية في ضوء علم التشكيل الصوتي، ط1، دار المحمدية للطباعة 1979م.
21. **عبد الصبور شاهين:**
   - المنهج الصوتي للبنية العربية، القاهرة 1977م.
   - القراءات القرآنية في ضوء علم اللغة الحديث، دار القلم، 1966م.
22. **عوض المرسي جهاوي:** ظاهرة التنوين في اللغة العربية، مكتبة الخانجي بالقاهرة، ط1، عام 1403هـ.
23. **غانم قدوري الحمد:** أبحاث في علم التجويد، دار عمار للنشر والتوزيع، الأردن.
24. **فؤاد سزكين:** تاريخ التراث العربية، ترجمة عرفة مصطفى، مراجعة مازن عماوي، مطبوعات جامعة محمد بن سعود ، الرياض، 1988م.
25. **كمال محمد بشر:** علم اللغة العام (الأصوات)، طبعة دار غريب، ط2، 1971م.

26. **مازن الوعر:** صلة التراث اللغوي العربي باللسانيات، النسخة الإلكترونية من مجلة التراث العربي– مجلة فصلية تصدر عن اتحاد الكتاب العرب – دمشق العدد 48 – السنة 12 – تموز "يوليو" 1992.

27. **محمد بن أبي بكر الرازي:** مختار الصحاح، مكتبة لبنان، 1989م.

28. **محمد بن محمد بن محمد ابن الجزري:** التمهيد في علم التجويد، ط1، القاهرة، 1908م.

29. **محمد حسن عبد العزيز:** مصادر البحث اللغوي، دار الثقافة العربية، 2000م.

30. **محمد صالح الضالع:** التجويد القرآني "دراسة صوتية فيزيائية"، دار غريب 2002 م.

31. **محمد صلاح الدين بكر:** الوصفيَّة في الدراسات العربيَّة القديمة والحديثة، مجلة التراث.

32. **محمد علي الخولي:** الأصوات اللغوية، مكتبة الخريجي، الطبعة الأولى 1987م .

33. **محمد فتيح:** الأصوات العامة والأصوات العربية، دار الثقافة العربية، القاهرة.

34. **محمد مكي نصر:** نهاية القول المفيد في علم التجويد، مكتبة الحلبي، 1349هـ.

35. **محمد يوسف حبلص:** مقدمة في علم اللغة، دار الثقافة العربية، 1997.

36. **محمود السعران:** علم اللغة مقدمة للقارئ العربي، دار المعارف بمصر 1962م.

37. **محمود خليل الحصري:** أحكام قراءة القرآن الكريم، مكتبة السنة، ط1، 2000م.

38. **مصطفى زكي التوني:** النون في اللغة العربية "دراسة لغوية في ضوء القرآن الكريم"، حوليات كلية الآداب جامعة الكويت، الحولية السابعة عشرة، 1416–1417هـ، 1996– 1997م.

39. **مناف الموسوي:** علم الأصوات اللغوية، عالم الكتب، بيروت، ط1، 1998م.

40. **منصور بن محمد الغامدي:** الصوتيات العربية، مكتبة التوبة، ط1، 2000م.

41. **نعيم عبد البَاقي:** قواعد تشكل النغَم في مُوسيقى القرآن، مجلة التراث العربي، النسخة الإلكترونية، العدد25، "أكتوبر"1986م.

42. **يحيى بن علي المباركي:** الكم الزمني لصويت الغنة في الأداء القرآني، دوريات جامعة أم القرى، النسخة الإلكترونية.

# Metaphor Interpretation Mechanisms

**Farzaneh Salehi**
*Najaf Abad University, Isfahan, Irn*
*farzaneh2006_S_K@yahoo.com*

**Gholam Reza Zarei**
*Isfahan University of Technology*
*grzarei@cc.iut.a.ir*

## Abstract

Understanding and comprehension of English texts, especially literary ones, has always been problematic for foreign language learners. Metaphors, as difficult devices, create major comprehension problems for these learners. This study aims at discovering the most current mechanisms employed by the foreign language learners while engaged in the process of interpreting metaphors taken from the domains of: ‘land mammals’, ‘birds’ and ‘sea creatures’, the subparts of the larger, inclusive domain of ‘animals’, together with those from inanimate domains of ‘vehicles’ and ‘the things that fly’. The unconventionality of these domains lead to the instigation of the subjects’ creativity and thus the emergence of novel features. The results of self-reporting technique suggest that reliance on the appearance, manner of behaving/doing action or the behavior or action itself, as well as focusing on the intuition, semantic features, experience and social culture constitute the mechanisms applied by the foreign language learners. The findings of this study, if properly applied, can greatly affect the whole process of learning/teaching metaphors.

# Metaphor Interpretation Mechanisms

**Farzaneh Salehi**
*Najaf Abad University, Isfahan, Irn*
*farzaneh2006_S_K@yahoo.com*

**Gholam Reza Zarei**
*Isfahan University of Technology*
*grzarei@cc.iut.a.ir*

## Abstract

Understanding and comprehension of English texts, especially literary ones, has always been problematic for foreign language learners. Metaphors, as difficult devices, create major comprehension problems for these learners. This study aims at discovering the most current mechanisms employed by the foreign language learners while engaged in the process of interpreting metaphors taken from the domains of: ‘land mammals’, ‘birds’ and ‘sea creatures’, the subparts of the larger, inclusive domain of ‘animals’, together with those from inanimate domains of ‘vehicles’ and ‘the things that fly’. The unconventionality of these domains lead to the instigation of the subjects’ creativity and thus the emergence of novel features. The results of self-reporting technique suggest that reliance on the appearance, manner of behaving/doing action or the behavior or action itself, as well as focusing on the intuition, semantic features, experience and social culture constitute the mechanisms applied by the foreign language learners. The findings of this study, if properly applied, can greatly affect the whole process of learning/teaching metaphors.

# LEARNING GENERATOR: NEURO-LINGUISTIC PROGRAMMING AND LEARNING STYLES IN ENGLISH TEXT BOOKS

Dr. Mrs. EVA ZANUY PASCUAL
PhD on Linguistics
Universidad Nacional de Educación a Distancia
Spain
evazanuy@hotmail.com

## Abstract

The study of learning styles tries to explain the individual differences in the way people use their cognitive resources, learning best depending on the way they both perceive and process information. After analysing the proportion of the learning styles in the text books used to learn English as a second language, changes should be made to improve the quality of education of books as they only benefit a small percentage of students. The excessive representation of exercises that benefit the Learning Style with less students and the small representation of exercises which benefit students with majority styles demonstrate that text books follow a mistaken tendency. The higher representation of exercises that benefits the Reflector Style demonstrates that all the editorials, without exception, follow the natural method. The natural method fails because it has an excess of representation of a single Learning Style, which is the one with the smallest representation among the students (Reflector Style). Text books act like a tool that generates learning, and if we perfected it considering the different Learning Styles, we would be creating a Learning Generator: an optimal tool of learning.

## Keywords

Neuro-Linguistic Programming, Teaching and Learning Styles, Text books

### 1. COGNITIVE STYLES

Cognitive styles refer to the preferred way an individual processes information. Unlike individual differences in abilities (e.g., Gardner, Guilford, Sternberg) which describe peak performance, styles describe a person's typical mode of thinking, remembering or problem solving.

Styles are usually considered to be bipolar dimensions whereas abilities are unipolar (ranging from zero to a maximum value). Having more of an ability is usually considered beneficial while having a particular cognitive style simply denotes a tendency to behave in a certain manner. Cognitive style is a usually described as a personality dimension which influences attitudes, values, and social interaction.

A large number of cognitive styles have been identified and studied in the past. Field independence versus field dependence is probably the most well known style.

It refers to a tendency to approach the environment in an analytical, as opposed to global, fashion. At a perceptual level, field independent personalities are able to distinguish figures as discrete from their backgrounds compared to field dependent individuals who experience events in an undifferentiated way. In addition, field dependent individuals have a greater social orientation relative to field independent personalities. Studies have identified a number connections between this cognitive style and learning. For example, field independent individuals are likely to learn more effectively under conditions of intrinsic motivation (e.g., self-study) and are influenced less by social reinforcement.

Individuals learn best in many different ways, sometimes using a variety of learning styles, but teachers and trainers may not always present information and learning experiences in the ways that best suit you. Forms of learning through workshops, practical activities or through informal methods may suit some people more than others. Sometimes, people feel they are not good at learning when it may be just that they don't know their own learning styles.

### 1.1. NEURO-LINGUISTIC PROGRAMMING

Whether you realise it or not, we all have preferences for how we absorb information, analyse it and make decisions. Some people like to see what you mean and make decisions based on how things look. Some people like to hear your ideas and decide based on what they sound like. Some people like to experience what you are talking about and decide by how things feel to them.

The Human Brain is an incredible computer. However, it does not come with an instruction manual! Understanding your personal learning style can improve your own learning process, and assist you in communicating with other people.

**Visual system.**

- Need to see it to know it.
- Have strong sense of colour.
- May have artistic ability.
- Often have difficulty with spoken directions.
- Might over-react to sounds.
- Might have trouble following lectures.
- Often misinterprets words.

**Auditory system.**

- Prefer to get information by listening-needs to hear it to know it.
- May have difficulty following written directions.
- Difficulty with reading.
- Problems with writing.
- Inability to read body language and facial expressions.

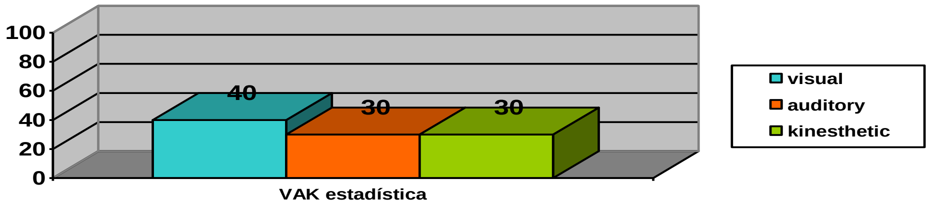
**Kinesthetic system.**

- Prefer hands-on learning.
- Often can assemble parts

without reading directions.
• Have difficulty sitting still.
• Learn better when physical activity is involved.
• May be very well coordinated and have athletic ability.

### 1.1.4. Percentage of the representational systems

According to Pau Cazau (2002):
40% of the people are visual
30% auditory
30% kinesthetic



Graphic 1. Percentage of representational systems according to Pau Cazau

## 1.2. LEARNING STYLES

Honey and Mumford postulate that people prefer different methods of learning, depending upon the situation and their experience level, thus they move between the four modes of learning, rather than being dominantly locked into one mode.
Honey and Munford's learning cycle also slightly differs from Kolb's:
1. Having an experience
2. Reflecting on it
3. Drawing their own conclusions (theorizing)
4. Putting their theory into practice to see what happens

**Activists (Do)**
Immerse themselves fully in new experiences
Enjoy here and now
Open minded, enthusiastic, flexible
Act first, consider consequences later
Seek to centre activity around themselves

**Reflectors (Review)**
Stand back and observe
Cautious, take a back seat
Collect and analyze data about experience and events, slow to reach conclusions
Use information from past, present and immediate observations to maintain a big

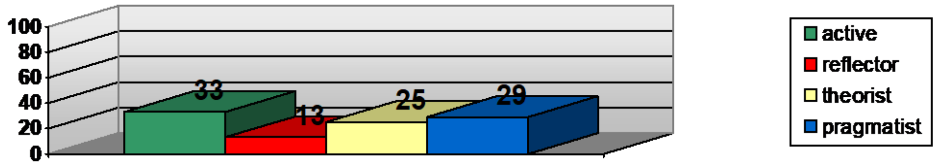
picture perspective.

**Theorists (Conclude)**
Think through problems in a logical manner, value rationality and objectivity
Assimilate disparate facts into coherent theories
Disciplined, aiming to fit things into rational order
Keen on basic assumptions, principles, theories, models and systems thinking

**Pragmatists (Plan)**
Keen to put ideas, theories and techniques into practice
Search new ideas and experiment
Act quickly and confidently on ideas, gets straight to the point
Are impatient with endless discussion

### 1.2.5. Percentage of Learning Styles

According to Honey-Alonso the percentage of  Learning Styles are:
Activist; 33 %
Reflector; 13 %
Theorist; 25 %
Pragmatist; 29 %



Graphic 2. Percentage of  Learning Styles

## 2. GENERAL ANALYSIS OF EDITORIALS

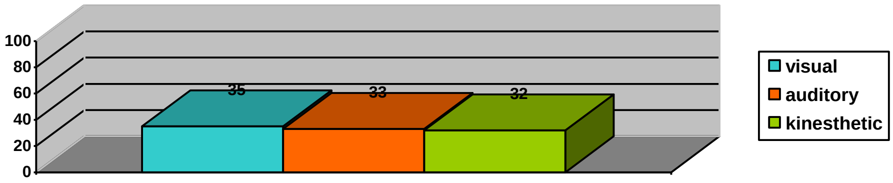
Not only do students have their preferences and their style of learning. All teachers have their own style when giving class, and that style is also reflected when we use the different representational systems. Most of us tend to use a system more often than the others when we teach. In order to detect what our tendencies are, we need to analyse our way of teaching from the point of view of the NLP. Generally, in all the groups of students we will find different types of learning styles. If our teaching style is the same as that of our students, learning will be easier for them than if it is not the same one, and with a book using all the different styles we will be benefiting all our students.

### 2.1. Analysis according to the Neuro-Linguistic Programming

Each editorial has common characteristics and differential characteristics. One of the main common characteristics that has been found after analysing different editorials belonging to the same level is the great numerical equality of exercises that benefit the different systems of neurolinguistic representations. The neurolingistic representation in the editorials would be;

Visual; 35 %

Auditory; 33 %
Kinesthetic; 32 %

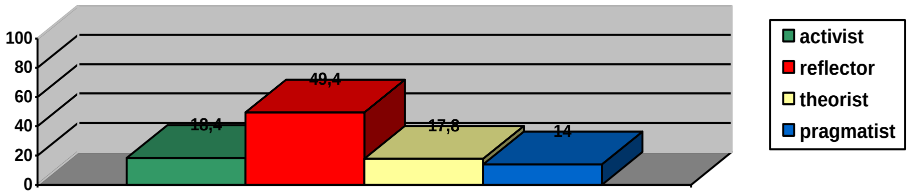


Graphic 3. Percentage of NLP in the editorials

The Oxford publishing house turns out to be the one that benefits more the students with predominance in Visual style (50 %). Cambridge is second (41.5 %), Pearson occupies third (38.5 %), whereas Heinemann (24.1 %) and Richmond (20.8 %) includes a smaller representation of exercises that benefit this group of students. The Visual style is the one that has the greatest representation in three of five editorials, although not by much from the second predominant style, the Auditory style. The one with the greatest percentage is Heinemann (44.1 %), followed by Richmond (40.8 %) and Cambridge (30.5 %). Those that have a smaller percentage are Oxford (26 %) and Pearson (24.2 %). The Kinesthetic style is the least represented in two of five editorials although not by a remarkable amount from the other representational systems, and varies between the greatest representation of Richmond (38.4 %) and the representations of Pearson (37.3 %), Heinemann (31.8 %), Cambridge (28 %) and Oxford (24 %). This analysis demonstrates that the books of the most sold and used editorials are near being learning generators. The percentage of visual children habitually is very superior to the auditory and Kinesthetic children, for that reason many activities are prepared for these children.

## 2.2. Analysis according to the Learning Styles

One of the main common characteristics after analysing different editorials on the same level is the great representation of exercises that a Learning Style has over other Styles. The average representation in percentages of the Learning Styles in the analysed editorials would be;
Activist; 18.4 %
Reflector; 49.4 %
Theorist; 17.8 %
Pragmatist; 14 %



Graphic 4. Percentage of Learning Styles in the editorials

The Reflector Style, with a representation of 49.4 %, is the Style which all editorials benefit to. This data is common in all the analysed editorials. The Activist Style occupies second position if we consider the average, with an 18,4 %

representation, but it has only been the second more represented Style in three of the five editorials. The third most represented Style is the Theorist Style, with 17,8 %, that is also the second most represented Style in three of the five analysed editorials. The Pragmatist Style, with a representation of 14 %, has been the least represented Style in three of the five editorials, and it is, the Style with the smallest representation in general.

The Richmond publishing house turns out to be the one that most benefits the students with predominance in Activist Style (30 %). The Pearson publishing house is second (23 %) and Cambridge and Heinemann occupy third (17 %), whereas Oxford has the smallest representation of exercises that benefit this group of students.

The Reflector Style is the one that has the greatest representation in all the editorials, and with a clear advantage in percentage from the second predominant Style. The publishing houses with the highest percentage (56 %) are Oxford, and on the other hand, Heinemann is the one that has the lowest percentage (43 %). As it can be verified, the highest score and the lowest do not distant to a great extent. Heinemann is also the publishing house with the greatest percentage in representation of exercises with Theorist Style (29 %). Oxford occupies the second position (22 %). Cambridge (17 %) and Pearson (14 %) occupy the following positions and Richmond has the lowest percentage (7 %). The Pragmatist Style is the least represented style and varies between Pearson and Heinemann (11 %) and Oxford, Cambridge and Richmond (16 %).

## 3. THE IMPROVEMENT OF THE LEARNING STYLES

### Grammar

A- For Activist students; to make theoric questions, to solve to problems in small groups, to make representations, competitions so that they interact with other companions

B- For Theorist students; to face systems and concepts that present/display a challenge, grammar competitions, to remind them that the activities that they are doing serve to reach concrete goals, to elaborate a notebook of schemes and exercises, to explain the theory or to summarize concepts after the end of the class

C- For the improvement of the Pragmatist students; representations of dialogues that work on a concrete grammar structure

### Speaking

A- For Activist students; activities that present/display a challenge, relative brevity and immediate result activities. We must try emotion, drama and crisis with them. Begin the class announcing that a variety of activities will be done and what new things will be learned. Parallelly, we should avoid these students adopting a passive role, so as to analyse and process data, not working alone. Proposed activities; to draw faces on their fingers so that they speak, the corner of humour, to create a time for humour, to compete in teams and scoring the results and having debates

B- Theorist students learn better if they are taught with general rules, so using posters with general indications makes it easier for them to have conversation activities

C- Pragmatist students learn better with useful activities and things they need in their daily tasks. Proposed activities; to study structures, vocabulary, etc. related to the daily life, to repeat phrases and structures after the teachers, to have a model to imitate, to mark real and daily situations on a map, to have role plays, to practice things with clear useful advantages

**Reading comprehension**

A- Activist students; Reading a second language can also be worked with activities that benefit the Activist students, as the following ones: treasure hunting, reading instructions so as to get a treasure or prize, the newspaper: to elaborate a newspaper with different articles written by groups of students

B- For the improvement of the reading comprehension for students with predominance in Theorist Style we could use the following activities; filling the blanks: to fill in the blanks in a text, cross out the extra word: words have been added to a text and the student must eliminate the extra words, alter texts: to change the order of paragraphs or words so that the student orders them correctly.

C- Pragmatist students; the activities for the reading comprehension in a second language for students of this Style are: treasure hunting; the students divided into groups have to find missing objects, or follow the instructions and the winning group is the one that obtains the object in less time, to make up a comic; after creating a comic strip and is corrected by the teacher, comics are distributed to other students so that they are read by their classmates and this activity also serves to work in the writing, to sail in the network; Internet offers a multitude of possibilities to work on reading comprehension in English.

**Listening comprehension**

A- So as the Activist students understand English we can use the following activities; to have debates, to practice the initiation of conversations with simulated strangers, to take part actively and to compete in equipment; the game of Bingo is an example: to adapt the classic game to the necessities of the class, that is to say, with numbers, objects, offices, foods, animals..., to use songs in the studied language to work the listening comprehension.

B- For the student with predominance in Theorist Style to improve their listening comprehension, the activities must allow them to analyse what is said, we should repeat it over and over, and even analyse it deeply... Taking this into account, to make activities with video is very appropriate; watching films, documentaries, ...

C- For the Pragmatist students we could prepare these activities; to study structures, vocabulary, etc. related to daily life, interviewing a native person: after the elaboration of questions, a student acts as a famous person, and we work on the vocabulary, the colloquial expressions...

**Writing**

A-For Activist students; to chat using Messenger.

B-For Theorist students; to translate subtitles: to watch a scene of a film in original version with subtitles in Spanish and translate it into the second language, working the vocabulary, the grammar... and verify the original version

C-For Pragmatist students; to write e-mails

## 4. CONCLUSIONS

Our students have a preference for certain learning styles, making obsolete the old system centred on the teacher and in order to make learning effective, each student requires of a style of education adapted to his own way of learning. It also has the additional problem that not all teachers have much knowledge of that variety and do not know the strategies to follow according to the theories of the Learning Styles.

The importance of this investigation is to try to optimise the education and practice of a foreign language, increasing the level of knowledge of all the students using a Learning Generator or common text book for all the students of a definite English level, organizing it previously so that it teaches up to the maximum capacity of each student, considering their Learning Style and thus eliminating the teaching style of each teacher.

In order to identify learning styles we must take into account the investigations made by David Kolb and Peter Honey. Both investigations are complementary and they help us to identify the different learning styles and to see the different ways of learning that each individual has.

Dr. Catalina Alonso maintains that "it is frequent that a teacher tends to teach as he would like to be taught, that is to say, he teaches as he would like to learn, he really teaches according to his own learning style". It is clear that we cannot choose our students and, consequently, the learning styles of our students, but we can choose a teaching method that benefits all our students.

In this investigation the methodology of the main editorials in English teaching text books has been analysed (in general and by units) in order to see what percentage of quantitative representation they have in the different learning styles corresponding to the theories of Honey and Mumford (Activist, Reflectors, Theorist and Pragmatist) and we have seen that books do not follow the theories of the Learning Styles.

The excessive representation of exercises that benefit the Learning Style with less students and the small representation of exercises which benefit the students with majority styles demonstrate that the text books follow a mistaken tendency. The higher representation of exercises that benefits the Reflectors Style demonstrates that all the editorials, without exception, follow the communicative or natural the method.

The editorials do not consider the different Learning Styles of the students, and they are centred in a method that will soon be obsolete because the academic results do not reflect good results. After analysing the main deficiencies, some activities were created so as to deal with the deficiencies of the analysed text books (schemes, additional material for the teacher...), and verified if the modifications previously mentioned were effective as far as the attainment of the objectives proposed by each book, using a control group to which these modifications were not applied to. The results were highly encouraging since the students with Learning Styles with smaller representation in text books obtained better results than those than did not do the activities, since they belonged to the control group. This demonstrated that the complementary activities that had been prepared to replace the deficiencies of books, adding exercises and activities that benefited students from no-Reflectors Learning Style, were positive.

There seemed to be a connection between certain Learning Styles and certain linguistic aspects, seeming to have a relation between the oral abilities and the Activist and Pragmatist Styles, and between the written abilities with the Reflector and Theorist Styles, since they improved parallelly according to the linguistic area worked.

Those students that obtain worse academic results, perhaps by the format generally used in examinations are the students with Activist Style, those of Theorist Style, being the students with better academic results to whom the format of the examinations benefits, and the students of Reflectors Style, probably due to the insistence in text books to work this cognitive facet.

The use of a pedagogical approach and the elaboration of the learning programming of a second language must respond to several considerations. In a deductive presentation one begins with axioms, principles or rules. A great percentage of the class is deductive, probably being an elegant and efficient way of introducing what it is taught. Nevertheless, it is evident that to incorporate an inductive component in education promotes effective learning. Thus, inductive education has to have its place just like the deductive.

Connecting this to the education of second languages, we could say that, at the moment, the deductive method would be the classic one or taylorist and the inductive one would correspond to the natural method, so fashionable nowadays. For this last one, to acquire a language means a gradual learning, obtaining the ability of communication without the necessity of using the rules that a teacher explains, which benefits the students with a predominant Reflector Style, since they are observers, compilers and assimilators.

Different to other subjects, the teaching of English as a second language is very poor in deductive techniques, which makes learning for students with predominant Theorist Style quite difficult. If we have to balance deduction and induction, the text books used in English language teaching follow the wrong methodology, since they benefit a single style, the Reflexive, making learning difficult for students with other Learning Styles. This happens because the editorials follow the natural method. Thus, we must conclude that this method does not benefit the great majority of students and, consequently, we should eradicate it, or, at least, modify it.

We can conclude with clear evidence that the editorials do not consider the different Learning Styles at the time of programming their books. On the one hand, they do not seem to consider the percentage of representation of the pupils pertaining to each Learning Style. But on the other hand, they seem to consider the present tendency in the methods of education of the foreign languages, since they are centred in natural and communicative methods, leaving aside, for example, the grammar explanations that would benefit students with Theorist Style. Paradoxically, they do not turn out to be very communicative since they do not include a great variety of communicative exercises, that would benefit the students from Activist and Pragmatist Style. This must be because the text books are designed considering educative contexts where classes have a large number of students, which makes the accomplishment of these activities difficult.

It is obvious that the general implantation of the very fashionable natural or communicative method in the teaching of English does not give the corresponding results. Students who finish obligatory education do not end up with a level of English that allows good oral and written communication.

This investigation analysing the most widely used text books in the classrooms could discover the reason. Although it may seem excessive, this investigation exceeds expectations since the initial intention was only to see which editorial was better in quality, taking into account the diversity of learning styles, but this investigation has ended up finding the main failure of the tendency in education in second languages; the communicative method fails because it has an excess representation of a single Style, which is the one of smallest representation among the students (Reflectors Style).

Now it is time for the editorials to pay greater attention to the theories on Learning Styles than to the present educative tendencies, as the communicative and natural methodologies in foreign languages do not benefit all the students. Text books act as a tool which generates learning, and if we improved them taking into account the different Learning Styles, we would be working with a real Learning Generator for all the students, without any exception at all. Can we imagine a learning system where all the students learned at their best? What degree of knowledge could those students end up reaching if this system were implanted in a generalized manner? It seems utopia, but it is an attainable utopia if we prepare text books that benefit all students. If we used a method which benefited all our students, we would be creating students who would learn with the maximum of their capacities and all society would benefit from that.

## 5. BIBLIOGRAPHICAL REFERENCES

ALONSO, C. , GALLEGO, D. Y HONEY, P. 1997. *Recursos e instrumentos Psico-pedagógicos. Los estilos de aprendizaje. Procedimientos de diagnóstico y Mejora*. España*:* Universidad de Deusto, Ediciones Mensajero.

CHOMSKY, N. 2005. *L´educació.* Barcelona: Columna Edicions.

CHOMSKY, N. 1970. *Aspectos de la teoría de la sintaxis* Madrid: Aguilar

DAMASIO, A. R. Y DAMASIO, H. 1997. *Cerebro y Lenguaje. Fundamentos Biológicos II.* Madrid: UNED.

DEPARTAMENT D ´ENSENYAMENT 2002. *Autonomia a l´aula de Llengues Extrangeres: aula espais diversificats.* Barcelona: Generalitat de Catalunya.

DESPINS, J.P. 1985. *Connaitre les styles d´aprendissage pour mieux prespecter les Façons d´apprendere des enfanst.* Vie Pédagogique, 39, 10-16

DUNN, R. Y DUNN,K. 1984. *La Enseñanza y el Estilo Dindividual de Aprendizaje.* Madrid: Anaya

GARDNER, H. 1993. *Multiple intelligences: the theory in practice.* New York: Basic Books.

HADFIELD, J. Y HADFIELD, C. 2000. *Simple Speaking Activities.* Oxford: Oxford University Press.

KOLB, D. 1982. *Psicología de las organizaciones.* Madrid: Prentice Hall

KOLB, D. 1984. *Experimental learning: Experience as the source of Learning and Development.* New Jersey: Prentice Hall

PIAGET, J. 2001. *Psicología y pedagogía*. Barcelona, Editorial Crítica.

PUEYO, A. A. 1999. *Psicología Diferencial.* Barcelona, McGraw-Hill.

SEGOVIA, S. y GUILLAMÓN, A. 1997. *Psicobiología del Desarrollo.* Barcelona: Ariel Psicología.

SKINNER, B.F. 1992. *Verbal behaviour*. Massachusetts: Copley

VARELA, R. 1998. *Estrategias de enseñanza-aprendizaje de idiomas extranjeros.* Madrid: UNED.

VEZ, J.M. 2002. *El aula de lenguas extranjeras: umbral para una sociedad de la Cultura.* Barcelona: Graó.

# THE END OF THE NATURAL SYSTEM

**Dr. Eva Zanuy**
**PhD on Applied Linguistics**
**evazanuy@hotmail.com**

Not only do students have their preferences and their style of learning. All teachers have their own style when giving class, and that style is also reflected when we use the different representational systems. Most of us tend to use a system more often than the others when we teach. In order to detect what our tendencies are, we need to analyse our way of teaching from the point of view of the NLP. Generally, in all the groups of students we will find different types of learning styles. If our teaching style is the same as that of our students, learning will be easier for them than if is not the same one, and with a book using all the different styles we will be benefiting all our students.

Each editorial has common characteristics and differential characteristic. One of the main common characteristics that has been found after analysing different editorials belonging to the same level is the great numerical equality of exercises that benefit the different systems of neurolinguistic representations. The neurolingistic representation in the editorials would be; Visual; 35 %, Auditory; 33 % and Kinesthetic; 32 %

The Oxford publishing house turns out to be the one that benefits more the students with predominance in Visual style (50 %). Cambridge is second (41.5 %), Pearson occupies third (38.5 %), whereas Heinemann (24.1 %) and Richmond (20.8 %) includes a smaller representation of exercises that benefit this group of students. The Visual style is the one that has the greatest representation in three of five editorials, although not by much from the second predominant style, the Auditory style. The one with the greatest percentage is Heinemann (44.1 %), followed by Richmond (40.8 %) and Cambridge (30.5 %). Those that have a smaller percentage are Oxford (26 %) and Pearson (24.2 %). The Kinesthetic style is the least represented in two of five editorials although not by a remarkable amount from the other representational systems, and varies between the greatest representation of Richmond (38.4 %) and the representations of Pearson (37.3 %), Heinemann (31.8 %), Cambridge (28 %) and Oxford (24 %). This analysis demonstrates that the books of the most sold and used editorials are near being learning generators. The percentage of visual children habitually is very superior to the auditory and Kinesthetic children, for that reason many activities are prepared for these children.

One of the main common characteristics after analysing different editorials on the same level is the great representation of exercises that a Learning Style has over other Styles. The average representation in percentages of the Learning Styles in the analysed editorials would be; Activist; 18.4 %, Reflector; 49.4 %, Theorist; 17.8 % and Pragmatist; 14 %

The Reflector Style, with a representation of 49.4 %, is the Style which all editorials benefit to. This data is common in all the analysed editorials. The Activist Style occupies second position if we consider the average, with an 18,4 % representation, but it has only been the second more represented Style in three of the five editorials. The third most represented Style is the Theorist Style, with 17,8 %, that is also the second most represented Style in three of the five analysed editorials. The Pragmatist Style, with a representation of 14 %, has been the least represented Style in three of the five editorials, and it is, the Style with the smallest representation in general.

The Richmond publishing house turns out to be the one that most benefits the students with predominance in Activist Style (30 %). The Pearson publishing house is second (23 %) and Cambridge and Heinemann occupy third (17 %), whereas Oxford has the smallest representation of exercises that benefit this group of students.

The Reflector Style is the one that has the greatest representation in all the editorials, and with a clear advantage in percentage from the second predominant Style. The publishing houses with the highest percentage (56 %) are Oxford, and on the other hand, Heinemann is the one that has the lowest percentage (43 %). As it can be verified, the highest score and the lowest do not distant to a great extent. Heinemann is also the publishing house with the greatest percentage in representation of exercises with Theorist Style (29 %). Oxford occupies the second position (22 %). Cambridge (17 %) and Pearson (14 %) occupy the following positions and Richmond has the lowest percentage (7 %). The Pragmatist Style is the least represented style and varies between Pearson and Heinemann (11 %) and Oxford, Cambridge and Richmond (16 %).

Our students have a preference for certain learning styles, making obsolete the old system centred on the teacher and in order to make learning effective, each student requires of a style of education adapted to his own way of learning. It also has the additional problem that not all teachers have much knowledge of that variety and do not know the strategies to follow according to the theories of the Learning Styles.

The importance of this investigation is to try to optimise the education and practice of a foreign language, increasing the level of knowledge of all the students using a Learning Generator or common text book for all the students of a definite English level, organizing it previously so that it teaches up to the maximum capacity of each student, considering their Learning Style and thus eliminating the teaching style of each teacher.

In order to identify learning styles we must take into account the investigations made by David Kolb and Peter Honey. Both investigations are complementary and they help us to identify the different learning styles and to see the different ways of learning that each individual has.

In this investigation the methodology of the main editorials in English teaching text books has been analysed (in general and by units) in order to see what percentage of quantitative representation they have in the different learning styles corresponding to the theories of Honey and Mumford (Activist, Reflectors, Theorist and Pragmatist) and we have seen that books do not follow the theories of the Learning Styles.

The excessive representation of exercises that benefit the Learning Style with less students and the small representation of exercises which benefit the students with majority styles demonstrate that the text books follow a mistaken tendency. The higher representation of exercises that benefits the Reflectors Style demonstrates that all the editorials, without exception, follow the communicative or natural the method.

The editorials do not consider the different Learning Styles of the students, and they are centred in a method that will soon be obsolete because the academic results do not reflect good results. After analysing the main deficiencies, some activities were created so as to deal with the deficiencies of the analysed text books (schemes, additional material for the teacher...), and verified if the modifications previously mentioned were effective as far as the attainment of the objectives proposed by each book, using a control group to which these modifications were not applied to. The results were highly encouraging since the students with Learning Styles with smaller representation in text books obtained better results than those than did not do the activities, since they belonged to the control group. This demonstrated that the complementary activities that had been prepared to replace the deficiencies of books, adding exercises and activities that benefited students from no-Reflectors Learning Style, were positive.

There seemed to be a connection between certain Learning Styles and certain linguistic aspects, seeming to have a relation between the oral abilities and the Activist and Pragmatist Styles, and between the written abilities with the Reflector and Theorist Styles, since they improved parallelly according to the linguistic area worked.

Those students that obtain worse academic results, perhaps by the format generally used in examinations are the students with Activist Style, those of Theorist Style, being the students with better academic results to whom the format of the examinations benefits, and the students of Reflectors Style, probably due to the insistence in text books to work this cognitive facet.

The use of a pedagogical approach and the elaboration of the learning programming of a second language must respond to several considerations. In a deductive presentation one begins with axioms, principles or rules. A great percentage of the class is deductive, probably being an elegant and efficient way of introducing what it is taught. Nevertheless, it is evident that to incorporate an inductive component in education promotes effective learning. Thus, inductive education has to have its place just like the deductive.

Connecting this to the education of second languages, we could say that, at the moment, the deductive method would be the classic one or taylorist and the inductive one would correspond to the natural method, so fashionable nowadays. For this last one, to acquire a language means a gradual learning, obtaining the ability of communication without the necessity of using the rules that a teacher explains, which benefits the students with a predominant Reflector Style, since they are observers, compilers and assimilators.

Different to other subjects, the teaching of English as a second language is very poor in deductive techniques, which makes learning for students with predominant Theorist Style quite difficult. If we have to balance deduction and induction, the text books used in English language teaching follow the wrong methodology, since they benefit a single style, the Reflexive, making learning difficult for students with other Learning Styles. This happens because the editorials follow the natural method. Thus, we must conclude that this method does not benefit the great majority of students and, consequently, we should eradicate it, or, at least, modify it.

We can conclude with clear evidence that the editorials do not consider the different Learning Styles at the time of programming their books. On the one hand, they do not seem to consider the percentage of representation of the pupils pertaining to each Learning Style. But on the other hand, they seem to consider the present tendency in the methods of education of the foreign languages, since they are centred in natural and communicative methods, leaving aside, for example, the grammar explanations that would benefit students with Theorist Style. Paradoxically, they do not turn out to be very communicative since they do not include a great variety of communicative exercises, that would benefit the students from Activist and

Pragmatist Style. This must be because the text books are designed considering educative contexts where classes have a large number of students, which makes the accomplishment of these activities difficult.

It is obvious that the general implantation of the very fashionable natural or communicative method in the teaching of English does not give the corresponding results. Students who finish obligatory education do not end up with a level of English that allows good oral and written communication.

This investigation analysing the most widely used text books in the classrooms could discover the reason. Although it may seem excessive, this investigation exceeds expectations since the initial intention was only to see which editorial was better in quality, taking into account the diversity of learning styles, but this investigation has ended up finding the main failure of the tendency in education in second languages; the communicative method fails because it has an excess representation of a single Style, which is the one of smallest representation among the students (Reflectors Style).

Now it is time for the editorials to pay greater attention to the theories on Learning Styles than to the present educative tendencies, as the communicative and natural methodologies in foreign languages do not benefit all the students. Text books act as a tool which generates learning, and if we improved them taking into account the different Learning Styles, we would be working with a real Learning Generator for all the students, without any exception at all. Can we imagine a learning system where all the students learned at their best? What degree of knowledge could those students end up reaching if this system were implanted in a generalized manner? It seems utopia, but it is an attainable utopia if we prepare text books that benefit all students. If we used a method which benefited all our students, we would be creating students who would learn with the maximum of their capacities and all society would benefit from that.

Dr. Mrs. Eva Zanuy
UNED
Spain.

# A corpus based linguistic tool for machine translation:
# A case study of '~INGs' in English and its equivalents in Tamil through grammatical association approach

**Dr.S.Kamakshi Devi**
**ICFAI University, Regional Office, TN, East, Chennai**
**dr.kamakshidevi@gmail.com**
**09940593079**

## Abstract

The research paper aims to discuss about the techniques that can be used for compiling a Machine Tractable Dictionary (MTD) for machine translation based on parallel corpora. And also it explains the exhaustive analysis of translation equivalents ranging from morphemes to words and to sentences by having a typological study of English and Tamil languages.

## I. Introduction:

Traditionally, linguistic analyses have emphasized structure - identifying the structural units and classes of a language. (e.g. morphemes, words, phrases, grammatical classes) and describing how smaller units can be combined to form larger grammatical units (e.g. how words can be combined to form phrases, phrases can be combined to form clauses, etc.)

A recent perspective - corpus based statistical approach has been adopted here to investigate how speakers and writers exploit the resources of their language rather than looking at what is theoretically possible in a language, we study the actual language used in naturally occurring texts.

This research paper aims at investigating specifically the grammatical association of *~ing* words in English and its possible translation equivalents found in Tamil Language, since seemingly similar structures occur in different contexts and serve different functions based on its grammatical associations that can help to prepare a Bi-lingual Machine Tractable Linguistic Tool for Machine Translation.

## II. Modus operandi of investigating the right and left collocates of - ings in English to make the machine recognize

- It can be assured that one of the left collocates of *~ing* is verb that is grouped as a Open Parts of Speech in English and it has grammatical association with the present participle suffix i.e~*ing*. Hence a Corpus Based Statistical Approach can be adopted to database all the words found in the dictionary or in any tagged corpus as a verb as its principle entry.

- Since it is one of the trends in language that the Nouns can be made use of Verbs to shrink the language instead of populating the parts of speech, one can investigate the nouns which are already being made use of verbs or likely to be made use of verbs in the present day language to enrich the lexical resource of a machine readable dictionary which is an ultimate linguistic tool for machine translation.

- During present-participiling, some of the English verbs which possess *-ing* with the root verb take another *ing*. So it is necessary to make distinction between the present participle verbs which has an *ing* and an *ing* as found in the following samples like *singing, ringing, singing etc.*, since the machine is incapable of understanding the process of present participiling automatically.

- Analysis can be made on the problematic verb forms such as will since it does not have past tense form *willed** and past participle form *willed/willen** but it has present participle form *willing* but functioning as an adjective in the example `*Lakshmi was willing to join with us* ’, and the III person singular present tense ‘s’ form that is `*will*s’ is available in rare context, examples like `*if god wills…*’

- Since the following samples are overlapping the word-formation rules in English it is very interesting to note that even the adjectives like `*ready’* is made use of as verbs as found below

    a. `Theni readying for CM’s Visit’ found in the popular daily THE HINDU dated 23rd January 2004.
    b. `..officials are readying venues here and…’ found in the same daily dated Jan 31,2004.

    So this kind of analysis would be a worth experimenting research on corpus to find out the new word-formations and usages of such words to instance it in the lexical resources like Machine Readable Dictionary (MRD)

- Context Free Grammar (CFG) techniques can be adopted to train the machine to understand the various grammatical functions of the ~ing words as follows

    a. adjective in the example `*Smiling beauty* ’
    b. gerund in the example ` *Smiling rapports everbody*’
    c. present participle in the example `*She is smiling*’

    depends upon the concordance.

- Statistical analysis can be made to database and distinguish the *verbless-ing* words under its various grammatical functions as follows:

| | |
|---|---|
| Evening | [Noun/Time adverb] |
| King | [Pure Noun] |

| | |
|---|---|
| During | [Preposition] |
| According to | [Adverb] |
| Something | [Pronoun] |

This analysis would be useful for preparing a linguistics tool that would recognize and tag the various grammatical categories of ~ing words in the given corpus.

**III. The case and correlative study of *ings* in English and its equivalents in Tamil brings the following issues to be kept in mind to build a Machine Tractable Linguistic TOOL (MTLT).**

1. It is important to make an analysis whether all gerundial nouns in English are bringing *'thal*' or *` ththal'* suffixes in Tamil since it is highly productive in nature and can we add some more suffixes as gerundial suffixes in Tamil like *ippu* etc.,

   1 (I). The gerund `doing' is neutral in English in the following examples a), b), and c) and demonstrate the challenges in providing the translation equivalents since the relativity of tense is being maintained or expressed in Tamil

   Examples:

   a) By doing it himself, he is saving lot of money (Present time reference)
   *avane athai ceykirathan/ceyvathan muulamaaka avan athika panaththai ceemiththukkontirukkiraan*

   b) By doing it himself, he saved a lot of money (Past time reference)
   *avane athai ceyththan muulamaaka avan athika panaththai cemithaan*

   c) By doing it himself, he will save a lot of money (Future time reference)
   *avane athai ceyvathan muulamaaka avan athika panaththai cemippan*

2. The ~*ing* ending words that are functioning as an adjectives take participial noun suffixes in Tamil based on tense, person, number, gender. or take *akaa* or *aana* suffixes in Tamil based on the concordance found in the following examples

   e.g. a)`The story is interesting' *intha kathai aarvamaaka inrukkirathu*. In this example the word *interesting* is functioning as an adjective, but it takes *`aaka'* suffix *(*which is popularly noted as adverbial suffix) in Tamil since the copula verb *`is'* is there.

   e.g. b) `It is an interesting story' *ithu oru arumaiyaana kathai*. In this example the word *interesting* is functioning as an adjective, and it takes '*aana*' suffix in Tamil (which is popularly noted as adjectival suffix) since an article follows it.

3. Whether the *–ing* fronted with *a* verb followed by a `be' form take kontiru as its translation in Tamil. (be +verb+ing = -----+*kontiru* +--------+-----)

4. Since the systematic usages like ‘neighbouring house’ or ‘neighbouring country’ are being translated as *pakkathil ulla viitu* or *pakkatthu viitu* and *pakkaththil ulla nadu* or *pakkaththu nadu*, the frequency of lexical – lexical association or concordance listings from a parallel corpus can be considered for providing translation equivalents in the machine understandable linguistic tools such Arbitrarily Reordered Dictionaries (ARD) or Machine Readable Dictionaries (MRD).

## IV Conclusion:

The scope of the present research paper is a wider spectrum of identifying translation equivalents ranging from morphemes to words and to sentences and transferring the English Sentences into Tamil. The exhaustive correlative study of English and Tamil which amalgamates both contrastive and typological studies of the two languages will be a future documentation for one who will attempt for a full fledged Machine Tractable Dictionary (MTD) for Machine Translation (MT) system for transferring English into Tamil.

## References

1. The Oxford book of Computational Linguistics by Ruslan Mitkov, Oxford University Press. 2003
2. Modern English: A book of Grammar, Usages and Composition. By N.Krishnamoorthy CIEFL, Hyderabad, 1975
3. Modern English A Practical reference guide by Marcella Frank, New Jersey
4. Corpus Linguistics Investigating Language Structure and use Douglas Biber Susan Conrad & Randi Rappan (Cambridge approaches to Linguistics)
5. World Lexicon of Grammaticalization Cambridge – Bend Heine. Tania Kuteva.
6. The BNC Handbook exploring the British National Corpus with Sara Edinburgh Textbooks in Empirical Linguistics.
7. Linguistic Structures in Tamil – A Historical Study. A. Athithan, Publications Division, Madurai Kamaraj University.

*Dr.Kamakshi –a pioneering linguist earned her doctoral degree in Linguistics –Machine Translation from Tamil University in 2001.*
*She has authored a book entitled “Preliminaries to the preparation of a Machine Aid to Translate Linguistics Books written in English into Tamil” 2003, 455, p and published as DLA (Dravidian Linguistics Association-Trivandrum) publications. She has been into teaching Applied Linguistics and English since 1994. She has been rendering her service for ICFAI University as a Soft skill trainer and establishing media relations. Writing Linguistic research articles and giving guidance to the research scholars in Applied Linguistics is her passion*

# Proposed Biometric key for DES Scheme Applying Formant's Arc-tangents

**Zaki. T. Fayed**
**Faculty of Computers & Information Sciences, Ain Shams University**
**Dep. of Computer Science**
**Cairo, Egypt**
**ztfayed@hotmail.com**

## Abstract

This work explores the possibility of using a proposed biometric feature based on sound formants as differentiating parameters in speaker verification. It is claimed to be feasible by applying these parameters to substitute the typical keys in symmetric encryption algorithm as in data encryption standard (DES). The procedures for preparing the suggested key in this algorithm constitute a set of steps. First, to choose the candidate formants; F1, F2, and F3. Second to pre-permute these frequencies then to compute the arc-tangents of (*F1-F2, F1-F3, F2-F3, F3-F1, F2-F1,F3-F2*). Formants which are embedded in the computed arc-tangents are considered as signatures in DES with the purpose of adding more complication in encryption process for less cryptanalysis probabilities. The proposed digital signatures are chosen randomly in a pre-determined time slices, to be applied to the key schedule, KS in DES algorithm. Complexity issue due to these additional processes has to be discussed. Formant's arc-tangent is recommended to be of 64-bit block from which the 56-bit key is extracted as usual in DES algorithm. Trade-off of security needs and algorithm complexity are biased towards providing more encryption and authentication sophistication. The motivation to apply the proposed algorithm is the susceptibility of DES to analytical attack, in addition to the limitation of the maximum cryptographic security (56 bits) of DES. This paper focuses on vowels because it is characterized, mainly by formants. Also most often the two first formants, *F1*and *F2* are enough to disambiguate the vowels. The uniqueness of formants as biometric parameters reveals robust authentication for messages via networks. More add-on sophistication is due to the underlined procedures in DES-algorithm. The first three formants are extracted from pre-determined vowel-based sounds of five adult speakers. These formants are permuted for computing the arc-tangent values which then are undertaken the basic operations on the keys of DES-algorithm. Two nested loops for generating time slices and the formants permutation in addition to the linear order of computing the arc-tangent of the mentioned formants are added to the algorithm of the traditional DES.. The overall cost of the added algorithm is nearly $N^2$. The additional complexity is accepted compared with the great cryptanalysis resistance of the proposed biometric-based DES algorithm.

**Keywords:** **Data Encryption Standard (DES), Formants, Arc-tangent, Algorithm complexity**

## 1. Introduction

Data encryption is nothing new, but when it is used in conjunction with high performance, high-volume enterprise storage, it poses some legitimate challenges. Data encryption is being implemented for stored data in addition to the traditional use of encrypting data in transit. Most of the encryption focus had been on data transmission, but the risk of compliances are moving the topic of encrypting data at rest, or stored data, much higher on the priority list of leading edge data protection strategies. DES is now considered to be insecure for many applications. This is chiefly due to the 56-bit key size being too small as DES keys have been broken 24 hours or less as microprocessor speeds increase. Computer chips currently can test 200 million DES keys/second [1]. Other encryption algorithms have been in use for years and include Secure Sockets Layer (SSL) for internet transaction, Pretty Good Privacy (PGP), and Secure Hypertext Transfer protocol (S-HTTP). The additional proposed biometric layer aims at enhancing the performance of DES specially at the recipient. Hashing encryption is a cryptographic algorithm that takes data input of any length and produces an output of a fixed length which is denoted "Digital Signature" and is used for data integrity. Digital signatures typically range from 128 bits using the MD5 algorithm (message Digest 5) to 160 bits in size using the more secure SHA1 ( Secure Hash Algorithm 1) [1].

In this paper, a proposed additional layer which is considered as a digital signature like hash (One Way) algorithm is complemented with DES algorithm. The suggested features add biometric privilege to DES encryption capabilities to make it more secure. The robustness of applying biometric parameters in encryption systems is to provide unique authentication, well verification, and recipient privacy. The uniqueness of formants in the human voice adopts them to be applied in the security algorithms. Although the extraction of formants is not so simple and will increase the whole complexity of the intended systems, it on the other hand, will add value to these systems. Numerous symmetric cipher have been developed since the introduction of data encryption standard (DES). Although DES is replaced by the (AES), DES remains the most important such algorithm. [2]. DES is considered, now, insecure because a brute force attack is possible. The best analytical attack is linear cryptanalysis and has a time complexity of $2^{34-39}$.

The paper is organized as follows. In section II, an overview of DES is presented. A brief symbolic description of DES is explored in section III. In section IV, the formants as biometric metrics are discussed. Section V illustrates the proposed algorithm. Results are explored in section VI. Finally Conclusion and future work are presented in section VII.

## II. Data Encryption Standard (DES): Overview

Cryptography is the art of secret writing, or devising ways of transmitting messages so that others cannot read them. DES is a "private key" system; that communicants share a secret key, and the eavesdropper will succeed if he can guess this key among its quadrillions of possibilities [4].

In contrast, the security of a typical "public key" is based on the difficulty of taking "discrete logarithms" (reversing the process of exponentiation in a finite field); that is the Deffie Heliman key exchange protocol. The simplified DES (S-DES) has similar properties and structure to DES with much smaller parameters. S-DES takes an 8-bit block of plaintext and a 10-bit key as input and produces an 8-bit block of ciphertext as output. In the basic structure of DES, there is a secrete 56-bit key under whose influence a 64-bit plaintext (input) is transformed into 64-bit ciphertext (output). The input message is broken into two halves, "left" and "right" during the first of sixteen "rounds," the 32-bit right half along with 48 of the 56 key bits, is fed into a nonlinear function ***F***. The 32-bit output of this function , added to the left message, becomes the new right half message. Meanwhile the old right half message is funneled forward to become the new left half message. Thus ends one round. The process is repeated sixteen times, using a different selection of 48 key bits each time. The final left half and right half message become the ciphertext [5],[6]. Strength of the typical DES fall into two areas : key size and the nature of the algorithm. With the length of 56-bits, there are $2^{56}$ possible keys, which is approximately $7.2x10^{16}$ keys. The increase of the parallel hardware speed make the key detection and consequently, DES attack easier and in less time. Fortunately, there are a number of alternatives to DES, the most important of which are AES ( Advanced Encryption Standard) and triple DES (3DES) [7] 3DES has three times as many rounds as DES. AES is a symmetric block cipher with a block length of 128 bits and support for key length of 128, 192, and 256 bits [8]. DES encryption algorithm involves four functions for computation as indicated in **figure -1** [9].

1. an initial permutation, IP
2. a complex function, $f_k$ which constitutes both permutation and substitution operations depending on the key input.
3. a simple permutation that switches (SW) the two halves of the data $f_k$ once more, and
4. the inverse permutation function $IP^{-1}$ which inverse the IP-operation.

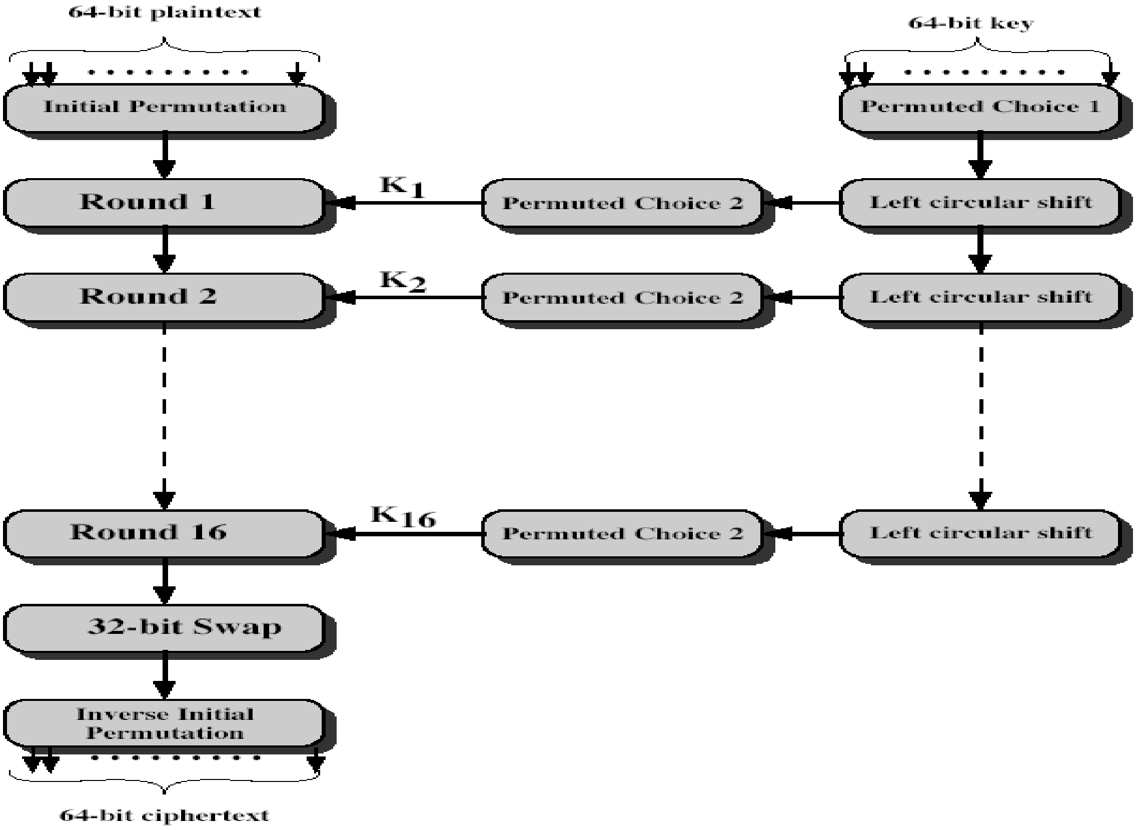


**Figure-1 Data Encryption System (DES) Algorithm [9]**

As with any encryption scheme, there are two inputs to the encryption function: The plaintext to be encrypted and the key. In DES-algorithm, the processing of the plaintext proceeds in three phases. First, the 64-bit plaintext passes through an initial permutation (IP) that rearranges the bits to produce the permuted input. This is followed by a phase constituting of 16 rounds of the same function, which involves both permutation and substitution functions. The output of the last (sixteen) round consists of 64 bits that are a function of the input plaintext and the key. The left and right halves of the output are swapped to produce the pre-output. Finally the pre-output is passed through a permutation (***IP*** $^{-1}$); the inverse of the initial permutation function, to produce the 64-bit ciphertext. The 56-bit key is passed through a permutation function. Then, for each of the 16 round a sub-key ($K_i$) is produced by the combination of a left circular shift and permutation. The permutation function is the same for each round, but a different sub-key is produced because of the repeated iteration of the key bits.

## III Brief Symbolic Description of DES

The algorithm is designed to encipher and decipher blocks of data consisting of 64 bits under control of 64-bit key (blocks are composed of bits numbered from left to right). The deciphering process is the reverse of the enciphering process. A block to be enciphered is subjected to an initial permutation ***IP***, then to a complex key-dependent computation and finally to permutation which is the inverse of the initial permutation ***IP*** $^{-1.}$ The k-function-dependent computation can be simply defined in terms of a function ***f***, called the cipher function, and a function ***KS***, called the key schedule. The function, F is defined in terms of primitive functions which are called the selection functions, $S_i$ and the permutation function P. Let the 64 bits of the input block to an iteration consists of a 32 bit block L followed by a 32 bit block R, then the output block is then LR. The output L'R' of an iteration with input LR is defined by:

$$\mathbf{L' = R}$$
$$\mathbf{R'\ L = \quad f(R,K)}$$

where **f (R,K)** denotes bit-by-bit addition modulo 2.

If ***L'R'*** is the output of the 16th iteration then ***R'L'*** is the pre-output block. At each iteration a different block ***K*** of key bits is chosen from the 64-bit key designated by ***KEY.*** Let KS be a function which takes an integer n in the range from 1 to 16 and a 64-bit block KEY as input and yields as output a 48-bit block $K_n$ which is a permuted selection of bits from KEY.. That is:

$$\mathbf{K_n = KS(\ n,KEY)}$$

Let the permuted input block be LR and let $L_0$, $R_0$ be respectively L and R and let $L_n$, $R_n$ be respectively L' and R' when L, R are respectively $L_{n-1}$ and $R_{n-1}$ and K is $K_n$; that is when n is in the range from 1 to 16,

$$\mathbf{L_n = R_{n-1}}$$
$$\mathbf{R_n = L_{n-1} \qquad f(R_{n-1},K_n)}$$

The pre-output block is then $\mathbf{R_{16}L_{16}}$**.**

The key schedule **KS** produces 16 $\mathbf{K_n}$ which are required for the algorithm**. Figure-2** explores the enciphering computation process [2].

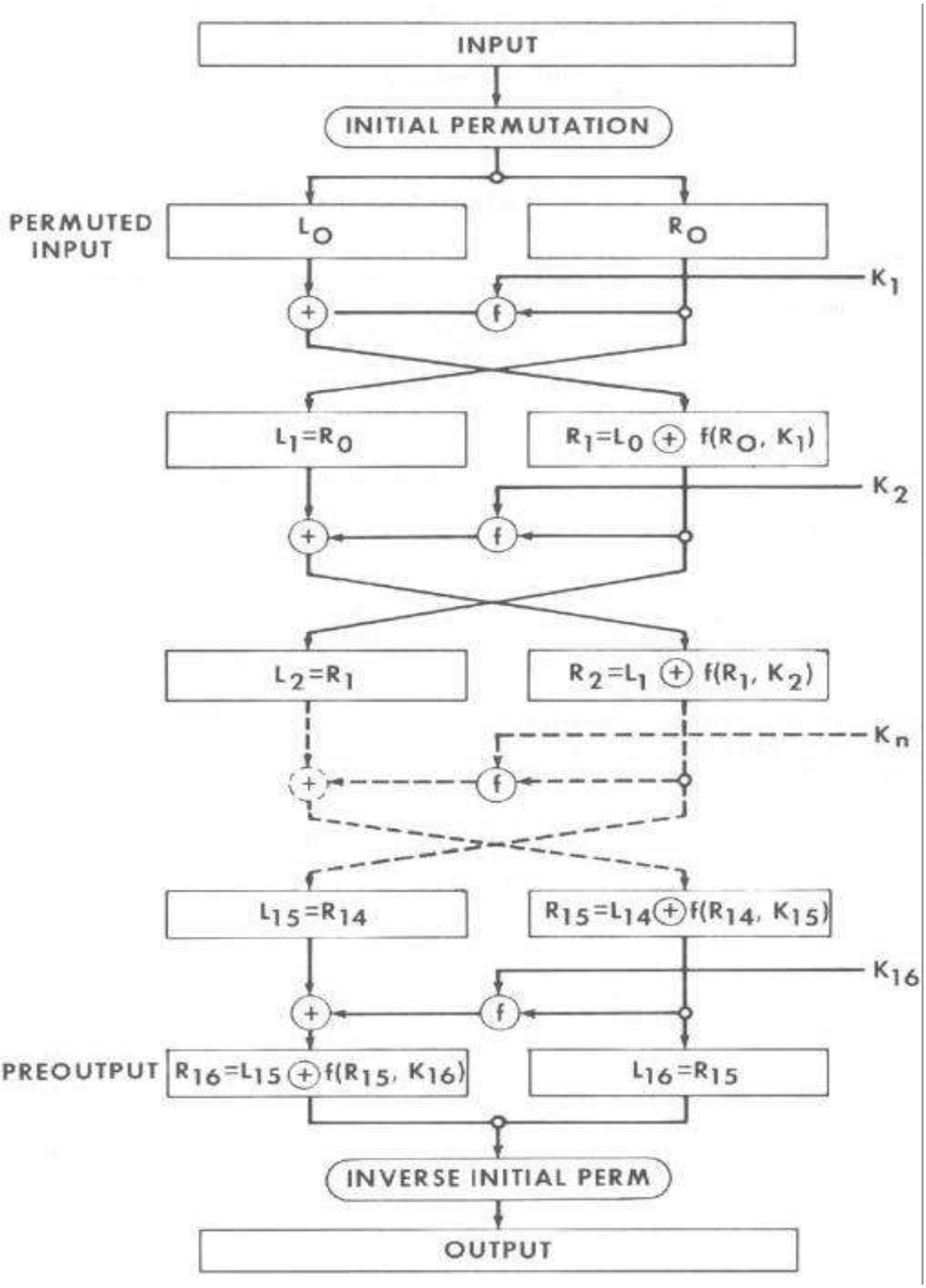


**Figure-2 Enciphering Computation process [2]**

Consequently, to decipher it is only necessary to apply the very same algorithm to an enciphered message block, takencare that at each iteration of the computation the same block

In deciphering, the permutation $\mathbf{IP^{-1}}$ applied to the pre-output block is the inverse of the initial permutation IP applied to the input.  It follows that:

**R =L'**
**L =R'  f(L',K)**

of key bits K is used during decipherment as is used during the enciphering of the block.

The permutation $\boldsymbol{IP^{-1}}$ applied to the pre-output block is the inverse of the initial permutation ***IP*** applied to the input. A sketch of the calculation of ***f(R,K)*** is given in **figure-3.** [10].

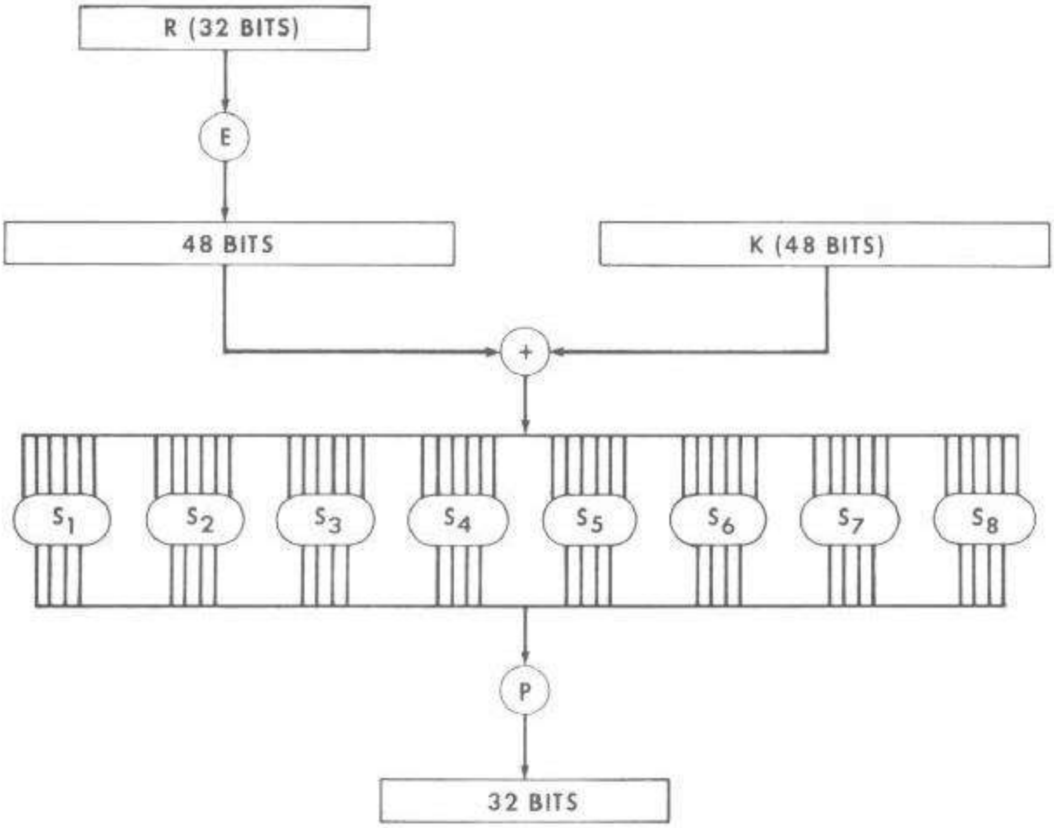


**Figure-3  Sketch of the F(R,K) Calculation [10]**

Let E denotes a function which takes a table of 32 bits as input and yields a block of 48 bits as output.  Let E be such that the 48-bits of the output, written as 8 blocks of 6 bits each, are obtained by selecting the bits in its inputs in order according to the following table:

**E bit-selection table**

| | | | | | |
|---|---|---|---|---|---|
| 32 | 1 | 2 | 3 | 4 | 5 |
| 4 | 5 | 6 | 7 | 8 | 9 |
| 8 | 9 | 10 | 11 | 12 | 13 |
| 12 | 13 | 14 | 15 | 16 | 17 |
| 16 | 17 | 18 | 19 | 20 | 21 |
| 20 | 21 | 22 | 23 | 24 | 25 |
| 24 | 25 | 26 | 27 | 28 | 29 |
| 28 | 29 | 30 | 31 | 32 | 1 |

It is noted from the above table that the first three bits of **E(R)** the bits in the position 32, 1 and 2 of **R** while the last 2 bits of **E(R)** are the bits in the position 32 and 1.

Each of the unique selection functions **$S_1$, $S_2$, ……,$S_8$** takes a 6-bit block as input and yields a 4-4-bit block as output and is illustrated by using a table containing the recommended **$S_i$**, see Appendices in [10]

## IV. Formants: The Biometric Security Metrics

A formant is a peak in an acoustic frequency spectrum which results from the resonant frequencies of any acoustic system. It is most commonly invoked in phonetics or acoustics involving the resonance frequencies of the vocal tract or musical instruments [13].

Formants are distinguishing or meaningful frequency components to human speech. A speech sound wave does not actually travel through the vocal tract and out into the air. Rather, the air in the vocal tract behaves like a spring that vibrates back and forth in standing wave. Resonant frequencies match the frequencies of the waves that will fit the tube. The general form for calculating formants is illustrated as follows [14]:

**$F_n = (2*N-1) * C /4L$** where
N = resonance number,
C = Speed of sound,
L = length of the vocal tract.

By definition, the information that humans require to distinguish between vowels can be represented quantitatively by the frequency content of the vowel sounds. The formant with the lowest frequency, F1, F2, F3 are enough to disambiguate the vowel. The first two formants, F1, F2 are primarily determined by the position of the tongue. F1 has a higher frequency when the tongue is lowered and F2 has higher frequency when the tongue is forward. Each formant corresponds to a resonance in the vocal tract. Vowels will almost have four or more distinguishable formants. Formants move about in a range of approximately 1000 Hz for a male adult. Nasals usually have an additional formant around 2500 Hz. The liquid usually has an extra formant at 1500 Hz, while the repetitive sound is distinguished by virtue of a very low third formant (below 2000 Hz). Plosives modify the placement of formants in the surrounding vowels. Bilabial sounds cause a lowering of the formants. Alveolar sounds cause less systematic changes in neighboring vowel formants, depending partially on exactly which vowel is present. The time-course of these changes in vowel formant frequencies are referred to as 'formant transitions' [11]. Each vowel can be placed on a graph, where F1, F2 are represented on the horizontal and the vertical dimension respectively. **Figure-4** depicts the formants, F1 vs. F2 for different vowels [12].

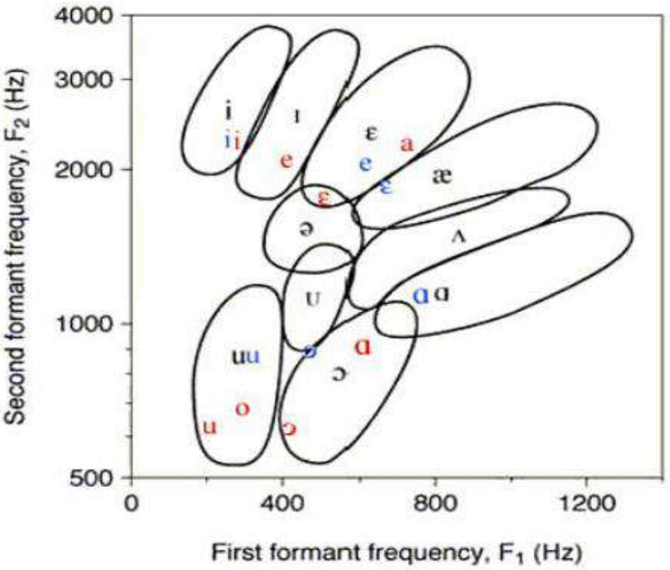


**Figure-4 F1 vs. F2 for different vowels [12]**

The energy contents in each formant is represented as dark band in wideband spectrogram as shown in **figure-5** [14] Formants are existing in both vowels and consonants and they are seen on the spectrogram around the frequencies that correspond to the resonances of the vocal tract. In case of consonants, the oral constriction in the vocal tract results in anti-resonances at one or more frequencies. Consequently, they attenuate or eliminate formants at or near these frequencies, as it is seen below 3000-4000Hz for the two instances of [s] in the above spectrogram. All vowels are characterized by F1 and F2, but more complete description of front vowels requires at least F3 as well, which differentiates between [i] and [y] etc.

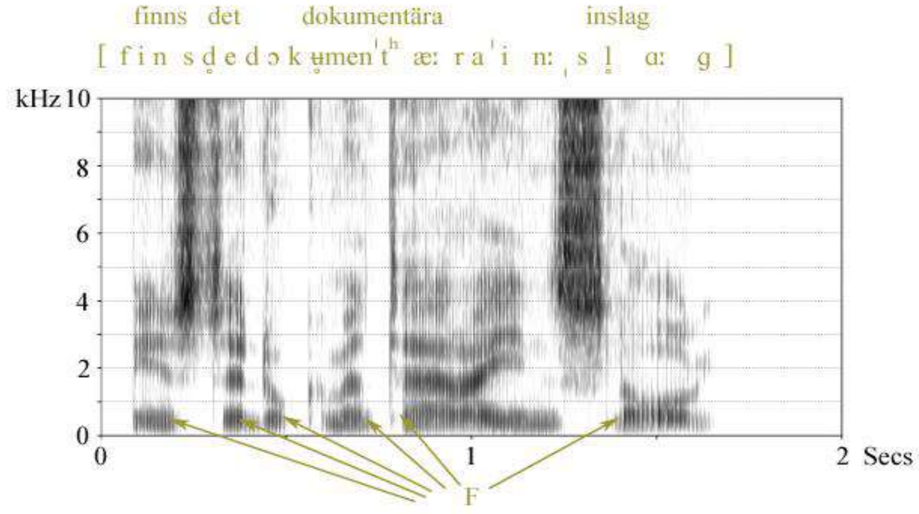
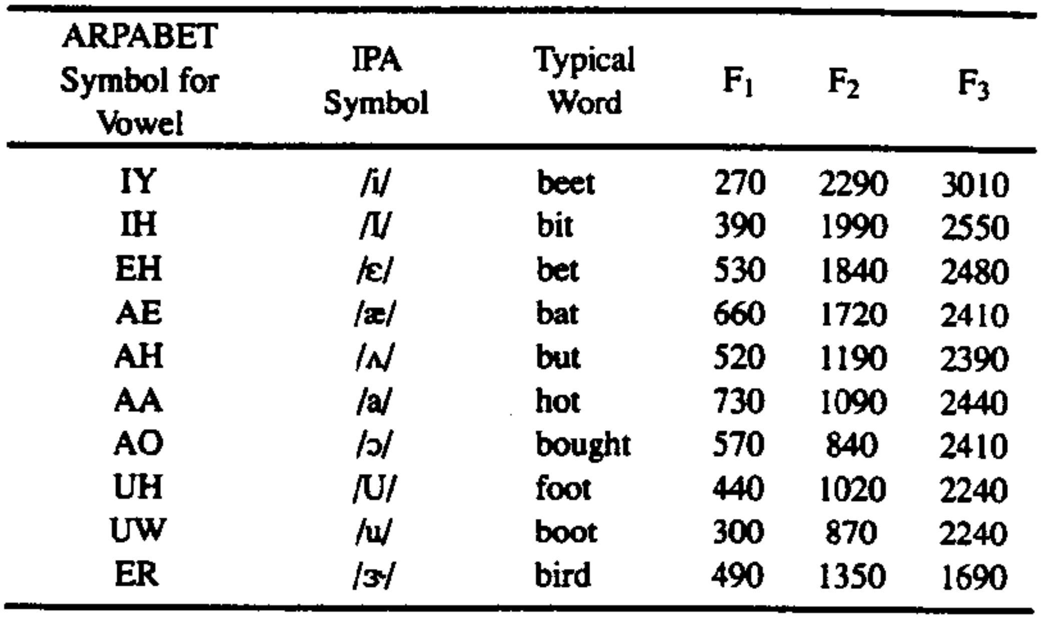


**Figure-5 Wide-banded formants of different sounds [14]**

**Table-1** declares the range of vowel formants to be checked during formants extraction in the proposed algorithm.

Table (1): F1, F2, F3 for different vowels for men with typical words [14]



| ARPABET Symbol for Vowel | IPA Symbol | Typical Word | $F_1$ | $F_2$ | $F_3$ |
|---|---|---|---|---|---|
| IY | /i/ | beet | 270 | 2290 | 3010 |
| IH | /I/ | bit | 390 | 1990 | 2550 |
| EH | /ε/ | bet | 530 | 1840 | 2480 |
| AE | /æ/ | bat | 660 | 1720 | 2410 |
| AH | /ʌ/ | but | 520 | 1190 | 2390 |
| AA | /a/ | hot | 730 | 1090 | 2440 |
| AO | /ɔ/ | bought | 570 | 840 | 2410 |
| UH | /U/ | foot | 440 | 1020 | 2240 |
| UW | /u/ | boot | 300 | 870 | 2240 |
| ER | /ɝ/ | bird | 490 | 1350 | 1690 |

## V. Proposed Algorithm:

**As mentioned above, the advancement in computer technology helps very much the cryptanalysis of DES algorithm. It is an added value to apply one of the biometrics to DES algorithm which is already considered as one of the most resistive to cryptanalysis techniques. The first three formants are formatted in a more complicated and difficult fashion. The suggested transformation of formants is to compute the arc-tangent of the permuted F1, F2, and F3 in pre-determined time slices. The suggested patterns are applied as DES keys in combination with the standard function, f(R,K). Key blocks are chosen from the computed arc-tan angents of F1, F2 and F3 (arc-tanangents are F1/F2, F1/F3, and F2/F3) respectively. An algorithm is designed to interchange the sixteen 64-bit key. Five male persons are the candidate speakers uttering five utterances. The chosen utterances are vowels like 'o', 'e', 'a', 'i', and 'u' which are rich with the recommended formants. These vowels are extracted from the mentioned typical utterances as shown in table-2, in results. For more complication, both speakers and utterances are swapped respectively. Two algorithms are developed to perform the mentioned swapping, one for speakers and the other for utterances. All the mentioned procedures of DES is performed in straight forward fashion in addition to the suggested formant-based key algorithm. The complexity of the added new algorithm is computed to get the overall big- "O" of the proposed DES system. The formant key-based algorithm is depicted in figure-6. Praat with Matlab is the candidate software that are used as speech analyzer for formants extraction mentioned above.**

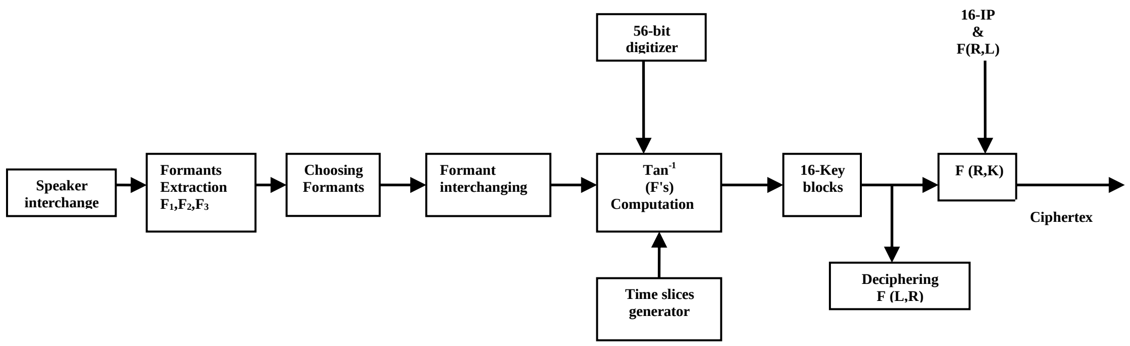


**Figure-6 Block diagram of the formant-based key in DES Scheme**

## VI. Results

The recommended utterances uttered by one speaker of five is considered as a sample. The values of the first the three formants are extracted from typical words. The ac-tan angles of F1-F2, F1-F3, and F2-F3 are computed and tabulated in table-2 and table-3 respectively.

**Table (2) Formants of a speaker uttering characters with typical words**

| utterances | F1 | F2 | F3 | Typical words |
|---|---|---|---|---|
| a | 810 | 1200 | 2390 | hot |
| e | 480 | 1950 | 2450 | bet |
| i | 220 | 2600 | 3110 | beet |
| o | 310 | 680 | 770 | Foo |
| u | 350 | 890 | 2230 | Boot |

**Table (3) arc-tan angles of F1, F2, F3**

| utterances | F1/F2 | F1/F3 | F2/F3 |
|---|---|---|---|
| a | 0.675 | 0.339 | 0.502 |
| e | 0.246 | 0.196 | 0.796 |
| i | 0.846 | 0.071 | 0.836 |
| o | 0.0456 | 0.403 | 0.883 |
| u | 0.393 | 0.157 | 0.399 |

The complexity of the new algorithm is computed as that with the algorithm of DES technique with addition of the proposed algorithm. The most dominant order of DES algorithm is that of the permutation ( order of the factorial computation) added to the order of F(k,R) with that of n- For looping. The add-on algorithm is for formant extraction accompanied with the following computations: swapping of speakers and utterances, time slicing, and arc-tangent calculation and digitizing. All of these algorithms has the most order of almost $N^2$. The trade-off is directed towards the security demand emphasizing the proposed algorithm. The uniqueness of formants as biometric feature make it impossible to be attacked otherwise the formant-based keys are stolen knowing all the pre-described processes. How it comes?. I claimed it is so sophisticated.

## VII. Conclusion & future work

**DES is considered to be insecure due to the 56-bit key size being too small (DES keys have been broken in less than 24 hours). TDES are theoretically attacked, ADES is more robust to cryptanalysis. The degree of susceptibility to attack DES-based algorithms is overcome by the proposed biometric algorithm. It is claimed to be more secured and robust due to the uniqueness for the speaker who sends the message and the extracted formants. To attack this algorithm, it is required to overcome multi-layers: speakers, utterances, formants, digitization technique, sequence of permutation, and the predefined time slices in addition to the core algorithm of DES. The big "O" of the proposed algorithm can be neglected if one considers the trade-off between security and the complexity of the algorithm especially for the critical systems. Military applications, high category E-commerce, E-banking, and political issues can benefit very much by applying this algorithm. Speech mining can be applied in such proposed algorithms to investigate the optimal algorithm for effective cryptanalysis.**

## References

**[1] Fred Moore, "Encryption high-performance high-volume storage"; Computer Technology Review; Oct/Nov 2005, ABI/INFOM Global; pp-17.**

**[2 William Stallings; "Cryptography and Network Security; Principles and Practice"; Prentice Hall; 2003.**

**[3] Data Encryption Standard Wikipedia, the free encyclopedia., 2004, (Knudsen and Mathiassen 2000).**

**[4] D. Coppersmith,"IBM journal of Research and Development; jan/Mar 2000; 44, 1/2 ABI/ C.**

**[5] H. Meyer and S. M. Matyas, "Cryptogrphy: A new dimension in Computer Data Security", John Wiley, New York, 1982. pp 246.**

**[7] Crackin, " DES: Secrets of Encryption Research, "Wiretape Politics, and Chip Design." EFF98, Electronic Frontier Foundation 1998.**

**[8] NIST97 (National Institute of Standard and Technology), "Request for Candidate Register, September 12, 1997. [9]- Bruce Schneier, "Applied Cryptography", John Wiley &Sons, Inc, 2004, ISBN: 471128457.**

**[10] Reaffirmed, "Federal Information Processing Standards Publication", FIPS PUB 46-3 Category: computer Security, Subcategory: Cryptography, 1999 October 25.**

**[11] http://en.wikipedia.org/wiki/Formant, Formant- Wikipedia: the free E ncyclopedia,, 2007.**

**[12] George Musser, "Forming Formants", Scientific American, National Geographic Channel, October 20, 2007.**

**[13] http://person.sol.lu.se/SidenyWood/prrate/whatform.html, Tutorial, Beginners guide to Praat, March 2008**

**[14] Stanely. Chen, Ellen Eide and Michaael A. Picheny, "Advanced Speech Recognition, Topics in signal Processing", September, 22, 2005., ELEN E6884.**

# A Bilingual Approach for Arabic Paraphrases Acquisition: Preliminary Experiments

Rania Al-Sabbagh
re832003@yahoo.com
Faculty of Al-Alsun (Languages)
Ain Shams University, EGYPT

Khaled Elghamry
elghamryk@ufl.edu
College of Liberal Arts and Science
University of Florida, USA

## Abstract

This paper presents preliminary experiments on a bilingual approach for Arabic paraphrase acquisition; a research which is motivated by the importance of paraphrasing for overcoming sparseness of data and its importance for many NLP applications such as Question Answering (QA) and Information Retrieval (IR). The proposed approach develops an unsupervised bilingual algorithm to acquire Arabic paraphrases at the phrase level which is rather more challenging than the elementary word-level paraphrasing and is less efficiently handled by current Arabic paraphrasing systems. Preliminary results show that our approach manages to get term variations – orthographic, lexical and syntactic – for ~ 70% of 4000 randomly selected phrases.

## I. Introduction

To paraphrase is to restate the same information using different lexical and/or syntactic structures. According to Callison-Burch (2007), paraphrasing proves to be an effective technique to overcome the inherent problem of Statistical Natural Language Processing (SNLP), namely sparseness of data. Moreover, it is an essential intermediate task for many Natural Language Processing (NLP) applications such as Question Answering (QA) – discovering paraphrased answers may provide additional evidence that an answer is correct (Ibrahim et al. 2003) – and Machine Translation (MT) (Elghamry 2007).

Paraphrasing is classified into word-based, phrase-based, sentence-based, paragraph-based and text-based paraphrasing. Current experiments focus on phrase-based paraphrasing for two main reasons. First, it is more challenging than lexical paraphrasing (i.e. synonymy identification) which is relatively simple due to the widespread of machine-readable thesauri. Second, the performance of current Arabic paraphrasing systems on phrase-based paraphrasing still needs improvement. Experiments focus on two types of phrases: named entities (e.g. names of organizations, locations, persons ... etc.) and common noun phrases.

According to Callison-Burch (2007), bilingual paraphrasing approaches outperform monolingual ones for many languages including Arabic. However, being based on parallel and/or comparable corpora, these approaches might not be practical for languages with scarce resources like Arabic. Therefore, the proposed bilingual approach tries to dispense with such corpora, meanwhile go unsupervised and robust. Preliminary experiments show promising results about acquiring orthographic, lexical and syntactic phrase-based paraphrases.

The rest of this paper falls in four parts. The first part reviews related work to bilingual paraphrase acquisition. The second part explains the proposed approach, its tools and implementation. The third part shows the used evaluation methodology and results. Finally, the paper ends with a conclusion of the main findings of the preliminary experiments and future work for a full-scale application of the proposed approach.

## II. Related Work

Previous bilingual approaches to paraphrasing relay on one of three resources: multiple translations, comparable corpora (Quirk et al. 2004) and parallel corpora (Callison-Burch 2007). Multiple translations approaches – which are applied to

English and French (Barzilay and McKeown 2001) and English and Chinese (Pang et al. 2003) – assume that different translations of the same source text paraphrase one another. In spite of the promising results achieved by such approaches, the scarcity of multiple translations and the fact that developing them manually is time and effort consuming are obstacles for a full-scale coverage.

Approaches using comparable and parallel corpora achieve better results than multiple translations in terms of coverage, especially for such languages with available corpora such as English (Quirk 2004, Callison-Burch 2007). Callison-Burch (2007) used parallel corpora for Arabic paraphrase acquisition using the only available source for Arabic parallel corpora, namely the LDC Arabic/English Parallel News Text[1]. No clear results are reported on applying this approach to Arabic; however, the approach is used to build the freely available Arabic paraphrase systems Linear B (http://linearb.co.uk/) and Lingo24 (www.lingo24.com).

Practical experience shows that these two systems perform better on the word-based paraphrasing than phrase-based paraphrasing for two reasons. First, parallel and comparable corpora for Arabic, though available, are still scarce. To the best of the authors' knowledge, the only ones available are LDC Arabic/English Parallel News Text[1] and ISI Arabic-English Automatically Extracted Parallel Text[2]. Second, using parallel and/or comparable corpora entails using alignment techniques, which pose

[1] A corpus of Arabic news stories and their English translations collected via Ummah Press Service from January 2001 to September 2004. It totals 8,439 story pairs, 68,685 sentence pairs, 2M Arabic words and 2.5M English words. The corpus is aligned at sentence level. It is available through Linguistic Data Consortium (LDC) catalog number LDC2004T18, URL: http://www.ldc.upenn.edu/

[2] An Arabic-English comparable corpus which is automatically extracted from news articles published by Xinhua News Agency and Agence France Presse. It is obtained using the automatic parallel sentence identification method described in Stefan, D. and Marcu, M. (2005). Machine Translation Performance by Exploiting Non-parallel Corpora, *Computational Linguistics*, Vol. 31. pp. 477-504. The corpus contains 1,124,609 sentence pairs; the word count on the English side is approximately 31M words.

another source of errors. Therefore, the proposed approach tries to avoid alignment and to find an alternative for both parallel and comparable corpora.

## III. The Bilingual Paraphrasing Approach

The proposed approach is based on the same hypothesis previously used by Barzilay and McKeown (2001): different translations of the same source text are paraphrases of one another. However, instead of using corpora of multiple translations, our approach generates necessary multiple translations using current Machine Translation (MT) systems such as Microsoft Translator, Google and Golden Al-Wafi (ATA 2002). It is also assumed that different MT systems use different dictionaries and are trained on different corpora; thus they are likely to yield different translations based on their different dictionaries, corpora and rules.

The approach is straightforward; it does not require any corpus preprocessing tasks and it does not rely on intermediate NLP tools such as POS taggers, NP chunkers or parsers. Therefore, the authors save time and effort; and minimize the sources of errors to one source only, namely the problems of the MT systems used.

Due to lexical and syntactic MT problems, a necessary phase of the proposed approach is MT output validation; that is, to validate the output against documents originally written in the target language (here Arabic). However, even with using Web documents, many rare, yet correct, translations yield zero search hits. For instance, the "National Center for Environmental Research" is translated by Golden Al-Wafi as "المركز الوطني للبحث البيئـي" /*Almrkz AlwTny llbHv Alby}y*/[3], which is a correct translation yet it gets zero search results on Google search engine. Therefore, relying on the regular Web validation technique, which uses the entire phrase as a search query, might not be helpful.

[3] Buckwalter's transliteration scheme (www.qamos.com)

Alternatively, we used a bigram-based term validation technique. We divide each translated phrase into consecutive bigrams and check the validity of each bigram independently on the Web. Each valid bigram is given a score of 1 and each invalid bigram a score of 0 (zero). The validity of the phrase is, therefore, measured as:

$$\textit{Phrase Validity} = \frac{\textit{Sum of valid bigrams}}{\textit{Total number of bigrams}}$$

A score of 1 is the maximum attained for an entirely valid phrase and a 0 score indicates an invalid translation. Intuitively, only phrases giving scores ≥ 0.8 are considered as valid.

Accordingly, given the aforementioned example of the "National Center for Environmental Research", Golden Al-Wafi translates it as "المركز الوطني للبحث البيئي" /*Almrkz AlwTny llbHv Alby}y*/,Microsoft Translator as"المركز الوطني للبحوث البيئية" /*Almrkz AlwTny llbHwv Alby}yp*/and Google as "المركز الوطنى لبحوث البيئة" /*Almrkz AlwTnY lbHwv Alby}p*/. The validation of each translation is measured as such:

| **Al-Wafi** | **Bigram 1** | **Bigram 2** | **Bigram 3** | **Phrase Validity Score** | **Result** |
|---|---|---|---|---|---|
| | المركز الوطني<br>/*Almrkz AlwTny*/ | الوطني للبحث<br>/*AlwTny llbHv*/ | للبحث البيئي<br>/*llbHv Alby}y*/ | 1 | Valid Translation |
| | 1 | 1 | 1 | | |

| **Google** | **Bigram 1** | **Bigram 2** | **Bigram 3** | **Phrase Validity Score** | **Result** |
|---|---|---|---|---|---|
| | المركز الوطني<br>/*Almrkz AlwTny*/ | الوطني لبحوث<br>/*AlwTny lbHwv* / | لبحوث البيئة<br>/*lbHwv Alby}p*/ | 1 | Valid Translation |
| | 1 | 1 | 1 | | |

| | **Bigram 1** | **Bigram 2** | **Bigram 3** | **Phrase Validity Score** | **Result** |
|---|---|---|---|---|---|
| **Microsoft Translator** | المركز الوطني<br>*/Almrkz AlwTny/* | الوطني للبحوث<br>*/AlwTny llbHwv/* | للبحوث البيئية<br>*/llbHwv Alby}yp/* | 1 | Valid Translation |
| | 1 | 1 | 1 | | |

*Table (1): Example of Phrase Validation Process*

Given three valid translations of the same source phrase – "National Center for Environmental Research", the three translations are considered as paraphrases of one another.

To sum up, the algorithm informally goes in four phases:

1. First, compiling source phrases: the source language for the present study is English whose resources – basically parsers and annotated corpora – are quite available.
2. Second, submitting source phrases to MT systems.
3. Third, implementing the bigram-based term validation.
4. Finally, selecting phrases with a score ≥ 0.8.

In spite of the problems of MT systems, using them is expected to achieve better coverage rates than parallel corpora, especially in terms of term variations. Meanwhile, using MT systems makes the proposed approach language independent and thus more applicable. The evaluation methodology and the results of our preliminary experiments and an error analysis are presented in the following subsections.

## IV. Evaluation and Results

In order to test our approach, a list of 2000 named-entities (i.e. names of organizations and locations) is compiled using Google search engine. Another list of 2000 common NPs is extracted from the British National Corpus (BNC). Each list is submitted to each of the used MT systems: Microsoft Translator, Google and Golden Al-Wafi (ATA 2002).

For evaluation, a human rater is used for two purposes: first, to evaluate the MT output (i.e. to judge it as valid/invalid and) and to decide whether valid translations are paraphrasing; second, to measure the agreement rate between the human rater and the results of the bigram-based term validation according to the Kappa Coefficient.

Kappa Coefficient is a statistical measure for the agreement between two raters, taking into consideration the difference between actual or observed agreement and agreement given by chance. It is defined as:

$$Kappa = \frac{P(o) - P(e)}{1 - P(e)}$$

Where
*P(o)* is the probability of observed agreement
*P(e)* is the probability of expected agreement

The human rater and the bigram-based term validation achieve a good kappa rate of ~80%. The main differences between the two raters are among the phrases scoring around 0.8. For instance, the "National Center on Addiction and Substance Abuse" is translated given the following three translations:

1. "المركز القومي للادمان وتعاطي المخدرات" (Google)

   /*Almrkz Alqwmy llAdmAn wtEATy AlmxdrAt*/

2. "المركز الوطني على سوء إستخدام المادة والإدمان" (Al-Wafi)

   /*Almrkz AlwTny ElY sw' <stxdAm AlmAdp wAl<dmAn*/

3. "المركز الوطني بشأن والادمان تعاطي المواد" (Microsoft Translator)

   /*Almrkz AlwTny b$>n wAlAdmAn tEATy AlmwAd*/

According to the abovementioned bigram-based term validation techniques, these translations are given the scores of 1, 0.83 and 0.83, respectively. Thus they considered as valid by the bigram-based term validation, yet the second and third translations are invalid according to the human rater being semantically and syntactically incorrect.

Final results of our approach can be summarized as follows:

| **Paraphrases of score 1 each** | **Recall** | **Precision based on human rater's evaluation** | **F-measure** |
|---|---|---|---|
| | ~ 47.5% | ~ 86% | ~ 61% |
| **Paraphrases of score ≥ 0.8 each** | **Recall** | **Precision based on human rater's evaluation** | **F-measure** |
| | ~ 5% | ~ 50% | ~ 9% |

*Table (2): Final Results*

Generated paraphrases can be divided into three classes: orthographic, lexical and syntactic paraphrases. Orthographic paraphrases are paraphrases with the same lexical and syntactic structures yet with different orthographic forms for such letters as ء /'/ (*hamza*) and ة /*p*/ (*teh marbuta*). Examples of orthographic paraphrases are given in table (3) below:

| **Source Phrase** | **Paraphrase 1** | **Paraphrase 2** | **Orthographic Difference bet. the Two Paraphrases** |
|---|---|---|---|
| National Center for Simulation | المركز الوطني **للمحاكاة**<br>*/Almrkz AlwTny **llmHAkAp**/* | المركز الوطني **للمحاكاه**<br>*/Almrkz AlwTny **llmHAkAh**/* | The last word is written with ة /*p*/ (*teh marbuta*) first and then with ه /*h*/ (*heh*) |
| National Center for Theoretical Sciences | المركز الوطني للعلوم **النظرية**<br>*/Almrkz AlwTny llElwm **AlnZryp**/* | المركز الوطني للعلوم **النظريه**<br>*/Almrkz AlwTny llElwm **AlnZryh**/* | The last word is written with ة /*p*/ (*teh marbuta*) first and then with ه /*h*/ (*heh*) |
| National Bank of Egypt | البنك **الأهلي** المصري<br>*/Albnk **Al>hly** AlmSry/* | البنك **الاهلي** المصري<br>*/Albnk **AlAhly** AlmSry/* | The second word is written with أ />/ (*alef with hamza above*) first and |

| | | | |
|---|---|---|---|
| | | | then with ا /*A*/ (*bare alef: alef with no hamza*) |
| The Egyptian Organization for Human Rights | المنظمه المصرية لحقوق **الانسان** /*AlmnZmh AlmSryp lHqwq* ***AlAnsAn***/ | المنظمة المصرية لحقوق **الإنسان** /*AlmnZmp AlmSryp lHqwq* ***Al<nsAn***/ | The last word is written with ا /*A*/ (*bare alef: alef with no hamza*) and then with إ /*<*/ (*alef with hamza under*) |

*Table (3): Examples of Orthographic Paraphrases*

The second category of the generated paraphrases is lexical paraphrases. These are paraphrases that contain synonymous words like the ones in table (4) below:

| **Source Phrase** | **Paraphrase 1** | **Paraphrase 2** | **Lexical Difference bet. the Two Paraphrases** |
|---|---|---|---|
| National Center for Higher Education Management Systems | المركز الوطني **لأنظمة** إدارة التعليم العالي /*Almrkz AlwTny* ***l>nZmp*** *<dArp AltElym AlEAly*/ | المركز الوطني **لنظم** إدارة التعليم العالي /*Almrkz AlwTny* ***lnZm*** *<dArp AltElym AlEAly*/ | The two synonymous words are: أنظمة /*>nZmp*/ and نظم /*nZm*/; both of which mean systems |
| International Organization for Conservation of Cultural Heritage | المنظمة الدولية **لحماية** التراث الثقافي /*AlmnZmp Aldwlyp* ***lHmAyp*** *AltrAv AlvqAfy*/ | المنظمة الدولية **للحفاظ على** التراث الثقافي /*AlmnZmp Aldwlyp* ***llHfAZ ElY*** *AltrAv AlvqAfy*/ | The two synonymous words are: حماية /*HmAyp*/ and حفاظ /*HfAZ*/; both of which mean conservation |
| Egyptian Association against Torture | الجمعية المصرية **ضد** التعذيب /*AljmEyp AlmSryp* ***Dd*** *AltE*yb*/ | الجمعية المصرية **لمناهضه** التعذيب /*AljmEyp AlmSryp* ***lmnAhDh*** *AltE*yb*/ | The two synonymous words are: ضد /*Dd*/ and مناهضه /*mnAhDh*/; both of which mean against |

*Table (4): Examples of Lexical Paraphrases*

The last category of the resulting paraphrases is the syntactic paraphrases. This means that the same phrase is given in different syntactic structures as in table (5):

| **Source Phrase** | **Paraphrase 1** | **Paraphrase 2** | **Syntactic Difference bet. the Two Paraphrases** |
|---|---|---|---|
| National Center for Environmental Research | المركز الوطني **للبحوث البيئية**<br>*/Almrkz AlwTny **llbHwv Alby}yp**/* | المركز الوطنى **لبحوث البيئة**<br>*/Almrkz AlwTnY **lbHwv Alby}p**/* | The first ends with an Adjectival Phrase (ADJP) whereas the second with a NP |
| National Association of Social Workers | الجمعية الوطنية **لموظفي الخدمات الاجتماعية**<br>*/AljmEyp AlwTnyp **lmwZfy AlxdmAt AlAjtmAE**yp/* | الرابطه الوطنية **للاخصائيين الاجتماعيين**<br>*/AlrAbTh AlwTnyp **llAxSA}yyn AlAjtmAEyyn**/* | The first ends with a NP whereas the second with a ADJP |
| further information | **مزيد من المعلومات**<br>*/**mzyd mn AlmElwmAt**/* | **معلومات أخرى**<br>*/**mElwmAt >xrY**/* | The first is a compound NP which includes an ADJP whereas the second is a simple NP |

*Table (5): Examples of Syntactic Paraphrases*

There are paraphrases that include more than one difference like the phrases included in table (6):

| **Source Phrase** | **Paraphrase 1** | **Paraphrase 2** | **Types of Difference bet. the Two Paraphrases** |
|---|---|---|---|
| National Center for the Preservation of Democracy | المركز الوطني **لحفظ الديمقراطية**<br>*/Almrkz AlwTny lHfZ AldymqrATyp/* | المركز الوطني **للحفاظ على الديمقراطيه**<br>*/Almrkz AlwTny llHfAZ ElY AldymqrATyh/* | Orthographic & Syntactic |
| National Association for Retired Firefighters | الجمعية الوطنية **لرجال الإطفاء المتقاعدين**<br>*/AljmEyp AlwTnyp lrjAl Al<TfA' AlmtqAEdyn/* | الرابطه الوطنية **للمتقاعدين من رجال الاطفاء**<br>*/AlrAbTh AlwTnyp llmtqAEdyn mn rjAl AlATfA'/* | Orthographic & Syntactic |

*Table (6): Examples of Multiple Differences between Paraphrases*

The 30% loss of the performance rate is attributed to two main reasons. First, the three MT systems yielded exactly the same translation for 7% of the tested phrases; and thus no paraphrases were available. Second, 23% of the output translations are linguistically unacceptable; that is, they include lexical and/or syntactic errors.

## V. Conclusion and Future Work

This paper presented the initial experiments for an unsupervised bilingual approach for Arabic paraphrases acquisition. Managing to extract different term variations (i.e. term paraphrases) – orthographic, lexical and syntactic – for 71% of the tested phrases shows that it is a promising approach. It deals with phrase-based paraphrasing which is poorly handled by current Arabic paraphrasing systems and is not limited to the phrases present in parallel corpora. Meanwhile, it does not require much preprocessing or NLP tools.

The main problem of the present approach was the law recall rates. Approximately, 7% of input phrases were given the same translation by all the used MT systems like "World Health Organization"; it was translated by all systems as منظمـة الصـحة العالميـة /*mnZmp AlSHp AlEAlmyp*/ and no system translates it as المنظمـة العالميـة للصـحة /*AlmnZmp AlEAlmyp llSHp*/ which is a correct translation that gets 7,720 search hits on Google search engine. In order to get such paraphrases, the authors expect for future work to integrate the proposed bilingual algorithm with monolingual paraphrasing rules automatically acquired from the bilingually generated paraphrases. Such rules might also contribute to finding paraphrases for the terms mistakenly translated by MT systems.

Although the bigram-based term validation achieves a good kappa rate with human evaluation, there should be more variables to test. For example, trigrams are to be compared with bigrams in terms of precision. The threshold of ≥ 0.8 gets a rather poor precision rate and thus higher thresholds are to be tested together with their effect on recall rates.

## References

ATA Software Technology Ltd. (2002). Golden Al-Wafi Translator Software. Version 1.12 [Online]. Available: www.**ata**soft.com Accessed: 21 April 2001.

Barzilay, R. and McKeown, K. (2001). Extracting Paraphrases from a Parallel Corpus. In *Proceeding of ACL-2001*.

Callison-Burch, C. (2007). Paraphrasing and Translation. Unpublished PhD. School of Informatics, University of Edinburgh, UK.

Elghamry, K. (2007). Machine Translation Oriented Syntactic Normalization of Noun Phrases in Arabic. *Proceedings of Information and Communication Technologies International Symposium (ICTIS07): Workshop on Arabic Natural Language Processing*, Morocco, 2007.

Ibrahim, A., Katz, B., and Lin, J. (2003). Extracting Structural Paraphrases from Aligned Monolingual Corpora. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)*.

Pang, B., Knight, K., and Marcu, D. (2003). Syntax-Based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of HLT/NAACL*

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual Machine Translation for Paraphrase Generation. In EMNLP-2004.

## Appendix: Sample Arabic Paraphrases & their Scores Generated by Our Approach

| Source Phrase | Paraphrase 1 | Score 1 | Paraphrase 2 | Score 2 | Paraphrase 3 | Score 3 |
|---|---|---|---|---|---|---|
| National Center for Public Policy Research | المركز الوطني لبحوث السياسات العامة | 1 | المركز الوطني للبحوث في مجال السياسة العامة | 1 | ------- | |
| National Center for Atmospheric research | المركز الوطني لابحاث الغلاف الجوي | 1 | المركز القومي لبحوث الغلاف الجوي | 1 | ------- | ------- |
| National Center for Higher Education Management Systems | المركز الوطني لنظم إدارة التعليم العالي | 1 | المركز الوطني لأنظمة إدارة التعليم العالي | 1 | ------- | ------- |
| National Center for Health Statistics | المركز الوطني للاحصاءات الصحية | 1 | المركز الوطني لإحصائيات الصحة | 1 | ------- | ------- |
| National Center for Public Productivity | المركز الوطني لمعدل الإنتاج العام | 1 | المركز القومي للانتاجيه العامة | 1 | ------- | ------- |
| The Mobilization of Muslim Women in Egypt | تعبئة النساء المسلمات في مصر | 1 | تعبئة المرأة المسلمة في مصر | 1 | ------- | ------- |
| Federation of Egyptian Chambers of Commerce | إتحاد غرف التجارة المصرية | 1 | اتحاد الغرف التجارية المصرية | 1 | ------- | ------- |
| European Bank for Reconstruction and Development | المصرف الأوروبي للإنشاء والتعمير | 1 | البنك الاوروبي للاعمار والتنمية | 1 | ------- | ------- |

| | | | | | | |
|---|---|---|---|---|---|---|
| social workers | الاخصائيون الاجتماعيون | 1 | موظفو الخدمات اجتماعية | 1 | ------- | ------- |
| Whole World | كل عالم | 1 | العالم كله | 1 | ------- | ------- |
| European cup | الكأس الأوروبي | 1 | كأس اوروبا | 1 | ------- | ------- |
| National Center for Health Education | المركز الوطني للتثقيف الصحي | 1 | المركز الوطني لتعليم الصحة | 1 | المركز الوطني للتثقيف في مجال الصحة | 1 |
| National Center for Environmental Research | المركز الوطني للبحوث البيئية | 1 | المركز الوطنى لبحوث البيئة | 1 | المركز الوطني للبحث البيئي | |

# Improving Tokenization of Clitics in Some Statistical Processing Tools for Arabic: AlwAw Coordinating Conjunction as a Case Example

**Nahed Abul-Hassan**
**Faculty of Arts, Ain Shams University**
**010/8851053**
**nahed.salma@yahoo.com**

## Abstract

Morphological segmentation of clitics is a key first step in syntactic disambiguation in Arabic. Therefore, in this paper, we present a method for improving morphological segmentation, and hence POS tagging, of Arabic words containing the ambiguous form الواو /AlwAw/ ('and'), using *ASVMTools*. Our hypothesis enhances accuracy rate to 97.4% by a single preprocessing step in input text.

**Index terms*:*** Morphological Segmentation, POS Tagging, Clitics, Coordinating Conjunctions, ASVMTools.

## I Introduction

Morphological Segmentation is the process of segmenting clitics from stems. Prepositions, conjunctions, and some pronouns are cliticized onto stems in Arabic [3]. This paper focuses on the morphological segmentation of الواو/AlwAw/ ('and') as a case example. الواو/AlwAw/ ('and') is the most commonly used coordinating conjunction in Arabic and a common source of morphological ambiguity. According to a manual evaluation of a random sample of 100k Arabic word tokens from newswire articles [1](2006), it has been found that الواو/AlwAw/ ('and') alone accounts for approximately 8.6% of any written text.

Unlike the English coordinator *and*, الواو /AlwAw/ can be morphologically ambiguous: it can function as a coordinating conjunction or as part of a word. For example, وحدة/whdp/ can be either وحدة /whdp/ ('unity') or و+ حدة/w + hdp/ ('and intensity'). It is worth noting that / الواوalwaw/ ('and') can be distinguished phonologically to be part of the word or a coordinating conjunction. However, when dealing with written text ambiguity arises.

The rest of this paper is divided as follows. Section 2 gives a brief background about different approaches to Arabic morphological segmentation. The hypothesis and our tools are given in section 3. Section 4 presents an evaluation of our work according to standard evaluation metrics. The conclusion and further suggestions for future work are given in section 5.

## II Related Work

This section represents a literature survey of different approaches to Arabic morphological segmentation and POS tagging, with an emphasis on Automatic Tagging of Arabic Text Using SVM (*ASVMTools*), upon which the work in this paper is based.

### 1 AraMorph

Buckwalter (2002) has introduced AraMorph[2] which applies a dictionary-based approach to Arabic morphological segmentation and POS tagging. In AraMorph, morphological analysis depends on a dictionary of prefixes, a dictionary of suffixes, a stem dictionary, and three checking tables for testing the validity of a word analysis. The system uses Latin characters, as input Arabic words are transliterated, and the linguistic data inside the system are represented in Latin characters as well (using Buckwalter transliteration system) [1].

### 2 Language Model Based Arabic Word Segmentation

Lee et al (2003) have presented a statistical approach for Arabic morphological analysis. They segment a word into prefix- stem-suffix sequence. This system

[1] Alahram Newspaper: http://www.ahram.org.eg

[2] http://www.nongnu.org/aramorph/english/download.html

depends on three linguistic resources: a small corpus manually segmented, a large unsegmented corpus, and a table of Arabic prefixes and suffixes. The authors choose to use the stem, not the root, in their approach. They believe that the stem as a morpheme is more suitable than the root in their applications (information retrieval and translation). A trigram language model is used to segment a word into its component. Their Arabic word segmentation system has achieved an accuracy rate of 97% on a test corpus containing 28,449 word tokens provided by LDC Arabic Treebank (*http://www.ldc.uppen.edu*) [5].

**3 Nizar and Rambow**

Nizar and Rambow (2005) have presented an approach in which they use a morphological analyzer for morphological segmentation and POS tagging of Arabic words. In this approach, morphological segmentation and POS tagging are considered the same operation, which consists of three phases. First, they obtain from their morphological analyzer *(i.e. Almorgeana)* a list of all possible analyses for the words in a given sentence. Then, they apply classifiers for ten morphological features to the words of the text. Then, they choose among the analyses returned by the morphological analyzer by using the output of the classifier [4]. It has been reported that this approach achieves a precision rate of 98.6% (token-based) in morphological segmentation and 94.3% (word-based) in POS tagging.

**4 Automatic Tagging of Arabic Text using SVM (*ASVMTools*)**

*ASVMTools* (2004) have been developed by Diab et al. They provide solutions to fundamental NLP problems such as Morphological Segmentation, Part-Of-Speech (POS) Tagging and Base Phrase (BP) Chunking. Morphological Segmentation (section I) is the process of segmenting clitics from stems, such as separating "ها" /ha/ ('her') from "كتابها" /kitAbahA/ ('her book'). In POS tagging, segmented words have been annotated with parts of speech drawn from the "collapsed" Arabic Penn Treebank POS tag set (e.g. *CC* stands for coordinating conjunctions). BP chunking is the process of creating non-recursive base phrases such as noun phrases, adjectival phrases, etc.

Diab et al have adopted a statistical approach using a language- independent algorithm trained on Arabic Penn Treebank. Arabic Penn Treebank is a modern standard Arabic corpus containing 734 news articles from *Agence France Presse* and covering various topics such as sports, politics, news, etc. Using standard evaluation metrics, they have reported that the Morphological Segmentation has achieved an accuracy of 99.12%, the POS Tagger yields 95.49%, and the BP Chunker has a precision of 92.08%. Morphological ambiguity is not taken into consideration during evaluation.

*ASVMTools* achieve a precision rate of 83.5% in the morphological segmentation of الواو /AlwAw/ ('and'). This is according to a random sample consisting of 3k Arabic word tokens extracted from newswire articles (1999) and processed by *ASVMTools*. See the following example;

| | **Coordinator** | **2nd conjunct** |
|---|---|---|
| Arabic: | اعتقد | و |
| Translit: | /w/ | /AEtqd/ |
| Gloss: | and | he thought |

ASVMTools'output: < wAEtqd/JJ>

In fact, incorrect morphological segmentation produces incorrect part-of-speech tags.

## III Experimental Setup

We assume that by segmenting clitics in input text before being submitted to the *ASVMTools*, we improve both morphological segmentation and POS tagging. This assumption has been applied to الواو /AlwAw/ clitic. Using *Perl* script language, we separate every initial واو /wAw/ in input text, except those that are in lexica. Our hypothesis is that every واو /wAw/ is a coordinating conjunction unless it is part of an entry in lexica, such as الواو /AlwAw/ in وفاة /wfAp/ ('death'), for instance.

We build our lexicon by referring to the following two lexica:

**1 Al-mawrid Lexicon:**

It contains 13553 stems including those for الواو /AlwAw/. It is found within Buckwalter's package for morphological segmentation (2002). Short vowels and diacritics are included in this lexicon.

**2 A Lexicon of proper names & country names:**

The lexicon of proper names is extracted from Al-alasmaa website [3]and consists of a list of 1682 male and female names which are alphabetically arranged. Regarding that of country names, it is acquired through a second language (English). First, it is

[3] http://www.alasmaa.com

extracted from a geography website[4]. Then, the output is submitted to Golden Al-Wafi [5]English-Arabic Machine Translation system, resulting in 477 possible country names.

## V Evaluation

Our hypothesis has been tested on a random sample of 10k tokens from newswire articles (1998). In this sample, one detects 832 instances of الواو /AlwAw/. This sample has been manually evaluated. Table1 shows the results of using our hypothesis and compares them with Diab’s. Improving morphological segmentation has reduced error rate in POS tagging by approximately 7%. Examining errors in our output, we have found that they are due to the fact that Al-mawrid lexicon does not include all word’s derivatives. For example, it does not contain the broken plural وزراء /wzrA}/ (‘ministers’), although it includes the single form وزير /wzyr/ (‘minister’).

| | Precision | Recall | F-measure |
|---|---|---|---|
| Diab’s Tokenizer | 83.5% | 100% | 91% |
| Our hypothesis | 97.4% | 100% | 98.7% |
| Diab’s POS Tag | 87.2 % | 100% | 93.2% |
| Our hypothesis | 93.6% | 100% | 96.7% |

**Table1: Results of our hypothesis**

## IV Conclusion and Future Directions

In this paper, we introduce a preprocessing procedure that would help improve the processing of Arabic. It focuses on the identification of الواو /AlwAw/ through a morphological segmentation of this clitic. Our hypothesis is that every واو /wAw/ is a coordinating conjunction unless it is part of a word that is found in a dictionary of words or of proper names. For future work, we suggest applying this hypothesis to other clitics, such as other coordinating conjunctions, prepositions, pronouns, etc. Moreover, a comparison with other morphological analyzers developed for Arabic can be provided.

## IIV References

1 Anbar, T. **Current Trends in Processing Arabic Morphology**. *In the Proceedings of the Sixth Conference of Language Engineering*, pp.1-15, Cairo, December 2006.

2 Buckwalter, T. **Arabic Morphology Analysis**. http://www.qamus.org/morphology.html, 2002.

3 Diab, M., Hacioglu, K., Jurafsky, D**. Automatic Tagging of Arabic Text: From Raw text to Base Phrase Chunks**. *In the Proceedings of Human Language Technology/North American chapter of the Association for Computational Linguistics*, pp. 1-4, Boston, May 2004.

4 Habash, N., Rambow, O. **Arabic Tokenization, Morphological Analysis, and Part-of-Speech tagging in One Fell Swoop**. *In Proceedings of the **43rd** Annual Meeting of the Association for Computational Linguistics*, pp. 573-580, Ann Arbor, June 2005.

5 Lee, Y., Papineni, K., Roukos, S., Emam, O., Hassan, H. (2003). **Language Model Based Arabic Word Segmentation**. *In the proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 399-406, Ann Arbor , June 2003.

[4] http://geography.about.com/od/countryinformation/a/capital.htm

[5] http://www.atasoft.com

بسم الله الرحمن الرحيم

# معالجة المحتوى المعجمي الدلالي في المعجم العربي الحاسوبي
# مقاربة لغوية حاسوبية

**أ.د. وفاء كامل - أ.د. محسن رشوان- د. عبد العاطي هواري**

## ملخص الورقة:

موضوع الورقة هو **مقترح لمنهجية معالجة المحتوى المعجمي الدلالي للمعجم العربي الحاسوبي** (في صورة تعريفات )؛ إذ يعد الجانب الدلالي أهم جوانب صناعة مصدر معجمي، لعدة أسباب هي:

. كونه الغايةَ الأساسية من أي مصدر معجمي، سواء الذي يستهدف المستعمل البشري، أو المعالجة الحاسوبية.

. تعَقّد جوانبَ المحتوى المعجمي واتساعها لتشمل جانبا من المحتوى الصرفي الدلالي، والجانب التركيبي، إضافة إلى الجانب الدلالي نفسه.

. أن المحتوى المعجمي الدلالي في المعجم هو مصدر المعالجات الحاسوبية، التي تهدف إلى الفهم الآلي للغة، ومعالجة جوانبها الدلالية.

وقد اقتُرِحت تصورات متعددة لصناعة مصادر معجمية عربية حاسوبية[1]، غير أن مقارباتِها الجوانبَ المعجمية الدلالية جاءت تقليدية وغير معمقة، بل إنها تكاد تكون نَسْخا لمقاربات المعجم العربي الأساسي والمعجم الوسيط[2]، فضلا عن المعاجم العربية التراثية.

وما من شك في أن اللسانيين الحاسوبيين يقعون دائما في مشكلات عند تعاملهم الحاسوبي مع المصادر المعجمية العربية. كما أن المصادر الحاسوبية منها تقليديةٌ في معالجتها لهذا الجانب أيضا؛ لذا كان لزاما– عند إرادة البدء في صناعة مصدر معجمي عربي– البدء ببناء منهجية لمعالجة المحتوى المعجمي الدلالي، الذي يعد أهم مشكلات المعجم العربي عموما؛ إذ إن القضايا الدلالية في هذا المعجم قد عولجت دون منهج واضح.

وينبني **التصور النظري للمقترح** على عدة جوانب:

1. المنطلقات النظرية : الصرفية والتركيبية والدلالية.
2. التحليل : مدونته، وإجراءاته، ومخرجاته.
3. التمثيل : لغته، وأبنيته.

[1] يوجد اهتمام في الوقت الحالي بإنجاز معجم عربي حاسوبي تتبناه منظمة الأليكسو بالتعاون مع مدينة الملك عبد العزيز للعلوم والتقنية بالسعودية ومجمع اللغة العربية بدمشق. ورغم طموح المشروع في جوانب عديدة نجد الأوراق البحثية المقدمة في المؤتمرين، الذين جعلا لأوراق خبراء المعجم الحاسوبي، تهمل- أو تقصر- في معالجة الجوانب الدلالية للمعجم. وستعرض الورقة لذلك في موضعه.

[2] يراجع تقرير المعجم العربي التفاعلي: http://www.almuajam.org/index2.htm

**مدخل:**

تُعَد المصادر المعجمية Lexical Resources ، على تنوعها، البنيةَ التحتيةَ Infrastructure لأنظمة معالجة اللغات الطبيعية Natural Language Processing واللسانيات الحاسوبية Computational Linguistics[3]؛ لأنها تقدم لها المعارف اللغوية التي تمثل لبنات الجملة والنص اللغوي. وقد "أصبح للمعجم، في مجال اللسانيات الحاسوبية، موقع محوري بل أصبح المعجم مفتاح الدلالة، ومحور التوليد اللغوي. وقد اقتضاه هذا الموقع أن يقنن منهج الشرح المعجمي على نحو يتجاوز ما كان متعارفا من أمر المعجمات التقليدية"[4]. فالمعجم يمثل عنق الزجاجة: من وجهة نظر جل العاملين بحوسبة اللغة.

وبالرغم من تنوع أشكال المصادر المعجمية، واختلافها في الأطر النظرية الموجهة لها، إلا أنها تتفق جميعها في الاهتمام بجانب المحتوى المعجمي الدلالي، مع اختلاف في طريقة تمثيله. إذ يقع المحتوى المعجمي الدلالي في قلب اهتمام معالجة اللغات الطبيعية باعتباره غايةَ أيِّ تواصلٍ لغوي، كما أن غاية أي مشروع معجمي، مهما كانت أسسه النظرية وغاياته العملية، هي تمثيل هذا المحتوى.

**مبررات المقترح:**

تتعدد المبررات التي تجعل من اقتراح منهجية لتحليل المحتوى المعجمي الدلالي وتمثيله، ومن طبيعة المعالجة المتبناة فيها، ضرورة بحثية. وهذه المبررات هي:

- وجود مشكلات في معالجة الجوانب المعجمية الدلالية في المعجم العربي التقليدي والحاسوبي، مردها غياب التصور النظري المتكامل للإشكالية، الذي يكون أساسا للتطبيق المعجمي. وهو ما لاحظه أحد الباحثين إذ يرى أن " قراءة أي مقدمة من مقدمات المعاجم المعاصرة، تجعلنا لا نقف على إشارة إلى المنهج المراد استثماره في تعريف المداخل [...] مما يجعل إعادة صياغة التعريف في المعاجم المعاصرة عملية نادرة أو غير معممة"[5]، ويرى بعض الباحثين أن من أهم أسباب التخبط في التطبيق المعجمي العربي (سذاجة المعجم في التهاون بقضايا شديدة التأثير، لاسيما قضية التعريف وتفاصيلها المتعددة)[6].
- هذه المشكلات لا تكاد دراسةٌ توجهت إلى المعجم العربي إلا أشارت إلى نقائص معالجتها المحتوى المعجمي، وقد خلصت معظم هذه الدراسات إلى ما يشبه التوجيهات والنصائح أو الاقتراحات، قدمتها لكل من يَرْغب في تأليف معجم. وكان من بين هذه النصائح أو التوجيهات ما

[3] N. Calzolari: Semantic web vision

[4] نهاد الموسى؛ العربية: نحو توصيف جديد في ضوء اللسانيات الحاسوبية. ص 252

[5] حلام الجيلالي: تقنيات التعريف في المعاجم العربية المعاصرة. ص:111

(6) محمد رشاد الحمزاوي؛ المعجم العربي إشكالات ومقاربات. ص 286

يخص التعريف[7]. على أنه لا يمكن الاكتفاء بالنقد دون التصدي لإنشاء معجم موسع، مبني على تحليل معجمي دلالي معمق، وتمثيل مصوغ بلغة تعريف منضبطة، يصلح أساسا معتمدا للبحث والتوظيف في معالجة اللغات الطبيعية.

- افتقار المجتمع الحاسوبي اللساني إلى مصادر معجمية عربية[8] تواكب الطفرة الحادثة في مجال الويب الدلالي Semantic Web ، ومرد ذلك إلى طريقة معالجة المحتوى المعجمي الدلالي العربي تحليلا وتمثيلا.
- عدم جدوى توظيف منهجية مقترحة في بيئة لغوية مختلفة عن العربية، إذ إن التعريف يحتوي جل المعرفة المعجمية بوصفه تمثيلا لمحتوى الوحدة المعجمي، الصرفي والتركيبي والدلالي. لذا فإن عملية التأسيس النظري لتعريفات معجم، وإجراءات ضبط محتواها ولغته لابد أن تنبع من تصور نظري خاص بالطبيعة الصرفية والدلالية والتركيبية للغة: فلكل لغة طبيعتها الصرفية الدلالية التي تفرض طبيعة المعالجة النظرية لتعريفات معجمها. وتفرض اللغة العربية، بطبيعتها الصرفية الاشتقاقية، تأسيسا نظريا لتعريفات معجمها يتلاءم وطبيعتها من جهة، ويكون قابلا للتطبيق بفاعلية على مستوى المعجمين التقليدي، والحاسوبي من جهة أخرى.
- غياب دراسة عربية- على حد علمنا- تتصدى لمعالجة المحتوى المعجمي الدلالي العربي تحليلا وتمثيلا، في صورة تعريفات معجمية.

**جدوى الدراسة:**

- على المستوى النظري:

إن فحص المعجم العربي في ضوء التصورات النظرية الغربية كفيل بأن يقدم خدمات جليلة للجانبين المعجميين الغربي والعربي. فنظرية مثل نظرية بوسطيوفسكي Pustejovsky (وهي نظرية- شأنها شأن معظم التصورات المقترحة في الإطار اللساني التوليدي- تدعي العالمية universality، أو قابليتها للسحب على اللغات المختلفة؛ لأنها تتحدث عن وسائط Parameters عوض الحديث عن قيم معينة Attributes؛ مما يجعلها، في تصور أصحابها، صالحة للتطبيق على أي لغة؛ لتوجهها إلى الجانب الذهني أكثر من توجهها إلى التحققات الواقعية لمعجم بعينه) لم تثبت صلاحيتها عند درس المعجم العربي بها ودرسها بالمعجم العربي؛ فقد وجدت فجوات في كلا التصورين؛ مما جعل الدراسة

[7] حسين نصار: المعجم العربي نشأته وتطوره، تمام حسان: اللغة العربية معناها ومبناها ص: 320 ، محمد فتيح: في الفكر اللغوي. ص:270، محمد أحمد أبو الفرج: المعاجم العربية في ضوء علم اللغة الحديث ص: 178.

[8] أشار تقرير: "نملار" NEMLAR في دراسته المسحية للمصادر المعجمية العربية الحديثة، في إطار معالجة اللغات الطبيعية، إلى وجود نقص حاد في المصادر المعجمية العربية المعمقة دلاليا التي يمكن استثمارها في تطبيقات معالجة اللغة الطبيعية Natural Language Processing.

تقترح البنية التعريفية الصرفية[9].

- على مستوى معالجة اللغات الطبيعية:

يساعد تعميق المحتوى المعجمي الدلالي وضبط لغته في إنجاز كثير من التطبيقات اللسانية الحاسوبية، مثل فك اللبس المعجمي الدلالي[10] Disambiguation، والتوليد Generation، والتحليل Analysis.

**أهداف الدراسة:**

1. اقتراح منهجية لمعالجة المحتوى المعجمي الدلالي العربي (تحاول أن) تفي بمتطلبات المعالجة الآلية للغة العربية[11]، وتشمل:
   - جوانب التحليل، لاستخلاص المحتوى المعجمي الدلالي وتعميقه.
   - جوانب التمثيل، لصياغة معالم لغة للتعريف المعجمي، تكون منضبطة وقابلة للتعامل الحاسوبي معها.
2. اقتراح تطبيق لساني حاسوبي يفيد من نتائج المنهجية المقترحة في تطوير المعجمية الحاسوبية ( نظام محرر التعريفات العربية Arabic Lexical Definition Editor System).

**المنهج وطبيعة المقاربة:**

تتبنى الدراسة، في معالجتها المحتوى المعجمي الدلالي العربي، المنهج التوليدي Generative، حسب بوسطيوفسكي Pustejovsky. وينبغي هنا أن تقدم الدراسة احترازين:

- أنها أفادت من مناهج معجمية دلالية سابقة أخرى: في بعض فروضها ومقولاتها النظرية، وفي توظيف بعض إجراءاتها، ومن هذه المنهجيات المقاربة العلاقية Relational

[9] ظهرت الجوانب الصرفية في النظريات المعجمية الدلالية الغربية محكومة بطبيعة لغاتها وإشكالاتها، فخلا تصور بوسطيوفسكي من الجانب الصرفي تقريبا، كما أن معالجة ميلتشوك للجانب الصرفي في نموذجه محدود بإشكالات اللغة الفرنسية (اللواصق والتصريف) وقد نوقشت هذه الجوانب في متن الدراسة.

[10] تقصد الدراسة بمصطلح "المعجمي الدلالي" المحتوى الصرفيَّ الاشتقاقي الدلالي والمحتوى التركيبي الدلالي، إضافة إلى المعنى.

[11] يمكن هنا إيراد نتائج إحدى الدراسات التي كان موضوعها علاقة المحتوى المعجمي بمعالجة اللغات الطبيعية: "ينبغي أن تتضمن معالجة للمستوى الدلالي في صناعة معجم محوسب على:إنشاء روابط دلالية جلية، ضمان الاتساق في محتوى المداخل، في لغة نصوص التعريف، في تقسيم المعاني"
Nancy Ide & Jean Véronis : Machine readable dictionaries: What have we learned, where do we go ?

Approach، وإجراءات التحليل التكويني Componential Analysis، ووظفت كل ذا في إطـار تصـور خـاص للمعجم العربـي ومنهجيـة تحليلـه وتمثيلـه. كمـا أن الدراسـة أفـادت مـن المنجز المعجمي العربي (من خلال تحليل نصوص مدونة تعاريف المعجم الوسيط والمعجم العربي الأساسي)، والمنجز العربي النقدي والتنظيري في مجالات المعجم والاشتقاق والدلالة.

- أنَّ تبنـي المنـهج التوليدي فـي المقاربـة لا يعنـي التطبيق الكامل لأفكاره وإجراءاتـه، بل يعنـي اقتناعا بفكرته الأساسية، وهي كيفية توصيف الوحدات المعجمية توصيفا يسمح برصد آليات تـآلف الكلمـات فيمـا بينهـا علـى المستوى الـدلالي؛ للسيطرة علـى المعـاني الإبداعيـة الجديـدة (غير المنصوص عليها في المعاجم ). فالاهتمام يكون بأنماط الدلالة الممكنة وآليات وقوعها أكثر من الاهتمام بالمعاني الكائنة والاعتناء بسردها في المعجم. غير أن الدراسة قد أضـافت بنية أخرى على تصور بوسطيوفسكي لأبنية التمثيل المعجمي الدلالي وهي البنية الاشتقاقية، كما أعادت النظر في بنية التوارث لتقصيرها عن تمثيل البنية المعجمية الدلالية العربية.

طبيعة المقاربة اللسانية:

هي مقاربة قائمة على المدونة النصية  Corpus based ، تستخلص معارفها اللغوية بتحليلها.

طبيعة المقاربة اللسانية الحاسوبية:

المجتمـع اللسـاني الحاسـوبي يعـرف مقـاربتين: الأولـى إحصـائية Statistical Based، وأخـرى قواعدية Rule Based أو معرفيةKnowledge Based . والدراسة تنتمي إلى النوع الثاني، إذ يذهب أصـحاب هذا التوجـه إلـى أنـه يمكن صـياغة المعـارف اللغويـة أو تمثيلها فـي نماذج أو قوالـب Modules منتظمة للحاسوب، بطريقة تحاكي الطريقة التي يتصرف بها الإنسان لغويا فهما وتوليدا. وهذه المقاربـة تعتمـد– أساسـا– علـى بنـاء مصـادر لغويـة، معجميـة أو نحويـة، تحتـوي توصـيفا مفصـلا وجليا للغة.

- **أهم معالم التصور المقترح:**

لا يرفض التصورُ المقترحُ الأعمال المعجمية والمناهج المقترحة المنجزَة رفضا تاما بل  يفيد منها لتحقيق غايات الدراسة من تعميق التحليل للبنية المعجمية الدلالية. وتدقيق لغة التمثيل، بما يجعله صالحا للتناول الحاسوبي. ويمكن إيجاز أهم معالم المنهجية المقترحة فيما يلي:

- يتوافق التصور المنهجي المقترح والفكر البرمجي والتطويري السائد في الفترة الأخيرة : فكر البرمجة الكائنية التوجه  Object Oriented Programming في قيامها على فكرة الفئة Class والكيان Object ، فالأول يحمل كل سمات النمط، والثاني يُشْتَق منه لكونه يعد واحدا من عدد لا نهائي من تحققاته.

وقد ظهر هذا الجانب في الدراسة في افتراض نمطين للوحدات المعجمية الاشتقاقية الأصول والفروع أو المشتقات المطردة، ومن ثم لتعريفاتهما بافتراض نمطين من التعريفات: تعريفات أصلية للأصول، وأخرى فرعية لمشتقاتها تُشتق منها.

- حاولت الدراسة الالتزام بمجموعة من الأفكار الحاسوبية التي تُكَوِّن الذهنية الحاسوبية أو الفكر الحاسوبي، مثل القول بالذرية في تحليل الوحدات المعجمية (السمات الدلالية)، وفي تمثيلها (الوحدة التعريفية Defimes)؛ لاتساق ذا مع أفكار التركيب والتوليد والقابلية للتحليل، والصياغة البنيوية، والتقليص في المعلومات بعدم الاضطرار للجوء لتكرارها، والقابلية لإعادة الاستخدام Reusability .

**إجراءات التحليل ومدونته وأدواته:**

اتخذت الدراسة مدونة لتطبيق مقولات التحليل عليها[12].

- **بنية المدونة:**

اعتمدت الدراسة مدونتين للتحليل:

- مدونة لغوية نصية Textual Corpus.
- ومدونة نصوص تعريفات معجمية، من المعجم الوسيط والمعجم العربي الأساسي.

وتكونت مدونة النصوص من حوالي 80000 ألف كلمة وشملت نصوصا عربية فصيحة حديثة ومعاصرة، وتوزعت على مجالات متنوعة كالأدب والدين والسياسة والطب والرياضة والعلوم الإنسانية والعلوم الطبيعية[13].

وقامت الدراسة بمعالجة مدونة فرعية Sub corpus من المدونة النصية العامة[14] هي الكلمات التي يبدأ جذرها بحرف الجيم[15] باعتبارها نموذجا للمدونة النصية الكبرى، كما عالجت مدونة من نصوص التعريفات المعجمية العربية.

- **التحليل وأدواته:**

[12] اعتمدت الدراسة الجذر في تحديد المدونة الفرعية، وذلك لأنها لا ترفض فرضية الجذر باعتبارها فرضية تصنيفية، لكنها ترفض اعتدادها فرضية تفسيرية صالحة لتأسيس التنظيم الداخلي للمعجم العربي عليها.

[13] اعتمدت الدراسة مدونة محدودة لكن عند التطبيق الفعلي للعمل يجب تحقيق أكبر حجم ممكن من النصوص.

[14] اكتفت الدراسة بعينة من المدونة بما يتسق وموضوع الدراسة وأهدافها، وكانت الغاية من هذه الدراسة أن تكون عينة للتحليل وتطبيق التصور النظري عليها.

[15] اعتمدت الدراسة الجذر في تحديد المدونة الفرعية؛ وذلك لأنها لا ترفض فرضية الجذر باعتبارها فرضية تصنيفية، لكنها ترفض اعتدادها فرضية تفسيرية صالحة لتأسيس التنظيم الداخلي للمعجم العربي عليها.

اتخذت الدراسة أربعة مستويات من التحليل تتآزر جميعها لتحليل المحتوى المعجمي الدلالي لوحدات المعجم، وهذه المستويات هي:

- المستوى الصرفي(الاشتقاقي): وشمل تحليل العلاقات الاشتقاقية الدلالية، واقتراح اتخاذ أصل الاشتقاق كلمة متحققة لا مجرد جذر، وتبع ذلك اقتراح لإعادة تنظيم البنية الصغرى Microstructure للمعجم العربي.
- المستوى التركيبي: ويشمل الوسم التركيبي الذي يحدد قسم الكلام الذي تنتمي إليه الوحدة المعجمية، وتوضيح طبيعة البنية المعجمية للحدث Event، وتوضيح القيود الانتقائية، وطبيعة المكملات التي تحتاج الوحدة المعجمية المعرفة إليها.
- المستوى الدلالي: ويشمل توضيح الحقل الدلالي للوحدة المعجمية، والعلاقات الدلالية التي تشترك فيها، واستخلاص السمات الدلالية.
- المستوى المعجمي: ويشمل تحديد أنماط الوحدات المعجمية بوصفها معرَّفات.

وقد استخدمت الدراسات أدوات حاسوبية للتحليل هي المحلل الصرفي Morphological Analyzer، والمفهرس الآلي Concordancer، وبرنامج لاستخلاص التصاحبات اللفظية N-gram إضافة إلى التحليل اليدوي.

**جوانب التمثيل (لغة التعريف):**

تتكون لغة التعريف – قياسا على اللغة الطبيعية – من معجم لوحدات التعريف Defimes، ونحو لجملة التعريف Definition Syntax، ومجموعة من أبنية التعريف وتقنياته.

فقامت الدراسة بوضع معالم **معجم لوحدات التعريف** Defimes ، يمكن أن يستخدم في تعريف مفردات المعجم.

**ونحو التعريف** Definition Syntax يعني اقتراح أنماط تركيبية لجملة التعريف بحيث تحقق غايتين: الأولى هي تحقيق الاتساق في التحرير، والثانية مترتبة عليها وهي أن يكون نص التعريف قابلا للتحليل الجملي الآلي Parsing.

**وأبنية التعريف** Definition Structure هي مكونات التعريف الفرعية، وتتوازى مع مستويات التحليل. وخرجت الدراسة بالتصور التالي لبنية التعريف:

وقد اعتَبَرَتْ الدراسةُ صياغةَ البنية المعجمية الدلالية Lexical Semantic Structure موضوعا مركزيا في التمثيل، فتتبعت عناصر صياغتها على المستويات المختلفة للتحليل والتمثيل.

## أولا: أهم مقومات تصور نظري لتحليل المحتوى المعجمي وتمثيله:

تقترح الدراسة مجموعة من الشروط التي ينبغي أن تتوفر في أي تصور نظري يحاول أن يقارب المحتوى المعجمي الدلالي بغرض تمثيله في صورة تعريفات معجمية. فينبغي أن تتوافر فيه المقومات الآتية:

- تصور شامل لمنظومة اللغة، ومنظوماتها المكونة، وموقعية المعجم بينها.
- تصور واضح لبنية معجم اللغة كياناتٍ وعلاقات، وآليات؛ لكيفية صياغة هذه البنية وتحقيقها في المتن المعجمي.
- تصور لطبيعة المحتوى المعجمي الدلالي (الصرفي والدلالي والتركيبي) والعلاقة بين مكونات هذا المحتوى، وكيفية تعاونها في تمثيل المحتوى0 وتصور لطبيعة المعنى.
- تصور لمنهجية التحليل: منطلقات نظرية وإجراءات تطبيقية، ووحدات التحليل لكل مستوى من مستويات اللغة.
- تصنيف لأنماط المعرفات، باعتبارها أولى خطوات تنظيم مادة المعجم، وتصنيف كياناته وأنماطه المَعرِفية.
- تحديد لمتطلبات كل صنف من أنماط المعرفات، وطبيعة كل نمط، وسماته الشكلية والمضمونية.
- إجراءات لتحليل المحتوى: سواء أكانت إجراءات يدوية أو إجراءات حاسوبية، تكون متسقة مع المنطلقات النظرية من جهة، وفعالة منتجة من جهة أخرى.
- استراتيجيات للتمثيل: فالمعلومات المعجمية الدلالية ليست صنفا واحدا بل أصناف متنوعة متباينة، كما أن الوحدات المعجمية نفسها أنماط متباينة، لكل نمط منها خصائصه.
- معجم للتمثيل: يشتمل على الوحدات التي يمكن تمثيل الوحدة المعجمية بها.
- قوالب أبنية تركيبية تمثيلية ( مستوعبة، ممثلة، نسقية ).
- رؤية واضحة لحدود التصور وأهدافه، وموقفه من التطبيقات الحاسوبية.
- آليات تكون الكلمة/ الوحدة المعجمية: الاشتقاق (وأثره في بناء المعجم والبنية الصغرى)، النحت، الاقتراض، الاصطلاح.
- الاطرادات الصرفية والدلالية، وكيفية توظيفها في تقليص حجم المعجم (Dictionary).
- تصور لوحدة التحليل في كل مستوى.
- تصور لوحدة التمثيل.
- تصور لبنية التعريف ومكوناته الجزئية.

## أولا: المنطلقات النظرية: الصرفية والتركيبية والدلالية:

تعددت التصورات النظرية التي قاربت الدلالة المعجمية بغرض تحليلها وتمثيلها، وهي في تعددها لم تبدأ واحدة منها من فراغ، بل أفاد كل تصور لاحق من التصورات السابقة. وقد أفاد التصور النظري المقترح من تصورات نظرية متعددة؛ بغرض توظيفها في إطار التصور، مع المحافظة على الاتساق الداخلي للتصور. فالتصور متأثر – في كثير من منطلقاته وإجراءاته – بالتوجه التفكيكي بدءا من كاتز وفودر في تصورهما عن المدخل المعجمي(16)، وأفكار تفتيت المعنى لدى بولينجر(17)، وأفكار تحليل المكونات لدى يوجين نايدا(18).

### 1. المنطلقات النظرية:

أهم المنطلقات النظرية للدراسة:

**مفهوم الدراسة للتعريف:** هو توصيف المحتوى المعجمي الدلالي للوحدة المعجمية في ذاتها، وفي علاقاتها داخل بنية المعجم.

**البنية المعجمية الدلالية:** هي إحدى غايات التصور النظري، ويمكن تعرفها في مرحلة التحليل عن طريق التوصل إلى الأنماط التي تحكم اللغة، مثل: أنماط الوحدات المعجمية، وأنماط الصيغ الصرفية. كما يمكن صياغتها على مستوى التمثيل عن طريق مجموعة من التقنيات:

. **تنميط المُعَرَّفات:** أول عمل ينبغي أن ينجز في سبيل تنميط التعريفات، والصياغة البنيوية للتعريف.

. **فصل المشترك اللفظي** Homonyms بعد تمييزه عن متعدد المعنىPolysemy.

. **تصنيف الوحدات المعجمية** في حقول دلالية.

. **تجلية العلاقات الدلالية:** الأفقية (مثل: الترادف والتضاد)، والرأسية أو الهرمية (مثل: علاقة الاشتمال، والتضمن).

### 2. في طبيعة المحتوى المعجمي الدلالي:

[16] Kats & Fodor, (1963), the structure of semantic theory .P: 178.

[17] Dwight Bolinger, The Atomization of Meaning. P:22.

http://www.jstor.org/pss/411524

[18] Nida, Eugene, A Componential Analysis of Meaning. An introduction to semantic structures p: 13.

**الصرف:** تعد الصيغة الصرفية البنية التحتية للمحتوى الدلالي، كما أن للمقولة الصرفية دورا كبيرا في تحديد المعنى. وفي مستوى الصرف اعتمدت الدراسة الأفكار التالية:

. **الوحدات المعجمية** نوعان: اشتقاقية لها عمق صرفي دلالي، وأخرى غير اشتقاقية لا تحتوي دلالة صرفية.

. **الوحدات المعجمية الاشتقاقية** نوعان: أصول وهي الوحدات المعجمية الأبسط والتي يشتق منها وحدات معجمية أخرى، وفروع وهي الوحدات المعجمية التي تشتق من الأصل.

. ينبغي أن يعاد النظر في **بنية المداخل المعجمية العربية**، وبالتالي طريقة تنظيم بنية المعجم الكبرى.

. **الجذر** فرضية تصنيفية لا تفسيرية.

. **الصيغة الصرفية** ( للوحدات الاشتقاقية) هي الوحدة الدلالية الصرفية الصغرى.

. **المشتقات الاسمية** صنفان: مشتقات مطردة، ومشتقات متحولة معجميا.

. **المقولة الصرفية** هي الوظيفة الصرفية الدلالية (مثل اسم الفاعل، اسم المفعول... الخ).

**الدلالة:** هي مركز المعالجة المعجمية الدلالية.

وقد أفادت الدراسة من مقولات الحقول الدلالية، وتحليل المكونات، وتصورات التوجه العلاقي، في بيان البنية المعجمية الدلالية عن طريق العلاقات الدلالية.

**التركيب:** يرتبط التركيب بالدلالة ارتباطا شديدا؛ نظرا لتوقف بعض الدلالات على المحتوى التركيبي، والعكس أيضا صحيح: فالدلالة تتحكم في النمط التركيبي الممكن للوحدة المعجمية. واتخذت الدراسة في هذا الجانب توجها سياقيا أفقيا Syntagmatic، فحاولت تنميط القيود الانتقائية، واستقصت أنماط بنية الحدث للفعل.

## ثانيا: التحليل : مدونته، وإجراءاته، ومخرجاته.

### مدونة الدراسة:

المدونة اللغوية Corpus[19] مجموعة من نصوص متجانسة، تُجمع على أسس معينة يحكمها غاية المدونة نفسها واستعمالها، وتطبيقاتها المحتملة. تنظم في شكل معين (شكل حاسوبي في الغالب)، تكون مادة للتحليل اللغوي ( أو الثقافي والحضاري والنفسي)، واستخلاص الظواهر حاسوبيا.

والمدونة النصية لا تقدم معلومات مباشرة عن اللغة، ولكن التحليل ( اليدوي أو الحاسوبي) هو ما يستخلص هذه المعلومات؛ لذا يستعين العاملون في مجال المدونات الحاسوبية بأدوات حاسوبية تقوم بإجراءات التحليل: من تصنيف وإحصاء وتصنيف للمادة. والمدونات المحوسبة من حيث المعالجة نوعان: مدونة ذات تعليقات (Annotated)[20]، وأخرى خام Raw.

ولقد غدت المدونات اللغوية المحوسبة، بعد الثورة اللسانية الحاسوبية، إحدى العلامات المميزة للسانيات عموما، واللسانيات الحاسوبية على وجه الخصوص؛ حتى ظهر مجال لساني جديد أُطلِقَ عليه لسانيات المدونة Corpus Linguistics، " لقد حَوَّل الحاسوب المدونةَ النصيةَ (التقليدية) من مجرد الاهتمام بموضوع بحثي تخصصي واحد إلى كونها أصبحت مفتوحة على كل الموضوعات إلى أن أحدثت المدونات المحوسبة ثورة في دراسة اللغة "[21].

والمدونة في أحد تصوراتها هي اللغة من خلال نماذج ممثلة لها، لذا ينعم المتخصصون النظر في اختيار نصوصها، وتصميم قاعدة بياناتها. فالمدونة النصية يمكن دراسة مستويات اللغة من خلالها، كما يمكن دراسة ظاهرة أو موضوع من خلال مادتها.

وقد ارتبطت المدونات النصية المحوسبة، أكثر ما ارتبطت، بصناعة المعاجم؛ فلم يعد من المستساغ (في الغرب) الحديث عن معجم دون البدء بالحديث عن مدونته النصية.

### تصميم المدونة النصية:

[19] تعددت المقابلات العربية لهذا المصطلح الإنجليزي ومنها المكنز، والذخيرة اللغوية، والمدونة الذي اعتمدته الدراسة، وقد اعتمدته لكونه المقابل العربي الأكثرشيوعا واستعمالا في أوساط اللسانيين المختصين، وكأنه نوع من الإجماع. إضافة إلى أن المصطلحات الأخرى المستخدمة لها دلالات أخرى في حقل اللسانيات، فمثلا مصطلح مكنز قد يقابل Corpus وقد يقابل Thesaurus.

[20] تفضل الدراسة المصطلح العربي "التحشية" مقابلا للمصطلح الإنجليزيAnnotation ، ويسوغ هذا الاختيار مناسبة المحتوى الدلالي للكلمة العربية لمحتوى المفهوم ( تدوين التعليقات والتصنيفات والملاحظات على متن المدونة)، وما للمصطلح من معرفة واستعمال في الثقافة العربية، مع اتفاق كبير في المضمون مع المصطلح العربي التراثي، إضافة إلى كون المقابل العربي غير مستقر أو متفق عليه.

[21] Susan Hunston: Corpora in Applied Linguistics p:1

تنوعت مادة المدونة من حيث المجال: (أدب، دين، سياسة واقتصاد، إنسانيات، فنون، رياضة، علوم وتقنية، جغرافيا وفلك)، ومن حيث الجنس الأدبي للمادة الأدبية: (رواية، قصة، مقالة)، ومن حيث تخصص الكُتَّاب (رجال، إناث)، ومن حيث التوزيع الجغرافي (مصر ، الجزائر ، المغرب، سوريا)، ومن حيث طبيعة الوسيط الإعلامي: (كتب، صحف، مجلات، مواقع إلكترونية)، ومن حيث طبيعة النص (عربي، مترجم إلى العربية).

## وصف المدونة النصية:

ليس من غاية الدراسة إنشاء مدونة، غير أن اعتمادها على المدونة أمر ضروري لازم على المستوى المنهجي، فكان من الممكن، أن تعتمد الدراسة إحدى المدونات العربية المنجزة بالفعل، وتُجري عليها تحليلاتها. غير أن الواقع العملي يتناقض مع ذلك: فمعظم من يمتلكون مدونات لا يتيحونها للاستعمال العام، وما يتاح منها يتاح بمقابل مالي باهظ، مع عدم قيامه على أسس تتوافق مع منهجية البحث[22].

على أن قيمة أي مدونة تعود، بعد دقة تصميمها وعمق تأسيسها اللساني، إلى التحليلات المجراة عليها: من وسم تركيبي Part of speech tagging ووسم دلالي Sense tagging ، وتعليق وإضافات Annotation صرفية دلالية وتركيبية.

والدراسة الحالية لا تحتاج لمدونة ذات تعليقات وحواش ، بل تحتاج إلى مدونة غير معالجة Raw Corpus ، فكان قرار إنشاء مدونة نصية (مبسطة) تؤدي المهام التي تفرضها إجراءات الدراسة في التحليل والتمثيل.

ويمكن توصيف المدونة بأنها مدونة للغة العربية الفصيحة في العصر الحديث، من خلال مجموعة من الأعمال الأدبية والدراسات العلمية والإنسانية. حاولت الدراسة أن تكون ممثلة للغة العربية في الاستعمال الطبيعي العام؛ فقد تنوعت بين الكتب والمقالات. بين التأليف والترجمة.

| | كلمات | مداخل |
|---|---|---|
| حجم المدونة | 80633 | 41451 |
| عدد المجالات | 8 | |
| التغطية الزمنية | من 1950 وما بعدها | |

ويمكن أن نقدم مخططا عاما لمدونة الدراسة في الجدول التالي:

[22] تتخذ معظم الجامعات مدونات لغوية محوسبة ضخمة، وتضعها في خدمة الباحثين من مختلف التخصصات، وهذا الأمر غير موجود في جامعة القاهرة، مع تاريخها ومكانتها فيما يخص الدراسات العربية. فهل من مدونة خاصة بجامعة القاهرة؟.

| المجال | المجال الفرعي | النص | صاحبه |
|---|---|---|---|
| الأدب | رواية وقصة قصيرة | أصداء السيرة الذاتية | نجيب محفوظ (مصر) |
| | | جلنار | زكرياء أبو ماريا (المغرب) |
| | | العمامة والقبعة | صنع الله إبراهيم (مصر) |
| | | ذاكرة الجسد | أحلام مستغانمي (الجزائر) |
| | دراسات أدبية | في الشعر الجاهلي | طه حسين (مصر) |
| | | فصول في علم الاقتصاد الأدبي | حنا عبود |
| | | تمظهرات التشكّل السير ذاتي قراءة في تجربة محمد القيسي السير ذاتيّة | محمد الباردي |
| الدين | | جَدِّدْ حياتَك | محمد الغزالي |
| | | اَلتَّصْوِير اَلْفَنِّيّ فِي اَلْقُرْآن | سيد قطب |
| | | مَعَانِي اَلْقُرْآن بَيْن اَلرِّوَايَة وَالدِّرَايَة | أَحْمَد حَسَن اَلْبَاقُورِيّ |
| اقتصاد وسياسة | | مقالات | |
| | | تاريخ الجزائر المعاصر دراسة ( الجزء الأول ) | محمد العربي الزبيري |
| طب وصيدلة | | مقالات | |
| رياضة | | مقالات | |
| فنون | سينما | مقالات | |
| | مسرح | مقالات | |
| | موسيقى | مقالات | |
| علوم إنسانية | انثروبولوجي (إناسة ) | مدخل إلى علم الإنسان (الأنثروبولوجيا ) | عيسى الشمّاس |

| | | | |
|---|---|---|---|
| | علم نفس | فن الإصـغاء | إريك فروم<br>ترجمة: محمود منقذ الهاشمي |
| علوم وتقنية | | مقالات | |
| علوم طبيعية | جغرافيا | مقالات | |
| | فلك | مقالات | |

## مدونات التعريفات المعجمية:

حاولت الدراسة الإفادة من المنجز العربي في مجال التعريفات المعجمية، فاتخذت الدراسة مدونة أخرى عبارة عن نصوص تعريفات المعجم الوسيط، وباب الجيم في كل من المعجم الوسيط والمعجم العربي الأساسي[23].

## معالجة المدونة النصية:

تشمل معالجة المدونة جانبين: حاسوبي ولغوي، فأما الحاسوبي فهو تهيئة البرامج التي تحتوى متن المدونة تخزينا وتعاملا معها. وأما الجانب اللغوي فيشمل التحليلات على المستويات الصرفية والتركيبية والدلالية والاستعمالية.

ويمكن أن نوجز الإجراءات التي تعاملت الدراسةُ بها مع مدونتها النصية فيما يلي:

تحليلات **N-gram** وهي تحليلات آلية لتبيان درجة تكرار الكلمةUni gram، أو العِبارة Tri gram & bi gram ، أو التصاحبات اللفظية Collocations في المدونة، مع عرض النتائج مرتبة حسب عدد مرات التكرار مع ذكر النسبة إلى عدد كلمات المدونة.

استخلاص مدونة فرعية sub corpus (الوحدات المعجمية التي تنتمي إلى الباب المعتمد عينة في الدراسة وهو باب الجيم).

التحليل الصرفي Morphological Analysis (وناتج هذه المرحلة وحدات معجمية مصنفة حسب المقولة الصرفية بالإضافة إلى سياقاتها ) أو Morphological clustering .

التحليل الدلالي.

التحليل التركيبي.

[23] كان التعامل الإحصائي مع متن الوسيط كاملا بغرض استكشاف لغة الوسيط في التعريف. في حين كان التركيز على باب الجيم في المعجمين (الوسيط والأساسي) باعتبارهما عينة الدراسة التي عليها التركيز في اختيار أمثلة للتطبيق عليها في مرحلة التمثيل.

التحليل المعجمي.

تحديد رؤوس المداخل Lemmatization .

الربط بين مدونة تعريفات الوسيط والأساسي ومداخل المدونة النصية الفرعية.

Matching between lexical entries in dictionaries and sub corpus

### معالجة مدونة التعريفات:

. التحليل الإحصائي للغة التعريف (كلمات وتصاحبات لفظية).

. التحليل الصرفي والدلالي.

. الربط بالمدونة النصية.

. الاستعانة بها في الوسم الدلالي والتركيبي في المدونة النصية.

. التوثيق.

## منهجية الدراسة

### طبيعة المعالجة:

. تنتمي الدراسة– منهجيا– إلى المقاربة القواعدية Rules Approach ، وهي مقاربة معرفية الأساس Knowledge based ، وإن استعانت ببعض جوانب الإحصاء الحاسوبي باعتبارها إجراءات تخدم مرحلة التحليل[24].

. كما يمكن تصنيف الدراسة– من حيث اعتمادها مدونة لغوية أو عدم اعتمادها[25]– باعتبارها دراسة معتمدة على المدونة النصية Corpus-based approach.

. أما في التحليل الدلالي فقد أفادت الدراسة من التحليل السياقي الأفقيSyntagmatic Analysis ويعتمد على التحليل السياقي للتوصل إلى المحتوى الدلالي، والتحليل الرأسي الاستبدالي Paradigmatic Analysis في تحليل العلاقات الدلالية والشبكة الصيغية.

. أما انتماء الدراسة فهو إلى الدلالة المعجمية الحاسوبية، حيث تقترب في جانب التحليل من علم الدلالة المعجمية، كما تنتمي في جانب التمثيل إلى المعجمية الحاسوبية.

[24] توجد مقاربتان في حقل اللسانيات الحاسوبية: مقاربة قواعدية تهدف إلى وضع قواعد لسانية للحاسوب، ومقاربة إحصائية تترك للحاسوب استكشاف الاطرادات الموجودة في المصدر اللغوي أو المدونة المحوسبة، وصنع قواعده.

[25] تقسم المقاربات اللسانية الحاسوبية من هذه الناحية مقاربتين: الأولى معتمدة على المدونة (Corpus-based approach)، والثانية مقاربة لسانية تعتمد المنجز اللساني بالأساس في استخلاص المعارف اللغوية.

### مستويات التحليل:

قدمت الدراسة تحليلا في أربعة مستويات هي:

- المستوى الصرفي.
- المستوى التركيبي.
- المستوى الدلالي.
- المستوى المعجمي.

### الإجراءات:

إجراءات التحليل هي المعالجة اليدوية أو الحاسوبية للمدونة استكشافا للظواهر، وتحليلا لها. وقد كان للدراسة توجهان في التحليل: الأول تحليل يشمل المدونة كلها، ويكون تحليلا حاسوبيا غايته مرحلة التحليل، والثاني تحليل يقتصر على باب الجيم (المدونة الفرعية لكلمات الجيم)؛ لأغراض التحليل المعمق بغرض التمثيل ، ويكون تحليلا يدويا في الغالب.

### إجراءات التحليل الدلالي:

. **الوسم الدلالي** Sense Tagging: ويعني تحديد قسم الكلام التي تنتمي إليه الوحدة المعجمية.

. **التحشية الدلالية** Sense Annotation : أي تسجيل المعاني المستنتجة من خلال السياق، وتدوين الحواشي الدلالية مثل العلاقات الدلالية، والحقول الدلالية، والسمات الدلالية.

. **الترابطات الدلالية** Sense Clustering ويقصد به تجميع المعاني المترابطة للوحدة المعجمية.

### إجراءات التحليل الصرفي:

. **التحليل إلى جذوع** Stemming : وهو إجراء غايته التوصل إلى الجذوع Stems أو الوحدات المعجمية الأبسط في المدونة مفهرسة، بحيث تصبح المخرجات قائمة من المداخل.

. **تحليل صيغ الوحدات المعجمية الاشتقاقية** Patterns Analysis: ويعنى به محاولة الوسم الصرفي الدلالي للصيغ الصرفية Morphological Patterns ، وتحديد الدلالة الصرفية لها، وتحديد المقولة الصرفية للمشتقات.

### إجراءات التحليل التركيبي:

. الوسم المقولي التركيبي POS tagging : وهو الوسم التركيبي للوحدات المعجمية من حيث مقولتها التركيبية، وانتماؤها إلى أحد أقسام الكلام ( الفعل، المشتق، الاسم، الأداة).

. تعيين فئات الأفعال Verb classes.

. تصنيف الفعل حسب نمط بنية الحدث Event Structure.

. الأنماط التركيبية للاسم المركب.

**إجراءات التحليل المعجمي:**

. تعيين رؤوس المداخل Lemmatization: ويعني تجميع التصريفات المترابطة، وتحديد رأس لها head word يتخذ عنوانا لها في المعجم، أو في قاعدة البيانات.

. فصل المشتركات اللفظية Homonymy distinctions.

. التصنيف حسب المقولة المعجمية (كلمة، تعبير اصطلاحي، فعل عباري ...).

**الأدوات الحاسوبية لتحليل المدونة:**

الأدوات الحاسوبية التي استخدمتها الدراسة في التحليل أدوات جاهزة أتيحت للباحث وهي:

**المفهرس الآلي** Concordancer [26]:

يُعَد المفهرس الآلي واحدا من الأدوات الأساسية الأكثر أهمية في التعامل مع المدونات الحاسوبية؛ إذ تعتمد عليه كثير من المشروعات المعجمية في تحليل مدونتها. فهو يقوم بتصنيف للكلمات، وفهرستها، وتحديد مرات التكرار، ونسبته مقارنا بعدد كلمات المدونة كلها، وتحديد التصاحبات اللفظية، وعرض المخرجات، أي أنه يقوم بتجهيز المدونة للمراحل التالية من التحليل. والمفهرس الذي اعتمدت عليه الدراسة هو: "ConcApp".

**خوارزم التصاحبات ( N-Gram ):**

وهو عبارة عن خوارزم Algorithm لتحديد التصاحبات اللفظية Collocations جُعلت له واجهة للمستخدم User Interface[27]، وهو يستكشف الكلمات التي تقع متصاحبة، ويرتبها تصاعديا أو تنازليا، مع ذكر نسبة ورود كل تكرار. وقد استخدم في تحديد التصاحبات في المدونة النصية، والتصاحبات في مدونة التعريفات.

وإلى جانب اعتماد الدراسة في التحليل على أدوات حاسوبية، اعتمدت أيضا التحليل اليدوي في كثير من جوانب التحليل؛ وذلك إما لتحقيق قدر أكبر من الدقة، أو لعدم توفر الأداة الحاسوبية التي تصلح لذلك.

[26] رابط المفهرس المعتمد في التحليل: http://www.edict.com.hk/PUB/concapp/

واسم الشركة المنتجة له: Educational Services Consultants

والنسخة المعتمدة: Version=4.00.000

[27] الأداة من إعداد المبرمج المصري: حاتم مصطفى .وهي موجودة على الشبكة الدولية للمعلومات من خلال الرابط التالي:

http://www.n-gram\CodeProjectN-gramandFastPatternExtractionAlgorithm

وفي الشكل التوضيحي التالي تستعرض الدراسة جوانب منهجية العمل على مستوى التحليل والتمثيل[28]:

| الموضوع | التحليل | التمثيل |
|---|---|---|
| **الغاية** | التوصل إلى العنصر الذري في:<br>الوحدات المعجمية.<br>المعاني. | صياغة المعنى الذري (ناتج التحليل) في صورة تعريف معجمي. |
| **المدخلات** | مدونة نصية محوسبة.<br>مدونة للتعريفات المعجمية العربية<br>مصادر التوثيق | وحدات معجمية مصنفة.<br>معانٍ. |
| **المخرجات** | وحدات معجمية مصنفة.<br>معانٍ.<br>سمات دلالية. | تصنيفات.<br>أبنية تعريفية. |
| **الإجراءات** | – التحليل اليدوي.<br>– التحليل الحاسوبي. | تنميط المعرفات.<br>تصميم بنية التعريف ومكوناتها.<br>صياغة معجم مفردات التعريف.<br>قواعد التعريفات. |
| | – التحليل السياقي.<br>– تحليل التصاحبات. | |
| **المستويات** | المستوى الصرفي.<br>المستوى التركيبي.<br>المستوى الدلالي.<br>المستوى المعجمي.<br>مدونة التعريفات المنجزة. | أنماط المعرفات.<br>معجم مفردات التعريف.<br>أنماط التعريفات<br>بنية التعريف Definition Syntax<br>البنية المعجمية الدلالية. |
| **المحتوى الصرفي** | مشتق / غير مشتق.<br>الأصول والفروع.<br>المقولة الصرفية.<br>مشتق مطرد/ مشتق متحول معجميا | البنية التعريفية الصرفية:<br>تعريف الأصول، والفروع.<br>تعريف المشتق المطرد.<br>تعريف المتحول معجميا.<br>تمثيل محتوى الصيغة الصرفية. |

[28] هذه الإجراءات لها جانبان، فالأول بكونه إجراءات الدراسة للتحليل، والثاني بكونه جزءا من التصور المقترح، وهو الجزء المتعلق بإجراءات التحليل.

|  |  |  |
|---|---|---|
| **المحتوى التركيبي** | المقولة التركيبية.<br>البنية المحمولية. | البنية التعريفية التركيبية:<br>التصنيف. |
| **المحتوى الدلالي** | الحقل الدلالي.<br>البنية العلاقية الدلالية.<br>السمات الدلالية.<br>نمط بنية الحدث. | البنية التعريفية الدلالية:<br>صياغة الحقول الدلالية.<br>التصنيف.<br>السمات الدلالية.<br>جملة التعريف. |
| **المحتوى المعجمي** | المقولة المعجمية.<br>الوحدة المعجمية. | أنماط المعرفات. |
| **المنطلقات النظرية** | الجذر وإشكالية أصل الاشتقاق.<br>المعنى الصرفي: طبيعته.<br>الصيغة الصرفية وحدة دلالية. | طبيعة المحتوى المعجمي الدلالي.<br>أبنية التعريف.<br>معجم وحدات التعريف (الديفيمات).<br>اشتقاق تعريفات المشتقات من تعريفات الأصول.<br>البنية المعجمية الدلالية. |
|  | المقولة التركيبية.<br>البنية الحملية.<br>القيود الانتقائية. | أقسام الكلم.<br>الأنماط التركيبية للفعل.<br>أنماط القيود الانتقائية. |
|  | السمات الدلالية.<br>العلاقات الدلالية الأفقية، والرأسية.<br>القيود الانتقائية. | صياغة قائمة السمات الدلالية ونحوها.<br>صياغة العلاقات. |
|  | الوحدة المعجمية العامة، والموسوعية، والمصطلحية.<br>الوحدة المعجمية البسيطة، ومتعددة الكلمات | بنية الوحدة العامة.<br>بنية الوحدة الموسوعية.<br>بنية الوحدة المصطلحية. |
| **المشترك اللفظي المشترك** | تحديد المعايير المميزة، والتي تفصله عن متعدد المعنى.<br>الفصل العملي. | اتخاذ كل وحدة رأسا لمدخل معجمي. |

| | | |
|---|---|---|
| **اللفظي** | تحديد قائمة المداخل النهائية للمعجم. | |
| **متعدد المعنى** | فصله عن المشترك اللفظي.<br>استخلاص المعاني المستقلة.<br>التعامل مع المعنى الواحد باعتباره وحدة المعجم الدلالية. | اتخاذ موقف من تنظيم المعاني لمتعدد الدلالة.<br>تبين العلاقات الدلالية داخل المدخل الواحد – لكل معنى تعريف. |
| **العلاقات الدلالية** | نمطان من العلاقات الدلالية: رأسية أفقية | صياغتها في جملة التعريف، وفي البنية الدلالية للتعريف. |
| **تنميط المعرفات** | تحديد أنماط المعرفات. | تحديد متطلبات كل نمط. |
| **البنية المعجمية** | تحديد متطلباتها | صياغتها على المستويين اللغوي والحاسوبي |

**ثالثا: التمثيل : لغته، وأبنيته.**

**لغة التعريف:** المقصود بلغة التعريف اللغة المستخدمة في تمثيل الوحدة المعجمية. واللغة هنا يقصد بها الوحدات التعريفية والعبارات والجمل وقواعد التركيب.

- **معجم وحدات التعريف:** ويعني قائمة الكلمات المستخدمة في تحرير التعريف، وهي مختارة على أسس من نتائج التحليل الصرفي والتركيبي والدلالي والمعجمي.
- **نحو التعريف:** هو قواعد تآلف الوحدات التعريفية في جملة التعريف، بحيث يمكن تحليل جملة التعريف (آليا) إلى مكونات يمكن الإفادة منها دلاليا في التعرف البنية الدلالية.
- **محددات التعريف:** هي كلمات أو عبارات أو تصنيفات تأتي مع تعريف وحدة معجمية؛ كي تميزه عن تعريف آخر لوحدة معجمية ذات دلالة متقاربة أو متداخلة معها.
- **بنية التعريف:** هي بنية نموذجية للتعريف، تعد بمثابة قالب يقوم المحررون بالبناء عليه، واستكمال محتواه لكل وحدة معجمية.

## الوحدة التعريفية (الديفيم DEFIME )

**المفهوم، الفروض، الطبيعة، الأصول، الإجراءات، الكفاية، قواعد التآلف:**

في الوقت الذي طغى فيه التفكيك على مقاربة الفعل والمشتقات ومحاولة تمثيله دلاليا، أو دلاليا تركيبيا، لم توجد في المقابل جهودٌ تحاول تمثيل هذه المقاربات أو نتائجها على المستوى المعجمي الخالص في صورة تعريفات معجمية، تفيد من إجراءات التفكيك والتذرية ونتائجها.

1. **المفهوم**[29]: هو الوحدة الدنيا التي تمثل الذرة البنائية التي يتكون منها التعريف، وهي وحدة تنتمي إلى لغة التعريف (التي هي لغة شارحة، تصف اللغة meta language) يمكن بناء التعريف بها، ويمكن تفكيكه إليها).

فالديفيم: نظرية في لغة التعريف، غير أن هذه اللغة– بوصفها تمثيلا للمعنى أو المحتوى– تقتضي الدراسة المفصلة للمحتوى؛ للتوصل إلى أولياته التي تتطلب اصطناع أوليات (ديفيمات) تعريفية توازيها.

وتقوم الفكرة على افتراض مفاده: إن المعنى بنية مركبة أو معقدة ينبغي تحليلها أو تفكيكها لإعادة توصيفها وتمثيلها ديفيميا، إذ "هناك تصور دقيق مؤداه أن كل لفظة معجمية عبارة عن حزمة من الملامح .[30]

2. **الفروض والمنطلقات** (التي تقوم عليها فكرة الديفيمات).

- يمكن افتراض مجموعة منتهية من الوحدات الدلالية يبنى منها التعريف المعجمي.
- تقوم الفكرة على إمكانية تذرية لغة التعريف في مقابل " تذرية المعنى"[31]. ويمكن صياغة ذا كما يلي:-

. المعنى مركب من عدة أولويات.

. يمكن اصطناع لغة واصفة (Meta language) للتعريف المعجمي، تعبر – بالتوازي– عن أوليات تمثل تلك الأوليات.

. ينتج تمثيل المعنى في صورة تعريف من تراكب هذه الأولويات.

. لتراكب هذه الأوليات قواعد rules of syntax تحكم تآلفها في بنية التعريف.

3. **الطبيعة: خصائص:**

[29] التعريف: هو تتميم الوصف الدلالي والتركيبي والصرفي للكلمة المعرفة، أو التعبير المعرف، لمتطلبات المستخدم البشري ومتطلبات برامج الحاسوب.

[30] جوديث جرين: علم اللغة النفسي، تشومسكي وعلم النفس. ترجمة وتعليق مصطفى التوني ص84 .

[31] وردت فكرة تذرية المعنى في ورقة بولينجر Dwight Bolinger: The Atomization of Meaning

- تنتمي الوحدة التعريفية إلى اللغة الواصفة (Meta Language) المعجمية ولكنها وثيقة الصلة– أيضا– بتحليل اللغة / المعجم ؛ استخلاصا وتمثيلا.
- وحدة بناء التعريف المعجمي ليست دائما كلمة بل قد تكون عبارة أو تركيبا أو جملة.
- للوحدة التعريفية علاقة وثيقة ببنية المعجم التصورية Conceptual Structure [32]؛ إذ هي تمثلها باحتوائها الوحدات المعجمية التي تكاد تمثل نموذجا لمعجم مصغر للغة.
- لها بنية (شكل) ودلالة (محتوى).
- يمكن التعامل معها تحليلا وتركيبا.

### 4. **الأصول** (الأصول الإبيستيمولوجية ):

ظهر مصطلح الفونيم Phoneme في أدبيات الدرس الصوتي، (علم وظائف الأصوات Phonology)[33]،وكان لظهور هذا المفهوم أثر في إحداث ثورة في مجالات البحث اللغوي كلها، باعتبار الفونيم وحدةً للتحليل يمكنها أن تقدم صياغة بنيوية أكثر كفاءة للبنية الصوتية للغة في علاقتها بالجانب الدلالي.

ثم ظهر في مستويات التحليل اللغوي مصطلحات تعبر عن مثل هذا المفهوم فيما يخص طبيعة مادتها. فظهر في مجال علم الصرف مفهوم الوحدة الصرفية (المورفيم morpheme). وتتابع ظهور المفاهيم والمصطلحات المماثلة في بقية فروع التحليل اللغوي.

ويمكن تحديد المصطلحات في مستويات اللغة من خلال الجدول التالي:

| المصطلح | المفهوم | المجال |
|---|---|---|
| phoneme | الوحدة الصوتية الدنيا التي يؤثر تغيرها في تغير معنى الكلمة التي تدخل في تكوينها | علم وظائف الأصوات<br>Phonology |
| morpheme<br>Lexeme | الوحدة الصرفية الدنيا التي تحمل معنى، ويمكنها الدخول في تكوين كلمة أو الدخول على كلمة<br>الوحدة المعجمية الدنيا | علم الصرف<br>Morphology<br>المعجمية<br>Lexicology |
| | | |

[32] مُقترِح المفهوم هو جاكندوف (Ray Jackendoff) ويذهب إلى أن البنية اللغوية جزء من بنية كبرى هي البنية المفهومية التي تجتمع فيها البنى الإدراكية للعالم.

[33] ويُقْصَد به الوحدة الدنيا التي تؤثر في تحديد معنى كلمة وتمييزه عن كلمة أخرى.

| | | |
|---|---|---|
| sememe | الوحدة الدلالية الدنيا | علم الدلالة<br>Semantics |
| synteme | الوحدة التركيبية الدنيا | التركيب<br>Syntax |

وهذه المفاهيم– على تعددها– يجمعها عدد من الخصائص:

- أنها تعبر عن "وحدات" للتحليل، لها دورها الحاسم في عملية التحليل اللغوي في مستوياته المختلفة، وكان الملجئ إليها إنما هو الحاجة إلى "وحدة " تحليلية، يمكن التعامل معها باعتبارها مقولة تصلح للتحليل والتركيب.
- ارتباطها الشديد بالدلالة.
- أنها وحدات تنتمي إلى اللغة الطبيعية[34].
- أنها توظف في التحليل والتركيب.

على أنه يمكن القول إن هذه الوحدات لها تَحَقُّقُها الشكلي فيما عدا السيميم (الوحدة الدلالية)، وإن الديفيم المفترض ليس التحقق الفعلي للسيميم بل هو وحدة معجمية.

### 5. الجدوى:

**في التأليف:**

- هي الوحدات الذرية التي ينبني منها التعريف في جملة دالة مؤدية لمحتوى الوحدات المعجمية .
- الالتزام بقائمة محددة من الديفيمات وبقواعد تآلفها يؤدي إلى الالتزام بأنماط تعريفية ثابتة (هي بدورها محددة قبلا).
- يؤدي الالتزام بأنماط تعريفية محددة إلى تحقيق الاتساق والبنية السليمة لمادة المعجم.
- يمكّن من اشتقاق تعريفات المشتقات من تعريفات أصولها.

**في التحليل:**

- يمكن استخدامها في تطبيقات معالجة اللغات الطبيعية من خلال محلل التعريفات، حيث يكون التعامل مع كيانات Entities يمكن معالجتها آليا عن طريق محلل جُمَلي آلي Parser؛ استخلاصا وتصنيفا.

وتنتمي الوحدات التعريفية Defimes إلى إحدى المجموعات التالية:

**المجموعة** **تحققاتها : أنماطها**

[34] بخلاف تلك المفاهيم المشابهة التي لا تنتمي إلى اللغة الطبيعية، كوحدات اللغات الاصطناعية مثل لغات البرمجة، وكالديفيم الفرضية موضوع هذه الفقرة لانتمائها إلى اللغة الشارحة Meta Language.

| | |
|---|---|
| المجموعة الصرفية | سمات صرفية دلالية |
| | مقولات صرفية دلالية |
| | علاقات اشتقاقية دلالية |
| المجموعة التركيبية | قيود انتقائية( فاعلية/ طبيعتها...الخ) |
| | بنية تركيبة للمحمولات |
| | بنية الحدث والبنية |
| المجموعة الدلالية | سمات دلالية / ذرات دلالية |
| | علاقات دلالية ( أفقية – رأسية) |
| | مجال دلالي |
| | هل يمكن توظيف بنية الكواليا هنا |

المجموعة الدالة على محددات التعريف

الروابط hinges التي تربط المعرف بالتعريف

يمكن تصور الديفيم باعتباره مفردات لغة شارحة، تعبر عن مدلولات وراء اللغة meta- language ، تستعمل في وصف البنية المعجمية الدلالية لمعجم. يضبط الديفيم اللغةَ ويضبط الأنماطَ بحيث تبقى هناك فراغات (غير ديفيمية، أو غير محددة، يمكن للمعجمي أن يملأها من معجم آخر). وعلى الاحتمال الثاني فإن ذلك يفترض وجود قائمتين معرفتين قبلا، هما: قائمة الديفيمات، وقائمة كلمات التعريف. ويكون التحليل عندئذ متوجها إليهما على الترتيب. وعلى ذلك يكون النمط التعريفي مرتبطا بمجموعة من الديفيمات. وعندئذ سيبرز التساؤل: وما جدوى التصور عندئذ ما لم يكن مستوعبا ؟

وهذه الفرضية يمكن أن تكون الرابط الوثيق بين محتوى التعريف ولغة التعريف، إذ هى الوحدة اللغوية التعريفية الحاملة لمحتوى تعريفي. على أن هذا المحتوى قد يكون تركيبا أو دلاليا أو صرفيا.

6. **الإجراءات:** (إجراءات التوصل إلى قائمة الوحدات التعريفية الديفيمات واستخلاصها).

التوصل إلى الوحدات التعريفية أهم نتائج التحليل للمدونتين المعتمدتين (مدونة النصوص، ومدونة التعريفات المعجمية)

- تحليل عينات تعريفية لاستخلاص الديفيمات المستخدمة من خلال استراتيجيات التعريف، وكان ذلك باتخاذ مدونة من تعريفات أشيع معجمين عربيين حديثين هما المعجم الوسيط والمعجم العربي الأساسي.
- استخلاص الديفيمات من المقاربات الدلالية السابقة، وقد أفادت الدراسة من تصورات كاتز ، وبوسطيوفسكي، وميلتشوك في تمثيل المحتوى المعجمى الدلالي.

- التحليل الدلالي التركيبي للعينة، لاستخلاص الوحدات التي تؤدى بها المعلومات التركيبية الدلالية.
- التحليل الدلالي، لاستخلاص السمات الدلالية التي تحكم بنية المعجم والوحدات التي يمكن تمثيلها بها.
- التحليل الصرفي الدلالي: (المعجم أو الشبكة الصيغية)، وغايته استخلاص السمات الصرفية الدلالية للصيغ المزيدة، والمقولات الصرفية الاشتقاقية، إضافة إلى الوحدات التي ستستخدم في اشتقاق تعريفات المشتقات من تعريفات أصولها.
- صياغة القائمة النهائية للديفيمات المستخلصة.
- صياغة قواعد التراكب والتآلف (وحدات واجبة، اقتراحات لوحدات ممكنة، احتمالات ممتنعة).
- اقتراح أنماط رئيسة وفرعية للتعريف.
- اقتراح آليات توليد التعريفات الفرعية من أصولها.

### 7. قواعد التآلف: Defime Syntax

- للديفيمات قواعد تحكم تراكبها عند بناء التعريف، أو جزء التعريف، لتكوين نَوْعَي التعريف الرئيسي والفرعي[35].
- توجد علاقات داخلية بين قائمة الديفيمات، لكونها تعد بمثابة معجم مصغرSub Lexicon ، به من الكيانات (الدلالية والتركيبة) والعلاقات (الاشتقاقية والدلالية) ما بالمعجم العام.
- من هذه العلاقات– ولعله أشدها تأثيرا في العمل– التوارث Inheritance[36]، ويعني الترابط الكلي لبنية المعجم، من خلال العلاقات التي تربط كياناته: مداخل ودلالات.
- قواعد التراكب والتآلف تبين(القيود الواجبة، الاقتراحات الممكنة، الاحتمالات الممتنعة).
- يمكن افتراض برنامج يكون بمثابة مدقق تركيبي Definition Checker للتعريفات أثناء عملية التحرير، ويكون جزءا من برنامج محرر التعريفات المقترح.

**تصور عام للغة التعريف**

| قائمة الوحدات |  | أنماط التعريفات ( أبنية التعريف ) | آليات توليد |
|---|---|---|---|

[35] التعريف الرئيسي يكون للأصول، والفرعي يكون للمشتقات.

[36] التوارث هنا مستخدم بمفهوم جيمس بوسطيوفسكي J. Pustejovsky بكونه البنية الكلية للمعجم والعلاقات التي تحقق هذا. J. Pustejovsky: Generative Lexicon.p:24

| | | | | |
|---|---|---|---|---|
| التعريفية ( الديفيمات) | قواعد التآلف<br>نحو التعريف | نمط رئيسي | نمط فرعي | تعريفات المشتقات<br>(المطردة) |
| **البنيـة المعجميـة الدلاليـة** | | | | |

### 8. الكفاية:

مثل أي تصور ينبغي قياس درجة كفايته مقارنا بالتصورات المقترحة المنافسة (السابقة). والكفاية تعني مدى ما يحققه التصور النظري من دقة في تمثيل الظواهر اللغوية: ملاحظة ووصفا وتفسيرا. وقد تعددت معايير اختبار الفرضيات اللغوية لتغطي كل مجالات الظواهر المدروسة. فقد اقترح تشومسكي[37] مستويات ثلاثة للكفاية في مجال النظرية النحوية، وأعاد جاكندوف توظيفها في مجال المعجمية ونظرية المعجم[38]، وهي:

- الكفاية الملاحظية Observational Adequacy.
- الكفاية الوصفية ( Descriptive adequacy).
- الكفاية التفسيرية ( explanatory adequacy).

وقد تعرضت هذه الأفكار للمراجعة والتطوير. ومما ينبغي ذكره في هذا السياق إضافة بوسطيوفسكي[39] مستوى آخر، هو:

- الكفاية الإمبيريقية (التجريبية) (Empirical adequacy).

ويمكن أن تقترح **معايير لقياس الكفاية اللسانية الحاسوبية لمنهجية تمثيل التعريفات**، وهي:

. التحليلية: قابلية التعريفات للتحليل الآلي: (Parsing).

. التركيبية: قابليتها التركيب.

. بساطة الوحدات التعريفية، ومخرجات التحليل، في مقابل التعقيد.

. الجدوى والإفادة في تطبيقات معالجة اللغات الطبيعية.

. واقعية التصور وقابليته للتطبيق.

. الاستيعاب للظاهرة اللغوية.

## التصور والتطبيقات الحاسوبية

[37] Chomsky, N. (1965) *Aspects of the Theory of Syntax*.

[38] Ray Jackendoff: Morphological and Semantic Regularities in the Lexicon p: 639.

[39] J. Pustejovsky: Generative Lexicon p: 28.

غاية التصور معالجة المحتوى المعجمي الدلالي بغرض تطبيق المخرجات في أنظمة معالجة اللغات الطبيعية وتطبيقاتها. فهذه الأنظمة والتطبيقات لها متطلبات في المحتوى، وأخرى في اللغة: فمشكلة معالجة اللغات الطبيعية مع المصادر المعجمية تنحصر في أمرين: المعارف اللغوية التي تحتاج إليها هذه المعالجة في تطبيقاتها، ولغة التمثيل المنضبطة التي يمكن التعامل معها حاسوبيا.

ولقد غدت القابلية للتوظيف في التطبيقات الحاسوبية معيارا من معايير الحكم على كفاية تصور نظري يهدف إلى تمثيل المحتوى المعجمي الدلالي.

**التعريفات مصدرا للمعارف المعجمية الدلالية:**

يحدد بوجريف (Bran Boguraev) وتيد بريسكو (Ted Briscoe) المعارف اللازمة لأنظمة معالجة اللغات الطبيعية Natural Language Processing بأنها: المعارف النطقية (الفونولوجية) والصرفية والتركيبية والدلالية والذرائعية[40]. وموضوع الدراسة يستبعد المعلومات النطقية لضآلة ارتباطه بالمحتوى المعجمي الدلالي، في حين يركز على الجوانب التركيبية والصرفية والدلالية. وبذا يكون التعريف المعجمي مصدرا لهذه الأصناف من المعلومات اللغوية بالنسبة لأنظمة معالجة اللغات الطبيعية وتطبيقاتها.

**جوانب فك اللبس في التصور المقترح:**

من أهداف التصور المقترح تقديم المعارف اللغوية التي تلزم للمساعدة في فك اللبس[41].

والمعلومات التي اهتم بها المقترح لتحقيق ذلك هي:

**على المستوى الصرفي:**

. فك اللبس المقولي (مثلا: اسم الفاعل، الصفة المشبهة، صيغة المبالغة، اسم المهنة[42])

. فك اللبس الصيغي الممكن بين الكلمات ذات الصيغ المتشابهة.

. فصل المعاني عن طريق فصل الأسر الاشتقاقية على أسس من أصل الاشتقاق.

. المشتق المطرد والمتحول معجميا.

**على المستوى التركيبي:**

[40] Bran Boguraev & Ted Briscoe: Computational lexicography for natural language processing. P: 4:5

[41] تجدر الإشارة هنا أن المقترح لا يدعي فك اللبس تماما، فاللبس جزء أساسي من سمات اللغة الطبيعية ولكن اقتراح إجراءات لغوية أو لسانية، وتقديم معلومات لسانية، كل ذلك يساعد في فك اللبس.

[42] تقصد الدراسة باسم المهنة: قسما من أسماء الفاعلين، وصيغ المبالغة تدل على حرفة أو مهنة أو وظيفة (مثل النجار، السقاء، الممرضة، الزجاج ... الخ).

. بيان أنماط المقولة التركيبية.

. بيان أنماط البنية الحملية.

. توضيح قيود الانتقاء Selection restriction .

. توضيح أبنية الحدث Event structures.

**على المستوى الدلالي:**

. تعد محددات التعريف نوعا من توصيف الوحدة المعجمية، وتفريق الدلالات وفصل اشتباكاتها.

. التصاحبات مُكَمِّلا للتعريف: يمكن توظيفها في مجال تعلُّم الآلة Machine Learning.

. العلاقات الدلالية.

. السمات الدلالية والفروق اللغوية.

. اقتراح معايير لتمييز أنماط الوحدات المعجمية من حيث طبيعة محتواها الدلالي ( العامة، الموسوعية، المصطلحية) واتخاذ ذا معيارا من معايير الاستقلال بمدخل.

**في التحليل والتركيب: الفهم والتوليد:**

وذلك من خلال مفهوم الوحدة التعريفية التي تعد وحدة للتحليل والتركيب، يسهم المقترح– في حالة تطبيقها– في تسهيل تحويل المعارف المعجمية الدلالية إلى كيانات مفهومية Ontologies، وهو توجه ذو أهمية كبيرة في تطوير الويب الدلالي.

**في المعجمية الحاسوبية:**

. اقتراح محرر لتعريفات المعجم العربي يحتوي قواعد المنهجية المقترحة، ويحكم عمل محرري المعجم.

. اقتراح مولد آلي لتعريفات المشتقات، يشتق مادته من تعريفات الأصول.

لا يمكن إنجاز مولد التعريفات بدون إنجاز هذه الفرضية على المستوى النظري، ثم تطبيقها عمليا بحيث تتحدد قائمة منتهية final Set من الديفيمات وتنوعاتها الشكلية Allo-Defime، بحيث يصبح بالإمكان عمل الدالة التي تكون أساس التوليد في المجال الحاسوبي.

## الجديد في التصور المقترح

في تصور طبيعة المحتوى المعجمي الدلالي موضوع التمثيل في صورة التعريفات.

**في لغة التعريف:**

في اقتراح معجم وحدات التعريف وتصميم جوانبه وأسس بنائه.

اقتراح نحو للتعريف يمكّن من التحليل والتركيب.

اقتراح جوانب تحقق بنية المعجم الدلالية.

**في الجانب الصرفي:**

- افتراض نمطين من المشتقات: أصول وفروع، واقتراح إعادة بناء المدخل المعجمي العربي على أساس ذلك.
- اقتراح آليات لتوليد تعريفات المشتقات من تعريفات أصولها.
- توظيف المحتوى الصرفي في بناء المعجم،

**على المستوى المعجمي:**

تعيير أنماط المداخل المعجمية.

توسيع معايير فصل المشتركات اللفظية.

**التوليد الآلي لتعريفات المشتقات:**

يجد توليد المشتقات كثيرا من الطروحات الحاسوبية للحديث عنه: مفهومه وآلياته، وفي المقابل لا نجد حديثا عن توليد مواز لتعريفات المشتقات (المطردة )، وهو ما يعد ضرورة في اقتراح لأي مصدر معجمي موسع.

**مشروع محرر التعريفات:**

يدخل هذا المقترح ضمن اهتمام الصناعة المعجمية الحاسوبية Computational Lexicography بما هو أداة لضبط لغة التعريف وتنميطه أثناء مرحلة التحرير.

**المراجع**

- **جوديث جرين:** علم اللغة النفسي، تشومسكي وعلم النفس. ترجمة وتعليق مصطفى التوني.
المفهومية التي تجتمع فيها البنى الإدراكية للعالم.
- **وفاء كامل:** معجم التعابير الاصطلاحية العربية المعاصرة. أبو الهول للنشر – القاهرة 2007.
- **حسين نصار:** المعجم العربي نشأته وتطوره, ط 4، 1988 دار مصر للطباعة، القاهرة.
- **حلام الجيلالي:** تقنيات العريف في المعاجم العربية المعاصرة منشورات اتحاد الكتاب العرب، دمشق ، سوريا 1999.
- **مجمع اللغة العربية بالقاهرة:**المعجم الوسيط ط2, القاهرة, سنة 1972. (إعداد لجنة من المجمع).

- Chomsky, N. (1965) Aspects of the Theory of Syntax
- Kats & Fodor 1963 the structure of semantic theory
- Dwight Bolinger: The Atomization of Meaning

**http://www.jstor.org/pss/411524**

- Nida , Eugene, *A componential Analysis of Meaning. An introduction to semantic structures.*
- J. Pustejovsky: *Generative Lexicon*
- Ray Jackendoff: *Morphological and Semantic Regularities in the Lexicon.*
- Bran Boguraev & Ted Briscoe: *Computational lexicography for natural language processing*
- Susan Hunston: *Corpora in Applied Linguistics*
- Ayto. J. R. (1983). *Semantic Analysis & Dictionary Definitions. In* R.R.K Hartmann (Ed.), *Lexicography: Principles & Practice.* Exeter
- Bogreav, B. & Levin, B. (1993). *Models for Lexical knowledge Bases*. In J. Pustejovsky (Ed.), *Semantics and the Lexicon*. Kluwer Academic Publishers
- Litkowski, K. C. (2005). *Computational Lexicons and Dictionaries.*
- *Encyclopedia of Language and Linguistics* (2$^{nd}$ ed). Oxford: Elsevier Publishers
- Svensén, B. (1993). *Practical lexicography :Principles and Methods of Dictionary-Making.* Translated from the Swedish. By Sykes, J. and Schofield, K. Oxford University Press

# English to Arabic Hybrid Machine Translation System for Languages with Scarce Resources

Eng. Ahmed Hatem, Prof. Dr. Amin Nassar

Elect. & Comm. Eng. Dept. Faculty of Engineering. Cairo University, Giza, Egypt
+2010 1634 196, +2012 3418 225
amhatem@gmail.com, aminassar@maktoob.com

**Abstract.** In this paper we are introducing a hybrid English to Arabic MT system. The system combines between rule based and example based machine translation techniques. The system main aim is to provide low cost and time implementation of machine translation systems. The system uses an English to Arabic dictionary, stemmer, retrieval engine and Arabic corpus without parallel corpus. A modified Dijkstra's shortest path algorithm [1] is introduced to detect the target language phrases. We list the indexes of the source sentence's words which were found in the target language corpus and create a directed graph to detect the phrases that form a shortest path walk in the graph. The system was examined and was found that results were promising to be used for domain specific and scarce resources translation.

**Keywords:** English to Arabic Hybrid Machine Translation. Directed Graph Decoder.

## I Introduction

Machine translation (MT) had passed through multiple stations since its introduction in 1950. MT used to depend on language based tools like dictionaries. A word to word translation was first introduced then moved to more language independent rule based machine translation (RBMT) methods. RBMT dominate till the mid of the 1990th [2]. Nagao introduced the first translation with analogy that became lately the Example Based Machine Translation (EBMT) method. EBMT depends on building two huge parallel language corpus of sentence that each sentence represents a translation of its analog sentence [3]. Brown et al., introduced the statistical machine translation (SMT) which is also as the EBMT depends on two parallel corpus of analog sentences. Calculate the n-gram relations for the source language (SL) model and the translation model with the fundamental SMT equation [4].

$$\operatorname*{argmax}_{S} \Pr(S) \Pr(T|S) \qquad \textbf{(1)}$$

Dologlou et al., introduced a hybrid mono corpus MT system "METIS-I". This system used a tagged and lemmatized target language (TL) corpus without the Source Language (SL) corpus. The SL corpus was replaced by a bilingual lexicon [5]. Vandeghinste et. al., with the European consortium upgraded the project to be "METIS-II". This hybrid solution aims to provide MT systems for the languages with little resources. The consortium built separate systems that use the Dutch, Greek, German and Spanish languages as the SL and English as the TL corpus [6].

We built a hybrid English to Arabic MT system similar to the "METIS-II" system. Our aim was to provide MT system for languages with little resources. The system takes less time and cost to implement and produces accepted translation compared to the effort done in building it.

## II System Data Flow

Our implemented system use hybrid technique combines between RBMT and EBMT to provide English to Arabic translation. The system consists of a tokenizer, English to Arabic (E/A) dictionary, Arabic stemmer, retrieval engine, phrase decoder, Arabic target language corpus and its inverted indexes as shown in figure 1.

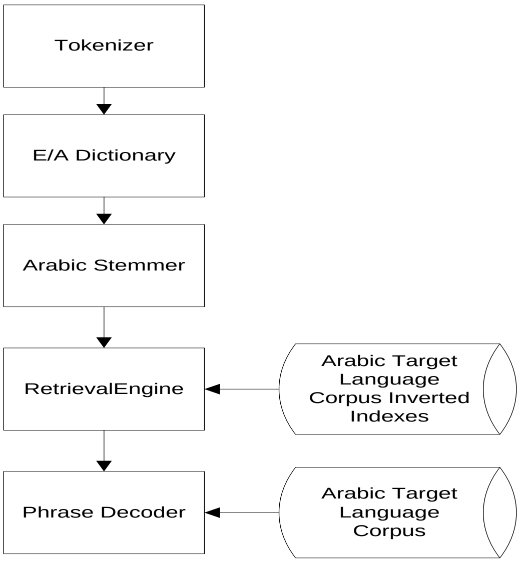

Tokenizer
E/A Dictionary
Arabic Stemmer
RetrievalEngine
Arabic Target Language Corpus Inverted Indexes
Phrase Decoder
Arabic Target Language Corpus

**Fig. 1.** System Components

The system was organized into three stages as the CLC METIS-II system described by Dirix et al., [7] "the SL Stage (SLS), the Translation Stage (TS) and the TL Stage (TLS)".

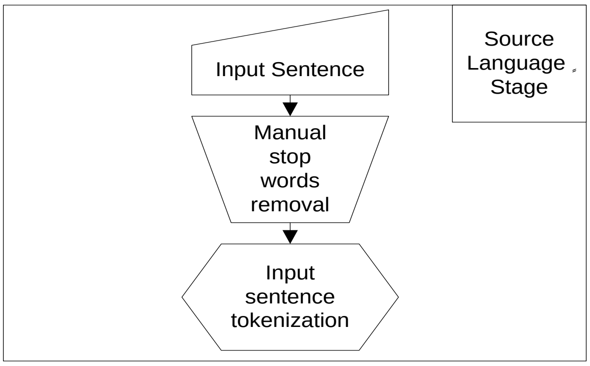


**Fig. 2.** Source language Stage

SLS, (figure 2), is concerned about handling the input English sentence. It consists of two steps. The first step is a manual stop words removal by detecting words in the English stop words list. We created a list of the top 20 English stop words that match the top 20 high frequency Arabic words. We calculated the frequency of these stop words and found that they represent 19% of the Arabic corpus size. These words have a cross meaning which makes some of the words based on its context map to more than one word from the Arabic high frequency words, like “on, in, of, at” which may map to single Arabic word “fee, في” based on its context.

A tokenizer step is used to parse the SL input sentence and generate tokens when a white space, punctuation or non alphabet character are encountered.

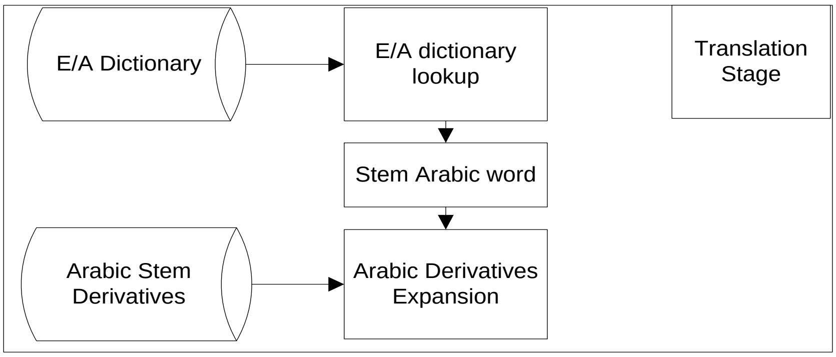


**Fig. 3.** Translation Stage

TS, (figure 3), provide word to word translation. It uses the E/A dictionary to get all the possible Arabic meanings for the English token.

All the produced Arabic meanings from the dictionary are stemmed by the stem stage. The stemmer is built with the same technique used by Chen et al., [8]. It removes the word affixes using a predefined letters set and provide the stem. The Arabic derivatives expansion stage lookup all the available Arabic derivatives for each single stem. The Arabic light stemmer is used to normalize the Arabic words returned from the E/A dictionary and to overcome the Arabic grammatical, morphological and writing challenges described by Chen et. al., [8], Larkey et. al., [9], and Attia [10]. Figure 4 illustrates a complete cycle for the input sentence data flow and how it is tokenized. It then illustrates the dictionary word translation and finally the stemmer derivatives expansion.

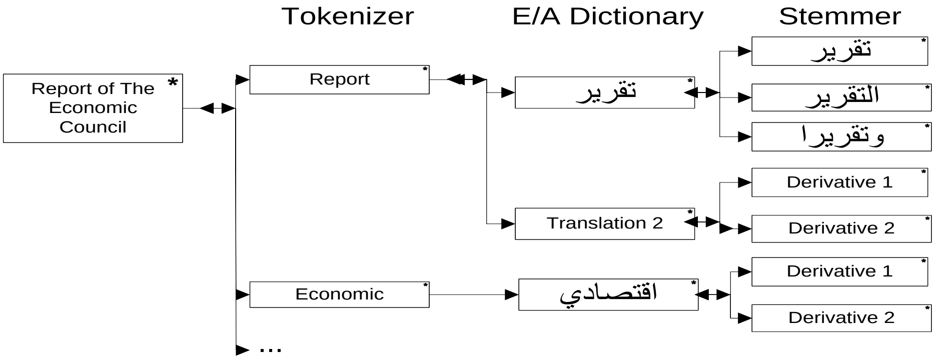


**Fig. 4.** Input sentence data flow

TLS, (figure 5), is the step that forms the TL phrase. It takes the Arabic derivatives list produced by the TS and searches the TL Arabic corpus inverted indexes for the derivatives occurrences. The Arabic TL corpus inverted indexes saves the TL corpus's word information "sentence's number and word's offsets" for each Arabic word in the TL corpus. We used the same inverted indexes structure used by Manning et al., [11]. The retrieved indexes are represented as a list of sentence's numbers and the derivative orders in each sentence. Each list produces a directed words' graph as in figure 6.

The phrase detection step is used to detect the adjacent words "phrase" from the directed graph. The phrase alignment stage tries to align phrases together to produce a longer phrase. A final list of phrases is produced with the TL sentence numbers and the start and end words' offsets in each sentence. All these phrases are passed to the phrase ranking step. The phrase ranking is based on an empirical rank equation to find the highest rank phrase.

$$S - \frac{7}{4} S' - 7 \frac{T'}{S} \qquad \textbf{(2)}$$

S is the sources span in the target phrase, S' is the missing source words in the target phrase and T' is the count of words in the target phrase that don't match any of the source words. We used multiple trials to calculate the equation (2) parameters. The main objectives while forming this equation is to increase the effect of the missing source words on the phrase rank and decrease the missing target slots effect on the rank. The highest rank will be for the phrases with more source words found.

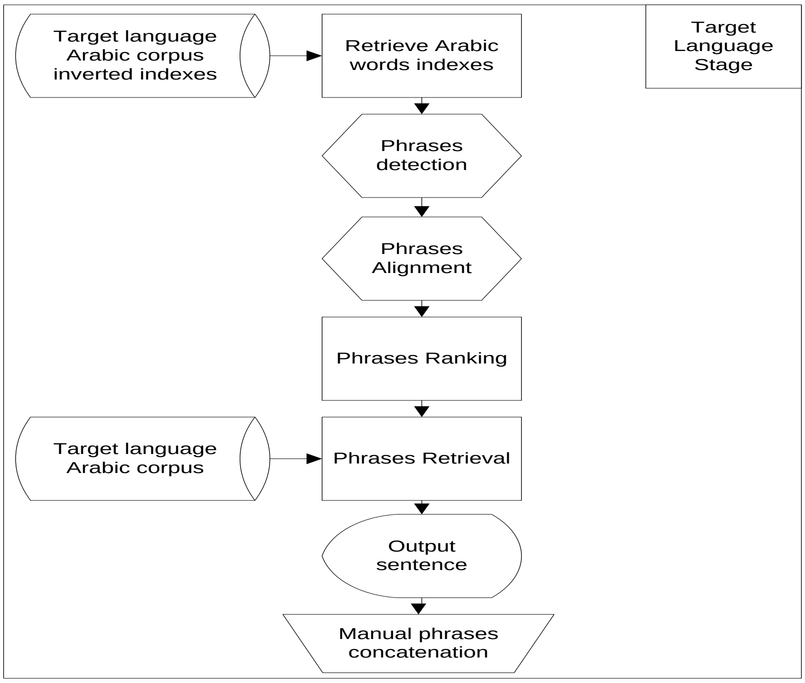

Target language Arabic corpus inverted indexes
Retrieve Arabic words indexes
Target Language Stage
Phrases detection
Phrases Alignment
Phrases Ranking
Target language Arabic corpus
Phrases Retrieval
Output sentence
Manual phrases concatenation

**Fig. 5.** Target language Stage

The phrase with the highest rank is retrieved from the TL corpus. A TL corpus of 29,233Arabic sentences is used. The average sentence length is 10 words. The TL Corpus is collected from the United Nation sessions Arabic translated documents. The TL corpus is used with the light stemmer to provide context based translation. Context based translation with word's stem search helps overcome some language features like definitive case, gender, count (single, double, plural) and linked pronouns [8], [10]. Retrieving a larger stream of words' meanings from certain sentence increases the word's context accuracy. The longer the retrieved sentence the more likely the correct context is detected [12].

A manual phrase concatenation can be done later to produce a full sentence translation.

## III Directed Words' Graph and Phrase Detection

We use directed words' graph, see figure 6, to represent SL and TL word relation. We are introducing the following notation as a way to represent the SL and TL sentences. We presented the SL as S = $\{s_1, s_2, \ldots, s_n\}$. $s_n$ is the word s with order n in the SL sentence S. The TL output sentence is represented as T = $\{t_1, t_2, \ldots, t_m\}$. $t_m$ is the word t with order m in the TL sentence T.

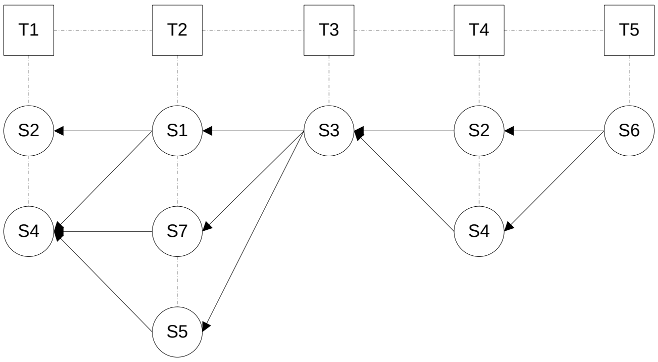


**Fig. 6.** Source to target directed graph

For each SL sentence S there is a TL sentence T where $t_i$ maps to the SL set of words $\{s_x, s_y, \ldots\}$ and $s_x \in S$. The final translated set is $T = \{t_{1S3}, t_{2S2}, \ldots, t_{mSn}\}$ where $t_{mSn}$ represent the TL word $t \in T$ with order m in the TL sentence T that corresponds to the SL word $s \in S$ with order n in the SL sentence S.

Our directed graph model has four dimensions (Node, Arc, Distance and Directed Walk/Phrase):

- Node: TL word $t_m$ corresponds to a single SL word order $s_x$ in the source sentence S. If the TL word $t_m$ maps to two different SL words $s_x$, $s_y$ in the input sentence S each SL word will be considered as a separate node. Nodes are represented in the form $t_{mSx}$.
- Arc: A connection between two adjacent word's translations in the TL sentence of two source words. It is represented as $(t_{mSx}, t_{nSy})$
- Arc Distance (length): The absolute difference between the source words order plus the absolute difference between target words order for two adjacent TL sentence words.

$$D(t_{mSx}, t_{nSy}) = |n-m| + |x-y| \qquad (\mathbf{3})$$

- Arc distance threshold (broken connection):

$$\text{threshold}(D(t_{mSx}, t_{nSy})) \leq 2\ . \qquad (\mathbf{4})$$

We chose the threshold to be 2 to make sure that only target words that correspond to two adjacent source words are connected. For one word that translates to two words or more, arc distance D will be equal to one between each two nodes.

- Allowed Directed Walk/Phrase: Is a set of connected nodes ordered from the highest TL order to the lowest TL order. The set is either has an ascending or descending source words order. The longer the detected phrase the more natural and accurate translation is generated [12].

Figure 6 shows that the retrieval engine returned the target sentence T={$t_1$, $t_2$, …, $t_5$}. Each target sentence word maps to a certain SL word. We can find that $t_1$ has the same stem that maps to $s_2$, $s_4$ translations, $t_1$ = { $s_2$, $s_4$ }. $t_2$ has the same stem that maps to $s_1$, $s_5$, $s_7$ translations, $t_2$ = { $s_1$, $s_5$, $s_7$ }. We should then calculate the arc distance and apply the threshold to detect the adjacent phrases.

Each phrase is defined by the following attributes: phrase minimum source word order, phrase maximum source word order, phrase minimum target word order, phrase maximum target word order, phrase count of source words and phrase count of target offsets.

## 1 Modified Dijkstra's Algorithm for Phrase Detection

We developed the following algorithm to detect the adjacent phrases. This process is based on Dijkstra's [1] shortest path algorithm. However in our algorithm we don't have a start and end points to get the minimum distance between them but we traverse all the target words found. Also we don't consider the shortest path arcs only but all arcs that have distance less than the identified threshold.

We traverse the target sentence word graph from its end to start to build the phrase list Ψ = {P1, P2, …Pn}. Where Pi = {tmSx, tnSy, …} is a phrase with internal distance D ≤ distance threshold between each adjacent target words of Pi. We can formalize our detection algorithm as follow: For each target sentence T that maps to the source sentence S we try to find Ψ. We use the following algorithm to get Ψ.

1. Start with an empty phrases list (Ψ = Φ).
2. Traverse the target sentence graph from the last target word to the first target word. An initial phrase set $P_1$ = {$t_{tailsx}$: $s_x \in t_{tail}$}, Ψ = {$P_1$}.
3. If the tail node $t_{tail}$ maps to more than one source word then Ψ = {$P_1$,$P_2$, …} where $P_i$ = { $t_{tailSx}$ }
4. Move to the previous target word $t_{i-1}$ and traverse all the source words $s_n \in t_{i-1}$ against all $P_i \in \Psi$ and $t_i \cap P_i \neq \Phi$. $P_i = P_i \cup \{t_{i-1sn}\}$ $\forall\, t_{i-1sn} \in t_{i-1}$ and distance $D(t_{i-1sn}, P_i) \leq$ threshold.
5. The nodes sources can either have an increasing or decreasing sequence. Otherwise the detection is branched and we have two phrases ($P_i$, P′). Ψ = Ψ ∪ P′. Where P′ = { $t_{i-1sx}$ , $t_{isn}$}.
6. If a broken connection is found, D > threshold. The detection algorithm add new phrase P′, Ψ = Ψ ∪ P′ and P′ = { $t_{i-1sx}$ }
7. If all nodes are traversed break. Else go to step number 4.

The above algorithm allows detecting all adjacent words of both source sentence and target sentence with the same or reversed word mapping order.

### 2 Phrase Alignment

The same directed graph technique used by the phrase detection is used for the phrase alignment. The following rules are used to consider a connected arc between two adjacent phrases.

1. Phrase 1 max target word order < Phrase 2 min target word order.
2. Phrase 1 max source word order < Phrase 2 min source word order.
3. Phrases distance is the distance between phrase 1 and phrase 2 = |Phrase 2 min target word order - Phrase 1 max target word order| + |Phrase 2 min source word order - Phrase 1 max source word order|.
4. Phrase distance threshold is 3: If two phrases have an arc with distance > 3 then this arc will be considered a broken connection. This threshold will allow having one missing source or one extra target word between each two phrases.

We redraw the graph with consideration that nodes represents phrases rather than words then run the same phrase detection algorithm with the above extra constrains.

## IV System Accuracy

The system was set under test to evaluate its accuracy. As the system is a hybrid system for specific domain. The system was tested with the United Nation English sessions. This will make more likely to find the corresponding translations with the same domain context as the TL Corpus. We used 100 input English sentences and got 68 accurate translations while the other 32 incorrect results are due to the missing dictionary translations or deficiencies in the phrase detection algorithm for Arabic language flexible order of words.

## V Conclusion

Our hybrid MT system can be used with the Arabic language. The hybrid system takes less cost and time to implement. The Arabic stemmer overcomes some of the morphological challenges that face the Arabic language translation. The TL corpus provides a context based translation guidance for the Arabic sentence. The hybrid system can be used when translating for languages with scarce resources. The hybrid system can’t replace the RBMT or the SMT at this stage.

Although the system produced translation with the current corpus state, but we found high need for a huge tagged Arabic Corpus to provide better translation. The tagged corpus should contain at least the following tags for each word (part of speech, gender, count, pronouns, and stem).

## VI Future work

Modify the phrase detection algorithm to handle the Arabic language flexible order of words. Study the algorithm performance and complexity with bigger size corpus and larger dictionary. Use the SMT methods to detect and align the phrases. Use the semantic features of the SL and TL words. Use tagged TL corpus.

## References

1. E.W. Dijkstra, "A note on two problems in connection with graphs". Numerische Math. Vol. 1, 269--271, (1959)
2. J. Hutchins, "Towards a definition of example-based machine translation". MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation; pp.63-70. (2005)
3. M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle". In: A.Elithorn and R.Banerji (eds.) Artificial and human intelligence, Amsterdam, North-Holland, pp.173-180, (1984)
4. Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to machine translation", Computational Linguistics, Volume 16, Issue 2: pp. 79–85, (1990)
5. Y. Dologlou, S. Markantonatou, G. Tambouratzis, O. Yannoutsou, A. Fourla, N. Ioannou, "Using Monolingual Corpora for Statistical Machine Translation: The METIS System". Proceedings of EAMT - CLAW 2003, Dublin, pp. 61--68, (2003)

6. V. Vandeghinste, I. Schuurman, M. Carl, S. Markantonatou, T. Badia, "METIS-II: Machine Translation for Low Resource Languages". Proceedings of the 5th international conference on Language Resources and Evaluation, Genoa, Italy, pp. 24--26, (2006)

7. P. Dirix, V. Vandeghinste, I. Schuurman, "A new hybrid approach enabling MT for languages with little resources", Proceedings of the 16th Meeting of Computational Linguistics in the Netherlands, pp. 117-132, (2006)

8. A. Chen, F. Gey, "Building an Arabic stemmer for information retrieval", Proceedings of the Eleventh Text Retrieval Conference (TREC 2002), NIST (2002)

9. L.S. Larkey, L. Ballesteros, M.E. Connell, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", In Proceedings of ACM SIGIR, pp. 269-274, (2002)

10. M.A. Attia, "Developing Robust Arabic Morphological Transducer Using Finite State Technology", the 8th Annual CLUK Research Colloquium, (2005)

11. C.D. Manning, P.Raghavan, H. Schutze, "Introduction to information retrieval", Cambridge University Press, pp.3--9, (2008)

12. T. Doi, H. Yamamoto, E. Sumita, "Example-Based Machine Translation Using Efficient Sentence Retrieval Based on Edit-Distance", ACM Transactions on Asian Language Information Processing (TALIP), Volume 4, Issue 4, pp.377--399, (2005)

13. V. Vandeghinste, P. Dirix, I. Schuurman, "Example-based Translation without Parallel Corpora: First experiments on a prototype", Proceedings of the Second workshop on EBMT, pp. 135--142 (2005)

Title: Towards a prototype intonational transcription system for Egyptian Arabic: testing the local f0 contour properties of intonational pitch accents in spontaneous speech.

Author: Dr Sam Hellmuth

Affiliation: University of York

Address: Department of Language & Linguistic Science
University of York, Heslington, York, UK YO10 5DD

Tel: +44 1904 432657

Fax: +44 1904 432673

Email: sh581@york.ac.uk

Abstract:

This paper sets out the properties of a prototype notational system for the transcription of Egyptian Arabic (EA) intonation, and tests the system by comparing the results of transcription of a small corpus of spontaneous conversational speech with known facts about EA intonation from experimental studies on 'laboratory' (read) speech. In particular the transcription system is used here to test the claim that a single pitch accent type is observed in EA. Specifically, a local phonetic contour annotation tier, adapted from that used in the IViE notation system for English, is used to establish the shape of the local f0 contour in the vicinity of the stressed syllable of a subset of target words identified in a small corpus of spontaneous speech. The results indicate that all of the variation in the local f0 contour can be explained from properties of the local prosodic context (target syllable type, prosodic context) and thus support the claim that a single accent type is sufficient for the description of spontaneous speech. The paper discusses potential future development and applications of the proposed transcription system for EA intonation.

Keywords: intonation, transcription, Egyptian Arabic, alignment, pitch accents

Running head: Towards a prototype intonational transcription system for Egyptian Arabic.

## I Introduction

This paper sets out the properties of a first prototype notational system for the transcription of Egyptian Arabic (EA) intonation, and tests the system by comparing the results of transcription of a small corpus of spontaneous conversational speech with known facts about EA intonation from experimental studies on ‘laboratory’ (read) speech. In particular the transcription system is used here to test the claim that a single pitch accent type is observed in EA. This is achieved by using a ‘local phonetic contour’ tier, adapted from that used in the IViE [1,2] and IVTS [3] notation systems, proposed for English and French respectively. The paper explores one way in which this annotation tier might be implemented for useful transcription of the intonation of spontaneous EA speech. The background to the study is set out in section II; section III describes the properties of the proposed transcription system; section IV describes the results of the transcription and compares them to known facts from experimental studies; the paper concludes with a discussion in section V.

## II Background to the study

Egyptian Arabic (EA) is defined here as the colloquial dialect of Arabic spoken in Cairo, Egypt, and by educated middle class Egyptians throughout Egypt. The segmental and metrical phonology of EA is well-described, and has been the subject of much research (see [4] for a summary). In contrast, EA intonation has received comparatively less attention, and this situation is paralleled across most spoken dialects of Arabic. Recent studies have however established some key properties of EA intonation, including the typical shape of global contours in different utterance types[5], the lack of complete deaccentuation after a focus[6,7], the alignment patterns of pitch peaks occurring on pre-nuclear (non-phrase-final) accented words[8] and the frequent distribution of these pitch peaks, such that one occurs on every content word [9]. This array of facts have led the author to propose (in other work[10,11]) that only a single pitch accent category is necessary for the phonological description of EA pitch accents (L+H*). Other authors have reached a similar conclusion for non-phrase-final accents in the Egyptian pronunciation of Modern Standard Arabic[12,13] (although they assign a different phonological label to the single accent in their studies). In general, these claims are based on analysis of either scripted or broadcast speech, which might be expected to be uniform in character. The aim of the small pilot study reported here was to test whether a single accent category is sufficient for the description of non-phrase-final accents as they are realised in fully spontaneous colloquial EA speech. It is however difficult to obtain quantitative generalisations from spontaneous speech due to the inevitable variation in segmental and prosodic contexts of potential target accented words (though not impossible[14]). In order to solve this problem the present paper proposes a prototype notation system for fine-grained prosodic transcription of EA intonation. Specifically, a local phonetic contour annotation tier is used to establish the shape of the local f0 contour in and around the stressed syllables of a subset of target words identified in a small corpus of spontaneous speech. The properties of the local f0 contour in these words is compared to known facts about EA intonation established experimentally, and the range of variation in the observed shape of pitch accents is discussed.

## III Methodology

The Intonational Variation in English (IViE) labelling system was designed to facilitate the creation of a corpus of "directly comparable transcriptions of several varieties of English in a single labelling system"[15: p1]. It is similar to the Tones and Break Indices (ToBI [16]) labelling system in that phonological pitch targets ('tones') are labelled on a tier separate from other aspects of the transcription. ToBI comprises: a tone tier, an orthographic tier, a break index tier and a miscellaneous tier. In ToBI, "the tone and break index tiers represent the core prosodic analysis" [16: p8]. The innovation in the IViE labelling system was the addition of a tier for the labelling of 'acoustic-phonetic structure'; this tier comprises a labelling of the "shape and alignment of f0 patterns relative to the location of (accented) strong syllables" [17: p2]. With the advent of technology such that labelled transcriptions are almost invariably presented alongside a spectrogram and f0 pitch trace of the speech fragment in question, one could argue, as Wightman[18] has done, that the local f0 contour does not need to be labelled, since users of the transcription can themselves see and interpret the f0 contour. However, the local f0 contour tier in IViE can play an important role whilst a fully fledged phonological analysis of the language or varieties concerned is being developed. It is this strength of the IViE approach that is exploited here, in order to test the (phonological) claim that a single accent type is sufficient for the description of EA accents in prenuclear position, in spontaneous as well as scripted speech. It is thus the necessary properties of a local f0 contour tier for EA that are explored in the present paper; future work will explore the required properties of other tiers for accurate transcription of EA intonation.

The dataset transcribed for the present study is a set of 100 target words from a spontaneous speech telephone conversation (between two female speakers, from the Call Home corpus[19]). The relevant portions of the conversation were transcribed by the author with reference to the f0 contour and spectrogram using Praat version 4.6.10[20], on three tiers: i) *words*, broad phonetic transcription, on separate tiers for each speaker, ii) *tones* (phonological labels: pitch accents and boundary tones), and iii) *local f0 contour* (phonetic-acoustic structure tier). The criterion for selection of target words was the presence of little or no perturbation of the f0 contour during the word. The position of the pitch peak (or valley) within each target word was identified automatically using the pitch maxima (minima) function within Praat, and a dummy "H*" or "L*" label assigned to the peak on the tones tier.

The shape of the f0 contour immediately before and after the pitch peak/valley was then annotated on the local f0 tier, as follows. Firstly, a capital letter (*L*, *M* or *H*) was assigned to denote the height of the f0 contour during the accented syllable, according to whether the height of the pitch peak/valley was low, mid or high in the speaker's pitch range. Next, the relative height of the f0 contour on unaccented syllables occurring immediately before and after the peak/valley was described: adjacent pitch low in the speaker's pitch range was transcribed '*l*' , adjacent pitch in middle of the speaker's pitch range was transcribed '*m*' and adjacent pitch high in the pitch range (and thus in most cases level with the pitch on the accented syllable itself) was transcribed '*h*'. The speakers' pitch range throughout the whole conversation had previously been calculated for each speaker independently (speaker A: 120-520Hz; speaker B: 80-520Hz). Finally, the position of the pitch peak/valley within the accent syllable was checked; if the peak/valley occurred very early or very late in the accented syllable this was noted, by insertion of a line '|' immediately before the capital letter (for an early peak/valley) or immediately after the capital letter (for a late peak/valley). This additional labelling convention is proposed here and is not part of the original IV notation systems. The set of 100 labels annotated on the local f0 contour was then harvested, along with relevant information about the syllable structure, stress position and prosodic context of each target word.  A sample annotation grid is provided in Figure 1.

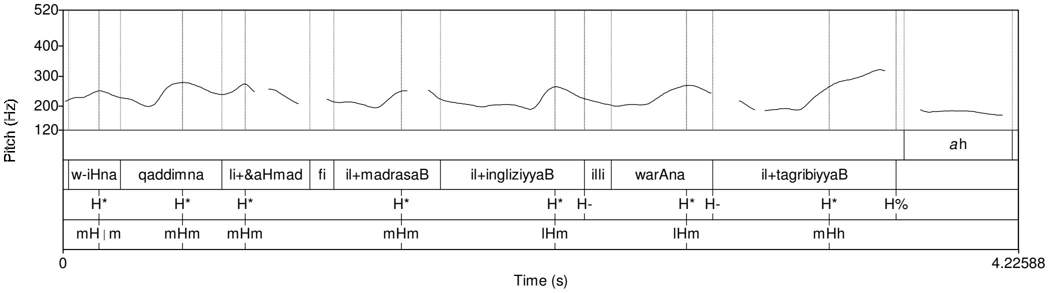


Figure 1: Sample of annotated text (*tiers top to bottom: words B, words A, tones, local f0*).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| w-iHna | qaddimna | li+&aHmad | fi | il+madrasaB | il+ingliziyyaB | illi | warAna | il+tagribiyyaB |
| ˈwiħna | ʔadˈdimna | ˈlaħmad | fi | l-madˈrasa | l-ingiliˈzijja | ˈilli | waˈraːna | t-tagriˈbijja |
| and we | applied | for-Ahmed | in | the-school | the-English | REL | behind-us | the-near |

## IV Results

All of the accents were transcribed with an ‘H’ on the accented syllable (and none with L), that is, no pitch valleys were found, only pitch peaks (note that the conversation does include declarative questions). This finding suggests that no ‘L*’ type phonological labels are required for the description of pitch accents in EA. The next two sections explore the results of the transcription with respect to the alignment of H peaks within the accented syllable (section IV.1) and the shape of the preceding and following f0 contour (section IV.2).

### IV.1 Position of the H peak within the accented syllable

In the great majority of cases (80%) the peak was observed to occur within the accented syllable. Consistent alignment of the high peak within the accented syllable is a good reason to propose that the H tone is the phonologically strong tone (the ‘starred tone’) of the EA pitch accent. The exact position of the H peak within the accented syllable was found to co-vary with the position of the stressed syllable within the word and with the position of the word within the phrase, factors known to affect peak alignment, particularly in pre-nuclear rising pitch accents [21,22]: the nearer the accented syllable is to a prosodic boundary (at the word or phrase level) the earlier the peak will be aligned. For example, there were just two cases where the pitch peak was observed to be very early in the accented syllable (denoted by ‘|H’); both were words occurring at the end of an intonational phrase and with stress on the final syllable: [inʃaʔalˈlaah], [tuˈreːl]. In addition, peak alignment was observed to vary with the type of syllable bearing stress in the word (CV, CVV or CVC) [cf. 8]. In phrase-initial words the H peak of the accent was observed to be at the very end of the stressed syllable in all cases (‘H|’), regardless of the position of stress in the word, however in phrase-medial words, the H peak was at the end of the syllable only in words bearing initial-stress, e.g.

[ˈɣaːja]. In sum, there is no evidence in the present (small) corpus for variation in the position of the H peak within the accented syllable other than that caused by local prosodic factors.

### IV.2 Shape of the local f0 contour before and after the H peak

Turning to the local pitch contour immediately before and after the H peak, the most common transcription used in this pilot transcription set was '*mHm*', observed on 55% of target words. Overall, the level of preceding pitch was very consistently observed to be '*m*' (on 82% of target words). There are three types of context in which preceding pitch was labelled 'l', indicating that the f0 contour rose to the H peak from a somewhat lower level in the speakers pitch range. One such context was on words which were the first accented word in the utterance, preceded only by an unaccented function word (e.g. [wi hijja…] 'and she…'), and in these cases the rise in pitch starts already at the onset of the utterance (suggesting a possible %L initial boundary tone). Preceding local pitch also seems to be somewhat lower when there are relatively large number of unstressed syllables between successive accents. This can result in a short 'low plateau' between two accents [cf. 23], and was observed in polysyllabic words such as [ingiliˈzijja] 'English' which contains three unstressed syllables before the stressed syllable, as in Figure 1. Finally, preceding local pitch seemed also to be lower when the word is followed by relatively high pitch, such as in words followed by a high phrase tone (H-), as observed on the word [waˈraːna], also in Figure 1. There were no instances at all of preceding high pitch ('*h*'). In sum then, the level of preceding pitch seems to be stable, and any variation can be ascribed to factors in the local prosodic context.

Turning to the level of following pitch, this was also most often observed to be '*m*' but was more variable, with following 'm' observed in 66% of target words only. Occurrence of a following ‘h’ or ‘l’ was however also found to be dependent on local prosodic factors such as other tonal events following the accented word. For example, all words followed by a H% boundary tone showed a continuous rise throughout the accented syllable (thus ‘*h*’ following the H peak), as in the word [tagriˈbijja] in Figure 1. In contrast, words followed by a L- phrase tone or L% boundary tone all show either mid ‘*m*’ or low ‘*l*’ following pitch. A word was observed to be more likely to have a ‘*mHl*’ shaped accent (than ‘*mHm*’) if the next stressed syllable was relatively distant.

## IV.3 Summary

This detailed transcription of target words found in a small corpus of naturally occurring speech suggests that observed variation in the local pitch contour both preceding and following the accented syllable in EA can be described as a function of the surrounding tonal environs. Such variation is arguably therefore predictable from the prosodic context, and does not constitute evidence for additional pitch accent types in EA of any kind. Nonetheless it is interesting to note that the relative height of the following local pitch contour was observed to be more variable than that of the preceding local pitch contour. This matches findings regarding the alignment properties of L turning points preceding and following the H peak in EA, whereby it is the preceding L tone that is stably aligned at the onset of the stressed syllable [8,24].

## V Discussion and conclusion

This paper has made use of the labelling conventions of a local f0 contour tier, in order to test a claim regarding the best phonological representation of the range of phonetic realisations of accented syllables in a limited corpus of target words found in spontaneous colloquial EA speech. The labelling system makes it possible to categorise potential candidates for membership of a single phonological category according to slight variations in the actual local f0 contour, and thus to determine whether any of these variation should in fact be considered as evidence of membership of some other phonological category. In the present study, it was found that all of the observed variations in the local f0 contour could be attributed to factors in the local prosodic environment, and it is thus argued that all of the 100 accented syllables should be assigned membership of a single phonological category. This was labelled with a dummy 'H*' in the annotation. We propose that the correct phonological designation is L+H*, for two reasons, both of which have previously been established experimentally [8] but which also become evident from the current study: firstly, the accented syllable is invariably marked by a pitch peak (H) which is positioned consistently within the stressed syllable, and any exceptions can be explained with reference to factors in the prosodic context; secondly, the level of pitch preceding the peak is more stable than the level of pitch following the peak, suggesting the presence of a leading L tone. Use of a prosodic annotation tier at the local f0 level is thus shown to be a useful tool in establishing phonological categories in spoken EA.

The labelling conventions of the local f0 contour tier represent only a small part of the range of tools available within the IV transcription systems; other tools include the use of a rhythmic tier (and corresponding Implementation Domains) to capture the relationship between rhythmic structure and intonational pitch events[15] and a global f0 contour tier proposed (in [3]) for transcription of declination and downstep. In addition we believe that it would be profitable to define and test a set of Break Indices for EA intonation[cf.16], in order better to determine the prosodic phrasing properties of spontaneous speech, and implementation of all these additional tools is planned. A crucial test of any successful labelling system however is the extent to which independent transcribers are able to use it to reach consensus about how a particular stretch of speech should be labelled. A study of intertranscriber agreement using the proposed labelling system is thus also necessary, to test the robustness of the prototype annotation guidelines for the local f0 contour tier demonstrated here, and to develop robust transcription norms for EA on rhythmic, global f0 contour and break indices tiers.

Establishment of an agreed intonational transcription system for EA will facilitate future interdisciplinary research between phonological and speech technology research teams. In addition a working transcription system will permit better descriptions of aspects of EA intonational phonology which remain as yet undescribed as well as paving the way for future documentation of intonational variation in Egypt, and comparison of EA intonation to that of other varieties of Arabic.

## References

1. E. Grabe, F. Nolan, & K. Farrar. (1998) IViE - a comparative transcription system for intonational variation in English. Proceedings of the 5th International Conference on Spoken Language Processing (Sydney, Australia)., 1259-1262.
2. E. Grabe. (2001) The IViE Labelling Guide. (Version 3.0) http://www.phon.ox.ac.uk/IViE/guide.html.
3. B. Post & E. Delais-Roussarie. (2006) Transcribing intonational variation at different levels of analysis. *In* Proceedings of Speech Prosody 2006. R.Hoffmann and H.Mixdorff, eds. Dresden, TUD Press Verlag der Wissenschaften GmbH.
4. J.C.E. Watson. (2002) The phonology and morphology of Arabic. Oxford, OUP.
5. O.A.G. Ibrahim, S.H. El-Ramly, & N.S. Abdel-Kader. (2001) A model of F0 contour for Arabic affirmative and interrogative sentences. 517-524. Mansoura University, Egypt. 18th National Radio Science Conference.

6. K. Norlin. (1989) A preliminary description of Cairo Arabic intonation of statements and questions. Speech Transmission Quarterly Progress and Status Report, 1, 47-49.

7. S. Hellmuth. (2006) Focus-related pitch range manipulation (and peak alignment effects) in Egyptian Arabic. *In* Proceedings of Speech Prosody 2006. R.Hoffmann and H.Mixdorff, eds. pp. 410-413. Dresden, TUD Press Verlag der Wissenschaften GmbH.

8. S. Hellmuth. (2007) The foot as the domain of tonal alignment of intonational pitch accents. Proceedings of the 16th ICPhS, Saarbruecken, Germany.

9. S. Hellmuth. (2007) The relationship between prosodic structure and pitch accent distribution: evidence from Egyptian Arabic. The Linguistic Review, 24, 289-314.

10. S. Hellmuth. (2006) Intonational pitch accent distribution in Egyptian Arabic., Unpublished PhD thesis, SOAS. PhD.

11. D. Chahal & S. Hellmuth. (2009) Comparing the Intonational Phonology of Lebanese and Egyptian Arabic. *In* Prosodic Typology. Volume 2. Oxford, Oxford University Press.

12. D. Rastegar-El Zarka. (1997) Prosodische Phonologie des Arabischen., Unpublished PhD thesis, Karl-Franzens-Universität Graz. PhD.

13. K. Rifaat. (2005) The structure of Arabic intonation: a preliminary investigation. *In* Perspectives on Arabic Linguistics XVII-XVIII: Papers from the seventeenth and eighteenth annual symposia on Arabic linguistics. M.T.Alhawary and E.Benmamoun, eds. pp. 49-67. Amsterdam/Philadelphia, John Benjamins.

14. S. Hellmuth, F. Kügler, & R. Singer. (2007) Quantitative investigation of intonation in an endangered language (and how this methodology could help us understand the intonation of better-studied languages). Proceedings of the Language

Documentation and Linguistic Theory Conference (SOAS, London, December 2007).

15. E. Grabe. (2001) The IViE Labelling Guide. (Version 3.0). University of Cambridge.

16. M. Beckman & G.A. Elam. (1993) Guidelines for TOBI Labelling (version 3.0 1997). The Ohio State University Research Foundation.

17. E. Grabe. (2001) The IViE Labelling Guide. (Version 3.0) [http://www.phon.ox.ac.uk/IViE/guide.html].

18. C.W. Wightman. (2002) ToBI or not ToBI? Proceedings of Speech Prosody 2002.

19. K. Karins, M. Liberman, C. McLemore, & E. Rowson. (2002) CallHome Egyptian Arabic Speech Supplement LDC2002S37. Philadelphia, Linguistic Data Consortium.

20. P. Boersma & D. Weenink. (2007) Praat: doing phonetics by computer (Version 4.6.10) [http://www.praat.org].

21. P. Prieto, J. van Santen, & J. Hirschberg. (1995) Tonal alignment patterns in Spanish. Journal of Phonetics, 23, 429-451.

22. D. Chahal. (2001) Modeling the intonation of Lebanese Arabic using the autosegmental-metrical framework: a comparison with English., Unpublished PhD thesis, University of Melbourne. PhD.

23. D. El Zarka & S. Hellmuth. (2008) Variation in the intonation of Egyptian Formal and Colloquial Arabic. Langues et Linguistique.

24. S. Hellmuth & D. El Zarka. (2007) Variation in phonetic realisation or in phonological categories? Intonational pitch accents in Egyptian Colloquial Arabic and Egyptian Formal Arabic. Proceedings of the 16th ICPhS, Saarbruecken, Germany.

# المعالجة الحاسوبية للتطور الدلالي للمشتقات

# في العربية

د. مدحت يوسف السبع

كلية دار العلوم / جامعة القاهرة

والأستاذ المساعد بجامعة الإمام محمد بن سعود الإسلامية بالرياض

## تقديم

الحمد لله رب العالمين ، والصلاة والسلام على نبيه الأمين، وبعدُ ...

فهذه ورقة بحثية تدخل في مجال علم اللغة الحاسوبي، وموضوعها هو: **المعالجة الحاسوبية للتطور الدلالي للمشتقات في العربية**، وقد حاولت هذه الورقة من الأمر عسيراً؛ إذْ إنها تدخل في منطقة ليست سهلة الارتياد، فهي تناقش الدلالة من منظور حاسوبي، فلئن كانت الأولى (الدلالة) يتحاماها عدد غير قليل من الباحثين فإن الثانية (المعالجة الحاسوبية) مما يندر وجود باحثين مجيدين لها، عارفين مبادئها.

وقد يسّر الله لصاحب هذه الورقة العمل في مجال حوسبة العربية خمسة عشر عاماً، توجت بالحصول على درجة الدكتوراه في تخصص حوسبة العربية[1]، وليس في هذا كبير نفع إلا أنه كان من تيسير الله أن جُمع أمام ناظريَّ الجانبان: التنظيري والتطبيقي، ومن ثم فهذه الورقة نتاج عمل اقتضى علماً.

[1] الرسالة بعنوان : العلاقات التركيبية في الجملة الفعلية القرآنية: دراسة نحوية حاسوبية، مقدمة إلى كلية دار العلوم، جامعة القاهرة، بإشراف سعادة الدكتور علي أبوالمكارم ، وسعادة الدكتور نبيل علي ، وقد أجيزت عام 2004 م بتقدير مرتبة الشرف الأولى.

# المعالجة الحاسوبية للتطور الدلالي للمشتقات فى العربية

**د. مدحت يوسف السبع**

**كلية دار العلوم- جامعة القاهرة**

**و الأستاذ المساعد بجامعة الإمام محمد بن سعود الإسلامية بالرياض**

## ملخص الورقة:

ناقشت هذه الورقة موضوع (**المعالجة الحاسوبية للتطور الدلالي للمشتقات في العربية**) في أربع قضايا كبرى، هي: المشتقات المتطورة دلاليًّا: تعريف وتمثيل وتحليل، أنماط التطور الدلالي في المشتقات، أسباب ظاهرة التطور الدلالي في المشتقات، متطلبات المعالجة الحاسوبية للتطور الدلالي في المشتقات.

وتشمل كل قضية كبرى من هذه الأربعِ مسائلَ فرعيةً ستبين عنها هذه الورقة.

## هدف هذه الورقة:

تهدف هذه الورقة إلى تقديم دراسة لغوية حاسوبية لإشكالية التطور الدلالي للمشتقات، بحيث تقدم منهجاً يصلح للبحث في علم اللغة الحاسوبي العربي، وتتضح فيها سمات المنهج وأسس المعالجة الحاسوبية للعربية.

وقد اعتمد البحث فكرة أن الجذر اللغوي هو أصل الاشتقاق، وما عداه مشتق منه، ومن ثم فليس الفعل هو أصل المشتقات، وليس المصدر كما يرى آخرون.

والمقصود بالتطور الدلالي للمشتقات في هذا البحث هو تحولها من الاشتقاق إلى الاسمية عن طريق التطور الدلالي الذي له أسباب ستفرد لها هذه الورقة مساحة لعرضها، ومن ثم أصبح المشتق لا عمل نحويا أو تركيبيا له؛ لأنه صار اسماً له ما للأسماء، وعليه ما عليها، فكيف يدرك هذا الحاسوب، وكيف يتعامل معه أثناء المعالجة الحاسوبية؟

وقد قامت هذه الدراسة على مدونة لغوية ضخمة لكبرى الشركات ذات الاهتمام الخاص بمعالجة العربية حاسوبيا، وقد اتسع مجال هذه المدونة الزمني فغطى العربية قديما وحديثا، وتنوعت الفنون التي تشملها فضمت علوما مختلفة، وإنها كذلك استوعبت كتابات عدد من المؤلفين متنوعي المشارب.

# المعالجة الحاسوبية للتطور الدلالي للمشتقات في العربية

## أولا – المشتقات المتطورة دلاليًا

من الظواهر اللغوية التي لها وجود واضح في مجال المعالجة الحاسوبية للغة تطور المشتقات دلاليا، وقد تعددت المشتقات التي دخلها التطور الدلالي. وبتتبع الظاهرة استطاع البحث أن يحصرها في عدد من المشتقات، هي : المصدر، و اسم المصدر، و المصدر الميمي، و المصدر الصناعي، و اسم المرة، و اسم المكان، و الفعل، و اسم الفاعل، و اسم المفعول، و الصفة المشبهة، و صيغة المبالغة، و أفعل التفضيل، و الصفة المنسوبة. وهذا توضيح:

### 1- تطور المصدر دلاليًا:

قد يتطور المصدر دلاليا فيتحول إلى اسم، ومن ثم لم يعد مجال لعمل تركيبي، فقد صار اسما لا يتطلب عملا نحويا من فواعل ومفاعيل أو مكملات، ومما حدث فيه تحول عن المصدر:

كلمة (التنوين): فكما أنها تستعمل مصدرًا تستعمل اسمًا «والتنوين وهو في الأصل مصدر نونت؛ أي: أدخلت نونًا، ثم غلب حتى صار اسمًا لنون تلحق الآخر لفظًا لا خطًا لغير توكيد»[1]. أي يستعمل مصدرا واسما متحولا عن المصدر.

ومنه كذلك: مباراة، جواز، وسؤال وغيرها، فكل منها لا عمل تركيبيًا له في جملته؛ لأنه لم يعد مشتقا عاملا، بل تحول إلى اسم.

### 2- تطور اسم المصدر دلاليًا:

قد يتحول اسم المصدر إلى الاسمية، ومن ذلك لفظتا : النكرة والمعرفة، يقول صاحب (شرح التصريح): «وهما في الأصل اسما مصدرين لنكرته وعرفته، فنقلا وسمي بهما الاسم المنكر والمعرف»[2]. أي استعملا اسمي مصدر ومتحولين عنه إلى الاسم.

وإن كان الشيخ يس قد اعتدهما مصدرين، يقول: «النكرة والمعرفة مصدران في الأصل ثم نقلا وسمي بهما نوعان من الأسماء»[3].

[1] شرح الأشموني: (30/1).

[2] شرح التصريح (91/1).

[3] حاشية يس: (91/1).

والأولى أنهما اسما مصدرين لقلة عدد حروفهما عن حروف الفعل.

### 3- تطور المصدر الميمي دلاليًّا:

قد يتحول المصدر الميمي دلاليا إلى اسم، ومن ذلك:

مفاد: مفاد الأمر محتواه ، وما يستفاد منه. [1]

ومنه كذلك مأخذ : المأخذ المنهج، وما يعاب من العمل والعامل ، وجمعها مآخذ.[2]

ومؤتمر: مجتمع للتشاور والبحث في أمر ما. [3]

وواضح من صياغة المعاجم لمعاني هذه المصادر الميمية أنها تنص على تحولها من المصدر الميمي إلى الاسم الذي لا يبحث له عن تعلق تركيبي أو عمل نحوي ينشئه في جملته.

### 4- تطور المصدر الصناعي دلاليًّا:

قد يتحول المصدر الصناعي إلى اسم، ومن ذلك:

دورية: العسس يطوفون ليلا (محدثة). [4]

جاهلية: ما كان عليه العرب قبل الإسلام من الجهالة والضلال. [5]

فكل من اللفظتين (دورية وجاهلية) لا يتضح في أي منهما معنى المصدر الصناعي، وإنما تحولتا إلى اسمين.

### 5- تطور اسم المرة دلاليًّا:

قد يتحول اسم المرة دلاليا إلى اسم، ومن ذلك:

نزلة: التهاب في الأنف والمسالك الهوائية، وتطلق على ما يطرأ على الصحة من وعكة أو مرض.

جولة: جمعها جوْلات، زيارة أو تجول لغرض من الأغراض. [6]

### 6- تطور اسم المكان دلاليًّا:

قد يتحول اسم المكان إلى الاسم، من ذلك كلمة (مسجد) فإنها تطلق ويقصد بها أمران، هما:

أ ـ مكان السجود.

ب ـ مكان معين مخصص لأداء الصلاة بقيامها وركوعها وسجودها.

[1] المعجم العربي الأساسي: فود.

[2] المعجم الوسيط: أخذ.

[3] المعجم الوسيط: أمر.

[4] المعجم الوسيط: دور.

[5] المعجم الوسيط: جهل.

[6] المعجم العربي الأساسي: جول.

وأعده الأشموني اسمًا، يقول: «إضافة الاسم إلى الصفة، نحو: مسجد الجامع»[1]، أضاف الصفة (الجامع) إلى الاسم (مسجد). أي استخدم اسم مكان ومتحولا عنه إلى الاسم.

ومنه كذلك :

مبدأ: (ج) مبادئ ، مبادئ العلم أو الفن أو الخلق أو الدستور أو القانون: قواعده الأساسية التي يقوم عليها ولا يخرج عنها. [2]

مبرة وملجأ ومستشفى: مصدر ميميّ، وموضع البر، كالملجأ والمستشفى (محدثة).[3]

منزلة: المنزلة المكانة والرتبة.[4]

## 7- تطور الفعل دلاليًّا:

قد يتطور الفعل دلاليا فينتقل إلى مجال العلمية والتسمية به، وهذا يؤدي إلى إعرابه "والفعل إذا نقل إلى العلمية وجب أن يعرب، نحو: كعب وتغلب واضرب». [5]

## 8- تطور اسم الفاعل دلاليًّا:

قد يتطور اسم فاعل دلاليا ويتحول إلى اسم، ومن ذلك كلمة (دابة) وجمعها دواب، فإنها تحولت عن اسم الفاعل إلى الاسم «وترى العرب يقولون: ثلاثة دواب (بالتاء)، إذا قصدوا ذكورًا؛ لأن الدابة، وهي لغة كل ما يدب على الأرض، صفة في الأصل غلبت عليها الاسمية، فكأنهم قالوا ثلاثة أحمرة جمع حمار، وسمع من كلامهم: ثلاث دواب ذكور، بترك التاء، لأنهم اعتبروا تأنيث اللفظ، وأجروا الدابة مجرى الاسم الجامد نظرا إلى الحال، فلا يجرونها على موصوف، قاله ابن مالك أخذا من قول ابن عصفور، وأما ثلاث دواب فعلى جعل الدابة اسمًا» [6].فكلمة (دابة) تستعمل استعمالين، هما: صفة واسم. ومنه كذلك:

حاجب: الحاجب البواب (صفة غالبة). [7]

الداهية: رجل داهية: بصير بالأمور، والأمر المنكر العظيم. [1]

[1] شرح الأشموني :(242/2)

[2] المعجم الوسيط: بدأ

[3] المعجم الوسيط: برر

[4] المعجم الوسيط: نزل

[5] شرح المفصل (28/4) بتصرف.

[6] شرح التصريح :(271/2، 272)، شرح الأشموني: (63/4).

[7] المعجم الوسيط: حجب

### 9 – تطور اسم المفعول إلى اسم:

قد يتحول اسم المفعول إلى اسم ، ومن ذلك:

مؤخر: نهاية الشيء من الخلف ، يقال مؤخر السفينة ومؤخر البناء. ومؤخر الدين والصداق. [2]

مدرج: مكان ذو مقاعد متدرجة (محدثة) [3]

معدّل: جمعه معدلات ، متوسط (معدل الإنتاج)، (معدل السرعة) ، (معدل الدخل). [4]

### 10– تطور الصفة المشبهة دلاليًّا:

قد تطور الصفة المشبهة دلاليًّا فتتحول إلى اسم، ومن ذلك:

أجير : من يعمل بأجر. [5]

الكبيرة: الإثم الكبير المنهي عنه شرعا، وجمعها كبائر. [6]

### 11– تطور صيغ المبالغة دلاليًّا:

قد تطور صيغ المبالغة دلاليًّا فتتحول إلى اسم، ومن ذلك:

حضّانة: وسيلة صناعية يتم فيها اكتمال نمو الجنين. [7]

مرسال: الرسول ، والناقة السهلة السير السريعته، وجمعه مراسيل. [8]

### 12– تطور أفعل التفضيل دلاليًّا:

قد يتحول أفعل التفضيل إلى اسم مذكر أو مؤنث.

من ذلك نقل (الأخضر) للتعبير بها عن الطحلب «قوله: للأخضر الذي يعلو الماء، أي: الشيء الأخضر، لا الوصف الأخضر، وعبر بعض الشافعية بقوله: النبت الأخضر، وعبارة القاموس: الطحلب خضرة تعلو الماء. وصفها بأنها تعلو الماء يقتضي أنه أراد الجرم الأخضر،

[1] المعجم الوسيط: دهي.

[2] المعجم الوسيط: أخر.

[3] المعجم الوسيط: درج.

[4] المعجم العربي الأساسي: عدل.

[5] المعجم الوسيط: أجر.

[6] المعجم الوسيط: كبر.

[7] لم أعثر عليها في المعاجم اللغوية العربية الحديثة.

[8] المعجم الوسيط: رسل.

لا الوصف، لأنه لون قاتم بالماء، ولا يقال إنه يعلوه»[1].أي تحول من الصفة إلى الاسم، فأصبح يستعمل بصورتين: صفة، واسم.

ومنه كذلك :

أشد: الأشد الاكتمال... وأشد في صيغة الجمع ومعناه، ولم يسمع لها مفرد.[2]

## 13 – تطور المنسوب دلاليًّا:

قد يحدث تطور دلالي في الصفة المنسوبة فتتحول إلى اسم، ومن ذلك: لفظة (العربية) استخدمت متحولة، وإن لم يُنص على ذلك، وقد سبق أن ذكر الخلاف حولها.

ومنه كذلك:

اختصاصي: جمعها اختصاصيون : مختص بالشيء متخصص فيه. [3]

مسماري: نوع من الخط يرجع فضل وضعه إلى السومريين. [4]

[1] حاشية يس: (356/2).

[2] المعجم الوسيط: شدد

[3] المعجم العربي الأساسي: خصص

[4] المعجم العربي الأساسي: مسماري.

## ثانياً – أنماط التطور الدلالي في المشتقات

التحول الدلالي في المشتقات من حيث تصوّر المتحول عنه نوعان: تام، وغير تام.

### 1. التحول التام:

إن ثمة شروطا في المتحول إذا تحققت أصبح متحولاً تحولاً تامًا:

وهذه الشروط هي:

- ألا يتصور معه وجود موصوف.
- لا يعمل عمل الصفات.
- لا يتحمل ضميرا.

يقول الشاطبي: « والدليل على أن هذه الأسماء انسلخ عنها الوصفية أنها لا تجري صفات على موصوف، ولا تعمل عمل الصفات، ولا تتحمل ضميرًا» [1]. وهذا توضيح:

**أ– المتحول التام لا يقدر معه موصوف:**

الصفة المتحولة اسما لا يقدر معها موصوف «لم يتعرض لوجوب حذف المنعوت مع أنه قد يجب، تقول: جاء الفارس؛ أي الرجل الراكب الفرس، ولا تقول جاء الرجل الفارس، وتقول: جاء الصاحب؛ أي الرجل الصاحب، ولا تقول: جاء الرجل المصاحب» [2].

لا يجوز ظهور الموصوف في مثل ما سبق، ولكن الأولى أن يقال إنه تحول دلاليا فأصبح يعبر عن الذات نفسها، وليس عن صفة من صفاتها.

يقول الشيخ يس: «والصفة تجري على موصوف لا محالة إلا أن يغلب عليها الاسمية كصاحب وركب» [3].

**ب . لا يقدر مع المتحول التام ضمير:**

لا يقدر مع التحول التام ضمير يعود على متقدم لأنه صار اسما: «فكل من: زيد وأسد وصاحب عندهم من قبيل الجوامد فلا يتحمل ضمير المبتدأ نحو: هذا زيد، وهذا أسد، وهذا صاحب، فليس في شيء منها ضمير يعود على المبتدأ» [4].

**ج – المتحول التام لا يعمل عمل الصفات:**

[1] حاشية الصبان: (164/1) ، شرح التصريح: (142/1).

[2] حاشية يس: (118/2).

[3] حاشية يس : (34/2).

[4] شرح التصريح : (160/1).

إذا تحولت الصفة إلى اسم فقدت عملها «وصاحب يقبل (أل) المؤثرة لتعريف، فنقول: الصاحب، وليس أل فيه موصولة؛ لأنه قد تنوسي فيه معناه الأصلي بحسب الاستعمال، وصار من قبيل الجوامد؛ ولذلك لا يعمل، لا تقول: مررت برجل صاحب أخوه عمرًا. »[1].

ويقول عباس حسن: «كلمة صاحب هنا . يقصد عندما تكون (ذو) نكرة تحل محلها (صاحب) التي تقبل (أل) لأنها بمعناها – ليست اسم فاعل معناه مصاحب؛ لأن معناها الأصلي الدال على التجدد والحدوث قد أهمل، وغلبت عليها (الاسمية) المحضة، فألحقت بالأسماء الجامدة؛ ولذلك لا تعمل، فـ (أل) الداخلة عليها للتعريف، وليست بالموصولة التي تدخل على اسم الفاعل ونحوه من المشتقات التي تعمل»[2].

ومما سبق يفهم أن (صاحب)، قد تستخدم مشتقًا ومن ثم تتحمل ضميرًا، وقد تستعمل اسمًا فلا تتحمل ضميرًا، وليس لها عمل في جملتها.

فما تحقق فيه الشروط الثلاثة السابقة فهو متحول تحولا تاما، ومن ثم فلا عمل له، وإن كان له مكمل استغنى عنه؛ فمأذون اسم المفعول عندما تحولت إلى اسم حذف مكملها الجار والمجرور (لـ + كذا)، وكذلك (مستند) عندما تحولت حذف منها مكملها الجار والمجرور (إلى + كذا).

### 2 . التحول غير التام:

إذا فقد شرط من شروط التحول التام أصبح التحول غير تام. فمما تصور معه أنه صفة تتحمل ضميرًا وتطلب عملا : أبطح وأجرع وأبرق وأدهم وأسود وأرقم «إنما منع صرف باب: أبطح، وهو المكان المنبطح من الوادي، وأجرع وهو المكان المستوي، وأبرق وهو المكان الذي فيه لونان، وباب أدهم للقيد وأسود للحية السوداء، وأرقم للحية التي فيها نقط سود وبيض كالرقم مع أنها أسماء لأنها وضعت صفات فلم يلتفت إلى ما طرأ لها من الاسمية»[3].

## ثالثاً – أسباب ظاهرة التطور الدلالي للمشتقات

### 1– كثرة الاستعمال:

[1] شرح التصريح : (92/1).

[2] النحو الوافي : 1 : (209، 210) هامش : 4

[3] شرح التصريح (213/2)، شرح الأشموني: (236/3).

للاستعمال كثرة وقلة أثر على اللغة بالغ «وما كثر دوره في الكلام كثر فيه الحذف والتغيير».[1] وورد عن عدم عمل كلمة (صاحب): «غلبت عليها الاسمية؛ أي بسبب كثرة استعمالها في الذات بقطع النظر عن الصفة،» [2].

## 2- تطور اللغة:

إن تطور اللغة مستمر ما بقيت حياة، ولا يمكن دفعه، ولكن تعرف آثاره ويقعد لها «الكلمة في تطورها لا تقف في دلالتها عند حدود مصدرها الأصلي، بل قد تتعداه إلى أمر لا صلة له بذلك المصدر وإلى معنى جديد لا يكاد يمت إلى الدلالة الأصلية بصلة وثيقة» [3] وذلك لأن « اللغة كائن حي لأنها تحيا على ألسنة المتكلمين بها، وهم من الأحياء، وهي لذلك تتطور وتتغير بفعل الزمن» [4] ومن ثم «فإذا سمى بصفة رجل، نحو: أحمد وأسعد صار اسمًا جامدًا وجمع الأسماء، نحو: أحامد وأساعد ... لأنه بالتسمية زال معنى الوصف عنه، ولم يبق يفيد من المعنى ما كان يفيد قبل التسمية» [5]

## 3- الترجمة:

ساعدت الترجمة عن اللغات الأخرى في وضوح ظاهرة التحول الدلالي للمشتقات؛ إذ إنها تتطلب مصطلحات لنقل المستحدثات العلمية، وليس أوسع من مجال المشتقات لمواجهة ذلك، ومنه لفظة (الحاسب) ولفظة (الحاسوب) إذ تحولت كل منهما دلاليا من المشتق إلى الاسم. ومنه كذلك لفظة (ألسنية) حيث اشتقت وتحولت اسما للتعبير عن هذا المفهوم الجديد على العربية: علم اللغة. [6]

[1] أمالي السهيلي: ص: (55)، ط2، 1970م.

[2] حاشية الصبان: (164/1)، شرح التصريح : (142/1).

[3] دلالة الألفاظ د.أنيس 21.

[4] بحوث ومقالات في اللغة – رمضان عبد التواب: 57، الخانجي، القاهرة، دار الرفاعي بالرياض، ط1، 1982.

[5] شرح المفصل 63/5.

[6] المعجم العربي الأساسي: لسن

## رابعاً – متطلبات المعالجة الحاسوبية للتطور الدلالي في المشتقات

تلزم متطلبات لإمكان المعالجة الحاسوبية للمشتقات المتطورة دلاليًا؛ تعرفًا عليها صرفيا، وتحديدًا لمتعلقاتها تركيبيا، ومن ثَم تحديد دلالتها. وهذه المتطلبات هي:

### 1– حصر المشتقات المتطورة دلاليًا:

حصر المشتقات المتطورة دلاليًا هو خطوة مهمة تساعد على معالجتها حاسوبيًا، ويكون هذا الحصر باتباع عدد من الخطوات، هي:

أ– الرجوع إلى معاجم اللغة، واستقراء مادتها، وتحديد التطور الدلالي الذي حدث للمشتق من خلال الشواهد والنصوص الواردة بها.

ب . تتبع الاستخدام اللغوي، فالاستخدام اللغوي قد يطور بعض الألفاظ دلاليًا، ولكن تتأخر المعاجم عن رصده لسبب ما، ومثال ذلك لفظة (العربية)، فالمعاجم لم تسجل تحولها الدلالي إلى اسم، ويتضح هذا من استقراء كتب اللغة، يقول صاحب (شرح التصريح): «لأنه ليس في العربية اسم معرب آخره واو مضموم ما قبلها»[1]. ففي هذا النص استعمل لفظ "العربية" متحولا دلالياً من الصفة إلى الاسم.

وإن كان الشيخ يس لم يرضه هذا، ونقل عن الدنوشري قوله عن استخدام صاحب (شرح التصريح) للفظ العربية السابق : بل هو «صفة لموصوف محذوف تقديره: في اللغة العربية»[2].

وقد اتضح من هذه النصوص استخدام لفظ (العربية) متحولة عن معنى المشتق إلى الاسمية.

وعن طريق حصر المشتقات المتطورة دلاليًا ودعم الحاسوب بها نكون قد وافيناه بأحد متطلبات الحوسبة أو ما يعتمد عليه الحاسوب في الفهم الآلي للغة أو الحوسبة.

فإن ورد له لفظ مشتق كان أمامه طريقان، إما أن يسير في طريق عدِّ هذه اللفظة مشتقًا، ومن ثم فيجب أن يبحث عن عمل تركيبي لها إذا توافرت شروط العمل. وإما أن يسير في طريق عدِّ هذه اللفظة اسمًا متحولاً عن المشتق، ومن ثم فلا بحث عن عمل تركيبي

[1] شرح التصريح : (188/2، 189).

[2] حاشية يس: : (188/2).

لها، وإنما يكتفي بتحديد موقعها في سياقها دون محاولة إيجاد فواعل لها أو مفعولات أو غيرهما.

ومن نافلة القول أن يقال إن هذا الحصر اللغوي للمشتقات التي تطورت دلاليًا حصر متنامٍ؛ لأن اللغة بطبيعتها متنامية، ومعاني ألفاظها لا تفتأ تزيد أو تنقص «الكلمة في تطورها لا تقف في دلالتها عند حدود مصدرها الأصلي، بل قد تتعداه إلى أمر لا صلة له بذلك المصدر وإلى معنى جديد لا يكاد يمت إلى الدلالة الأصلية بصلة وثيقة»[1].

## 2- تحديد مستوى الاستعمال اللغوي:

تحديد مستوى الاستعمال اللغوي للمشتق الذي تطور دلاليًا يساعد في معالجته حاسوبيًا؛ فأمام الناظر إلى حقل اللغة عدة صور لهذا اللفظ المتطور دلاليًا، وهي:

(أ)إهمال الأصل المتطور عنه، والاكتفاء باللفظ حاملاً دلالته الجديدة. من مثل: (قارب) فاستعمالها اسم فاعل من الفعل (قرب) يكاد يكون منعدمًا، ولكن استعماله بمعنى (الزورق) أو صحفة على هيئة القارب يؤكل عليها هو الشائع[2].

(ب) عدم إهمال الأصل المتطور عنه، وبقاؤه مستعملاً في وجود الدلالة الجديدة المتطور إليها، ومن ثم فيجب ترتيب هذه الدلالات بدءًا بالأشهر للمساعدة في سرعة فك اللبس، وفي هذه الحالة إما أن يسبق الأصل الدلالة الجديدة أو تسبق الدلالة الجديدة الأصل، أو يكونا سواءً في الشيوع.

فمما شاع فيه الأصل عن الدلالة الجديدة: لفظة (قريب) فإن الإطلاق الأول لها هو «قريب بمعنى: الداني في المكان أو الزمان، والإطلاق الثاني هو (قريب بمعنى: الداني في النسب»[3]، وقد شاع الإطلاق الأول الذي هو الأصل.

ومما شاعت فيه الدلالة الجديدة عن الأصل كلمة (اللفظ): فقد تستخدم مصدرا، وقد تستخدم اسماً ، ولعل هذا ما جعل الشيخ الصبان ينبه عليه عند الاستخدام لقسمي الكلمة أو قل للطورين من أطوار الكلمة «قوله: (جعلوه) أي هذا اللفظ بدلاً أو عوضًا من اللفظ، أي من التلفظ بالفعل»[4]. أي يستعمل مصدرا واسما متحولا عن المصدر.

[1] دلالة الألفاظ، 21

[2] راجع : المعجم الوسيط: قر ب.

[3]السابق نفسه.

[4]حاشية الصبان: (181/1) ، (188/3).

ومما استوى فيه الأمران: كلمة (التنوين): فإنها تستعمل اسمًا كما تستعمل مصدرًا « والتنوين وهو في الأصل مصدر نونت؛ أي: أدخلت نونًا، ثم غلب حتى صار اسمًا لنون تلحق الآخر لفظًا لا خطًا لغير توكيد» [1].

وعندما يستوي الإطلاقان يصبح التفضيل بينهما صعبا ومن ثم نلجأ إلى وسيلة أخرى للحكم على الإطلاق المقصود في السياق.

### 3- تحديد مجال الاستعمال:

مما يفيد في المعالجة الحاسوبية للعربية تحديد مجال استعمال اللفظ، وذلك إذا كان له من خصوصية الاستعمال ما يسمح بقصره على مجال دون غيره.

ومثال ما يعدّ استعمالا قرآنيا: ديَّار، وهو صيغة مبالغة تحولت إلى اسم، فالديّار: الديراني، وهو صاحب الدَّير الذي يعمره[2].

ومما يعدّ استعمالاً حديثيا: حليلة: حليلة الرجل زوجته[2]، فإن البحث لم يهتد لاستعمال هذا اللفظ مفردا مؤنثا في قرآن ولا شعر ولا نثر ، ومن ثم فهو استعمال حديثي.

على أنه ينبغي ألا يبالغ في قصر استعمال لفظ على مجال بعينه؛ وذلك لأن العربية تعتمد المجاز، وقد يحسن كاتب يومًا استعمال كلمة من مجال في غير مجالها ويستسيغ ذلك القراء – فتشبع، وتتجاوز مجالها.

### 4- تحديد القيود الدلالية :

مما يلزم لتمام المعالجة الحاسوبية للمشتقات التي تطورت معانيها تحديد قيود الاستعمال؛ وذلك عن طريق النصِّ مع كل لفظ، بطريقة أو بأخرى، على معلومات صرفية وتركيبية ودلالية.

فصرفيًا ينص على العدد والنوع، فإذا كان اللفظ مما يقتصر على حالة دون أخرى من حيث العدد: إفرادًا وتثنية وجمعًا أو من حيث النوع: تذكيرًا وتأنيثًا ينص على هذا، فإنه يفيد في المعالجة الحاسوبية، ويقلل من احتمالات اللبس.

وتركيبيًا ينص على سلوك اللفظ التركيبي، فيوضح إذا كان يلزم حالة إعرابية خاصة أو لا (الموقعية)، وإذا كان يلزم رتبة دون غيرها أو لا.

[1] شرح الأشموني: (30/1).

[2] المعجم الوسيط. دير.

[2] المعجم الوسيط. دير.

ودلاليًا: ينص على ما يقلل احتمالات اللبس، كأن يكون يقترن بلفظةٍ ما، أو يلزم مجالاً دلاليًا محددًا.

فالقيود الدلالية التي توضع على كل مشتق تحدد الصورة اللفظية التي يستعمل فيها ، وتحدد دلالته:

هل يستعمل : مفردا ومثنى وجمعا أو يقتصر من ذلك على شيء دون غيره؟

وهل يستعمل : معرفا بـ(أل) أو بالإضافة أو يستعمل نكرة؟

وهل يستعمل : مذكرا أو مؤنثا أو يقتصر على نوع دون الآخر؟

فالمتحول دلاليا قد يستعمل منه المفرد والمثنى والجمع ، وقد يقتصر من ذلك على بعضه، فمما يستخدم مفردا ومثنى وجمعا :تصنيف وتصنيفان وتصانيف ، وتركيب وتركيبان وتراكيب، وتجربة وتجربتان وتجارب، ومما يستعمل مفرداً فحسب إعلام ومفاد وحرمة وغيرها، ومما يستعمل متحولا جمعَ تكسير فحسب (رقائق) و(مسامع) و(أقاويل) و(سفاسف)، ومما يجمع جمعَ سلامة (مخابرات) و (محفوظات) و(لزوميات).

المتحول قد يستعمل مذكرا ومؤنثا، وقد يقتصر من ذلك على بعضه؛ فمما استعمل مذكرا ومؤنثا (حادث) ومؤنثه (حادثة) ، ومما استعمل مذكراً فحسب (الحاسب)، ومما استعمل مؤنثا فقط (منزلة) و (بادرة) و(ضائقة) وغيرها.

قد يلزم المتحول دلاليا الاقتران بسابقةٍ ما ، ولا يعد متحولا دونها ، ومن ذلك (اليابسة) و(المعمورة) ، فكي تعد متحولة يجب اقترانها بـ(أل) ، وإذا خلت منها فهي صفة غير متحولة. إن الاقتران بـ(أل) ليس قاطعا بالتحول، ولكن إذا وجد احتملت اللفظة أن تكون متحولة أو أن تكون صفة، لكن إذا عريت من (أل) وجب كونها صفة.

# المراجع والمصادر

- **جمال عبد الناصر** التعدد الوظيفي للصيغة الصرفية في القرآن الكريم، مخطوطة، كلية دار العلوم، جامعة القاهرة.
- **داود عبده** الإطار النظري للمعالجة الآلية من الإنكليزية إلى العربية، ومن العربية إلى الإنجليزية، المؤتمر الثاني حول اللغات الحاسوبية، الكويت، تشرين الثاني: 1985م.
- **رمضان عبد التواب** بحوث ومقالات في اللغة، الخانجي، القاهرة، ودار الرفاعي بالرياض، ط1، 1982.
- **شعبان صلاح** الجملة الوصفية في النحو العربي، دار غريب للطباعة والنشر والتوزيع،2004م.
- **الصبان، محمد بن علي** حاشية الصبان على (شرح الأشموني)، دار إحياء الكتب العربية.
- **علاء على إسماعيل** التعبير الاصطلاحي في الأمثال العربية: دراسة تركيبية دلالية، رسالة دكتوراه ، مخطوطة، كلية دار العلوم ، جامعة المنيا
- **على أبو المكارم** الحذف والتقدير ، مخطوطة، مكتبة كلية دار العلوم، جامعة القاهرة
- **فاضل إبراهيم السامرائي** الجملة العربية والمعنى، دار ابن حزم ، بيروت ، لبنان ، ط1، 1421 هـ / 2000 م
- **قاسم المقداد** هندسة المعنى، دار السؤال، دمشق، 1984م.
- **مدحت يوسف السبع** العلاقات التركيبية في الجملة الفعلية القرآنية: دراسة نحوية حاسوبية، مخطوط بكلية دار العلوم، جامعة القاهرة.
- **مروان البواب وآخرون** إحصاء الأفعال العربية في المعجم الحاسوبي، مكتبة لبنان، 1996م.
- **مصطفى حميدة** نظام الارتباط والربط في تركيب الجملة العربية، الشركة المصرية العالمية للنشر: لونجمان، بدون تاريخ.
- **المنظمة العربية للتربية والثقافة والعلوم** استخدام اللغة العربية في المعلوماتية، تونس ، 1996 م

- **نبيل علي** اللغة العربية والحاسوب، عالم المعرفة، تعريب، 1988م
- **نهاد الموسى** العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية، المؤسسة العربية للدراسات والنشر، ط1 ، 2000م
- **يس بن زين الدين العليمى** حاشية يس على (شرح التصريح)، مطبعة الحلبى، القاهرة
- **وفاء كامل فايد** المتطلبات اللغوية لمعالجة التعابير الاصطلاحية العربية معالجة آلية، الندوة الدولية الأولى عن الحاسب واللغة العربية، الرياض، 1428 هـ.
- **ابن يعيش، موفق الدين يعيش بن على** شرح المفصل، مكتبة المتنبى، القاهرة

# Arabic Information Retrieval: How to Get “Good” Results at a Lower Cost?

Claude Audebert*[1], André Jaccarini [2], Christian Gaubert**[3]

*Maison méditerranéenne des sciences de l’homme (MMSH)*
*5 rue du Château de l'Horloge BP 647 13094 Aix-en-Provence, France*
[1]claudeaudebert@yahoo.fr
[2]jaccarini@mmsh.univ-aix.fr

** *Institut français d’archéologie orientale du Caire (IFAO),*
*37 el Cheikh Aly Yousef Str., Cairo, Egypt*
[3]cgaubert@ifao.egnet.net

*Abstract*—**The method we apply in order to retrieve information at the lowest cost in order to help sort and characterize texts is described. Its four characteristics: non-dictionary based, restricted to surface structure, based on the used to tokens i.e. tools words represented by automata. Arabic texts are examined in regard to token density and categories, which have an impact on their nature. The results of our experiments are presented.**

## 1 Introduction

This question sounds like a commercial add and seems, at first sight, inappropriate in a scholarly assembly. But we will nevertheless consider it since the research we have been presenting so far does not belong in the mainstream, for many reasons. The first reason is that our work is based on the concept that, in order to reach a good linguistic resolution level (say 80%) one does not need to feed the machine a huge amount of rules. In other words with a number of limited rules, one can get a great deal of information. One could even go further and say that beyond a certain level, adding extra rules would not increase significantly the quality of results.

## 2 Principles

But lower cost does not mean cheap cost: as matter of fact, we have based our research on morpho-syntaxic analysis of Arabic, in other words, a linguistic approach, rather than other approaches such as the statistical approach, treebanks,...
We will characterize our approach in four main points:

- 1. It is not dictionary-based
- 2. It is based on the surface structure, made possible by:
- 3. The use of what we have called « tokens »
- 4. Which can very well be described by automata.

Some explanations are required.

Any Arabic text (and for that matter, it holds true for other languages) appears as formed by a group of sentences. These sentences are structured by words that have been spotted by grammarians as having a particular influence on the sentence structure. They can be prepositions, conjunctions whose function is to coordinate or subordinate clauses and so on. Arab grammarians have called them: *hurūf*, *asma’*, *ẓurūf* and so on, and classified them under: آدوات الربط   حروف المعاني

Such as *akhawāt Inna*, *akhawāt kana, ḥurūf al-garr, ...*

These particles induce expectations of various kinds and levels.

For instance ’*inna* introduces a nominal sentence, which means that expectations are high and on the level of the sentence. In turn this means that after that token we expect a *mubtada’* and a *khabar*. This also leads to the fact that after *inna* we are bound to find a (definite) noun. Just as after *lam* we will necessarily find a verb. So expectations are global or local.

- These particles or *token* are in *limited number* three to four hundreds.

- They do not - for the most part- obey to the root derivation that applies to lexical items.

- They are thus very informative.

- Just as they induce expectations on the syntactic level, they induce expectations on the semantic level. We have shown this fact in our paper delivered at ALTIC 2011 [1] as far as text characterization is concerned.

We precisely intend to show through the use of tokens, that a very restricted group can bring a very high degree of information and answer the question underlying the present paper :*I.R.*: *a good level at a lower cost*. That is: can we restrict our use to tokens in order to get a reasonably acceptable text characterization without the intervention of other elements?

Needless to say that this choice represents an extreme case for a first step and is perfectly compatible with adding other elements (other criteria such as articles, roots, repetitions and other we mentioned in [1]).

They can be represented by automata that, on the other hand, turned out to be very fit for the description of the Arabic language, in view of its high grammaticality and algorithmicity, directly perceptible on the surface structure.

## 3 WHY AUTOMATA?

Tokens are syntactic structures indicators. These structures are of different levels. Tokens may reveal complete sentence trees (syntagmatic indicators) or partial ones. They also may simply induce binds on the nearest neighborhood. Anyway, if we consider the informational flow of the text, they may be viewed as triggering expectations. This is the very reason why we have made the choice to represent them by automata.

Automata are at the core of the theory of informatics. They are abstract machines constituted by a reading head that moves on along the text, plus a control unit that can only be in a finite number of states, in the simplest cases. While reading each symbol, the control unit moves from one state to another – sometimes itself – in case of a loop. If, after having read the last symbol, the machine finally gets to a state belonging to a subset of states allowing a possibility of exit, then the sentence is accepted.

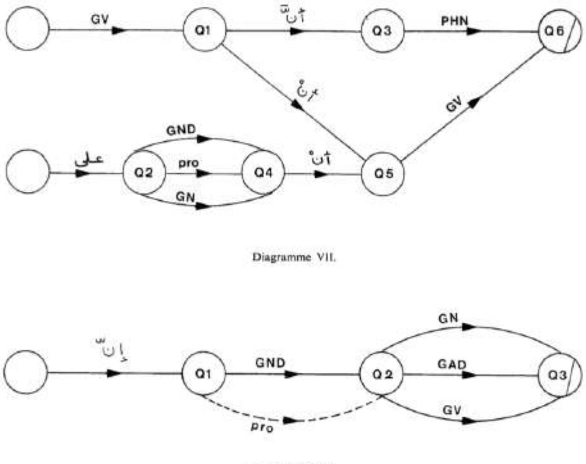


Diagramme VII.

Diagramme VIII.

**Figure 1: High level Automata**

The above examples represent some high-level automata [2]. Arcs are tagged by the symbols representing the constituting elements that are fundamental and cannot be suppressed: Nominal defined group (GND) and GN (nominal group) to which are attached grammatical functions (*mubtada'* and *khabar*).

In our top-down approach we consider these symbols as calling sub-automata until we get to symbols belonging to the terminal vocabulary that will be Arabic characters, if we choose to work without a dictionary, or lexical items, if we choose to use a dictionary.

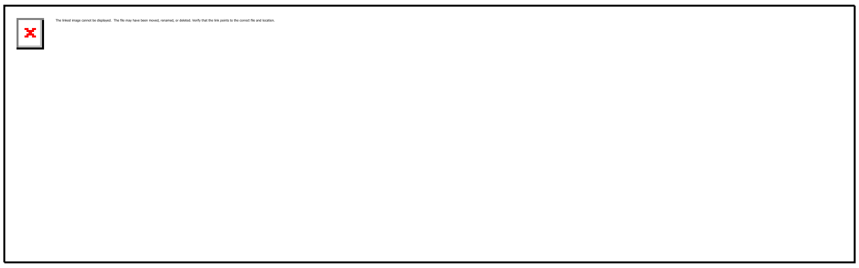
The possibility offered to operate without a dictionary, since there exists an internal grammar of the word, is a strong specificity of the linguistic system of Arabic. And it is useful to make it explicit not only to be able to show the structure of this language, but also in order to benefit from it on the algorithmic level.

Moreover, automata offer the theoretical specificity of being easily set into a hierarchy. Besides, programming automata can be very rigorous if their construction obeys to the mathematical bases of the theory. This hierarchy of such programs that can mathematically be specified can in turn lead to a hierarchy of *tokens*, considered as operators.

This formalization offers the following benefits:

- Automata are relatively simple programs that can mathematically be specified as we previously mentioned. They can imbricate and call one another. They can lead to modular analyzers whose complexity can be controlled.

They constitute a universal model of calculus, and are particularly fit for the structure of the Arabic language. For instance, the possibility of expressing by one non-deterministic automaton of six states, the morphological structure of Arabic, as described by D. Cohen in his pioneering study [3] is in itself remarkable. It reveals both the specific structure of Arabic morphology and the descriptive ability of the theory of automata.



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

**Figure 2: Arabic morphology in a 6 states automaton**

These automata which will be augmented i.e. to which will be attached tests and programs at each transition, can also be used to describe other operators in the field of I. R. as we have shown in our paper delivered at ALTIC 2011. In this paper, we described the choice of criteria for text characterization such as quotations' retrieval, marks of temporality, and so on and their description by automata in some cases.

Another automaton, that of quotations (one of the devices we have designed in order to extract quotations and reported discourse in an Arabic text) has 31 states; it operates on characters with no need of a dictionary, minimizing memory use [4] [5].

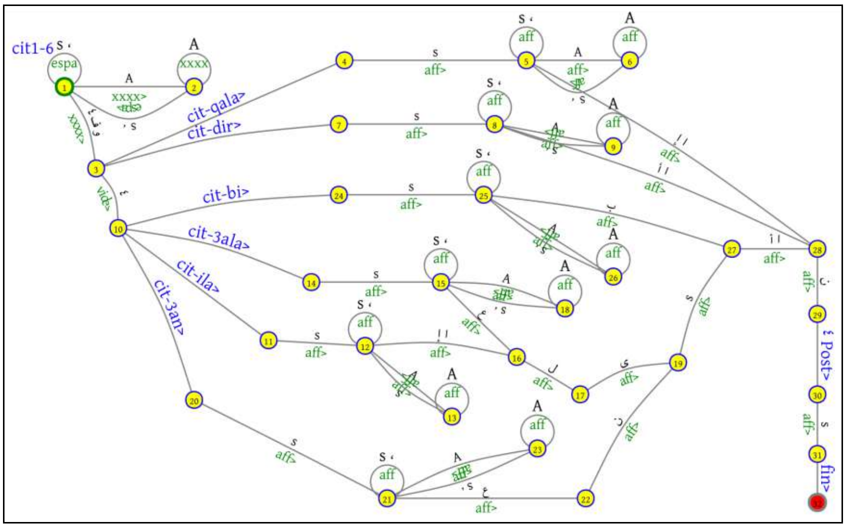


**Figure 3: Arabic quotations retriever**

## 4 EXPERIMENTS WITH THE TOKEN

Experimenting is also part of our approach. We can do so thanks to the software *Kawâkib* we developed and whose functions we described in [1].

A. Is token density meaningful for I.R.?

In his Phd, C. Gaubert [6] has shown that Arabic texts contained various amounts of tokens. *Al-ayyām* for instance (over 40%) as opposed to Arabic newspapers (mostly between 20% and 32%).

The global density of token seems very significant to determine the nature of texts. Not only their period. On the whole one could say that classical Arabic prose texts contain more token than newspaper Arabic. This is verified in the analysis of the results displayed in Table 1.

| TEXT Nature | TEXT | TOK % | TEMP |
|---|---|---|---|
| Philosophical | فصل في أقسام المعلومات | 48,82 | 5,49 |
| Fable | كليلة ودمنة | 47,22 | 7,51 |
| Novel | الأيام | 43,62 | 17,7 |
| Press | الفلسطينيون.. في صميم | 41,18 | 10,3 |
| Qur'an | سورة الكهف | 40,45 | 6,85 |
| Historical study | وجوب القتال عند المفسرين | 37,97 | 6,87 |

| | | | |
|---|---|---|---|
| Adab - portrait | أخلاق الوزيرين | 37,2 | 8,9 |
| Historical study | الزلازل | 36,3 | 9,1 |
| Novel | بداية ونهاية | 36,3 | 18,8 |
| Historical study | مذكرات ونستون تشرشل | 34,94 | 18,2 |
| Novel | قنديل أم هاشم | 34,88 | 9,65 |
| Press | قلب نظام الحكم | 34,5 | 3,8 |
| Novel | عودة الروح | 33,09 | 11,9 |
| Press | القمة العربية | 32,6 | 7,58 |
| Press | الأهرام 20 م | 29,92 | 6,13 |
| Law | قوانين الصحافة في مصر | 28,84 | 8,54 |
| Press | العد التنازلى لمفاوضات السلام | 27,98 | 3,33 |
| Press | الشعب يريد تطهير البلاد | 19,59 | 3,13 |

**TABLE 1 : GLOBAL TOKEN AND TEMPORALITY PROPORTIONS**

This table goes to show that the density of tokens can be found both in classical and modern texts. One of the highest percentages is found in a philosophical text (*Faṣl*) and in a modern newspaper text (*al-Filastiniyun*). How is that so? We suggest an answer in B.

B. The nature of tokens also is of the utmost importance to characterize a text. Certain tokens reveal, more than others, the nature of the structure of a text. For instance, those texts where there is a debate, in which a subject is discussed on its different aspects is bound to contain reasoning, argumentation where tokens governing subordinate clauses are more likely to appear.

In this respect, negation appears as a very discriminative set of tokens. In our corpus *FaSl* not only displays 46,82% of global tokens, but also 14,04% of negations and 13,19% of relative pronouns from which we can deduce relative clauses which can be interpreted as a part of reasoning. The specification a relative clause brings in while opposing entities is very remarkable. Texts that recount events are likely to display tokens in relation with time and/or aspect, i.e. what we call *temporality*.

This newspaper text happens to be also an argumentative and very polemic text, accounting for its high percentage of tokens. A study of their semantic nature would be of the highest interest to give tokens a value attached to their semantic and to the notions they are hereby linked to [4].

It does not come as a surprise to find low percentages of temporal tokens in some of the first texts either because they are on the whole dealing with a subject in a series of logical steps as opposed to narrative texts recounting events such as chronicles, novels or short stories, not to mention other genres.

At first sight it is surprising to find in a fable telling stories such as *Kalila wa Dimna* such a low rate of temporal tokens. But it shows in fact the level of the political and wisdom developments in this masterpiece.

We cannot express ourselves without the proper tools to do so. It is important to pay attention to the means offered by a language to allow expression.

Although we do not develop from the start, an *a priori* theory of texts characterization, it is nevertheless indispensable to ask these questions in order to be able to set percentages and then when needed, in a further step, ask the text the proper ones towards an interpretation of the phenomenon.

In the meantime, it is obvious that our approach to token analysis in regard to text specification must be homogenized in order not to misinterpret figures and facts.

A scale has to be established between the various categories of tokens in order to be able to measure measurable objects. For instance what is the average occurrence of prepositions since no text is devoid of this category?

In an attempt of the use of these ratios to classify texts, we show in Figure 4 the relative positioning of some texts in regard to four different criteria: global percentages of tokens (TOK), of temporal tokens (TMP), of consecutive determination marks (DET) and indication of the root dispersal and richness (ROOT). The two last criteria have been discussed in [1]. All the values have been scaled from 0 to 100 in order to shown trends.

One can observe that the philosophical text presents a high level of tokens combined with low temporality and the use of a reduced set of roots. It can be opposed to novels which use many more roots and temporality marks; studies show an intermediate case, but some of them (*al-zawâj*) can use few roots and many temporal tokens among very few tokens. A completely opposed case is that of a press article very poor in roots, tokens and temporality but using a lot of nominal determined groups.

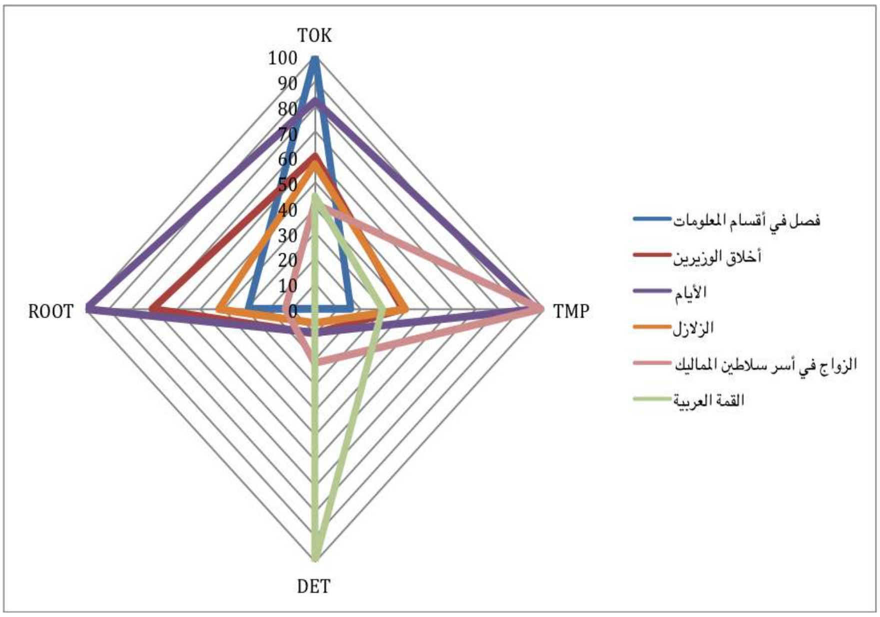


**Figure 4: Some texts in regards to 4 criteria**

## 5 CONCLUSION

This deliberate cursory use of tokens as developed in this paper show how far we can go in I.R. using only tokens represented by automata, that is a very restricted amount of objects that organize the language and thus are utterly informative at various degrees. Even a preposition is more informative about the structure than any lexical item.

The cost is that of a previous linguistic research to detect tokens and then a linguistic study of the semantic expectations they induce and of the notions to which they are attached.

We have shown some figures and statistics to illustrate the main ideas we have exposed in this paper. Fore more extended results the reader can consult our web site http://automatesarabes.net.

## REFERENCES

[1] Audebert, Gaubert, Jaccarini, « A *Flexible* Software Geared Towards Arabic Texts I.R And Evaluation : Kawâkib », ALTIC 2011, (Alexandria, Egypt), à paraître dans ALTEC 2011 (to be published at http://www.altec-center.org/conference/), 2011.

[2] Audebert C, Jaccarini A., À la recherche du *Ḫabar,* outils en vue de l'établissement d'un programme d'enseignement assisté par ordinateur, *Annales islamologiques* 22, Institut français *d'archéologie* orientale du Caire. 1986.

[3] Cohen, D., Essai d'une analyse automatique de l'arabe. In: *David* Cohen. E*tudes de linguistique sémitique et arabe*. Paris:Mouton, pp. 49-78. 1970.

[4] Audebert, Gaubert, Jaccarini, Minimal Ressources for Arabic Parsing/ an Interactive Method for the Construction of Evolutive Automata. *MEDAR* 09, (http://www.elda.org/medar-conference/summaries/37.html) , 2009.

[5] Audebert, Gaubert, Jaccarini, «Linguistique arabe. Programme de traitement par automates de la langue arabe (Tala)» ISBN 978-2-7247-0561-4, *Annales islamologiques* 44, Institut français *d'archéologie* orientale du Caire. 2010.

[6] Gaubert Chr., *Stratégies et règles pour un traitement automatique minimal de l'arabe*. PhD. Département d'arabe, Université d'Aix-en- Provence. 2001.

# An Empirical Analysis of Clustering Techniques for Arabic Documents

Doaa Farag Ibrahim F. Imam

Department of Computer Science
College of Computing and Information Technology
Arab Academy for Science, Technology and Maritime Transport
Masaken Sheraton, Cairo, Egypt
doaafarag@medstar-misr.com, ifi05@yahoo.com, imam@vt.edu

***Abstract-*** Documents clustering is the process of grouping electronic text documents into groups of similar documents. This paper presents an empirical analysis of utilizing different clustering techniques and algorithms with different similarity criteria. For all experiments the entropy measure and f-measure are used to calculate the clusters' quality. The experiments are performed on 300 Arabic news documents representing five classes. The results show that using different similarity functions does not affect much the results except for the Euclidean criterion which gives the worst result. While for the clustering techniques; the hybrid technique has the best results in most of the experiments. The single link agglomerative algorithm shows better results than the other two agglomerative algorithms complete link and UPGMA.

## 1. Introduction

Document clustering approaches can be classified into hierarchical and partitional. Hierarchical clustering is divided into agglomerative (bottom-up) and divisive (top-down) algorithms. While partitional clustering includes many algorithms such as k-means, k-mediods, probabilistic, relocation and density-based algorithms [2]. Hierarchical algorithms produce clusters gradually, partitional algorithms produce clusters directly. In other words; hierarchical algorithms builds a tree of clusters known as a dendrogram. Each node (cluster) in the tree is the union of its children (sub-clusters) and the root of the tree is the cluster containing all the documents. Most of the times, but not always, each leaf node is the cluster containing only one document. An agglomerative clustering begins with one cluster for each document and recursively merges the two (or more) most similar clusters. A divisive clustering begins with one cluster of all documents and recursively splits the most suitable cluster. The process continues until a stopping criterion (frequently, the required number *k* of clusters) is achieved. The observable disadvantage of hierarchical algorithms is that they do not revisit the constructed clusters for the purpose of improvement. K-means [7] is the most famous partitional algorithm and thus been investigated by many researchers. K-means algorithm begins with choosing initial *k* number of centroids and iteratively assigns each document in the corpus to the closest centroid and recomputes the centroids. The process continues until a stopping criterion (frequently, centroids don’t change or for *l* number of iterations.) is achieved. Although k-means algorithm is simple and straightforward, its results mainly depend on the initial centroids randomly chosen in the start of the algorithm. Many refinements can be done to k-means algorithm to improve its quality [3] [17]. Hierarchical and partitional approaches can be combined to form another approach of clustering which is called hybrid clustering .The Scatter/Gather system [5] uses hierarchical clustering to produce centroids for a final k-means phase.

This paper presents empirical analyses of different clustering algorithms for Arabic text documents. Arabic text documents were analyzed using the RDI morphological analyzer [1]. Stop words and verbs were removed. For each noun, the morphological analyzer retrieved its stem (denoted as term). Terms were weighted using the term frequency (*tf*) and inverse document frequency (*idf*). Three agglomerative algorithms (single link, complete link and UPGMA) and two k-means algorithms (traditional k-means, incremental k-means) were implemented in this research. In addition, the hybrid algorithm which combines both agglomerative and k-means was also implemented. The experiments were performed on 300 Arabic news documents. These news documents were extracted from Google Arabic news group. These documents represent news from five classes (arts, economy, politics, sports, and science – 60 documents per class). The documents were partitioned randomly four times with different number of documents in each partition. The experiments were repeated with different number of required clusters (*k*) and different similarity criteria. The results show that single link agglomerative algorithm is better than the other two agglomerative algorithms. The incremental k-means produces better quality than the traditional k-means. The hybrid technique mostly has the best results but its computational complexity is the union of that of both the agglomerative and the k-means.

## 2. Documents Clustering

### 2.1. Document Representation

Documents cannot be directly interpreted by a clustering technique. As a result, document representation phase is needed to map the document into another interpretable form. For term definition; the most frequent choice is to define terms either with the words (Bag Of Words approach) occurring in the documents after eliminating stop words and prepositions in preprocessing phase or with their stems [15].

A document is usually represented by a vector, called term weight vector [14]. The term weight vector correspond to the terms exist in all training documents. The contents of the term weight vector equal to the weighted value which that term has for the document.

Term weight is famous with its equation term frequency multiplied by the inverse document frequency (tf*idf).This function denotes that the more often a term occurs in a document, the more it is representative of that document and the more documents the term occurs in, the less representative it is.

### 2.2 Similarity Criteria

A similarity criterion must be stated and measured before applying a clustering algorithm. This similarity criterion computes the degree of similarity between two term vectors. There are many examples of similarity criteria including cosine, Jaccard, Dice, Euclidean distance [11]. But the most popular one is the cosine criterion [14] which measures the cosine angle between two vectors. These criteria calculate similarity based on the co-occurrences of a term in the documents.

### 2.4. Clustering Techniques

Clustering techniques is mainly classified into hierarchical and partitional clustering. Hierarchical clustering produces nested clusters while partitional clustering produces un-nested ones [9]. Agglomerative hierarchical and k-means partitional approaches are two clustering techniques that are commonly used for document clustering. Sometimes K-means and agglomerative hierarchical approaches are combined to produce a hybrid technique which has the best from both sides. There are many agglomerative algorithms which always vary in how to compute the similarity between two clusters. Examples of these algorithms are single link, complete link, UPGMA, centroid similarity, Intra-cluster similarity [12] [13] [17]. The k-means algorithm [7] is the most famous clustering tool used in many applications. Its name comes from representing each of *k* clusters by the mean of its data points also called centroid. Many variants of the k-means have been investigated as to improve its clustering quality. A procedure for computing a refined starting condition from a given initial one that is based on expectation maximization (EM) is presented in [3]. Another variant is the bisecting k-means [17] which starts with one cluster containing all documents and iteratively split a cluster into two sub clusters using basic k-means algorithm. Bisecting k-means show better results than agglomerative algorithms. Incremental K-means is another variant of the basic k-means. In the basic k-means recomputing of the centroids are taken place after assigning all documents to the closest centroid, while for the incremental k-means recomputing of the centroids are repeated after each document is assigned to the closest centroid. The partitional clustering algorithms are more suitable for clustering large documents datasets due to their relatively low computational time. In [10] [18] partitional algorithms actually produce inferior and less effective clustering quality than the agglomerative algorithms. All of these studies did not examine the effect of the criterion functions. Criterion functions optimize the entire clustering process for both approaches. A recent study reported by Zhao and Kapyris [19][20] investigated the effect of the criterion functions to the problem of partitional and hierarchical documents clustering and the results showed that different criterion functions lead to significantly different results.

Hybrid technique is not investigated as much as the hierarchical and partitional approaches. May be this was a result of its complexity and longer computational time than other techniques. But, the Scatter/Gather system [5] implements a hybrid technique that uses hierarchical clustering to produce centroids for a final K-means phase.

### 2.5 Measures of Cluster Quality

There are two approaches for measuring clustering quality which are the internal and the external quality measures. The internal approach measures the quality of clustering without the aid of any external knowledge. It uses the consistency of clusters to measure the clustering quality. An example of the internal approach is the overall similarity measure used in [17]. The external approach measures the clustering quality by comparing the clusters produced from the clustering technique to known classes. The two famous measures of this approach are the entropy measure [16] and the f-measure [10] which are applied in this paper.

## 3. The Proposed System

Each document in the dataset is mapped to a new representation to be interpretable by the different clustering techniques. Each document passes through two steps: text manipulation and weight matrix creation. Figure 1 shows the steps of the proposed system.

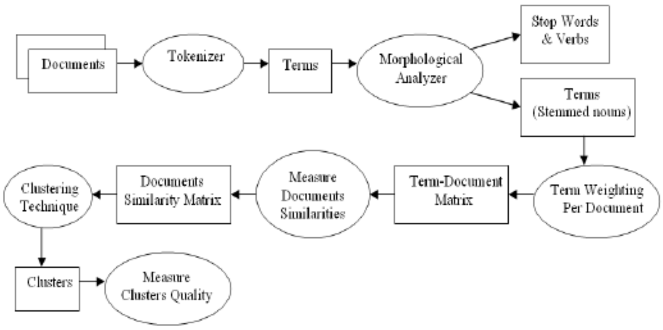


Figure 1: Clustering System.

### 3.1. Text Manipulation

The RDI morphological analyzer is used to analyze each document [1]. First, RDI tokenizes each document into a set of terms (words). Then each term is analyzed and information about this term is produced including prefix, suffix, stem, root, and type. For each document the following three steps are performed:

1- Removing stop words and verbs.

2- Term Stemming: Is the process of removing the prefix and suffix of the term. In Arabic language a word stem may be more representative than the root. Stem reserves the meaning and the type of the word. For some Arabic words, root can be used to produce verbs and nouns with different meanings.

3- Removing diacritics: In Arabic language diacritics are marks (ُّ,َ,ْ ,ٍ,ِ,ٌّ,ِ ِ) above or below letters used in orthography. Most people rely on their knowledge of the Arabic language and the context while writing the Arabic text. As a result the Arabic texts may be fully, partially or even free from diacritics which may make the written Arabic words ambiguous. Thus, removing diacritics is a way to uniform Arabic text.

### 3.2. Term Weighting

Term Weighting is the process of calculating weights for the terms with respective to each document in the training set. These weights are used to build the term-document matrix that is interpreted by the different clustering techniques. Term weighting is applied using the equation of term frequency multiplied by the inverse document frequency. This function denotes that the more often a term occurs in a document, the more it is representative of that document and the more documents the term occurs in, the less representative it is**.**

$$tfidf\,(t_i,d_k) \;\;=\;\; \#(t_i,d_k)\,.\log\frac{|\,Tr\,|}{\#_{Tr}\,(t_i)} \qquad (1)$$

where *#(t*$_i$ *, d*$_k$*)* is the number of times term *t*$_i$ occurs in document *d*$_k$, |*Tr*| is the number of the document in the training set and *#*$_{Tr}$ *(t*$_i$*)*is the number of documents in *Tr* in which *t*$_i$ occurs at least once .

A normalization function is then applied to the weighted value in order to make weights fall in the range [0, 1] and to represent documents by vectors of equal length.

$$w_{ik} \;\;=\;\; \frac{tfidf\,(t_i,d_k)}{\sqrt{\sum_{s=1}^{|T|}(tfidf\,(t_s,d_k))^2}} \qquad (2)$$

where |*T*| is the number of all terms that occur at least once in *Tr*.

### 3.3. Similarity Criteria

A similarity criterion is calculated to each pair of documents in the corpus. The output from this step is the document-document similarity matrix. This step is essential before applying any clustering algorithm. In this paper four similarity criteria are used Cosine, Jaccard, Dice and Euclidean distance criteria.

#### 3.3.1. Cosine Criterion

Considering the *tfidf* in equation 1 and equation 2, measuring similarity between documents j and k using Cosine criterion [14] is calculated as follows

$$Cos(D_j,D_k)=\frac{\sum_{i=1}^{n}w_{ij}\times w_{ik}}{\sqrt{\sum_{i=1}^{n}w_{ij}^2}\sqrt{\sum_{i=1}^{n}w_{ik}^2}} \qquad (3)$$

The value of equation 3 ranges from 0 which means dissimilar (orthogonal vectors) to 1 which means similar (identical vectors).

#### 3.3.2. Jaccard Criterion

Considering the *tfidf* in equation 1 and equation 2, measuring similarity between documents j and k using Jaccard criterion [8] is calculated as follows

$$Jac(D_j,D_k)=\frac{\sum_{i=1}^{n}w_{i,j}\times w_{i,k}}{\sum_{i=1}^{n}w_{i,j}^2+\sum_{i=1}^{n}w_{i,k}^2-\sum_{i=1}^{n}w_{i,j}\times w_{i,k}} \qquad (4)$$

The value of equation 4 ranges from 0 which means dissimilar (orthogonal vectors) to 1 which means similar (identical vectors).

#### 3.3.3. Dice Criterion

Considering the *tfidf* in equation 1 and equation 2, measuring similarity between documents j and k using Dice criterion [6] is calculated as follows

$$Dice(D_j,D_k)=\frac{2\sum_{i=1}^{n}w_{i,j}\times w_{i,k}}{\sum_{i=1}^{n}w_{i,j}^2+\sum_{i=1}^{n}w_{i,k}^2} \qquad (5)$$

The value of equation 5 ranges from 0 which means dissimilar (orthogonal vectors) to 1 which means similar (identical vectors).

#### 3.3.4. Euclidean Distance

Considering the *tfidf* in equation 1 and equation 2, measuring similarity between documents j and k using Euclidean distance [4] is calculated as follows

$$Euc(D_j,D_k)=\sqrt{\sum_{i=1}^{n}(w_{i,j}-w_{i,k})^2\,/\,n} \qquad (6)$$

The value of equation 6 ranges from 0 which means similar documents to 1 which means dissimilar documents.

### 3.4. Clustering Techniques

Three clustering techniques are applied with different similarity criteria, Agglomerative hierarchical technique, Basic K-means partitional technique and Agglomerative K-means hybrid technique. A variant of K-means technique is also applied which is called incremental K-means technique. Figure 2 shows the different techniques and algorithms applied in this paper.

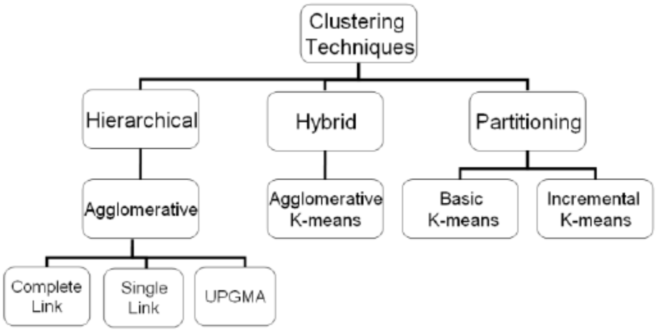


Figure 2: Clustering Techniques.

#### 3.4.1. Agglomerative Clustering Technique

A simple Agglomerative Clustering Algorithm is applied as follows

1. Compute the similarity between all pairs of clusters and build a similarity matrix

2. Merge the closest two clusters (with the highest similarity).

3. Update the similarity matrix to reflect the similarity between the new cluster and the other clusters.

4. Repeat steps 2 and 3 until stopping criteria is reached (here it is the requested number *k* of clusters).

Three different agglomerative hierarchical algorithms are applied; complete link, single link and UPGMA. The main difference between these algorithms is how they define the similarity between clusters (step 1).

**Complete Link** defines the similarity between any two clusters as the maximum distance between any two documents in the two different clusters. Similarity between two clusters *c*$_i$ and *c*$_j$ is defined as follows

$$sim(c_i,c_j)=\max_{x\in c_i\;y\in c_j} sim(x,y) \qquad (7)$$

**Single Link** defines the similarity between any two clusters as the minimum distance between any two documents in the two different clusters. Similarity between two clusters *c*$_i$ and *c*$_j$ is defined as follows

$$sim(c_i,c_j)=\min_{x\in c_i\;y\in c_j} sim(x,y) \qquad (8)$$

**UPGMA (Group Average)** defines the similarity between any two clusters as the average pairwise distance among all pairs of documents in the two different clusters. This is an intermediate approach between complete and single approaches. Similarity between two clusters *c*$_i$ and *c*$_j$ of sizes *n* and *m* is defined as follows

$$sim(c_i,c_j) = -\frac{\sum\limits_{x\in c_i y\in c_j} sim(x,y)}{n*m} \qquad (9)$$

### 3.4.2. K-Means Clustering Technique

The basic (traditional) K-means Algorithm for finding *K* clusters is applied as follows

1. Select randomly *K* documents as the initial centroids.
2. Assign all documents to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change or for *l* number of iterations.

A variant algorithm of k-means is also applied which is called **incremental k-means**. The main variance between this algorithm and the basic k-means is the time when to recompute the centroids and hence assigning documents to the recomputed centroids. In the basic k-means recomputing of the centroids are taken place after assigning all documents to the closest centroid, while for the incremental k-means recomputing of the centroids are repeated after each document is assigned to the closest centroid.

### 3.4.3. Hybrid Clustering Technique

This technique combines the previous two techniques the hierarchical and the partitional. The agglomerative technique is first applied and gives output of *k* clusters with *k* centroids. These *k* centroids are taken as an input to the k-means algorithm. In the experiments of the hybrid technique the single link agglomerative algorithm is applied in the step of agglomerative technique and the incremental k-means is applied in the step of the k-means technique.

## 4. Experiments

The dataset contains 300 documents representing five different classes. Each class contains 60 documents. There are mainly three groups of experiments. The first group is for comparing the agglomerative different algorithms using different similarity criteria and different *k* values. The second group is for comparing the traditional and the incremental k-means algorithms also using different similarity criteria and different *k* values. The third group is for comparing hybrid technique with the agglomerative and k-means techniques.

For each group of experiments the experiment is repeated four times using different number of documents to be clustered. The first experiment utilizes 300 documents, the second experiment utilizes 100 documents, the third experiment utilizes 50 documents and the fourth experiment utilizes 10 documents. All documents were selected randomly. The number of documents per class in each training data is the same.

The experiments of the basic and the incremental k-means are repeated 10 times and the average of the results is taken to be the final result.

### 4.1 Evaluation of clustering quality

The quality of the clustering solution is measured by means of external quality measure. Two external measures are used which are the Entropy and the F-measure.

### 4.1.1. Entropy

The Entropy measure concerns mainly at how the different classes of documents are distributed within each cluster. For a given cluster $S_r$ of size $n_r$, the entropy of that cluster is calculated as follows

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r} loq \frac{n_r^i}{n_r} \qquad (10)$$

where *q* is the number of classes in the dataset and $n^i_r$ is the number of documents of the *i*th class that were assigned to the *r*th cluster. The entropy of the clustering solution is then calculated by taking the sum of the individual cluster entropies weighted according to the cluster size.

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \qquad (11)$$

A perfect clustering solution is the one that leads to clusters containing documents from only a single class, in this case the entropy will be zero. In general, the smaller the value of the entropy, the better the clustering solution is.

### 4.1.2. F-measure

F-measure depends on the precision and the recall ideas from the information retrieval. First the recall and the precision of each cluster is calculated for each given class. For cluster *j* and class *i* the precision and the recall are calculated as follows

$$\mathrm{Pr}\,ecision(i,j) = \frac{n_{ij}}{n_j} \qquad (12)$$

$$\mathrm{Re}\,call(i,j) = \frac{n_{ij}}{n_i} \qquad (13)$$

where $n_{ij}$ is the number of the members of class *i* in cluster *j*, $n_j$ is the number of members of cluster *j* and $n_i$ is the number of members of class *i*. The F-measure of cluster j and class i is calculated as follows

$$F(i,j) = \frac{(2*\mathrm{Pr}\,ecision(i,j)*\mathrm{Re}\,call(i,j))}{\mathrm{Pr}\,ecision(i,j)+\mathrm{Re}\,call(i,j)} \qquad (14)$$

The overall F-measure for any class is calculated by taking the average of all values for F-measure as follows

$$F = \sum_i \frac{n_i}{n}\ \max\{F(i,j)\} \qquad (15)$$

where the max is taken over all clusters and *n* is the number of documents. In general, the larger the value of the f-measure, the better the clustering solution is

## 5. Results and Analysis

The results are divided into three groups. The first group is for the comparison of different agglomerative algorithms and different similarity criteria. The second group is for the comparison of different K-means algorithms. The third group is for the comparison of hybrid, agglomerative and k-means algorithms. The output value from any experiment represents the **Entropy** or **F-measure** of the clustering solution.

### 5.1. Comparison of the Agglomerative algorithms

This group of experiments aims to find the best agglomerative algorithm from the three applied ones which are single link, complete link and UPGMA. As mentioned in section 3 the main difference between these algorithms is how they define the similarity between clusters. For most of the experiments the single link has the best results.

UPGMA's results are near from that of the single link especially for low number of documents. The complete link has the worst results compared to the previous two algorithms. Figure 3 and figure 4 show the **entropy** results of the three algorithms for different number of documents (N=100, N=300).

Another aim is to find which similarity criteria is the best for document clustering. Table 1, table 2 and table3 show some of the results of different clustering algorithms using different similarity criteria. It is observable that changing the similarity criterion does not change the results except for the Euclidean distance which lead to the worst results. As a result the cosine criterion is used as a similarity measure in all of the next experiments.

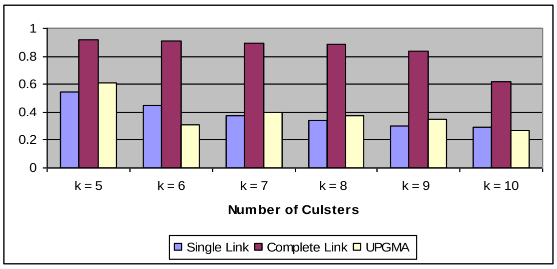


Figure 3: Comparison between Single Link, Complete Link and UPGMA Algorithms for N=100

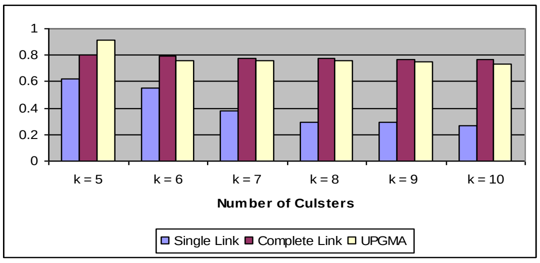


Figure 4: Comparison between Single Link, Complete Link and UPGMA Algorithms for N=300

Table 1: Clustering quality for N=100 and *k* =5

| Similarity | Single Link | | Complete Link | | UPGMA | |
|---|---|---|---|---|---|---|
| Criteria | entropy | f-measure | entropy | f-measure | entropy | f-measure |
| Cosine | 0.541 | 0.24 | 0.919 | 0.151 | 0.611 | 0.243 |
| Dice | 0.299 | 0.24 | 0.919 | 0.151 | 0.611 | 0.243 |
| Euclidean | 0.886 | 0.118 | 0.954 | 0.135 | 0.954 | 0.135 |
| Jaccard | 0.541 | 0.24 | 0.914 | 0.151 | 0.611 | 0.243 |

Table 2: Clustering quality for N=100 and *k* =7

| Similarity | Single Link | | Complete Link | | UPGMA | |
|---|---|---|---|---|---|---|
| Criteria | entropy | f-measure | entropy | f-measure | entropy | f-measure |
| Cosine | 0.375 | 0.278 | 0.896 | 0.153 | 0.395 | 0.270 |
| Dice | 0.375 | 0.278 | 0.896 | 0.153 | 0.395 | 0.270 |
| Euclidean | 0.867 | 0.118 | 0.932 | 0.137 | 0.916 | 0.152 |
| Jaccard | 0.375 | 0.278 | 0.896 | 0.153 | 0.395 | 0.270 |

Table 3: Clustering quality for N=100 and *k* =9

| Similarity | Single Link | | Complete Link | | UPGMA | |
|---|---|---|---|---|---|---|
| Criteria | entropy | f-measure | entropy | f-measure | entropy | f-measure |
| Cosine | 0.299 | 0.286 | 0.841 | 0.158 | 0.350 | 0.270 |
| Dice | 0.299 | 0.286 | 0.841 | 0.158 | 0.350 | 0.270 |
| Euclidean | 0.853 | 0.118 | 0.912 | 0.133 | 0.898 | 0.1475 |
| Jaccard | 0.299 | 0.286 | 0.841 | 0.158 | 0.350 | 0.270 |

### 5.2. Comparison of the K-means algorithms

This group of experiments compares the results of the traditional and incremental k-means. As mentioned in section 3, the main variance between the incremental and the traditional k-means is the time when to recompute the centroids and hence assigning documents to the recomputed centroids. The incremental algorithm tries to improve the centroids and recompute them as soon as any document been assigned to a cluster. This may improve the weak point of the traditional k-means which is the bad choice of the initial centroids leading to a bad clustering quality. The results show that the incremental k-means truly improves the clustering quality. Although the incremental k-means take more computational time complexity but it is still a linear in the number of documents. Figure 5 , figure 6 and figure 7 show the **entropy** result of the comparison between the traditional and the incremental k-means for different number of documents (N=50, N=100 and N=300).

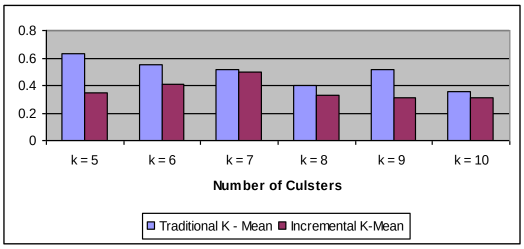


Figure 5: Comparison between the traditional and the incremental K-means for N=50

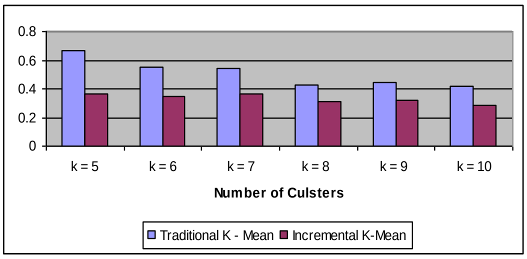


Figure 6: Comparison between the traditional and the incremental K-means for N=100

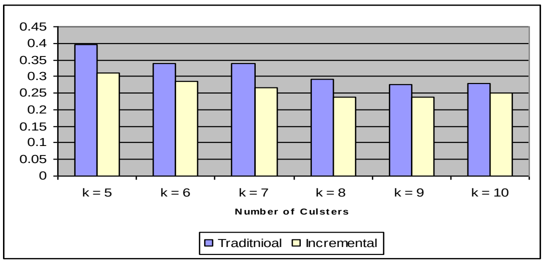


Figure 7: Comparison between the traditional and the incremental K-means for N=300

### 5.3. Comparison of the four clustering algorithms

This group of experiments compares the results of the traditional k-means, incremental k-means, single link agglomerative and hybrid algorithms. The hierarchical clustering is always been described as the best quality algorithm. But the results show that the incremental k-means which is a variant of the k-means produces better clustering quality. Traditional k-means produces the worst clustering quality. The hybrid algorithm produces in most of the experiments the best clustering quality especially for high number of documents. Although it's best quality, the hybrid algorithm combines the time complexity of both agglomerative and incremental k-means algorithms. Figure 8 and figure 9 show the **entropy** result of the comparison between the four algorithms for different number of documents (N=100 and N=300).

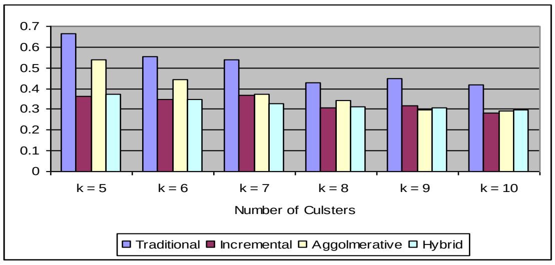


Figure 8: Comparison between the four algorithms for N=100

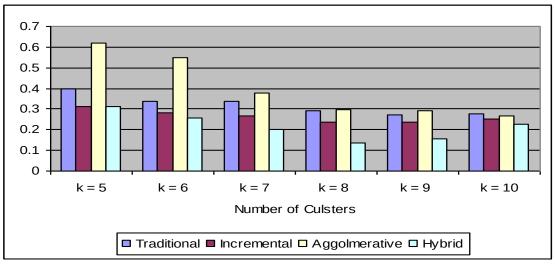


Figure 9: Comparison between the four algorithms for N=300

## Conclusion

The paper presents an empirical analysis of utilizing different clustering techniques and algorithms with different similarity criteria. The experiments were performed on 300 Arabic news documents representing five classes. The news documents were extracted from Google news group. All documents were morphologically analyzed using the RDI system. Three agglomerative algorithms, two k-means algorithms and a hybrid algorithm were implemented. Each experiment is repeated four times with different number of documents and four different similarity criteria. The results show that using different similarity functions does not affect much the results except for the Euclidean criterion which gives the worst result. The single link agglomerative algorithm is better than the other two agglomerative algorithms. The incremental k-means produces better quality than the traditional k-means. The incremental k-means produces as good as or better quality than the single link agglomerative. The hybrid technique mostly has the best results but its computational complexity is the union of that of both the agglomerative and the k-means.

## References

[1] Attia, M. "A Large-Scale Computational Processor of the Arabic Morphology, and Applications", *A Master's Thesis, Faculty of Engineering, Cairo University*, Cairo, Egypt, 2000.

[2] Berkhin, P. "Survey of Clustering Data Mining Techniques", 2002.

[3] Bradley, P. and Fayyad, U. "Refining initial points for k-means clustering", *In Proceedings of the 15th ICML*, pp. 91--99, Madison, WI.,1998.

[4] Cui, X., Potok, T. and Palathingal, P. "Document Clustering using Particle Swarm Optimization", *In the Proceedings of the IEEE Swarm Intelligence Symposium*, Pasadena, CA., June, 2005.

[5] Cutting D. R., Karger D. R., Pedersen J. O., and Tukey J. W. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", *SIGIR '92*, pp. 318--329, 1992.

[6] Dice, L. R. "Measures of the Amount of Ecologic Association between Species", *Journal of Ecology*, Vol.26, No. 3, pp. 297—302, 1945.

[7] Hartigan, J. and Wong, M. "Algorithm AS136: A k-means clustering algorithm", *Applied Statistics*, Vol. 28, pp. 100--108, 1979.

[8] Jaccard, P. "The Distribution of Flora in the Alpine Zone", *The New Phytolgist*, Vol. 11, No. 2, pp. 37—50, 1912.

[9] Jain, A. K., Murty M. N. "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, September, 1999

[10] Larsen, B. and Aone, C. "Fast and Effective Text Mining Using Linear-time Document Clustering", *In the Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, San Diego, California, 1999.

[11] Lee, L. "Measures of Distributional Similarity", *in the Proceedings of the 37$^{th}$ Annual Meeting of the Association for Computational Linguistics*, pp. 25--32, 1999.

[12] Murtagh, F. "Multidimensional Clustering Algorithms", *Physica-Verlag*, Vienna, 1985.

[13] Olson, C. "Parallel algorithms for hierarchical clustering", *Parallel Computing*, Vol. 21, pp. 1313--1325, 1995.

[14] Salton, G. "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", *Addison-Wesley*, 1989.

[15] Sebastiani, F. and Debole, F. "Supervised Term Weighting for Automated Text Categorization" *in the Proceedings of 18th ACM Symposium on Applied Computing (SAC-03)*, pp. 784--788, Melbourne, US, 2003.

[16] Shannon, C. E., "A mathematical theory of communication", *Bell System Technical Journal*, Vol. 27, pp. 379--423 and 623--656, July and October, 1948.

[17] Steinbach, M., Karypis, G. and Kumar,V. "A comparison of document clustering techniques", *KDD Workshop on Text Mining*,2000.

[18] Willett, P. "Recent Trends in Hierarchical Document Clustering: A Critical Review". *In Information Processing and Management*, Vol.24, No. 5, pp. 577--597, 1988.

[19] Zhao, Y. and Karypis, G. "Criterion functions for document clustering: Experiments and analysis", *Technical Report TR #01–40, Department of Computer Science, University of Minnesota*, Minneapolis, MN, Feb 2002.

[20] Zhao, Y. and Karypis, G. "Hierarchical Clustering Algorithms for Document Datasets", *Data Mining and Knowledge Discovery*, Vol.10, pp.141--168, 2005.

# Bilingual Dictionaries: From Theory to Computerization.

Sherif A. Sattar Okasha
Advanced Multilingual Systems Co./Cairo
Sherifokasha2006@yahoo.com

## Abstract

This paper suggests a computationally-enhanced model of an English –Arabic dictionary based on a systematically empirical linguistic analysis of the SL and TLsystems rather than the introspective intuitions of bilingual lexicographers. In this model, computerized text corpora and bilingual semantic concordances play a key role in turning out a reliable bilingual dictionary that does not only serve the purposes of all types of BD users but will also be a robust bilingual repertoire in bilingual NLP systems such as Machine Translation.

**Key Words: Machine Translation, Bilingual Dictionaries, Natural Language Processing**

### 1.0 Introduction

Notwithstanding the great advances in the fields of lexical semantics and computational lexicology, bilingual lexicography (BL) is still a far cry from being a scientific discipline per se .Bilingual comparative analysis of the source language and the target language has not yet built itself into the toolkit of the bilingual lexicographer. Computerization as far as bilingual lexicography is concerned is still restricted to such surface-level automation as can be sufficient to transform a book dictionary into a computerized form. This attitude is definitely oblivious to what potentialities artificial intelligence and smart computation can have for updating the linguistic content of bilingual dictionaries beyond what mere CD –Rom churning can. On the other hand, linguistic theories on bilingual lexicography have been governed-somewhat unconsciously-by commercial considerations. Still in the literature on bilingual dictionaries we can read something about the “purpose” of the dictionary and whether it is targeted for production of the TL by SL users or comprehension of an SL by certain TL users, depending on the direction of the SL-TL pair. This view has always governed such critical issues as sense discrimination in both the source language and target language, rendering the need for semantic disambiguation in a bilingual dictionary (BD) subject to the pre-determined purpose of the dictionary. This paper tries to expose the shortcomings of this view ,adopting a different theoretical position which sees the unity of purpose as the basis of building the architecture of the bilingual dictionary so that it becomes suited to the needs of all users ,be they average users ,specialized ones ,language learners or translators and be they native speakers of the source language or the target language. At the same time, it would be fair to argue that that eclectic view of the bilingual dictionary can be attributed to the limited space made available in paper dictionaries. However, such an argument, one can contend, is no longer valid once we have adopted full-fledged computerization- with its immense potential for storage and retrieval of large chunks of data- as an irrecoverable substitute for the paper dictionaries. In this way, an integrated bilingual dictionary which unifies purpose and content may well come into existence.

### 1.1 A Theoretical Framework

Bilingual contrastive analysis can be done in two stages which can also be regarded as two paradigms for this type of analysis. The first stage is the **preparatory stage**, which involves a thoroughgoing comparative and contrastive analysis of the two language systems and the relative position syntactic categories occupy in both of them before embarking on the compilation work. The second stage is the **compilation stage** in which contrastive analysis focuses on the lexical transfer part of the process .In this part, the lexicographer will select lexicographical equivalents for SL words from a repertoire of translational equivalents provided by bilingual text corpora. Most existing BDs reflect a level of contrastive analysis based on either of the two stages just mentioned. This is why BD theorists classify bilingual dictionaries correspondingly into two broad categories which reflect either one or the other of these two paradigmatic stages .These two categories are: the segmental BD and the idiomatic BD (Piotrowski: 1994.p.148).A segmental BD contains decontextualized lexemic equivalents which are supposed to be substitution forms to be used by bilingually competent users such as translators. An idiomatic dictionary contains highly contextualized lexemic equivalents together with preconstructed expressions. Thus it is best suited for production of Sl texts by non-native speakers of the TL or for comprehension of L1 by L1 learners when they are native speakers of L2 .It can be suited also for communication based on comprehension by, say ,tourists or businessmen, but not so much for translation. This is because translators need ready lexemic equivalents which they can substitute for the source words in the target text at hand rather than idiomatic paraphrases since they are supposed to be already aware of the semantic subtleties of both languages.

It seems, then, that segmental BDs are the most suitable for the purposes of translators.However, segmental BDs usually contain lexical equivalents which serve as contrastive lexical components in the TL system rather than "real" translational equivalents that can be substituted automatically for SL words. For example, all known French-English dictionaries supply the quantitative adjective *some* as a direct equivalent of *de*, despite the fact that a corpus-based statistical study conducted by Catford

(1965) had found that the actual translational equivalent of *de* in English is (0), that is, it is not translated. Yet this equivalent was motivated by a belief that the two words occupy the same position in their respective language systems. . At any rate, there is certainly a difference between using translation as a paradigm against which we model our BD and considering it the be-all and end-all target of the BD.

To use translation as a paradigm in BL is to consider it as a *tertium comparation*, that is a third model against which the other two approaches of the segmental and idiomatic BD are compared with a view to integrating them into a single approach. This segmental-idiomatic approach assumes that translational equivalents can be included in a BD as lexicographical equivalents if they follow a regular pattern of occurrence .The pattern should be so regular that translation equivalents can be reduced to a definite and at the same time variegated number of lexicographical equivalents which represent this pattern in a balanced manner. At the same time they are to be excluded by the lexicographer when they are irreducibly irregular or infrequent and randomly dispersed in TL stretches of discourse. We cannot hope that the successful lexicographical equivalents will be fit for substituting SL words in all relevant contexts but we can expect them to be so for the greatest number of contexts in which SW is likely to occur. It should be also noted that this substitutability presupposes an unchanged SW status on the morphological and syntactical levels and that any change at these levels may affect this substitutability so that the one remaining constant will be: meaning.

This integrational approach cannot be fully realized in a paper dictionary because in such a case translational equivalents will stand as segmental equivalents which, due to considerations of space, will not be accompanied by a representatative variety of expressions in which they occur in TL texts and thus will serve only one purpose ,that of translation. However, in a corpus-based bilingual dictionary, these expressions will serve two purposes: to show the validity of the translational equivalents as lexicographical equivalents, account for their diversity and to be explanatory examples for unsophisticated users. The electronic dictionary seems to be the optimum solution for implementing this view. It should be noted that this solution cannot come out in the form of an automatic acquisition of the lexicographical equivalence data provided by the lexicographer as an output from the compilation stage but rather in the form of this data linked to the natural contexts from which the equivalents were derived. This will require building bilingual semantic concordances, a possibility which will be discussed in section 5.

### 1.2 Translation Vs Lexical Transfer

Before we start, a certain stumbling block has to be removed which has often stood in the way of compiling a bilingual dictionary based on a sound linguistic basis: that is lexicographers' inattention to the difference between lexical transfer involved in translation and that involved in bilingual dictionary-making. Bilingual dictionaries may go to extremes in stating what should remain implied, which results in an explanatory equivalent rather than a lexical one. Such a kind of equivalent will soon prove to be a fiasco once we encounter the SL word in a different context than that which the lexicographer had in mind while lexically transferring it into the TL. For example, the English noun *abortionism* is translated by ***Al-Nafees*** **English- Arabic** dictionary as تأييد حرية الإجهاض (*ta'yid Hurreyat il-ijhaD)* (literally: supporting the freedom of abortion). When this noun occurs in a sentence like: *The US supports abortionism,* it becomes easy to see how erratic such an equivalent is, due to the lexical tautology it causes when we use it in translating this sentence into Arabic. It transpires that the more terminological equivalent حرية الإجهاض *Hurreyat al-ijhaD* "*freedom of abortion*" is the proper one, for it serves both the purpose of comprehension and that of production and would cover a wider spectrum of the contextual occurrence of the SL term than the explanatory equivalent.

### 2.0 Contrastive Semantic Analysis:

#### *2.1 Polysemy in the source language*

Perhaps the most important challenge for a bilingual dictionary user, be he a reader or a translator of a text written in the source language, is to figure out the meaning of the lexical unit for which he seeks a lexical equivalent from between the lines of the source language text. The next step is to spot the nearest equivalent to that meaning from the “map” of lexical equivalents listed by bilingual dictionaries for that lexical unit. If the reader or translator is already familiar with all the senses of the source word he will not make a hard job of "recognizing" the proper TL equivalent as he goes through his bilingual checklist. Otherwise, the practiced user, say a translator or a specialized reader, will perhaps first resort to a SL monolingual dictionary, in order to compare the different meanings listed under the entry for the SL word with the contextualized lexical unit, as it occurs in the source text at hand, till he settles on a satisfactory sense mapping. Then he may consult a bilingual dictionary in search of an exact TL equivalent. As for the language learner or the general user, they may well dispense with the SL dictionary intermediation simply by browsing all the lexical equivalents catalogued by the bilingual dictionary for the source word. The browsing will continue till they find an approximation which they think is the closest thing to the meaning of the source word in the given text, which is an even harder task.

It is our contention that the bilingual dictionary should reduce these steps to a minimum and save its users all this trouble by stating the various meanings of the source language word. Most bilingual dictionary theorists argue that the bilingual dictionary should not state the different meanings of the SL polysemous word unless there is semantic ambiguity in both languages. That is, when there is a polysemous target word for each meaning of a polysemous source word.

The problem with such views is that they restrict comprehension and production to the limited area of temporary users such as language learners and general readers. What about advanced bilingual dictionary users like translators and academic writers? A translator, for example, would want to use the dictionary for comprehension and production at the same moment: comprehension of the SL and production of the TL. Therefore, he would like to have a well defined SL meaning linked to an accurate TL equivalent, regardless of whether he is a native or non-native speaker of the source language, and to the elimination of SL dictionary intermediation.

There are two models of the monolingual lexicon which the bilingual lexicographer can choose from when he sets about the task of incorporating the SL meanings into his dictionary. These two models are: the sense enumerative lexicon and the generative lexicon. The former assumes that a multi-sense word has a definite number of meanings which may be unified under one sense spectrum, a phenomenon which lexical semanticists call *polysemy*, or they may not be unified by the same sense spectrum, a phenomenon traditionally known as *homonymy*. A prototypical example of polysemy is that of the noun *bank*, which could mean a ‘financial institution’ or the ‘building’ used by that institution. The same word can also provide us with a typical example of homonymy when it means ‘side of a river’, a meaning which has nothing to do with the previous ones. As for the generative lexicon, it rejects the idea of a word having a pre-determined set of meanings on grounds that word meaning is affected by the context, the linguistic and the non-linguistic one, and is constantly subject to change in such a way that the sense enumerative lexicon cannot track.

Thus meaning, according to this model, is generated from usage. Let's take the example of an adjective like *fast*. According to the sense enumerative lexicon, three sense spectrums can be tracked of this word within which any subsequent usage of it has to be understood. The word *fast* may indicate the speed of an event or an action as in *fast trip,* or it can indicate the speed of an object when it is the initiator of the speed as in: *fast runner* and *fast car*. Finally, it can indicate the speed of an object when this object, which is expressed by the noun the adjective qualifies, is the product of the speed rather than the producer or initiator of it as in *fast meal*. When an expression like *fast road* occurs, it is automatically mapped, according to the sense enumerative lexicon, to the second meaning. This will be rejected by the generative lexicon model on the grounds that what is being described as "fast" here is not the road, but, rather, the cars speeding on it, which is a new meaning generated from the context and other meanings can be generated from other contexts if we have a reliable corpus.

In order for the generative lexicon model to be implemented in a bilingual dictionary, this will require computerized bilingual text corpora where SL meanings are generated from the contextual co-occurrences of SWs and then mapped to their TL equivalents. The computational paradigm can provide us with a means to integrate the two models of the generative and sense enumerative lexicons. This comes about by extending the repertoire of the sense enumerative lexicon beyond a finite list through comparing the already given meanings against corpus sense-in-text and generating new meanings to be constantly added to the list of meanings.

### *2.2 Lexical Equivalents in the Target Language*

One can argue that bilingual dictionary theories focus mainly on word-to-word equivalence and sense–to-word equivalence and don't give due attention to meaning-to-meaning equivalence. Before carrying the discussion a step further, I would first like to make clear what I mean by these three terms. Word-to-word equivalence is the simplest form of lexical equivalence; it exists when there is a monosemous source word mapped to a monosemous target word. Sense-to-word equivalence occurs when there is a polysemous source word for each meaning of which there is a separate lexical item in the TL lexicon, which does not intersect semantically with it except in respect of that meaning. In other words, the target word in such a case could be monosemous or polysemous. If it is monosemous, there will naturally be semantic equivalence between it and the particular SW meaning for which it was selected. If it is polysemous, the semantic equivalence will hold only between one of its meanings and the meaning of the SW for which it was selected, while other SW meanings will be covered by other, different TWs and so on.

Meaning-to-meaning equivalence, on the other hand, occurs when all the senses of a SW can be mapped to all the senses of a TW without need to go to different TWs to translate the different SW senses. From now onwards I will give a lexical equivalent resulting from meaning-to-meaning equivalence the term *semantic equivalent* while a lexical equivalent resulting from sense-to-word equivalence or word-to-word equivalence will be assigned the term *lexical-word equivalent*.

### *2.2.1 Semantic Equivalents*

A semantic equivalent in the sense just defined could be isomorphic or non-isomorphic, depending on the degree to which the meanings of both the source word and the target word are identical. An isomorphic semantic equivalent occurs when there is a source word which has a certain number of senses or semantic extensions, linked by the same semantic spectrum, and a corresponding target word, having the same number of senses and the same collocation range. Therefore, the TW is said to represent an isomorphic semantic equivalent of the SW if (1) the meanings of the TW are linked by the same semantic spectrum as that whereby the SW meanings are linked; (2) the TW is valid as a lexical equivalent of the SW in all of the latter's contextual co-occurrences (i.e. its immediate collocation range, which the lexicographer discovers through a thorough-going corpus investigation of the word). In such a case, the lexicographer, and often the translator as well, will not need, as we have noted, to go to a separate lexical item in the target language lexicon for each meaning of the source word and will use the same isomorphic TW for all meanings. For example, the English verb *collapse* has three meanings linked by the semantic spectrum of “falling down". This “falling down” could be literal, figurative or psychological, as illustrated below by 1. (a), (b) and (c) respectively:

*I . (a) The building collapsed*

*(b) Negotiations collapsed*

*(c) The man collapsed*

It is to be observed that the Arabic verb ينهار (*yanhār*) has the same three meanings of the English verb and in this way there will be no need to use a lexical-word equivalent pertaining to a different semantic spectrum or an explanatory equivalent which, in addition to being lexically clumsy, does not communicate the SW meaning precisely, as we find in ***Al-Mawrid* English-Arabic dictionary**. In this dictionary, we encounter the Arabic verb يخفق (*yukhfiq),* which means: *to fail*, as the equivalent of the second sense of *collapse*. For the third sense, the dictionary supplies a paraphrase: يصاب بضعف شديد,

(yuSābu biD$^{c}$afin shadid), literally: *to be affected by severe weakness*. This means that the isomorphic semantic equivalent is the ideal lexical equivalent not only on account of its broad semantic coverage but also for its semantic exactitude. One can argue that behind this bilingual semantic isomorphism are macro-level universal principles underlying human cognition. To verify this claim no doubt requires detailed empirical research into many translational language pairs. It can be noticed that the second and third senses exemplified by 1(b) and (c) are a metaphoric extension of the first concrete sense exemplified by 1(a). The comparative corpus analysis of the Arabic translation of *collapse* in different texts where it occurs, in these three senses, reveals that translators favor the bilingual cognitive metaphor of *falling down*, lexically realized in the Arabic verb ينهار **(**anchor), over a lexical-word equivalent pertaining to a different semantic spectrum. This reveals that the semantic equivalent ينهار (yanhār) is the absolute equivalent of the word due to its semantic comprehensiveness and the diversity of the SW contextual co-occurrences it covers (about 50 out of 50 occurrences found in one computerized bilingual corpus); it therefore qualifies as an isomorphic semantic equivalent.

By a non-isomorphic semantic equivalent is meant a polysemous target word semantically identical with a polysemous source word in respect of some senses only, or in respect of all senses, but not all contextual co-occurrences. According to this definition, a non-isomorphic semantic equivalent is produced in either of two cases:

( 1) the source word and the target word are identical in respect of some of their senses, but not all of them. For example, the Arabic verb يكسر (yaksar) is fit as an equivalent of the English verb *break* in almost all its senses which are related by the sense spectrum of ‘splitting in a harsh manner’; yet it is not a correct equivalent for one of these senses – that of ‘cutting’ as it occurs in a sentence like: *The dog broke the girl’s skin,* in which case the proper TL equivalent is the Arabic verb يقطع (yaqT$^{c}$a) *to cut*.

(2) The source word and the target word are identical in respect of all their senses yet the target word cannot cover all the collocational co-occurrences of the source word in one or more of these senses (in this case, it is sufficient for a target word to cover only one contextual co-occurrence of each sense of the source word in order to say that there is a non-isomorphic semantic equivalence between the source word and the target word). To illustrate this case, we can return to the example of the adjective *fast* we mentioned before with its three sense sub spectra of *event-speed, agent-speed* and *patient-speed* in a

sense enumerative lexicon as has been demonstrated before. We find that the English-Arabic lexicographer and/or translator will often use one Arabic word – سريع (sarī$^{c}$) – to express the three broad meanings of the English *fast.* It so happens that the Arabic adjective سريع has these three major senses or, rather, sense sub spectra: Arabic native speakers say: ولد سريع *(*paladin sarī$^{c}$un) a *fast boy,* جري سريع (jaryun sarī$^{c}$un) *fast run,* قطار سريع (qiTārun sarī$^{c}$un) *fast train*.

Yet this Arabic semantic equivalent is still non-isomorphic because it does not cover all the contextual co-occurrences of the source word. For example *fast café* will not be translated into standard Arabic as مقهى سريع (maqha sarī$^{c}$), because سريع does not collocate with مقهى**SA** for *(coffee shop)* in this variety of the Arabic language. The translator or the lexicographer will therefore paraphrase the English NP rendering it as: مقهى للمشروبات السريعة (maqha lil-mashrubāti-s-sarī$^{c}$a) *a café for fast drinks*. Hits of the Arabic monolingual corpus for this Arabic adjective tell us that مقهى سريع *maqha sarī$^{c}$a is* mostly used informally to mean: a high-speed cyber café!

#### *2.2.2 Lexical-word Equivalents*

The lexical-word equivalent is used in either of two cases: the first case occurs when the SW is polysemous; here it is used either to fill in inadequate coverage gaps left by a non-isomorphic semantic equivalent or as the sole type of equivalent when there is no semantic equivalent. The second case is encountered when the SW is monosemous, in which case the lexical-word equivalent is naturally the only choice available.

##### *2.2.2.1 Lexical-word Equivalents When SW is Polysemous*

When the SW is polysemous, the lexical-word equivalent is relevant only in either of two cases: (1) when there is no semantic equivalent, isomorphic or non-isomorphic, for the source word. For example, the English adjective *fat*, has two senses related by the same sense spectrum, i.e. that of size. The first one falls within the semantic field of human body adjectives as in the nominal compound *fat man*, while the second one falls within the semantic field of adjectives that describe inanimate objects as in the nominal phrase: a *fat book.* In modern standard Arabic, there is no single adjective lexeme that combines these two senses precisely and so the lexicographer finds himself forced to resort to discrete lexical items as lexical-word equivalents in the target language: بدين (Baden), literally: *large-bodied* for the first sense and ضخم (Dakhm), *large-sized* for the second.

(2) There is only a non-isomorphic semantic equivalent for the source word and so either the semantic coverage gaps or the collocational coverage gaps have to be filled by lexical-word equivalents in the manner described before. For example, the Arabic verb يشق *(*yashuq), literally: *to split,* could also be suggested as a possible lexical-word equivalent for that sense of *break* uncovered by the non-isomorphic equivalent يكسر (yaksar), i.e. *break* in the sense of *breaking the skin*, as illustrated above in the discussion of the non-isomorphic semantic equivalent. As for gaps resulting from the inadequacy of collocational coverage by a non-isomorphic equivalent, such gaps are also filled by lexical-word equivalents, as exemplified earlier.

#### *2.2.2.2 Lexical-word Equivalents When SW is monosemous:*

When the source word is monosemous, the dichotomy of the semantic equivalent and lexical-word equivalent disappears and only the second pole of it survives – i.e. the lexical-word equivalent. Strikingly enough, the relationship between the two poles is not one of binary opposition but rather one of complementarity: The lexical-word equivalent, when properly employed, fills in gaps left by a non-isomorphic semantic equivalent. For a monosemous source word, the situation is different: there is no scope for such gaps since the source word has a single meaning and the lexical-word equivalent is the only lexical equivalent possible. There are three cases for the lexical-word equivalent when the source word is monosemous:

A) The lexical-word equivalent is monosemous and its meaning is identical to that of the source word. Examples of this phenomenon abound in all language pairs and it is indeed one of the reasons why lexical transfer between languages is possible. It can be observed among abstract lexical items as well as concrete lexical items. Nouns indicating plants and animals in English, for example, are mostly monosemous words for which there are equally monosemous nouns in Arabic. A word like *bravery* in English has many synonymous lexical-word equivalents in Arabic, all of which are single-meaning words.

B) The lexical-word equivalent is monosemous yet its meaning is not identical to that of the source word. The result is that the source word meaning is acquired by the target word and added to its already existing single meaning. For example, the Arabic noun أصالة (‘sale), which originally meant *antiquity or* *precedence of occurrence of something*, came to acquire the meaning of ‘creative thinking’ when it was used as a translation of the English noun *originality* which means ‘creative thinking’ or ‘newness based on creative thinking’. What happened is that the English source word extended the Arabic sense spectrum of the Arabic word-equivalent so that it means also ‘precedence of thinking’, a sense unfamiliar to the word before this translation came into existence. .

C) The lexical-word equivalent for the monosemous source word is polysemous. Here the politely problem is transferred from the source language to the target language and in this case it ceases to be a comparative problem of lexical equivalence between the source language and the target language, but rather one of comprehension related only to the target language. To explain this point, let us pick an example. The English noun *science* has a single meaning – i.e. that of ‘experimental study of the natural world’. The Arabic target word علم ($^c$ilm) has two meanings: the first one refers to knowledge in general and the noun in this sense behaves as a deverbal noun which inherits the argument structure of the verb from which it is derived – the Arabic verb يعلم (ya$^c$lam), *to know*. The second meaning refers to ‘experimental science’. Having selected this Arabic equivalent, it will then be the task of the lexicographer to select from its two meanings the one which can be mapped to the source word *science* – in this case the second meaning, of course – since the target language speaker certainly needs this mapping in order to “comprehend” the meaning of the source word.

The common mistake which bilingual lexicographers inadvertently make is that they usually fail to recognize the significance of differentiating between the semantic equivalent and the lexical-word equivalent. They tend to introduce lexical-word equivalents for the different meanings of the source word without making sure that there is one lexical equivalent which can be suitable as a TL semantic equivalent to all or most of these senses, which could be the first lexical equivalent introduced. In this way, they bar the target language from revealing its semantic richness on the one hand and a considerable part of its expressive force is lost in the translation on the other hand, as we have seen in the case of *collapse*.

### 2.3 Grammar and Meaning in a BD

There is a systematic relationship between meaning and grammar which affects the choice of lexical equivalents in a BD. We will restrict the concept of grammar in this section to that common sense found in traditional textbooks which focuses on basic syntactic and grammatical properties of words.

.Substitutability of a given TL equivalent is not a given .It depends on many factors. One of these factors is the variability of the syntactico -semantic properties of the Sl word. For example, the English noun *suicide* can be countable or uncountable. The conceptual lexical equivalent of this English noun is intiha<r, which is lexically substitutable for the SL noun only when the latter is uncountable. When *suicide* behaves syntactically as a countable noun, this equivalent should be changed into *ha<latu Intiha<r, or 'amal Intiha<ry.*

The countable-uncountable alternations turn out to be responsible for many semantic alternations between an abstract concept and an abstract entity within the same lexical unit. As an example, there is the alternation between *abortion* (uncountable, abstract concept) and an *abortion* (particular event, countable) .As we mentioned earlier, An English Arabic dictionary has to provide two different equivalents for the two variants of the English noun, Ighad {for the uncountable variant and *'analyst Ighad {*for the countable one. It's only when such variations show a regular, systematic pattern that reflects on TW substitutability that they have to be tackled by a BD at all. One way to do this in a paper dictionary is to list them as subentries under their lemmatized forms and list the lexicographical equivalents in the opposite direction.

Shifting the focus to adjectives, we can say that, in some cases the syntactical position of the adjective either before or after the noun can have some bearing on its semantic interpretation in a way which affects the choice of lexical equivalents in Arabic. It should be noted first that we do not mean by the syntactic position of adjectives those cases in which the adjectives is grammatically fixed in one position only ,either attributively or predicatively. This having been said, we can proceed. When a regular adjective is used attributively, its meaning may be slightly different than when it is used predicatively after a copulative verb. For example, in 2a and 2b below

2a He is a tense person

2b H is/looks tense

It is easy to notice that *tense* in 2a expresses a rather stable trait in the noun described by the adjective while in 2b it refers to a temporary state of affairs.

Generally speaking, lexical equivalents of adjectives will not be affected by their mobility.However, when the meaning alternation resulting from this mobility is not reflected by the corresponding position of the regular adjectival equivalent; the alternation has to be preserved in the target language with lexical means by introducing a semantically different adjective for each position. So it seems that one

Arabic equivalent for *tense* in both its syntactical positions is unlikely. The Arabic adjective *mutawatir* ,supplied by three English-Arabic dictionaries, is a stative adjective and so will be fit to substitute for *tense* in the predicative position illustrated by 2b.For the attributive position exemplified by 2a,we suggest قلوق *qaluq* ,which is an inherent adjective in Arabic and is therefore more semantically felicitous in this position.

In order for the lexicographer to make precise predictions of this kind, he has to restrict his test criteria to two variables only: the syntactical position of the adjective and its meaning and neutralize any other variables that may influence his decision such as the communication situation in the texts he is examining. .To achieve this end, test sentences of a simple structure like that of 11 and 12 above should be gleaned out of text and analyzed.

**3. Contrastive Morphological Analysis**

.Arabic is often described as a non-concatenative language. This is because word formation in Arabic is based on the derivation of various morphological patterns from a single root rather than a concatenation of affixes to a stem. Each morphological pattern reflects a set of semantic patterns. But this does not mean that there is no affixation in Arabic morphology. In modern Arabic morphology, concatenation and affixation play a central role in word formation and coinage in order to cope with the terminological needs of the language in the different domains .However, progress in Arabic morphology has been very slow and random in terms of extending the semantic applicability of already existing morphological patterns .

Such slow and random progress has had negative influence on lexical transfer from foreign languages, especially English, into Arabic. This influence consists in using certain Arabic morphological patterns as equivalents to some derivational patterns in English without careful study based on contrastive analysis on the morpho-semantic level. For example, The Arabic nominal category known as almas{dar als{ina<'i (adjectival masdar) is often used both in the translation of English “isms” and names of sciences which end with the suffix "ics".To give but a few examples, there is Ishtra<kiya for socialism,*ma’lumatiya and uslubyia* for *informatics and stylistics,* respectively.

A careful contrastive analysis of the Arabic adjectival masdar and the equivalence patterns based on it reveals that it is not an accurate choice for translating science names which end in *ics*.The line of reasoning on which we base our argument is as follows. The adjectival masdar in Arabic is semantically parallel to a relational adjective. A relational adjective is an adjective which indicates a relation to a noun and ascribes the attributes of this noun to the noun which it qualifies. It may be used as an inherent adjective as in mu'amala Insanyia (human treatment) and hajama<t wah{shyia (brutal attacks) wherein the attributes of *a human* and those of *wah{sh* (brute) are ascribed to the deverbal noun *mu'amala* (treatment) and the plural noun hajamat (attacks),respectively. Or it may be used to indicate the mere existence of a relation as in I'tiba<ra<t syia<syia (political considerations), that is, considerations related to politics. In this way this noun-related adjective in Arabic serves a twofold function: it can be used subjectively as an inherent adjective and objectively as a relational adjective. By analogy, the *adjectival masdar* can be used to do these functions nominally.; For example, The nouns Insanyia (humaneness) ,wah{shyia (brutality) and hamajyia refer to subjective personal traits ,while *uluhyia* (divinity) refers to a relation as in the phrase *uluhyiat al-masdar* (divinity of origin).However, the latter,relationl use of the adjectival masdar is very rare in Arabic.

In English, isms can also be used objectively as names of doctrines or subjectively to name individual intellectual attitudes .In this way there is semantic symmetricality between the Arabic adjectival masdar and an English *ism*, which makes the former a suitable pattern for translating such *isms*. On the other hand, names of sciences are characterized by a neutral degree of objectivity since they refer to disciplines of knowledge which are concerned with objective realities. .Therefore, their lexical equivalents have to be as neutrally objective, which the adjectival masdar is not.

It is to be observed that using the adjectival masdar in the translation of names of sciences, whether natural or human sciences, is a relatively new trend. The more established one is the use of a pluralized relational adjective on the grounds that the noun which it qualifies is elliptically slashed. On this assumption, a noun like riyad {yiat (mathematics) is a reduced form of umur Ryad {iya (mathematical matters) in such a way that the plural noun *umur* (matters) is slashed and replaced by the plural morpheme ات. .What has been said of mathematics can also be said of linguistics, which is often translated as *Lesanyia<t*

We conclude thus far that the pluralized relational adjective is more appropriate, from the semantic point of view, for the translation of science names since it is elliptically derived from a semantically neutral nominal compound. The adjectival mascarpone the contrary, is less appropriate due to the fact that it is often used to label personal traits or value-laden doctrines, which all runs counter to the objective nature of science. Shifting the focus again to the English-Arabic BD, we find that we cannot burden the bilingual lexicographer with finding solutions to such complicated problems in Arabic morphology. It is the role of Arabic-language academies to solve these problems. Then, lexicologists can receive the results of their research and use them in their arduous contrastive analysis which is essentially related to the *preparatory stage. Later* on, it will be the task of lexicographers to put such results into practical application in the *compilation stage*. Without parallel tagged text corpora, no such comparative morpho-semantic analysis of the lexical categories in both languages can be hoped for.

## 4. Contrastive Syntactic Analysis

. In a corpus-linked bilingual dictionary syntax acquires a particular importance due to the interdependent relationship between syntax and semantics in general. There are already many theories which try to frame the relationship between syntax and semantics, the most important of which, in my view, as far as bilingual lexicography is concerned, is the valency grammar theory, which was developed by the French linguist Lucien Tesniere (1893-1954). The valency metaphor is derived from chemistry and refers to the tendency of an atom to acquire or lose a certain number of electrons while it forms a bond with the atom of another chemical element. In language, the atoms are the syntactic categories and electrons are the arguments which they acquire or lose in their interaction with other syntactic elements. Syntactic valencies represent the argument structures of the lexical items. The syntactic valencies of a verb are the subject, object or complement arguments and those of a noun or adjective are the phrasal complements which are attached to them and tied to their semantic representation.

Such quantitative specification of syntactic valencies suits the segmental nature of the lexicon and makes it easier for computers to deal with them as minimum coded units, such as V, which stands for a univalent (i.e. intransitive) verb, *Vn* which stands for a bivalent verb whose argument structure consists of a subject and a direct object, *Vpr* for a bivalent verb with a subject and a prepositional complement forming its argument structure and so on.

Semantic valencies represent the semantic content of the syntactic arguments in the form of semantic features and taxonomies,as we will see in the next section.

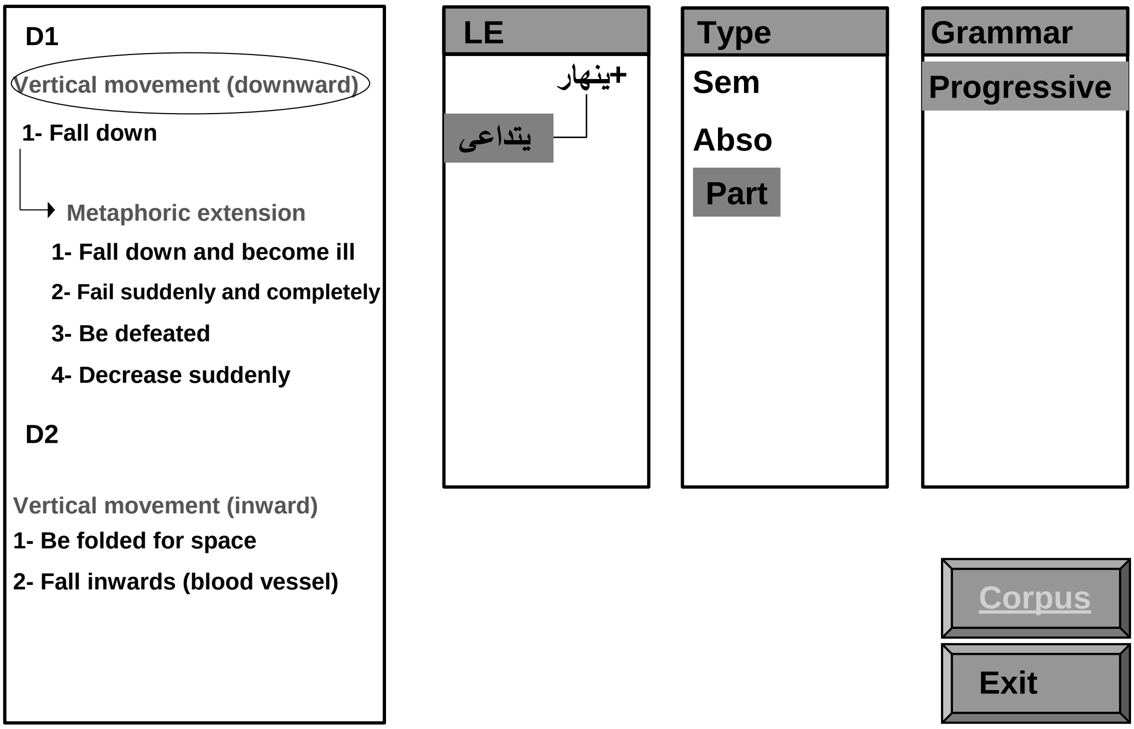
## 5. Implementation Mechanisms &the Role of Computers

In a semantically organized computerized English-Arabic dictionary, syntactic valency (SVL) is the 'blade' whereby a lexical entry is divided into lexemes and the conceptual content of each lexeme into lexical units. Each set of lexical units is unified by a semantic spectrum, which could be a semantic extension, a semantic field or a cognitive metaphor. Semantic extension is a method of relating senses of a polysemous word semantically rather than at a level of semantic organization. A set of senses unified by semantic extension of a core concept usually have a semantic equivalent in the target language. For example, *love* in the sense of 'strong liking' as in *love of horses* is a semantic extension of the primary sense of *love* as 'warm affection'. In Arabic both senses will have the semantic equivalent حب (Hub). Semantic field is a broad term for taxonomy, a feature or a dimension. Senses grouped under a given syntactic valency can be divided into taxonomic subsets. For example, the noun *bed* has several senses that can be divided taxonomically. The first sense is assigned the taxonomy *furniture* while the other two senses are grouped by the taxonomy land *surface* (sea bed, bed of roses, a bed of rock). Needless to say, it is sufficient to attach the taxonomy name to the first sense of the subset unified by the same taxonomy. However, when a semantic extension leads to a change of taxonomy the sense generated by extension should be assigned its own taxonomical label if it happens to have the same semantic equivalent in the TL. As an example, the first sense of *bed* is semantically extended to mean 'a state of sleep', as in the sentence: *she put the child to bed.* The latter sense has to be assigned the taxonomy *state*. A semantic feature can be used to group senses in a manner which shows a certain contrastive value. For example, the semantic feature 'inchoative' (i.e. gradual) can be assigned to the first three senses of the verb *decline* (decrease gradually, go into a worse condition and slope downwards). For these three senses there is an inchoative verbal equivalent in Arabic, that is, the semantic equivalent ينحدر (yanHadir).

It is important to note that these levels of semantic organization are not mutually exclusive in theory. A feature, in principle, can well be combined with a taxonomy (e.g. to narrow down its applicability). Semantic extension, far from being a level – as we have just noted – is a technique which can permeate all levels. The message is that we use the single semantic spectrum which is most suitable to highlight contrastive properties of the two languages in so much as they affect our choice of lexicographical equivalents, and not to show the semantic features of each language separately. The taxonomy 'decrease verbs', for example, does not bring into focus the contrastive inchoative feature of the English verb *decline* and the Arabic verb ينحدر since 'decrease verbs' in English and Arabic could be inchoative or non-inchoative. This is why we use the semantic feature inchoative+ on its own for grouping the above-mentioned senses of *decline* into one set, rather than the taxonomy.

Unlike a feature, a dimension represents a concept on a scale of continuous, graded properties rather than a set of binary, discrete ones. For example, in the semantic representation of the verb *collapse* as a univalent verb *V*, the dimension of movement grades from vertical downward movement to vertical inward movement. Between these two dimensional spectra stands the cognitive metaphor of *falling down*. A cognitive metaphor is a semantic extension of a dimension or a dimensional spectrum. The dimension is conceptually more comprehensive than a cognitive metaphor. The latter generates from the concept several senses on the same point of the dimensional scale.

Senses unified by a cognitive metaphor will mostly have one isomorphic semantic equivalent while senses unified by a dimension could be covered by a non-isomorphic semantic equivalent and lexical-word equivalents that fill the non-isomorphic gaps. In our would-be **English–Arabic Bilingual Dictionary** there is a separate screen for each syntactic valency. Figures 1 and 2 show a semantic representation of the English verb collapse as a univalent verb (V) together with its Arabic equivalents in a linguistically-based, corpus-based and corpus-linked electronic English-Arabic dictionary. To simulate the mouse shifts in the original prototype, the V screen of *collapse* is split here into two screens.



D1

Vertical movement (downward)

1- Fall down

Metaphoric extension

1- Fall down and become ill

2- Fail suddenly and completely

3- Be defeated

4- Decrease suddenly

D2

Vertical movement (inward)

1- Be folded for space

2- Fall inwards (blood vessel)

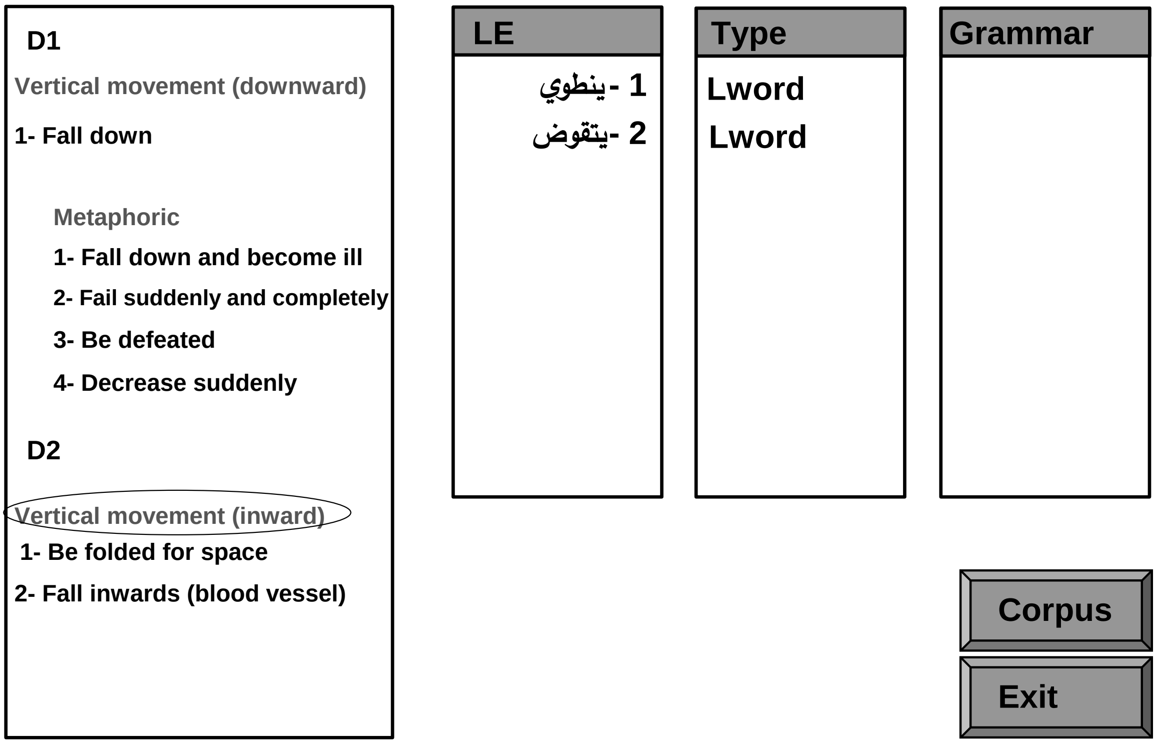
LE: ينهار+ ; يتداعى

Type: Sem ; Abso ; Part

Grammar: Progressive

**Fig 1: Dimension 1 (D1) of *collapse*-V (encircled): downward vertical movement**

The first text box to the left in Fig 1 shows the first dimension of the verb collapse, which covers the concrete concept of falling down and is metaphorically extended to cover four other related senses, all of which are linked in the data set to their relevant semantic equivalents as shown in the first list box to the left (where **LE** stands for Lexicographical Equivalent). The next list box shows the type of equivalent. **Abso** stands for absolute equivalent, i.e. an equivalent which covers a great number of contexts; **Part** is short for partial equivalent, i.e. an equivalent which covers a limited number of contexts. The partial equivalent يتداعى is linked to a special grammatical feature in the third list box which specifies that it can be used only as an equivalent of the source verb when the latter occurs in a progressive aspect. This is because يتداعى is an inchoative verb while *collapse* is a terminative verb and so it cannot be an equivalent for it when it occurs in the past or present simple tenses.



D1

Vertical movement (downward)

1- Fall down

Metaphoric

1- Fall down and become ill

2- Fail suddenly and completely

3- Be defeated

4- Decrease suddenly

D2

Vertical movement (inward)

1- Be folded for space

2- Fall inwards (blood vessel)

| LE | Type | Grammar |
|---|---|---|
| 1 -ينطوي | Lword | |
| 2 -يتقوض | Lword | |

**Fig 2: Dimension 2 (D2) of *collapse* -V: Vertical inward movement.**

Fig 2 shows the second dimensional spectrum D2 which relates to downward vertical movement. It covers two senses which are completely different in meaning and register yet are related by the same dimension. They are linked to two different lexical-word equivalents in the second list box. L-word in the type box stands for Lexical-word equivalent.

To link a bilingual corpus properly to the bilingual database we need to build a bilingual semantic concordance (BSC). A semantic concordance (SC) is defined by Miller et al (1993, 303) as "a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon". A bilingual semantic concordance can then be defined as "a bilingual textual corpus and a lexicon so combined that every substantive word in the SL text is linked to its appropriate sense in the SL lexicon and its TL equivalents in the parallel corpus and the TL lexicon"

Building a BSC as such from scratch is both costly and time-consuming. Using commercially-available tools will make our job much easier and more cost-effective. These tools are: a bilingual machine-readable dictionary, a part-of-speech-tagged bilingual corpus and a grammatically annotated computerized English dictionary.

Syntactic categories and their valencies in the form of V, Vpr, Adj:pr, V.to.inf ..etc can be extracted from an English electronic dictionary which has such tags for each lexical unit. Then they can be mapped manually to their lexicographical equivalents in the Bilingual Dictionary. The syntactic tags of the part-of-speech tagger are also to be mapped to the part of speech tags of the English lexicon (V, adj, N etc). In this way we can build a crude English parser which we can use to do an automatic syntactic tagging of the corpus texts. Then human syntactic and semantic taggers will have to improve automation results by manual bootstrapping. This will involve correcting errors of automatic syntactic tagging by linking corpus lexemes to their correct syntactical valencies provided by the SL lexicon. It will involve also semantic tagging of corpus words by linking them to their proper senses of the lexicon. Thanks to the close relationship between semantics and syntax, we assume that most of the words that were correctly syntactically tagged by the parser are also semantically tagged in a correct way. Of course if we had a semantically disambiguated parser, this would save a lot of manual tagging. Finally the Arabic hits in the TL side of the bilingual corpus will appear with the SVL-linked lexicographical equivalents. Now that the bilingual corpus has been linked to a bilingual dictionary, the lexicographer becomes ready to embark on his arduous task of compiling his own linguistically-based, corpus-based bilingual dictionary. Among the myriad tasks he will have to undertake is that of updating the lexicographical equivalents of the traditional Bilingual Dictionary, classifying them semantically and adding new ones based on extensive corpus research.

## Conclusion

The formulation of a linguistic framework as well as an empirical model for a corpus-based English Arabic electronic database is not a luxury with which we can afford to dispense. Rather it is an exigency dictated by an increasingly globalized world in which cultural contact and linguistic pluralism are the norm rather than the exception. The major points which we need to re-emphasize in conclusion are: First, the importance of selecting a computationally-tractable model for a monolingual dictionary to be used as an input for the bilingual dictionary. Second, the need to focus on the semantic expansion of the Arabic lexicon not just its the lexical word power so as to provide the lexicographer with a repertoire of word-senses that ultimately extend the applicability of already existing lexemes. This can be achieved through compiling an Arabic dictionary in which semantic generation is based on extensive corpus-based analysis not just on the intuitions of lexicographers. Third, the integrational approach to the BD suggested by the author cannot be achieved without a parallel computationally integrative approach. Such an approach certainly draws heavily on state-of-the-art techniques in Natural Language Processing and data mining as well as the traditional interface-oriented software mechanisms in revolutionizing the content and structure of the Electronic BD. In this way it exacts a radical change in the non-linguistically -minded interface culture propagated by current computerized BDs.

**References**

1-Alberton, D.J: .2003 http://www.linguistik.uni-erlangen.de/~msbethke/papers/Valency.pdf

2-Al-Kasimi, M. Ali. 1977. Linguistics and bilingual dictionaries. Leiden. E. J. Brill.

3-Bell, Roger.T.1995.Translation and Translating: Theory and Practice. Longman, London and New York.

4-Catford, John C. (1965).A linguistic Theory of Translation. London: Oxford University Press

5-Crystal, David. 2003. A dictionary of linguistics and phonetics. Oxford

6-Fellbaum, Christiane. 1998. A Word Net Electronic Lexical Database. MIT

7- Larson, Mildred L.1984.Meaning-based translation University Press of America.

8-Lyons, John (1982) . Language, Meaning and Context. London: Fontana.

9-Miller, G. A. (1995).Word net: A lexical database for English. Communications of the ACM.

10-Piotrowky,Tadeuz. 1994. Problems in Bilingual Lexicography. PHD, Wroclaw.

11-Pustejovsky, J.James. 1998. The Generative Lexicon. Massachusetts Institute of Technology.

12-Zgusta, Ladislav (1988). "Translational Equivalence in Bilingual Lexicography", Bahukosyam dictionaries. vol. 9.

13-Ba'labaky, Munir.2003, Al-Mawrid English-Arabic Dictionary. Dar Al-I'lm lilmala<yin

14- Wahba, Magdy 2003.Al-Nafees English-Arabic Dictionary. Beirut for Publishing and Distribution.

15-Sakhr Online Dictionary, www.ajeeb.com

# Experiments for Automatic Arabic Diacritization

Mohsen A. A. Rashwan[1, 2], Mohammad Al-Badrashiny[1], Mohamed Attia[1]

[1] The Engineering Company for the Development of Computer Systems; RDI, Egypt www.RDI-eg.com
[2] professors in the dept. of Electronics and Electrical communications, Faculty of Engineering, Cairo University
{ Mohsen_Rashwan, Mohammed.Badrashiny, m_Atteya }@RDI-eg.com

## Abstract

*A hybrid system for automatic Arabic diacritization for the raw Arabic text is introduced here. The system consists of two layers; the first layer deals with the complete form of the Arabic words based on the m-gram technique and A$^*$ search. If the word is out of vocabulary then we back off to the second layer. The second layer is a complete system (based also on m-gram and A$^*$ search) that deals with the word factorization by factorizing the word into lexemes (prefix, root, form and suffix). The second layer has the advantage of coverage but suffers from relatively low accuracy especially for the syntactical diacritization. The accuracy of the hybrid system is much better than any of the two systems. The errors of the factorized system are (7.8% for the morphological diacritization and 32.5% for the syntactic diacritization). The errors of the hybrid system are (3.6% for the morphological diacritization, and 13% for the syntactic diacritization). The fully diacritized words are only available in certain domain so we studied the effect on this domain and others.*

## 1. Introduction

Automatic words diacritization is one of the NLP challenges in the languages that have diacritics especially in Arabic since the change of one diacritic can change the meaning completely (e.g. رِجْلُ has the meaning of foot but رَجُلٌ has the meaning of man). Automatic words diacritization is important for Text-To-Speech (TTS), Word sense disambiguation, Machine translation, and Part-of-Speech tagging (PoS-tagging) applications [2]. The main challenge in Arabic is its rich derivative and inflective nature, so it is very difficult to build a complete vocabulary that can cover all possible words. The importance of language factorization gets more and more crucial as the vocabulary of the subject language gets richer [2]. In fact, while Arabic is on the extreme of richness as per its vocabulary when regarded as full-form words, this language is also on the extreme of compactness of atomic building entities due to its very systematic and rich derivative and inflective nature [1], [2], [5], [7], [13]. But the main problem of dealing with the language in the factorization form is that it has moderate performance (accuracy) with the need of fully POS annotated training data (costly data); while the problem of unfactorizing systems is the low coverage ratio of the language [6], [7]. So, the represented system here is a hybrid system between the factorizing and unfactorizing systems to achieve both the highest probability estimation through the unfactorizing system and the highest coverage ratio through the factorizing system. Section two will discuss the factorizing system. Section three will discuss the hybrid system. Section four will discuss the word diacritics disambiguation in both techniques. Section five will represent the experimental results and the evaluation, and finally section six will be the conclusion.

## 2. Arabic factorizing system

The Arabic word has two types of diacritizations (morphological and syntactical). The morphological diacritization is affecting the meaning of the word (e.g. رِجْل has the meaning of foot but رَجُل has the meaning of man) while the syntactical diacritization is affecting the syntactic meaning of the word (e.g. is it a subject رَجُلُ? or an objective رَجُلَ). In *Figure 1* below the input Arabic text is sent to the Arabic *Lexical Analyzer* to hopefully get the most likely morphological diacritization, and morphemes sequence.

From Arabic morphemes, an Arabic *PoS Tagger* is needed to extract the Arabic PoS-tags, and then a trained statistical Arabic *Syntactic Diacritizer* is deployed to infer the most likely syntactic diacritization to complement the lexical diacritization of the input text [2].

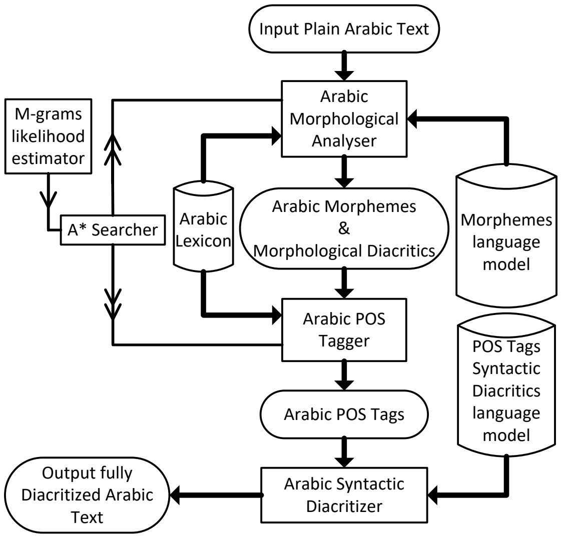

Input Plain Arabic Text
Arabic Morphological Analyser
M-grams likelihood estimator
A* Searcher
Arabic Lexicon
Arabic Morphemes & Morphological Diacritics
Morphemes language model
Arabic POS Tagger
POS Tags Syntactic Diacritics language model
Arabic POS Tags
Output fully Diacritized Arabic Text
Arabic Syntactic Diacritizer

*Figure 1*: The factorizing system architecture for Arabic diacritization.

### 2.1. Morphological model: [2], [5]

Due to the highly derivative and inflective nature of the Arabic language, it is much more comprehensive, effective, and economic to deal with its compact set of basic building entities; i.e. morphemes, than its unmanageably huge generable vocabulary. Following that morpheme-based approach, the canonical lexical structure of any Arabic word **w** has been formulated as a quadruple;

$$\boldsymbol{w} \rightarrow \underline{q} = (t : p, r, f, s) \quad \_\_ \ (1)$$

Where t is Type Code (with possible types are Regular Derivative, Irregular Derivative, Fixed, Arabized), p is Prefix Code, r is Root Code, f is Pattern Code, and s is Suffix Code.

*Table 1* below shows this model in application on few sample Arabic words.

| Sample word | Type | Prefix & Prefix Code | Root & Root Code | Pattern & Pattern Code | Suffix & Suffix Code |
|---|---|---|---|---|---|
| اَلْكِتَابَات | Regular Derivative | ال<br>**9** | ك ت ب<br>**3354** | فِعَال<br>**684** | ـات<br>**27** |
| اَلْعِلْمِيَّة | Regular Derivative | ال<br>**9** | ع ل م<br>**2754** | فِعْل<br>**842** | ـِيَّة<br>**28** |
| مِنْ | Fixed | –<br>**0** | مِنْ<br>**63** | مِنْ<br>**118** | –<br>**0** |
| مَوَاضِيع | Regular Derivative | –<br>**0** | و ض ع<br>**4339** | مَفَاعِيل<br>**93** | –<br>**0** |

*Table 1*: Canonical lexical structure of sample words.

### 2.2. Syntactical model:

The syntactical diacritization in this model is depending on the PoS-tagging. But composing Arabic PoS-tags set necessitates scanning the lexico-syntactic features of each possible word of the Arabic vocabulary which is apparently infeasible. Instead, thanks for the morpheme-based approach, the features of each morpheme in the relatively compact knowledge base have been scanned, then digested through several iterations of decimation into a non redundant compact Arabic PoS-tags set [2], [3].

During that scanning process the following criteria has been adhered to:

1- All the existing lexico-syntactic features must be named and registered, which aims to the completeness of the resulting PoS-tags set.

2- All the named and registered features must be atomic, which aims to compactness and avoids redundancy in the resulting tags set. This in turn is vital for the effectiveness of the based upon PoS-tagging process - which is essentially an abstraction process - and all higher processing layers as well.

3- All the named and registered features can be ensured upon the PoS labeling of the morphemes in our Arabic lexical knowledge base.

The total size of the PoS-tags set is 62 tags [2], [3].

## 3. Arabic hybrid system

In this system we depend on the full form of the word (i.e. with its all diacritics either the morphological diacritics or the syntactical diacritics) to build an m-grams language model during the offline phase to be used in the probability estimation of the words during the runtime phase to find the most relevant diacritics for the words.

The system architecture of the hybrid system is shown in *Figure 2*. During the offline phase the training data is passed to the "*Dictionary builder*" module to collect all unique words from the text to build the language dictionary and also to give an index to each unique word. We have to note here that the word is considered unique if it has any difference from its similar word (e.g. كُرَةُ and كُرَةَ are two unique words). After the dictionary is built the training data is then passed to the "*Text to index converter*" module that uses the already built dictionary to give an index for each word in the training data in order to decrease the memory usage and to ease the process of words counting. Now the converted data is passed to the "*Words m-grams language model builder*" module to build a statistical language model that will be used in probability estimation after that [2], [4], [8], [9], [10], [11]. So at the end of the offline phase we have a dictionary that carries the unique words and their indices and we have the statistical language model that will be used for the probability estimation.

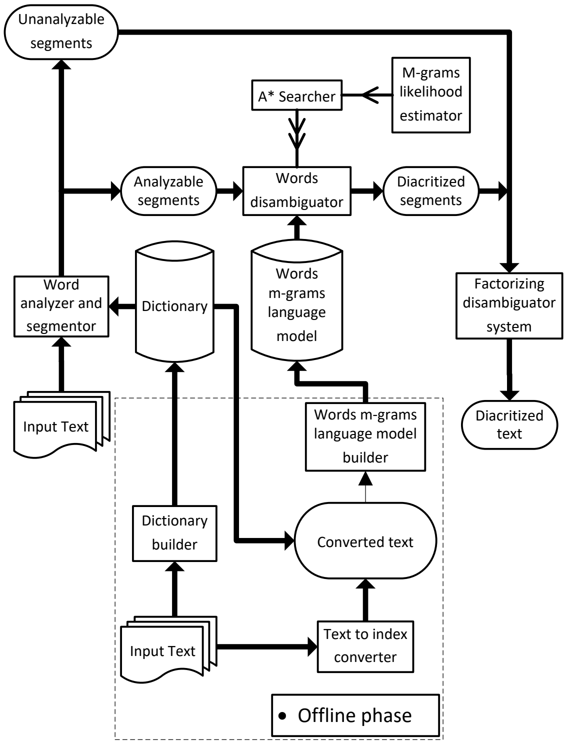


*Figure 2*: The hybrid system architecture for Arabic diacritization.

In the runtime phase the input text is passed to the "*Word analyzer and segmentor*" module that uses the dictionary to get all possible analyses for the given word (e.g. the analyses of كرة is كُرَةٌ, كُرَةَ, and كُرَةِ). If the word is not found in the dictionary the module separates the phrase into an analyzable segment and an unanalyzable segment (see *Figure 3*). The analyzable segment is then passed to the "*words disambiguator*" module that generates the solution trellis for the given analyses (see *Figure 4*) [2], [4], [8], [9], [10], [11]. Then using the $A^*$ search algorithm [2], [9], the most relevant solution according to the given statistical language model is generated. The analyzable segment is diacritized using the 3-gram language model and $A^*$ search. We apply the back-off technique on the unanalyzable segment but since the two segments (the analyzable and the unanalyzable segments) were originally from the same sentence, we send the output from the "*words disambiguator*" module (the unanalyzable segment) to the "*factorizing disambiguator system*" module. The solution of the analyzable segment that came from the "*words disambiguator*" module will be used to help the "*factorizing disambiguator system*" module to find the most relevant solution for the unanalyzable segment according to the solution of the analyzable segment.

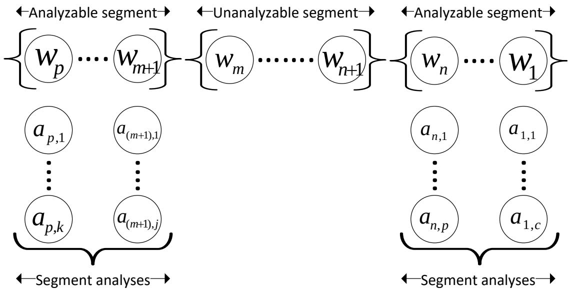


*Figure 3*: phrase segmentation process.

$$\begin{array}{cccc}
w_1 & w_2 & \cdots & w_L \\
\hline
a_{1,1}\bullet & a_{2,1}\bullet & \cdots & a_{L,1}\bullet \\
a_{1,2}\bullet & a_{2,2}\bullet & \cdots & a_{L,2}\bullet \\
\vdots & \vdots & & \vdots \\
a_{1,J_1}\bullet & a_{2,J_2}\bullet & \cdots & a_{L,J_L}\bullet \\
\hline
\underline{a}_1 & \underline{a}_2 & \cdots & \underline{a}_L
\end{array}$$

*Figure 4*: The input sequence $\underline{W}$ and the solution trellis of possible analyses ($\underline{a}_1$, $\underline{a}_2$, ..., $\underline{a}_L$).

## 4. Word diacritics disambiguation

This section will discuss the diacritics disambiguation technique in both the factorizing and the unfactorizing systems.

In both systems the main idea is to build an m-gram model for the entity that is required to be disambiguated (the full form word in the case of the unfactorizing system and the morphemes & the Pos-tags in the case of the factorizing system).

### 4.1. Disambiguation problem in the factorizing system

As mentioned in section two the Arabic word has two types of diacritization (morphological and syntactical). The same word structure (i.e. the same letters) has more than one morphological analysis that is required to be disambiguated. The morphologically diacritized word has more than one syntactical diacritics that are also required to be disambiguated. So the word diacritization is done in two levels: The first one is morphological diacritization disambiguation and the second one is syntactical diacritization disambiguation for the morphologically disambiguated word.

*Table 2* below, shows an example for the possible morphological and syntactical analysis for an Arabic word "الكتابات".

| الكتابات | | | | | | |
|---|---|---|---|---|---|---|
| Possible morphological analyses | | | | | Corresponding morphological diacritics | Possible syntactical diacritics |
| t | p | r | f | s | | |
| مُصَرَّفة<br>مُنْتَظِمَة | ال | كتب | فِعَال | ات | الْكِتَابَات | الْكِتَابَاتِ |
| | | | | | | الْكِتَابَاتُ |
| مُصَرَّفة<br>مُنْتَظِمَة | ال | كتب | فَعَال | ات | الْكَتَابَات | الْكَتَابَاتِ |
| | | | | | | الْكَتَابَاتُ |
| مُصَرَّفة<br>مُنْتَظِمَة | ال | كتب | فَعَّال | ات | الْكَتَّابَات | الْكَتَّابَاتِ |
| | | | | | | الْكَتَابَاتُ |
| مُصَرَّفة<br>مُنْتَظِمَة | ال | كتب | فُعَال | ات | الْكُتَابَات | الْكُتَابَاتِ |
| | | | | | | الْكَتَابَاتُ |

*Table 2*: Example for morphological and syntactical analyses.

### 4.2. Disambiguation problem in the unfactorizing system

In the unfactorizing system the word is completely diacritized (morphologically and syntactically) at the same time so the same word has more than one complete diacritization that is required to be disambiguated.

*Table 3* below shows an example for the possible diacritization for an Arabic word "الكتابات".

| The word | Possible diacritization |
|---|---|
| الكتابات | الْكِتَابَاتِ |
| | الْكِتَابَاتُ |
| | الْكَتَابَاتِ |
| | الْكَتَابَاتُ |
| | الْكَتَّابَاتِ |
| | الْكَتَابَاتُ |
| | الْكُتَابَاتِ |
| | الْكَتَابَاتُ |

*Table 3*: Example for word analyses.

### 4.3. Disambiguation technique

The example in *Table 2* and *Table 3* shows the possible diacritics of the word when the word is not in a context, but when the word becomes in a certain context it will give a higher probability to one solution over the others.

One of the effective techniques widely adopted today, namely "Bayes', Good-Turing Discount, and Back-Off" which is used here to estimate the most likely solution [2], [9].

Searching for the best path (the best sequence of analysis) through the solution space can easily be perceived as a tree search problem. Obviously, exhaustive search methods through this tree are infeasible due to its exponential complexity.

Fortunately, the well known best-first strategy can be employed that a list of open nodes (that are candidates of expansion on the next step) is maintained while the search, and the node to be expanded next, is the one that maximizes some likelihood function. So the A$^*$ search algorithm is applied here.

## 5. Results and evaluation

### 5.1. Data Base description:

- The training data we have consists of two parts:

  a. 800K words fully diacritized and fully POS tagged data. About 500k words of this 800k words is the NEMLAR annotated data [12] sold in ELDA (www.elda.org). This piece of data is mostly news domain. And the remaining part, from the internal RDI data, is mostly religious domain.

  b. About 2.5M words that are fully diacritized but they are not POS tagged. This data is totally religious domain.

- The test data consists of 11,079 words that are fully diacritized and fully POS tagged. This data consists of:

  a. 5422 words of religious domain.

  b. 3424 words news domain.

  c. 2233 words other different domains (science, sports, social, .. etc.).

- We calculated all our results by automatically matching between the results and the annotated test data. A check by human (linguist) was done to verify the results. The role of the linguist is just to be sure that the errors assigned by the automatic tool are correctly assigned the error (the mismatch of the automatic diacritics versus the previously annotated data) as morphological or syntactic. However, neither the human nor the automatic tool allow for the possibility of having more than one correct solution.

- We have designed many experiments by dividing the test data into 3 sets (Religious, News, and total) and three sizes of training data: the 800K only, the 800k + 1.2M words (fully diacritized but not POS tagged, religious domain), and 800k+ 2.5M words (fully diacritized but not POS tagged, religious domain).

- It is to be noted that the factorizing system needs fully annotated (POS tagging) data and cannot benefit from the only diacritized data.

- The factorizing system uses 10-gram with the A$^*$ search. The unfactorizing (UF) word analysis in the hybrid system uses 3-gram with A$^*$ search.

### 5.2. Results:

We have run few experiments. Each experiment will be described below with its motivation and results. Conclusions will be deducted from each experiment.

**Experiment 1:**

**Motivation:** we like to test the concept of the hybrid system versus the factorizing one.

We have already built a complete system based on factorizing the Arabic word. The results of this system were good in the morphological diacritization but not good at all for the syntactical diacritization. *Table 4* shows the comparative results for the two systems.

These conclusions could be deducted from this experiment:

i. The two systems have better performance with the increase of the data when we look to the morphological diacritization.

ii. The factorizing system did not improve with the syntactical diacritization even when we increased the size of the training data.

iii. The hybrid system outperforms the factorizing system in all the cases. The performance was much better in case of the syntactical diacritization. The hybrid system did improve with the increase of the training data.

iv. From our previous experience, there is always more than one solution in some words in both morphological and syntactical diacritization. We did not make this kind of revision for the results because this is a very time consuming process, and there is a good possibility to have this advantage for the two systems. This explains the difference between the results that is reported here and in other references for our factorized system [2] (the reported results were about 5% for our factorized system). Listening experiments confirms the superiority of the hybrid system over the factorizing one alone. (This site contains some TTS samples for the two systems http://www.RDI-eg.com/RDI/TTS-Samples).

| training Data size | Morphological Errors | | Syntactical Errors | |
|---|---|---|---|---|
| | Factorizing system | Hybrid system | Factorizing system | Hybrid system |
| 128k | 11.6% | 9.2% | 30.7% | 21% |
| 256k | 12.1% | 7.9% | 29.9% | 18.7% |
| 512k | 10.1% | 6.5% | 30.9% | 16.8% |
| 800k | 7.8% | 7% | 32.7% | 16% |

*Table 4*: the comparative results for the factorizing and the hybrid systems

**Experiment 2**

**Motivation:** from the last experiment we noticed that the performance is affected by the OOV percentage, so we have shown the performance of systems versus the OOV for the different domains.

From *Table 5* we can deduce that:

i. There is a clear correlation between better results and lower OOV.

ii. If we imagine that we could get enough diacritized data with negligible OOV (which might not be easy, but we like to predict the asymptotic performance of the system), the results will approach 1.5% (or a little less) morphological diacritization errors and 5% syntactical diacritization errors. (The performance of the unfactorized for the seen vocabulary only).

iii. The OOV could be considered a good reference for the unfactorized system performance; i.e. if we build a system that diacritize the complete words directly without any need to back off to the factorizing system, the errors of this system is partially from the OOV (the higher percentage) and from the internal errors for the seen vocabulary.

| training Data size | OOV | Test data domain | Morphological Errors | | Syntactical Errors | |
|---|---|---|---|---|---|---|
| | | | Unfactorizing system (seen vocabulary) | Hybrid system | Unfactorizing system (seen vocabulary) | Hybrid system |
| 800k + 2.5M | 13.3% | Religious domain | 1.8% | 3.7% | 5.5% | 11.3% |
| | 17.9% | News domain | 1.1% | 3.7% | 5% | 15.6% |
| | 13.7% | Total | 1.5% | 3.6% | 4.9% | 13.4% |

*Table 5*: studying the effect of the OOV on different domains

**Testing the systems performance:**

We have recorded the memory (for the language models) and the processing time needed for each system to evaluate the cost of the gain in results shown above.

- As shown in *Table 6*, there is some increase in the memory for the hybrid system compared to the factorizing one. The more data used by the hybrid system the more memory size is needed. The size of the memory increases linearly with the increase of the data size. This increase of the required memory is not that serious with nowadays computer resources. It is worth mentioning that the OOV percentage is decreasing by the increasing of the training data. The size of the dictionary formulated for the language model is shown in *Table 6*; it is clear that it is far from saturation.

| training Data size | Dictionary size (words) | OOV | Language model size (byte) | | |
|---|---|---|---|---|---|
| | | | Factorizing system | Unfactorizing system | Hybrid system |
| 64k | 21k | 37.4% | 15.7M | 2.3M | 18M |
| 128k | 34. 8k | 32.5% | 33.3 M | 4 M | 37.3M |
| 256k | 60.5k | 26.2% | 60.3 M | 8 M | 68.3M |
| 512k | 97.4k | 21.8% | 113M | 15.7 M | 128.7M |
| 800k | 136.6k | 20.1% | 167M | 24.2 M | 191.2M |
| 800k+1.2M | 216.8k | 15.8% | 167M | 46.3 M | 213.3M |
| 800k+2.5M | 260.9k | 13.7% | 167M | 60 M | 227M |

*Table 6*: studying the effect of the increase of the training data on the dictionary, the OOV, and the memory size.

- Regarding the time needed by the two systems; the hybrid system outperforms the factorizing system considerably. This is noticed in all the experiments; however we recorded one of these experiments as shown in *Table 7*. Our explanation for that is as follows: The hybrid system uses the unfactorized words which form a more compact set for the A$^*$ search than the factorizing system.

| | Testing time(min.)<br>(11079 words) | |
|---|---|---|
| training Data size | Factorizing system | Hybrid system |
| 800k | 21.5 | 9.7 |

*Table 7:* studying the time consumption by the factorizing and the hybrid systems

- It is worth mentioning that the system is built to allow the linguists, after analyzing the errors, to suggest rules from their experiences to help reducing the whole system errors to a minimum without any need to increase the resources needed (training data, memory, and with negligible increase in the processing time). This is the future work for the hybrid system.

## 6. Conclusion

It is clear from the above analysis that the hybrid system outperforms the factorizing alone in both the accuracy and the processing time with relatively little increase in the memory needed for the language models.

## Acknowledgement

The authors would like to thank RDI research lab for this research and all the other researchers in this lab for their continuous cooperation and support.

## References

### I- References in English

[1] M. Attia, M. Rashwan, A. Ragheb, M. Al-Badrashiny, H. Al-Basoumy, S. Abdou, *A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields*, Lecture Notes on Computer Science (LNCS): Advances in Natural Language Processing, Springer-Verlag Berlin Heidelberg; www.SpringerOnline.com, LNCS/LNAI; Vol. No. 5221, Aug. 2008.

[2] M. Attia, *Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications*, PhD thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, 2005.

[3] M. Attia, M. Rashwan, *A Large-Scale Arabic PoS Tagger Based on a Compact Arabic PoS Tags- Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words,* Proceedings of the Arabic Language Technologies and Resources Int'l Conference; NEMLAR, Cairo, 2004.

[4] M. Attia, M. Rashwan, G. Khallaaf, *On Stochastic Models, Statistical Disambiguation, and Applications on Arabic NLP Problems*, The Proceedings of the 3rd Conference on Language Engineering; CLE'2002, by the Egyptian Society of Language Engineering (ESoLE); www.ESoLE.org.

[5] M. Attia, *A Large-Scale Computational Processor of the Arabic Morphology, and Applications*, M.Sc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

[6] V. Cavalli-Sforza, A. Soudi, T. Mitamura, *Arabic Morphology Generation Using a Concatenative Strategy*, ACM International Conference Proceeding Series; Proceedings of

the first conference on North American chapter of the Association for Computational Linguistics (ACL), 2000.

[7] Asd J. Dichy, M. Hassoun, the *DINAR.1 (DIctionnaire INformatisé de laree, version 1) Arabic Lexical Recourse, an outline of contents and methodology*, The ELRA news letter, April-June 2005, Vol.10 n.2, France.

[8] D. Jurafsky, J. H. Martin, *Speech and Language Processing; an Introduction to Natural Language Processing, Computational Linguistics, and Speech Processing*, Prentice Hall, 2000

[9] S. M. Katz, *Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-35 no. 3, March 1987.

[10] A. Ratenaparkhi, *Maximum Entropy Models for Natural Language Ambiguity Resolutions*, PhD thesis in Computer and Information Science, Pennsylvania University, 1998.

[11] H. Schütze, C. D. Manning, *Foundations of Statistical Natural Language Processing*, the MIT Press, 2000.

[12] M. Yaseen, et al., *Building Annotated Written and Spoken Arabic LR's in NEMLAR Project*, LREC2006 conference http://www.lrec-conf.org/lrec2006, Genoa-Italy, May 2006.

**II- References in Arabic**

[13] (A. Arragehy, 1993) التَّطبيقُ الصَّرْفيُّ، عَبْدُهُ الرَّاجِحيّ، دارُ المَعْرِفَةِ الجَامِعِيَّةِ، الإسْكَنْدَرِيَّةُ، □□□□م.

# Word Sense Disambiguation in Machine Translation Using Monolingual Corpus

**Ola M. Ali**
Faculty of Computer and Information Sciences,
Department of Scientific Computing,
Ain Shams University
olaforn@yahoo.com

**Mahmoud GadAlla**
Military Technical Collage,
Computer science Department
mahmoud_mtc@yahoo.com

**Mohammad S. Abdelwahab**
Faculty of Computer and Information Sciences,
Department of Scientific Computing,
Ain Shams University
m_s_wahab@fcisainshams.edu.eg

## Abstract

This paper presents dictionary-graph based Word Sense Disambiguation in Machine Translation system which translates Arabic Noun Phrase into English. This system uses Arabic-English dictionary and monolingual corpus of the target language to solve the ambiguity in the translation process. The bi-grams of the target language monolingual corpus was first computed and then was used with Viterbi search algorithm to find the appropriate translation of the Arabic Noun Phrase. This work compares the results of bi-gram with Simple Interpolation Smoothing as a baseline with the results of five methods of statistical measures of association which is used to rank bi-grams. The experiments show an improvement in accuracy when using the methods of statistical measures of association. The experiments are discussed with its results.

**Keywords**: Word Sense Disambiguation, Machine Translation, Arabic Noun Phrase Monolingual corpus.

## I. Introduction

Translation of Arabic sentences is a difficult task. The Arabic sentence is complex and syntactically ambiguous due to the frequent usage of grammatical relations, order of words and phrases, conjunctions, and other constructions. One of the most difficult problems in developing high-accuracy translation systems for Arabic is the predominance of non-diacritized text material. The absence of diacritics, which represent most vowels, in the written text creates ambiguity which hinders the development of Arabic natural language processing applications such as Machine Translation (MT). Consequently, most of the researches in Arabic (MT) mainly concentrated on the translation from English to Arabic. Word Sense Disambiguation (WSD) in (MT) is required to carry out the lexical choice in the case of semantic ambiguity during the translation, i.e., the choice for the most appropriate translation for a source language word when the target language offers more than one option, with different meanings, but the same part-of-speech.

WSD approaches are classified into supervised, dictionary based and unsupervised approaches [1]. The distinction is that with *supervised* approaches we know the actual status (here, sense label) for each piece of data on which we train, whereas with *unsupervised* approaches we do not know the classification of the data in the training sample. Because the production of labeled training data is expensive, people will often want to be able to learn from unlabeled data but will try to give their algorithms a head start by making use of various *knowledge sources*, such as dictionaries, and preprocessed bilingual or monolingual corpus. *Dictionary-based* approaches rely on the definition of senses in dictionaries and thesauri. Most WSD approaches disambiguate each word in isolation. Graph-based WSD is one of unsupervised approaches which connects words in a sentence to solve disambiguation. Graph is a natural way to capture connections between entities.

This paper investigates a dictionary-graph based approach which based on translations in a second-language corpus. The second language here is the target language of an Arabic-English translation system. Viterbi algorithm is used to search the graph of all possible translation words to find an appropriate translation of the input Arabic NP. This investigation is part of an on-going MT system, which uses minimal resources for both the source and the target language to translate Arabic NP into English. Figure1 shows the structure of the translation system.

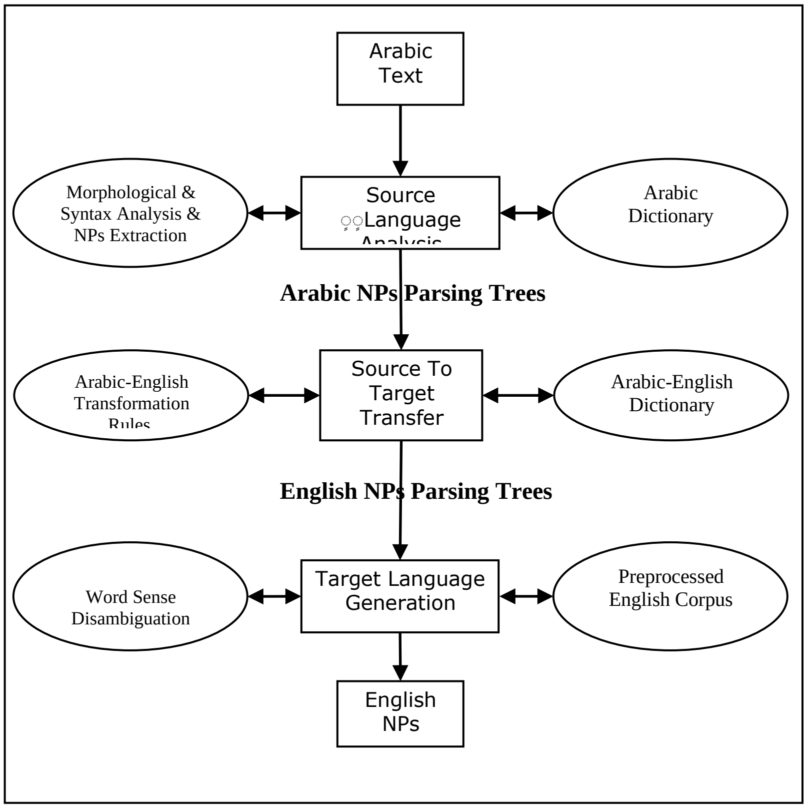

Arabic Text
Morphological & Syntax Analysis & NPs Extraction
Source Language Analysis
Arabic Dictionary
Arabic NPs Parsing Trees
Arabic-English Transformation Rules
Source To Target Transfer
Arabic-English Dictionary
English NPs Parsing Trees
Word Sense Disambiguation
Target Language Generation
Preprocessed English Corpus
English NPs

Figure 1 the architecture of the Arabic to English MT system

The system consists of three components: Source Language Analysis, Source to Target Transfer, and Target Language Generation. The WSD approach is part of the third component of the translation system.

The rest of the paper is structured as follows. In, Section II, we present an overview of the related works. Section III we describe our disambiguation approach. In Section V, we describe our experiments and results,  In Section IV, we give some concluding remarks and future directions.

## II. Related work

Usually, the larger the available training corpus, the better the performance of a translation system. Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and for some language pairs is not even possible. In addition, small corpora have certain advantages: the possibility of manual creation of the corpus, possible manual corrections of automatically collected corpus, low memory and time requirements for the training of a translation system, etc. Therefore, the strategies for exploiting limited amounts of bilingual data are receiving more and more attention.

Statistical machine translation of Spanish-English and Serbian-English language pairs were investigated in [9]. They used sparse training data in translation process. The morpho-syntactic transformations have been implemented as a preprocessing step, therefore modifications of the training or search procedure were not necessary.

For MT purposes, however, the context may include the translation in the target language, i.e., words in the text which have already been translated. Although intuitively plausible, this strategy has not been explored specifically for WSD. On the other hand, some related approaches have exploited similar strategies for other purposes. For example, some approaches for MT which are similar to ours, make use of the words which have already been translated as context, implicitly accomplishing basic WSD during the translation process [3].

[8] investigated the sense inventory discrepancies for English-Italian,[10] for English-Hungarian, [12] for English-Hindi, and [7] for English-Portuguese. They show that there is not a one-to-one relation between the number of senses in the source language and their translations into another language. More specifically, they show that many source language senses are translated into a unique target language word, while some senses need to be split into different translations, conveying sense distinctions that only exist in the target language.

Some approaches for monolingual WSD use techniques to gather co-occurrence evidence from bilingual corpus in order either to carry out WSD [14] ,[7], or to create monolingual sense tagged corpora [4]. Other monolingual related approaches somehow explore the already disambiguated or unambiguous words by taking into account the senses of the other words in the sentence in order to disambiguate a given word [11], [2],and [5].

## III. Disambiguation Approach

Disambiguation in MT aims to select the correct translation in the target language for an ambiguous item in the source language, based on its context in the translation unit. The proposed system is a dictionary-graph based approach which is a combination of dictionary-based and graph-based WSD approaches. It makes use of words which have already been translated as context, implicitly accomplishing basic WSD during the translation process. It uses a one to many Arabic-English dictionary which gives all possible translation for each word in the input Arabic NP. Statistical analysis of the target language corpus is used to get the bi-grams of the words of the English corpus. In order to solve the ambiguity in the translation of the Arabic NP , the translation system identify the ambiguous words and use the viterbi search algorithm to find the appropriate translation of the Arabic words to generate the target English NP. Section 1 describes the dictionary-based approach which is used. Section 2 discusses the graph-based approach.

### 1 Disambiguation based on translations in a second-language corpus.

The Disambiguation based on translations in a second-language corpus is one of dictionary-based WSD approach[1]. The basic idea of this approach is best explained with the example in table1.

Table1 How to disambiguate *interest* using a second-language corpus.

| | Sense1 | Sense2 |
|---|---|---|
| Definition | legal share | Attention, comcern |
| Translation | Beteiligung | *Interesse* |
| English Collocation | *Acquired an  interest* | *show interest* |
| Translation | *Beteiligung  erwerben* | *Interesse zeigen* |

English word *interest* has two senses with two different translations in German. Sense 1 translation is *Beteiligung* and Sense 2 translation is *Interesse.* Suppose *interest* is used in the phrase *showed interest.* The German translation of *show,* ‘zeigen,’ will only occur with *Interesse* since “legal shares” are usually not shown. We can conclude that *interest* in the phrase *to show interest* belongs to the sense *attention, concern.* On the other hand, the only frequently occurring translation of the phrase *Acquired an  interest* is *eine erwerben Beteiligung*, since *interest* in the sense ‘attention, concern’ is not usually acquired. This tells us that a use of *interest* as the object of *acquire* corresponds to the second sense, “legal share”. A simple implementation of this idea is shown in [1] as:

```
Comment: Given: a context $c$ in which $w$ occurs in relation $R(w,v)$
for all senses $sk$ of $w$ do
    score($sk$)= $|\{c \in S \mid \exists w' \in T(sk), v' \in \tau(v) : R(w'.v') \in c\}|$
end
choose s’ = argmax ,, score(sk)
```

For the above example the relation *R* is ‘is-object-of’ and the goal would be to disambiguate *interest* in *R (interest, show).* To do this, we count the number of times that translations of the two senses of *interest* occur with translations of *show* in the second language corpus. The bi-gram of *R(Interesse, zeigen)* would be higher than the bi-gram of *R (Beteiligung, zeigen), so we* would choose the sense ‘attention, concern,’ corresponding to Interesse.

**2 Graph-based WSD.**

Graph-based WSD was introduced in [15].The goal of graph-based approach to WSD is to utilize relations between senses of various words that represented in a graph. Given a sequence of words W = {w1,...,wn},and a set of admissible labels  Lwi ={ $l^{1}_{wi}$,…., $l^{Nwi}_{wi}$ }. These words and labels are defined in a weighted graph G(V,E) such that V is the set of nodes in the graph, where each node corresponds to a word/label assignment $l^{j}_{wi}$ and E is the set of weighted edges that capture dependencies between labels. These weights in our approach are the bi-grams of each two consecutive words. Figure2 shows an example of constructed graph.

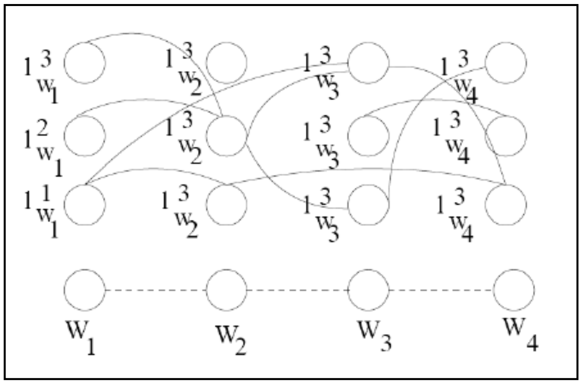


Figure2 shows Example of Constructed Graph

This method disambiguates the words by computing the most likely sequence of words that gives the maximum probability. Viterbi algorithm is a technique which efficiently computes the most likely state sequence [1]. Viterbi algorithm for finding optimal sequence of senses described in [18] as:

```
function VITERBI(observations of len T,state-graph) returns best-path

    num-states←NUM-OF-STATES(state-graph)
    Create a path probability matrix viterbi[num-states+2,T+2]
    viterbi[0,0]← 1.0
    for each time step t from 1 to T do
       for each state s from 1 to num-states do
          $viterbi[s,t] \leftarrow \max_{1\le s'\le num-states} [viterbi[s',t-1]*a_{s',s}]*b_s(o_t)$
          $back-po\text{int}er[s,t] \leftarrow \arg\max_{1\le s'\le num-states} [viterb[s',t-1]*a_{s',s}]$
    Backtrace from highest probability state in final column of viterbi[] and return path
```

Given a graph of nodes and weighted edges the algorithm returns the state-path through the graph which assigns maximum likelihood to the observation sequence. $a[s', s]$ is the transition probability from previous state $s'$ to current state $s$, and $b_s(o_t)$ is the observation likelihood of $s$ given $o_t$ . Note that states 0 and N+1 are non-emitting start and end states.

## IV. Experiments and Results

A combination of Brown and English Treebank corpus which are available in the Natural Language Toolkit (NLTK) [17] is used as a target language corpus. A corpus of about 1 million words are used and analyzed statistically to get the bi-grams of the words of the corpus. We used the Ngram Statistical Package (NSP) [16]. This package is a set of perl programs that analyze Ngrams in text files. One of these programs takes as input a list of Ngrams with their frequencies and runs a user-selected statistical measure of association to compute a "score" for each Ngram. The Ngrams, along with their scores, are output in descending order of this score. The statistical score computed for each Ngram can be used to decide whether or not there is enough evidence to reject the null hypothesis (that the Ngram is not a collocation) for that Ngram.

The statistical measures of association (SMA) which provided are: dice, log-likelihood , mutual information, t-score, and the left-fisher test of associativity.

Table2 the accuracy of the statistical measures of association method.

| Corpus size | Bi-gram baseline | Dice | Log-likelihood | Mutual information | T-score | Left-Fisher |
|---|---|---|---|---|---|---|
| 53649 words | 54.9% | **55.4%** | **55.4%** | 54.9% | **55.4%** | **55.4%** |
| 138205 words | **57.3%** | 56.6% | **57.3%** | 55.9% | **57.3%** | 56.6% |
| 1171868 words | 60.1% | **63.8%** | 63.4% | 62.0% | 62.9% | 62.0% |

Automatic evaluation systems are often criticized for not capturing linguistic subtleties. This is clearly apparent in the field’s moving back toward using human evaluation metrics. We conducted a human evaluation of nouns and adjectives realization in a document contained 190 noun phrases. These noun phrases consist of 969 words from them 213 words are ambiguous. We compared bi-gram with Simple Interpolation Smoothing as a baseline with five bi-gram scoring methods, dice, log-likelihood, mutual information, t-score, and the left-fisher test of associativity. The evaluation was conducted using one bilingual Arabic-English speaker (native Arabic, almost native English). The task is to determine for every ambiguous word that appears in the Arabic input NP whether it is realized or not in the English translation with the correct sense. The results are presented in Table 2 .

The results show that the accuracy in WSD increases with the incensement of the corpus size in all methods. This means that the size of corpus is a very important factor.  The results of SMA was better than the results of bi-gram only. Dice method shows the higher accuracy in the final experiments.

## V. Conclusion and Future Work

In this paper we defined a dictionary-graph based Word Sense Disambiguation approach whish is uses in a Machine Translation system that translates Arabic Noun Phrase into English. This approach makes use of words which have already been translated as context, implicitly accomplishing WSD during the translation process. Virirbi search algorithm was used to get the appropriate translation of the input Arabic NP.

We compared the results of bi-gram with Simple Interpolation Smoothing as a baseline with five bi-gram scoring methods, dice, log-likelihood, mutual information, t-score, and the left-fisher test of associativity. Manual evaluation of  213 ambiguous words in 190 NPs was accomplished for all scoring methods.We fined that the accuracy of all methods increased when the size of the corpus increased. The final experiment with the largest corpus showed that dice method is the statistical measures of association method which gave the highest accuracy. It improved the WSD accuracy by 3.6%.

To improve the translation process in future work we need larger corpus. Using the morphological features with the translated words may also improve the out put.An interesting question is whether these results will improved if similarity between words is used in stead of bi-gram and statistical approaches. In the forthcoming work we will investigate their validity in Word-Net Similarity modules. The complete translation system will be accomplished and evaluated using automated evaluation metrics like BLEU [6] and NIST [13].

## References

[1] C. D. Manning and H. Schutze, “Foundations of Statistical Natural Language Processing,” the MIT press, 1999.

[2] G. Hirst. (1987). Semantic Interpretation and the Resolution of Ambiguity. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

[3] I. Dagan and A. Itai . (1994). Word Sense Disambiguation Using a Second Language MonolingualCorpus. *Computational Linguistics*, 20, pp. 563-596.

[4] J. Fernández, M. Castilho, G. Rigau, J. Atserias and J. Turmo. (2004). Automatic Acquisition of Sense Examples using ExRetriever. LREC, Lisbon, pp. 25-28.

[5] J. Cowie, J.A. Guthrie, and L. Guthrie. (1992). Lexical Disambiguation Using Simulated Annealing. COLING, Nantes, pp. 359-365.

[6] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. (2002). BLEU: a method for automatics evaluation of machine translation. In *Proceedings of the 40th ACL meeting*, pp. 311-318.

[7] L. Specia, G. Castelo-Branco, M.G.V. Nunes and M. Stevenson. (2006). Multilingual versus Monolingual WSD. To appear in Workshop Making Sense of Sense, EACL, Trento.

[8] L. Bentivogli, P. Forner, and E. Pianta. (2004).Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus. COLING, Geneva, pp. 364-370.

[9]M. Popović and H. Ney. (2006) Statistical Machine Translation with a Small Amount of Training Data. In *Proceedings of the 5th LREC SALTMIL Workshop on Minority Languages*, pages. 25-29, Genoa, Italy.

[10] M. Miháltz..(2005). Towards A Hybrid Approach to Word-Sense Disambiguation in Machine Translation. Workshop Modern Approaches in Translation Technologies, RANLP, Borovets.

[11] M. Lesk. (1986). Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. SIGDOC, Toronto, pp. 24-26.

[12] N. Chatterjee, S. Goyal and A. Naithani. (2005). Pattern Ambiguity and its Resolution in English to Hindi Translation. RANLP, Borovets, pp. 152-156. Co

[13] National Institute of Standards and Technology. (2001). *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. http://www.nist.gov/speech/tests/mt/doc/ngramstudy.pdf

[14] R. Mihalcea and D.I. Moldovan (1999). A Method for Word Sense Disambiguation of Unrestricted Text. 37th ACL, Maryland.

[15] R. Mihalcea. (2005). Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Human Language Technology / Empirical Methods in Natural Language Processing conference,* Vancouver.

[16] S. Banerjee and T. Pedersen.( 2003). The Design, Implementation and Use of the Ngram Statistics Package. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, ,2003February, Mexico City.

[17] Bird, Steven and E. Loper (2006). Natural Language Toolkit. http://nltk.sourceforge.net/

[18] D. Jurafsky and J. H. Martin, “Speech and Language Processing: An introduction to natural language processing,computational linguistics, and speech recognition”. Prentice Hall, 2000, 934 pages.

# Designing and Implementing Arabic Sign Language (ArSL) for Deaf Peoples

**Hassanin M. Al-Barhamtoshy, Sami M. Halawani and Sakher F. Ghanem**

King Abdel Aziz University, Jeddah, Saudi Arabia

## Abstract

The purpose of this paper is to design and implement a signed model to understand and deal with those individuals, who are deaf. Therefore, one of the important point to use this paper is to develop dictionaries (Arabic to Sign), and the representation of information to understand the words automatically as the key to resolving many issues of deaf peoples. Consequently, the paper will present an interactive software, using avatars (3D character module), to process and translate free Arabic words/sentences to Arabic Sign Language (ArSL). The ArSL is a visual natural language model to be used by many deaf people in Arabic countries. Such natural language is based on morphological, and partially syntactic and semantic analysis.

A Prototype for teaching Arabic numerals, basic mathematical operations, some basic words and Arabic text translation to Arabic sign language will be designed, developed and tested on deaf students. This paper could be valuable, as a teaching tool, by increasing: (1) the opportunity for deaf children to learn Arabic numbers and letters via interactive media; (2) the effectiveness of ArSL teachers.

Consequently, facilitating communication with deaf people, using Computer Technology is the main goal of this paper and it is necessary to give deaf people some of their rights in their society.

## 1 Introduction

Sign languages can be recognized into four elements [Abdel-Fattah, 2005]: Hand shape (Configuration of hands), Position of hand with body, Hand movement, and Non-manual features (face, mouth and tongue).

Due to the specialty and variety of Arabic languages, additional feature be included: Time, tense and duration of verbs (past, present, and imperative), Gender (male, female, and neutral), and Number (single, double, and plural). Consequently, these additional features must be considered during the design and the implementation of the proposed model.

This research describes a system for generating natural-language sentences from syntax and lexical structures, taken into our point of view an internal (or interlingual) representation. Such model will be developed as part of an English-Arabic Machine Translation (MT) system; however, it is designed to be used for many other MT language pairs and natural language applications.

The following subsections introduce to the previous works in sign languages. Therefore British Sign Language (BSL), American Sign Language (ASL), French sign language, and Arabic sign language will be introduced.

ASL is linguistic structure distinct from English – used for communication to approximately one half million deaf people in the United States (Neidle et al., 2000, Liddell, 2003; Mitchell, 2004). Technology for the deaf rarely addresses this field; so, many deaf people find it difficult to read text on electronic devices. Software tools for translating English text into animations of a computer-generated character performing ASL can make a variety of English text sources accessible to the deaf, (Huenerfauth, 2004). Language processing and machine translation (MT) can also be used in educational software for deaf children to help them improve their English literacy skills.

There are papers describe the design of English to- ASL MT system (Huenerfauth, 2004, 2003), describing ASL generation. This overview illustrates important correspondences between the problem of ASL natural language generation (NLG) and related research in Multimodal NLG.

The objective of this proposal report is designing and implementing TRGM model to accept an Arabic text, analyze and parse such input, represent this analysis into an interlingua semantic representation, then employee NLP techniques and therefore generate sign language synthesis stages.

Section 2 briefly describes relevant aspects of sign languages, which challenge a translation system. Sections 3 and 4 are devoted to the overall text processing architecture, text analysis and text understanding. Consequently, the proposed model will be presented. The subsections describe the syntactic parsing, translation to the interlingua semantic representation, and the pronoun resolution stage respectively. Current progress in the realization of the natural language component is also outlined in Section 3. During the designing and the implementation of the proposed model, we will review our experience with constructing dictionary, corpus and lexicon. Also, the research discusses the part of speech tags (POS).

## 1.1.1. British Sign Language (BSL)

The following subsections describe the BSL features.

### (a) Sign Order

BSL has a topic-comment structure, in which the subject or topic is signed first. The topic is the framework within which the predication takes place. After the topic has been identified, the rest of the sentence is the comment, the new information on it. Furthermore BSL has no fixed order of basic elements (Subject, Verb, Object). This flexibility is due to the extra information carried in the directional verbs (see later) and eye-gaze (Eva S´af´ar  and Ian Marshall, 2002).

### (b) Signing Space, Placement and Pronouns

In many sign languages and in BSL, signers exploit the signing space in front of their body. In a discourse components of a description can be situated in that space: first the area is defined and then all items or actions are related to that area. Thus, BSL has more pronouns than English, which are articulated by pointing to a location previously associated with a noun. This means that English is underspecified when using plural pronouns.

### (c) Sign Agreement Verbs

Agreement verbs include the information about person and number of the subject and object. This is realized by moving the verb in the syntactic space, in which the subject and the object are placed around the signer (Eva S´af´ar  and Ian Marshall, 2002). The signing of the verb begins at the position of the subject and ends at the position of the object (GIVE, TELL, etc), some verbs begin at the object and finish at the subject (BORROW).

### (d) Sign Classifiers

Classifiers are hand-shapes that can denote an object from a group of semantically related objects. They are used with verbs which require a classifier so that when combined with location, orientation, movement and non-manual features the composite forms a predicate. The hand-shape is used to denote a referent from a class of objects that have similar features (J. R. Kennaway. 2001 and Kopp, S., Tepper, P., and Cassell, J. 2004).

### (e) Sign Tenses

BSL has no tense system. Rather than express temporal information by morphological or syntactic features associated with verbs, it is expressed with the help of four time lines in the signing space or by the ordering of the propositions in the discourse.

## 1.1.2. American Sign Language (ASL) and English Linguistic

In many of sign languages, especially in ASL, several parts of the body convey meaning in parallel: hands (location, orientation, shape), eye gaze, mouth shape, facial expression, head-tilt, and shoulder-tilt.

Signers may also interleave lexical signing (LS) with classifier predicates (CP) during a performance. During LS, a signer builds ASL sentences by syntactically combining ASL lexical items (arranging individual signs into sentences). The signer may also associate entities under discussion with locations in space around their body; these locations are used in pronominal reference (pointing to a location) or verb agreement (Huenerfauth, 2004).

During generation, signers' hands draw a 3D scene in the space in front of their torso. One could imagine invisible placeholders floating in front of a signer representing real-world objects in a scene. To represent each object, the signer places his/her hand in a special hand shape (used specifically for objects of that semantic type: moving vehicles, seated animals, upright humans, etc.). The hand is moved to show a 3D location, movement path, or surface contour of the object being described.

The following sub sections describe four systems to translate from English text to American Sign Language (ASL). Therefore, the following subsections introduce the four systems under consideration:

- The VisiCAST translator (Marshell & Safar, 2002) and (Bangham, 2000),
- The ZARDOZ system (Veale eta al., 1998),
- The Workbench (Speers, 2001), and;
- The TEAM system (Zhao et al., 2000).

Such systems take into consideration the following terms: Machine Translation (MT) architecture, Grammar formalisms, Linguistic representations, Lexicon Format, Grammatical rules, and; Developing time.

### (a) The VisiCAST Translator

The VisiCAST introduced as a part of European Union's (EU), the university of East Anglia implemented a system for translating from English text into British Sign Language (BSL: Marshell, 2002). The approach CMU link parser to analyze an input English text and uses prolog grammar rules to convert this output into discourse representation structure. Therefore, head driven phrase structure rules are used to produce symbolic sign language representation script. This script is defined as "signing gesture markup language", and it is based on scheme of movement required to perform natural sign language (Kennaway, 2001).

The ViSiCAST is an EU Framework V supported project which builds on work supported by the UK Independent Television Commission and Post Office. The project develops virtual signing technology in order to provide information access and services to Deaf people (´Eva S´af´ar and Ian Marshall, 2002).

### (b) The ZARDOZ system

This system was proposed to translate English text to sign language using set of hand coded schema, as an Interlingua representation (Huenerfauth, 2003). The authors were developing their framework with British, Irish and Japanese sign language.

### (c) The ASL Workbench

This system is based on lexical functional grammar (LFG) to analyze English text, then transfer rules is used to converting an English f-structure into ASL output. Sometimes, the system encounters difficulties in analysis or other translation tasks; an additional step is employed to ask the user of the system for advice.

### (d) The TEAM Project

At university of Pennsylvania, TEAM project is employed to build an ASL syntactic structure from English text depending on tree during analysis (Zhao, 2000).

## 3 ArSL System Architecture

This section presents Arabic Sign Language (ArSL) design model, and discusses its structures and related grammars. Due to peculiarities and specialties of some grammar rules in Arabic language [Al-Barhamtoshy, 1992; El-Samahy, 1993], ArSL basically includes grammars of spatial signs. As a matter of fact, many grammatical concepts used to describe written text and/or spoken language may be inadequate to describe a sign language.

In many sign languages, they do not follow the same order of their spoken or written patterns. In general, a reversed order is used, that is because sign languages are thematized and more pragmatic than the spoken ones [Abdel-Fattah, 2005].

An Arabic sign language is based on content signs (phrase or sentence contents: those representing nouns, verbs, adverbs, prepositions …etc). As described in many literatures [Abdel-Fattah, 2005; Kamel, 1999], Arabic words can be classified into many types: noun, verb, special characters, determiner, preposition, adjective, ant etc.

The following figures show ArSL signs: noun sign in figure (1), verb sign in figure (2), preposition in figure 3, adjective in figure (4)… etc. There are relations between some of these types, according to the position, place and/ or frequency.

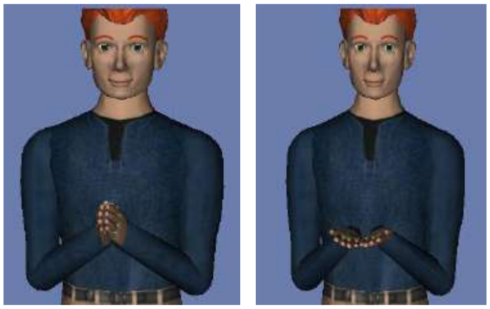

a b

Fig 1: Noun ( كتاب: Book).

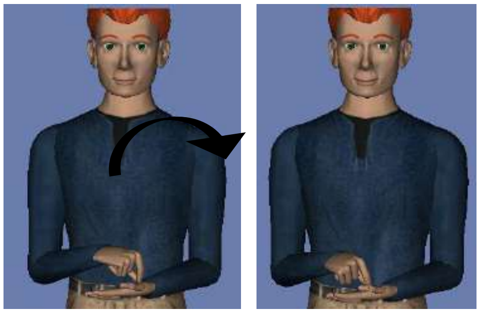

a b

Fig 2: Verb ( يكتب: Write).

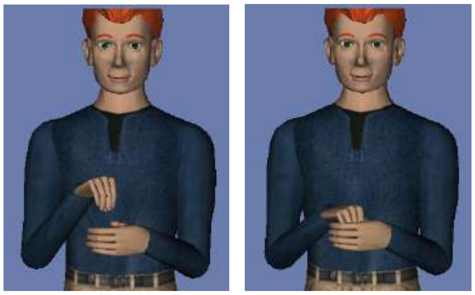

a b

Fig 3: Special Character (فى: in).

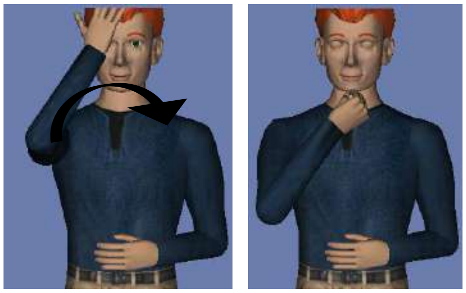

a b

Fig .4: Adjective (جميل: beautiful).

Some of such types can be included into intensifier (adverb) model, (e.g.; every day: repeating the sign to show frequency), as shown in Figure 5. However, the relationships and concepts can be represented by prepositions and intensifiers [Abdel-Fattah, 2005; Kamel, 1999].

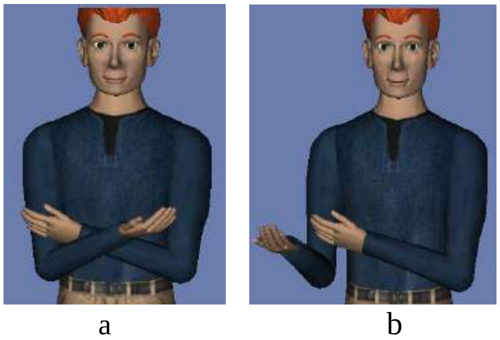


Fig 5: Repeating (كل صباح: every morning).

### 3.1 Verbs in ArSL

Verbs are the heart of a statement, in many languages. Consequently, we can say that verbs are used to encode meanings related to actions and states. Orientation and location of signs may be combined to contribute information about subject and/or object of verbs.

In the case of an Arabic verb, tense is used, therefore, past, present and future times are indicated at the beginning/ ending of a conversation scene. Therefore, the rule of a signed verb can be expressed as (see Figure 6):

Verb-sign ➔ <tense><signed-verb> | <signed-verb><tense>

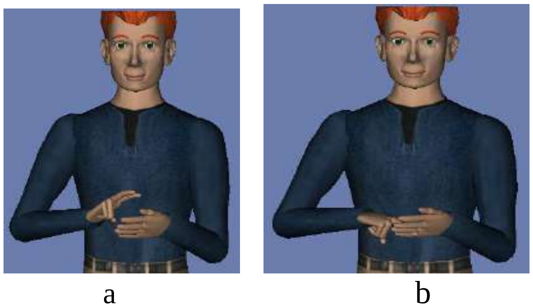


**Fig 6: Sign of verb (إقرأ: read).**

In many sign languages, there is no difference between a verb in the past, present and future tenses. As an example, Sign Smith Studio shows the sign of write, wrote and written as shown in Figures 7, 8 and 9, respectively.

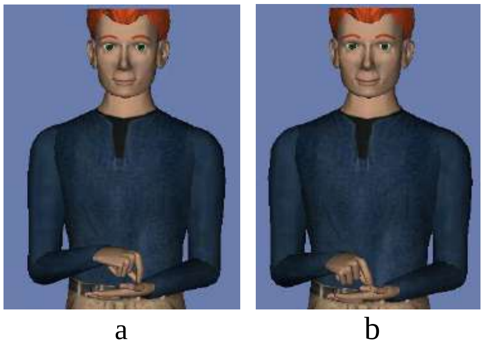


Fig 7: Sign of Verb ( يكتب: Write).

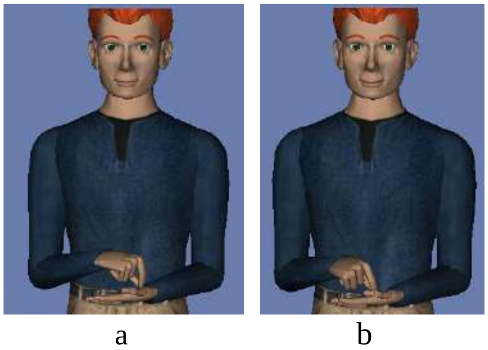


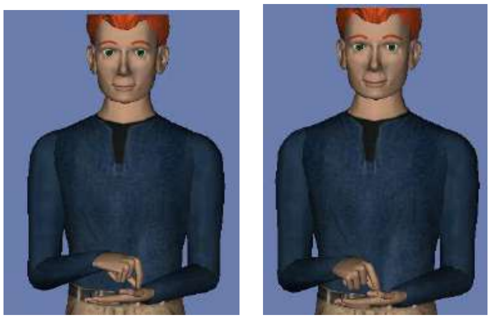
Fig 8: Sign of Verb ( كتب: Wrote).

For the past verb, an additional sign (like yesterday) is used. Therefore, the verb (Wrote كتب) is expressed as the following rule:

Verb-sign (كَتَبَ) ➔ <signed-verb (كتب) > <signed-word (أمس) >

Also, the present verb uses signed words to indicate present (like: now الآن , will سوف). Consequently, the verb (go يذهب) can be ruled by:

Verb-sign (يذهب) ➔ <signed-verb (ذهب) > <signed-word (الآن) >



a b

Fig 9: Sign of Verb ( كُتِبَ: Written).

In case of imperative tense, the signed verb is correlated with the pronouns (you أنتن – أنتم, you أنتَ أنتِ –, I أنا). In such cases, the verb (go إذهب) can be expressed as:

Verb-sign (إذهب) ➔ <signed-verb (يذهب) > <signed-word (أنتَ) >

## 3.2 Rules of Negative / Interrogatives

In case of negative and interrogatives, the sign can be expressed in more than one way of expression, as shown in Figures 10 and 11. Therefore, the negative rule can be expressed as:

Negative-sign ➔ <negative><signed-verb>

Note that <negative> can be {(لا , لم , لن , … )}. And, the interrogatives rule can be expressed as the following:

Intro-sign ➔ < ? ><signed-word> | <signed-word>< ? >

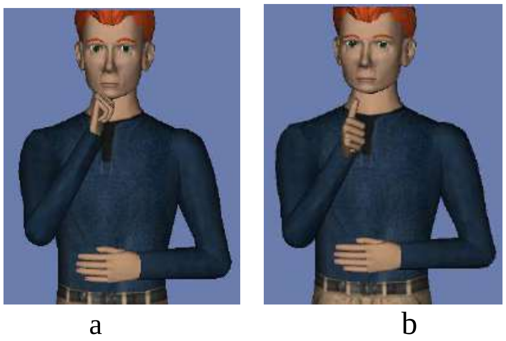


Fig 10: Sign of Negative ( لا: not).

In case of present, the sign of word (لا) usually is used. But, in the past tense, the sign of word (لم) is used. Whereas, the sign of word (لن) is used for future. See Figures 3.10 and 3.11.

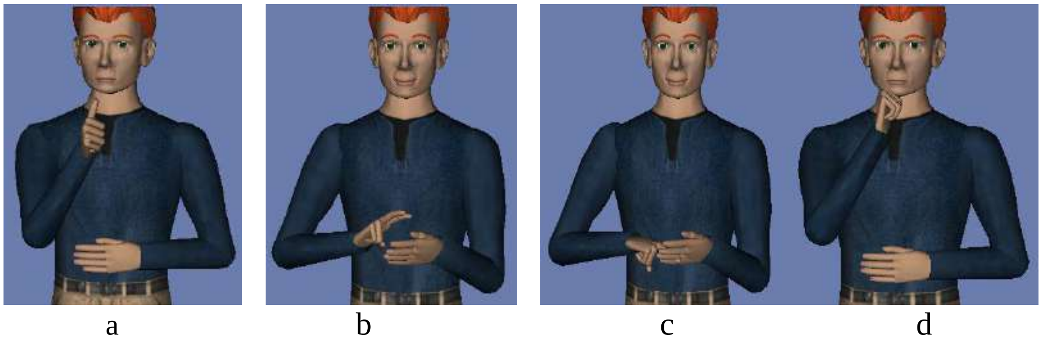


**Fig 11: Sign of Negative ( لا يقرأ : not read).**

## 3.3 Rule of Nouns

In many signs, nouns can be expressed as a word-to-sign image, see Figures 3.12-3.17.

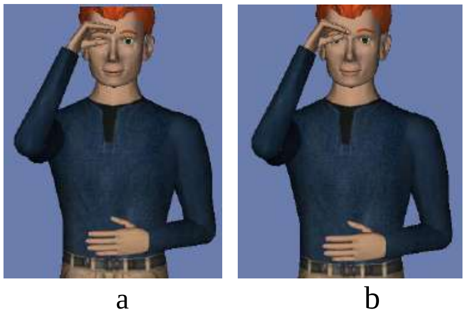


Fig 12: Sign of Noun ( ولد: Boy).

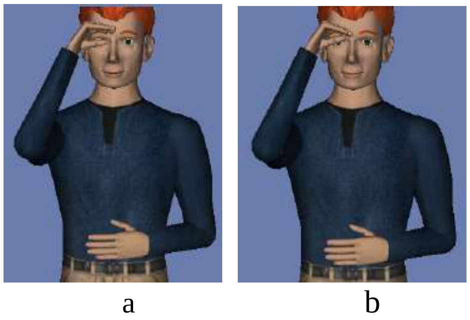


Fig 13: Sign of Noun ( أولاد: Boys).

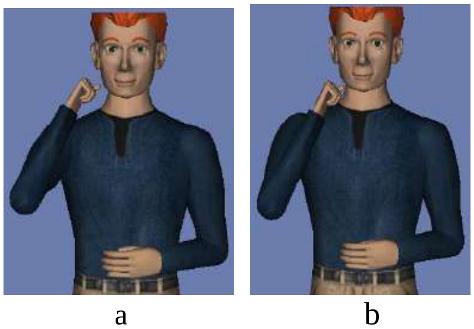


Fig 14: Sign of Noun ( بنت: Girl).

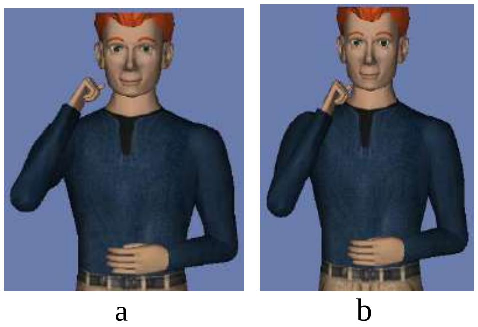


Fig 15: Sign of Noun ( بنات: Girls).

As a matter of fact, plural cannot be expressed in many sign languages. In Arabic, there are single, dual (double) and plural. Consequently, ArSL needs to express these rules.

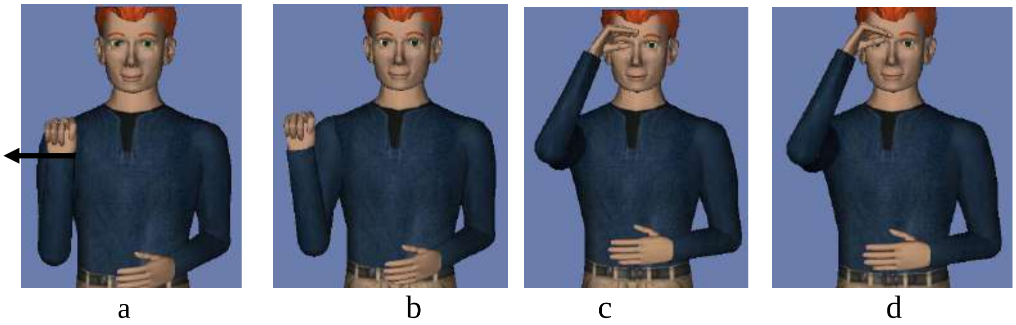


Fig 16: Sign of Noun ( الولد : The Boy).

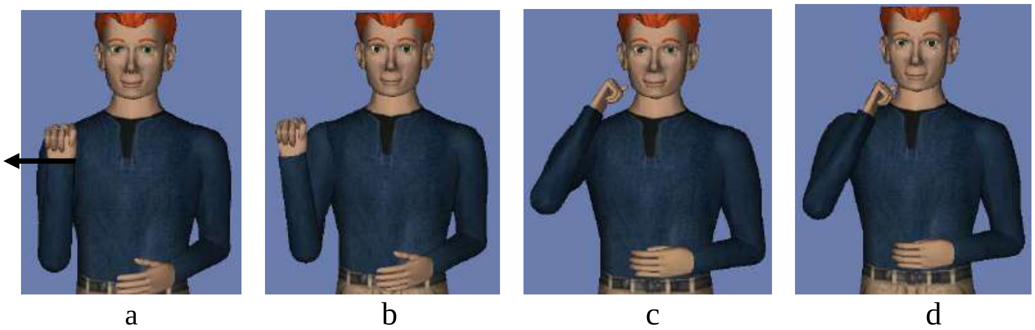


Fig 17: Sign of Noun ( البنت : The Girl).

Therefore, the general rule for nouns in ArSL is as the following:

Noun-sign ➔ <signed-noun><number> | <number><signed-noun> | <signed-noun>

Generally, plural can be expressed as a lot of things; therefore, the previous rule can be explained as the following rule:

Noun-sign (أسماك) ➔ <signed-noun (سمك)>< signed-word (كثير)>

## 3.4 Rule of Adjectives

Adjectives are also expressed by a word-to-sign image, see Figures 18 and 19.

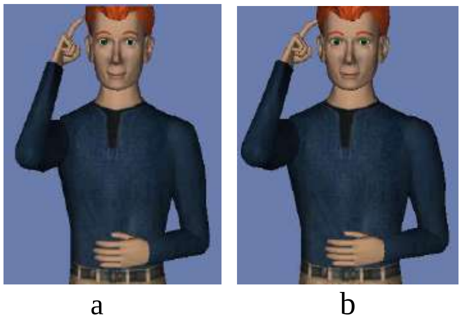


Fig 18: Sign of Adjective
( ذكي: Intelligent).

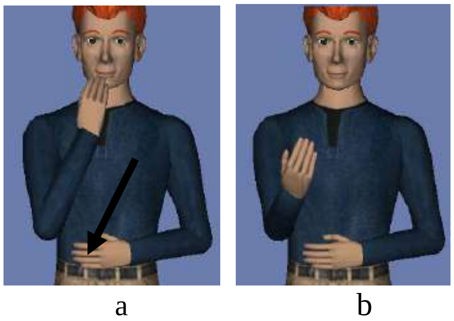


Fig 19: Sign of Adjective
( جيد: Good).

Therefore, the general rule for adjectives in ArSL is as the following:

Adjective-sign ➔ <signed-adjective>

## 3.5 Rule of Adverbs

In the same way of adjectives, an adverb is also expressed by a word-to-sign image, see Figures 20 and 21.

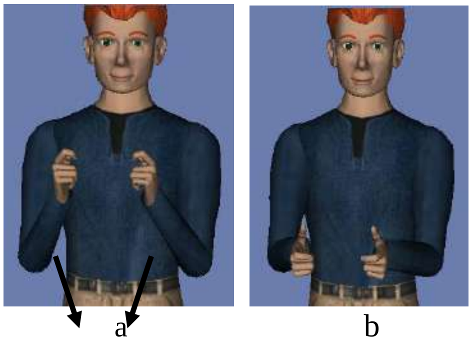


Fig 20: Adverb ( سريعاً: Quickly).

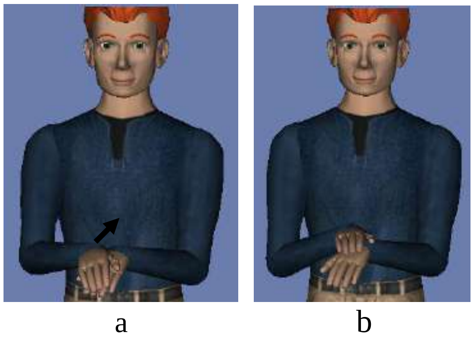


Fig 21: Adverb ( بطيئاً: Slowly).

Therefore, the general rule for adverbs in ArSL is as the following:

Adverb-sign ➔ <signed-adverb>

## 3.6 Rule of Special Character

### a) Rule of Prepositions

ArSL expresses the preposition characters like other sign languages. Its grammar rule is simplified as the following (Figures 22-25 show some prepositions' sings):

Prep-sign ➔ <signed-prep>

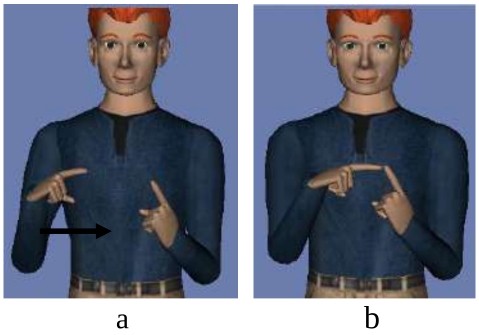


Fig 22: Special Character (إلى: to).

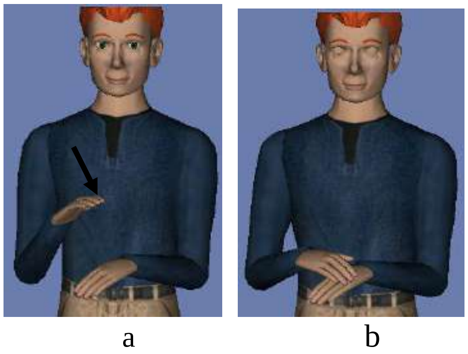


Fig 23: Special Character (على: on).

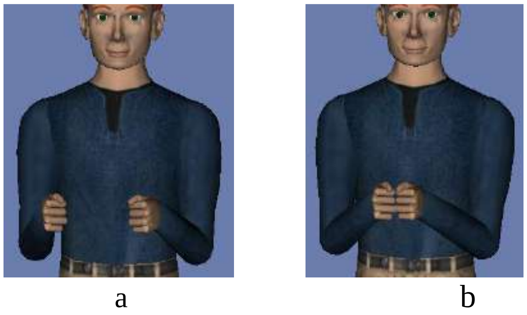


Fig 24: Special Character (مع: with).

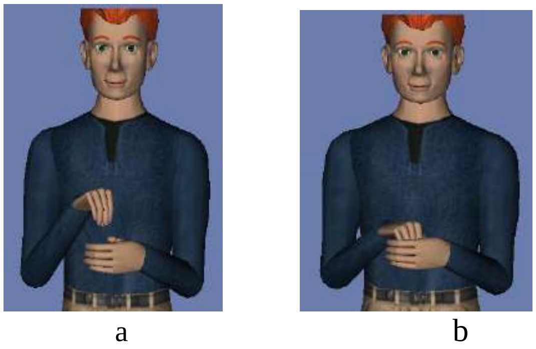


Fig 25: Special Character (في: in).

### b) Rule of Sisters of Kana

The following rule gramatizes the special grammar of Kana and Kana sisters.

Kana-sign ➔ <Signed-Kana>

where <Signed-Kana> ➔ كان | أمسى | أصبح | أضحى | بات | ظل | صار | ليس | مازال | مادام

### c) Rule of Sisters of Ena

Such rule can be summarized as the following:

Ena-sign ➔ <Signed-Ena>

where <Signed-Ena> ➔ إن | أن | لعل | ليت | كأن | لكن | لم

### d) Rule of Sisters of Exception Character

Like other special characters, a simple rule can be summarized as the following:

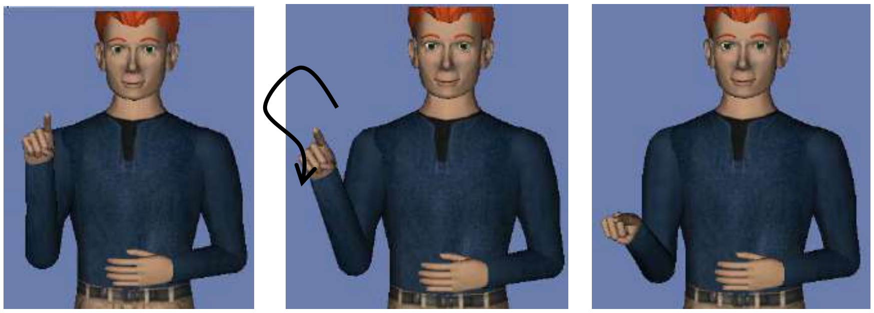
Exp.-sign ➔ <Signed-Exp.>

where <Signed-Exp.> ➔ ماعدا | غير | إلا | خلا | ما خلا | عدا | حاشا | ما حاشا

## 3.7 Rule of Question

In case of question or asking for something, this can be represented using a wonder facial expiration or by using sign of question mark (?).

Question ➔ <Sentence | Word><Facial Exp.> | <Sentence | Word> < ? > | < ? > <Sentence | Word>



a b c

Fig 26: Sign of Question (?).

There are many question keywords like: (How many كم, When متى, What ماذا, Where أين, How كيف, Who من). Generally, the keywords of question are used before, but in sign language the sign of question is used after the word. Therefore, the rule of question can be simplified according to the following rule:

Question-sign ➔ <signed word><question signed>

## 3.8 Rule of Pronouns

The Arabic personal pronouns can be classified into the following: Singular, Plural, and Double (Dual).

### a) Singular Pronouns

The groups of singular pronouns are shown in Table 3.1.

**Table (1): Arabic Singular Personal Pronouns**

| Pronoun | | Examples | |
|---|---|---|---|
| Arabic | English Equivalent | Arabic | English Equivalent |
| أنا | I | أنا كتبتُ (anaa katabto) | I wrote |
| أنتَ | You (singular masculine) | أنتَ كتبتَ (anta katabta) | you wrote |
| أنتِ | You (singular feminine) | أنتِ كتبتِ (anti katabti) | you wrote |
| هو | He | هو كتَبَ (huwa kataba) | he wrote |
| هي | She | هي كتَبَتْ (heya katabat) | she wrote |

## b) Plural Pronouns

Table (3.2) describes the plural pronouns of Arabic language with examples.

**Table (2): Arabic Plural Personal Pronouns**

| Pronoun | | Examples | |
|---|---|---|---|
| Arabic | English Equivalent | Arabic | English Equivalent |
| نحن | We | نحن كتبنا (nahnu katabna) | we wrote |
| أنتم | You | أنتم كتبتم (antum katabtum) | you wrote |
| أنتن | You | أنتن كتبتن (antunna katabtunna) | you wrote |
| هم | They | هم كتبوا (hum katabuu) | they wrote |
| هن | They | هن كتبن (hunna katabna) | they wrote |

## c) Double Pronouns

Double (dual) pronouns of Arabic language are described with examples in Table 3.3.

**Table (3): Arabic Double Personal Pronouns**

| Pronoun | | Examples | |
|---|---|---|---|
| Arabic | English Equivalent | Arabic | English Equivalent |
| أنتما | You | أنتما كتبتما (antuma katabtuma) | you wrote |

| | | | |
|---|---|---|---|
| هما (مذكر) | They (masculine) | هما كتبا (huma katabaa) | they wrote |
| هما (مؤنث) | They (feminine) | هما كتبتا (huma katabata) | they wrote |

## 3.4 Numbers and Numeral Incorporation

Signs are composed of movements, holds and information about hand shape, location and orientation. For example, the sign of (أسبوع; Week) would be represented as hand shape without movement. However, the concept of (أسبوعان; two weeks) can be expressed in the shape of hand shapes of this sign (i.e.; change 1 to 2). The location and orientation remain the same. Consequently, this process is known as numeral incorporation, and it is has been described by Scott Liddell and Robert E. Johnson [Valli et al., 2005].

We can say that the sign (أسبوعان; two weeks) has two morphemes. The first one that includes (أسبوع; Week)- the segmental structure holds and movement and the location and orientation. The second part is the hand-shape, which has the meaning of the specific number; as shown in Figure 27. A rule of the two morphemes would look like this:

أسبوعان ➔ < أسبوع > + <Duo Sign.>

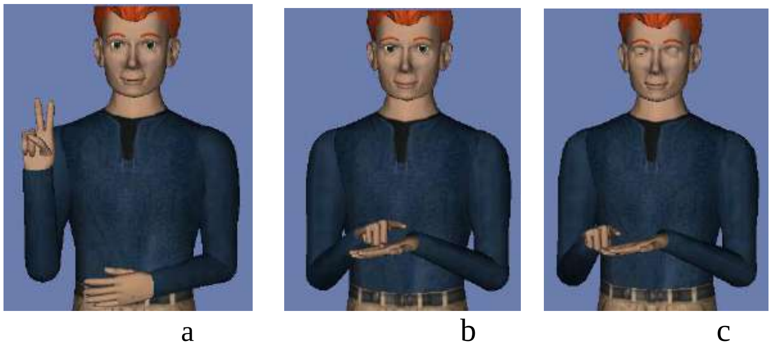


**Fig 27: Sign of Two Morphemes (أسبوعان : two weeks).**

### 3.4.1 Rule of Plural and Repetition

Plural and repetition can be represented by including signed number after the sign of the word, by repeating the sign of the word many times or by moving the finger inside.

Plural & Repetition ➔ <Sign><Number> | < Sign >* | < Sign > <Finger inside>

where * means zero or more than one time.

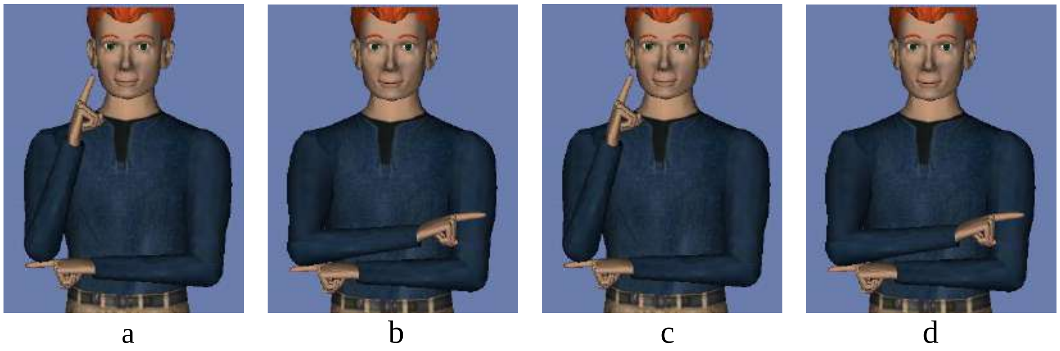


**Fig 28: Sign of Plural (أيام : days).**

### 3.4.2 Rule of Emphasis

In case of emphasis, it can be done by repetition, longer signing time and sometimes it uses facial expression and dramatization.

emph-sign ➔ <emph-verb>* + <facial exp.>

where * means zero or more than one time.

But, adverbs are expressed manually, by one hand position in relation to the other.

## 3.5 Compounds ArSL

The Arabic nouns can be created from Arabic verbs. In this section, we will look at a grammar way by which ArSL can create new signs.

Sometimes, a language creates new words by taking two words (free morphemes, [Valli et al., 2005]) that it already has and putting them together, i.e.; this process is called compounding, English language have many compounds.

When nouns are derived from verbs in Arabic, a regular pattern (or template) can be described. A pattern or template can also be described for the formation of compounds. In Arabic, when two morphemes come together to form a compound, the following grammar rule takes place: Noun ➔ < ال > + <Verb>

A new meaning is created when two different morphemes come together to form a compound. For example, هندرة is composed from Arabic morpheme <هند> from <هندسة>, and Arabic morpheme <رة> from <إدارة>.

## 5. ArSL System Structure

The proposed architecture for the desired system contains three major steps: (1) Grammatical: morphological analysis, syntax, lexical/ semantic analysis, (2) transformation and (2) generation of script and avatar showing scene. First of all, the written text passes to the tokenizer to separate the words of the Arabic sentence. Words after separation go to the morphological analysis phase. Morphological analysis using Arabic template grammar is done with each word and, as a result of this operation, each word with its attributes passes to the next phase [Al-Barhamtoshy et al., 2007]. The most important attributes needed are: root, word type, tense (for verbs) and the count of words. These attributes proceed to lexical and semantic analysis phases.

The grammatical analyzer was developed on the basis of a Arabic Template Grammar (ATG), adapted for effective natural language processing [Al-Barhamtoshy et al., 2007]. It was implemented into the morphological and partial syntactic system as its grammatical component, supporting manifold morphological, derivational and syntactic analyses of texts. The main advantage of the grammatical component is that it selects those parses of items which are pertinent to the respective outer units. Thus high ambiguity of the morphological structures is solved (or significantly reduced) by the syntactic information supplied.

The derivation procedure is time-consuming (decelerating analysis for 10 %), so stems from vocabulary are checked first, before generated ones. Proper names and abbreviations are processed using template matching of the ATG. An output of the grammatical modules includes: • lemma and its POS tag for each wordform in the Arabic text; a list of grammatical tags for each word form (root, tense (for verb), case, number, animate, etc.); • a dependency tree for each sentence (a set of syntactically linked word pairs with established relations "head-daughter"); • a phrase structure in terms of semantic-syntactic functions, such as "proposition", "subject", "object", etc.

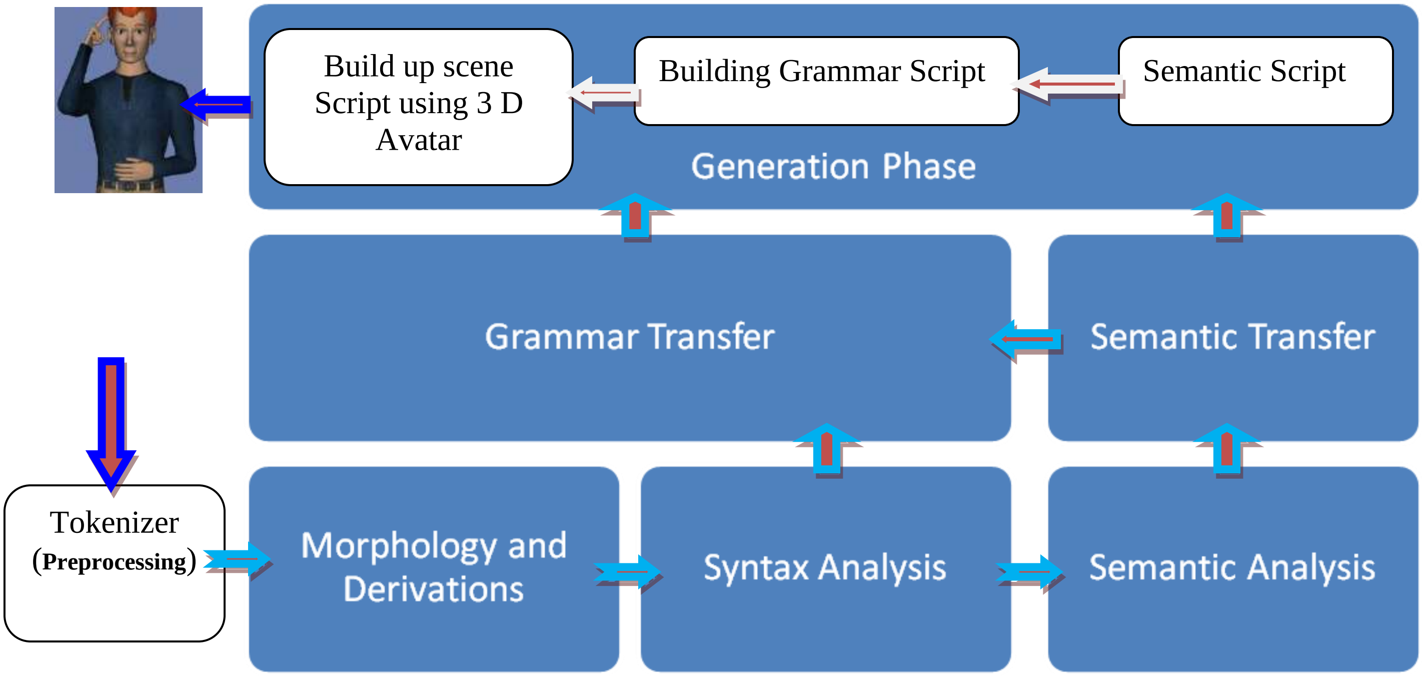


Fig 30: ArSL System Architecture

In this step, many rules are checked, to customize the associated lexical meaning: synonym, antonym, homophony, compounds, hyponymy and hypernym. Transformational and generation script phase is used to generate the output script. Some rules must apply before translation. These rules are about choosing the best meaning of the word to translate and also to reorder words in sentence to be easier to understand by deaf people.

The output script is input for the 3D Avatar system to generate the Arabic signs, using the signs database. Figure 30 shows the proposed system architecture.

## 5.1 ArSL Testing and Discussion

In this section, selected sentences were used for testing the proposed model. Such sentences include all various possible analyzed verbs, nouns, adjectives, adverbs, etc. and their various combinations of using affix rules. The following tables (Table 1, 2, 3 and 4) are examples of these selected sentences.

**Table (1): Selected Example 1**

| مؤنث | مذكر | |
|---|---|---|
| ذاكرت الطالبة | ذاكر الطالب | مفرد |
| الطالبة ذاكرت | الطالب ذاكر | |
| ذاكرت الطالبتان | ذاكر الطالبان | مثنى |
| الطالبتان ذاكرتا | الطالبان ذاكرا | |
| ذاكرت الطالبات | ذاكر الطلاب | جمع |
| الطالبات ذاكرن | الطلاب ذاكروا | |

**Table (2): Selected Example 2**

| مؤنث | مذكر | |
|---|---|---|
| أكلت الطفلة | أكل الطفل | مفرد |
| الطفلة أكلت | الطفل أكل | |
| أكلت الطفلتان | أكل الطفلان | مثنى |
| الطفلتان أكلتا | الطفلان أكلا | |
| أكلت الطفلات | أكل الأطفال | جمع |
| الطفلات أكلن | الأطفال أكلوا | |

**Table (3): Selected Example 3**

| مؤنث | مذكر | |
|---|---|---|
| داومت الموظفة | داوم الموظف | مفرد |
| الموظفة داومت | الموظف داوم | |
| داومت الموظفتان | داوم الموظفان | مثنى |
| الموظفتان داومتا | الموظفان داوما | |
| داومت الموظفات | داوم الموظفون | جمع |
| الموظفات داومن | الموظفون داوموا | |

**Table (4): Selected Example 4**

| مؤنث | مذكر | |
|---|---|---|
| كتبت المعلمة | كتب المعلم | مفرد |
| المعلمة كتبت | المعلم كتب | |
| كتبت المعلمتان | كتب المعلمان | مثنى |
| المعلمتان كتبتا | المعلمان كتبا | |
| كتبت المعلمات | كتب المعلمون | جمع |
| المعلمات كتبن | المعلمون كتبوا | |

The testing sample contains complete sentences with their included. The results of this experiment are presented in Table 5.5.

The sample is composed of 20 sentences, 10 of which are analyzed to verb phrases and 10 sentences belong to noun phrases.

**Table (5): Results of testing the proposed model**

| | Total No of hits | Correct Ratio | Error Ratio |
|---|---|---|---|
| Verb Phrases | 24 | 92% | 0.08 |
| Noun Phrases | 24 | 95% | 0.05 |
| Total | 48 | 93% | 0.07 |

Due to the proposed model complexity, we turn our attention to see how analysis is conducted by the proposed model, the running time cost is determined by components of the following algorithm:

Step 1: Morphological, syntax and syntactic analysis.
Step 2: Transfer the sentence to signed sentence.
Step 3: Generate the sign script.
Step 4: Avatar engine show the signed sentence.

Therefore, for the first step; checking the existence of the entire word and employing the match with the Arabic dictionary. Since the comparison is carried out character by character, we should assume that the number of comparisons would be: $\mathbf{T_1 = n,}$ where $\mathbf{n}$ is the length of the entire Arabic word ($\mathbf{n=3}$, $\mathbf{n=4}$ or $\mathbf{n=5}$).

At the second step, if the entire word exists in a proper sequence, after validating prefixes and suffixes that are checked against a list of stored prefixes and suffixes, the number of comparisons is determined as follows: $\mathbf{T_2 = \log N_{ps}}$ **,** where $\mathbf{N_{ps}}$ is the number of prefixes and suffixes.

The validation of word infixes depends on two factors [Suleiman H. Mustafa, 2003]: the size of the difference between positions of letters of root in the entire word, and the list of infix letters to be checked. Accordingly, the number of comparisons would be calculated as follows:

$$\mathbf{T_3 = D + I\ ,}$$

where $\mathbf{D}$ is the number of comparisons for checking the difference and $\mathbf{I}$ is the number of character comparisons to match an infix against the set of infixes.

Consequently, the overall running time for the proposed model is computed as the sum of the three factors listed above:

$\mathbf{T = T_1 + T_2 + T_3 = n + (\log N_{ps}) + (D + I)}$ **(** per word in worst case.)

## 5.2 Screenshots

The designed system contains four basic links. As shown in Figure 31, the links are: Arabic Numbers "الأرقام العربية", basic math learning"تعليم الحساب", chosen words"كلمات مختارة" and text translation"ترجمة نص". The first link, Arabic numbers, serves to see the sign of each Arabic number. By clicking the number and then clicking the sign button ✊ at the bottom of page, so the sign of the selected number will appear in the character side.

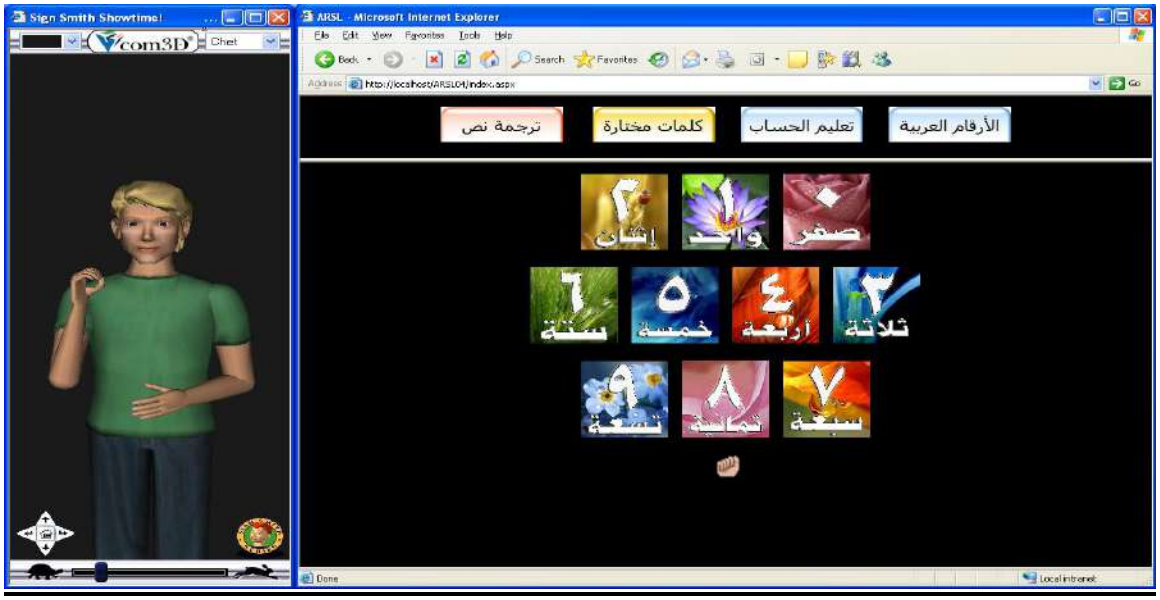


**Fig 31: Arabic numbers page**

The second link serves to teach the children some basic mathematical operations such as addition, subtraction, multiplication and division. Firstly, choose the operation and then, an example of that operation will appear. The sign of that operation can be shown, by pressing the sign button✊. By clicking the button Next "التالي", another example of the same operation will be displayed, see Figure 32.

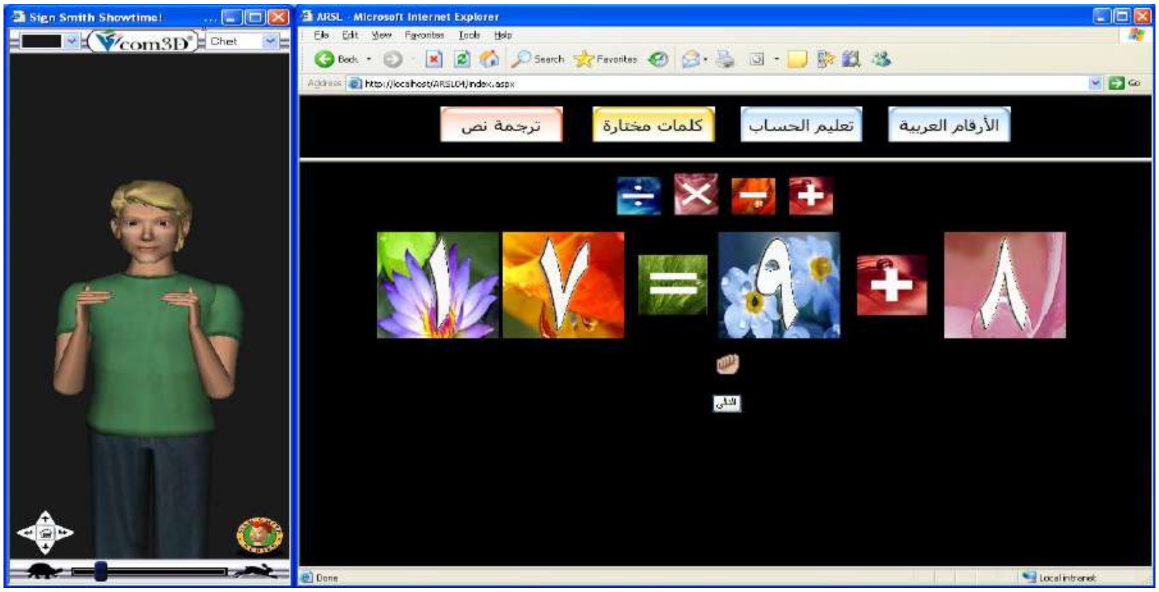


**Figure 32: Basic math learning page.**

In the chosen words page, there are four categories: pronouns "الضمائر", family "الأقارب", nouns "الأسماء" and verbs "الأفعال". In each category, there are some examples covering the four categories. By choosing one of the examples and pressing the sign button✊, the sign of the selected example will appear. Figure 33 shows the pronoun (أنا).

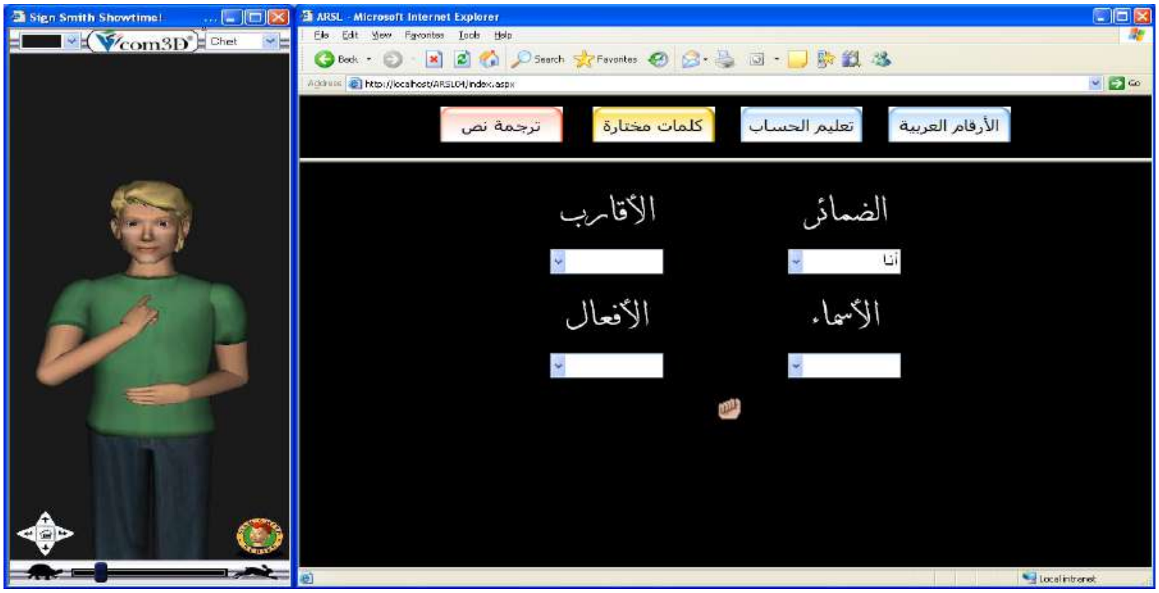
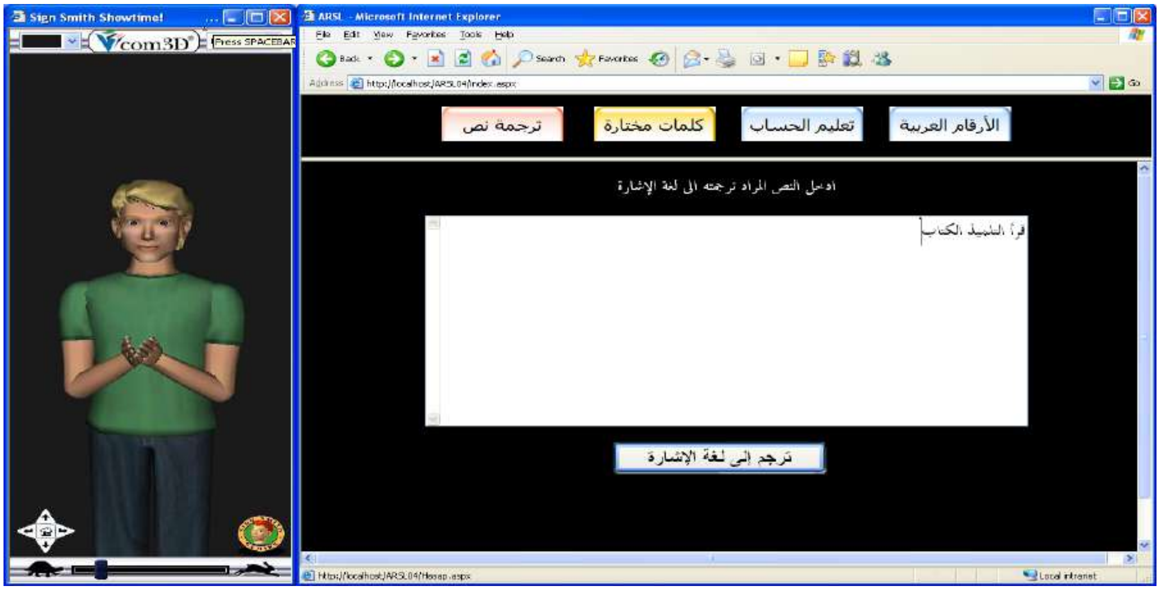


**Figure 33: Chosen words page.**

The last page is text translation page that translates any Arabic text to Arabic sign language (ArSL). The page contains a text box to enter the Arabic text, a button to translate the text and also the sign button ✊ to show the translated text with sign language. See Figure 34.


Sign Smith Showtime!
Vcom3D
Press SPACEBAR
ARSL - Microsoft Internet Explorer
File Edit View Favorites Tools Help
http://localhost/ARSL04/index.aspx
ترجمة نص
كلمات مختارة
تعليم الحساب
الأرقام العربية
ترجم إلى لغة الإشارة
http://localhost/ARSL04/Hesap.aspx
Local intranet

ادخل النص المراد ترجمته الى لغة الإشارة

قرأ التلميذ الكتاب

**Figure 34: Arabic Text Translation**

## 7. Conclusion

The main goal of this paper is to design a tool to facilitate communication with deaf people, using computer technology. Deaf people are small part of the society but, it is important to merger them with normal people.

To achieve this goal, a prototype system for translating from written natural Arabic language to Arabic sign language is presented. This system needs to design and implement architecture to produce Arabic Sign Language (ArSL), design and implement a dictionary for ArSL and generate the required scripts for ArSL, to be represented using a computerized 3D Avatar system. Therefore, the paper has shown the need to increase the effectiveness of ArSL system and to apply it for teaching deaf children, via interactive media. Everybody will support this kind of systems and it will be very a useful system for deaf people at all ways.

The proposed ArSL model is just a small part of a biggest work that deaf people need, which a represents natural translation system between deaf and normal people. Deaf people need more attention, more paper and economic and moral support.

The future development of this work is to make the second way of translation from sign language to written Arabic text. Some paper in that area has begun, but it needs more ideas and hard work. Also, it is important to merge a speech engine with these systems to be more natural. The new generation of all new systems is to be on portable devices – like PDA and mobile devices, so life will be easier and more helpful.

## References

M. A. Abdel-Fattah. 2005, "Arabic Sign Language: A Perspective", Journal of Deaf Studies and Deaf Education vol. 10, no. 2, Oxford University Press.

Al-Barhamtoshy Hassanin M., Khalid O. Thabit and Basil Ba-Aziz. 2007, "Arabic Morphology Template Grammar - Based", the 7th Conference on Language Engineering, Cairo, Ain Shams University, December 2007.

Al-Barhamtoshy Hassanin M. and Sami Halawani. 2000, "Computer-based Arabic Sign Language Recognizer", AUEJ: Al-Azhar University Engineering Journal, vol.4, no.4, October 2000.

Al-Barhamtoshy. Hassanin M. 1992, "Text Understanding of Arabic Application", Ph.D Thesis, System and Computers Dept., Faculty of Engineering, Al-Azhar University, Cairo.

Moh'd Belal Al- Daoud. 2004, "A Simple Sign Language Finger Spelling System", August 2004.

Omar Al-Jarrah, Alaa Halawani. 2001, "Recognition of gestures in Arabic sign language using neuro-fuzzy systems", November 2001.

Arabic Sign Language Dictionary 2008, web site: http://www.deaf.4t.com .

J. A. Bangham, S. J. Cox, R. Elliot, J. R. W. Glauert, I. Marshall, S. Rankov, and M. Wells. 2000, "Virtual signing: Capture, animation, storage and transmission – An overview of the ViSiCAST project", IEEE Seminar on Speech and language processing for disabled and elder people.

A. M. El-Samahy, Hassanin M. Al-Barhamtoshy and M. A. Madkour. 1993, "An Arabic Morphological Analyzer", 3rd International Conference, pp. 294-304, Faculty of Engineering, Al-Azhar University, Cairo.

Matthew P. Huenerfauth. 2004, "A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation", August 2004.

Matthew P. Huenerfauth. 2003, "A Survey and Critique of American Sign Language Natural Language Generation and Machine Translation Systems", September 2003.

Mohammad A. Kamel. 1999, "Sign Language for Deaf Teachers", First Edition, Tanta University- Egypt. محمد على كامل، لغة الإشارة للقائمين على رعاية الصم، الطبعة الأولى 1999، جامعة طنطا – مصر.

J. R. Kennaway. 2001, "Synthetic animation of deaf signing gestures", 4th International Workshop on Gesture and Sign Language Based Human-Computer Interaction, London. Lecture Notes in Artificial Intelligence vol. 2298 (eds. Ipke Wachsmuth and Timo Sowa).

Kuwaiti Sign Language Dictionary 2008, web site: http://d2000.4mg.com .

S. Liddell and R. Johnson. 1989, "American Sign Language: The Phonological Base, Sign Language Studies", Volume 64, pp 195-277.

Ian Marshall, ´ Eva S´af´ar. 2004, "Sign Language Generation in an ALE HPSG", September 2004.

Mohamed Mohandes and Jamil Bakhshawai 2006, "Translation of the Arabic Text to Arabic Sign Language Using Computers", Forth Saudi Technical Conference and Exhibition, Saudi Arabia.

Suleiman H. Mustafa (2003), "A Morphology driven string matching approach to Arabic text searching", The journal of systems and software, vol. 67 (2003) 77-87.

É. Sáfár and I. Marshall. 2001, "The architecture of an English-text-to-sign-languages translation system", In G. Angelova, editor, Recent Advances in Natural Language Processing (RANLP), pages pp223--228. Tzigov Chark, Bulgaria.

VCom3D Inc, 2008. http://www.vcom3d.com .

L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, and M. Palmer. 2000, "A Machine Translation System from English to American Sign Language", Association for Machine Translation in the Americas.

# Diacritization and Transliteration of Proper Nouns from Arabic to English

Hamdy S Mubarak Mohamed Al Sharqawy Esraa Al Masry

Arabic NLP Dept
Sakhr Software
Cairo, Egypt
{hamdys,mas,emasry}@sakhr.com

## ABSTRACT

This paper proposes a complete system for the automatic diacritization and transliteration of proper nouns from Arabic to English. The system consists of three phases: Correction, Diacritization, and Transliteration.
Our results show an average accuracy of 89% on blind test sets with forced spelling mistakes (and 95% for correct input).

## KEYWORDS

Diacritization, Transliteration, Machine Translation, Arabic

## 1. INTRODUCTION

Transliteration is the task of transcribing a word or text written in one writing system into another writing system. Person names, locations and organizations as well as words borrowed into the language are the most frequent candidates for transliteration.

In Information Retrieval (IR), the most important query words in are often proper names. In cross language retrieval, a user issues a query in one language to search a collection in a different language. If the two languages use the same alphabet, the same proper names can be found in either language. However, if the two languages use different alphabets, the names must be transliterated or rendered in the other alphabet.

As mentioned in (Al-Onaizan and Knight, 2002), two types of transliteration exist, *forward transliteration* and *backward transliteration*.

Forward Transliteration is the transliteration of a foreign name (in the case of our system, Arabic) into English. Typically, there are several acceptable transliteration candidates. For example, the Arabic name “محمد” (mhmd in Buckwalter encoding might correctly be transliterated into Mohamed, Mohammed, Mohammad, etc. In fact, the many types of name variation commonly found in databases can be expected.
A recent web search on Google for texts about “Muammar Qaddafi” (spelled in Arabic as معمر القذافي mEmr AlqdAfy in Buckwalter encoding), for example, turned up thousands of relevant pages under the spellings Qathafi, Kaddafi, Qadafi, al-Qaddafi, Al-Qaddafi, Al Qaddafi, etc (and these are only a few of the variants of this name known to occur).

Backward Transliteration is the reverse transliteration process used to obtain the original form of an English name that has already been transliterated into the foreign language. In this case, only one transliteration is retained.

For the transliteration from Arabic to English, some observed problems are:
- Both Arabic and English lack some sounds and letters from the other language. For example, there is no perfect match for “ ع” in English and “P” in Arabic. This leads to ambiguities in the process of transliteration.

Another problem associated with Arabic is the omission of diacritics and vowels (fatha, dumma, kasra, shadda,sokoon) in almost all the Arabic writings. The information contained in unvocalized Arabic writing is not sufficient to give a reader, who is unfamiliar with the language, sufficient information for accurate pronunciation. Diacritics are considered to be one of the main causes of ambiguity when dealing with Arabic proper nouns.

Another observed problem in Arabic is the existence of Common Arabic Mistakes (CAM) in which different characters are used interchangeably; like Hamza errors (ا،أ،آ،إ), Yaa errors (ى،ي), and Taa Marbuta errors (ة،هـ).

In this paper, we present a complete system for Correction, Diacritization, and Transliteration of Arabic proper nouns using a database of name pairs in Arabic and English languages.

The system searches for the normalized form of user input (in Arabic) in a dictionary of proper names, and if found it returns the most frequent transliteration, otherwise it suggests the appropriate diacritization based on its Morphological Analysis if analyzed or the best matching with patterns obtained from the diacritized proper names. And in case of mismatch with patterns, it uses a probability matrix for diacritization of any consecutive characters. Finally, the system applies transliteration rules to obtain the English equivalent.

## 2. PROPER NAMES DICTIONARY

For training purposes, we needed a list of name pairs, i.e. names written in Arabic and correctly transliterated into English. We used Sakhr Proper Names Database which consists of different types of proper names (Human, Location, Organization, etc) . For each proper name, information about Diacritization, Transliteration, Gender, Theme, etc are provided manually.

From the total of 171K Human names (26K Arabic, and 94K non Arabic), we used only 51K names for Transliteration (13K Arabic, and 38K non Arabic) which contain the transliteration information, and use the entire list for Correction and Diacritization processes.

Examples for Arabic Human names: آمنة بنت وهب، أبو الأسود الدؤلي، الملك عبد الله الثاني, and for Non Arabic Human names: آبل غومبا، آدم سميث، جورج بوش.

In the case of Forward Transliteration (where we want to convert a name originally from Arabic into English), there is usually more than one acceptable corresponding name in English. For example, the name “ طارق” has 4 different equivalents in our database: “Tarek”, “Tareq”, “Tarik”, and “Tariq”.

The distribution of names with different number of alternatives is summarized in table 1.

Table 1: Number of Alternative Names.

| # | One | Two | Three | Four | Five+ |
|---|---|---|---|---|---|
| % | 76% | 16% | 4% | 2% | 2% |
| Ex: | مدحت Medhat<br>إسلام Islam | أحمد Ahmed, Ahmad | أسامة Ossama, Osama, Usama | إلياس Elias, Elyas, Ilyas, Alias | محمد Mohamed, Mohamad, Mohammad, Muhammad, Mohammed,… |

It’s also observed that the maximum number of alternatives was 11 (ex: the name “يوسف” has these equivalents: Youssef, Yusuf, Yousef, Yusef, Youssif, etc.)

## 3. CORRECTION

From a random sample of 1000 Queries to Sakhr Arabic Search Engine on the Web (http://johaina.sakhr.com/), about 70% of these queries are Proper Names, and 13% of them have spelling mistakes in Hamza, Yaa, and Taa Marbuta.
To solve this problem, we use Text Normalization for both the input and the Proper Names Dictionary.

### 3.1 Text Normalization

Text Normalization is a process by which text is transformed in some way to make it consistent in a way which might not have been before, and it’s often performed before comparison. For Arabic, all shapes of Hamza (إ،آ،أ،ا) are converted to Plain Alef (ا), Yaa (ي) is converted to Alef Maqsura (ى), and Taa Marbuta (ة) is converted to Haa (هـ). This makes both اسامه and أسامة match.

### 3.2 Concatenation Errors

Another type of spelling errors is the concatenation of two or more tokens in the user input which leads to the need for splitting mechanism to obtain correct tokens. Tokens end with any letter that doesn’t cause visual ambiguity - i.e. has isolated and final forms only (no initial or middle forms) like the letters (د،ذ،ر،ز،و،ا) – when concatenated to other tokens, they need a Simple Splitting; otherwise a Complex Splitting is needed.
Examples: محمداحمدعبدالله (Simple) and منالحسنمحمد (Complex).

## 4. DIACRITIZATION

Diacritization of proper names is used in many applications like Address Book, Text to Speech, and also as an intermediate step in Transliteration which is used in Machine Translation Systems (MT), and Multilingual Personal Information (Banking, ID Cards, Passports, etc.)

In Arabic, words consist of prefixes, suffixes, and stem; the stem can be determined by a root and a morphological pattern pair. The root represents the stem original letters while the morphological pattern decides how the stem will be pronounced.
Because short vowels (diacritics) are commonly not presented in Arabic orthography, this creates a problem in transliterating unknown proper names (Out Of Vocabulary) since these missing diacritics should be deduced before transliteration to obtain the appropriate pronunciation.

### 4.1 Proper Name Patterns

To diacritize unknown proper name, we use a set of rules to deduce its “pattern”, and search in the proper names database to retrieve a list of proper names with the same candidate pattern. To select the best match (minimum number of character substitutions), the "Hamming distance" values are used in addition to statistical information for patterns frequencies. Suggested diacritics are the same as this favored pattern.

While deducing proper name pattern, some issues should be taken into consideration for Arabic names only (for non Arabic names, only Long vowels (ا،و،ي) are preserved):

1. Long vowels (ا،و،ي), Ending characters (ى،ة), different Hamza shapes (أ،إ،آ،ء،ؤ،ئـ) are preserved.

2. Definite article (ال), some initial characters (م،بو), and some ending characters or suffixes (اني،ني،ان،اوي..) are also preserved .
3. Other characters (consonants) are replaceable and can match with any other consonant.
4. Special cases for preserving some consonants (م،ت) are applied.
5. The “Sun Letters” (other than the letters in string "أبغ حجك وخف عقيمه") should have an extra Shadda.

For 171K diacritized Human names (26K Arabic, and 94K non Arabic), we have generated about 3K and 10K patterns for Arabic and non Arabic respectively. Examples for these patterns are shown in Table 2 and Table3:

Table 2: Arabic Proper Names Patterns.

| Pattern | Diac1 | % | Examples1 | Diac2 | Examples2 | % |
|---|---|---|---|---|---|---|
| ـ ـ اء | ـَـَ اء | 61 | عَلَاء، بَهَاء | ـِـَ اء | فِدَاء، لِقَاء | 28 |
| ال ـ ـ ـ و ـ ي | اَلْ ـَـْـُو ـِيّ | 70 | اَلْبَرْغُوثِيّ، اَلْجَنْزُورِيّ | اَلْ ـَـَـُو ـِيّ | اَلْبَرَمُوسِيّ | 10 |

Table 3: Non Arabic Proper Names Patterns.

| Pattern | Diac1 | % | Examples1 | Diac2 | Examples2 | % |
|---|---|---|---|---|---|---|
| ـ ي ـ و ـ ـ | ـِ ي ـُو ـْـ | 69 | رِيمُونْد، هِيرُولْد | ـِ ي ـُو ـَـ | شِيرُودَر، نِيكُولَس | 14 |
| ـ و ـ ـ ـ ي | ـُو ـِـْـِ ي | 44 | شُومِسْكِي | ـُو ـْـَـِ ي | كُورْدَرِي | 28 |

For the unknown Arabic name “سماء” which matches the 1st Arabic pattern, the system suggests its correct diacritization “سَمَاء” after getting the nearest known name “سناء”. Similarly, it diacritizes the name “البعقوبي” as “اَلْبَعْقُوبِيّ”.

When the input pattern doesn’t match any of the existing patterns, the system recursively splits it into smaller patterns and searches for them. The overall diacritization is the concatenation of all these diacritized sub patterns after applying some concatenation rules.
Example: The name “كرومازوف” is splitted into 2 parts “كروما” and “زوف” which match the patterns “ـ ـ و ـ ا” and “ـ و ـ”, which leads to the final diacritization “كُرُومَازُوف”.

## 4.2 Bigram Diacritization Matrix

After pattern splitting, if the name pattern doesn’t match any of existing patterns, the system uses a Bigram Diacritization Matrix to diacritize the name. This matrix has all the consecutive characters associated with the diacritics probabilities obtained from Proper Names database. The names “اَلْعُوِيرَان” and “كُونْدِلِيزَا” are sample outputs using this matrix.

## 4.3 Morphological Analysis

Morphological Analyzer is also used for diacritizing normal words that exist in Arabic proper names, examples: هاني اَلضَّابِط، أحمد عَقْل، صلاح فُولاذ

The decision of being Arabic or non Arabic name is important for Diacritization and hence Transliteration, and it’s done through these surface rules:

1. If the name exists in Proper Names database, the Arabic and non Arabic probabilities are taken into consideration, i.e. some names tend to be an Arabic name (ex: محمد with probability of 75%), some names tend to be non Arabic name (ex: جون with probability of 98%), while others are neutral (ex: آدم with probability of 50%).

2. If the name contains one of the following characters “ح، ظ، ع، ص، ط، ض، ق”, it tends to be an Arabic name. These characters are obtained statistically from Proper Names database (ex: Probability of character ‘ض’ in Arabic name is 84% and 16% for non Arabic name, like: حامد قرضاي).
3. If the name matches an Arabic pattern and doesn’t match any of the non Arabic patterns, it’s considered an Arabic name (and vice versa) and gets diacritized consequently.
4. Unless user input is provided, the name is considered non Arabic name and diacritized therefore.

## 5. TRANSLITERATION

Transliteration is a mapping from one system of writing (alphabet) into another, word by word. Transliteration attempts to be exact, so that an informed reader should be able to reconstruct the original spelling of unknown transliterated words. To achieve this objective, transliteration may define complex conventions for dealing with letters in a source script which do not correspond with letters in a goal script.

Transliteration is opposed to transcription, which specifically maps the sounds of one language to the best matching script of another language. Also, transliteration should not be confused with translation, which involves a change in language while preserving meaning.

### 5.1 Transliteration Standards

Although there is no Universal Transliteration system from Arabic to English, there are some common systems like Buckwalter, ISO 233, Qalam, etc. Any transliteration system should consider the following issues:

- Transliteration ignores assimilation of the article before the “sun letters”, and may be easily misread by non-Arabs. For instance the proper name word “An-nour” would be more correctly transliterated to “Al Nour”.
- A transliteration is ideally fully reversible: a machine must be able to transliterate it into Arabic and back.
- Rendering several Arabic phonemes with an identical transliteration, or digraphs for a single phoneme (such as sh) may be confused with two adjacent phonemes.
- ASCII transliterations using capital letters to disambiguate phonemes are easy to type but may be considered unaesthetic.

Examples of character mapping in different systems are shown in Table 4.

Table 4: Transliteration Examples

| Letter | Name | ISO | Qalam | Buckwalter |
|---|---|---|---|---|
| ث | Thaa | t̠ | Th | v |
| خ | Khaa | h̠ | Kh | x |
| س | Seen | s | S | s |

We used a manual character mapping similar to Qalam, with some enhancements, to preserve the spelling rather than the pronunciation. The system has these additional features:

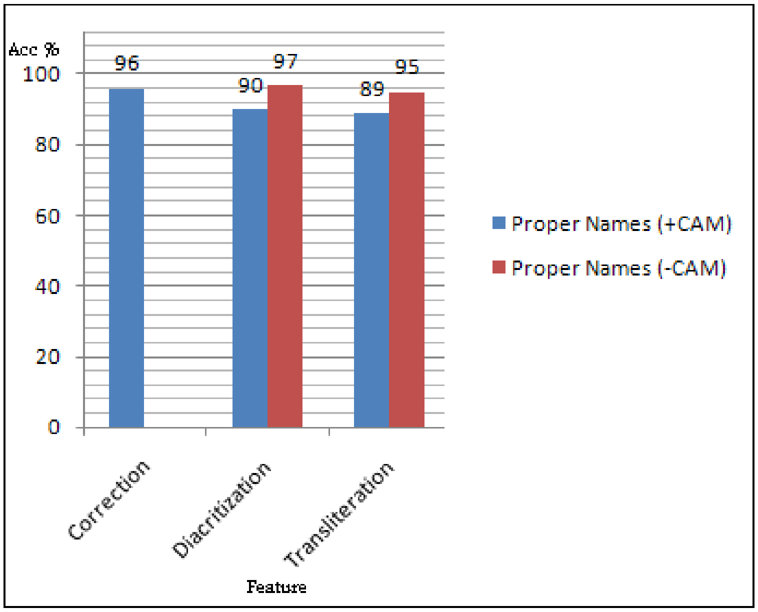
1- A customized mapping of letters. For example the mapping of letter ‘ج’ may be to ‘g’ in Egypt or ‘j’ in Gulf.
2- Restoring the single character abbreviation (like transliterating “دبليو” to “W.”).

3- Fine tuning of the Transliteration based on statistics obtained from the database (for consecutive vowels, or special patterns and conditions).

## 6. EVALUATION

As described before, in case of forward transliteration, there is more than one acceptable transliteration. Ideally, our gold standard should maintain a set of equivalent English names for each Arabic entry, but it is not possible to gather all the possible transliterations for all the Arabic names. So, we evaluated the system accuracy manually through experienced linguists.

From a random sample of 1000 proper names (Arabic and non Arabic) with a total of 2200 tokens, we have normalized all the inputs (i.e. forcing Common Arabic Mistakes CAM) and evaluated the accuracy of Correction, Diacritization, and Transliteration. Manual assessment shows accuracies of 96%, 90%, and 89% for Correction, Diacritization, and Transliteration in order (and raise to 97% in Diacritization, and 95% in Transliteration if the input is correct.)
Results are shown in Graph 1:



Graph 1: Correction, Diacritization, and Transliteration Accuracy

### 6.1 Transliteration Sample Output

Transliteration sample outputs of blind inputs are shown in Table 5:

Table 5: Sample Outputs

| Proper Name (+CAM errors) | Transliteration |
|---|---|
| ابوالوكل مؤنس بن فرحان الرويلي | Abu Al Wakl Monis Bin Farhan Al Rwiely |
| وضحه مساعد عبدالرحمن ابراهيم الجليبي | Wadha Musaed Abdul Rahman Ibrahim Al Joliby |
| شاديه علي وصل مرشود الحازمي | Shadia Ali Wasl Marshod Al Hazmi |
| | |
| هينار كي آر هولتيت | Hinar Ki R. Holtit |
| شانتال ميلون دلسول | Shantal Meillon Dalsol |

## 7. CONCLUSION

We have introduced a complete system for Correction, Diacritization, and Transliteration of names from Arabic to English with an accuracy of 89% on blind test-data. The system uses bilingual training data, along with morphological analysis (Sakhr's Morphological Analyzer), some heuristic rules and observations to achieve these results in combination with traditional statistical language processing and machine learning algorithms.

## 8. REFERENCES

[1] Mehdi M. Kashani, Fred Popowich, and Fatiha Sadat (2005). "Automatic Transliteration of Proper Nouns from Arabic to English", School of Computing Science, Simon Fraser University, National Research Council of Canada.

[2] AbdulJaleel, Nasreen and Leah S. Larkey (2003). 'Statistical Transliteration for English-Arabic Cross Language Information Retrieval', CIKM 2003: Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA.

[3] Al-Onaizan, Yaser and Kevin Knight (2002). 'Machine Transliteration of Names in Arabic Text', ACL Workshop on Computational Approaches to Semitic Languages.

# SPOKEN TERM DETECTION FOR ARABIC EDUCATIONAL MEDIA

M. Hesham* and M. F. Abu-EL-Yazeed**

mhesham@eng.cu.edu.eg

Faculty of Engineering, Cairo University, 12613, Egypt

## Abstract

Automatic media content analysis and understanding for efficient topic searching and browsing are current challenges in the management of e-learning content repositories. This work discusses the results of Arabic media searching using pure audio information. A hidden-Markov modeling is used in the training of an automatic-speech recognition (ASR) engine for modern standard Arabic (MSA). The phoneme-based ASR engine is then used to model searching lattices for word searching. The results show lower word-error rates (WER) which can be compared to corresponding results in recent literatures. The work also reveal that more work is needed to model the linguistic architecture of Arabic in more details.

**Keywords** Acoustic Modeling using Hidden Markov Models, Keyword Spotting from Speech Utterances, Arabic Speech Recognition.

* M. Hesham, a professor at Engineering Math. & Physics Dept., Cairo University

* *M. F. Abu-El-Yazeed, a professor at Electronics Engineering Dept., Cairo University

## I. Introduction

Ever increasing computing power and connectivity bandwidth together with falling storage costs result in an overwhelming amount of data of various types being produced, exchanged, and stored. Consequently, search has emerged as a key application as more and more data are being saved. Text search in particular is the most active area, with applications that range from web and private network search to searching for private information.

Speech search has not received much attention due to the fact that large collections of un-transcribed spoken material have not been available, mostly due to storage constraints. Recently, the availability and usefulness of large collections of spoken documents is limited strictly by the lack of adequate technology to exploit them. Manually transcribing speech is expensive and sometimes outright impossible due to privacy concerns. This leads to exploring an automatic approach to searching and navigating spoken document collections. In order to deal with limitations of current automatic speech recognition (ASR) technology, many recent works deal this topic. Another important aspect of such a work, is the performance measure, it may be taken as reference the output of a text retrieval engine that runs each query on the manually transcribed documents, rather than the spoken ones.

Spoken document collections usually have metadata as text information alongside the spoken documents. On one hand, the text metadata is deterministic, very limited in size, and very likely differs from the actual spoken transcription which may limit its relevance to the content of the document. On the other hand, the speech recognition output is a noisy representation of the underlying lexical content and therefore we need to deal with content document uncertainty. Consequently, an approach that optimally integrates these two sources of information is desirable. Some researchers propose a simple method for integrating text metadata and speech content for the spoken document retrieval problem and they investigate how much performance gain is provided by the spoken document material with respect to a baseline system that uses only the text-metadata for document search.

The main research effort aiming at spoken document retrieval (SDR) was centered around the SDR-TREC evaluations [1], although there is a large body of work in this area prior to the SDR-TREC evaluations, as well as more recent work outside this community. Most notable are the contributions of [2] and [3]. One problem encountered in work published prior or outside the SDR-TREC community is that it does not always evaluate performance from a document retrieval point of view—using a metric like Mean Average Precision (MAP) or similar—but rather uses word-spotting measures, which are more technology- rather than user-centric. The TREC–SDR 8/9 evaluations—([2], Section 6)—focused on using Broadcast News speech from various sources: CNN, ABC, PRI, Voice of America. About 550 h of speech were segmented manually into 21,574 stories each comprising about 250 words on the average. The approximate manual transcriptions— closed captioning for video—used for SDR system comparison with text-only retrieval performance had fairly high word error rate (WER): 14.5% for video and 7.5% for radio broadcasts. ASR systems tuned to the Broadcast News domain were evaluated on detailed manual transcriptions and were able to achieve 15–20% WER, not far from the accuracy of the approximate manual transcriptions.

In order to evaluate the accuracy of retrieval systems, search queries—‘‘topics’’—along with binary relevance judgments were compiled by human assessors for each of the 21,574 retrieval units—‘‘documents’’. SDR systems indexed the ASR 1-best output and their retrieval performance—measured in terms of MAP—was found to be flat with respect to ASR WER variations in the range of 15–30%. Simply having a common task and an evaluation-driven collaborative research effort represents a huge gain for the community. There are shortcomings, however, to the SDR-TREC framework. The recognizers are heavily tuned for the domain leading to very good ASR performance. It is well known that ASR systems are very brittle to mismatched training/test conditions and it is unrealistic to expect error rates in the range 10–15% when decoding speech mismatched with respect to the training data. It is thus very important to work at an ASR operating point which has higher WER. This work used a standard dictation ASR engine whose language model has been trained on newswire text and the acoustic model was trained on wide-band continuous speech, resulting in an ASR operating point of 50% WER.

Also, the out-of-vocabulary (OOV) rate was very low, below 1%. Since the ‘‘topics’’/queries were long and stated in plain English rather than using the keyword search scenario, the query-side OOV (Q-OOV) was very low as well, an unrealistic situation in practice. Woodland et.al. (2000) [4] evaluates the effect of Q-OOV rate on retrieval performance by reducing the ASR vocabulary size such that the Q-OOV rate comes closer to 15%, a much more realistic figure since search keywords are typically rare words. They show severe degradation in MAP performance—50% relative, from 44 to 22.

The ability to deal in an effective way with OOV query words is an important issue. The most common approach is to represent both the query and the spoken document using sub-word units—typically phones or phone n-grams—and then match sequences of such units. In his thesis, Ng (2000) [5] shows the feasibility of sub-word SDR and advocates for tighter integration between ASR and IR technology. His approach was to index phone n-grams appearing in ASR N-best lists. This work also focused on Broadcast News speech, thus benefiting from unrealistically superior ASR performance. Similar conclusions are drawn by the work in [6].

As pointed out in [7], word level indexing and querying is still more accurate and thus more desirable, were it not for the OOV problem. The authors argue in favor of a combination of word and subword level indexing. Another problem pointed out by the paper is the abundance of word-spotting false-positives in the sub-word retrieval case, somewhat masked by the MAP measure.

Similar approaches are taken by [8]; one interesting feature of this work is a two-pass system whereby an approximate match is carried out on the entire set of documents after which the costly detailed phonetic match is carried out on only 15% of the documents in the collection.

More recently, Saraclar and Sproat in [9] propose an approach that builds an inverted index from ASR lattices—word or phone (sub-word) level—by storing the full connectivity information in the lattice; retrieval is performed by looking up strings of units. This approach allows for exact calculation of n-gram expected counts but more general proximity information (distance-k skip n-gram, $k > 0$) is hard to calculate. No compression of the original lattice is achieved. Their evaluation is focused on word-spotting rather than document retrieval performance.

Siegler (1999) [6] and Saraclar and Sproat (2004) [8] show that making use of more than just the 1-best information— N-best lists, and full ASR lattices, respectively— improves retrieval accuracy.

In [10] Chelba et al. present the Position Specific Posterior Lattice (PSPL), a novel lossy representation of automatic speech recognition lattices that naturally lends itself to efficient indexing and subsequent relevance ranking of spoken documents. This technique explicitly takes into consideration the content uncertainty by means of using soft-hits. Indexing position information allows one to approximate N-gram expected counts and at the same time use more general proximity features in the relevance score calculation. In fact, one can easily port any state-of-the-art text-retrieval algorithm to the scenario of indexing ASR lattices for spoken documents, rather than using the 1-best recognition result.

Experiments in [10] are performed on a collection of lecture recordings—MIT iCampus database—show that the spoken document ranking performance was improved by 17–26% relative over the commonly used baseline of indexing the 1-best output from an automatic speech recognizer (ASR).

One major problem with using speech recognizer for delivering sequence of words to other modules in spoken dialog systems is its robustness. Another problem is the presence of unpredictable or unexpected words in incoming utterances [11], [12] and [13]. Therefore, a speech recognizer must cover a large number of vocabularies, as well as grammars, in order to support every sentence possibly spoken by the users. This also includes un-precedently used words and non-speech sounds, such as fillers, that are rather common in spoken languages.

In this work, we begin with an annotated audio training set where pure audio concepts are manually transcribed. By pure concepts we imply instances of audio with only one concept (such as speech, silence and short-pause) in the audio track. Specifically, we manually transcribe speech, and silence. Regions corresponding to each concept are segmented from the audio and low-level features are extracted. One obvious modeling scheme uses these features to train a HMM for each concept. We use the HMMs to generate an 1-best word at each audio frame.

To date, most word spotters use a set of Hidden Markov Models (HMM) for their components. The models are modeled from combinations of sound units, which, may be phonemes or words. They are, then, trained using a target language corpus. Many HMM parameters still need to be re-estimated for Arabic speech via different applications.

In this work, the HMM parameters are estimated for efficient Arabic phoneme recognition using HTK [27]. The HMM phoneme models are, then, connected into word lattice models. Experiments are conducted through audio content of instructional videos used at Cairo university.

## II. Available Related Products:

To date, the most prominent and comprehensive effort to build a digital library (DL) of digital video is the Informedia Project [14]-[16]. Informedia uses a variety of visual features (e.g., color, faces, text superimpositions) as well as textual features (e.g., speech to text transcripts) to make a large volume of digital video retrievable. The project has demonstrated the efficacy of many technical processes for organizing, searching, and scaling video DLs. While there has been substantial research on particular aspects of digital video retrieval, e.g., segmentation and feature detection [17], Informedia addressed many of the integration challenges in incorporating different research products into a demonstration DL.

Other important projects include IBM's CueVideo, which has been integrating a variety of segmentation, indexing, and user interface techniques developed in the Almaden and Watson labs [19], and the Digital Video Multimedia Group at Columbia [20], which has been engaged in several streams of work including efforts to automate video summaries [17]. The Multimedia Information Retrieval Group at Dublin City University has been developing the Físchlár Project, which provides broadcast video for the university community. This group has developed innovative user interfaces for the video repository [21]. The European Union's ECHO Project [22] is developing archives of historical footage from different European countries and has focused on creating metadata schemes and cross-language access techniques. Each of these large-scale projects draws upon substantial efforts by the engineering communities devoted to finding effective signal-processing techniques for digital video.

The Open Video Digital Library [23] aims to capitalize on advances in engineering as well as in library and information science to create usable services for the research and educational communities. In this work they described the primary goals of the Open Video Digital Library, its evolution and current status. They also provided overviews of the user interface research and user studies.

For search engines, many companies are using link text and the text surrounding the links to podcasts and videoblogs as a means to index their contents. Now, optimized audio files can be used for searching a web page content. Some podcast –specific search engines seem to have solved some of the searching/indexing problems. *Podscope* searches for the spoken words within the podcasts themselves. *Signing Fish* provides results by typing in podcast or bring back audio results. Other search engines include EveryZing, Blinx, BlogDigger and Lycos Audio searches.

EveryZing (formerly known Podzinger) allows the user to search podcasts in the same way as in the web. A word or phrase typed in will find relevant broadcast and highlights the segment of the audio in which they occurred. It *is not perfect* but it serves the user need at this time. It works with a speech-recognition software which transforms audio into words. Podscope [24] and Blinx [25] are sites that work in a similar way. They scour audio content for keywords by translating the audio into text and creating an index for quick searching. This is a step ahead of traditional search engines that can only identify keywords in a podcats's metadata such as headline and introductory notes which describe the audio file's general content. Podscope searches podcasts but scans only for sounds of syllable rather than full words. It has operated a keyword search engine for video and radio broadcasts since 1999. By far, the Blinkx service is a bit more extensive as it scours thousands of podcsat and offers search for 1 million hours of TV news video and the content of academic hours. Unlike, Blinkx current technology, BBN's technology lets EveryZing extract high-level concepts that originally might not have been searched for.

EveryZing underlying technology is composed of two basic technologies from BBN. The core speech-to-text system, called Byblos, has been funded by $50 million of research money on a series on government grants over the past 5 years. Using probabilistic machine learning algorithm, the system takes  1-minute to convert each minute of audio content into text. The second part of the technology is the algorithms that process the content of the text. BBN's natural language technology contains huge stores of phrases and words for context searching. This system has about 80% accuracy. The accuracy drops when the background music is present and if there are multiple people talking at once [5].

## III. Experimental Work and Results:

The automatic speech recognition (ASR) engine is trained using an Arabic Globalphone database from ERLA [26]. The provided database has approximately 4908 speech sentences from about 84 speakers (1.5GB recorded size). The recording sampling frequency is16 ks/s with 16 bits quantization.

The database is divided into two subsets; one for training and the remaining part is preserved for testing. The training set is composed of 3069 recorded waves.

Each recording is associated with an Arabic transcription composed from about 35 phonemes. For simplicity and lower complexity, similar sounds (e.g., vowels, long vowels) are merged, and then the common phonemes are selected for use as garbage models. The HTK toolkit [27] is used with 16 bits encoding and 16 k sampling rate. The frame size is 25 ms with overlapping of  10 ms. The feature vector is taken 39 mel frequency cepstral coefficients and their derivatives (MFCC). The signal energy is normalized since the application domain is,  in general, a recording signal or offline processing. Each phoneme is modeled with 3 states. The searching lattice or network is defined in HTK format.

An Arabic phoneme recognizer is trained on this database. Different training parameters are tried, specifically, the HMM initialization and the number mixture-Gaussian components. Table 1 summarizes the results of phoneme recognizer training.

The average recognition rate of Spanish phonemes reached the order of 65% in [28]. The results of this work can partially be compared to that of [28].

Table 1. Phoneme recognition rate

| Condition | accuracy |
|---|---|
| Multi-gaussian mixture (16) phoneme hmm model, with uniform phoneme initialization, and E-normalize over limited recording size | 62.5% |
| 1-gaussian mixture phoneme hmm model, with uniform phoneme initialization, | 46.72% |
| 1-gaussian mixture phoneme hmm model, with isolated phoneme initialization, | 45.41% |
| Multi-gaussian mixture (8) phoneme hmm model, with isolated phoneme initialization, | 53.19% |
| Multi-gaussian mixture (6) phoneme hmm model, with isolated phoneme initialization, | 52.38% |
| Multi-gaussian mixture (8) phoneme hmm model, with uniform phoneme initialization, | 53.20% |
| Multi-gaussian mixture (4) phoneme hmm model, with isolated phoneme initialization, | 51.1% |
| Multi-gaussian mixture (8) phoneme hmm model, with uniform phoneme initialization, and E-normalize | 53.26% |

A word recognizer is built based on the obtained phoneme recognizer. The searching dictionary contains about 63 words as a trial test. The word error rate approaches 25% which can be compared to the results in [10]. This WER order is not far from the accuracy of the approximate manual transcriptions. These results need more tests for larger dictionary content.

## IV. Discussions and Conclusions:

This work considers the problem of Arabic Speech Recognition (ASR) and word searching. The well-known HMM toolkit (HTK) is used with different training and recognition parameters. The study reveals that an order of 50% Arabic phoneme recognition can be achieved. The Arabic word searching approaches comparable rates with that found in literatures. However, this rate can be improved by incorporating some linguistic rules into the system.

## Acknowledgment

The authors are fully indebted for the valuable discussions and help provided by Prof. Dr. Waleed Fakhr, Arab Academy of Sciences & Maritime and Dr. Ahmed Hussien, Electronic Dept., Fac. of Engineering, Cairo University. We also acknowledge with sincere appreciation the support provided by ITIDA via its research grant.

## V. References:

[1] Garofolo, J., Auzanne, G., Voorhees, E., “The TREC spoken document retrieval track: a success story”. In: Proceedings of the Recherche d’Informations Assiste par Ordinateur: ContentBased Multimedia Information Access Conference, 2000. Available from: <www.citeseer.ist.psu.edu/garofolo00trec.html>.

[2] Brown, M.G., Foote, J.T., Jones, G.J.F., Jones, K.S., Young, S.J., “Open-vocabulary speech indexing for voice and video mail retrieval”. In: Proceedings of the ACM Multimedia 96, Boston, pp. 307–316, 1996.

[3] James, D.A., “The application of classical information retrieval techniques to spoken documents”, Ph.D. Thesis, University of Cambridge, Downing College. 1995.

[4] Woodland, P.C., Johnson, S.E., Jourlin, P., Jones, K.S., “Effects of out of vocabulary words in spoken document retrieval”. In: Proceedings of SIGIR, Athens, Greece, pp. 372–374. 2000.

[5] Ng, K., Subword-based approaches for spoken document retrieval. Ph.D. Thesis, Massachusetts Institute of Technology. 2000.

[6] Siegler, M.A., Integration Of Continuous Speech Recognition And Information Retrieval For Mutually Optimal Performance, Ph.D. Thesis, Carnegie Mellon University.

[7] Logan, B., Moreno, P., Deshmukh, O., “Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio”. In: Proceedings of the HLT, 2002. Available from: <www.citeseer.nj.nec.com/585562.html>.

[8] Seide, F., Yu, P., “Vocabulary-independent search in spontaneous speech. In: Proceedings of ICASSP 2004”., Montreal, Canada.

[9] Saraclar, M., Sproat, R., Lattice-based search for spoken utterance retrieval. In: HLT-NAACL 2004, Boston, Massachusetts, pp. 129–136, 2004.

[10] Chelba, C. Silva, J. and Acero, A. "Soft indexing of speech content for search in spoken documents", Computer Speech and Language, vol. 21, pp. 458–478, 2007.

[11] K. Koumpis and Steve Renals, "Content-based access to spoken audio", IEEE signal processing magazine, pp. 61-69, Sep. 2005.

[12] Ying Li, and Chitra Dorai, “Instructional Video Content Analysis Using Audio Information”, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, No. 6, pp. 2264-2274, 2006.

[13] Atsushi F., Katunobu I., Tetsuya I., “LODEM: A system for on-demand video lectures”, Speech communication, El-Sevier S.N.H., vol. 48, pp. 516–531, 2006.

[14] Christel, M., Winkler, D. & Taylor, C.R. “Improving Access to a Digital Video Library”. INTERACT97, THE 6TH IFIP CONFERENCE ON HUMAN-COMPUTER INTERACTION, Sydney, Australia, July 14-18, 1997.

[15] Christel, M., Smith, M., Taylor, C.R., & Winkler, D. “Evolving video skims into useful multimedia abstractions”. PROCEEDINGS OF CHI '98: HUMAN FACTORS IN COMPUTING SYSTEMS, pp171-178, (Los Angeles, April 18-23, 1998).

[16] Witbrock, M. & Hauptmann, A. “Artificial intelligence techniques in a digital video library”. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, vol. 49, No. 7, pp. 619-632. (1998).

[17] Chang, S., Smith, J., Meng, H., Wang, H., & Zhong, D.”Finding images/video in large archives”, D-LIB MAGAZINE (February): 1997 <http://www.dlib.org/dlib/february97/columbia/02chang.html>.

[18] Benesty, J., Sondhi, M. and Huang, Y. (EDs), Springer Handbook of Speech Processing, Springer Berlin-Heidelberg 2008.

[19] Ponceleon, D., Amir, A., Srinivasan, S., Syeda-Mahmood, T., & Petkovic, D. (1999). “CueVideo: Automated multimedia indexing and retrieval”, ACM MULTIMEDIA '99 (Orlando, FL, Oct. 1999). p. 199.

[20] See the Digital Video Multimedia Group at Columbia University at <http://www.ctr.columbia.edu/dvmm>.

[21] Lee, H. & Smeaton, A. (2002). “Designing the User-Interface for the Físchlár Digital Video Library”, JOURNAL OF DIGITAL INFORMATION, 2(4), Special Issue on Interactivity in Digital Libraries, May 2002.

[22] The home page for the European Union's ECHO Project is at <http://pc-erato2.iei.pi.cnr.it/echo>

[23] The Open Video Digital Library (OVDL) is at <http://www.open-video.org>.

[24] <http://www.Blinx.com>

[25] D. Wirken, "Podcasting search engine", <http://www.theinternote.net>

[26] ERLA, ELRA-S0193, GlobalPhone Arabic, ELDA S.A., 2006.

[27] Young, S. Kershaw, D. Odell, J. Ollason, D. Valtchev, V. and Woodland, P. The HTK Book, (for HTK Version 4.30), Cambridge University Engineering Department, England, 2006. http://htk.eng.cam.ac.uk/docs/docs.shtml.

[28] Tejedor, J., Dong W., Joe F., Simon K. and Jose´ C.., “A comparison of grapheme and phoneme-based units for Spanish spoken term detection” ., Speech Comm. (2008), doi:10.1016/j.specom.2008.03.005

# من مشكلات التحليل النصي للمحتوى العربي على شبكة الإنترنت د.سلوى

السيد حمادة　　　　عمرو جمعة

معهد بحوث الإلكترونيات

hesalwa@hotmail.com　　　amr1979go@yahoo.com

## ملخص

تعد نظم البحث واسترجاع المعلومات أحد أهم الطرق التي نلجأ إليها للتغلب على مشكلة فوضى المعلومات أو ما يمكن أن نعبر عنه بتوافر وتضخم الإنتاج الفكري باللغة العربية على الشبكة العالمية. إذ إن هذا التضخم – إن لم نستطع أن نطوعه للاستفادة منه – لربما أدى إلى قتل المعرفة بدلاً من تنميتها.

ولن نعمل هنا على إلقاء الضوء على نظم البحث استرجاع المعلومات، أو تقديم وجهات النظر المختلفة لأنواع نظم استرجاع المعلومات، ولكن سنركز هنا على مشكلات البحث والاسترجاع على الشبكة العالمية للنصوص العربية من الناحية اللغوية وسنخص بالاهتمام قضية اللبس.

وهو موضوع تعرض له الكثيرون من الباحثين، إلا أننا في هذا البحث سنحاول شرح تأثير هذه المشكلة على أشهر وأهم التطبيقات اللغوية المتداولة والمعروفة، وتوضيح طرق الحل والمعالجة لكل واحدة منها على حدة.

## 1. مقدمة

من منا لم يستخدم مواقع الترجمة الآلية أو محركات البحث – على كثرتها – ومن منا لم يدرك الهوة الواسعة بين مخرجات هذه المواقع وتلك المحركات وبين ما يجب أن تكون عليه النتائج المرجوة – ذلك فيما يخص العربية . ولا شك أن هذه المواقع وتلك المحركات لا تزال في طور البداية في معالجة اللغة العربية، وأنها لا تزال تحاول حل مشكلات التحليل النصي لتلك النتائج التي أخرجتها محركات البحث ومواقع الترجمة والتي تمتد لتشمل مستويات اللغة جميعها: صرفا ونحوا ودلالة .

والحق أنها مشكلات تتسم بالتعقيد؛ نظرا لاختلاف العقل الذي يتعامل مع اللغة هذه المرة، فالعقل البشري – حين القيام بالترجمة من لغة إلى أخرى – يتحرك في إطار معرفته اللغوية السابقة أو ما أطلق عليه تشومسكي – المقدرة اللغوية – فضلا عن الخلفية الثقافية التي يعتمدها العقل البشري حين التعامل مع اللغة والتي يفتقد الحاسوب إليها، كذا الأمر في محركات البحث واعتمادها على حروف الكلمات وأشكالها المطابقة لا على دلالاتها اللغوية.

## 2. من المشكلات اللغوية لتطبيقات اللغة العربية على شبكة الإنترنت

### 1. محركات البحث ونظم استرجاع المعلومات في المحتوى العربي

تعد المطابقة هي السمة المميزة للبحث في محركات البحث المختلفة الموجودة على شبكة الإنترنت، فالبحث عن كلمة "مطعم" يؤدي بنا إلى نتائج هذه الكلمة مطابقة بنفس حروفها التي كتبت بها، وهي في ذلك تهمل بشكل كبير إحدى سمات العربية المهمة؛ التصريف ، فالكلمة الواحدة لها أكثر من شكل، فقد تكون مجردة عن أية سوابق أو لواحق، وقد تلحقها السوابق واللواحق، ومحرك البحث يعتمد في نتائجه على المطابقة في الشكل، ففي حال الكلمة المجردة عن أية سوابق أو لواحق "مطعم" مثلا، تأتي النتائج مهملة الصور الأخرى مثل المطعم، المطاعم، مطاعمنا، مطاعمهم، ... إلى آخر تلك النتائج التي قد تكون مهمة لبعض الباحثين.

لكن الأمر لم يدم على ذلك طويلا، فقد تطور الأمر بعد ذلك عن طريق برامج التحليل الصرفي ليتم البحث بالجذر لا بشكل الكلمة المطابق، ثم بالكلمة وسوابقها ولواحقها، وخرجت علينا بعض البرامج والمواقع التي تعتمد الجذر في أساسا في عملية البحث، والذي يثري بدوره عملية البحث ويعمل على زيادة النتائج.

ثم إن الأمر لم يقف عند هذا الحد بل تعداه إلى الصور الكتابية الخاطئة، بل إلى التدقيق اللغوي للمدخلات، وهو ما نلحظه في محرك البحث الشهير جوجل، فهو إما أن يقترح أشكالا أخرى صحيحة للبحث أو يورد في نتائج البحث كل الاحتمالات الصحيح منها والخطأ.

فكلمة: القاهره – بالهاء لا بالتاء المربوطة إذا ما تم الاستعلام عنها في محرك البحث جوجل، فإنه يخرج النتائج التي تخص"القاهره" بالهاء و"القاهرة" ، بالتاء المربوطة .

لكن مشكلة كبيرة لم تحل إلى الآن، ألا وهي مشكلة الغموض أو اللبس فالكلمة المدخلة قد تتضمن أكثر من معنى، ومن الواجب تحديد المعنى المقصود بواسطة المستخدم، ويقع العبء الأكبر بعد ذلك على محرك البحث في اختيار النتائج وعرضها تبعا للمعنى المقصود، وهو أمر غير موجود في كل المحركات الموجودة حاليا، ويأمل مشروعنا في تحقيق هذا الأمر، فيستطيع التمييز بين احتمالات كلمة "حامل" والتي تحمل معاني كثيرة: امرأة حامل أي: صفة للمرأة – حامل الملابس أي: شماعة – حامل أي: اسم الفاعل من "حمل" .

هذا ونعرض فيما يلي بعض المشكلات التي تواجهها العربية في نظم البحث واسترجاع المعلومات:

### 2. من المشكلات التي تواجهها اللغة العربية في نظم البحث واسترجاع المعلومات:[1]

إن من أهم العلاقات التي يجب أن تنبني عليها آلية البحث باللغة العربية: الترادف والمشترك اللفظي والتي يمكن أن تثري البحث باللغة العربية إذا ما تمت معالجتها.

**1] الترادف**

فالبحث عن كلمة ما في محركات البحث لا يتضمن البحث عن مرادفاتها.

ومن أمثلة ذلك: قوانين ، تشريعات ، أنظمة ، أحكام

حقول البترول ، آبار البترول

البترول ، النفط

البنوك ، المصارف

الحاسوب ، الحاسب ، الحاسب الآلي ، الكمبيوتر

الأسرة ، العائلة

الغيث ، المطر

النبأ ، الخبر

لهذا فإننا نريد عند البحث واسترجاع أيا من المفاهيم السابقة أن نسترجع أيضا الوثائق التي تحتوي مترادفاتها.

**2] المشترك اللفظي**

وهو أن يكون للفظ الواحد أكثر من معنى، و هو قليل جداً في اللغة، ومثاله: العين، التي هي في الأصل عضو الإبصار؛ لأن الدمع يجري منها كما يجري الماء، أو أنها للمعانها وما يحف بها من أهداب تشبه عين الماء التي تحف بها الأشجار، والعين من أعيان الناس، وهم وجهاؤهم، والعين بمعنى الإصابة بالحسد؛ لأن العين هي المتسببة في هذه الإصابة ... وما إلى ذلك من معان أخرى.

لكن المشترك اللفظي في علم اللغة الحاسوبي قد يأخذ شكلا مختلفا، فالكلمة نفسها تأخذ تعريفا مختلفا، **حيث يمكن تعريف الكلمة عند الحاسوبيين:** بأنها مجموعة من الحروف المتراصة الواقعة بين مسافتين ، سواء أفادت معنى مركبا أو معنى مفردا أو لم تفد ، أو كانت صحيحة أو خطأ.

في الوقت الذي يعرفها اللغويين العرب بأنها على حد قول الزمخشري ( ت 538 هـ) : «الكلمة هي اللفظة الدالة على معنى مفرد بالوضع » [2] ،

وتعريف ابن معطي لها ( ت 628 هـ ) بأنها « اللفظ المفرد الدال على معنى مفرد » [3] .

فالمشترك اللفظي هنا سيختلف وفق تعريف الكلمة عند الحاسوبيين، يقول [1] المشترك اللفظي: يقصد به اللفظ الواحد الدال على معنيين مختلفين أو أكثر ، بحيث تأخذ الكلمة الواحدة في اللغة الطبيعية عدة معان ومفاهيم بحيث لا يمكن تمييز المعنى الصحيح أو المستهدف للكلمة في السياق ، مثلاً :

الدين ( الإسلام )

الدين ( القرض )

العلم ( الحقل أو التخصص )

العلم ( الراية )

السنة ( مذهب )

السنة ( العام )

السنة ( برهة من النوم )

فنحن إذن نريد آلية لفك اللبس بين المعاني المختلفة للكلمة الواحدة حتى نستعيد الوثائق المتعلقة بموضوع المعنى المقصود فقط.

**3] التركيب :**

تكثر المفاهيم المركبة سواء في اللغات الطبيعية أو الاصطناعية ، ومن أمثلتها :

اختزان المعلومات واسترجاعها

الطب عند العرب

التأمين ضد الحريق

تعليم المرأة

صحة الطفل

كذلك يلاحظ هنا تغير دلالة المفهوم المركب بسبب تبادل مواقع بين المضاف والمضاف إليه، مثل :

إدارة العلوم . علوم الإدارة

كليات المكتبات . مكتبات الكليات

سجلات الإعارة . إعارة السجلات

تصنيف الوظائف . وظائف التصنيف

**ومن التراكيب كذلك العنونة بصيغة الفعل ( وبخاصة في عناوين الكتب وغيرها من مصادر المعلومات) :**

فقد تأتي المفاهيم أو التراكيب على صيغة أفعال ، مثل :

كيف تصوم ؟

هكذا نصوم .

كيف تدعو الناس ؟

كيف تحج ؟

فهنا يلاحظ أن الكلمات الدالة ، هي المشتقة من الأفعال : تصوم ، تدعو ، تحج . فالباحث الذي يرغب في كتب عن المفاهيم السابقة قد لا يتوقع بأنها تأتي في صيغ الأفعال .

**4] الإملاء والرسم الإملائي:**

تتفاوت صور كتابة بعض الكلمات وبخاصة المعرَّبة من مجتمع إلى مجتمع مما يؤثر على كفاءة البحث والاسترجاع ، ومنها مثلاً :

ببليوجرافيا . ببليوغرافيا

جوجل . غوغل

المسؤولية . المسئولية

شؤون . شئون

كمبيوتر . كومبيوتر

رومانتيكية . رومانسية . رومانطيقية

حاسب . حسّاب . حاسوب

فالمطلوب في الآلية مراعاة هذه الاختلافات والبحث عن أية وثائق تحتوى أيا من صورها.

ومن اختلاف صور الكتابة أيضا :

**5] تفاوت كتابة الأسماء المترجمة:**

فهناك تفاوت عند كتابة أسماء الهيئات والأشخاص الأجنبية بالحروف العربية، مثلا:

الاسم ( Michael ) يكتب بعدة صيغ هكذا :

مايكل ، ميكائيل ، ميشيل ، ميخائيل ...

وعند كتابة الأسماء العربية ، مثل الاسم "Mohamad " فإنه يكتب بعدة صيغ تكاد تكون مختلفة بالنسبة للحاسب الآلي .

فالمطلوب في الآلية مراعاة هذه الاختلافات والبحث عن أية وثائق تحتوى أيا من صورها.

**6] كلمات التوقف:**

من المتفق عليه في نظم الاسترجاع أن يتم استبعاد كلمات التوقف أو الكلمات غير الدالة موضوعياً ، مثل :

أدوات الربط وحروف المعاني ، مثل : حروف الجر ، وأدوات التوكيد ، والكلمات غير الدالة الموضوعيا كـ : نحو ، مدخل ، مقدمة ، أساسيات...

ولكن هنا ملحوظة مهمة تخص اللغة العربية، وهي : أننا نجد في بعض الأحيان أن الكلمة نفسها ترد دالة ، وغير ذات دلالة وفي أحيان أخرى ، مثل كلمة "نحو" في السياقات التالية :

نحو إدارة علمية جديدة . نحو اللغة العربية وقواعدها – المدخل إلي علم النحو والصرف .

**7] طريقة كتابة الأعداد والأرقام:**

. أربعون درساً في قواعد اللغة العربية

. العالم سنة 2003

. مائة سؤال عن الإعلام

كانت هذا عرضا لبعض المشكلات التي تعرض لمحركات البحث الموجودة على شكبة الإنترنت، ونقدم فيما يلي عرضا للمشكلات التي تواجه تطبيقا من التطبيقات اللغوية الموجودة على شبكة الإنترنت وهي برامج الإحصاء اللغوي، أو ما يسمى بـ **Concordance.**

### 3. مشكلات اللغة العربية على برامج الإحصاء اللغوي Concordance

**[2] <u>مشكلة الصور المختلفة للدلالة الواحدة</u>**

**1] الأعلام المترادفة**

فوجود أكثر من صورة للعلم الواحد ( الأعلام المترادفة ) في اللغة من المشاكل التي تواجهنا عند معالجة اللغة آليا ، و ( concordance ) يعتبر كل صورة من صور العلم علما منفصلا قائما بذاته. ومن أمثلة هذه المشكلة :

(أسامة بن لادن) ، (بن لادن) ، (ابن لادن) ، (قائد تنظيم القاعدة) ، (زعيم تنظيم القاعدة) ، (زعيم القاعدة)

( أمريكا ) ، ( الولايات المتحدة ) ، ( الولايات المتحدة الأمريكية ) ، (أميركا) ، ( الولايات المتحدة الأميركية )

( إنجلترا ) ، ( المملكة المتحدة ) ، ( بريطانيا )

( بوش ) ، ( جورج بوش ) ، ( جورج بوش الابن ) ، ( جورج دبليو بوش ) ، ( بوش الابن)

(2004 ) ، (عام 2004) ، (سنة 2004) ، (ألفين وأربعة)

(أحداث سبتمبر عام 2001) ، (أحداث الحادي عشر من سبتمبر) ، (أحداث 11 سبتمبر)

**2] مشاكل الأخطاء الإملائية**

للكلمات العربية صور مختلفة في الكتابة تبعا لمستوى الكتابة، حيث تأتي بعض الكلمات في صور إملائية مختلفة من بينها الصورة الصحيحة ،

وتتمثل الأخطاء في

- الكلمات التي تحتوي همزة الوصل : كـ استعلام ، إستعلام
- أو القطع – فوقية مفتوحة : أيمن – ايمن
- أو مضمومة : كـ أسامة – اسامة
- أو تحتية مكسورة – : كـ إسلام – اسلام

- أو كانت الهمزة متوسطة باختلاف حالاتها على السطر : كـ رءوف في الكتابة المصرية ، أو على واو : كـ رؤوف في الكتابة الشامية
- هذا خلاف اختلاف حال الهمزة باختلاف الحالة الإعرابية : كـ انتهاء – انتهائه – انتهاؤه
- وكذا الكلمات المنتهية بياء المنقوص الشامية المنقوطة : كـ القاضي ، و الكلمات المنتهية بياء المنقوص المصرية غير المنقوطة : كـ القاضى .
- أو الألف المقصورة غير المنقوطة والتي تلتبس مع الكلمات المنتهية بياء المنقوص المصرية غير المنقوطة كـ : مرتضى – مرتضي
- وكذا الكلمات المنتهية بالتاء المربوطة والكلمات المنتهية بالهاء : كـ مدرسة – مدرسه

### 3] السوابق واللواحق

اللغة العربية من اللغات التصريفية الاشتقاقية، فالصورة الواحدة للكلمة قابلة للظهور في صور أخرى جديدة تكون فيها هي نفسها الكلمة الأصلية متضمنة السوابق واللواحق والحواشي الاشتقاقية.

فالكلمات : إدارته ، إدارتها ، إدارتهما ، إدارتهم ، إدارتهن ، هي صورة للكلمة إدارة ، ويتضاعف العدد إذا وضعنا السوابق المختلفة مثل حروف العطف المتصلة وحروف الجر المتصلة، وإدارته، فإدارته، لإدارته ...

## [2] <u>مشكلة الصورة الواحدة للدلالات المختلفة</u> (ظاهرة اللبس الحاسوبي الصرفي والدلالي)

### 1] االلبس الصرفي :

وتتجلى هذه المشكلة عند استخدام Stop List : فكلمات مثل "بين" ، "عن" ، "من" ، "على" (في الكتابة المصرية) ، "قبل" ، "بعد" ، "عبر" ، "غير" ، "مهما" ، "ظل" ، "هم" ، "حول" ، "دون" ، "قد". فكل من هذه الكلمات لها تحليل صرفي آخر مغاير، وبالتالي لا يمكن وضعها في قائمة كلمات التوقف.

ويخلط Concordance بين التحليلات الصرفية المختلفة للكلمات الملتبسة صرفيا، فكلمات "عَامَ" ، "عَامٌ" و "كَتَبَ" ، "كُتُبٌ" وغيرها من الكلمات، حيث لا يستطيع البرنامج التفرقة بينها.

### 2] اللبس الدلالي

حيث يكون للكلمة الواحدة الموجودة في نتيجة بحث واحدة معاني ودلالات مختلفة مثل كلمات:

"حامل" : امرأة <u>حامل</u> ، رجل <u>حامل</u> سيفه

"دقيقة" : <u>جزء من الثانية</u> ، امرأة <u>دقيقة</u> في كلامها ، هذه جزئية <u>دقيقة</u>

"دقيق" : يصنع الخبز من <u>الدقيق</u> ، الباحث <u>الدقيق</u> يراجع مراجعه كلها

## [3] مشكلة فصل التراكيب المتلازمة

### 1] الفعل الذي يختلف معناه باختلاف حرف الجر الداخل عليه

تختلف معاني بعض الأفعال والمشتقات في العربية باختلاف متعلقاتها من حروف الجر ومفعولاتها :

ومن أمثلة ما تختلف باختلاف حروف الجر المتعلقة بها:

**2] الفعل (رغب) ومشتقاته:**

- رغب في كذا : أي حرص عليه ، وطمع فيه
- رغب عن كذا : أي تركه ، وزهد فيه
- ورغب فلان إلى فلان : أي تضرع إليه ، وطلب منه

3] وكذا **الفعل (مال) ومشتقاته:**

- مال إلى كذا : أي أحبه ، وانحاز إليه
- مال فلان عن كذا : أي حاد وعدل عنه
- مال فلان على فلان : أي جار عليه ، وظلمه

وكذا من الأفعال التي تختلف معانيها باختلاف مفعولاتها

**فلا شك أن ترجمات الفعل (شرب) ومشتقاته إلى الإنجليزية ستختلف باختلاف مفعولاته:**

شرب الماء ، شرب الخمر ، شرب سيجارة ، شرب المر .

4] **الأعلام والمسكوكات والمتلازمات:** فالعلم "جورج بوش" ، والمسكوك "ضرب أخماسا في أسداس"، والمتلازم "صلى الله عليه وسلم" ، لا شك أنها كل كلمة فيها ستفقد دلالتها إن فصلت عن سياقها.

**والخلاصة التي يمكن استنتاجها** : أن برامج الإحصاء اللغوي ليست ذات جدوى كبيرة أو فائدة، عند معالجتها للغة العربية، بل لابد من تزويد مثل هذه البرامج بالمحللات الصرفية وبقواعد البيانات الدلالية، والتي يمكن من خلالها معالجة اللغة العربية وحل مشكلات تطبيقاتها.

## [4] الترجمة الآلية من العربية وإليها

حيث يعد تعدد المعاني والدلالات للشكل الواحد أحد أهم مشكلات الترجمة الآلية، فالآلة تعجز عن تحديد المعنى المطلوب في النص من بين العديد من المعاني التي تحملها الكلمة في العربية وهوما يسمى باللبس الصرفي، فضلا عن معرفة رتبة الفاعل من المفعول، أو المبتدأ من الخبر، وهو ما نعني به اللبس التركيبي، وهو ما سنبينه بالتفصيل عند حديثنا عن اللبس التركيبي.

لا شك أنه بعد عرض هذه المشكلات جميعا من خلال تطبيقاتها، قد تبين لنا أن **من أهم أسباب هذه المشكلات اللغوية؛ اللبس أو الغموض اللغوي، وهو الأمر الذي حدا بنا إلى دراسة ظاهرة اللبس اللغوي بشيء من التفصيل .**

## 4. اللبس وكيفية علاجه في النصوص العربية

إن من خصائص أي لغة الوضوح والإبانة، لكن قد يحدث أن يعتريها اللبس في مستوياتها المختلفة (الصوت، والصرف، والنحو، والدلالة)، وما من لغة في العالم إلا تشتمل على تراكيب صالحة لتعدد المعنى والدلالات.

ومن هنا تأتي أهمية هذا الموضوع؛ معالجة اللبس أو الغموض في المحتوى العربي على الإنترنت، ودراسة أسبابه وأنماطه وآثاره المختلفة، ومعالجة مشكلاته لاسيما في اللغة العربية المعاصرة، ثم الاستفادة بذلك في برامج معالجة اللغة العربية آليا، كبرامج الترجمة الآلية ومحركات البحث التي تتعامل مع المحتوى العربي على الإنترنت، وكذا برامج الإحصاء اللغوي السياقي.

### [1] معالجة اللبس الصرفي الموجود في المحتوى العربي على الإنترنت (عند استخدام التطبيقات اللغوية)

يمكن للبس الصرفي أن يتمثل جليا على مستويين اثنين:

### 1] مستوى النص مضبوط البنية (وهو أقل وجودا من النوع الثاني) .

ففي النص المضبوط يتم تقييد الكلمة بالضبط، وهي بذلك لا تحتمل إلا المعاني التي تندرج تحت الضبط الواحد. نحو كلمة: أكل

| الكلمة | الاحتمالات | |
|---|---|---|
| أَكَلَ | فعل ماض | تناوَلَ طعامه |
| أَكْلَ | مصدر ثلاثي | تناوُل الطعام |
| أَكْلَ | اسم جامد | طعام |
| أَكُلٌّ | همزة استفهام + اسم بمعنى جميع | أ + جميع |
| أَكَلَّ | همزة استفهام + فعل ماض | هل + يئس |

لاحظنا تعدد المعاني الصرفية للكلمة الواحدة وهو ما يسمى اللبس الصرفي، وهو ما يزول نهائيا بضبط الكلمة بالشكل فتحتمل كلمة " أكل ": الصور التالية: أَكَلَ أو أَكُلّ أو ... إلخ .

ولكن ذلك الضبط الذي يساعد في عملية فك اللبس لا يوجد إلا في النصوص التراثية كالقرآن الكريم والسنة النبوية وكتب الفقه الإسلامي، ولا يوجد في النصوص المعاصرة ، وهو ما

تصعب معه عملية فك اللبس الصرفي لهذه النصوص، لذلك فإن البدء بمعالجة النصوص التراثية هذه تعد خطوة جيدة جدا .

**2] مستوى النص غير مضبوط البنية**

فيكثر فيه اللبس الصرفي وتتعدد صوره لعدم تقييد الكلمة بالضبط، وهو ما يؤدي إلى استنباط معانٍ أخرى للصيغة الصرفية المحتمِلة لأكثر من صيغة صرفية واحدة. فكلمة "كتب"

| الكلمة | التحليل الصرفي |
|---|---|
| كَتَبَ | فعل ماض مبني للمعلوم |
| كُتِبَ | فعل ماض مبني للمجهول |
| كُتُبٌ | جمع كتاب |
| كَتَّبَ | فعل ماضي مشدد |

فغياب التشكيل في النص العربي المكتوب المعاصر يؤدي إلى تضخيم كل أنواع اللبس خاصة اللبس الصرفي.

هذا كله على مستوى الكلمة المجردة من السوابق واللواحق، حيث يزداد اللبس في حال دخول السوابق أو اللواحق على الكلمة، مثل كلمة : "التهم" ، والتي يمكن تحليلها كاسم مرة وفعل مرة أخرى ، حيث يمكن أن تكون فعلا ماضيا أو أن تكون جمع تكسير لـ "تهمة" .

| | |
|---|---|
| اِلْتَهَمَ | فعل ماض |
| التُّهَم | ال التعريفية + جمع تهمة |

هذا بالإضافة إلى أن هناك مستوى آخر من اللبس الصرفي، وهو ما ينشأ عن شيوع الأخطاء الإملائية في الكتابات العربية المعاصرة، ومنه الفعل المضارع "ألعب" التي يمكن أن تكتب "العب" على سبيل الخطأ فيلتبس حينئذ بالفعل الأمر "العب".

وقد أردت فيما سبق أن أعرض لأمثلة توضح مواضع اللبس الصرفي، وهي كثيرة ، وكيف يمكن لهذه المعاني والدلالات المختلفة أن تؤثر على النتائج المرجوة في برامج الترجمة الآلية ومواقعها ومحركات البحث.

**(1)من أسباب اللبس في البنية ما يلي[2]:**

1) تعدد الاحتمالات الصرفية للكلمة الواحدة المجردة.

2) تعدد الاحتمالات الدلالية للتحليل الصرفي الواحد المجرد .

3) تعدد الاحتمالات الصرفية للكلمة الواحدة المزيدة بالسوابق واللواحق.

4) تعدد الاحتمالات الدلالية للتحليل الصرفي المزيد بالسوابق واللواحق.

5) تعدد معاني الأبنية الواحدة ( كتعدد معاني حروف الجر ).

6) التحول الدلالي المعجمي لمعاني الكلمات الاشتقاقية إلى كلمات جامدة.

**(2)فك اللبس**

ويهدف بحثنا إلى دراسة كيفية التخلص آليا من هذا اللبس الصرفي من خلال عنصرين أساسيين:

1) **صياغة قواعد لغوية صرفية في شكل رياضي رقمي يمكن الحاسوب من فهمها**، ومنها على سبيل المثال لا الحصر:

علامات الفعل : ومنها السوابق المختصة بالفعل ، مثل "قد" التي تسبق الفعل دائما ، ففي جملة: قد لعب بالكرة، يمكن أن يكون لـ "لعب" ثلاثة تحليلات

| | |
|---|---|
| لَعِبَ | فعل ماض مبني للمعلوم |
| لُعِبَ | فعل ماض مبني للمجهول |
| لُعَب | اسم – جمع لعبة |

ولدخول "قد" عليها ، ينحصر تحليل "لعب" الوحيد في كونها "فعلا ماضيا" وينتفي عنها تحليل الاسمية ، فعن طريق هذه العلامات تم نفي تحليل من التحليلات الموجودة وهو الاسمية وترجيح احتمال آخر عليه .

ومن ذلك "سوف" ، والسين ، و"لم" ، و"لن" ، فهذه العلامات توجب كون ما بعدها فعلا في حال اللبس .

بل يمتد الأمر من الجانب الصرفي إلى الجانب الدلالي ، حيث إن دخول قد على الفعل المضارع تفيد الشك ، ودخولها على الفعل الماضي تفيد التأكيد

قد + فعل مضارع = معنى الشك ( ربما حدث )

قد + فعل ماض = معنى التأكيد ( حدث بالفعل )

والأمر كذلك بالنسبة لعلامات الاسم، فالسوابق المختصة بالاسم توجب تحليل ما بعدها اسما وليس فعلا وتخليصه من الاحتمالات المتعددة .

فدخول "إلى" ، "في" ، "على" ، على الكلمات التي بعدها توجب تحليلها أسماء ، وليست أفعالا، مثل : سلمت على أفضل الناس ، "أفضل" هنا لها تحليلان :

| | |
|---|---|
| أُفَضِّل | فعل مضارع مشدد |
| أَفْضَل | اسم تفضيل |

ودخول "على" على "أفضل" يوجب تحليلها للاسمية، لا للفعلية.

هذا عن علامات الاسم والفعل التي توجب ترجيح بعض الاحتمالات على الأخرى وهو ما يؤدي إلى فك اللبس .

**2) الاعتماد على قاعدة بيانات معجمية تضم ما يلي :**

- الأفعال ومتعلقاتها من: الفاعلين والمفعولين ، أو كل ما هو ذو ارتباط في اللغة ، أو ما يعرف بالحقول الدلالية ، وهو ما يمكن الاستفادة فيه من معجم حاسوبي تم إعداده من قبل. مثل :

الأفعال : قبض – حبس – سجن

يستلزم

الفاعلين : ضابط – شرطي – شرطة – ضباط

والمفعولين : مقبوض عليه – محبوس – مسجون

فكل من الفعل والفاعل والمفعول يرجح تحليلا واحدا لكل عنصر من هذه العناصر .

- قاعدة بيانات للأعلام Proper noun والكيانات Entities

فـ " أحمد " علم ، وفي نفس الوقت هو فعل مضارع

وفي الجملتين :

أحمد يلعب بالكرة ، أحمد هنا علم وليس اسما لوجود ما يمنع توالي فعلين في العربية .

أحمد الله على الخير ، أحمد هنا فعل مضارع لعدم وجود تركيب علمي لها مع لفظ الجلالة، مثل عبد الله ، سيف الله .

- بناء قاعدة بيانات إحصائية تتضمن درجة تكرار وشيوع استخدام هذه الاحتمالات من أجل اختيار التحليل السليم، أو المفاضلة بينه وبين تحليل آخر ممكن ، فكلمة مثل : المسلم ، لها تحليلات كثيرة

| | |
|---|---|
| المُسْلِم | اسم فاعل من أَسْلَم |
| المُسَلِّم | اسم فاعل من سَلَّم |
| المُسْلَم | اسم مفعول من أَسْلَم |
| المُسَلَّم | اسم مفعول من سَلَّم |

لاشك أن التحليل الأكثر شيوعا هو اسم الفاعل من أسلم ، تليه الاستخدامات الأخرى .

- قواعد بيانات المسكوكات والمتلازمات والتراكيب الاصطلاحية :

إنه جانب آخر من الجوانب التي يمكن تحليل الكلمات فيها يدويا وفق تحليل واحد واحتمال واحد صحيح من عدة احتمالات، مثل جمل : صلى الله عليه وسلم ، ضرب أخماسا في أسداس، ...

ومن ذلك جملة : صلى الله عليه وسلم :

ففي هذه الجملة هناك لبس دلالي في التحليل الصرفي الوحيد لكلمة "صلى" كالتالي :

صلى : فعل ماض مشدد له معنيان :

صلى فلانا النار : عذبه بها ، صلى فلان العصر : أدى الصلاة ، صلى الله عليه : وجه إليه التحية ( حيّاه ) .

فيكون الاختيار الأخير هو الصحيح ، وهو التحليل الثابت لهذه الكلمة في حال ورودها في هذا التركيب .

وكذلك كلمة : وسلم، فتحليلها كالتالي :

وسلم : حرف عطف + فعل ماض مشدد ( بمعنى حيا ) ، وسلم : حرف عطف + اسم آلة يصعد عليها ، وسلم : حرف عطف + مصدر سلم ( السلم عكس الحرب )

فيكون الاختيار الأول هو الصحيح ، وهو التحليل الثابت لهذه الكلمة في حال ورودها في هذا التركيب .

وهكذا من خلال قاعدة بيانات المتلازمات وتحليلاتها يمكن فك لبس كثير من الكلمات وكذلك كثير من التراكيب التي تتكرر في السياقات المختلفة

### 3) الاستفادة من أنماط الجمل العربية في فك اللبس الصرفي

فتوالي فعلين في العربية غير ممكن إلا في أفعال معينة هي أفعال الشروع أو الأفعال الناسخة وهو ما يمكن استثناؤه .

**فعل (ما لم يكن من أفعال الشروع أو الأفعال الناسخة ) + فعل = نمط غير موجود في العربية**

ودخول حروف الجر على الأفعال كذلك غير ممكن .

**حرف جر + فعل = نمط غير موجود في العربية**

ودخول أل التعريفية غير ممكن على الأفعال .

أل + فعل = **نمط غير موجود في العربية**

فهذه الأنماط وغيرها أنماط غير موجودة في العربية ، وكل احتمال صرفي يؤدي إلى هذه الأنماط لا شك أنه غير صحيح بالمرة .

**طرق التخلص آليا من اللبس الصرفي/ فك اللبس الصرفي:**

- صياغة القواعد اللغوية الصرفية في شكل رياضي رقمي
- بناء قاعدة بيانات معجمية عن كل ما هو مستعمل من أوزان الكلمات العربية وقواعدها
- بناء قاعدة بيانات للأعلام

- بناء قاعدة بيانات لأنماط الأخطاء الإملائية والنحوية الشائعة في النصوص، كالخلط بين همزتي القطع والوصل، والخطأ في رسم الهمزات المتوسطة والمتطرفة، والخلط بين التاء المربوطة والهاء، والألف المقصورة والممدود
- بناء بناء قاعدة بيانات لاحتمالات استعمال الكلمات العربية ومدى دورانها وشيوعها.

## التطبيق عملي للبس الصرفي وكيفية التخلص منه:

**Table1**

| ID | Word | التحليل الأول | التحليل الثاني | التحليل الثالث | التحليل الرابع | الاختيار | طريقة فك اللبس الصرفي |
|---|---|---|---|---|---|---|---|
| 1 | مؤتمر | اسم فاعل من ائتمر | اسم مفعول | اسم جامد | | اسم جامد | قاعدة بيانات تتضمن الصفة والموصوف ، فالموصوف في التحليل الأول والثاني لا يتعلق بالصفة "علمي" ولكن التحليل الأخير هو الوحيد المتعلق بالوصف " علمي" |
| 2 | علمي | نسبة إلى علم | نسبة إلى علم | مصدر مضاف إلى ياء المتكلم | فعل أمر للمخاطبة | نسبة إلى علم | |
| 3 | بالقاهرة | ب +اسم علم | اسم فاعل من قهر | | | | |
| 4 | حول | ظرف | فعل ماضي معلوم مضعف | فعل ماضي مجهول مضعف | اسم جامد بمعنى عام | ظرف | الفعل "حول" يتعلق بحرفي جر : لـ أو إلى ، وهما غير موجودان في الجملة ، والتحليل الرابع تؤخره قواعد الإحصاء ، ويترجح تحليل الظرفية حينئذ . |
| 5 | أدب | اسم جامد | مصدر | فعل ماض معلوم | فعل ماض مجهول | اسم جامد | من القواعد النحوية الحاسوبية : أن حول يأتي بعدها اسم أو مصدر، وهو ما ينفي تحليل الفعلية ويحصر اللبس في الاسم أو المصدر ، ومن الممكن الاعتماد على قاعدة بيانات التراكيب الإضافية التي تحتوي على هذا التركيب " أدب الطفل " |
| 6 | الطفل | اسم جامد | | | | | لا لبس فيها |
| 7 | والإعلام | مصدر | اسم جامد بمعنى الميديا | | | اسم جامد بمعنى الميديا | المصدر " الإعلام " يتعلق بحرف جر الباء ، فإذا لم يوجد ترجح اختيار الاسم الجامد |
| 8 | جانب | ظرف | فعل ماض | | | ظرف | تحليل الفعل الماضي فاعل ينبغي أن يكون الصواب ، وينبغي أن يتعلق بضمير : جانبه الصواب ، ويمكن اعتبار "جانب من الحضور" مسكوكا أو تعبير متلازم يفك لبسه يدويا مرة واحدة ويعتمد فيه على قاعدة بيانات. |
| 9 | من | حرف جر | فعل ماض | مصدر | | حرف جر | |
| 10 | الحضور | مصدر | جمع اسم فاعل | | | جمع اسم فاعل | |
| 11 | الجزيرة) | اسم علم | | | | | علم مركب |

**Table1**

| ID | Word | التحليل الأول | التحليل الثاني | التحليل الثالث | التحليل الرابع | الاختيار | طريقة فك اللبس الصرفي |
|---|---|---|---|---|---|---|---|
| 12 | نت) | اسم علم | | | | | |
| 13 | بدر | اسم علم | اسم جامد | | | اسم علم | علم مفرد |
| 14 | محمد | اسم علم | اسم مفعول | اسم فاعل | | اسم علم | علم مفرد ( توالي الأعلام المفردة يرجح تحليلها كأعلام |
| 15 | القاهرة | اسم علم | اسم فاعل مؤنث | | | اسم علم | |
| 17 | ناقش | فعل ماض | فعل أمر | اسم فاعل | | فعل ماض | التحليل الثاني ينفيه المفعول به المؤتمر إذ لا يمكن أن يكون المؤتمر شخصا يُنَاقَش ، وكذلك لا يصح التحليل الثالث لعدم صحة الإضافة "ناقش المؤتمر" اسم فاعل (ناقش) + اسم (المؤتمر) . |
| 18 | المؤتمر | اسم فاعل من ائتمر | اسم مفعول | اسم جامد | | اسم جامد | |
| 19 | العلمي | نسبة إلى علم | نسبة إلى علم | مصدر مضاف إلى ياء المتكلم | | نسبة إلى علم | |
| 20 | الرابع | اسم فاعل صفة (رقم) | | | | | |
| 21 | لمركز | اسم مكان +ل | اسم فاعل +ل | اسم +ل مفعول | | اسم مكان +ل | مركز بحوث ( لقب ) قاعدة بيانات الألقاب |
| 22 | بحوث | جمع اسم جامد | جمع مصدر | | | جمع اسم جامد | |
| 23 | أدب | اسم جامد | مصدر | فعل ماض معلوم | فعل ماض مجهول | اسم جامد | |
| 24 | الأطفال | اسم جامد | | | | | |
| 25 | بجامعة | اسم جامد +ب | اسم فاعل +ب مؤنث | | | اسم جامد +ب | قاعدة بيانات الألقاب والأعلام |
| 26 | حلوان | اسم علم | مثنى الوصف حلو | | | | |
| 27 | بالقاهرة | اسم علم+ب | اسم فاعل من قهر | | | | |

**Table1**

| ID | Word | التحليل الأول | التحليل الثاني | التحليل الثالث | التحليل الرابع | الاختيار | طريقة فك اللبس الصرفي |
|---|---|---|---|---|---|---|---|
| 28 | على | حرف جر | | | | | |
| 29 | مدار | ظرف | اسم مفعول | | | ظرف | |
| 30 | يومين | زمان | | | | | |
| 31 | أكثر | فعل ماض | تفضيل | | | تفضيل | |
| 32 | من | حرف جر | مصدر | فعل ماض | | حرف جر | |
| 33 | عشرين | رقم | | | | | |
| 34 | بحثا | اسم جامد | مصدر | فعل ماض معلوم للمثنى | فعل ماض مجهول للمثنى | اسم جامد | قاعدة نحوية : التمييز اسم جامد وليس فعلا أو مصدرا |
| 35 | عبر | ظرف | فعل ماض | جمع اسم جامد | | ظرف | |
| 36 | ثلاثة | رقم | | | | | |
| 37 | محاور | جمع اسم جامد | اسم فاعل | اسم مفعول | | جمع اسم جامد | |
| 38 | أساسية | صفة منسوبة | | | | | |
| 39 | تناولت | تاء +فعل ماض الفاعل | تاء +فعل ماض التأنيث | | | | |
| 40 | أدب | اسم جامد | مصدر | فعل ماض معلوم | فعل ماض مجهول | اسم جامد | |
| 41 | الأطفال | اسم جامد | | | | | |
| 42 | ولغته | اسم +حرف عطف ضمير +جامد غائب | +فعل ماض ضمير غائب | | | | |

**[2] اللبس التركيبي في المحتوى العربي على شبكة الإنترنت**

**1] أسباب اللبس التركيبي في المحتوى العربي وجود الظواهر اللغوية التالية في اللغة العربية:**

1- التقديم والتأخير

2- الحذف

3- مرجعية الضمير

أ- في الإسناد الخبري

ب-في التركيب البدلي أو الوصفي

ت-في جملة صلة الموصول

ث-في أسلوب العطف

**فمن قواعد عود الضمير على مرجعه :** [3]

1-عود ضمير الغائب على أقرب مذكور :

" يقول علماء إنهم توصلوا إلى طريقة سريعة لتصنيع مصل أنفلونزا الطيور"

فالهاء في " إنهم " تعود إلى " علماء "

ما لم يكن هناك دليل آخر يرجح عوده على غير الأقرب من:

أ- دليل من اللغة ، ومنه عود ضمير التأنيث إلى المؤنث لعدم جواز عوده على المذكر

" وقال المتحدث إن الفحوصات هي "إجراء وقائي" وإنها سوف تستغرق عدة ساعات"

فضمير التأنيث في " إنها " يعود إلى " الفحوصات " المؤنثة ، ولا يعود إلى "إجراء وقائي" أو " المتحدث " المذكرتين .

ومنه كذلك :

"ووجد الباحثون أن فيروس الأنفلونزا الذي فحصوه كان قريبا جدا من تلك السلالة المميتة"

ب- دليل من المقام أو السياق ( من السمات الدلالية )

" أصابت مركبة أمريكية طفلة عراقية كان يقودها جنديان مخموران "

فالضمير في " يقودها " يعود إلى " مركبة أمريكية " لا إلى " طفلة عراقية.

إلى غيرها من القواعد التي يمكن من خلالها إرجاع الضمير إلى مرجعه.

4- الإسناد إلى الفاعل أو المفعول

5- الإضافة إلى الفاعل أو المفعول ( كإضافة المصدر إلى فاعله أو إلى مفعوله ، والإضافة إلى الموصوف قبل الإتيان بالصفة ، والإضافة إلى المعطوف عليه قبل تمام العطف )

6- تعدي الفعل ولزومه

7- الإطلاق والتقييد

8- تعلق الحدث ( الفعل – المشتق – المصدر ) بمكملاته أو بالمفاعيل من ( الحال أو الظرف أو المفعول به أو ... ) عند التقييد بها

9- المجاز

10-غياب علامات الترقيم

11-النبر والتنغيم ( كالأسلوب الخبري والإنشائي )

**2] معالجة اللبس التركيبي في المحتوى العربي على شبكة الإنترنت:**

اللبس التركيبي خاص بتقنيات الترجمة الآلية على وجه الخصوص، ويحدث عندما يكون هناك أكثر من معنى أو دلالة لتركيب واحد.

| التركيب | التحليلات المحتملة |
|---|---|
| أغنية الفتاة الجميلة | حيث يمكن أن تكون "الجميلة" صفة لل "الفتاة" أو لـ "أغنية". |
| قاض أمريكي يرفض دعوى فلسطينية ضد مسئول | إذ يمكن أن تكون "فلسطينية" صفة لـ دعوى – أي دعوى فلسطينية – أو اسما مضافا إليه لـ دعوى – أي دعوى امرأة |

| | |
|---|---|
| إسرائيلي | فلسطينية– . |
| فررت منه كما يفر الواحد من المجرمين | ف "من" هنا يمكن أن تكون ابتدائية – أي كما يفر أحدنا من المجرمين – أو تكون تبعيضية – أي كما يفر أحد المجرمين– . |
| إن زيارة الأصدقاء تسعد النفس | (فمن الزائر؟) ، أي زيارة الأصدقاء للمتكلم أو زيارة المتكلم للأصدقاء |
| ذهبت لزيارة أبناء زيد وعمرو | فهل ذهبت إلى عمرو أو إلى أبنائه |
| أعجبت بمعلمة اللغة العربية | فهل العربية هي المعلمة أو اللغة؟ |

**5. وسائل معالجة اللبس أو الغموض التركيبي (من خلال تطبيقات معالجة اللغات الطبيعية):**

**[1] القواعد النحوية لفك اللبس ومنها[4]:**

1. التضام (ويشمل الافتقار والاختصاص والاستغناء والمناسبة المعجمية بين المفردات)...

كافتقار حرف الجر إلى مجرور، وحرف العطف إلى معطوف، والفعل إلى فاعل، والموصول إلى صلة... إلخ. والاختصاص كاختصاص المضارع بدخول "لم" للنفي وتحويل الزمن النحوي إلى الماضي، واختصاص الفعل اللازم بواحد من حروف الجر من أجل التعدية.

والمناسبة المعجمية تتمثل في تقبل قولك: فهم التلميذ الدرس، ورفض عبارة: فهم الماء القمر؛ لأن الفعل " فهم" يتطلب فاعلاً عاقلا وليس الماء كذلك.

2. الرتبة وقد تكون محفوظة كرتبة الحروف والأدوات من مدخولاتها، وقد تكون غير محفوظة كرتبة الفعل والمفعول به، ورتبة المبتدأ أو الخبر.

3. الربط (إما بالمطابقة وإما بالإحالة). والربط قد يكون بالمطابقة في العدد (الإفراد والتثنية والجمع)، والشخص (المتكلم والمخاطب والغائب)، والنوع (التذكير والتأنيث)، والتعيين (التعريف والتنكير) وفي الإعراب. وقد يكون بالإحالة. والأصل في الإحالة إعادة الذكر نحو: رأيت سائلا فأعطيت السائل، ومنه إعادة صدر الكلام بعد طول الشقة، نحو: ﴿ثم إن ربك للذين هاجروا من بعد ما فتنوا ثم جاهدوا وصبروا إن ربك من بعدها لغفور رحيم﴾ (النحل 110). وقد تكون الإحالة بالضمير، نحو: ﴿ونادى نوح ابنه﴾ (هود 42)، وبالإشارة، نحو: ﴿ولباس التقوى ذلك خير﴾ (الأعراف 26)، وبالموصول، نحو: ﴿ولو نزلنا عليك كتابا في قرطاس فلمسوه بأيديهم لقال الذين كفروا إن هذا إلا سحر مبين﴾

(الأنعام 7)، أي لقالوا، وبالألف واللام، نحو: ﴿ومأواهم النار وبئس مثوى الظالمين﴾ (آل عمران151) أي مثواهم، وبالوصف، نحو: ﴿وإن نكثوا أيمانهم من بعد عهدهم وطعنوا في دينكم فقاتلوا أئمة الكفر﴾ (التوبة 12) أي فقاتلوهم.

4. البنية (وتشمل أقسام الكلم والصيغ الصرفية ومعانيها والأدوات والإجراءات التصريفية). وتتناول قرينة البنية دلالات أقسام الكلم والصيغ الصرفية (مبنى ومعنى) وإجراءات تحولاتها، وكذلك الأدوات وحروف المعاني ووظائفها في السياق.

5. الإعراب (ويكون بالعلامة أو بالمعاقبة).

فهو أشهر ما تناوله النحاة في منهجهم، فبدا كأن النحو هو الإعراب وأن ما يتناوله النحاة من القرائن الأخرى إنما يتوقف تناوله على المصادفة، هذا مع أن الإعراب لا يتناول من عناصر اللغة إلا ما كان منها صحيح الآخر، ويبقى غير الصحيح الآخر (المقصور والمنقوص) والمبنيات والجمل الفرعية غير ذات صلة مباشرة بالعلامة الإعرابية؛ وإنما يربطها بهذه القرينة عنصر المعاقبة.

فقولنا: هزم المناضلون المصريون اليهود، يوجب الإعراب كون المصريون صفة لـ "المناضلون" وينفي عنها أن تكون من باب المفعول به .

6. النغمة (في الكلام المسموع). والنغمة قرينة على المعنى في الكلام المنطوق، لأنها تختلف في الخبر عنها في الإنشاء، وفي الإثبات عنها في النفي، وفي الموافقة عنها في الإنكار، وفي الفرح عنها في الحزن، وفي الجد عنها في الهزل، وهكذا نجدها دليلا على المعنى في كل كلام مسموع. ومع أن النحاة العرب كانوا يدركون أن للنغمة أثرا في المعنى لم يكن لديهم من الحيل الفنية في التأليف ما يعينهم على نقل معلوماتهم عنها للأجيال القادمة، فعوضوا هذا القصور بالكلام في الوقت والسكت والفصل والوصل ونحو ذلك مما يتصل بإجراءات النطق. ومن واجب من يؤلف في حقل النحو أن يشمل هذه القرينة بالدرس بعد أن تعددت الوسائل الفنية لدراستها ونقل المعلومات عنها.

ويمكن الاعتماد على النغمة كوسيلة من وسائل فك اللبس من خلال علامات الترقيم ، فعلامة ! التعجب توجب كون أفعل للتفضيل وليس للفعلية مثلا.

7. دلالة السياق (وهي كبرى القرائن، وقد يدخل في تكوينها جميع ما تقدم وتنقسم إلى سياق النص وسياق الموقف) .

**[2] قواعد البيانات المعجمية التي تحتوي على:**

1- معجم الحقول الدلالية

2- معجم المتلازمات والقرائن المعجمية

**3**- معجم المسكوكات والتعابير الاصطلاحية

## 6. خطوات مشروع فك اللبس :

1- توافر مدونة لغوية ضخمة تمثل المحتوى العربي على الإنترنت ( حوالي 500 مليون كلمة ) في جميع مجالات المعرفة ( السياسة – الاقتصاد – الدين – الرياضة – الاجتماع – إلخ ) .

2- توافر محلل صرفي ومحلل نحوي تركيبي وكذلك محلل دلالي .

3- إعمال هذه المحللات في هذه المدونة .

4- استخراج مواضع اللبس المختلفة التي تعجز هذه المحللات عن حلها أو فهمها .

5- تصنيف ودراسة هذه المواضع ومحاولة اقتراح وسائل لحلها ، من خلال وسائل فك اللبس أو الغموض .

6- عمل تطبيق برمجي لنظريتنا في فك اللبس، حيث يمكن تركيبه على أي من هذه التطبيقات اللغوية (المحلل الصرفي، والمحلل النحوي (التشكيل الآلي أو الإعراب الآلي)، والفهم والترجمة الآلية، والقاموس الآلي، والمرادفات والأضداد والمصطلحات، وترجيح الاحتمالات الدلالية، والإملاء الصوتي، وتوليد الكلام، والقارئ الآلي، والبحث المعجمي ، والفهرسة الآلية) .

## 7. التعليق النهائي:

تمكنا في هذا البحث من عرض بعض مشكلات التحليل النصي للمحتوى العربي على شبكة الإنترنت – من خلال تطبيقاتها المختلفة؛ محركات البحث، وبرامج الإحصاء اللغوي، والترجمة الآلية. مع دراسة لأهم أسباب هذه المشكلات: اللبس، ومستوياته المختلفة، من حيث البنية والتركيب والدلالة. وقدمنا مقترحات عن كيفية علاجه في النصوص العربية، وقدمنا عرضا لبعض المشكلات التي تعرض لمحركات البحث الموجودة على شكبة الإنترنت، كما قدمنا عرضا للمشكلات التي تواجه تطبيقا من التطبيقات اللغوية الموجودة على شبكة الإنترنت وهي برامج الإحصاء اللغوي، أو ما يسمى بـ Concordance. ، وقد وتوصلنا إلى أن التطبيقات اللغوية تتضمن العديد من المشكلات وأنها في حاجة إلى مزيد من المعالجة الخاصة باللغة العربية، والملائمة لسمات وخصائص العربية. وختمنا البحث بمقترح مشروع لحل مشكلة اللبس في البحث على محركات البحث باستخدام اللغة العربية. ودعو للباحثين اللغويين والمطورين إلى ضرورة التعاون من أجل معالجة مثل هذه المشكلات، وكذلك ندعو الشركات التي قامت بجهد في هذا المجال إلى ضرورة تقديم الدعم لمثل هؤلاء الباحثين.

## 8. المراجع والمصادر والقراءات

[1] د. محمد سالم غنيم. النظم المحسبة للاسترجاع الموضوعي باللغة الطبيعية : دراسة تطبيقية على اللغة العربية .– إشراف محمد فتحي عبد الهادي .– القاهرة : م. س ، 2003 (أطروحة دكتوراه ، قسم المكتبات والوثائق والمعلومات ، جامعة القاهرة)

[2] سلوى حماده, "تمثيل المعلومات لفك اللبس", المؤتمر السادس لجمعية هندسة اللغة, سبتمبر، 2006،القاهرة.

[3] د. محمد حسنين صبرة، مرجع الضمير في القرآن الكريم، دار غريب ، القاهرة ، الطبعة الثانية ، 2001 .

[4] د. تمام حسان، تطوير التأليف في مجالات اللغة العربية، بحث منشور بمجلة مجمع اللغة العربية، القاهرة، 1998.

# Formal semantics: luxury or necessity?

Allan Ramsay
School of Computer Science,University of Manchester
Manchester M13 9PL, UK

**Abstract.** We will revisit the goals of natural language processing, and argue that thse goals cannot be achieved unless you construct fine-grained formal paraphrases (also known as 'logical forms'). We will then consider the problems involved in building such paraphrases, and review the connections between semantics and inference. We will close by seeing how natural language can be used for expressing detailed domain knowledge, thus avoiding the need for complex translations between formalisms.

## 1 What should a natural language understanding system do?

What, ideally, should a natural language understanding system be able to do? What tests would a system have to pass before you were prepared to accept that it 'understood' what you meant when you used natural language? What would it take for a system to pass the Turing test?

Roughly speaking, a reasonable test would be to see whether its view of what someone who produced a sentence in natural language could reasonably be expected to believe was the same as that of a native speaker. Suppose I said one of (1a) and (1b):

(1) a. I forgot to prepare any handouts for my lecture on Wednesday.
b. I forgot that I had prepared some handouts for my lecture on Wednesday.

Consider the sentences in (2):

(2) a. I had a lecture on Wednesday
b. I intended to prepare some handouts for this lecture
c. I did not prepare any handouts for this lecture
d. I prepared some handouts for this lecture

If a system did *not* say that someone who said (1a) would also be expected to believe (2a), (2b) and (2c), and that someone who said (1a) would also be expected to believe (2a), (2b) and (2d) then you would hardly want to say that it had understood what it was told. There may, in some ineffable way, be more to understanding than this, but it certainly provides a baseline. If a system does not agree with a native speaker about what follows from a given utterance then it cannot be said to understand natural language. If it did agree with a typical native speaker about what does and does not follow from a wide collection of utterances then it would be very hard to tell the difference between the system and the person–it would have passed a version of the Turing test. As noted, there may be something about human understanding of natural language which is not covered by this test, but in that case it is very hard to see what other test would capture the difference. If a system knew what follows from an arbitrary utterance, it would be hard to point to something that was lacking in its understanding.

This test does lead to blurring of the boundaries between language and background knowledge. Consider (3):

(3) I played tennis on Saturday.

Which of the sentences in (4) would a system have to agree with before it counted as 'understanding' (3)?

(4) a. I played a sport on Saturday.
b. I played a game where you use a racket to hit a ball on Saturday.
c. I played a game whose scoring system was devised in medieval France on Saturday.

Clearly, a system that did not know that (4a) followed from (4) would not be very impressive; and equally clearly, knowing that (4c) follows from (4) is specialist knowledge rather than general understanding. The boundary between language and general knowledge is evidently hard to draw–which side of this boundary does (4b), for instance, lie? Nonetheless, the idea that a system's level of understanding is displayed by its ability to recognise what does and does not follow from an utterance provides a good measure of understanding.

## 2 Entailment, logic and inference

But if understanding is characterised by an ability to determine whether one sentence follows from another then any system that is to understand natural language must include some treatment of entailment. In other words, it must exploit a 'logic'–a framework for determining the relationships between propositions.

This is a very weak claim: there are a wide variety of such frameworks, in which you can express a wide variety of propositions and relationships between propositions. However, there is one widely accepted desideratum. If we write $A \vdash B$ to mean that $B$ follows from $A$, then it should not be possible for $A$ to be true and $B$ to be false. If the notion of entailment is to be of any use at all, then it should not be possible to obtain a false conclusion from true premises. A logic that provides a set of inference rules that never produce a false conclusion from true premises is said to be 'sound'. Logics which are not sound are of very little practical use, since they will lead you to construct inaccurate pictures of the world and thence to make poor decisions. They are of even less theoretical use, since the relationships between different notions then become extremely unclear.

That is not to say that human beings infallibly perform correct inferences. There is a difference between saying that the meanings of natural languages are underpinned by a coherent notion of consequence and claiming that people reliably perform sound chains of inference. There is a strong analogy here with the distinction between competence/performance in syntax. A fluent speaker of a language has access to a system of rules and constraints about what constitutes a legitimate sentence of that language. Most people will make occasional errors when they are producing utterances, and they will be able to compensate for errors that other people make. They will also often have different sets of rules and constraints which they apply in different situations, to reflect the variety of dialects and registers that a typical speaker can switch between. Nonetheless, they will know whether a given sentence is legitimate according to the dialect and register they are currently using: any fluent English speaker can judge that *'I don't know much about art but I know what I like'* sounds better than *'I don't know much about art, but I know I like what'*.

It is the shared appreciation of the relevant rules and constraints that makes communication possible, even if individual utterances can break these rules (either accidentally, from production errors, or deliberately, for rhetorical effect). Similarly, language would not work as a vehicle for conveying ideas if it were not underpinned by a notion of consequence. If I could not assume that when you said *'I played tennis on Saturday'* you were also committed to *'I played tennis'* and *'I played something on Saturday'* then I would be unable to act on what you said. You might say something without realising all its consequences, or you might just say things that you don't believe. Nonetheless, once I point out the consequences of what you have said, you have to either accept them or withdraw your original statement. To use language is to make statements about the world knowing that those statements have consequences.

Once we accept that entailment is a critical aspect of natural language understanding, we have to decide what kind of logic we require. There is a clear trade-off between the expressive power of a logic and the difficulty of carrying out inference within that logic. Fig. 1 shows the complexity of inference for a number of well-known families of logics.

| | |
|---|---|
| Attribute:value pairs<br>Database languages<br>Sort logics | linear |
| Propositional logic<br>Description logic | NP-complete |
| Predicate logic<br>Modal logic<br>temporal logics, epistemic logics | Recursively enumerable (RE) |
| Default logic | Co-RE |
| Property theory<br>Set theory<br>Intensional logic | Incomplete |

**Fig. 1.** Expressive power vs. complexity

Clearly, the nearer the top of this diagram your logic is, the better your program will perform. Nonetheless, natural language supports a range of distinctions that force us to move quite a long way down. (5) provides a set of examples that show the difficulty of the problem.

(5) a. Every man is mortal, Socrates is a man ⊢ Socrates is mortal. [**Quantification**]
b. i. I ate a peach ⊢ there was none left
ii. I was eating a peach ⊢ I was part way through eating it. [**Time**](Moëns and Steedman, 1988; Reichenbach, 1958)
c. Is there any milk in the fridge? ⊢ speaker does not know whether there is any milk [**Knowledge and belief**](Austin, 1962; Searle, 1969; Allen and Perrault, 1980; Cohen and Perrault, 1979)
d. Most swans are white, Bruce is a swan, *I don't know that he's not white* ⊢ Bruce is white [**Default logic**](Reiter, 1980; McCarthy, 1980)
e. I wish I had a decent top-spin backhand: *'wish'* expresses a relationship between me and an 'event type' or a 'parameterised state-of-affairs' [**Intensionality**](Barwise and Perry, 1983)

There is nothing strange or esoteric about any of these examples. They are just examples of everyday constructions. They also support consequences which any native English speaker would accept: if I said *'I saw a swan in the park this morning'*, for instance, you would naturally picture a large white bird; but if I said *'I saw a black swan in the park this morning'* your picture would be quite different. You could hardly claim to understand English if you were not able to draw these consequences. How people do inference is still very unclear, and it seems unlikely that any implemented theorem provers match the human reasoning process in detail. Nonetheless, the ability to draw quite complex inferences, about time, about other people's knowledge and beliefs, about hypothetical states of affairs, and so on is a key part of language understanding. If you do not realise that *'I wish I had a decent top-spin backhand'* ⊢ *'I do not have a decent top-spin background'* then you do not understand English.

## 3 Construction of logical forms

If the ability to understand natural language is intimately bound up with the ability to carry out inference, and computational approaches to inference are very largely descended from attempts automate the rules of formal logic, then we need a connection between language and logic.

The work of Richard Montague (Montague, 1974) is crucial here. Montague's argument, which is effectively summarised by Dowty et al. (1981), is that you can obtain an expression in a formal logic by associating terms from that logic with individual words and grammatical constructions in natural language, and then combining these terms in a simple and systematic way. The aim is to construct a sentence in the logic which would be true under exactly the same circumstances as the natural language sentence. Montague's aim was to get a precise way of talking about the meanings of natural languages: because the semantics of a formal language is precisely defined, if we can show that an expression of the formal language is true in the same circumstances as a sentence of our natural language, we can get a handle on the semantics of the natural language.

The key to this is the 'Principle of Compositionality': the meaning of the whole is a function of the meanings of the parts and their mode of composition. In other words, if you know what each word means and you know the synactic relationships between them, you can work out the meaning of the whole sentence.

In some sense this is obvious: when you want to work out what a sentence means, all you have to go on is the words and the way they are arranged. So if the meaning is *not* encoded by the meanings of the parts and their mode of combination, then it is hard to know where you could look for it. The interesting thing about Montague's work was that he provided a very concrete way of thinking about this, namely by using the $\lambda$-calculus (Church, 1936).

There are many ways of thinking about the $\lambda$-calculus. For our purposes, it is convenient to think of it as a way of specifying sets by describing the entities that belong to them. Thus if $love(x, Mary)$ is a formula of some language, then $\lambda x(love(x, Mary))$ describes the set of all things that love Mary. We can then write $\lambda x(love(x, Mary)).John$ to mean that John is a member of this set (we will also read this as saying that Joh satisfies the property of being someone who loves Mary, or that this property applies to John).

Clearly, if John is a member of the set of things that love Mary then the claim *'John loves Mary'* is true, and *vice versa.* This equivalence is captured by the rule of $\beta$-reduction:

$$\texttt{lambda(x, A):t} \equiv \texttt{A}_{\texttt{t/x}} \qquad (\beta\text{-reduction})^{1}$$

Montague's insight was that if you assigned appropriate $\lambda$-terms as the meanings of words, then you could obtain meanings of sentences just by applying the meaning of one item to the meaning of another, as directed by the syntactic relationships between them. To take as simple example as possible, let the meanings of *'a'*, *'man'* and *'sleeps'* be as in Fig. 2.

[1] We write `lambda(X, A)` rather than the more usual $\lambda X(A)$ because many of the examples below are program input or output, and this format is easier for a program to manipulate.

```
a = lambda(P, lambda(Q, exists(X, (P:X & Q:X))))
man = lambda(B, B:lambda(A, man(A))
sleeps = lambda(W, W:lambda(V,exists(U,sleep(U)&agent(U, V)))
```

**Fig. 2.** Meanings of *'a'*, *'man'*, *'sleeps'*

Suppose we had a dependency tree for *'A man sleeps'* which said that *'man'* is a daughter of *'sleeps'* and *'a'* is a daughter of *'man'*, and that we decided the way to construct the interpretation of a phrase was by applying the meaning of its head to the meanings of its daughters.

Then the meaning of *'A man sleeps'* would be

```
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))
           :(lambda(B,B:lambda(A,man(A)):lambda(P,lambda(Q,exists(X, P:X & Q:X)))))
```

This looks completely horrible. But applying a series of $\beta$-reductions to it rapidly leads to something more manageable (the underlined items in this derivation show the variable that is to be bound, where it occurs in the term, and the item that it is to be substituted for it).

```
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))
           :(lambda(B,B:lambda(A,man(A)):lambda(P,lambda(Q,exists(X, P:X & Q:X)))))
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V))))
           :(lambda(P,lambda(Q,exists(X, P:X & Q:X))):lambda(A,man(A))))
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))):
           (lambda(Q,exists(X, lambda(A,man(A)).X & Q:X))))
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))):lambda(Q,exists(X, man(X) & Q:X))
(lambda(Q,exists(X, man(X)) & Q:X))
           :lambda(V,exists(U,sleep(U)&agent(U, V)))
exists(X, man(X) & lambda(V,exists(U,sleep(U) &agent(U, V))).X)
exists(X, man(X) & exists(U, sleep(U) & agent(U, X)))
```

The process of $\beta$-reduction looks complex when you first see it, but it can be carried out entirely mechanically, and has indeed been used as the basis for a number of programming languages (e.g. LISP, ML, Scheme, Python). Devising appropriate expressions to stand as the meanings of individual words is, indeed, rather tricky, and requires a certain amount of both insight and imagination. The pay-off is that once you have come up with an interpretation of a word, you can use that interpretation anywhere that the word occurs. A given word makes the same contribution, no matter what context it appears in[2]. Thus so long as you can parse a sentence and you know what each word in it means, you can obtain the meaning of the whole thing just by carrying out $\beta$-reduction (which is essentially just string-substitution, and *is* trivial to implement and execute). So the meaning of a complex sentence can be computed straightforwardly from the meanings of the words that appear in it once you have determined its syntactic structure–see Fig. 3 for the meaning of (6).

[2] This claim is slightly undermined by the fact that a given surface form may correspond to a number of different underlying words, and choosing between these senses may not be entirely trivial.

(6) I know you think the woman who I met in the pub yesterday is a fool.

```
utt(claim,
    exists(A,
           event(A, know)
           & theta(A,
                 event,
                 exists(B,
                        event(B, think)
                        & theta(B,
                              event,
                              exists(C,
                                     at(C,
                                        exists(D :: {fool(D) & NEW(D)},
                                               exists(E :: {woman(E)
                                                             & exists(F :: {past(now, F)},
                                                                    exists(G,
                                                                           exists(H :: {pub(H) & KNOWN(H)},
                                                                                  event(G,meet)
                                                                                  & theta(G, object, E)
                                                                                  & theta(G,
                                                                                        agent,
                                                                                        ref(lambda(I, speaker(I))))
                                                                                  & loc(in, G, H)
                                                                                  & yesterday(G))
                                                                           & aspect(F,simplePast,G)))
                                                             & KNOWN(E)},
                                                      E=D)))
                                     & aspect(now, simple, C)))
                        & theta(B, agent, ref(lambda(J, hearer(J))))
                        & aspect(now, simple, B)))
           & theta(A, agent, ref(lambda(K, speaker(K))))
           & aspect(now, simple, A)))
```

**Fig. 3.** Logical form of (6)

This looks extraordinarily complicated. Constructing it, however, is completely straightforward once you have the parse tree and the meanings of the indvidual words. Just apply the meaning of the head to the meaning of the daughters. There is therefore no need to be scared of building logical forms. So long as you can parse the input text and you have appropriate meanings for words then the logical form will just emerge.

This does, of course, beg two questions. What do you do if you cannot parse the input text, and where do you get appropriate meanings for words from?

There are three possible reasons why it may not be possible to parse a piece of text:

- Your description of the rules governing the relations between words and phrases may be inadequate.
- The text may just be so long and complex that your parser is swamped, and either takes an unacceptable amount of time or produces large numbers of analyses, with the intended one buried too deep to find.
- The text may not be well-formed.

These are indeed serious problems for anyone trying to produce logical forms on the basis of the syntactic relations between items. There are tricks and techniques for recovering from all of them (the first and third are very similar in practice). These problems, however, all concern the task of detecting the syntactic relationships between items, and as such they do not undermine the general principles discussed above, though they may make implementation difficult when trying to handle complex texts.

The task of trying to develop appropriate representations of lexical items is undoubtedly challenging. Roughly speaking, you have to start by deciding what you would like the meaning of a typical sentence to look like. If you can then attribute different aspects of this to individual words, you can usually divide the meaning of the whole into parts which can be glued together to make the whole (see (van Genabith and Crouch, 1997; Dalrymple et al., 1996) for an elaboration of the idea of 'glueing' fragments of a meaning representation together). Consider again *'A man sleeps'*. Suppose you had decided that

```
exists(X, man(X) & exists(U, sleep(U) & agent(U, X)))
```

was what you wanted as the meaning of *'A man sleeps.'*. It is fairly clear which elements of this are contributed by *'man'* and *'sleeps'*, namely the fact that there's a man and the fact that there's a sleeping event whose agent is the man.

If we take these prts out of the logical form, we are left with a skeleton like

```
exists(X, ...(X) & ...(X))
```

In other words, what *'a'* contributes to the meaning of *'A man sleeps'* is that there is something, and that we're going to supply two pieces of information about that thing, namely what kind of thing it is (which will come from the noun) and what it did (which will come from the verb).

The notation of the $\lambda$-calculus lets us specify the order in which these two pieces of information will be supplied. *'a'* will combine with the noun first and then with the verb. So the actual logical form is

```
lambda(P, lambda(Q, exists(X, P:X & Q:X)))
```

exactly as in Fig. 2.

Where did the idea that `exists(X, man(X) & exists(U, sleep(U) & agent(U, X)))` come from in the first place? That has to come from reflection on what kinds of information need to be encoded, and what kinds of inference need to be carried out. But once you have done that, determining what information is contributed by each word is fairly routine.

It is worth noting here that this process is easily transferrable between languages. If two languages share the same sets of lexical classes, then they will give rise to very similar dependency trees. If that is so, exactly the same mechanisms can be used for building logical forms. It is just as easy to build logical forms for Arabic, for instance, as it is for English[3]. Thus we can construct the logical form in Fig. 4 for the Arabic sentence (7) just by parsing it and reading the interpretation off the parse tree, in exactly the same way as we did for English examples.

(7) اعتقد الولد ن البنت التي درست في المدرست كتبت الدرس. (*āʿtqd ālwld n ālbnt ālty drst fy ālmdrst ktbt āldrs.*)

It is possible to obtain a labelled dependency tree from the output of any grammar that lets you identify the head of a structure (e.g. HPSG, LFG, GPSG, categorial grammar) so these techniques can be employed for any sentence for which you have an analysis within such a framework.

[3] With one minor caveat: many languages allow NPs without determiners, but the interpretation of this construction varies: English NPs with no determiner have a generic reading, Persian NPs with no determiners have definite readings, Arabic NPs with no determiners have indefinite readings. You therefore have to supply a language-specific interpretation for such cases.

```
utt(claim,
    exists(A :: {(w?l?d(A) & KNOWN(A))},
           exists(B :: {(d?r?s(B) & KNOWN(B))},
                  exists(C :: {(b?n?t(C)
                                & (exists(D :: {past(now, D)},
                                        exists(E,
                                               (event(E, d?r?s)
                                               & (theta(E, agent, C)
                                               & (fy(E, B) & aspect(D, simple, E))))))
                                & KNOWN(C)))},
                         exists(F :: {(d?r?s(F) & KNOWN(F))},
                                exists(G :: {past(now, G)},
                                       exists(H,
                                              (event(H, E?q?d)
                                              & (theta(H, agent, A)
                                              & (theta(H,
                                                     event,
                                                     exists(I :: {past(now, I)},
                                                            exists(J,
                                                                   (event(J, k?t?b)
                                                                   & (theta(J, agent, C)
                                                                   & (theta(J, object, F)
                                                                   & aspect(I, simple, J)))))))
                                              & aspect(G, simple, H))))))))))
```

**Fig. 4.** Logical form for (7)

## 4 Using logical forms

If you can parse a piece of text, then, you can construct a logical form for it. What can you do with it once you have done so?

Consider (8):

(8) I am allergic to eggs. Should I avoid eating pancakes?

This is a reasonably straightforward question. What resources would you need in order to answer it?

Firstly, you would need a great deal of basic information: pancakes contain eggs, eating things that contain foodstuffs that you are allergic to will make you ill, doing things that will make you ill is bad for you, if something is bad for you then you avoid doing it, . . .

Some of this information is reasonably easy to represent. The fact that pancaks contain eggs, for instance, can be captured by the formula in Fig. 5.

```
forall(X :: {sort(pancake, X)},
       exists(C :: {aspect(now, simple, C)},
              event(C, contain) & exists(Y, sort(egg, Y) & theta(C, object, Y)) & theta(C, agent, X)))
```

**Fig. 5.** If X is a pancake then it contains an egg

Other elements are extremely difficult to capture. Part of the problem here is that, as noted above, much of this information is intensional, but we have already seen using the $\lambda$-calculus allows use to state relationships between intensional objects. The logical form for *'Should I avoid eating pancakes?'*, for instance, is as in Fig. 6.

```
utt(query,
    exists(A,
           (should(A,
                   lambda(B,
                           (event(B, avoid)
                           & (theta(B,
                                  object,
                                  lambda(C,
                                         exists(D,
                                                (event(D, eat)
                                                & (exists(E :: {sort(pancake, E, F, G)},
                                                        theta(D, object, E))
                                                & theta(D, agent, C))))))
                           & theta(B, agent, ref(lambda(H, speaker(H)))!1)))))
           & aspect(now, simple, A))))
```

**Fig. 6.** Logical form for *'Should I eat pancakes?'*

The next problem is that writing rules in logic is very awkward. The rule we will need for expresing the fact that if something is bad for you then you should avoid doing it is given in Fig. 7.

```
forall(Y,
       forall(X,
              (exists(A,
                      (at(A, sort(Y, ~good))
                      & (for(A, X) & aspect(now, simple, A))))
              => exists(B,
                        (should(B,
                                lambda(C,
                                       (event(C, avoid)
                                       & (theta(C, object, Y)
                                       & theta(C, agent, X)))))
                        & aspect(now, simple, B))))))
```

**Fig. 7.** If something is bad for you then you should avoid it

Most people (even logicians) find it hard to write complex rules in logic, and the situation is made worse by the fact that the rules you have to write have to use the same terms and structures as the terms and structures used in logical forms. Because, in the examples above, we are using expressions like `event(X, ???)` to represent the fact that something is an eating event, we have to make sure that when we mention an event in a hand-coded rule we use the same kind of expression. The difficulty of bearing all the relevant conventions in mind when writing rules just makes this task harder.

The worst problem of all, however, is working out just what the relationships between the concepts denoted by various constructions in natural language are. Consider the word *'good'*: if you look this word up in a dictionary, you generally find some rather unhelpful circular definition (e.g. 'having the right qualities':

Pocket Oxford Dictionary, ‘of a favorable character or tendency’: Merriam Webster on-line dictionary) followed by a large collection of examples. Dictionary entries of this kind provide very little information about what follows from saying that something is good, or about the circumstances under which you would say that something is good.

One way to deal with this problem is to tackle it head on. What we want to know is when you should say that something is good or bad, and what else we would be inclined to accept once we know this. So why not write down, in natural language, the kinds of things that we want to be able to reason with?

(9) a. eating P will make X ill if X is allergic to P
b. X is dangerous for Y if X will make Y ill
c. X is bad for Y if X is dangerous for Y
d. X should avoid Y if Y is bad for X
...

The rules in (9) are basic commonsense knowledge about very common words. By and large, conclusions that someone arrived by using these rules would be sensible, and anyone who did not have access to this kind of information could hardly be said to understand English.

Using natural language to write down the necessary knowledge about how words and concepts are related helps overcome the problems listed above:

- They are fairly easy to write. You do not have to become fluent in some complex formal language, and you do not have to familiarise yourself with the way that that language is being used, in order to add new rules. Just write them down in natural language.
- There is no risk that the terminology used for the logical forms derived for rules will differ from the terminology used for logical forms derived from general discourse. The same mechanisms will be used for both, so there is no possibility of inconsistency in the use of terminology.
- The rules say what is needed. In most cases, these rules will not constitute a ‘definition’, in the sense of a set of necessary and sufficient conditions for the use of a word. Instead they license specific inferences. It may, in fact, be the case that there is nothing more to be said about what a word means: that the meaning of a word is determined by its relations with other words. These relations may or may not provide a complete and precise set of constraints. They may, rather, constitute a set of ‘semantic traits’ (Cruse, 1986). Whether or not there is, indeed, more to the meaning of a word than that, certainly providing rules like these will support the derivation of a range of appropriate inferences.

## 5 Conclusions

Is there any alternative? It is clear that the ability to respond to (8) by saying *‘Yes’* (or, better, by saying *‘Yes, because pancakes contain eggs, and if you are allergic to eggs then eating things which contain eggs will make you ill’*) is highly desirable. A system that cannot do this cannot be said to understand the sentences that make up (8). It is equally clear that in order to do this, you have to be able to extract the relevant elements of your knowledge and construct a chain of inference from them. So if there is to be an alternative to the approach outlined here, it has to involve an alternative notion of inference.

It may well be, of course, that the specific logical forms outlined above are inadequate to the task. Indeed, the forms given here are inadequate: they were obtained by a system that follows Konrad et al. (1996) in allowing a variety of different forms to be obtained from a given tree by attaching ‘codes’ to words and modes of combination, and then using different codebooks to obtain different interpretations. This approach lets you construct logical forms which contain different amounts of detail, depending on the task you want to perform. Simple, coarse-grained interpretations are often useful for exposition (as here). More importantly, it is easier to reason about simple interpretations than about complex ones. It might, for instance, be convenient to use a simple representation for disambiguating a sentence, and then build a more detailed one for using the chosen reading for some more complex task. The suggestion that the logical forms given above are not complex or detailed enough, however, does not undermine the thesis that building logical forms and reasoning about them is the right approach to semantics. It just indicates that we need to develop more detailed logical forms, probably alongside improvements in our inference engines to cope with the extra complexity.

The only alternative is to define rules of inference that operate directly on texts. The textual entailment program (Dagan et al., 2005) is an attempt to follow this path. Any such attempt will, inevitably, have to cope with quantification. If you cannot infer *'Socrates is mortal'* from *'All men are mortal'* and *'Socrates is a man'* then you are not going to be able to handle more interesting discourses such as (8). There will, inevitably, be a temptation to start introducing rules like Fig. 8.

$$\begin{array}{c}\text{All Ps are Qs}\\ \underline{\text{X is a P}}\\ \text{X is a Q}\end{array}$$

**Fig. 8.** Basic syllogism

Once you start introducing rules like this, the next obvious step is to normalise away various aspects of the surface form of the text. It would be extremely irritating, for instance, to have the rules in Fig. 9 as well.

$$\begin{array}{c}\text{All Ps are Qs}\\ \underline{\text{I am a P}}\\ \text{I am a Q}\end{array} \quad \begin{array}{c}\text{All Ps are Qs}\\ \underline{\text{You are a P}}\\ \text{You are a Q}\end{array} \quad \begin{array}{c}\text{All Ps are Qs}\\ \underline{\text{X was a P}}\\ \text{X was a Q}\end{array} \quad \dots$$

**Fig. 9.** Overspecific inference rules

The obvious way round this is to eliminate irrelevant aspects of the surface form, so that Fig. 9 would become Fig. 10.

$$\begin{array}{c}\text{All P be Q}\\ \underline{\text{X be P}}\\ \text{X be Q}\end{array}$$

**Fig. 10.** Generalised syllogism

But this is exactly the route that led from Aristotle to the development of modern formal logic. Constructing logical forms is, exactly, the process of abstracting away from elements of the surface form that do not help you to decide which patterns are applicable. It thus seems very likely that attempts to make textual entailment more general and robust will lead to exactly the same end result. The key to constructing appropriate (linguistic or extra-linguistic) responses is inference: the only real question is not whether you make meaning representations that allow you to perform inference, but whether you do it well or badly.

# Bibliography

Allen, J. F., Perrault, C. R., 1980. Analysing intention in utterances. Artificial Intelligence 15, 148–178.

Austin, J., 1962. How to Do Things with Words. Oxford University Press, Oxford.

Barwise, J., Perry, J., 1983. Situations and Attitudes. Bradford Books, Cambridge, MA.

Church, A., 1936. An unsolvable problem of elementary number theory. American Journal of Mathematics 58(2), 345–363.

Cohen, P. R., Perrault, C. R., 1979. Elements of a plan-based theory of speech acts. Cognitive Science 7(2), 171–190.

Cruse, D. A., 1986. Lexical Semantics. Cambridge University Press, Cambridge.

Dagan, I., Magnini, B., Glickman, O., 2005. The PASCAL recognising textual entailment challenge. In: Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment.

Dalrymple, M., Lamping, J., Pereira, F. C. N., Saraswat, V., 1996. A deductive account of quantification in LFG. In: Kanazawa, M., Piñón, C., de Swart, H. (Eds.), Quantifiers, deduction and context. pp. 33–58.

Dowty, D. R., Wall, R. E., Peters, S., 1981. Introduction to Montague Semantics. D. Reidel, Dordrecht.

Konrad, K., Maier, H., Milward, D., Pinkal, M., 1996. An education and research tool for computational semantics. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96). Copenhagen, pp. 1098–1102.

McCarthy, J., 1980. Circumscription: a form of non-monotonic reasoning. Artificial Intelligence 13, 27–39.

Moëns, M., Steedman, M., 1988. Temporal ontology and temporal reference. Computational Linguistics 14(2), 15–28.

Montague, R., 1974. The proper treatment of quantification in ordinary English. In: Thomason, R. (Ed.), Formal Philosophy: Selected Papers of Richard Montague. Yale University Press, New Haven.

Reichenbach, H., 1958. The Philosophy of Space and Time. Dover Books, New York.

Reiter, R., 1980. A logic for default reasoning. Artificial Intelligence 13(1), 81–132.

Searle, J. R., 1969. Speech Acts: an Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

van Genabith, J., Crouch, R., 1997. How to glue a donkey to an f-structure. In: Bunt, H. C., Kievit, L., Muskens, R., Verlinden, M. (Eds.), 2nd International Workshop on Computational Semantics. University of Tilburg, pp. 52–65.

# Formal semantics: luxury or necessity?

Allan Ramsay
School of Computer Science,University of Manchester
Manchester M13 9PL, UK

**Abstract.** We will revisit the goals of natural language processing, and argue that thse goals cannot be achieved unless you construct fine-grained formal paraphrases (also known as 'logical forms'). We will then consider the problems involved in building such paraphrases, and review the connections between semantics and inference. We will close by seeing how natural language can be used for expressing detailed domain knowledge, thus avoiding the need for complex translations between formalisms.

## 1 What should a natural language understanding system do?

What, ideally, should a natural language understanding system be able to do? What tests would a system have to pass before you were prepared to accept that it 'understood' what you meant when you used natural language? What would it take for a system to pass the Turing test?

Roughly speaking, a reasonable test would be to see whether its view of what someone who produced a sentence in natural language could reasonably be expected to believe was the same as that of a native speaker. Suppose I said one of (1a) and (1b):

(1) a. I forgot to prepare any handouts for my lecture on Wednesday.
b. I forgot that I had prepared some handouts for my lecture on Wednesday.

Consider the sentences in (2):

(2) a. I had a lecture on Wednesday
b. I intended to prepare some handouts for this lecture
c. I did not prepare any handouts for this lecture
d. I prepared some handouts for this lecture

If a system did *not* say that someone who said (1a) would also be expected to believe (2a), (2b) and (2c), and that someone who said (1a) would also be expected to believe (2a), (2b) and (2d) then you would hardly want to say that it had understood what it was told. There may, in some ineffable way, be more to understanding than this, but it certainly provides a baseline. If a system does not agree with a native speaker about what follows from a given utterance then it cannot be said to understand natural language. If it did agree with a typical native speaker about what does and does not follow from a wide collection of utterances then it would be very hard to tell the difference between the system and the person–it would have passed a version of the Turing test. As noted, there may be something about human understanding of natural language which is not covered by this test, but in that case it is very hard to see what other test would capture the difference. If a system knew what follows from an arbitrary utterance, it would be hard to point to something that was lacking in its understanding.

This test does lead to blurring of the boundaries between language and background knowledge. Consider (3):

(3) I played tennis on Saturday.

Which of the sentences in (4) would a system have to agree with before it counted as 'understanding' (3)?

(4) a. I played a sport on Saturday.
b. I played a game where you use a racket to hit a ball on Saturday.
c. I played a game whose scoring system was devised in medieval France on Saturday.

Clearly, a system that did not know that (4a) followed from (4) would not be very impressive; and equally clearly, knowing that (4c) follows from (4) is specialist knowledge rather than general understanding. The boundary between language and general knowledge is evidently hard to draw–which side of this boundary does (4b), for instance, lie? Nonetheless, the idea that a system's level of understanding is displayed by its ability to recognise what does and does not follow from an utterance provides a good measure of understanding.

## 2 Entailment, logic and inference

But if understanding is characterised by an ability to determine whether one sentence follows from another then any system that is to understand natural language must include some treatment of entailment. In other words, it must exploit a 'logic'–a framework for determining the relationships between propositions.

This is a very weak claim: there are a wide variety of such frameworks, in which you can express a wide variety of propositions and relationships between propositions. However, there is one widely accepted desideratum. If we write $A \vdash B$ to mean that $B$ follows from $A$, then it should not be possible for $A$ to be true and $B$ to be false. If the notion of entailment is to be of any use at all, then it should not be possible to obtain a false conclusion from true premises. A logic that provides a set of inference rules that never produce a false conclusion from true premises is said to be 'sound'. Logics which are not sound are of very little practical use, since they will lead you to construct inaccurate pictures of the world and thence to make poor decisions. They are of even less theoretical use, since the relationships between different notions then become extremely unclear.

That is not to say that human beings infallibly perform correct inferences. There is a difference between saying that the meanings of natural languages are underpinned by a coherent notion of consequence and claiming that people reliably perform sound chains of inference. There is a strong analogy here with the distinction between competence/performance in syntax. A fluent speaker of a language has access to a system of rules and constraints about what constitutes a legitimate sentence of that language. Most people will make occasional errors when they are producing utterances, and they will be able to compensate for errors that other people make. They will also often have different sets of rules and constraints which they apply in different situations, to reflect the variety of dialects and registers that a typical speaker can switch between. Nonetheless, they will know whether a given sentence is legitimate according to the dialect and register they are currently using: any fluent English speaker can judge that *'I don't know much about art but I know what I like'* sounds better than *'I don't know much about art, but I know I like what'*.

It is the shared appreciation of the relevant rules and constraints that makes communication possible, even if individual utterances can break these rules (either accidentally, from production errors, or deliberately, for rhetorical effect). Similarly, language would not work as a vehicle for conveying ideas if it were not underpinned by a notion of consequence. If I could not assume that when you said *'I played tennis on Saturday'* you were also committed to *'I played tennis'* and *'I played something on Saturday'* then I would be unable to act on what you said. You might say something without realising all its consequences, or you might just say things that you don't believe. Nonetheless, once I point out the consequences of what you have said, you have to either accept them or withdraw your original statement. To use language is to make statements about the world knowing that those statements have consequences.

Once we accept that entailment is a critical aspect of natural language understanding, we have to decide what kind of logic we require. There is a clear trade-off between the expressive power of a logic and the difficulty of carrying out inference within that logic. Fig. 1 shows the complexity of inference for a number of well-known families of logics.

| | |
|---|---|
| Attribute:value pairs<br>Database languages<br>Sort logics | linear |
| Propositional logic<br>Description logic | NP-complete |
| Predicate logic<br>Modal logic<br>temporal logics, epistemic logics | Recursively enumerable (RE) |
| Default logic | Co-RE |
| Property theory<br>Set theory<br>Intensional logic | Incomplete |

**Fig. 1.** Expressive power vs. complexity

Clearly, the nearer the top of this diagram your logic is, the better your program will perform. Nonetheless, natural language supports a range of distinctions that force us to move quite a long way down. (5) provides a set of examples that show the difficulty of the problem.

(5) a. Every man is mortal, Socrates is a man $\vdash$ Socrates is mortal. [**Quantification**]
b. i. I ate a peach $\vdash$ there was none left
ii. I was eating a peach $\vdash$ I was part way through eating it. [**Time**](Moëns and Steedman, 1988; Reichenbach, 1958)
c. Is there any milk in the fridge? $\vdash$ speaker does not know whether there is any milk [**Knowledge and belief**](Austin, 1962; Searle, 1969; Allen and Perrault, 1980; Cohen and Perrault, 1979)
d. Most swans are white, Bruce is a swan, *I don't know that he's not white* $\vdash$ Bruce is white [**Default logic**](Reiter, 1980; McCarthy, 1980)
e. I wish I had a decent top-spin backhand: *'wish'* expresses a relationship between me and an 'event type' or a 'parameterised state-of-affairs' [**Intensionality**](Barwise and Perry, 1983)

There is nothing strange or esoteric about any of these examples. They are just examples of everyday constructions. They also support consequences which any native English speaker would accept: if I said *'I saw a swan in the park this morning'*, for instance, you would naturally picture a large white bird; but if I said *'I saw a black swan in the park this morning'* your picture would be quite different. You could hardly claim to understand English if you were not able to draw these consequences. How people do inference is still very unclear, and it seems unlikely that any implemented theorem provers match the human reasoning process in detail. Nonetheless, the ability to draw quite complex inferences, about time, about other people's knowledge and beliefs, about hypothetical states of affairs, and so on is a key part of language understanding. If you do not realise that *'I wish I had a decent top-spin backhand'* $\vdash$ *'I do not have a decent top-spin background'* then you do not understand English.

## 3 Construction of logical forms

If the ability to understand natural language is intimately bound up with the ability to carry out inference, and computational approaches to inference are very largely descended from attempts automate the rules of formal logic, then we need a connection between language and logic.

The work of Richard Montague (Montague, 1974) is crucial here. Montague's argument, which is effectively summarised by Dowty et al. (1981), is that you can obtain an expression in a formal logic by associating terms from that logic with individual words and grammatical constructions in natural language, and then combining these terms in a simple and systematic way. The aim is to construct a sentence in the logic which would be true under exactly the same circumstances as the natural language sentence. Montague's aim was to get a precise way of talking about the meanings of natural languages: because the semantics of a formal language is precisely defined, if we can show that an expression of the formal language is true in the same circumstances as a sentence of our natural language, we can get a handle on the semantics of the natural language.

The key to this is the 'Principle of Compositionality': the meaning of the whole is a function of the meanings of the parts and their mode of composition. In other words, if you know what each word means and you know the synactic relationships between them, you can work out the meaning of the whole sentence.

In some sense this is obvious: when you want to work out what a sentence means, all you have to go on is the words and the way they are arranged. So if the meaning is *not* encoded by the meanings of the parts and their mode of combination, then it is hard to know where you could look for it. The interesting thing about Montague's work was that he provided a very concrete way of thinking about this, namely by using the $\lambda$-calculus (Church, 1936).

There are many ways of thinking about the $\lambda$-calculus. For our purposes, it is convenient to think of it as a way of specifying sets by describing the entities that belong to them. Thus if $love(x, Mary)$ is a formula of some language, then $\lambda x(love(x, Mary))$ describes the set of all things that love Mary. We can then write $\lambda x(love(x, Mary)).John$ to mean that John is a member of this set (we will also read this as saying that Joh satisfies the property of being someone who loves Mary, or that this property applies to John).

Clearly, if John is a member of the set of things that love Mary then the claim *'John loves Mary'* is true, and *vice versa*. This equivalence is captured by the rule of $\beta$-reduction:

$$\texttt{lambda(x, A):t} \equiv \texttt{A}_{\texttt{t/x}} \qquad (\beta\text{-reduction})^1$$

Montague's insight was that if you assigned appropriate $\lambda$-terms as the meanings of words, then you could obtain meanings of sentences just by applying the meaning of one item to the meaning of another, as directed by the syntactic relationships between them. To take as simple example as possible, let the meanings of *'a'*, *'man'* and *'sleeps'* be as in Fig. 2.

[1] We write `lambda(X, A)` rather than the more usual $\lambda X(A)$ because many of the examples below are program input or output, and this format is easier for a program to manipulate.

```
a = lambda(P, lambda(Q, exists(X, (P:X & Q:X))))
man = lambda(B, B:lambda(A, man(A))
sleeps = lambda(W, W:lambda(V,exists(U,sleep(U)&agent(U, V)))
```

**Fig. 2.** Meanings of *'a'*, *'man'*, *'sleeps'*

Suppose we had a dependency tree for *'A man sleeps'* which said that *'man'* is a daughter of *'sleeps'* and *'a'* is a daughter of *'man'*, and that we decided the way to construct the interpretation of a phrase was by applying the meaning of its head to the meanings of its daughters.

Then the meaning of *'A man sleeps'* would be

```
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))
           :(lambda(B,B:lambda(A,man(A)):lambda(P,lambda(Q,exists(X, P:X & Q:X)))))
```

This looks completely horrible. But applying a series of $\beta$-reductions to it rapidly leads to something more manageable (the underlined items in this derivation show the variable that is to be bound, where it occurs in the term, and the item that it is to be substituted for it).

```
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))
           :(lambda(B,B:lambda(A,man(A)):lambda(P,lambda(Q,exists(X, P:X & Q:X)))))
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V))))
           :(lambda(P,lambda(Q,exists(X, P:X & Q:X))):lambda(A,man(A))))
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))):
           (lambda(Q,exists(X, lambda(A,man(A)).X & Q:X))))
lambda(W,W:lambda(V,exists(U,sleep(U)&agent(U, V)))):lambda(Q,exists(X, man(X) & Q:X))
(lambda(Q,exists(X, man(X)) & Q:X))
           :lambda(V,exists(U,sleep(U)&agent(U, V)))
exists(X, man(X) & lambda(V,exists(U,sleep(U) &agent(U, V))).X)
exists(X, man(X) & exists(U, sleep(U) & agent(U, X)))
```

The process of $\beta$-reduction looks complex when you first see it, but it can be carried out entirely mechanically, and has indeed been used as the basis for a number of programming languages (e.g. LISP, ML, Scheme, Python). Devising appropriate expressions to stand as the meanings of individual words is, indeed, rather tricky, and requires a certain amount of both insight and imagination. The pay-off is that once you have come up with an interpretation of a word, you can use that interpretation anywhere that the word occurs. A given word makes the same contribution, no matter what context it appears in[2]. Thus so long as you can parse a sentence and you know what each word in it means, you can obtain the meaning of the whole thing just by carrying out $\beta$-reduction (which is essentially just string-substitution, and *is* trivial to implement and execute). So the meaning of a complex sentence can be computed straightforwardly from the meanings of the words that appear in it once you have determined its syntactic structure–see Fig. 3 for the meaning of (6).

[2] This claim is slightly undermined by the fact that a given surface form may correspond to a number of different underlying words, and choosing between these senses may not be entirely trivial.

(6) I know you think the woman who I met in the pub yesterday is a fool.

```
utt(claim,
    exists(A,
           event(A, know)
           & theta(A,
                 event,
                 exists(B,
                        event(B, think)
                        & theta(B,
                              event,
                              exists(C,
                                     at(C,
                                        exists(D :: {fool(D) & NEW(D)},
                                               exists(E :: {woman(E)
                                                             & exists(F :: {past(now, F)},
                                                                    exists(G,
                                                                           exists(H :: {pub(H) & KNOWN(H)},
                                                                                  event(G,meet)
                                                                                  & theta(G, object, E)
                                                                                  & theta(G,
                                                                                        agent,
                                                                                        ref(lambda(I, speaker(I))))
                                                                                  & loc(in, G, H)
                                                                                  & yesterday(G))
                                                                           & aspect(F,simplePast,G)))
                                                             & KNOWN(E)},
                                                      E=D)))
                                     & aspect(now, simple, C)))
                        & theta(B, agent, ref(lambda(J, hearer(J))))
                        & aspect(now, simple, B)))
           & theta(A, agent, ref(lambda(K, speaker(K))))
           & aspect(now, simple, A)))
```

**Fig. 3.** Logical form of (6)

This looks extraordinarily complicated. Constructing it, however, is completely straightforward once you have the parse tree and the meanings of the indvidual words. Just apply the meaning of the head to the meaning of the daughters. There is therefore no need to be scared of building logical forms. So long as you can parse the input text and you have appropriate meanings for words then the logical form will just emerge.

This does, of course, beg two questions. What do you do if you cannot parse the input text, and where do you get appropriate meanings for words from?

There are three possible reasons why it may not be possible to parse a piece of text:

- Your description of the rules governing the relations between words and phrases may be inadequate.
- The text may just be so long and complex that your parser is swamped, and either takes an unacceptable amount of time or produces large numbers of analyses, with the intended one buried too deep to find.
- The text may not be well-formed.

These are indeed serious problems for anyone trying to produce logical forms on the basis of the syntactic relations between items. There are tricks and techniques for recovering from all of them (the first and third are very similar in practice). These problems, however, all concern the task of detecting the syntactic relationships between items, and as such they do not undermine the general principles discussed above, though they may make implementation difficult when trying to handle complex texts.

The task of trying to develop appropriate representations of lexical items is undoubtedly challenging. Roughly speaking, you have to start by deciding what you would like the meaning of a typical sentence to look like. If you can then attribute different aspects of this to individual words, you can usually divide the meaning of the whole into parts which can be glued together to make the whole (see (van Genabith and Crouch, 1997; Dalrymple et al., 1996) for an elaboration of the idea of 'glueing' fragments of a meaning representation together). Consider again *'A man sleeps'*. Suppose you had decided that

```
exists(X, man(X) & exists(U, sleep(U) & agent(U, X)))
```

was what you wanted as the meaning of *'A man sleeps.'*. It is fairly clear which elements of this are contributed by *'man'* and *'sleeps'*, namely the fact that there's a man and the fact that there's a sleeping event whose agent is the man.

If we take these prts out of the logical form, we are left with a skeleton like

```
exists(X, ...(X) & ...(X))
```

In other words, what *'a'* contributes to the meaning of *'A man sleeps'* is that there is something, and that we're going to supply two pieces of information about that thing, namely what kind of thing it is (which will come from the noun) and what it did (which will come from the verb).

The notation of the $\lambda$-calculus lets us specify the order in which these two pieces of information will be supplied. *'a'* will combine with the noun first and then with the verb. So the actual logical form is

```
lambda(P, lambda(Q, exists(X, P:X & Q:X)))
```

exactly as in Fig. 2.

Where did the idea that `exists(X, man(X) & exists(U, sleep(U) & agent(U, X)))` come from in the first place? That has to come from reflection on what kinds of information need to be encoded, and what kinds of inference need to be carried out. But once you have done that, determining what information is contributed by each word is fairly routine.

It is worth noting here that this process is easily transferrable between languages. If two languages share the same sets of lexical classes, then they will give rise to very similar dependency trees. If that is so, exactly the same mechanisms can be used for building logical forms. It is just as easy to build logical forms for Arabic, for instance, as it is for English[3]. Thus we can construct the logical form in Fig. 4 for the Arabic sentence (7) just by parsing it and reading the interpretation off the parse tree, in exactly the same way as we did for English examples.

(7) اعتقد الولد ن البنت التي درست في المدرست كتبت الدرس. (*āˁtqd ālwld n ālbnt ālty drst fy ālmdrst ktbt āldrs.*)

It is possible to obtain a labelled dependency tree from the output of any grammar that lets you identify the head of a structure (e.g. HPSG, LFG, GPSG, categorial grammar) so these techniques can be employed for any sentence for which you have an analysis within such a framework.

[3] With one minor caveat: many languages allow NPs without determiners, but the interpretation of this construction varies: English NPs with no determiner have a generic reading, Persian NPs with no determiners have definite readings, Arabic NPs with no determiners have indefinite readings. You therefore have to supply a language-specific interpretation for such cases.

```
utt(claim,
    exists(A :: {(w?l?d(A) & KNOWN(A))},
           exists(B :: {(d?r?s(B) & KNOWN(B))},
                  exists(C :: {(b?n?t(C)
                                & (exists(D :: {past(now, D)},
                                        exists(E,
                                               (event(E, d?r?s)
                                               & (theta(E, agent, C)
                                               & (fy(E, B) & aspect(D, simple, E))))))
                                & KNOWN(C)))},
                         exists(F :: {(d?r?s(F) & KNOWN(F))},
                                exists(G :: {past(now, G)},
                                       exists(H,
                                              (event(H, E?q?d)
                                              & (theta(H, agent, A)
                                              & (theta(H,
                                                     event,
                                                     exists(I :: {past(now, I)},
                                                            exists(J,
                                                                   (event(J, k?t?b)
                                                                   & (theta(J, agent, C)
                                                                   & (theta(J, object, F)
                                                                   & aspect(I, simple, J)))))))
                                              & aspect(G, simple, H))))))))))
```

**Fig. 4.** Logical form for (7)

## 4 Using logical forms

If you can parse a piece of text, then, you can construct a logical form for it. What can you do with it once you have done so?

Consider (8):

(8) I am allergic to eggs. Should I avoid eating pancakes?

This is a reasonably straightforward question. What resources would you need in order to answer it?

Firstly, you would need a great deal of basic information: pancakes contain eggs, eating things that contain foodstuffs that you are allergic to will make you ill, doing things that will make you ill is bad for you, if something is bad for you then you avoid doing it, . . .

Some of this information is reasonably easy to represent. The fact that pancaks contain eggs, for instance, can be captured by the formula in Fig. 5.

```
forall(X :: {sort(pancake, X)},
       exists(C :: {aspect(now, simple, C)},
              event(C, contain) & exists(Y, sort(egg, Y) & theta(C, object, Y)) & theta(C, agent, X)))
```

**Fig. 5.** If X is a pancake then it contains an egg

Other elements are extremely difficult to capture. Part of the problem here is that, as noted above, much of this information is intensional, but we have already seen using the $\lambda$-calculus allows use to state relationships between intensional objects. The logical form for *‘Should I avoid eating pancakes?’*, for instance, is as in Fig. 6.

```
utt(query,
    exists(A,
           (should(A,
                   lambda(B,
                          (event(B, avoid)
                          & (theta(B,
                                 object,
                                 lambda(C,
                                        exists(D,
                                               (event(D, eat)
                                               & (exists(E :: {sort(pancake, E, F, G)},
                                                       theta(D, object, E))
                                               & theta(D, agent, C))))))
                          & theta(B, agent, ref(lambda(H, speaker(H)))!1)))))
           & aspect(now, simple, A))))
```

**Fig. 6.** Logical form for *‘Should I eat pancakes?’*

The next problem is that writing rules in logic is very awkward. The rule we will need for expresing the fact that if something is bad for you then you should avoid doing it is given in Fig. 7.

```
forall(Y,
       forall(X,
              (exists(A,
                      (at(A, sort(Y, ~good))
                      & (for(A, X) & aspect(now, simple, A))))
              => exists(B,
                        (should(B,
                                lambda(C,
                                       (event(C, avoid)
                                       & (theta(C, object, Y)
                                       & theta(C, agent, X)))))
                        & aspect(now, simple, B))))))
```

**Fig. 7.** If something is bad for you then you should avoid it

Most people (even logicians) find it hard to write complex rules in logic, and the situation is made worse by the fact that the rules you have to write have to use the same terms and structures as the terms and structures used in logical forms. Because, in the examples above, we are using expressions like `event(X, ???)` to represent the fact that something is an eating event, we have to make sure that when we mention an event in a hand-coded rule we use the same kind of expression. The difficulty of bearing all the relevant conventions in mind when writing rules just makes this task harder.

The worst problem of all, however, is working out just what the relationships between the concepts denoted by various constructions in natural language are. Consider the word *‘good’*: if you look this word up in a dictionary, you generally find some rather unhelpful circular definition (e.g. ‘having the right qualities’:

Pocket Oxford Dictionary, ‘of a favorable character or tendency’: Merriam Webster on-line dictionary) followed by a large collection of examples. Dictionary entries of this kind provide very little information about what follows from saying that something is good, or about the circumstances under which you would say that something is good.

One way to deal with this problem is to tackle it head on. What we want to know is when you should say that something is good or bad, and what else we would be inclined to accept once we know this. So why not write down, in natural language, the kinds of things that we want to be able to reason with?

(9) a. eating P will make X ill if X is allergic to P
b. X is dangerous for Y if X will make Y ill
c. X is bad for Y if X is dangerous for Y
d. X should avoid Y if Y is bad for X
. . .

The rules in (9) are basic commonsense knowledge about very common words. By and large, conclusions that someone arrived by using these rules would be sensible, and anyone who did not have access to this kind of information could hardly be said to understand English.

Using natural language to write down the necessary knowledge about how words and concepts are related helps overcome the problems listed above:

- They are fairly easy to write. You do not have to become fluent in some complex formal language, and you do not have to familiarise yourself with the way that that language is being used, in order to add new rules. Just write them down in natural language.
- There is no risk that the terminology used for the logical forms derived for rules will differ from the terminology used for logical forms derived from general discourse. The same mechanisms will be used for both, so there is no possibility of inconsistency in the use of terminology.
- The rules say what is needed. In most cases, these rules will not constitute a ‘definition’, in the sense of a set of necessary and sufficient conditions for the use of a word. Instead they license specific inferences. It may, in fact, be the case that there is nothing more to be said about what a word means: that the meaning of a word is determined by its relations with other words. These relations may or may not provide a complete and precise set of constraints. They may, rather, constitute a set of ‘semantic traits’ (Cruse, 1986). Whether or not there is, indeed, more to the meaning of a word than that, certainly providing rules like these will support the derivation of a range of appropriate inferences.

## 5 Conclusions

Is there any alternative? It is clear that the ability to respond to (8) by saying *‘Yes’* (or, better, by saying *‘Yes, because pancakes contain eggs, and if you are allergic to eggs then eating things which contain eggs will make you ill’*) is highly desirable. A system that cannot do this cannot be said to understand the sentences that make up (8). It is equally clear that in order to do this, you have to be able to extract the relevant elements of your knowledge and construct a chain of inference from them. So if there is to be an alternative to the approach outlined here, it has to involve an alternative notion of inference.

It may well be, of course, that the specific logical forms outlined above are inadequate to the task. Indeed, the forms given here are inadequate: they were obtained by a system that follows Konrad et al. (1996) in allowing a variety of different forms to be obtained from a given tree by attaching ‘codes’ to words and modes of combination, and then using different codebooks to obtain different interpretations. This approach lets you construct logical forms which contain different amounts of detail, depending on the task you want to perform. Simple, coarse-grained interpretations are often useful for exposition (as here). More importantly, it is easier to reason about simple interpretations than about complex ones. It might, for instance, be convenient to use a simple representation for disambiguating a sentence, and then build a more detailed one for using the chosen reading for some more complex task. The suggestion that the logical forms given above are not complex or detailed enough, however, does not undermine the thesis that building logical forms and reasoning about them is the right approach to semantics. It just indicates that we need to develop more detailed logical forms, probably alongside improvements in our inference engines to cope with the extra complexity.

The only alternative is to define rules of inference that operate directly on texts. The textual entailment program (Dagan et al., 2005) is an attempt to follow this path. Any such attempt will, inevitably, have to cope with quantification. If you cannot infer *'Socrates is mortal'* from *'All men are mortal'* and *'Socrates is a man'* then you are not going to be able to handle more interesting discourses such as (8). There will, inevitably, be a temptation to start introducing rules like Fig. 8.

$$\frac{\begin{array}{c}\text{All Ps are Qs}\\ \text{X is a P}\end{array}}{\text{X is a Q}}$$

**Fig. 8.** Basic syllogism

Once you start introducing rules like this, the next obvious step is to normalise away various aspects of the surface form of the text. It would be extremely irritating, for instance, to have the rules in Fig. 9 as well.

$$\frac{\begin{array}{c}\text{All Ps are Qs}\\ \text{I am a P}\end{array}}{\text{I am a Q}} \quad \frac{\begin{array}{c}\text{All Ps are Qs}\\ \text{You are a P}\end{array}}{\text{You are a Q}} \quad \frac{\begin{array}{c}\text{All Ps are Qs}\\ \text{X was a P}\end{array}}{\text{X was a Q}} \quad \ldots$$

**Fig. 9.** Overspecific inference rules

The obvious way round this is to eliminate irrelevant aspects of the surface form, so that Fig. 9 would become Fig. 10.

$$\frac{\begin{array}{c}\text{All P be Q}\\ \text{X be P}\end{array}}{\text{X be Q}}$$

**Fig. 10.** Generalised syllogism

But this is exactly the route that led from Aristotle to the development of modern formal logic. Constructing logical forms is, exactly, the process of abstracting away from elements of the surface form that do not help you to decide which patterns are applicable. It thus seems very likely that attempts to make textual entailment more general and robust will lead to exactly the same end result. The key to constructing appropriate (linguistic or extra-linguistic) responses is inference: the only real question is not whether you make meaning representations that allow you to perform inference, but whether you do it well or badly.

# Bibliography

Allen, J. F., Perrault, C. R., 1980. Analysing intention in utterances. Artificial Intelligence 15, 148–178.

Austin, J., 1962. How to Do Things with Words. Oxford University Press, Oxford.

Barwise, J., Perry, J., 1983. Situations and Attitudes. Bradford Books, Cambridge, MA.

Church, A., 1936. An unsolvable problem of elementary number theory. American Journal of Mathematics 58(2), 345–363.

Cohen, P. R., Perrault, C. R., 1979. Elements of a plan-based theory of speech acts. Cognitive Science 7(2), 171–190.

Cruse, D. A., 1986. Lexical Semantics. Cambridge University Press, Cambridge.

Dagan, I., Magnini, B., Glickman, O., 2005. The PASCAL recognising textual entailment challenge. In: Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment.

Dalrymple, M., Lamping, J., Pereira, F. C. N., Saraswat, V., 1996. A deductive account of quantification in LFG. In: Kanazawa, M., Piñón, C., de Swart, H. (Eds.), Quantifiers, deduction and context. pp. 33–58.

Dowty, D. R., Wall, R. E., Peters, S., 1981. Introduction to Montague Semantics. D. Reidel, Dordrecht.

Konrad, K., Maier, H., Milward, D., Pinkal, M., 1996. An education and research tool for computational semantics. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96). Copenhagen, pp. 1098–1102.

McCarthy, J., 1980. Circumscription: a form of non-monotonic reasoning. Artificial Intelligence 13, 27–39.

Moëns, M., Steedman, M., 1988. Temporal ontology and temporal reference. Computational Linguistics 14(2), 15–28.

Montague, R., 1974. The proper treatment of quantification in ordinary English. In: Thomason, R. (Ed.), Formal Philosophy: Selected Papers of Richard Montague. Yale University Press, New Haven.

Reichenbach, H., 1958. The Philosophy of Space and Time. Dover Books, New York.

Reiter, R., 1980. A logic for default reasoning. Artificial Intelligence 13(1), 81–132.

Searle, J. R., 1969. Speech Acts: an Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

van Genabith, J., Crouch, R., 1997. How to glue a donkey to an f-structure. In: Bunt, H. C., Kievit, L., Muskens, R., Verlinden, M. (Eds.), 2nd International Workshop on Computational Semantics. University of Tilburg, pp. 52–65.

# Statistical Machine Translation: Current Trends

Ahmed Rafea
Computer Science and Engineering Department
American University in Cairo

**Abstract:** The paper first introduces the basics of Machine Translation SMT and then explains the main ideas of the two currently used techniques for Statistical (SMT) namely: hierarchical phrase based and syntax based. The paper concludes by comparing the performances of the systems developed using different techniques presented.

## Introduction

This paper is a sort of a literature reviews paper with emphasis on the current techniques used in the Statistical Machine Translation (SMT) community. The main objective of presenting this paper is to give a brief and simplified introduction to the topic. The paper first introduces the basics of (SMT) and then explains the main ideas of the two currently used techniques for SMT namely: hierarchical phrase based and syntax based.

The SMT basics cover the theory on which SMT is based, the language model, the translation model and decoding. The language model is based on constructing n-grams from monolingual corpus while the translation model is based on computing sets of probabilities that relates the source and target languages. There is strong relationship between the computations of these sets of translation probabilities and word alignment of bilingual corpora. Word alignment is the corner stone of building translation models. There are two main categories for word alignment: heuristic and statistical models. The translation models started by word based models that are well known as IBM models. The phrase translation model was then introduced and over performed the word based models. Once the language and translation models are built, they are used by a program called decoder which uses Artificial Intelligence search techniques to find the best translation of a given source sentence into a target sentence.

Recently the hierarchical and syntax based models are the ones dominating the area of SMT. The hierarchical phrase based SMT is based on a synchronous CFG, also known as a syntax-directed transduction grammar. In a synchronous CFG the elementary structures are rewrite rules with aligned pairs of right-hand sides and the decoding process can be described as a derivation from the source sentence using a synchronous CFG rules extracted from aligned words of a bilingual corpus. The syntax based SMT idea is established on parsing the target side of a bilingual corpus, aligning the words of both sides, and then developing translation rules that relate words, phrases, and sentences of the source language with the parse trees of the target language. This set of translation rules is the translation model of the syntax based SMT. The decoding process of the syntax based SMT is building a parse tree for the target sentence given a source sentence using the translation rules.

The second section will present the SMT basics while the third and forth sections will describe briefly the two currently used techniques mentioned here above. The last section concludes the paper by comparing the performances of the systems developed using the techniques presented in the paper and provides public domain resources that can used by interested audiences.

## Statistical Based Machine Translation Basics [Brown et al., 1993]

Given a foreign sentence **f,** we seek the native sentence **e** that maximizes the probability P(e | f), the “most likely” translation.  This is written as:

$$\operatorname*{argmax}_{e} P(e \mid f) \qquad (1)$$

Read this argmax as follows: “the native sentence e, out of all such sentences, which yields the highest value for P(e | f).

Using Bayes Rule, we can rewrite the expression for the most likely translation:

$$\operatorname*{argmax}_{e} P(e \mid f) = \operatorname*{argmax}_{e} P(e) * P(f \mid e)/p(f) \qquad (2)$$

P(f) can be considered as a constant here and hence can be taken out. That means the most likely translation e maximizes the product of two terms:

1. The chance that someone would say e in the first place, p(e) and

2. if he did say e, the chance that someone else would translate it into f, p(f|e).

Handling statistical machine translation like this is inspired from the noisy channel metaphor used for a lot of engineering problems, like actual noise on telephone transmissions. The noisy channel works like this. We imagine that someone has e in his head, but by the time it gets on to the printed page it is corrupted by “noise” and becomes f. To recover the most likely e, we reason about (1) what kinds of things people say e in English for example, and (2) how e in English gets turned into f , in French for example.

If we reason directly about translation using P(e | f), then our probability estimates have to be very good. On the other hand, if we break things apart using Bayes Rule, then we can theoretically get good translations even if the probability numbers are not that accurate. Suppose we assign a high value to P(f | e) only if the words in f are generally translations of words in e. The words in f may be in any order: we don't care. For example, if the string “the boy runs” passes, then “runs boy the” will also pass. Some word orders will be grammatical and some will not. Now let's talk about P(e). Suppose that we assign a high value to P(e) only if e is grammatical. So, the factor P(e) will lower the score of ungrammatical sentences. In effect, P(e) worries about English word order so that P(f | e) doesn't have to. That makes P(f|e)  easier to build than p(e|f).  In effect, those two probabilities, p(e), and p(f|e), represent two of the challenges of statistical machine translation. The first challenge is estimating the language model probability. The second one is estimating the translation model probability. There is another challenge which is finding the translation that maximizes the product of those two probabilities which is called decoding in the SMT terminology.  The process of finding this translation is in fact an optimal search problem. This is not easy in real life translation so the target would be using a suboptimal search algorithm [Brown et al, 1993]. By this, the research challenges of statistical machine translation are building the language model, building the translation model, and decoding. Each of those challenges will be discussed in more details in later sections.

### Language Model [Manning and Schutze, 1999]

We need to build a machine that assigns a probability P(e) to each English sentence e. This is called a language model. A simple idea is just to record every sentence that anyone ever says in English, in a database say one billion utterances. If the sentence “how's it going?” appears 76,413 times in that database, then we say P(how's it going?) = 76,413/1,000,000,000 = 0.000076413. One big problem is that many perfectly good sentences will be assigned a P(e) of zero, because we have never seen them before.  People seem to be able to judge whether or not a string is belonging to a certain language without storing a database of utterances.

This seems to be done by breaking the sentence down into components. If the components are good, and if they combine in reasonable ways, then we say that the string is in this language. Word substring is called an n-gram. If n=1, we say unigram. If n=2, we say bigram. If n=3, n we say trigram. For example the trigram language model is the set of the probabilities for all words x, y, and z in a monolingual corpus computed using the following formula:

p(z | x y) = number-of-occurrences (“xyz”) / number-of-occurrences (“xy”) (3)

The probability p(e) of a sentence **e** composed of n-words $e_1 \ldots e_n$, to be grammatically correct based on the trigram model, is calculated using the following formula:

$$p(e) = \prod_{i=1}^{n} P(e_i \mid e_{i-2}, e_{i-1}) \qquad (4)$$

N-gram models can assign non-zero probabilities to sentences they have never seen before. The only way you'll get a zero probability is if the sentence contains a previously unseen trigram if we are 3-gram model. In that case, we can do smoothing. If “z” never followed “xy” in our text, we might further wonder whether “z” at least followed “y”. If it did, then maybe “xyz” isn't so bad. If it didn't, we might further wonder whether “z” is even a common word or not. If it's not even a common word, then “xyz” should probably get a low probability. Therefore, there is a need to combine more than one type of n-gram; 1-gram, 2-gram, 3-gram, and a constant in order to have a smoothed language model. Instead of using the simple trigram formula given in equation (3) we can use this smoothed formula to compute the language model:

p(z | x y) = 0.95*number-of-occurrences (“xyz”) / number-of-occurrences (“xy”) + 0.04*number-of-occurrences (“yz”) / number-of-occurrences (“z”)+ 0.008 * number-of-occurrences (“z”) / total-words-seen + 0.002 (5)

It's handy to use different smoothing coefficients in different situations. You might want 0.95 in the case of xy(z), but 0.85 in another case like ab(c). For example, if “ab” doesn't occur very much, then the counts of “ab” and “abc” might not be very reliable. Notice that as long as we have that “0.002” in there, then no conditional trigram probability will ever be zero, so P(e) will never be zero. That means we will assign some positive probability to any string of words, even if it's totally ungrammatical.

## Translation Model

In statistical machine translation it is necessary to model the translation probability P(fl e).

Most SMT models (Brown et al., 1993; Vogel et al., 1996) try to model word-to-word correspondences between source and target words using an alignment mapping from source position to target position. The word alignment models are often the basis and/or the starting point of all types of statistical machine translation systems: single-word-based statistical machine translation systems [Berger et al. 1994; Wu 1996; Wang and Waibel 1998; Nießen et al. 1998; Garc´ıa-Varea, Casacuberta, and Ney 1998; Och, Ueffing, and Ney 2001; Germann et al. 2001], phrase-based statistical machine translation [Och and Weber 1998; Och, Tillmann, and Ney 1999], example-based translation systems [Brown 1997], syntax based statistical machine translation [Yamada and Knight 2001], and hierarchical based translation [Chaing 2007].

## Word Alignment

There are two general approaches to computing word alignments: heuristic models and statistical alignment models. In the following, we describe both types of models and compare them from a theoretical viewpoint.

**Heuristic Models:** Considerably simpler methods for obtaining word alignments use a function of the similarity between the types of the two languages (Och and Ney 2003; Ker and Chang 1997). Frequently, variations of the Dice coefficient (Dice 1945) are used as this similarity function. For each sentence pair, a matrix including the association scores between every word at every position is then obtained:

$$\text{dice}(i, j) = 2 * C(e_i , f_j)/C(e_i) \cdot C(f_j) \qquad (6)$$

C(e, f ) denotes the co-occurrence count of e and f in the parallel training corpus. C(e) and C(f ) denote the count of e in the target sentences and the count of f in the source sentences, respectively. From this association score matrix, the word alignment is then obtained by applying suitable heuristics. One method is to align $f_j$ with $e_i$ with the largest association score.

**Statistical Alignment Models**: In statistical machine translation, we try to model the translation probability P(f |e) which describes the relationship between a source language string f and a target language string e. We can rewrite the probability P(f|e) by introducing the 'hidden' alignments **a** to be:

$$P(f,a|e)= \sum_{1}^{n} P(f,a_i|e) \qquad (7)$$

where n is the possible number of alignments of source to target sentences words

In general, the statistical model depends on a set of unknown parameters θ that is learned from training data. To express the dependence of the model on the parameter set, we use the following notation:

$$P(f,a|e) = p_\theta(f,a|e) \qquad (8)$$

In case of the statistical alignment model, the model has to describe the relationship between a source language string and a target language string adequately. To train the unknown parameters θ, we are given a parallel training corpus consisting of S sentence pairs $\{(f_s, e_s) : s = 1, . . . , S\}$. For each sentence pair $(f_s, e_s)$, the alignment variable is denoted by **a**. The unknown parameters θ are determined by maximizing the likelihood on the parallel training corpus:

$$\hat{\Theta} = \underset{\theta}{\text{argmax}} \prod_{s=1} \sum_{a} p_\theta(f_s,a|e_s) \qquad (9)$$

Typically, for this model, the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) or some approximate EM algorithm is a useful tool for solving this parameter estimation problem. Although for a given sentence pair there is a large number of alignments, we can always find a best alignment:

$$\hat{a} = \underset{a}{\text{argmax}}\, p_\theta(f_s,a|e_s) \qquad (10)$$

In effect the statistical model parameters represent the translation models of statistical machine translation systems such as those described in Brown, Della Pietra, Della Pietra, and Mercer (1993).

**Comparison of Heuristic Models and Statistical Models:** The main advantage of the heuristic models is their simplicity. They are very easy to implement and understand. Therefore, variants of the heuristic models described above are widely used in the word alignment literature. One problem with heuristic models is that the use of a specific similarity function seems to be completely arbitrary. The literature contains a large variety of different scoring functions, some including empirically adjusted parameters. In a comparative study of alignment models conducted by Och and Ney(2003), it was found that the approach of using statistical alignment models is more coherent as the general principle for coming up with an association score between words results from statistical estimation theory, and the parameters of the models are adjusted such that the likelihood of the models on the training corpus is maximized.

## Word Based Translation Models

The most widely used word based translation models are the five models introduced by IBM and described in [Brown et al. 1993]. The following paragraphs present briefly these five models.

The first statistical translation model introduced was IBM model-1 [Brown et al. 1993]. This model was based on building all possible alignments between words in the source sentence and words in the target sentence, computing the translation probabilities of foreign words given native words t(f|e), maximizing these probabilities iteratively using EM algorithm [Dempster et al. 1977].

IBM2 is similar to IBM1, they both use the same training algorithm but IBM2 use two probabilities instead of one. IBM2 depends upon the translation probability in addition to the distortion probability. The distortion probability is the probability that the source language word at position j is aligned to the target language word at position I given the lengths of both the source and target languages sentences d(j|I,m, l) where m, and l are the lengths of the source and target languages.

IBM3 uses the same probabilities as model2 in addition to the fertility probability n(⍰|ei ) which means that the word ei in the target sentence is mapped to ⍰⍰words in the source sentence. Another difference is how the distortion and fertility probabilities for e0 are treated. The e0 purpose is to account for those words in the source string that cannot readily be accounted for by other words in the target string.

Model 4 is modifications of model 3 to account for translating words in a target string constituting phrases as units into source string. Sometimes, a translated phrase may appear at a spot in the source string different from that at which the corresponding target phrase appears in the target string. The distortion probabilities of Model 3 do not account well for this tendency of phrases to move around as units. Movement of a long phrase will be much less likely than movement of a short phrase because each word must be moved independently.

Both Model 3 and Model 4 ignore whether or not a source position has been chosen. In addition, probability mass is reserved for source positions outside the sentence boundaries. For both of these reasons, the probabilities of all valid alignments do not sum to unity in these two models. Such models are called deficient [Brown, et al., 1993). Model 5 is a reformulation of Model 4 with a suitably refined alignment model to avoid deficiency.

## Phrase Based Translation Model

The basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations. Given a sentence pair and a corresponding word alignment, phrases are extracted following the criterion in [Och and Ney 2004]. A phrase (or bilingual phrase) is any pair of m source words and n target words that satisfies two basic constraints:

1. Words are consecutive along both sides of the bilingual phrase,

2. No word on either side of the phrase is aligned to a word out of the phrase.

It is infeasible to build a dictionary with all the phrases. That is why the maximum size of any given phrase is limited. Also, the huge increase in computational and storage cost of including longer phrases does not provide a significant improve in quality [Kohen et al. 2003] as the probability of reappearance of larger phrases decreases. The phrases of length X or less (usually X equal to 3 or 4) . Then, phrases up to length Y (Y greater than X) are added if they cannot be generated by smaller phrases [Crego et al. 2005]. Given the collected phrase pairs, the phrase translation probability distribution is estimated by relative frequency:

$$P(f|e) = N(f, e)/ N(e) \qquad (11)$$

Where N(f,e) means the number of times the phrase f is translated by e.

## Decoding

The decoding process is actually a search problem. In general, the search problem for statistical MT even using only Model 1 of Brown et al. (1993) is NP-complete (Knight 1999). Therefore, it is not expected to develop efficient search algorithms that are guaranteed to solve the problem without search errors. Hence, the art of developing a search algorithm lies in finding suitable approximations and heuristics that allow an efficient search without committing too many search errors. It should be possible to translate a sentence of reasonable length within a few seconds of computing time. The search algorithm should be able to scale up to very long sentences with an acceptable computing time.

To meet these aims, it is necessary to have a mechanism that restricts the search effort. Och and Ney [2004] accomplished such a restriction by searching in a breadth-first manner with pruning: beam search. In pruning, the set of considered translation candidates (the “beam”) are only to the promising ones. [Och and Ney 2004]

Many of the other search approaches suggested in the literature do not meet the described aims. According to Och and Ney [2004]:

- "Neither optimal A* search (Och, Ueffing, and Ney 2001) nor optimal integer programming (Germann et al. 2001) for statistical MT allows efficient search for long sentences.
- Greedy search algorithms (Wang 1998; Germann et al. 2001) typically commit severe search errors.
- Other approaches to solving the search problem obtain polynomial time algorithms by assuming monotone alignments (Tillmann et al. 1997) or imposing a simplified recombination structure (Nießen et al. 1998). Others make simplifying assumptions about the search space (Garc´ıa-Varea, Casacuberta, and Ney 1998; Garc´ıa-Varea et al. 2001), as does the original IBM stack search decoder (Berger et al. 1994). All these simplifications ultimately make the search problem simpler but introduce fundamental search errors."

## Hierarchical Phrase Based Statistical Machine Translation

Chiang [2005] modeled the hierarchical phrase based SMT based on a synchronous CFG, elsewhere known as a syntax-directed transduction grammar (Lewis and Stearns 1968). In a synchronous CFG the elementary structures are rewrite rules with aligned pairs of right-hand sides:

$$X \rightarrow \_\gamma, \alpha, \sim\_$$

where X is a nonterminal, γ and α are both strings of terminals and nonterminals, and ~ is a one-to-one correspondence between nonterminal occurrences in γ and nonterminal occurrences in α. For example, if we have the following Arabic sentence[1]

استراليا هي واحدة من البلاد القلائل التي لها علاقات دبلوماسية مع كوريا الشمالية

and its English translation is

Australia is one of the few countries that have diplomatic relations with North Korea

We can formalize the following synchronous CFG rule:

X → < $X_1$ هي $X_2$ التي لها $X_3$ مع $X_4$ , $X_1$ is $X_2$ that have $X_3$with $X_4$> (12)

Indices are used to indicate which nonterminal occurrences are linked by ~. The conventional phrase pairs would be formalized as:

X → < استراليا, Australia> (13)

X → < كوريا الشمالية, North Korea> (14)

X → < علاقات دبلوماسية, diplomatic relations> (15)

X → < واحدة من البلاد القلائل , one of the few countries > (16)

Two more rules complete our example:

S → < $S_1 X_2$, $S_1 X_2$ > (17)

S → < $X_1$, $X_1$ > (18)

A synchronous CFG derivation begins with a pair of linked start symbols or a single start symbol. For an example using these rules, the following is the derivation for the above sentence and its translation:

<S > → < $X_1$, $X_1$>

→ < $X_1$ هي $X_2$ التي لها $X_3$ مع $X_4$ , $X_1$ is $X_2$ that have $X_3$with $X_4$>

→ < استراليا, هي $X_2$ التي لها $X_3$, مع $X_4$ , Australia is $X_2$ that have $X_3$ with $X_4$>

→ < استراليا, هي واحدة من البلاد القلائل التي لها $X_3$, مع $X_4$ ,Australia is one of the few countries that have $X_3$ with $X_4$>

→ < استراليا, هي واحدة من البلاد القلائل التي لها علاقات دبلوماسية مع $X_4$ , Australia is one of the few countries that have diplomatic relations with $X_4$>

→ < استراليا, هي واحدة من البلاد القلائل التي لها علاقات دبلوماسية مع كوريا الشمالية , Australia is one of the few countries that have diplomatic relations with North Korea>

The bulk of the grammar consists of automatically extracted rules. The extraction process begins with a word-aligned corpus: a set of triples (f, e, ~) where f is a source sentence, e is a target sentence, and ~ is a (many-to-many) binary relation between positions of f and positions of e. The word alignments can be obtained by running GIZA++ (Och and Ney 2000) on the corpus in both directions, and forming the union of the two sets of word alignments. A set of rules can then be extracted from each word-aligned sentence pair. These rules are consistent with the word alignments. This can be thought of in two steps. First, initial phrase pairs are identified using the same criterion as most phrase-based systems (Och and Ney 2004), Second, in order to obtain rules from the phrases, phrases that contain other phrases are recognized and the subphrases are replaced with nonterminal symbols.

Given a French sentence f , a synchronous CFG will have, in general, many derivations that yield f on the French side, and therefore (in general)many possible translations e. Chiang [2007] defined a model over derivations D to predict which translations are more likely than others. Following Och and Ney (2002), he departed from the traditional noisy-channel approach and used a more general log-linear model over derivations D:

$$P(D) \propto \Pi_{\varphi i}(D)^{\lambda i} \qquad (19)$$

where the ϕi are features defined on derivations and the λi are feature weights. One of the features is an m-gram language model PLM(e); the remainder of the features are defined as products of functions on the rules used in a derivation. The factors other than the language model factor can be put into a particularly convenient form. A weighted synchronous CFG is a synchronous CFG together with a function w that assigns weights to rules. This function induces a weight function over derivations:

$$w(D) = \prod_{(X \rightarrow \_\gamma,\alpha\_) \in D} w(X \rightarrow <\gamma,\alpha>)\_ \qquad (20)$$

If we define

$$w(X \rightarrow <\gamma,\alpha>) = \prod_{I \;\; /= LM} \varphi i(X \_\_\gamma,\alpha\_)\lambda_i \qquad (21)$$

then the probability model becomes

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times w(D) \qquad (22)$$

It is easy to write dynamic-programming algorithms to find the highest-weight translation or k-best translations with a weighted synchronous CFG.

In brief, the decoder proposed by Chiang [2007] is a CKY (Cocke-Kasami-Younger) parser with beam search together with a postprocessor for mapping source language derivations to target language derivations. Given a source sentence f , it finds the target sentence yield of the single best derivation. For more details on the decoding process reader can refer to [Chaing 2007].

## Syntax Based Statistical Machine Translation

The syntax based SMT idea is established on parsing the target side of a bilingual corpus, aligning the words of both sides, and then developing translation rules that relate words, phrases, and sentences of the source language with the parse trees of the target language. This set of translation rules is the translation model of the syntax based SMT. The decoding process of the syntax based SMT is building a parse tree for the target sentence given a source sentence using the translation rules.

Yamada and Knight [200]) presented a syntax-based translation model that statistically models the translation process from an English parse tree into a foreign language sentence.

They conducted a small-scale experiment to compare the performance with IBM Model 5, and got better alignment results.

The GHKM extractor [Galley et al. 2001] learns translation rules from an aligned parallel corpus where the target side has been parsed. This corpus is conceptually a list of tuples of <source sentence, target tree, bi-directional word alignments> which serve as training examples. Now we have a precise problem statement: learn the set of rules □A(S; T). It is not immediately clear how such a set can be learned from the triple (S; T;A). Fortunately, these rules can be inferred directly from a structure called an alignment graph. Formally, the alignment graph corresponding to S, T, and A is just T, augmented with a node for each element of S, and edges from leaf node t □ T to element s □ S iff A aligns s with t. In the example presented in Galley et al. [2004] paper they assumed that the alignment graph is connected, i.e. there are no unaligned elements. It turns out that it is possible to systematically convert certain fragments of the alignment graph into rules of □A(S; T). A fragment of a directed, acyclic graph G, is defined to be a nontrivial subgraph G' of G such that if a node n is in G' then either n is a sink node of G' (i.e. it has no children) or all of its children are in G' (and it is connected to all of them). To demonstrate this idea I will give an example from English to Arabic which is the same one given in [Galley et al., 2004] but it was from English to French. Figure 1 describes the alignment graph for the sentence "He does not go" and its translation to Arabic

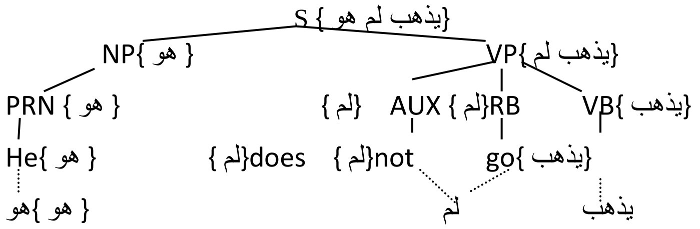
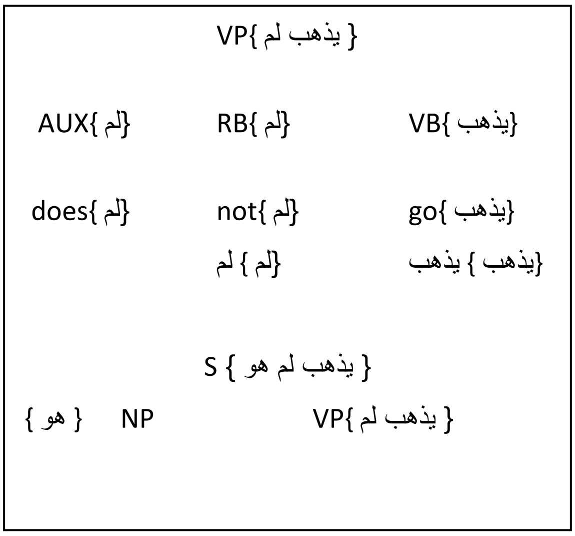
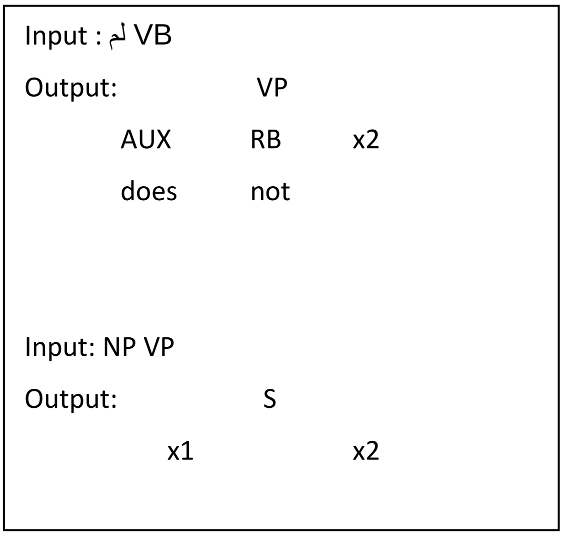


Figure 1- An alignment graph. The nodes are annotated with their spans

In Figure 2, we show two examples of graph fragments of the alignment graph of Figure 1.

The span of a node n of the alignment graph is the subset of nodes from S that are reachable from n. A span is said to be contiguous if it contains all elements of a contiguous substring of S. The closure of span(n) is the shortest contiguous span which is a superset of span(n). The alignment graph in Figure 1 is annotated with the span of each node. Take a look at the graph fragments in Figure 2. These fragments are special: they are examples of frontier graph fragments [Gallet et. al. 2006]. A frontier graph fragment of an alignment graph G is defined to be a graph fragment such that the root and all sinks are in the frontier set. Nodes of G whose spans and complement spans are non overlapping form the frontier set F ε G. The complement span of n is the union of the spans of all nodes n' in graph G that are neither descendants nor ancestors of n. Frontier graph fragments have the property





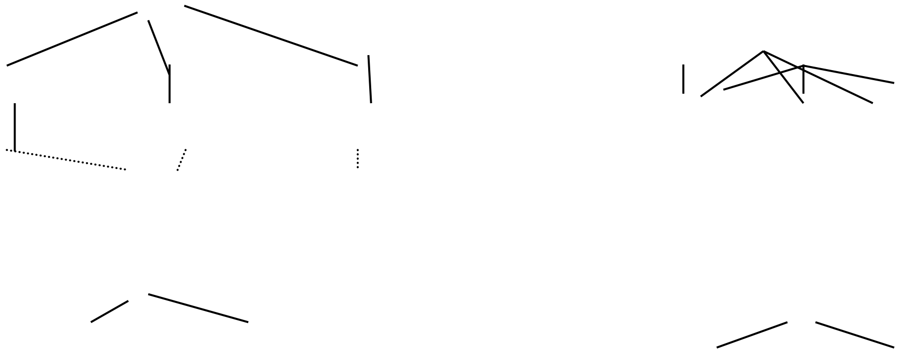


Figure 2 Two frontier graph fragments and the rules induced from them

that the spans of the sinks of the fragment are each contiguous and form a partition of the span of the root, which is also contiguous. This allows the following transformation process:

1. Place the sinks in the order defined by the partition (i.e. the sink whose span is the first part of the span of the root goes first, the sink whose span is the second part of the span of the root goes second, etc.). This forms the input of the rule.
2. Replace sink nodes of the fragment with a variable corresponding to their position in the input, then take the tree part of the fragment (i.e. project the fragment on T). This forms the output of the rule.

The syntax SMT translation model is the set of rules extracted from the alignment graph.

The decoding process is to construct a parse tree of the target sentence given the source sentence and the translation model. The rules extracted can be given certain probabilities. More than one tree can be generated and the most probable tree is to be selected by the decoder.

## Concluding Remarks

This paper was intended to give a general introduction to the topic of SMT with emphasis on the techniques that are currently used. The SMT community provides a lot of resources for interested scientists such that they can advance this area more.  I will present here the resources that are widely used by the SMT community. GIZA++[1] is an extension of the program GIZA (which was part of the SMT toolkit EGYPT) that was developed during a summer research workshop in 1999 at the Center for Language and Speech Processing at Johns Hopkins University (CLSP/JHU). Giza++ is used by many scientists to build the translation model of a word based SMT. It produces as output words alignment of parallel corpus that can be used further by other advanced systems. This tool can be downloaded freely from the web. Language models can also be built using free available software from Carnegie Melon University (CMU) and Stanford Research Institute (SRI).The CMU-Cambridge Statistical Language Modeling toolkit is a suite of UNIX software tools to facilitate the construction and testing of statistical language models[2].  SRILM is a toolkit for building and applying statistical language models (LMs), primarily for use in speech recognition, statistical

[1] http://www.fjoch.com/GIZA++.html

[2] http://mi.eng.cam.ac.uk/~prc14/toolkit.html

tagging and segmentation, and machine translation[3]. There are a set of decoders available for free. The word based decoder is REWRITE. The phrase based decoders are Pharaoh[4] and Moses[5]. Moses decoder is used extensively nowadays to build baseline systems for scientists working on enhancing SMT systems. The decoders for hierarchical based SMT and syntax based SMT have not been availed yet as open source.

For readers who are interested to have a look at the scores of SMT systems developed in the research and industrial community, the National Institutes of Standards and Technology (NIST) annual competition[6] .

There is a need to build capacity in this are at the national level for SMT as the research in this area is very expensive. It needs a lot of language, human, and computational resources.

## References

Adam L Berger., Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Harry Printz, and Lubos Ure ˇ s. 1994. The Candide system for machine translation. In Proceedings of the ARPA Workshop on Human Language Technology, pages 157–162, Plainsboro, NJ, March.

Peter E Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation”, Annual Meeting of the Association of Computing Linguistics, 1993.

David Chiang, "A hierarchical phrase-based model for statistical machine translation. 2005. In Proc. ACL, pages 263–270.

David Chiang, " Hierarchical Phrase-Based Translation" Computational Linguistics, Volume , 33 No.2, June 2007, pp 201-228

Josep M. Crego, Marta R. Costa-juss_a, Jos´e B. Mari.no, Jos´e A. R. Fonollosa, “Ngram-based versus Phrase-based Statistical Machine Translation”, International Workshop on Spoken Language Translation, 2005.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society,Series B, 39(1):1–22.

M. Galley, M. Hopkins, K. Knight, D. Marcu "What's in a Translation Rule?", Proc. NAACL-HLT, 2004.

Ismael Garc´ıa-Varea, Francisco Casacuberta, and Hermann Ney. 1998. An iterative, DP based search algorithm for statistical machine translation. In Proceedings of the International Conference on Spoken Language Processing (ICSLP’98), pages 1235–1238, Sydney, Australia, November.

Ismael Garc´ıa-Varea, Franz Josef Och, Hermann Ney, and Francisco Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL), pages 204–211, Toulouse, France,July.

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2001. Fast decoding and optimal decoding for machine translation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL), pages 228–235, Toulouse, France, July.

Knight, Kevin. 1999. Decoding complexity in word-replacement translation models. Computational Linguistics, 25(4):607–615.

[3] http://www.speech.sri.com/projects/srilm/

[4] http://www.isi.edu/licensed-sw/pharaoh/

[5] http://sourceforge.net/projects/mosesdecoder/

[6] http://www.nist.gov/speech/tests/mt/2008/doc/mt08_official_results_v0.html

P. Koehn, F. Och, and D. Marcu, .Statistical phrasebased translation,. Proc. of the Human Language Technology Conference, HLT-NAACL'2003, May 2003.

Christopher D. Manning and Hinrich Schutze, “Foundations of Statistical Natural Language Processing”. The MIT Press, Cambridge, Massachusetts, London, England, 1999. Chapter 6.

I. Dan Melamed. 2000. Models of translational equivalence among words. Computational Linguistics, 26(2):221–249.

Sonja Nießen, Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1998. A DP-based search algorithm for statistical machine translation. In COLING-ACL ’98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pages 960–967, Montreal, Canada, August.

Franz Josef Och. and Hans Weber. 1998. Improving statistical natural language translation with categories and rules. In COLING-ACL ’98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, pages 985–989, Montreal, Canada, August.

Franz Josef Och , Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20–28, University of Maryland, College Park, June.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the ACL, pages 440–447, Hong Kong.

Franz Josef Och,, Nicola Ueffing, and Hermann Ney. 2001. An efficient A* search algorithm for statistical machine translation. In Data-Driven Machine Translation Workshop, pages 55–62, Toulouse, France, July.

Franz Josef Och and Hermann Ney, “A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, Volume 29, No.1, March 2003, pp 19-51

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. Computational Linguistics, 30:417–449.

Christoph Tillmann, Stephan Vogel, Hermann Ney, and Alex Zubiaga. 1997. A DP-based search using monotone alignments in statistical translation. In Proceedings of the 35th Annual Conference of the Association for Computational Linguistics, pages 289–296, Madrid, July.

S. Vogel, H. Ney, and C. Tilhnann. 1996. HMMbased word alignment in statistical translation. In COLING '96: The 16th Int. Conf. on Computational Linguistics, pages 836-841, Copenhagen, August.

Ye-Yi Wang and Alex Waibel. 1998. Fast decoding for statistical machine translation. In Proceedings of the International Conference on Speech and Language Processing, pages 1357–1363, Sydney, Australia, November.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In Proceedings of the 34th Annual Conference of the Association for Computational Linguistics (ACL ’96), pages 152–158, Santa Cruz, California, June.

Kenji Yamada, and , Kevin Knight (2001) *A syntax-based statistical translation model.* In Proceedings of ACL01

# Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage

Sameh Alansary[*†] Magdy Nagi[*††] Noha Adly[*††]
Sameh.alansary@bibalex.org magdy.nagi@bibalex.org noha.adly@bibalex.org

[*] Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.
[†] Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University , El Shatby, Alexandria, Egypt.
[††] Computer and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria Egypt.

## Abstract:

*This paper sheds light on four axes. The first axis deals with the levels of corpus analysis e.g. morphological analysis, lexical analysis, syntactic analysis and semantic analysis. The second axis captures some attempts of Arabic corpora analysis. The third axis demonstrates different available tools for Arabic morphological analysis (Xerox, Tim Buckwalter, Sakhr and RDI). The fourth axis is the basic section in the paper; it deals with the morphological analysis of ICA. It includes: selecting and describing the model of analysis, pre-analysis stage and full text analysis stages.*

## 1. Introduction:

It can be said that corpus analysis highly depends on the availability of previous history of the analysis, because information with decisive solutions in one stage, are used in the next stages of the analysis . The major difference between creating and analyzing a corpus is that while the creator of a corpus has the option of adjusting what is included in the corpus to compensate for any complications that arise during the creation of the corpus, the corpus analyst is confronted with a fixed corpus, and has to decide whether to continue with the analysis, even if the corpus is not entirely suitable for analysis, or find a new corpus altogether (Meyer, 2002).

It is important, first of all, to begin the process with a very clear goal in mind; that the analysis should involve more than a simple (count) of linguistic features. Also, it is necessary to select the appropriate corpus for analysis: to make sure, for instance, that it contains the right types of texts for the analysis and that the samples to be examined are lengthy enough. Also, if more than one corpus is to be compared, the corpora must be comparable, or else the analysis will not be valid. After these preparations are made, the analyst must find the appropriate software tools to conduct the study, code the results, and finally subject these results to the appropriate statistical tests. If all of these steps are followed, the analyst can rest assured that the results obtained are valid and the generalizations that are made have a solid linguistic basis (Meyer, 2002).

## 2. Levels of corpus analysis:

Linguistic analysis has more than one level of analysis such as morphological analysis, lexical analysis, syntactic analysis (parsing) and semantic analysis. The focus of corpus analysis is empirical, whereas the interpretation can be either qualitative or quantitative.

**Morphological analysis** is the most basic type of linguistic corpus analysis because it forms the essential foundation for further types of analysis (such as syntactic parsing and semantic field annotation), and because it is a task that can be carried out with a high degree of accuracy by a computer. The aim of morphological analysis of corpora is not only to assign to each lexical unit in the text a code indicating its part of speech, but also to indicate other morphological information. There are many morphological dimensions for describing verbs, nouns and particles. Consequently, the morphological tag can either be extended to include all morphological features (including additional features such as transitivity, perfectness and voice for verbs, number, gender and derivation for nouns and agglutination for particles), or contracted to include only the main morphological tags and other morphological features are indicated separately (see Al-Sulaiti & Atwell, 2001).

There are two main approaches in morphological generation and analysis; namely**,** the Two-level approach (Non-concatenative approach) and the Concatenative approach. The two-level approach defines two levels of strings; lexical strings which represent morphemes, and surface strings which represent surface forms.

The two-level approach views the Arabic word vertically, as a composition of two layers; root and pattern. In Arabic, for instance, there is a clear sense that the forms in table 1 are morphologically related to one another, although they do not share isolable strings of segments in concatenated morphemes:

| Word | Gloss |
|---|---|
| (kataba) كَتب | He write |
| (makotuwb) مكتوب | Written |
| (kutub) كُتب | Books |
| (kutiba) كُتب | Be written |
| (kitab) كتاب | Book |
| (kut~Ab) كُتاب | Writers/Quran school |
| (kAtib) كاتب | Writer |

Table 1: variant words related to each other.

The Concatenative morphology, which appears almost exclusively in the more familiar languages, involves prefixation or suffixation only. In other words, morphemes are discrete elements linearly concatenated at the right or the left end of the base of the morphological operation (Hockett,1947). Although the concatenative approach cannot predict the word-pattern automatically, it compensates for this by keeping a large database of Arabic lexemes with their related information including word-patterns.

Hence, the input word passes through less complicated processing than in the two-level approach.

**Lexical analysis** is the process of taking an input string of characters and producing a sequence of symbols called "lexical tokens", which may be handled easily by lexical analyzers (parsers, programs of lexical analysis). These analyzers have two phases of analysis; i.e. the scanning phase and tokenization phase, the process of determining and classifying a clause into tokens.

In **Syntactic analysis** the linear sequence of tokens is replaced by a tree structure through building a parse tree in order to define the language’s syntax according to the rules of formal grammar , and generate, or transform the parse tree. Parsing is also crucial in various applications in natural language processing, including text-to-speech synthesis, and machine translation (Patten, 1992).

**Semantic analysis** is one of the most important levels of analysis. In this level, the semantic information is added into the parse tree, the symbol table is built, and finally semantic checks are performed. Logically, semantic analysis intermediates the parsing phase and the code generation phase because it requires a complete parse tree. In machine learning, the semantic analysis of a corpus is the task of building structures that capture concepts from a large set of documents. It does not generally involve prior semantic understanding of the documents.

## 3. Some attempts of Arabic corpora analysis:

**CLARA (Corpus Linguae Arabicae):** The ultimate goal of this project is building a balanced and annotated corpus. The annotation should be done for morphological boundaries and Part Of Speech (POS). Some tools and databases are built for the sake of the analysis stage; for instance, a training corpus with marked morphological boundaries consisting of 100,000 words, a database of strings with marked morphological boundaries and another training corpus with annotation of parts of speech. Currently, the analyzed size of this corpus is about 15,000 words. The parts of speech tagset is based on the EAGLES recommendations[1].

**The Penn Arabic Treebank:** is a corpus of one million words of Arabic. Treebank is designed to support the development of data-driven approaches to natural language processing (NLP), human language technologies, automatic content extraction (topic extraction and/or grammar extraction), cross-lingual information retrieval, information detection, and other forms of linguistic research on Modern Standard Arabic (MSA) in general. There are two distinct phases of analysis in the Penn Arabic Treebank; namely, Part-of-Speech (POS) tagging, and Arabic Treebanking (ArabicTB) (Abdelali, 2004).

**Prague Arabic Dependency Treebank**: is a project of analyzing large amounts of linguistic data in Modern Written Arabic in terms of the formal representation of language that originates in the Functional Generative Description (Sgall et al. 1986, Sgall & Hajičová 2003). Prague Arabic Dependency Treebank (PADT) does not only

[1] http://www.ilc.pi.cnr.it/

consist of multi-level linguistic annotations of the Modern Standard Arabic, but it even has a variety of unique software implementations, designed for general use in Natural Language Processing (NLP).

The linguistic analysis takes place in three stages: the morphological level (inflection of lexemes), the analytical level (surface syntax), and the tectogrammatical level (underlying syntax) (Smrž, 2004). The morphological level of PADT has for long been the same as that available in Penn Arabic Treebank, Part 2. However, PADT has adopted the way of Buckwalter Arabic Morphological Analyzer.

## 4. Existing Arabic Morphological analyzers:

There are many morphological analyzers for Arabic, some of them are available for research and evaluation while the others are proprietary commercial applications. Among those known in the literature are Xerox Arabic Morphological Analysis and Generation (Beesley, 1998a,2001), Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002), Sakhr and RDI Arabic Morphological Analyzer.

**Xerox Morphology:** is “based on solid and innovative finite-state technology” (Dichy & Fargaly, 2003). It adopts the root-and-pattern approach and includes 4,930 roots and 400 patterns, effectively generating 90,000 stems. Its main advantage is that it is rule based with wide coverage. It also reconstructs vowel marks and provides an English glossary for each word. At Xerox, the treatment of Arabic starts with a lexc grammar where prefixes and suffixes concatenate to stems in the usual way, and where stems are, similarly, represented as a concatenation of a root and a pattern (Beesley, 1998a & b).

The system includes more classical entries, and lacks more grammar-lexis specifications. Additional disadvantages of Xerox morphology are:

1. Overgeneration in word derivation, The distribution of patterns for roots is not even, and although each root was hand-coded in the system to select from among the 400 patterns, the task is understandably tedious and prone to mistakes as shown in table 2.

| Word | Transliteration | Root | Meaning |
|---|---|---|---|
| قال | qaal | qwl | Say (verb) |
| | | qlw | Fry (active participle) |
| | | qll | decrease (active participle) |

Table 2: Example of over generation.

The first root analysis is valid, while the other two are illegal derivations that have no place in the Arabic language, and not mentioned in classical dictionaries.

2. Underspecification: in POS classification, which makes it unsuited for serving a syntactic parser. Words are only classified into: (verbs, nouns which include adjectives and adverbs, participles and function words which, in turn, include prepositions, conjunctions, subordinating conjunctions, articles, negative particles…etc).

3. Increased rate of ambiguity: due to the above-mentioned factors, the system suffers from a very high level of ambiguity, as it provides so many analyses (many of them spurious) for most words (Attia , 2006).

**Buckwalter Arabic Morphological Analyzer:** It uses a concatenative lexicon-driven approach where morphotactics and orthographic rules are built directly into the lexicon itself instead of being specified in terms of general rules that interact to realize the output (Buckwalter , 2002). Buckwalter Morphology contains of 38,600 lemmas, and is used in LDC Arabic POS-tagger, Penn Arabic Treebank, and the Prague Arabic Dependency Treebank. It is designed as a main database of word forms and it interacts with other concatenation databases. Every word form is entered separately, Buckwalter's morphology reconstructs vowel marks and provides English glossary. It takes the stem as the base form and root information is provided (Attia , 2000). In Buckwalter analyzer, Arabic words are segmented into prefix, stem and suffix strings according to the following rules[2]:

- the prefix can be 0 to 4 characters long.
- the stem can be 1 to infinite characters long.
- the suffix can be 0 to 6 characters long.

**Sakhr Arabic Morphological Processor**: It is a morphological analyzer-synthesizer that provides basic analysis for a single Arabic word, covering the whole range of modern and classical Arabic. The analyzer identifies all possible stem forms of a word; i.e. extracting its basic form stripped from the affixes, , the morphological data such as root, the Morphological Pattern (MP), and its part of speech. The synthesizer works in a reverse mode to regenerate the word from its morphological forms (stem, root, morphological pattern, part of speech and/or affixes). Sakhr has designed the Morphological Processor to produce word level analysis through regeneration and comparison[3].

In Sakhr morphological processor each regular derivative root is allowed to be combined with a selected set of forms or patterns to produce words that can be found in standard Arabic dictionaries. Sakhr did not publish any technical documents about its Arabic morphological analyzer; no one knows how its model of Arabic morphology looks like. (Attia , 2000).

**RDI Arabic Morphological Analyzer:** The main RDI's NLP core engine is the basis of Arabic morphological analysis, Arabic POS tagging, and Arabic Lexical Semantic Analysis. ArabMorpho is a morpheme-based lexical analyzer/synthesizer which distinguishes it from its vocabulary-based rivals and boosts its flexibility. After morphological rules are exhausted, deep-horizon dynamic statistical analysis is employed to realize disambiguation; hence, word accuracy can reach up to 96%4. In RDI analyzer each regular derivative root is allowed to combine freely with any form

[2] http://www.ldc.upenn.edu/Catalog/docs/LDC2004L02/readme.txt

[3] http://www.sakhr.com/Technology/Morphology/Default.aspx?sec=Technology&item=Morphology

[4] http://www.rdi-eg.com/rdi/technologies/arabic_nlp.htm

as long as this combination is morphologically allowed. This allows the system to deal with all the possible Arabic words and eradicates the need to be tied to a fixed vocabulary (Attia, 2000)[5].

## 5. The International Corpus of Arabic (ICA) “Analysis stage”:

Alansary et al. (2007) surveyed the compilation of ICA, its design and the preliminary software used in interrogating the compiled corpus. This attempt can be considered one of the most successful approaches for building a representative corpus for MSA. It is important to realize that the creation of ICA is a "cyclical" process, requiring constant re-evaluation as the corpus is being compiled. Once the process of collecting and computerizing texts is completed, texts will be ready for the final stage of preparation; mark up, from there, it is easy to deal with texts in the analysis stage.

The process of analyzing a corpus is in many respects similar to the process of creating a corpus. Like the compiler, the corpus analyst needs to consider some factors such as: whether the corpus to be analyzed is lengthy enough for the particular linguistic study being undertaken and whether the samples in the corpus are balanced and representative (Meyer, 2002).

This section is devoted to describing the process of analyzing the ICA corpus. It will focus on selecting and describing the model of analysis, pre-analysis stage (data processing), full text analysis stages, adding root information and current state of ICA.

### 5.1 Selecting and describing the model of analysis:

According to our 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. Therefore, the stem-based approach (concatenative approach) is adopted as a linguistic approach to analyze the ICA. 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. The Arabic Morphology module uses a simple approach of dividing the Arabic word into three parts:

*Prefix: consist of as many as three concatenated prefixes, or could be null.*
*Stem: it is composed of root and pattern morphemes.*
*Suffix: consist of as many as two concatenated suffixes, or could be null.*

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 prefix or a null suffix are both possible. Table 3 shows example of valid word forms:

[5] http://www.rdi-eg.com/rdi/Downloads/Scientific%20Papers/M_Atiyya_MScThesis2000.pdf

| Suffix | Stem | Prefix |
|---|---|---|
| xxx | كتاب | الـــ |
| ان | كتاب | xxx |
| ين | كتاب | والـــ |
| xxx | كتب | يـــ |
| xxx | كتب | xxx |
| ين | كتب | تـــ |

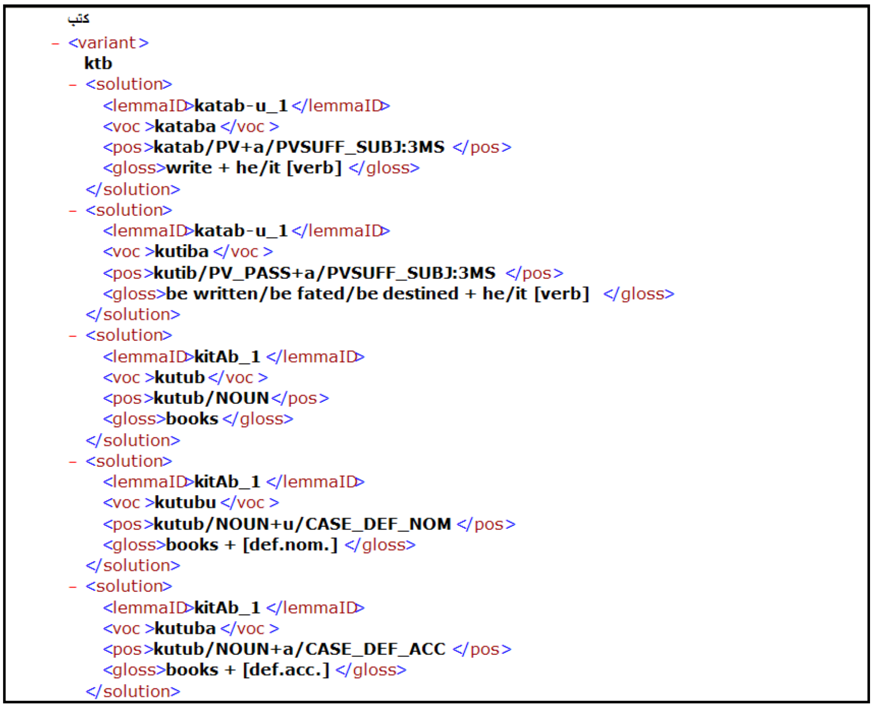
Table 3: valid word forms.

Not every Prefix-Stem-Suffix combination is necessarily a valid or a legal word. To confirm that the Prefix-Stem-Suffix composition is a valid Arabic word, morphological categories are assigned to each entry in the lexicons.

When trying to select the morphological analyzer system to be used in analyzing the ICA, Buckwalter morphological analyzer has been selected to analyze the ICA as it was found that to be the most suitable lexical resource to our approach.

The Buckwalter's morphological analyzer has many advantages such as its ability to provide a lot of information like Lemma, Vocalization, Part of Speech (POS) and Gloss. Also, Buckwalter is capable of supplying other information such as prefix(s), stem, word class, suffix(s), number, gender, definiteness and case. The output of Buckwalter appears in XML format.

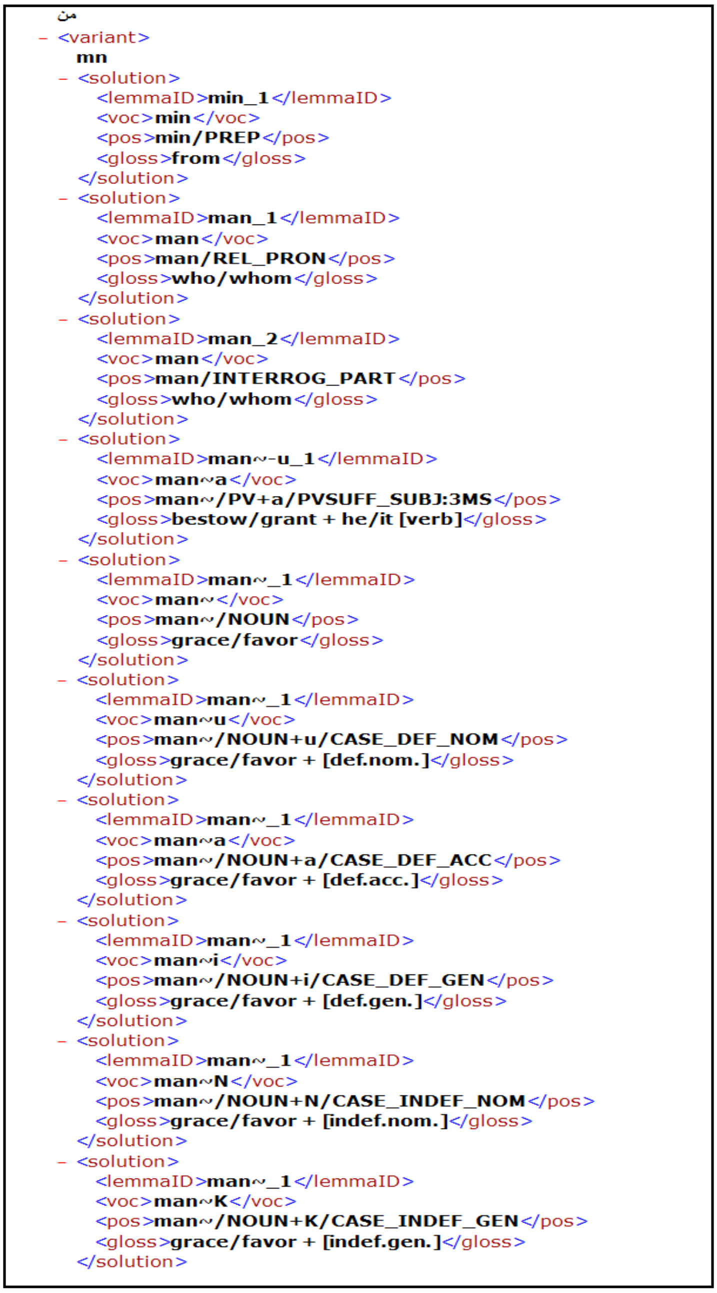
A single word may belong to more than one word class. For example the word "كتب" appears in Buckwalter output as noun or verb as shown in figure 1:



```
   كتب
 - <variant>
     ktb
   - <solution>
       <lemmaID>katab-u_1</lemmaID>
       <voc>kataba</voc>
       <pos>katab/PV+a/PVSUFF_SUBJ:3MS </pos>
       <gloss>write + he/it [verb] </gloss>
     </solution>
   - <solution>
       <lemmaID>katab-u_1</lemmaID>
       <voc>kutiba</voc>
       <pos>kutib/PV_PASS+a/PVSUFF_SUBJ:3MS </pos>
       <gloss>be written/be fated/be destined + he/it [verb] </gloss>
     </solution>
   - <solution>
       <lemmaID>kitAb_1 </lemmaID>
       <voc>kutub</voc>
       <pos>kutub/NOUN</pos>
       <gloss>books </gloss>
     </solution>
   - <solution>
       <lemmaID>kitAb_1 </lemmaID>
       <voc>kutubu</voc>
       <pos>kutub/NOUN+u/CASE_DEF_NOM </pos>
       <gloss>books + [def.nom.] </gloss>
     </solution>
   - <solution>
       <lemmaID>kitAb_1 </lemmaID>
       <voc>kutuba</voc>
       <pos>kutub/NOUN+a/CASE_DEF_ACC </pos>
       <gloss>books + [def.acc.] </gloss>
     </solution>
```

Figure 1: The word classes of "كتب"

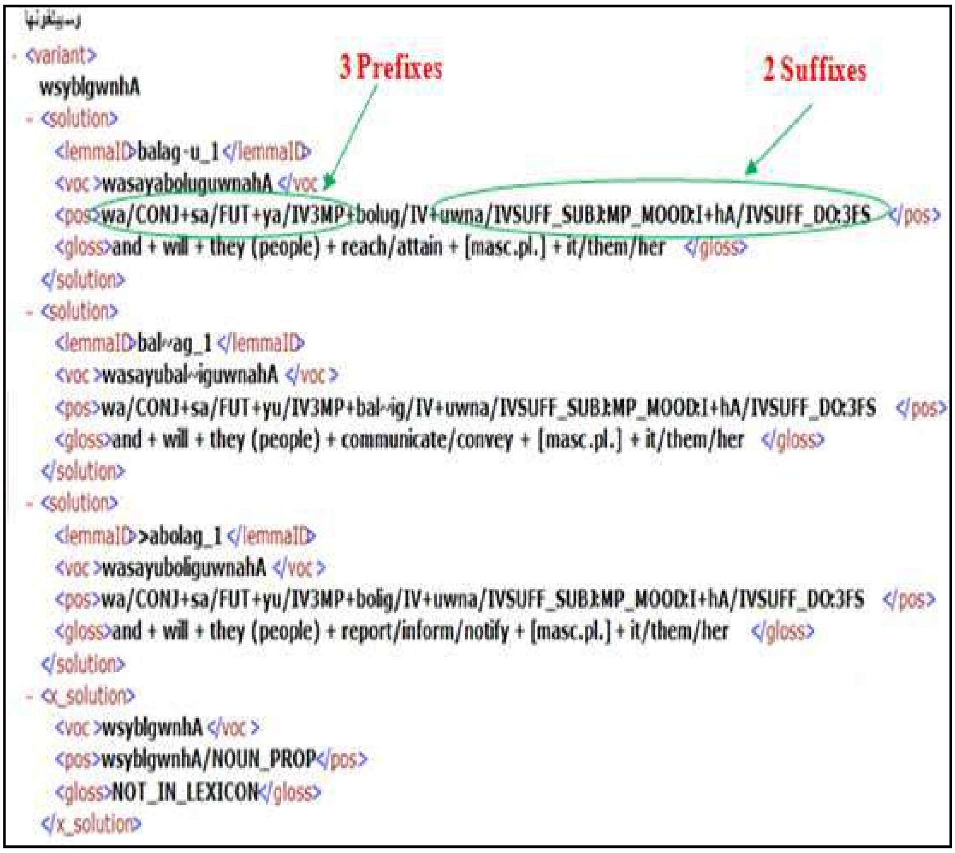
The word "من" appears in Buckwalter output as a Noun, verb, Preposition, Relative Pronoun or Interrogative part as shown in figure 2:



```
   من
 - <variant>
     mn
   - <solution>
       <lemmaID>min_1</lemmaID>
       <voc>min</voc>
       <pos>min/PREP</pos>
       <gloss>from</gloss>
     </solution>
   - <solution>
       <lemmaID>man_1</lemmaID>
       <voc>man</voc>
       <pos>man/REL_PRON</pos>
       <gloss>who/whom</gloss>
     </solution>
   - <solution>
       <lemmaID>man_2</lemmaID>
       <voc>man</voc>
       <pos>man/INTERROG_PART</pos>
       <gloss>who/whom</gloss>
     </solution>
   - <solution>
       <lemmaID>man~-u_1</lemmaID>
       <voc>man~a</voc>
       <pos>man~/PV+a/PVSUFF_SUBJ:3MS</pos>
       <gloss>bestow/grant + he/it [verb]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~</voc>
       <pos>man~/NOUN</pos>
       <gloss>grace/favor</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~u</voc>
       <pos>man~/NOUN+u/CASE_DEF_NOM</pos>
       <gloss>grace/favor + [def.nom.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~a</voc>
       <pos>man~/NOUN+a/CASE_DEF_ACC</pos>
       <gloss>grace/favor + [def.acc.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~i</voc>
       <pos>man~/NOUN+i/CASE_DEF_GEN</pos>
       <gloss>grace/favor + [def.gen.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~N</voc>
       <pos>man~/NOUN+N/CASE_INDEF_NOM</pos>
       <gloss>grace/favor + [indef.nom.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~K</voc>
       <pos>man~/NOUN+K/CASE_INDEF_GEN</pos>
       <gloss>grace/favor + [indef.gen.]</gloss>
     </solution>
```

Figure 2: The word classes of “من”.

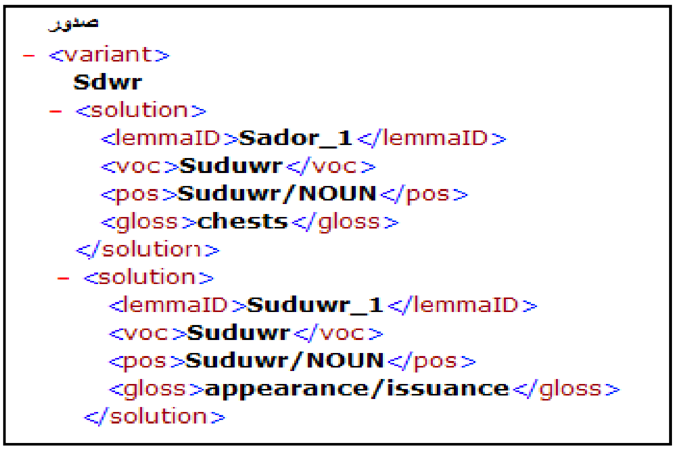
Buckwalter’s morphological analyzer can also determine the number of prefixes and suffixes in each word. For example the word **“وسيبلغونها”** has three prefixes and two suffixes as shown in figure 3:



```
وسيبلغونها
- <variant>
   wsyblgwnhA
  - <solution>
      <lemmaID>balag-u_1</lemmaID>
      <voc >wasayaboluguwnahA </voc
      <pos>wa/CONJ+sa/FUT+ya/IV3MP+bolug/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS </pos>
      <gloss>and + will + they (people) + reach/attain + [masc.pl.] + it/them/her  </gloss>
    </solution>
  - <solution>
      <lemmaID>bal~ag_1 </lemmaID>
      <voc >wasayubal~iguwnahA </voc >
      <pos>wa/CONJ+sa/FUT+yu/IV3MP+bal~ig/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS  </pos>
      <gloss>and + will + they (people) + communicate/convey + [masc.pl.] + it/them/her  </gloss>
    </solution>
  - <solution>
      <lemmaID>>abolag_1 </lemmaID>
      <voc >wasayuboliguwnahA </voc >
      <pos>wa/CONJ+sa/FUT+yu/IV3MP+bolig/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS  </pos>
      <gloss>and + will + they (people) + report/inform/notify + [masc.pl.] + it/them/her   </gloss>
    </solution>
  - <x_solution>
      <voc >wsyblgwnhA </voc >
      <pos>wsyblgwnhA/NOUN_PROP</pos>
      <gloss>NOT_IN_LEXICON</gloss>
    </x_solution>
```

Figure 3:The prefixes and suffixes of “وسيبلغونها”

Additionally, a single Arabic word may have more than one meaning according to its context. Buckwalter has the ability to indicate this feature by showing different glosses for the same word with the same word class. For example, the word **“صدور”** when classified as a noun it may have more than one gloss as shown in figure 4:



```
   صدور
 - <variant>
     Sdwr
   - <solution>
       <lemmaID>Sador_1</lemmaID>
       <voc>Suduwr</voc>
       <pos>Suduwr/NOUN</pos>
       <gloss>chests</gloss>
     </solution>
    - <solution>
        <lemmaID>Suduwr_1</lemmaID>
        <voc>Suduwr</voc>
        <pos>Suduwr/NOUN</pos>
        <gloss>appearance/issuance</gloss>
      </solution>
```

Figure 4: The prefixes and suffixes of “صدور”.

## 5.2 Pre-analysis stage:

The basic idea behind the rule-based approach to parts-of-speech tagging is to provide the analyzer software with three lexicons (a prefix lexicon, a stem lexicon and a suffix lexicon) and some sorts of internal grammar which use grammatical rules to disambiguate words.

Surely there must be some objective criteria that enable the analyst to decide to which class a word belongs in order to assign the part-of-speech class. Hence, if one word can be assigned to more than one class, this must be mentioned in the lexicon of the analysis system.

There is a number of general considerations to bear in mind when beginning the process of analyzing the ICA corpus. The pre-analysis stage is an important stage that includes:

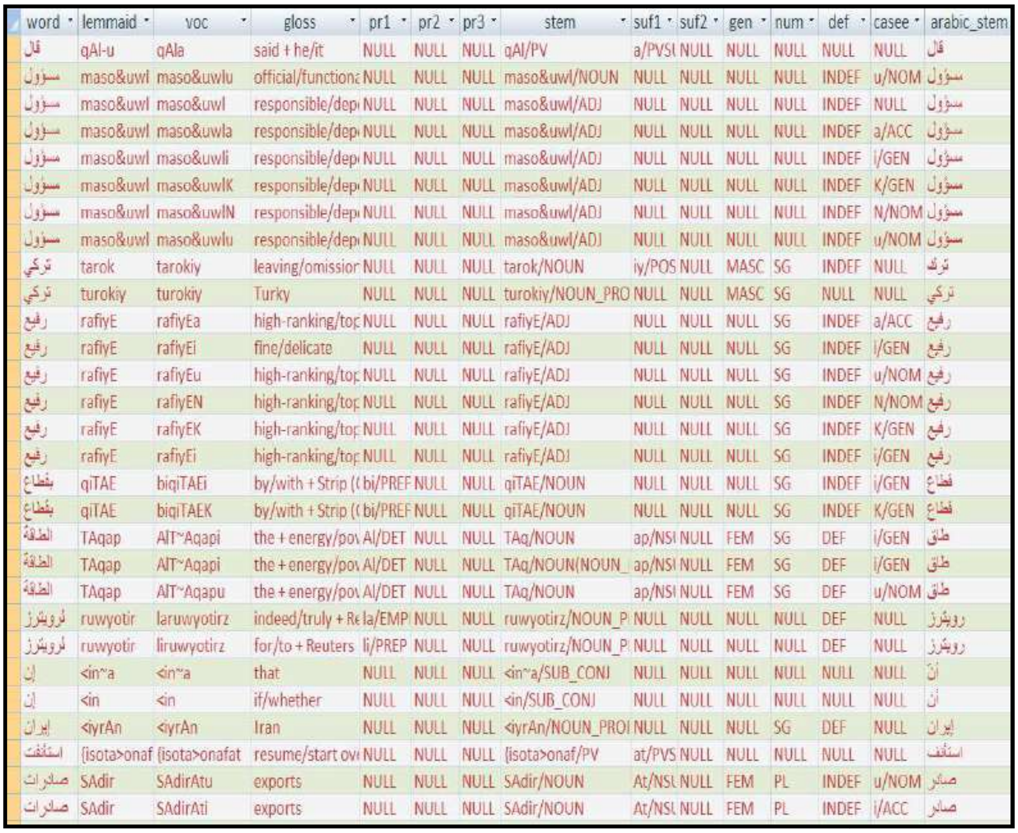
A. **Handling Buckwalter's output:** When dealing with texts and Buckwalter's output it was preferred to use a database format because it helps in capturing, editing and changing any part of the information easily. The conversion to database format caused a problem because Buckwalter's output is divided into three tables: A table for analyzed words with all possible solutions, a table for unanalyzed words that do not exist in the analyzer's lexicon and a third for punctuation marks found in the text being analyzed. However, this process results in the loss of the context of the text to be analyzed.

B. **Handling texts:** This stage includes transferring texts from 'plain text' horizontal format to database vertical format (from text to list). This process of handling texts helps in keeping the context of words in each text file to be analyzed in one hand, and enabling a list of features to be inserted horizontally besides each word in the list on the other hand.

C. **Mapping between Buckwalter's solutions and word list:** In this stage each word in the word list will be mapped with its suitable morphological solutions according to Buckwalter's output.

An interface has been used to map between Buckwalter's solutions and the word list. It leads to have a table containing 16 columns of information as follows: Word, Lemma, Vocalization, Gloss, Prefix1, Prefix2, Prefix3, Stem, word class, Suffix1, Suffix2, number, gender, definiteness, Arabic stem and case. Figure 5 shows the following:

- Each solution  appears in a separate row.
- Each solution has 16 types of information separated in an independent column.



| word | lemmaid | voc | gloss | pr1 | pr2 | pr3 | stem | suf1 | suf2 | gen | num | def | casee | arabic_stem |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| قال | qAl-u | qAla | said + he/it | NULL | NULL | NULL | qAl/PV | a/PVSU | NULL | NULL | NULL | NULL | NULL | قال |
| مسؤول | maso&uwl | maso&uwlu | official/functiona | NULL | NULL | NULL | maso&uwl/NOUN | NULL | NULL | NULL | NULL | INDEF | u/NOM | مسؤول |
| مسؤول | maso&uwl | maso&uwl | responsible/dep | NULL | NULL | NULL | maso&uwl/ADJ | NULL | NULL | NULL | NULL | INDEF | NULL | مسؤول |
| مسؤول | maso&uwl | maso&uwla | responsible/dep | NULL | NULL | NULL | maso&uwl/ADJ | NULL | NULL | NULL | NULL | INDEF | a/ACC | مسؤول |
| مسؤول | maso&uwl | maso&uwli | responsible/dep | NULL | NULL | NULL | maso&uwl/ADJ | NULL | NULL | NULL | NULL | INDEF | i/GEN | مسؤول |
| مسؤول | maso&uwl | maso&uwlK | responsible/dep | NULL | NULL | NULL | maso&uwl/ADJ | NULL | NULL | NULL | NULL | INDEF | K/GEN | مسؤول |
| مسؤول | maso&uwl | maso&uwlN | responsible/dep | NULL | NULL | NULL | maso&uwl/ADJ | NULL | NULL | NULL | NULL | INDEF | N/NOM | مسؤول |
| مسؤول | maso&uwl | maso&uwlu | responsible/dep | NULL | NULL | NULL | maso&uwl/ADJ | NULL | NULL | NULL | NULL | INDEF | u/NOM | مسؤول |
| تركي | tarok | tarokiy | leaving/omission | NULL | NULL | NULL | tarok/NOUN | iy/POS | NULL | MASC | SG | INDEF | NULL | ترك |
| تركي | turokiy | turokiy | Turky | NULL | NULL | NULL | turokiy/NOUN_PRO | NULL | NULL | MASC | SG | NULL | NULL | تركي |
| رفيع | rafiyE | rafiyEa | high-ranking/top | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | NULL | SG | INDEF | a/ACC | رفيع |
| رفيع | rafiyE | rafiyEi | fine/delicate | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | NULL | SG | INDEF | i/GEN | رفيع |
| رفيع | rafiyE | rafiyEu | high-ranking/top | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | NULL | SG | INDEF | u/NOM | رفيع |
| رفيع | rafiyE | rafiyEN | high-ranking/top | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | NULL | SG | INDEF | N/NOM | رفيع |
| رفيع | rafiyE | rafiyEK | high-ranking/top | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | NULL | SG | INDEF | K/GEN | رفيع |
| رفيع | rafiyE | rafiyEi | high-ranking/top | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | NULL | SG | INDEF | i/GEN | رفيع |
| بقطاع | qiTAE | biqiTAEi | by/with + Strip (( | bi/PREF | NULL | NULL | qiTAE/NOUN | NULL | NULL | NULL | SG | INDEF | i/GEN | قطاع |
| بقطاع | qiTAE | biqiTAEK | by/with + Strip (( | bi/PREF | NULL | NULL | qiTAE/NOUN | NULL | NULL | NULL | SG | INDEF | K/GEN | قطاع |
| الطاقة | TAqap | AlT~Aqapi | the + energy/pov | Al/DET | NULL | NULL | TAq/NOUN | ap/NSI | NULL | FEM | SG | DEF | i/GEN | طاق |
| الطاقة | TAqap | AlT~Aqapi | the + energy/pov | Al/DET | NULL | NULL | TAq/NOUN(NOUN_ | ap/NSI | NULL | FEM | SG | DEF | i/GEN | طاق |
| الطاقة | TAqap | AlT~Aqapu | the + energy/pov | Al/DET | NULL | NULL | TAq/NOUN | ap/NSI | NULL | FEM | SG | DEF | u/NOM | طاق |
| لرويترز | ruwyotir | laruwyotirz | indeed/truly + Re | la/EMP | NULL | NULL | ruwyotirz/NOUN_P | NULL | NULL | NULL | NULL | DEF | NULL | رويترز |
| لرويترز | ruwyotir | liruwyotirz | for/to + Reuters | li/PREP | NULL | NULL | ruwyotirz/NOUN_P | NULL | NULL | NULL | NULL | DEF | NULL | رويترز |
| إنّ | <in~a | <in~a | that | NULL | NULL | NULL | <in~a/SUB_CONJ | NULL | NULL | NULL | NULL | NULL | NULL | أنّ |
| إن | <in | <in | if/whether | NULL | NULL | NULL | <in/SUB_CONJ | NULL | NULL | NULL | NULL | NULL | NULL | أن |
| إيران | <iyrAn | <iyrAn | Iran | NULL | NULL | NULL | <iyrAn/NOUN_PRO | NULL | NULL | NULL | SG | DEF | NULL | إيران |
| استأنفت | {isota>onaf | {isota>onafat | resume/start ove | NULL | NULL | NULL | {isota>onaf/PV | at/PVS | NULL | NULL | NULL | NULL | NULL | استأنف |
| صادرات | SAdir | SAdirAtu | exports | NULL | NULL | NULL | SAdir/NOUN | At/NSU | NULL | FEM | PL | INDEF | u/NOM | صادر |
| صادرات | SAdir | SAdirAti | exports | NULL | NULL | NULL | SAdir/NOUN | At/NSU | NULL | FEM | PL | INDEF | i/ACC | صادر |

Figure 5: The database after mapping word list with Buckwalter's solutions.

## 5.3 Full text analysis stages:

The full text analysis stage includes: disambiguation of words that may have multiple solutions, modifying and adding extra linguistic information and manual analysis of unanalyzed words.

### 5.3.1 Disambiguating words:

The suitable analysis for each word is chosen according to its context. An interface is used to select the correct analysis solution. Figure 6 shows an example of disambiguating the word "كتب".

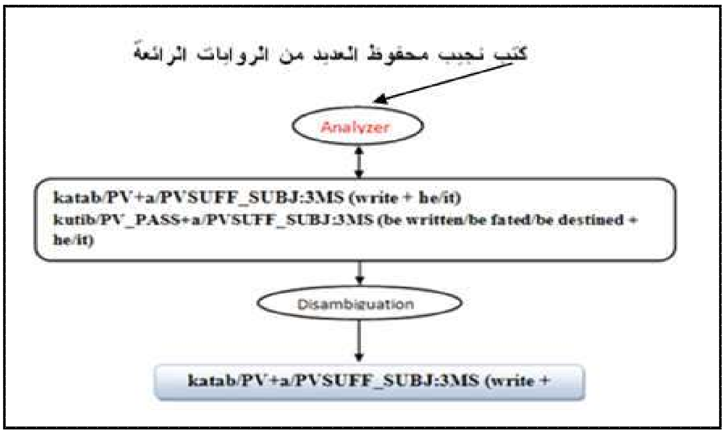
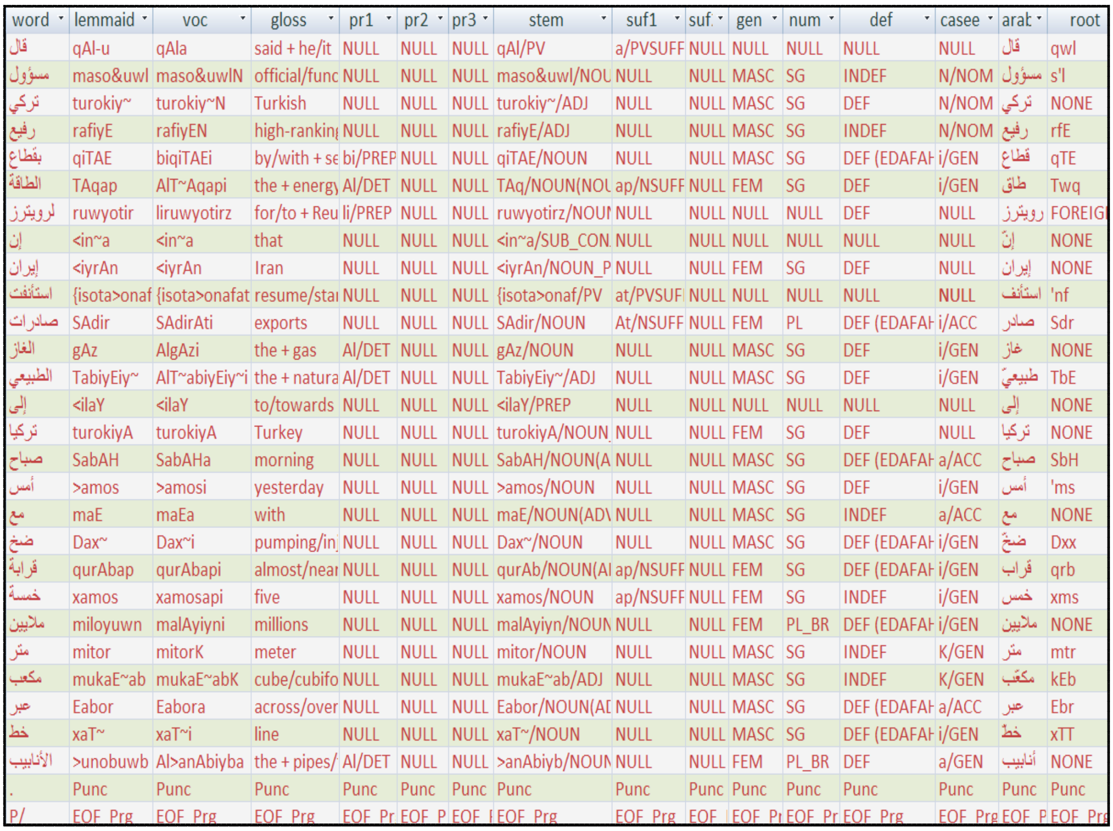

كتب نجيب محفوظ العديد من الروايات الرائعة
Analyzer
katab/PV+a/PVSUFF_SUBJ:3MS (write + he/it)
kutib/PV_PASS+a/PVSUFF_SUBJ:3MS (be written/be fated/be destined + he/it)
Disambiguation
katab/PV+a/PVSUFF_SUBJ:3MS (write +

Figure 6: An example of the disambiguation process.

Figure 7 shows one text after it was disambiguated:



| word | lemmaid | voc | gloss | pr1 | pr2 | pr3 | stem | suf1 | suf | gen | num | def | casee | arab | root |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| قال | qAl-u | qAla | said + he/it | NULL | NULL | NULL | qAl/PV | a/PVSUFF | NULL | NULL | NULL | NULL | NULL | قال | qwl |
| مسؤول | maso&uwl | maso&uwlN | official/func | NULL | NULL | NULL | maso&uwl/NOU | NULL | NULL | MASC | SG | INDEF | N/NOM | مسؤول | s'l |
| تركي | turokiy~ | turokiy~N | Turkish | NULL | NULL | NULL | turokiy~/ADJ | NULL | NULL | MASC | SG | DEF | N/NOM | تركي | NONE |
| رفيع | rafiyE | rafiyEN | high-rankin | NULL | NULL | NULL | rafiyE/ADJ | NULL | NULL | MASC | SG | INDEF | N/NOM | رفيع | rfE |
| بقطاع | qiTAE | biqiTAEi | by/with + se | bi/PREP | NULL | NULL | qiTAE/NOUN | NULL | NULL | MASC | SG | DEF (EDAFAH | i/GEN | قطاع | qTE |
| الطاقة | TAqap | AlT~Aqapi | the + energy | Al/DET | NULL | NULL | TAq/NOUN(NOL | ap/NSUFF | NULL | FEM | SG | DEF | i/GEN | طاق | Twq |
| لرويترز | ruwyotir | liruwyotirz | for/to + Reu | li/PREP | NULL | NULL | ruwyotirz/NOUI | NULL | NULL | NULL | NULL | DEF | NULL | رويترز | FOREIG |
| إن | <in~a | <in~a | that | NULL | NULL | NULL | <in~a/SUB_CON | NULL | NULL | NULL | NULL | NULL | NULL | إنّ | NONE |
| إيران | <iyrAn | <iyrAn | Iran | NULL | NULL | NULL | <iyrAn/NOUN_P | NULL | NULL | FEM | SG | DEF | NULL | إيران | NONE |
| استأنفت | {isota>onaf | {isota>onafat | resume/sta | NULL | NULL | NULL | {isota>onaf/PV | at/PVSUF | NULL | NULL | NULL | NULL | NULL | استأنف | 'nf |
| صادرات | SAdir | SAdirAti | exports | NULL | NULL | NULL | SAdir/NOUN | At/NSUFF | NULL | FEM | PL | DEF (EDAFAH | i/ACC | صادر | Sdr |
| الغاز | gAz | AlgAzi | the + gas | Al/DET | NULL | NULL | gAz/NOUN | NULL | NULL | MASC | SG | DEF | i/GEN | غاز | NONE |
| الطبيعي | TabiyEiy~ | AlT~abiyEiy~i | the + natura | Al/DET | NULL | NULL | TabiyEiy~/ADJ | NULL | NULL | MASC | SG | DEF | i/GEN | طبيعيّ | TbE |
| إلى | <ilaY | <ilaY | to/towards | NULL | NULL | NULL | <ilaY/PREP | NULL | NULL | NULL | NULL | NULL | NULL | إلى | NONE |
| تركيا | turokiyA | turokiyA | Turkey | NULL | NULL | NULL | turokiyA/NOUN | NULL | NULL | FEM | SG | DEF | NULL | تركيا | NONE |
| صباح | SabAH | SabAHa | morning | NULL | NULL | NULL | SabAH/NOUN(A | NULL | NULL | MASC | SG | DEF (EDAFAH | a/ACC | صباح | SbH |
| أمس | >amos | >amosi | yesterday | NULL | NULL | NULL | >amos/NOUN | NULL | NULL | MASC | SG | DEF | i/GEN | أمس | 'ms |
| مع | maE | maEa | with | NULL | NULL | NULL | maE/NOUN(AD\ | NULL | NULL | MASC | SG | INDEF | a/ACC | مع | NONE |
| ضخ | Dax~ | Dax~i | pumping/in | NULL | NULL | NULL | Dax~/NOUN | NULL | NULL | MASC | SG | DEF (EDAFAH | i/GEN | ضخّ | Dxx |
| قرابة | qurAbap | qurAbapi | almost/near | NULL | NULL | NULL | qurAb/NOUN(Al | ap/NSUFF | NULL | FEM | SG | DEF (EDAFAH | i/GEN | قراب | qrb |
| خمسة | xamos | xamosapi | five | NULL | NULL | NULL | xamos/NOUN | ap/NSUFF | NULL | FEM | SG | INDEF | i/GEN | خمس | xms |
| ملايين | miloyuwn | malAyiyni | millions | NULL | NULL | NULL | malAyiyn/NOUN | NULL | NULL | FEM | PL_BR | DEF (EDAFAH | i/GEN | ملايين | NONE |
| متر | mitor | mitorK | meter | NULL | NULL | NULL | mitor/NOUN | NULL | NULL | MASC | SG | INDEF | K/GEN | متر | mtr |
| مكعب | mukaE~ab | mukaE~abK | cube/cubifo | NULL | NULL | NULL | mukaE~ab/ADJ | NULL | NULL | MASC | SG | INDEF | K/GEN | مكعّب | kEb |
| عبر | Eabor | Eabora | across/over | NULL | NULL | NULL | Eabor/NOUN(AI | NULL | NULL | MASC | SG | DEF (EDAFAH | a/ACC | عبر | Ebr |
| خط | xaT~ | xaT~i | line | NULL | NULL | NULL | xaT~/NOUN | NULL | NULL | MASC | SG | DEF (EDAFAH | i/GEN | خطّ | xTT |
| الأنابيب | >unobuwb | Al>anAbiyba | the + pipes/ | Al/DET | NULL | NULL | >anAbiyb/NOUN | NULL | NULL | FEM | PL_BR | DEF | a/GEN | أنابيب | NONE |
| . | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc |
| P/ | EOF_Prg | EOF_Prg | EOF_Prg | EOF_Pr | EOF_P | EOF_ | EOF_Prg | EOF_Prg | EOF_ | EOF_Pr | EOF_Pr | EOF_Prg | EOF_Prg | EOF_P | EOF_Prg |

Figure 7: One of disambiguated texts.

### 5.3.2 Modifying and adding some linguistic information:

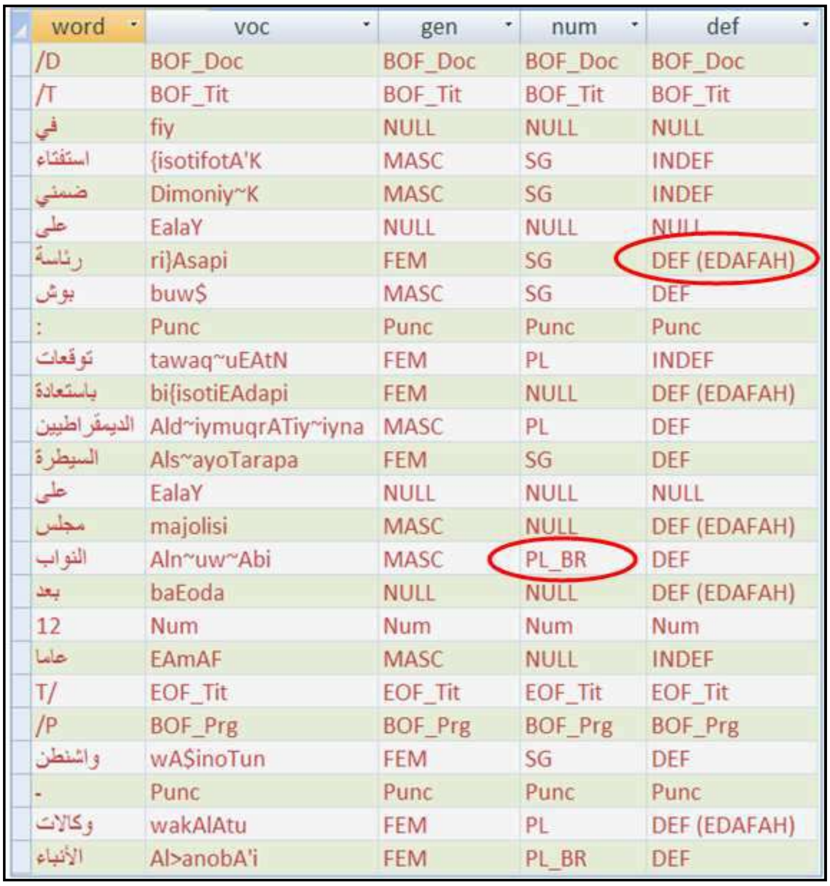
Some information in the output of Buckwalter's analyzer such as number, gender and definiteness needed modifications according to their morphosyntactic properties. These features can be explained as follows:

- ***Gender:*** Buckwalter's analyzer does not identify the gender of Arabic words in two case. The first, if a masculine word or a broken plural ends in "ة" e.g. "أسامة" and "أساتذة", it considers both of them as feminine. The second, if a feminine word or a broken plural does not end in "ة" e.g. "نساء ، أملاك ، أبواب ،......", the analyzer does not identify the gender and assigns "NULL" to the words under identification. In both cases, a manual intervention is used to fix the gender.

- ***Number:*** It has been noted that Buckwalter's analyzer has a problem with broken plurals; it deals with some of these words as singular, e.g. "أبخرة ، أحذية", and deals with others by assigning them (NULL), e.g. "أبواب ، أحوال ، أنحاء". This type of plural is given "PL_BR" for number manually. In addition all other nouns that do not end in any morpheme the denotes gender e.g. ,"أسمنت ، أبلغ ، أكبر", have been assigned "NULL". All number problems have been fixed manually.

- ***Definiteness:*** Buckwalter could detect the suitable definiteness for most words, however, there are some indefinite words that Buckwalter identified as definite words such as “التفاف ، التحاق ، التزام”, these words have been modified to be indefinite. In addition, the analyst added a new value for the feature of definiteness (DEF_EDAFAH), e.g. as in “مهاراته”, in order to make the feature o definiteness more expressive. " مهاراته .

Figure 8 shows the new modifications for Gender, Number and Definiteness according to their contexts:



| word | voc | gen | num | def |
|---|---|---|---|---|
| /D | BOF_Doc | BOF_Doc | BOF_Doc | BOF_Doc |
| /T | BOF_Tit | BOF_Tit | BOF_Tit | BOF_Tit |
| في | fiy | NULL | NULL | NULL |
| استفتاء | {isotifotA'K | MASC | SG | INDEF |
| ضمني | Dimoniy~K | MASC | SG | INDEF |
| على | EalaY | NULL | NULL | NULL |
| رئاسة | ri}Asapi | FEM | SG | DEF (EDAFAH) |
| بوش | buw$ | MASC | SG | DEF |
| : | Punc | Punc | Punc | Punc |
| توقعات | tawaq~uEAtN | FEM | PL | INDEF |
| باستعادة | bi{isotiEAdapi | FEM | NULL | DEF (EDAFAH) |
| الديمقراطيين | Ald~iymuqrATiy~iyna | MASC | PL | DEF |
| السيطرة | Als~ayoTarapa | FEM | SG | DEF |
| على | EalaY | NULL | NULL | NULL |
| مجلس | majolisi | MASC | NULL | DEF (EDAFAH) |
| النواب | Aln~uw~Abi | MASC | PL_BR | DEF |
| بعد | baEoda | NULL | NULL | DEF (EDAFAH) |
| 12 | Num | Num | Num | Num |
| عاما | EAmAF | MASC | NULL | INDEF |
| T/ | EOF_Tit | EOF_Tit | EOF_Tit | EOF_Tit |
| /P | BOF_Prg | BOF_Prg | BOF_Prg | BOF_Prg |
| واشنطن | wA$inoTun | FEM | SG | DEF |
| - | Punc | Punc | Punc | Punc |
| وكالات | wakAlAtu | FEM | PL | DEF (EDAFAH) |
| الأنباء | Al>anobA'i | FEM | PL_BR | DEF |

Figure 8 : Gender, Number and Definiteness.

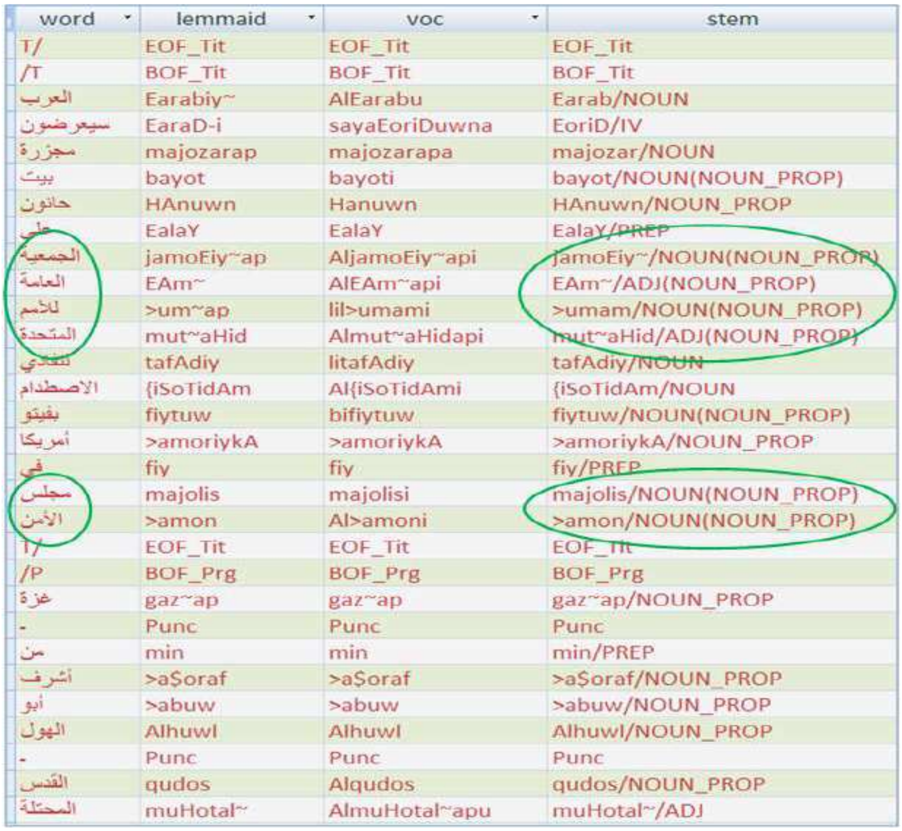
In order to make the morphological analysis more expressive, we have seen that the following extra information that exceed the scope of Buckwalter’s analyzer should be added:

A. ***Name entities:*** name entities are words that represent the title of an institute, ministry, association, compound name of a country, book, film, company or conference. Analysts identified these names by adding the feature (NOUN_PROP) right after the basic word class of these words. For example "الولايات المتحدة الأميركية" appears in analysis as shown in table 4:

| Word | Word Class |
|---|---|
| الولايات | NOUN(NOUN_PROP) |
| المتحدة | ADJ(NOUN_PROP) |
| الأميركية | ADJ(NOUN_PROP) |

Table 4 : An example of a name entity.

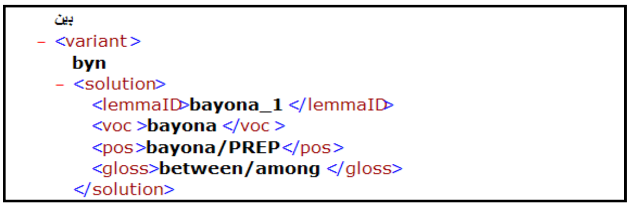
By adding the name entity feature, researchers can capture name entities easily in addition to capturing the word with respect to the part of speech. Figure 9 shows some examples of name entities within their contexts:



| word | lemmaid | voc | stem |
|---|---|---|---|
| T/ | EOF_Tit | EOF_Tit | EOF_Tit |
| /T | BOF_Tit | BOF_Tit | BOF_Tit |
| العرب | Earabiy~ | AlEarabu | Earab/NOUN |
| سيعرضون | EaraD-i | sayaEoriDuwna | EoriD/IV |
| مجزرة | majozarap | majozarapa | majozar/NOUN |
| بيت | bayot | bayoti | bayot/NOUN(NOUN_PROP) |
| حانون | HAnuwn | Hanuwn | HAnuwn/NOUN_PROP |
| على | EalaY | EalaY | EalaY/PREP |
| الجمعية | jamoEiy~ap | AljamoEiy~api | jamoEiy~/NOUN(NOUN_PROP) |
| العامة | EAm~ | AlEAm~api | EAm~/ADJ(NOUN_PROP) |
| للأمم | >um~ap | lil>umami | >umam/NOUN(NOUN_PROP) |
| المتحدة | mut~aHid | Almut~aHidapi | mut~aHid/ADJ(NOUN_PROP) |
| لتفادي | tafAdiy | litafAdiy | tafAdiy/NOUN |
| الاصطدام | {iSoTidAm | Al{iSoTidAmi | {iSoTidAm/NOUN |
| بفيتو | fiytuw | bifiytuw | fiytuw/NOUN(NOUN_PROP) |
| أمريكا | >amoriykA | >amoriykA | >amoriykA/NOUN_PROP |
| في | fiy | fiy | fiy/PREP |
| مجلس | majolis | majolisi | majolis/NOUN(NOUN_PROP) |
| الأمن | >amon | Al>amoni | >amon/NOUN(NOUN_PROP) |
| T/ | EOF_Tit | EOF_Tit | EOF_Tit |
| /P | BOF_Prg | BOF_Prg | BOF_Prg |
| غزة | gaz~ap | gaz~ap | gaz~ap/NOUN_PROP |
| - | Punc | Punc | Punc |
| من | min | min | min/PREP |
| أشرف | >a$oraf | >a$oraf | >a$oraf/NOUN_PROP |
| أبو | >abuw | >abuw | >abuw/NOUN_PROP |
| الهول | Alhuwl | Alhuwl | Alhuwl/NOUN_PROP |
| - | Punc | Punc | Punc |
| القدس | qudos | Alqudos | qudos/NOUN_PROP |
| المحتلة | muHotal~ | AlmuHotal~apu | muHotal~/ADJ |

Figure 9: Some name entities according to context.

B. One of the disadvantages of the Buckwalter's morphological analyzer is that it determines the word class of Arabic words according to their counterparts in English. For example, Buckwalter's has classified some adverbs in Arabic as prepositions. Figure 10 shows Buckwalter's analysis of "بين" which should be analyzed as an adverb.



```
  بين
- <variant>
    byn
  - <solution>
      <lemmaID>bayona_1 </lemmaID>
      <voc >bayona </voc >
      <pos>bayona/PREP</pos>
      <gloss>between/among </gloss>
    </solution>
```

Figure 10: The word "بين" as preposition.

According to Buckwalter's analysis of adverbs (figure 10), four observations can be noticed. First**,** the word "بين" should be analyzed as an adverb; it can be used to describe either a place, as in "بين الأشجار", or a time as in "بين الساعة الخامسة والخامسة والنصف". Second**,** Some adverbs are nominalized (no longer adverbs) if they occur after a preposition; in this case their case is genitive as shown in example (1):

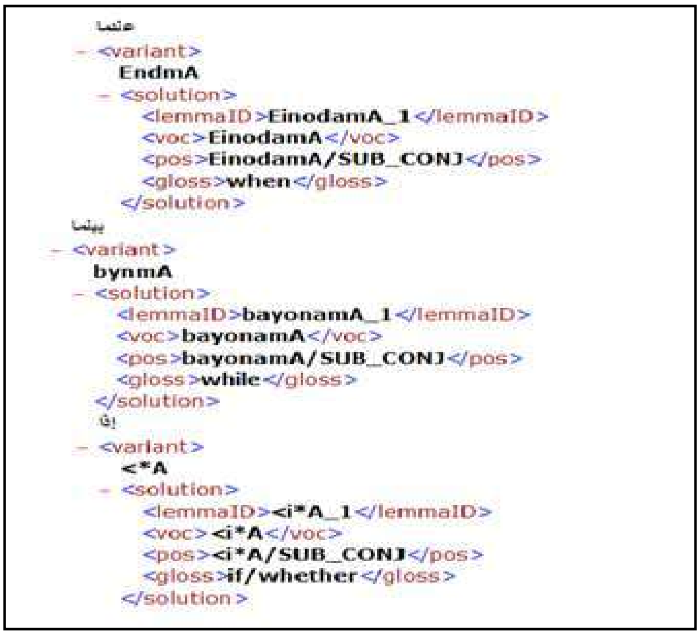
(1)

"ما زال تنظيم الأسرة من **بين** التحديات التي تواجه المجتمع"

↓

*(bayon/NOUN+i/CASE_DEF_GEN)*

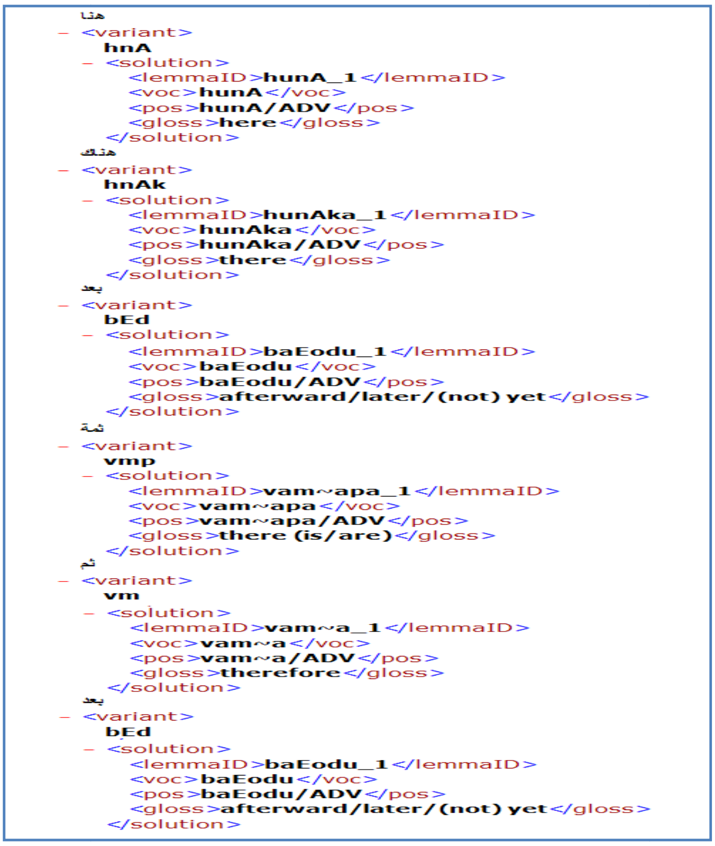
However, when Buckwalter's analyzer dealt with "بين" as a noun it gave out three possible cases, namely: nominative, accusative, and genitive (u/NOM, a/ACC, i/GEN, N/NOM and K/GEN), which is not correct. Third**,** Buckwalter's analyzer mistakenly analyzed some adverbs not only as prepositions but also as sub conjunctions (SUB_CONJ) as shown in figure 11.



```
     عندما
  - <variant>
      EndmA
    - <solution>
        <lemmaID>EinodamA_1</lemmaID>
        <voc>EinodamA</voc>
        <pos>EinodamA/SUB_CONJ</pos>
        <gloss>when</gloss>
      </solution>
  بينما
- <variant>
    bynmA
  - <solution>
      <lemmaID>bayonamA_1</lemmaID>
      <voc>bayonamA</voc>
      <pos>bayonamA/SUB_CONJ</pos>
      <gloss>while</gloss>
    </solution>
     إذا
  - <variant>
      <*A
    - <solution>
        <lemmaID><i*A_1</lemmaID>
        <voc><i*A</voc>
        <pos><i*A/SUB_CONJ</pos>
        <gloss>if/whether</gloss>
      </solution>
```

Figure 11: Example of Buckwalter output.

Forth**,** adverbs in Arabic are tagged with respect to two classes: adverbs which describe time (ADV_T) and adverbs which describe place (ADV_P). The same adverb may describe both time and place in different contexts. Buckwalter's analyzer can analyze some words as adverbs without determining the manner of that adverb (time or place) as shown in figure 12.



```
  هنا
– <variant>
    hnA
  – <solution>
      <lemmaID>hunA_1</lemmaID>
      <voc>hunA</voc>
      <pos>hunA/ADV</pos>
      <gloss>here</gloss>
    </solution>
  هناك
– <variant>
    hnAk
  – <solution>
      <lemmaID>hunAka_1</lemmaID>
      <voc>hunAka</voc>
      <pos>hunAka/ADV</pos>
      <gloss>there</gloss>
    </solution>
  بعد
– <variant>
    bEd
  – <solution>
      <lemmaID>baEodu_1</lemmaID>
      <voc>baEodu</voc>
      <pos>baEodu/ADV</pos>
      <gloss>afterward/later/(not) yet</gloss>
    </solution>
  ثمة
– <variant>
    vmp
  – <solution>
      <lemmaID>vam~apa_1</lemmaID>
      <voc>vam~apa</voc>
      <pos>vam~apa/ADV</pos>
      <gloss>there (is/are)</gloss>
    </solution>
  ثم
– <variant>
    vm
  – <solution>
      <lemmaID>vam~a_1</lemmaID>
      <voc>vam~a</voc>
      <pos>vam~a/ADV</pos>
      <gloss>therefore</gloss>
    </solution>
  بعد
– <variant>
    bEd
  – <solution>
      <lemmaID>baEodu_1</lemmaID>
      <voc>baEodu</voc>
      <pos>baEodu/ADV</pos>
      <gloss>afterward/later/(not) yet</gloss>
    </solution>
```

Figure 12: Buckwalter Adverbs analysis.

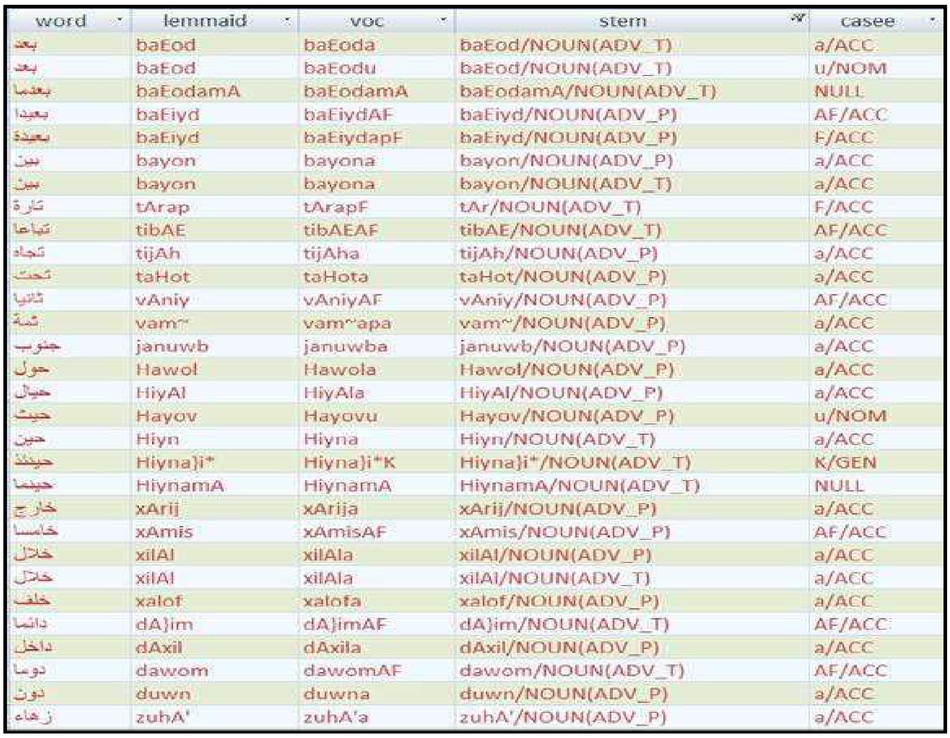
In retagging adverbs two criteria have been taken into account:

1. Separating the case tag from the stem; when Buckwalter analyzes the adverbs it considers the case as a part of the stem and consequently a part of lamma; for example, the stem of "هناك" is (hunAka/ADV) and the lemma is “hunAka”. So the case should be separated from stem and lemma.
2. In Arabic adverbs are nouns. Accordingly this has been tagged to every adverb. Consequently, the analysis of adverbs should contain three pieces of information: noun, adverb and time or place (T/P) as table 5 shows.

| Word | Buckwalter analysis | New analysis | Example |
|---|---|---|---|
| عند | Einoda/PREP | Einod/NOUN(ADV_T)<br>Einod/NOUN(ADV_P) | يرجى الاتصال **عند** حدوث أي مشكلة.<br>يلزم بناء سد **عند** مدخل الفيوم. |
| بعد | baEoda/PREP | baEod/NOUN(ADV_T)<br>baEod/NOUN(ADV_P) | سيتم تشغيلها **بعد** الحصول على الترخيص.<br>الشريك التجاري الثاني **بعد** تركيا. |
| بين | bayona/PREP | bayon/NOUN(ADV_T)<br>bayon/NOUN(ADV_P) | الفترة ما **بين** العامين الماضيين خلت من التطور.<br>إن التنسيق **بين** مصر وسوريا منتظم. |
| أمام | >amAma/PREP | >amAm/NOUN(ADV_P) | إننا **أمام** قضية خطيرة. |
| عبر | Eabora/PREP | Eabor/NOUN(ADV_P) | تم إرسال البيانات **عبر** شبكة المعلومات. |
| قبل | qabola/PREP | qabol/NOUN(ADV_T) | المبادرة التي اتخذها **قبل** بضعة أشهر. |
| فور | fawora/PREP | fawor/NOUN(ADV_T) | ستعود إلى القاهرة **فور** انتهاء أعمالها. |

Table 5: Example for adverbs.

Figure 13 shows the analysis of some adverbs which have been found in the ICA analyzed corpus:



| word | lemmaid | voc | stem | casee |
|---|---|---|---|---|
| بعد | baEod | baEoda | baEod/NOUN(ADV_T) | a/ACC |
| بعد | baEod | baEodu | baEod/NOUN(ADV_T) | u/NOM |
| بعدما | baEodamA | baEodamA | baEodamA/NOUN(ADV_T) | NULL |
| بعيدا | baEiyd | baEiydAF | baEiyd/NOUN(ADV_P) | AF/ACC |
| بعيدة | baEiyd | baEiydapF | baEiyd/NOUN(ADV_P) | F/ACC |
| بين | bayon | bayona | bayon/NOUN(ADV_P) | a/ACC |
| بين | bayon | bayona | bayon/NOUN(ADV_T) | a/ACC |
| تارة | tArap | tArapF | tAr/NOUN(ADV_T) | F/ACC |
| تباعا | tibAE | tibAEAF | tibAE/NOUN(ADV_T) | AF/ACC |
| تجاه | tijAh | tijAha | tijAh/NOUN(ADV_P) | a/ACC |
| تحت | taHot | taHota | taHot/NOUN(ADV_P) | a/ACC |
| ثانيا | vAniy | vAniyAF | vAniy/NOUN(ADV_P) | AF/ACC |
| ثمة | vam~ | vam~apa | vam~/NOUN(ADV_P) | a/ACC |
| جنوب | januwb | januwba | januwb/NOUN(ADV_P) | a/ACC |
| حول | Hawol | Hawola | Hawol/NOUN(ADV_P) | a/ACC |
| حيال | HiyAl | HiyAla | HiyAl/NOUN(ADV_P) | a/ACC |
| حيث | Hayov | Hayovu | Hayov/NOUN(ADV_P) | u/NOM |
| حين | Hiyn | Hiyna | Hiyn/NOUN(ADV_T) | a/ACC |
| حينئذ | Hiyna}i* | Hiyna}i*K | Hiyna}i*/NOUN(ADV_T) | K/GEN |
| حينما | HiynamA | HiynamA | HiynamA/NOUN(ADV_T) | NULL |
| خارج | xArij | xArija | xArij/NOUN(ADV_P) | a/ACC |
| خامسا | xAmis | xAmisAF | xAmis/NOUN(ADV_P) | AF/ACC |
| خلال | xilAl | xilAla | xilAl/NOUN(ADV_P) | a/ACC |
| خلال | xilAl | xilAla | xilAl/NOUN(ADV_T) | a/ACC |
| خلف | xalof | xalofa | xalof/NOUN(ADV_P) | a/ACC |
| دائما | dA}im | dA}imAF | dA}im/NOUN(ADV_T) | AF/ACC |
| داخل | dAxil | dAxila | dAxil/NOUN(ADV_P) | a/ACC |
| دوما | dawom | dawomAF | dawom/NOUN(ADV_T) | AF/ACC |
| دون | duwn | duwna | duwn/NOUN(ADV_P) | a/ACC |
| زهاء | zuhA' | zuhA'a | zuhA'/NOUN(ADV_P) | a/ACC |

Figure 13: Some adverbs in the ICA analyzed corpus.

***NOUN(ADV_M):*** This type of adverbs needs the context to be detected, but Buckwalter's did not identify this type of adverbs As shown in example (2):

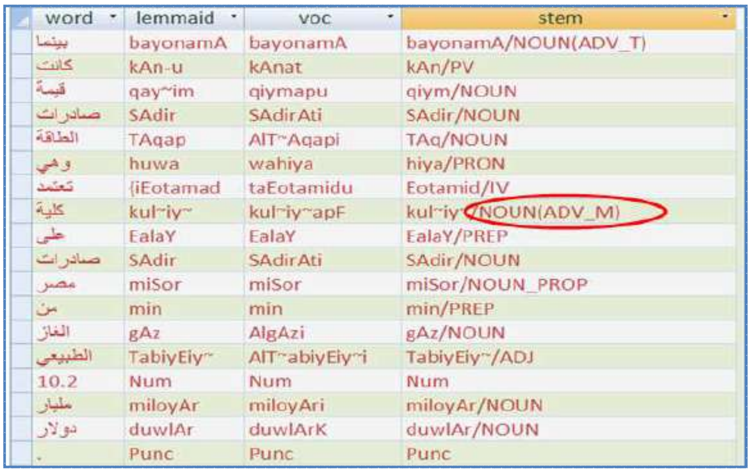
(2)

جاء الولد مسرعا

↓

**NOUN(ADV_M)**

Figure 14 shows an example of NOUN(ADV_M) within its context:



| word | lemmaid | voc | stem |
|---|---|---|---|
| بينما | bayonamA | bayonamA | bayonamA/NOUN(ADV_T) |
| كانت | kAn-u | kAnat | kAn/PV |
| قيمة | qay~im | qiymapu | qiym/NOUN |
| صادرات | SAdir | SAdirAti | SAdir/NOUN |
| الطاقة | TAqap | AlT~Aqapi | TAq/NOUN |
| وهي | huwa | wahiya | hiya/PRON |
| تعتمد | {iEotamad | taEotamidu | Eotamid/IV |
| كلية | kul~iy~ | kul~iy~apF | kul~iy~/NOUN(ADV_M) |
| على | EalaY | EalaY | EalaY/PREP |
| صادرات | SAdir | SAdirAti | SAdir/NOUN |
| مصر | miSor | miSor | miSor/NOUN_PROP |
| من | min | min | min/PREP |
| الغاز | gAz | AlgAzi | gAz/NOUN |
| الطبيعي | TabiyEiy~ | AlT~abiyEiy~i | TabiyEiy~/ADJ |
| 10.2 | Num | Num | Num |
| مليار | miloyAr | miloyAri | miloyAr/NOUN |
| دولار | duwlAr | duwlArK | duwlAr/NOUN |
| . | Punc | Punc | Punc |

Figure 14: An example of NOUN(ADV_M) within context.

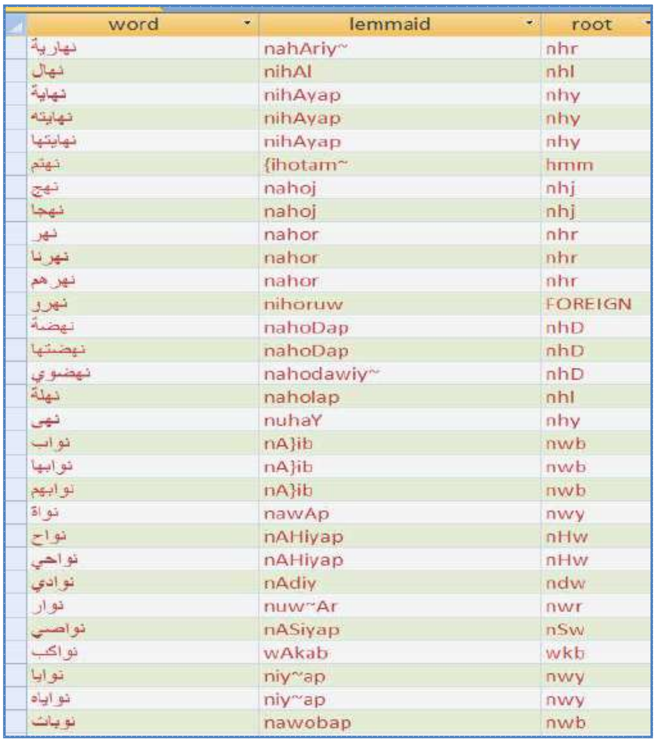
**C.** For more accuracy, analysts added new information that Buckwalter's analyzer does not provide; namely, root information.

The root of each word was detected according to its lemma. It was noted that some words have no root like "إذا، إفريقيا، أسفلت، ..." . Analysts gave such words the root "NONE". Also some foreign words were found in Arabic orthography such as, "إنترناشونال، سوستيه، شارون، ..." , analysts gave these words the root "FOREIGN". In addition, some words may have two roots as shown in table 6:

| Word | Lemma | Root |
|---|---|---|
| أبناء | {ibon | bnw/bny |
| أزال | >azAl | zwl/zyl |
| تنمية | tanomiyap | nmw/nmy |

Table 6: example of words may take two roots.

Figure 15 shows each word, lemma and its detected root:



| word | lemmaid | root |
|---|---|---|
| نهارية | nahAriy~ | nhr |
| نهال | nihAl | nhl |
| نهاية | nihAyap | nhy |
| نهايته | nihAyap | nhy |
| نهايتها | nihAyap | nhy |
| نهتم | {ihotam~ | hmm |
| نهج | nahoj | nhj |
| نهجا | nahoj | nhj |
| نهر | nahor | nhr |
| نهرنا | nahor | nhr |
| نهرهم | nahor | nhr |
| نهرو | nihoruw | FOREIGN |
| نهضة | nahoDap | nhD |
| نهضتها | nahoDap | nhD |
| نهضوي | nahodawiy~ | nhD |
| نهلة | naholap | nhl |
| نهى | nuhaY | nhy |
| نواب | nA}ib | nwb |
| نوابها | nA}ib | nwb |
| نوابهم | nA}ib | nwb |
| نواة | nawAp | nwy |
| نواح | nAHiyap | nHw |
| نواحي | nAHiyap | nHw |
| نوادي | nAdiy | ndw |
| نوار | nuw~Ar | nwr |
| نواصي | nASiyap | nSw |
| نواكب | wAkab | wkb |
| نوايا | niy~ap | nwy |
| نواياه | niy~ap | nwy |
| نوبات | nawobap | nwb |

Figure 15: Examples of root table.

### 5.3.3 Manual analysis of unanalyzed words:

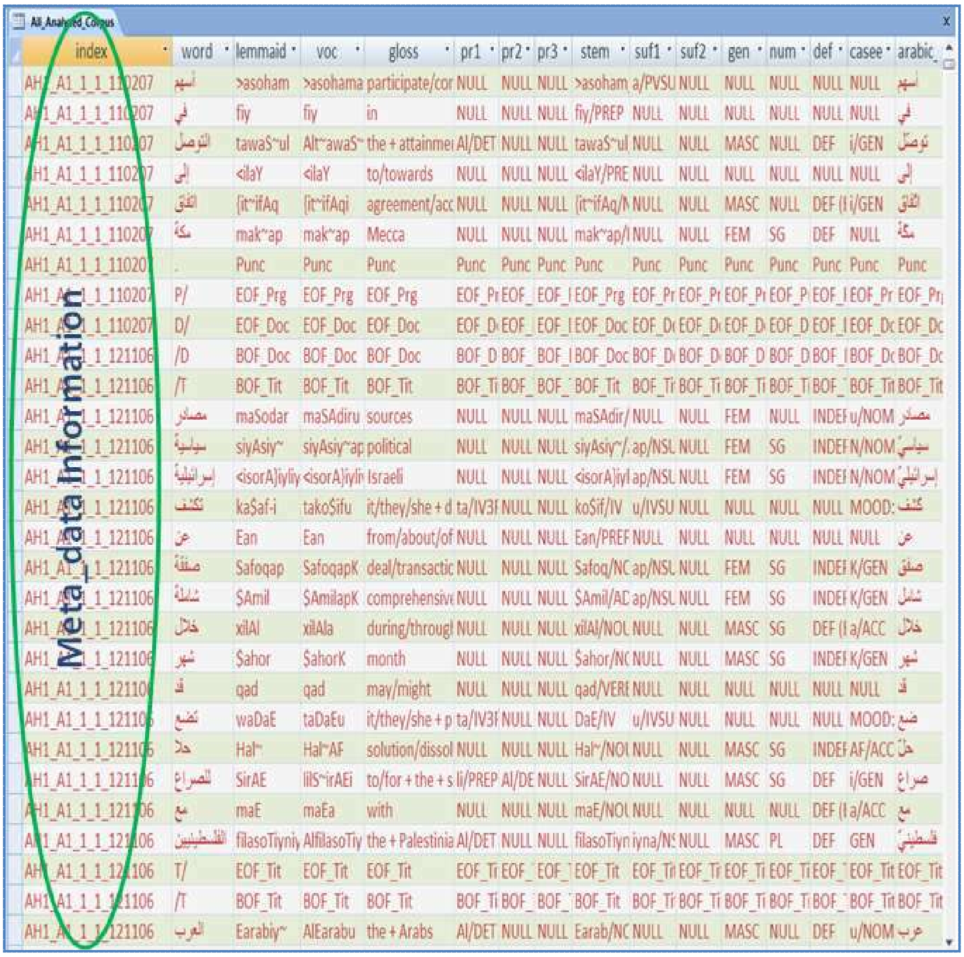
After choosing the suitable analysis for each word according to the context, some words were found to have no solution for one of two reasons. The First, some words have no analysis according to Buckwalter's analyzer. The Second, some words can be analyzed but no suitable analysis can be selected according to their context in the text. Therefore, these words have been analyzed manually according to their contexts as if they have been analyzed automatically.

It has been  noted that not all unanalyzed words were MSA Arabic words some of them are:

A. Colloquial words like "... جواهرجي – بتحبك – حنشوف – إزاي" which analysts tagged as (Colloquial).
B. Loan words like "... بلودوج – البرجماتية – تكنوكراتي ". These words have no counterpart in Arabic language and therefore have been tagged (Loan).
C. Non Arabic words that are used commonly like "... سنجل – ديكشنري " and also English words. These words have been tagged as (Not_Arabic).

## 5.4  ICA: A final analyzed view:

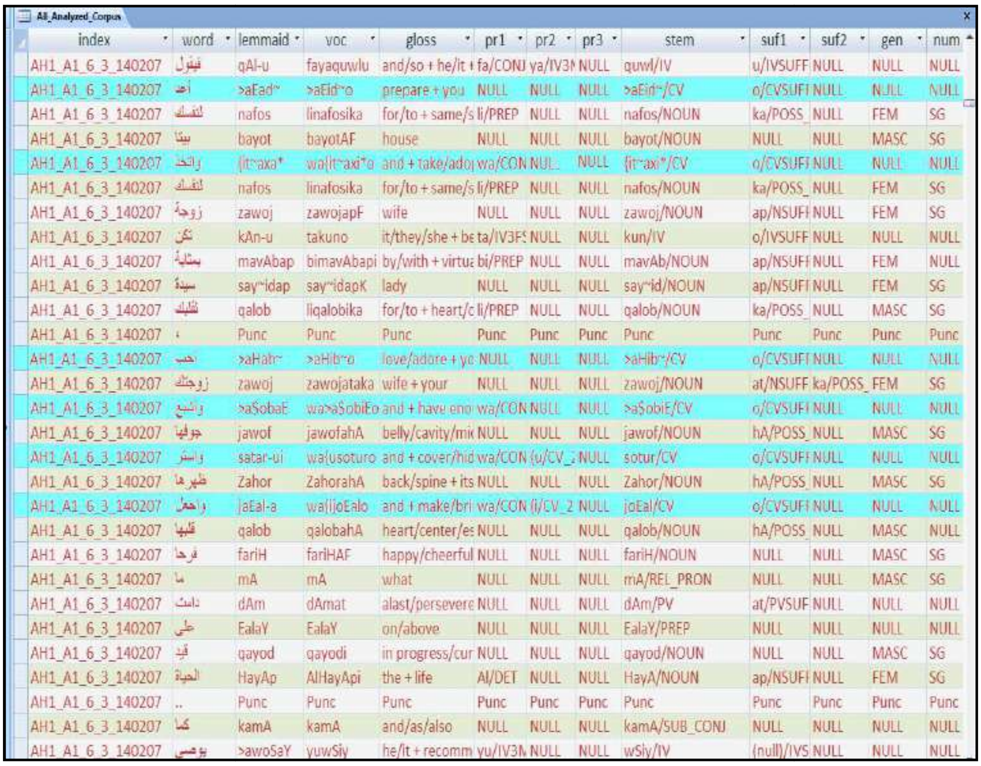
The current state of ICA analyzed corpus helps in interrogating a lot of phenomena since there is one database containing all analyzed words in their context and with their Meta data information. Each word has 17 pieces of information namely: Word, Lemma, Vocalization, Gloss, Prefix1, Prefix2, Prefix3, Stem, word class, Suffix1, Suffix2, number, gender, definiteness, Arabic stem, case and root as shown in figure 16.



| index | word | lemmaid | voc | gloss | pr1 | pr2 | pr3 | stem | suf1 | suf2 | gen | num | def | casee | arabic_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AH1_A1_1_1_110207 | أسهم | >asoham | >asohama | participate/cor | NULL | NULL | NULL | >asoham | a/PVSU | NULL | NULL | NULL | NULL | NULL | أسهم |
| AH1_A1_1_1_110207 | في | fiy | fiy | in | NULL | NULL | NULL | fiy/PREP | NULL | NULL | NULL | NULL | NULL | NULL | في |
| AH1_A1_1_1_110207 | التوصل | tawaS~ul | Alt~awaS~ | the + attainmei | Al/DET | NULL | NULL | tawaS~ul | NULL | NULL | MASC | NULL | DEF | i/GEN | توصّل |
| AH1_A1_1_1_110207 | إلى | <ilaY | <ilaY | to/towards | NULL | NULL | NULL | <ilaY/PRE | NULL | NULL | NULL | NULL | NULL | NULL | إلى |
| AH1_A1_1_1_110207 | اتفاق | {it~ifAq | {it~ifAqi | agreement/acc | NULL | NULL | NULL | {it~ifAq/N | NULL | NULL | MASC | NULL | DEF (I | i/GEN | اتفاق |
| AH1_A1_1_1_110207 | مكة | mak~ap | mak~ap | Mecca | NULL | NULL | NULL | mak~ap/I | NULL | NULL | FEM | SG | DEF | NULL | مكّة |
| AH1_A1_1_1_110207 | . | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc |
| AH1_A1_1_1_110207 | P/ | EOF_Prg | EOF_Prg | EOF_Prg | EOF_Pr | EOF_ | EOF_I | EOF_Prg | EOF_Pr | EOF_Pr | EOF_P | EOF_P | EOF_I | EOF_Pr | EOF_Pr |
| AH1_A1_1_1_110207 | D/ | EOF_Doc | EOF_Doc | EOF_Doc | EOF_D | EOF_ | EOF_I | EOF_Doc | EOF_D | EOF_D | EOF_D | EOF_D | EOF_I | EOF_Dc | EOF_Dc |
| AH1_A1_1_1_121106 | /D | BOF_Doc | BOF_Doc | BOF_Doc | BOF_D | BOF_ | BOF_I | BOF_Doc | BOF_D | BOF_D | BOF_D | BOF_D | BOF_I | BOF_Dc | BOF_Dc |
| AH1_A1_1_1_121106 | /T | BOF_Tit | BOF_Tit | BOF_Tit | BOF_Ti | BOF_ | BOF_ | BOF_Tit | BOF_Ti | BOF_Ti | BOF_Ti | BOF_Ti | BOF_ | BOF_Tit | BOF_Tit |
| AH1_A1_1_1_121106 | مصادر | maSodar | maSAdiru | sources | NULL | NULL | NULL | maSAdir/ | NULL | NULL | FEM | NULL | INDEF | u/NOM | مصادر |
| AH1_A1_1_1_121106 | سياسية | siyAsiy~ | siyAsiy~ap | political | NULL | NULL | NULL | siyAsiy~/. | ap/NSL | NULL | FEM | SG | INDEF | N/NOM | سياسيّ |
| AH1_A1_1_1_121106 | إسرائيلية | <isorA}iyliy | <isorA}iyli | Israeli | NULL | NULL | NULL | <isorA}iyl | ap/NSL | NULL | FEM | SG | INDEF | N/NOM | إسرائيليّ |
| AH1_A1_1_1_121106 | تكشف | ka$af-i | tako$ifu | it/they/she + d | ta/IV3I | NULL | NULL | ko$if/IV | u/IVSU | NULL | NULL | NULL | NULL | MOOD: | كشف |
| AH1_A1_1_1_121106 | عن | Ean | Ean | from/about/of | NULL | NULL | NULL | Ean/PREF | NULL | NULL | NULL | NULL | NULL | NULL | عن |
| AH1_A1_1_1_121106 | صفقة | Safoqap | SafoqapK | deal/transactic | NULL | NULL | NULL | Safoq/NC | ap/NSL | NULL | FEM | SG | INDEF | K/GEN | صفق |
| AH1_A1_1_1_121106 | شاملة | $Amil | $AmilapK | comprehensive | NULL | NULL | NULL | $Amil/AC | ap/NSL | NULL | FEM | SG | INDEF | K/GEN | شامل |
| AH1_A1_1_1_121106 | خلال | xilAl | xilAla | during/througl | NULL | NULL | NULL | xilAl/NOL | NULL | NULL | MASC | SG | DEF (I | a/ACC | خلال |
| AH1_A1_1_1_121106 | شهر | $ahor | $ahorK | month | NULL | NULL | NULL | $ahor/NC | NULL | NULL | MASC | SG | INDEF | K/GEN | شهر |
| AH1_A1_1_1_121106 | قد | qad | qad | may/might | NULL | NULL | NULL | qad/VERI | NULL | NULL | NULL | NULL | NULL | NULL | قد |
| AH1_A1_1_1_121106 | تضع | waDaE | taDaEu | it/they/she + p | ta/IV3I | NULL | NULL | DaE/IV | u/IVSU | NULL | NULL | NULL | NULL | MOOD: | ضع |
| AH1_A1_1_1_121106 | حلا | Hal~ | Hal~AF | solution/dissol | NULL | NULL | NULL | Hal~/NOI | NULL | NULL | MASC | SG | INDEF | AF/ACC | حلّ |
| AH1_A1_1_1_121106 | للصراع | SirAE | liS~irAEi | to/for + the + s | li/PREP | Al/DE | NULL | SirAE/NO | NULL | NULL | MASC | SG | DEF | i/GEN | صراع |
| AH1_A1_1_1_121106 | مع | maE | maEa | with | NULL | NULL | NULL | maE/NOL | NULL | NULL | NULL | NULL | DEF (I | a/ACC | مع |
| AH1_A1_1_1_121106 | الفلسطينيين | filasoTiyniy | AlfilasoTiy | the + Palestinia | Al/DET | NULL | NULL | filasoTiyn | iyna/N: | NULL | MASC | PL | DEF | GEN | فلسطينيّ |
| AH1_A1_1_1_121106 | T/ | EOF_Tit | EOF_Tit | EOF_Tit | EOF_Ti | EOF_ | EOF_ | EOF_Tit | EOF_Ti | EOF_Ti | EOF_Ti | EOF_Ti | EOF_ | EOF_Tit | EOF_Tit |
| AH1_A1_1_1_121106 | /T | BOF_Tit | BOF_Tit | BOF_Tit | BOF_Ti | BOF_ | BOF_ | BOF_Tit | BOF_Ti | BOF_Ti | BOF_Ti | BOF_Ti | BOF_ | BOF_Tit | BOF_Tit |
| AH1_A1_1_1_121106 | العرب | Earabiy~ | AlEarabu | the + Arabs | Al/DET | NULL | NULL | Earab/NC | NULL | NULL | MASC | NULL | DEF | u/NOM | عرب |

Figure 16: Final view of ICA analyzed corpus.

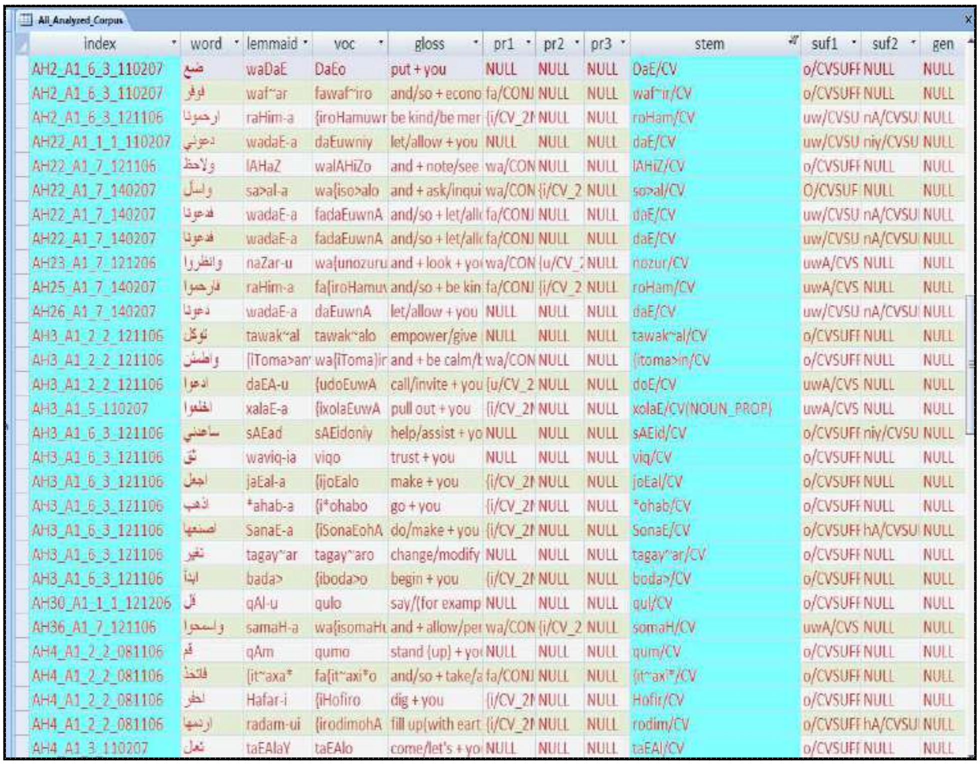
Through the analyzed ICA sample the analysts can capture any information easily. For example the analysts can capture all the imperative verbs whether in their contexts or without context as shown in figure 17 & 18. This can help in building a good search engine tool.



All_Analyzed_Corpus

| index | word | lemmaid | voc | gloss | pr1 | pr2 | pr3 | stem | suf1 | suf2 | gen | num |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AH1_A1_6_3_140207 | فيقول | qAl-u | fayaquwlu | and/so + he/it + | fa/CONJ | ya/IV3N | NULL | quwl/IV | u/IVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | أعد | >aEad~ | >aEid~o | prepare + you | NULL | NULL | NULL | >aEid~/CV | o/CVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | لنفسك | nafos | linafosika | for/to + same/s | li/PREP | NULL | NULL | nafos/NOUN | ka/POSS_ | NULL | FEM | SG |
| AH1_A1_6_3_140207 | بيتا | bayot | bayotAF | house | NULL | NULL | NULL | bayot/NOUN | NULL | NULL | MASC | SG |
| AH1_A1_6_3_140207 | واتخذ | {it~axa* | wa{it~axi*o | and + take/ado | wa/CON | NUL | NULL | {it~axi*/CV | o/CVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | لنفسك | nafos | linafosika | for/to + same/s | li/PREP | NULL | NULL | nafos/NOUN | ka/POSS_ | NULL | FEM | SG |
| AH1_A1_6_3_140207 | زوجة | zawoj | zawojapF | wife | NULL | NULL | NULL | zawoj/NOUN | ap/NSUFF | NULL | FEM | SG |
| AH1_A1_6_3_140207 | تكن | kAn-u | takuno | it/they/she + be | ta/IV3FS | NULL | NULL | kun/IV | o/IVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | بمثابة | mavAbap | bimavAbapi | by/with + virtu | bi/PREP | NULL | NULL | mavAb/NOUN | ap/NSUFF | NULL | FEM | NULL |
| AH1_A1_6_3_140207 | سيدة | say~idap | say~idapK | lady | NULL | NULL | NULL | say~id/NOUN | ap/NSUFF | NULL | FEM | SG |
| AH1_A1_6_3_140207 | لقلبك | qalob | liqalobika | for/to + heart/c | li/PREP | NULL | NULL | qalob/NOUN | ka/POSS_ | NULL | MASC | SG |
| AH1_A1_6_3_140207 | ، | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc |
| AH1_A1_6_3_140207 | احب | >aHab~ | >aHib~o | love/adore + yo | NULL | NULL | NULL | >aHib~/CV | o/CVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | زوجتك | zawoj | zawojataka | wife + your | NULL | NULL | NULL | zawoj/NOUN | at/NSUFF | ka/POSS_ | FEM | SG |
| AH1_A1_6_3_140207 | واشبع | >a$obaE | wa>a$obiEo | and + have eno | wa/CON | NULL | NULL | >a$obiE/CV | o/CVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | جوفها | jawof | jawofahA | belly/cavity/mi | NULL | NULL | NULL | jawof/NOUN | hA/POSS_ | NULL | MASC | SG |
| AH1_A1_6_3_140207 | واستر | satar-ui | wa{usoturo | and + cover/hid | wa/CON | {u/CV_2 | NULL | sotur/CV | o/CVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | ظهرها | Zahor | ZahorahA | back/spine + its | NULL | NULL | NULL | Zahor/NOUN | hA/POSS_ | NULL | MASC | SG |
| AH1_A1_6_3_140207 | واجعل | jaEal-a | wa{ijoEalo | and + make/bri | wa/CON | {i/CV_2 | NULL | joEal/CV | o/CVSUFF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | قلبها | qalob | qalobahA | heart/center/es | NULL | NULL | NULL | qalob/NOUN | hA/POSS_ | NULL | MASC | NULL |
| AH1_A1_6_3_140207 | فرحا | fariH | fariHAF | happy/cheerful | NULL | NULL | NULL | fariH/NOUN | NULL | NULL | MASC | SG |
| AH1_A1_6_3_140207 | ما | mA | mA | what | NULL | NULL | NULL | mA/REL_PRON | NULL | NULL | MASC | SG |
| AH1_A1_6_3_140207 | دامت | dAm | dAmat | alast/persevere | NULL | NULL | NULL | dAm/PV | at/PVSUF | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | على | EalaY | EalaY | on/above | NULL | NULL | NULL | EalaY/PREP | NULL | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | قيد | qayod | qayodi | in progress/cur | NULL | NULL | NULL | qayod/NOUN | NULL | NULL | MASC | SG |
| AH1_A1_6_3_140207 | الحياة | HayAp | AlHayApi | the + life | Al/DET | NULL | NULL | HayA/NOUN | ap/NSUFF | NULL | FEM | SG |
| AH1_A1_6_3_140207 | .. | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc | Punc |
| AH1_A1_6_3_140207 | كما | kamA | kamA | and/as/also | NULL | NULL | NULL | kamA/SUB_CONJ | NULL | NULL | NULL | NULL |
| AH1_A1_6_3_140207 | يوصي | >awoSaY | yuwSiy | he/it + recomm | yu/IV3M | NULL | NULL | wSiy/IV | (null)/IVS | NULL | NULL | NULL |

Figure 17: CV within context.



All_Analyzed_Corpus

| index | word | lemmaid | voc | gloss | pr1 | pr2 | pr3 | stem | suf1 | suf2 | gen |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AH2_A1_6_3_110207 | ضع | waDaE | DaEo | put + you | NULL | NULL | NULL | DaE/CV | o/CVSUFF | NULL | NULL |
| AH2_A1_6_3_110207 | فوفر | waf~ar | fawaf~iro | and/so + econo | fa/CONJ | NULL | NULL | waf~ir/CV | o/CVSUFF | NULL | NULL |
| AH2_A1_6_3_121106 | ارحموا | raHim-a | {iroHamuwr | be kind/be mer | {i/CV_2N | NULL | NULL | roHam/CV | uw/CVSU | nA/CVSU | NULL |
| AH22_A1_1_1_110207 | دعوني | wadaE-a | daEuwniy | let/allow + you | NULL | NULL | NULL | daE/CV | uw/CVSU | niy/CVSU | NULL |
| AH22_A1_7_121106 | ولاحظ | lAHaZ | walAHiZo | and + note/see | wa/CON | NULL | NULL | lAHiZ/CV | o/CVSUFF | NULL | NULL |
| AH22_A1_7_140207 | واسأل | sa>al-a | wa{iso>alo | and + ask/inqui | wa/CON | {i/CV_2 | NULL | so>al/CV | O/CVSUF | NULL | NULL |
| AH22_A1_7_140207 | فدعونا | wadaE-a | fadaEuwnA | and/so + let/all | fa/CONJ | NULL | NULL | daE/CV | uw/CVSU | nA/CVSU | NULL |
| AH22_A1_7_140207 | فدعونا | wadaE-a | fadaEuwnA | and/so + let/all | fa/CONJ | NULL | NULL | daE/CV | uw/CVSU | nA/CVSU | NULL |
| AH23_A1_7_121206 | وانظروا | naZar-u | wa{unozuru | and + look + yo | wa/CON | {u/CV_2 | NULL | nozur/CV | uwA/CVS | NULL | NULL |
| AH25_A1_7_140207 | فارحموا | raHim-a | fa{iroHamu | and/so + be kin | fa/CONJ | {i/CV_2 | NULL | roHam/CV | uwA/CVS | NULL | NULL |
| AH26_A1_7_140207 | دعونا | wadaE-a | daEuwnA | let/allow + you | NULL | NULL | NULL | daE/CV | uw/CVSU | nA/CVSU | NULL |
| AH3_A1_2_2_121106 | توكل | tawak~al | tawak~alo | empower/give | NULL | NULL | NULL | tawak~al/CV | o/CVSUFF | NULL | NULL |
| AH3_A1_2_2_121106 | واطمئن | {iToma>an~ | wa{iToma}ir | and + be calm/t | wa/CON | NULL | NULL | {itoma>in/CV | o/CVSUFF | NULL | NULL |
| AH3_A1_2_2_121106 | ادعوا | daEA-u | {udoEuwA | call/invite + you | {u/CV_2 | NULL | NULL | doE/CV | uwA/CVS | NULL | NULL |
| AH3_A1_5_110207 | اخلعوا | xalaE-a | {ixolaEuwA | pull out + you | {i/CV_2N | NULL | NULL | xolaE/CV(NOUN_PROP) | uwA/CVS | NULL | NULL |
| AH3_A1_6_3_121106 | ساعدني | sAEad | sAEidoniy | help/assist + yo | NULL | NULL | NULL | sAEid/CV | o/CVSUFF | niy/CVSU | NULL |
| AH3_A1_6_3_121106 | ثق | waviq-ia | viqo | trust + you | NULL | NULL | NULL | viq/CV | o/CVSUFF | NULL | NULL |
| AH3_A1_6_3_121106 | اجعل | jaEal-a | {ijoEalo | make + you | {i/CV_2N | NULL | NULL | joEal/CV | o/CVSUFF | NULL | NULL |
| AH3_A1_6_3_121106 | اذهب | *ahab-a | {i*ohabo | go + you | {i/CV_2N | NULL | NULL | *ohab/CV | o/CVSUFF | NULL | NULL |
| AH3_A1_6_3_121106 | اصنعها | SanaE-a | {iSonaEohA | do/make + you | {i/CV_2N | NULL | NULL | SonaE/CV | o/CVSUFF | hA/CVSU | NULL |
| AH3_A1_6_3_121106 | تغير | tagay~ar | tagay~aro | change/modify | NULL | NULL | NULL | tagay~ar/CV | o/CVSUFF | NULL | NULL |
| AH3_A1_6_3_121106 | ابدأ | bada> | {iboda>o | begin + you | {i/CV_2N | NULL | NULL | boda>/CV | o/CVSUFF | NULL | NULL |
| AH30_A1_1_1_121206 | قل | qAl-u | qulo | say/(for examp | NULL | NULL | NULL | qul/CV | o/CVSUFF | NULL | NULL |
| AH36_A1_7_121106 | واسمحوا | samaH-a | wa{isomaHu | and + allow/per | wa/CON | {i/CV_2 | NULL | somaH/CV | uwA/CVS | NULL | NULL |
| AH4_A1_2_2_081106 | قم | qAm | qumo | stand (up) + yo | NULL | NULL | NULL | qum/CV | o/CVSUFF | NULL | NULL |
| AH4_A1_2_2_081106 | فاتخذ | {it~axa* | fa{it~axi*o | and/so + take/a | fa/CONJ | NULL | NULL | {it~axi*/CV | o/CVSUFF | NULL | NULL |
| AH4_A1_2_2_081106 | احفر | Hafar-i | {iHofiro | dig + you | {i/CV_2N | NULL | NULL | Hofir/CV | o/CVSUFF | NULL | NULL |
| AH4_A1_2_2_081106 | اردمها | radam-ui | {irodimohA | fill up(with eart | {i/CV_2N | NULL | NULL | rodim/CV | o/CVSUFF | hA/CVSU | NULL |
| AH4_A1_3_110207 | تعال | taEAlaY | taEAlo | come/let's + yo | NULL | NULL | NULL | taEAl/CV | o/CVSUFF | NULL | NULL |

Figure 18: CV without context.

## 6. Conclusion:

This paper presented a road map of a trial for Arabic corpus analysis. The analysts followed a stem-based approach to be used in analyzing ICA. Buckwalter Morphological analyzer is the most suitable available lexical resource for our approach. The paper discussed a number of general considerations to bear in mind when beginning the process of analyzing the ICA corpus. This trial can be considered one of the most successful approaches for analyzing modern standard Arabic (MSA) in comparison with other trials of Arabic analyzed corpora.

This analyzed sample will be developed to be used as a training corpus to analyze the target size of ICA (100 million words). The ICA software will be developed to interrogate the analyzed version to help researchers to capture powerful textual search.

## 7. References:

Abdelali A. (2004), **Localization in Modern Standard Arabic**, Journal of the American Society for Information Science and technology (JASIST), Volume 55, Number 1, 2004. pp. 23-28.

Al-Sulaiti L. & Atwell E. (2001), **Extending the Corpus of Contemporary Arabic**, School of Computing, University of Leeds.

Attia M. (2000), **A large-scale computational processor of the Arabic morphology and applications**, Faculty of engineering, Cairo university.

Attia M. (2006), **An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks** , School of Informatics, The University of Manchester.

Beesley K. (1996), **Arabic finite-state morphological analysis and generation**, In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pages volume 1, 89–94, Copenhagen, Denmark.

Beesley K. (1998a.), **Arabic morphology using only finite-state operations**, **Computational Approaches to Semitic Languages**, Proceedings of the Workshop, pages 50–57, Montr´eal, Qu´ebec, August 16. Universit´e de Montr´eal.

Beesley K. (1998b.), **Arabic Linguistic Society**, Paper presented at the 12th Symposium on Arabic Linguistics, 6-7 March, Champaign, IL.

Buckwalter T. ( 2002), **Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium**, University of Pennsylvania, LDC Catalog No.: LDC2002L49.

Choukri K., Krawner S. (2004), **Arabic Language Resources and Tools**, Nemlar.

Choukri K., Krawner S., Maegaard B., The BLARK (2006), **concept and BLARK for Arabic**, Proceedings of the 5th International Conference on Language Resources and Evaluation. Genova.

Darwish K. (2002), **Building a Shallow Morphological Analyzer in One Day,** In Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA.

Dichy J. & Fargaly A. (2003), **Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?**, Proceedings of the MTSummit IX workshop on Machine Translation for Semitic Languages, New-Orleans.

Eriksson T. & Ritchey T. (2002), **Scenario Development using Computerised Morphological analysis**, Presented at the Winchester International OR Conference, England.

Habash N. & TALN J. (2004), **Scale Lexeme Based Arabic Morphological Generation**, Session Traitement Automatique de l'Arabe, Institute for Advanced Computer Studies, University of Maryland College Park College Park, Maryland, 20742.

Hajič O. & et al (2006), **THE CHALLENGE OF ARABIC FOR NLP/MT, Tips and Tricks of the Prague Arabic Dependency Treebank**, International Conference at The British Computer Society (BCS), 23 October, London.

Hilbert D. & Krenn B. (2006), **Computational Approaches to Collocations**, UCS toolkit v0.5 pre-release version fixes some compatibility issues (11-01).

Hockett C., 1947, **problems of morphemic analysis** , Linguistic Society of America, Language, Vol. 23, No. 4 (Oct. - Dec., 1947), pp. 321-343.

Hulstijg J. (1992), **Retention of inferred and given word meanings: experiments in incidental vocabulary learning**, In P.J.L Arnaud and H.bejoint (eds), vocabulary and applied linguistics. London: Macmillan, 113-25.

Kaplan J. & Holland V. (1995), **Natural language processing techniques in computer assisted language learning: status and instructional issues**, Springer, Instructional Science. 23,351-80.

Karttunen L. (2005), **Twenty-five years of finite-state morphology**, CSLI Publications.

Karttunen, Kaplan R., & Zaenen A. (1992), **Two-level morphology with composition**, In Proceedings of Fourteenth International Conference on Computational Linguistics (COLING-92), pages 141–148, Nantes, July 20–28, France.

Kiraz G.(1994), **Multi-tape Two-level Morphology: A Case study in Semitic Non-Linear Morphology**, In Proceedings of Fifteenth International Conference on Computational Linguistics (COLING-94), pages 180–186, Kyoto, Japan.

Krauwer S. (2003), **The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap**, Proceedings of 2nd International Conference on Speech and computer.

Landauer T., Foltz P., & Laham D. (1998), **Introduction to Latent Semantic Analysis**., Discourse Processes, 25, 259-284.

Lee Y. (2004), **Morphological Analysis for Statistical Machine Translation**, IBM T. J. Watson Research Center, Yorktown Heights, NY-10598.

Maamouri M., Bies A., Buckwalter T. & Mekki W. (2004), **The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus**, NEMLAR Conference on Arabic Language Resources and Tools.

Manning C. & Schütze H. (1999), **Foundations of Statistical Natural Language Processing**, MIT Press, Cambridge, Massachusetts.

Meyer C. (2002), **English corpus linguistics, an introduction**, Cambridge University Press.

Nerbonne J., Jager S. & Essen A. (1997), **Language Teaching and Language Technology**, the University of Groningen, April 28-29, 1997.

Resnik P. (1998), **Statistical Methods in NLP**, July 8-10, Short Course.

Ritchey T. (2002-2006), **General Morphological Analysis, A general method for non-quantified modeling**, Downloaded from the Swedish Morphological Societ, Adapted from the paper "Fritz Zwicky, Morphology and Policy Analysis".

Ritchey T. (2005-2008), **Wicked Problems, Structuring Social Messes with Morphological Analysis**, Swedish Morphological Society.

Ritchey, T. (1998), **General Morphological Analysis, A general method for non-quantified modeling**, "Fritz Zwicky, **'Morphologie' and Policy Analysis**", Presented at the 16th Euro Conference on Operational Analysis, Brussels.

Soudi A., Cavalli-Sforza V., & Jamari A. (2001), **A Computational Lexeme-Based Treatment of Arabic Morphology**, Proceedings of the Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001), Jul 6, Toulouse, France.

Soudi A., Bosch A. & Neumann G. (2007), **Arabic Computational Morphology, Knowledge-based and Empirical Methods**, Springer.

Swaab T. & Kaan E. (2003), **Repair, Revision, and, Complexity in Syntactic Analysis: An Electrophysiological Differentiation**, The MIT Press, Journal of Cognitive Neuroscience.

ZAUGUAGE S, Varga D. (1955), **Syntactic analysis in the case of highly inflecting languages, international conference on computational linguistics**, Computing Centre of the Hungarian Academy of Sciences, 53, Uri u., Budapest I., Hungary.A1

Zemanek P. (2001), **CLARA (Corpus Linguae Arabicae): An Overview**, Proceedings of ACL/EACL Workshop on Arabic Language.

# Egyptian License Plate Recognition System Using DWT and Template Matching

Ahmed R. EL-Barkouky
Engineering Physics and
Mathematics Department,
Faculty of Engineering,
Ain Shams University
barkouky@mathasu.edu.eg

Salwa H. El-Ramly
Electronics and Communication
Engineering Department,
Faculty of Engineering,
Ain Shams University
sramlye@netscape.net

Mohamed I. Hassan
Engineering Physics and
Mathematics Department,
Faculty of Engineering,
Ain Shams University
mihassan@hotmail.com

## Abstract

License Plate Recognition (LPR) Systems are of considerable interest because of their potential applications in many automated security and access control systems. This paper introduces an automatic license plate recognition system for the Egyptian license plates. The system receives the images of vehicles captured with a digital camera. Discrete Wavelet Transform is then used to find the location of the license plate followed by several image processing techniques for segmentation of characters. Finally template matching is used for the recognition of characters. The performance of the system has been investigated on real images of the three main types of license plates that represent the majority of license plates in Egypt.

## 1. Introduction

License Plate Recognition has a wide range of applications, which use the extracted plate number to create automated solutions for various problems. Examples of such applications include **Parking,** where the plate number is used to automatically enter pre-paid vehicles and calculate parking fees for non-members (by comparing the in & out times). **Access control,** controlling gate opening in a secured area for authorized vehicles only, this replaces or assists the security guard. In **Border crossings,** the plate number and a picture of the car are saved when entering or leaving the Country; this can be used to monitor the border crossings, shortens the turnaround time and eliminates the typical long lines. **Toll payment,** the plate number is used to calculate the road toll, or used to double-check the ticket. **Enforcement,** the plate number is used in traffic surveillance to produce a speed ticket or any violation ticket. For **Marketing,** the plate number can be used to make a database of frequent visitors for marketing purposes or to build a traffic profile, like the number of visits versus the time [1].

The Egyptian license plates had three main different types; one of them is recently introduced in August 2008 which will be referred to as “new license”, it contains letters and numbers all of them are written in the same font. While the two other types contain only numbers and will be referred to as “old license”. The algorithm can also work for other types of old licenses which look the same but with different font and contents; some of them are shown in figure 1.

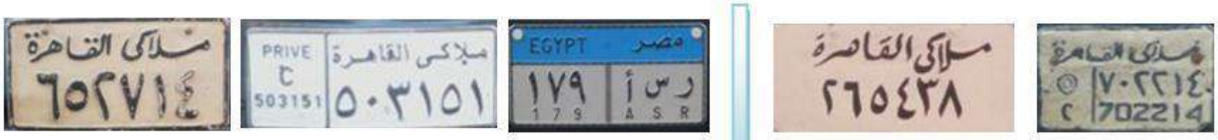


figure 1.(a) The three main types of Egyptian plates (b) Other types

LPR algorithms are generally composed of the following three processing steps: locating the license plate region, segmentation of the plate characters and recognition of each character [2].

The rest of the paper is organized as follows: Section 2 introduces the 2D wavelet decomposition. Section 3 describes the overall system, listing different stages in the LPR system. Section 4 discusses experimental results and the paper is concluded in Section 5.

## 2. The Discrete Wavelet Transform

When digital images are to be viewed or processed at multiple resolutions the Discrete Wavelet Transform (DWT) is the mathematical tool of choice [3]. The Fast Wavelet Transform (FWT) is a computationally efficient implementation of the DWT that uses filters and downsamplers.

We first start by the one dimensional scaling function $\varphi$ and the corresponding wavelet function $\psi$ which have the property that they can be expressed as linear combinations of double-resolution copies of themselves.

$$\varphi(x) = \sum_n h_\varphi(n)\sqrt{2}\varphi(2x-n) \qquad (1)$$

$$\psi(x) = \sum_n h_\psi(n)\sqrt{2}\varphi(2x-n) \qquad (2)$$

where $h_\varphi, h_\psi$ are called scaling and wavelet vectors & they are the filter coefficients of the FWT.

Now in two dimensions, a two-dimensional scaling function $\varphi(x,y)$, and three two-dimensional wavelets $\psi^H(x,y)$, $\psi^V(x,y)$, $\psi^D(x,y)$ are required as follows [4].

$$\varphi(x,y) = \varphi(x)\varphi(y) \qquad (3)$$

$$\psi^H(x,y) = \psi(x)\varphi(y) \qquad (4)$$

$$\psi^V(x,y) = \varphi(x)\psi(y) \qquad (5)$$

$$\psi^D(x,y) = \psi(x)\psi(y) \qquad (6)$$

These wavelets measure intensity or gray-level variations along different directions: $\psi^H$ measures variations along columns (horizontal edges), $\psi^V$ responds to variations along rows (vertical edges), and $\psi^D$ corresponds to variations along diagonal. The directional sensitivity is a natural consequence of the separability imposed by their definition.

Define the scaled and translated basis functions as:

$$\varphi_{j,m,n}(x,y) = 2^{j/2}\varphi(2^j x - m, 2^j y - n) \qquad (7)$$

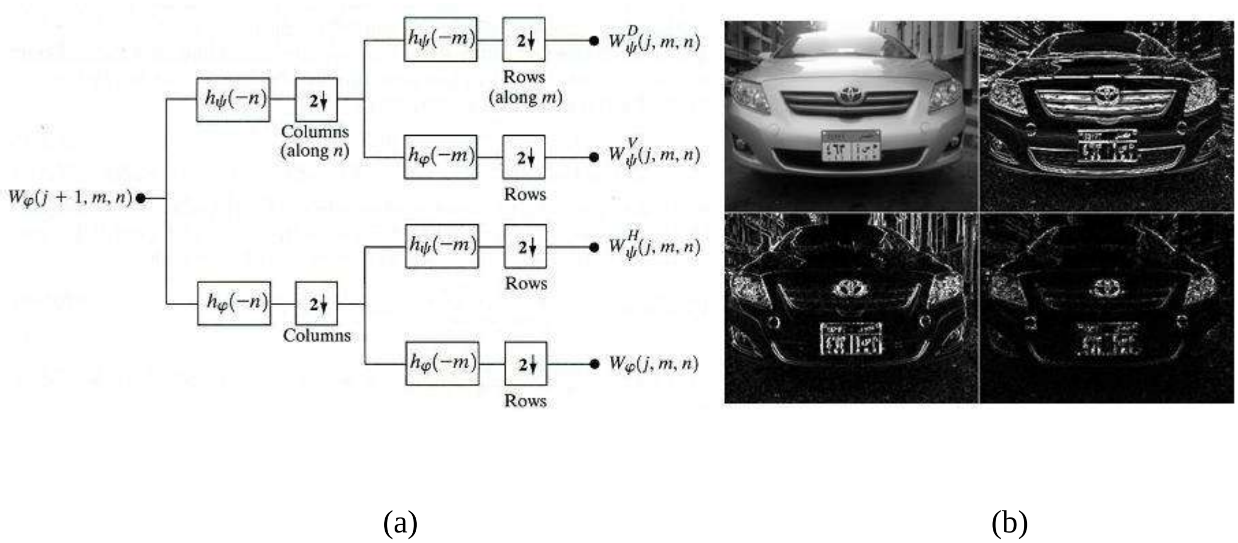
$$\psi^i_{j,m,n}(x,y) = 2^{j/2}\psi^i(2^j x - m, 2^j y - n) \qquad (8)$$
$$i = \{H, V, D\}$$

The DWT of an image $f(x,y)$ of size $M \times N$ is:

$$W_\varphi(j_o, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\varphi_{j_o,m,n}(x,y) \qquad (9)$$

$$W^i_\psi(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\psi^i_{j,m,n}(x,y) \qquad (10)$$
$$i = \{H, V, D\}$$

The two-dimensional DWT can be implemented using digital filters and down samplers, we simply take the one-dimensional FWT of the rows of $f(x,y)$, followed by FWT of the resulting columns as shown in figure 2(a) where Blocks containing time reversed scaling and wavelet vectors $h_\varphi(-n), h_\psi(-m)$ are low pass and high pass decomposition filters, respectively. Figure 2(b) illustrates how the wavelet decomposition coefficients are displayed into four sub images containing approximation, horizontal, vertical and diagonal details.



(a) (b)

Figure 2: (a) Block diagram for one stage of FWT. (b) Displaying the output, vertical and horizontal details (edges) are clear

### 3. License Plate Recognition system

The system is designed to recognize license plates from the front of the vehicle but it can also work from the rear. The input to the system is an image taken by a digital camera, containing the license plate and the output of the system is the characters on the license plate. The system consists of three stages: License plate extraction, character Segmentation and character recognition, as shown in figure 3.

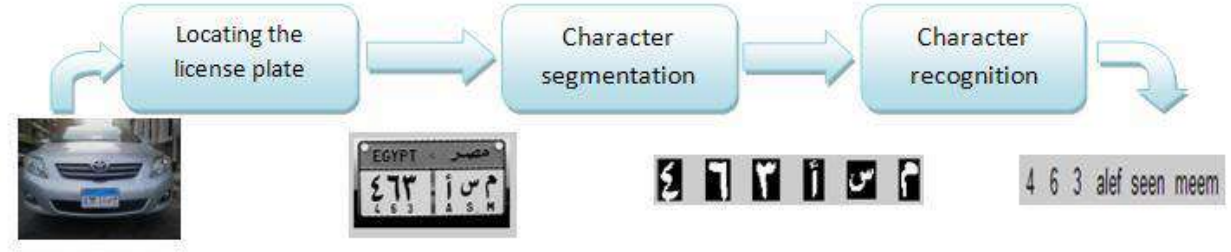


Figure 3. The typical structure of a LPR system

### 3.1 Locating the License Plate

Extracting the license plate from the vehicle image is considered the most challenging stage in the LPR system [5]. In this paper we depend on the directional sensitivity of the two-dimensional DWT for finding the license plate as follows:

1. The DWT transforms the image into 4 sub-images: approximation and three details horizontal, vertical and diagonal. We use three levels of decomposition, so the approximation of each level is replaced by the 4 sub-images of the next level as shown in figure 4.
2. The approximation coefficients will be replaced by zeros, and the details coefficients (edges) will be threshold, eliminating most of the horizontal and diagonal details and leaving the vertical details to make use of the characters inside the license plate which appear as adjacent vertical edges. We suppress the horizontal details to reduce the effect of the other edges in the vehicle.
3. Then we will compute the inverse DWT which will contain more vertical edges in the license plate part due to the presence of characters inside it.
4. Now to compensate the effect of reducing horizontal edges we will use image closing (dilation followed by erosion) with a horizontal long rectangle which will connect these vertical edges of the characters to make one big rectangle in the region of the license plate.

Finally we search the connected regions for the license, first arrange the connected objects by size then remove very large and very small objects. Finally, searching in a descending way, the remaining objects for a rectangle with suitable dimensions.

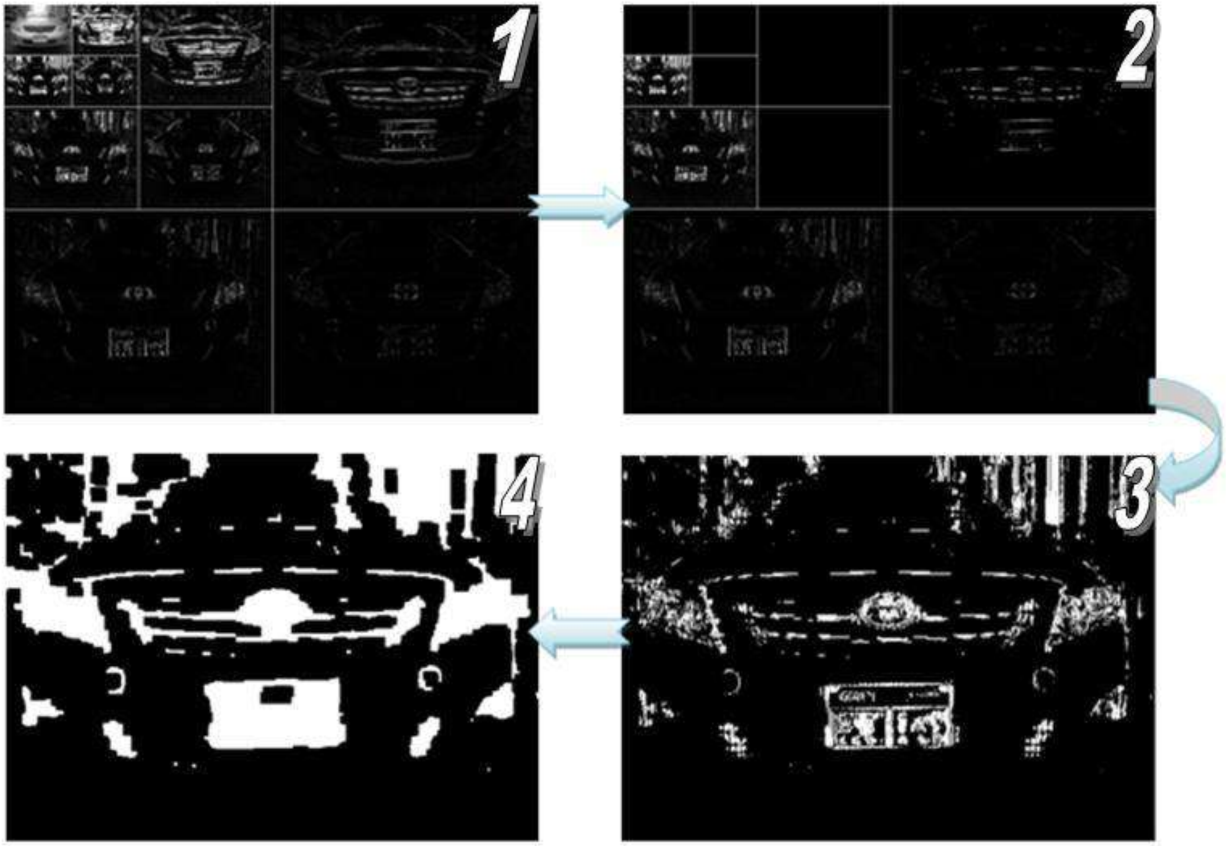


Figure 4. The 4 steps for locating the license plate

The advantage of using DWT in this way is that if we didn't find the license (which means no object in the image meets the dimensions of a license), we can repeat the process but adding more details. Now the process is adaptive, it repeats it self with different amount of details until the license is found. This way we can find the license even if the image has poor edges as a result of old bad licenses.

Figure 5. (a) Can't find the license (b) just add more wavelet coeffs. (c) and the license is found

### 3.2 Character segmentation

After finding the license plate we will use image filling which converts any isolated black region to white. Now subtracting the result from the license image will highlight the characters and cancel any extra parts that surround the license plate. This idea compensates the effect of having some vertical details adjacent to the license plate in the front of some cars which results in extra part of the car surrounding the license plate in the last stage.

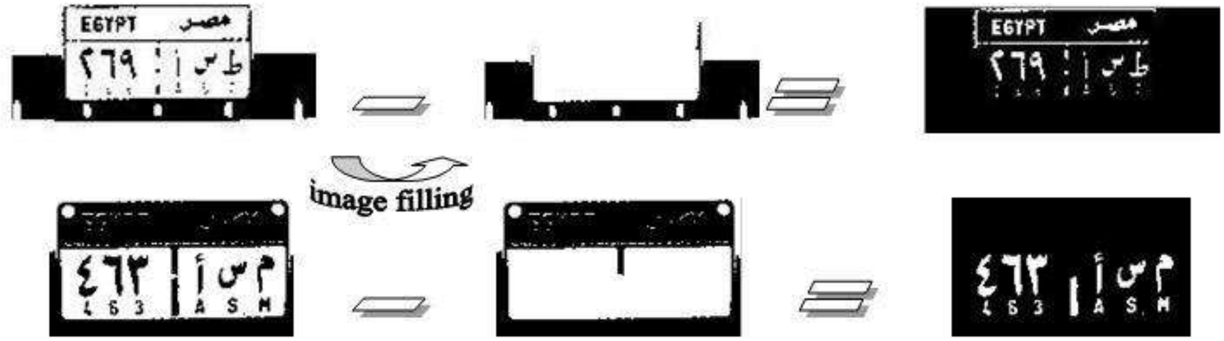


Figure 6. Image filling then subtracting to highlight the characters

The objects in the upper part of the license plate (EGYPT) are deleted if there's any, and  the image is dilated with a vertical rectangle to connect the Arabic and English characters then  any small objects are deleted (some times the license plate is dirty or scratched). After that the characters are separated.

### 3.3 Character recognition

For character recognition we used template matching just for simplicity because we have clear separated characters [6]. For the new Egyptian license plates this gives good results because all the characters are written in the same font and the plates are new. But for the old license plates some problems appeared due to bad conditions of some license plates or because of using different fonts (usually hand written).

To make the templates of each character we take only the largest object in each character image to be our characters and delete the others (the small English letters and any scratches), then we confine our image to only the borders of the letters, then we resize the image to 100x60 so that they all have the same size. We take care that the letter 'alef' will appear in this way like a big white rectangle so in the correlation it will always give the highest value. To avoid this we check first the width of the letter if it's smaller than a specified value we use zero padding from the left and right so that the letter 'alef' appears as a white strip in a black rectangle.

The available letters in the new license are only 17 till now, maybe they are avoiding the letters that differs only with a dot like ‘peh’, ‘teh’ & ‘seh’ we have only ‘peh’. The same for ‘seen’, ‘sheen’; we only got ‘seen’; also ‘saad’, ‘daad’ we only found ‘saad’, and so on.

Also for the numbers we still don’t have a zero, but in the old license we have zero. We used the blue part which contains Egypt in the new license to help the program decide whether this is a new license or an old one, because in the new ones we have the first three characters letters, while in the old all of them are numbers. Also the program is designed to check if this is a four, five or six digits numbers in the old license.

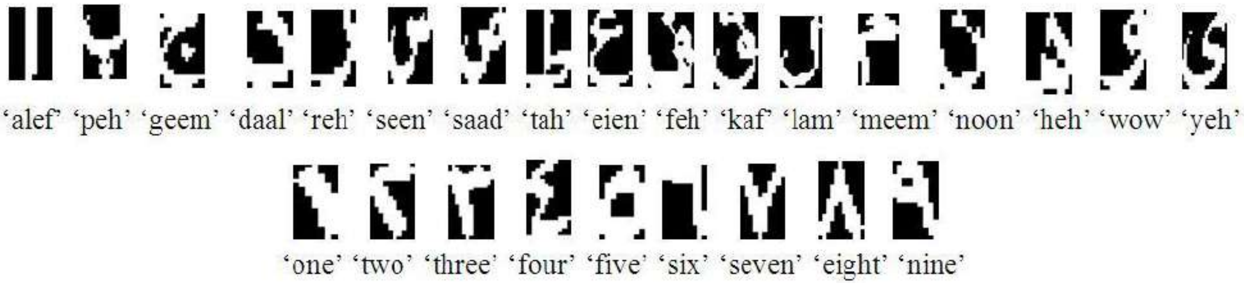

‘alef’ ‘peh’ ‘geem’ ‘daal’ ‘reh’ ‘seen’ ‘saad’ ‘tah’ ‘eien’ ‘feh’ ‘kaf’ ‘lam’ ‘meem’ ‘noon’ ‘heh’ ‘wow’ ‘yeh’
‘one’ ‘two’ ‘three’ ‘four’ ‘five’ ‘six’ ‘seven’ ‘eight’ ‘nine’

Figure 7. The new license templates

**4 Experimental results**

Experiments have been performed to test the above system. The system is designed in MATLAB 7.5 for recognition of Egyptian license plates using Pentium(R) M 1.8 GHz processor. The input image to the system is a gray scale of size 640x480 captured by a Carl Zeiss 3.2 megapixel digital camera.

The LPR system consists of three stages, using the DWT in locating the license plate gives great results. The code was tested on 200 images; the license plate was located correctly in all of them except only 2 images, achieving an impressive detection rate of 99%. Note that combining DWT with the simple but smart idea of filling the image then subtracting to highlight the characters enables us to avoid trying to locate the license plate precisely (without any small adjacent part from the car image). During the experiments some license plates was not detected at all, this was not a difficult problem cause we just added more coefficients as explained before and the problem was solved. The errors occurred when the program is deceived by another part of the car that has the same dimensions as the plate.

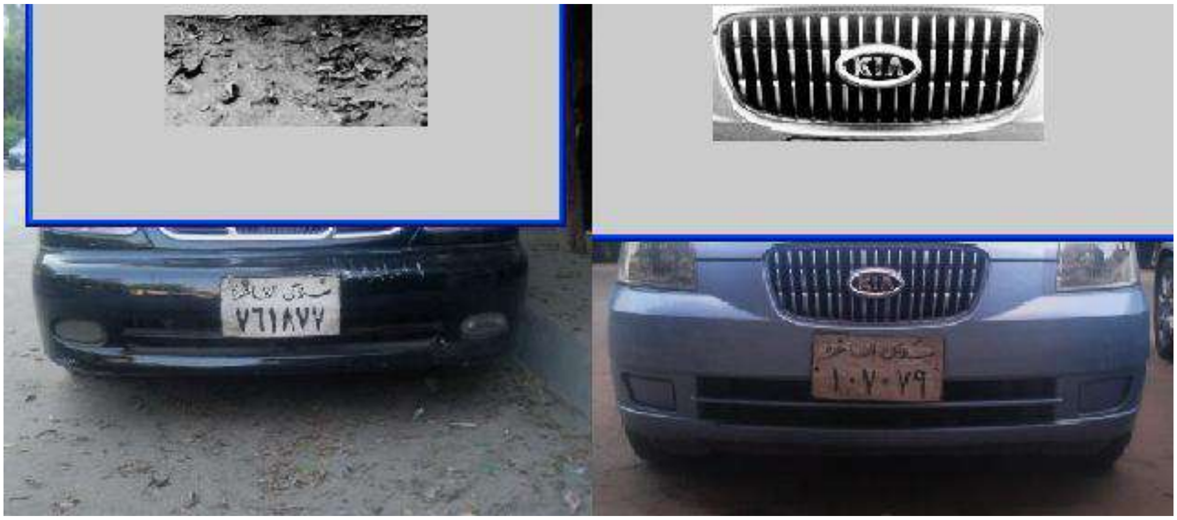


Figure 8. Errors in license plate detection

Segmentation of characters only faced problems on those images with bad lighting conditions like reflections of sun light if the image was taken in a sunny day. In the old licenses, we faced some problems in dirty or scratched license plates. From the 198 cars that pass the first stage correctly, the character segmentation was correct in 196 of them giving a 99% success rate. This makes the performance 98% for both the first and second stage together. Note that the images with license plates in very bad condition were removed from our statistics.

Recognition using template matching was good in both the new license plates and the old license plates of the first kind (shown in the 2nd column in table 1) but it was not suitable for old license plates of the 2nd kind (shown in the last column in table 1). This is because those license plates are old and most of them contain many scratches and bolts and written in different fonts so usually there is one wrong character in detection. So we will measure the performance of the system for only the 160 cars of the first two types: 13 cars had only one wrong character and 3 had more than one wrong character resulting in a 90% success rate. The main problem was in letters like ‘noon’ and ‘peh’ when their dot is not connected to them. Solving this dot problem will raise the success rates.

This makes the total detection rate for only those 160 cars 89.38%. The following table summarizes the results for the three different types of license plates with the success rates in each stage.

Table 1. Summary of the results

| | | | |
|---|---|---|---|
| First stage<br>License plate detection | 115/115<br>100% | 45/45<br>100% | 38/40<br>95% |
| Second stage<br>Character segmentation | 114/115<br>99.13% | 45/45<br>100% | 37/38<br>97.37% |
| Third stage<br>Character Recognition | 104/114<br>91.2% | 39/45<br>86.67% | ----- |
| The over all system | 104/115<br>90.43% | 39/45<br>86.67% | ----- |

In some cars the plate was not fixed perfectly horizontal, and the code detected it correctly. Also some cars had accessories in front of the license and it also works great, because the code doesn’t depend only on the edges of the plate; it also depends on the numbers inside the plate, so when the plate edges are not clear the numbers will do the job. Images captured from the back of the vehicle works just like from the front.

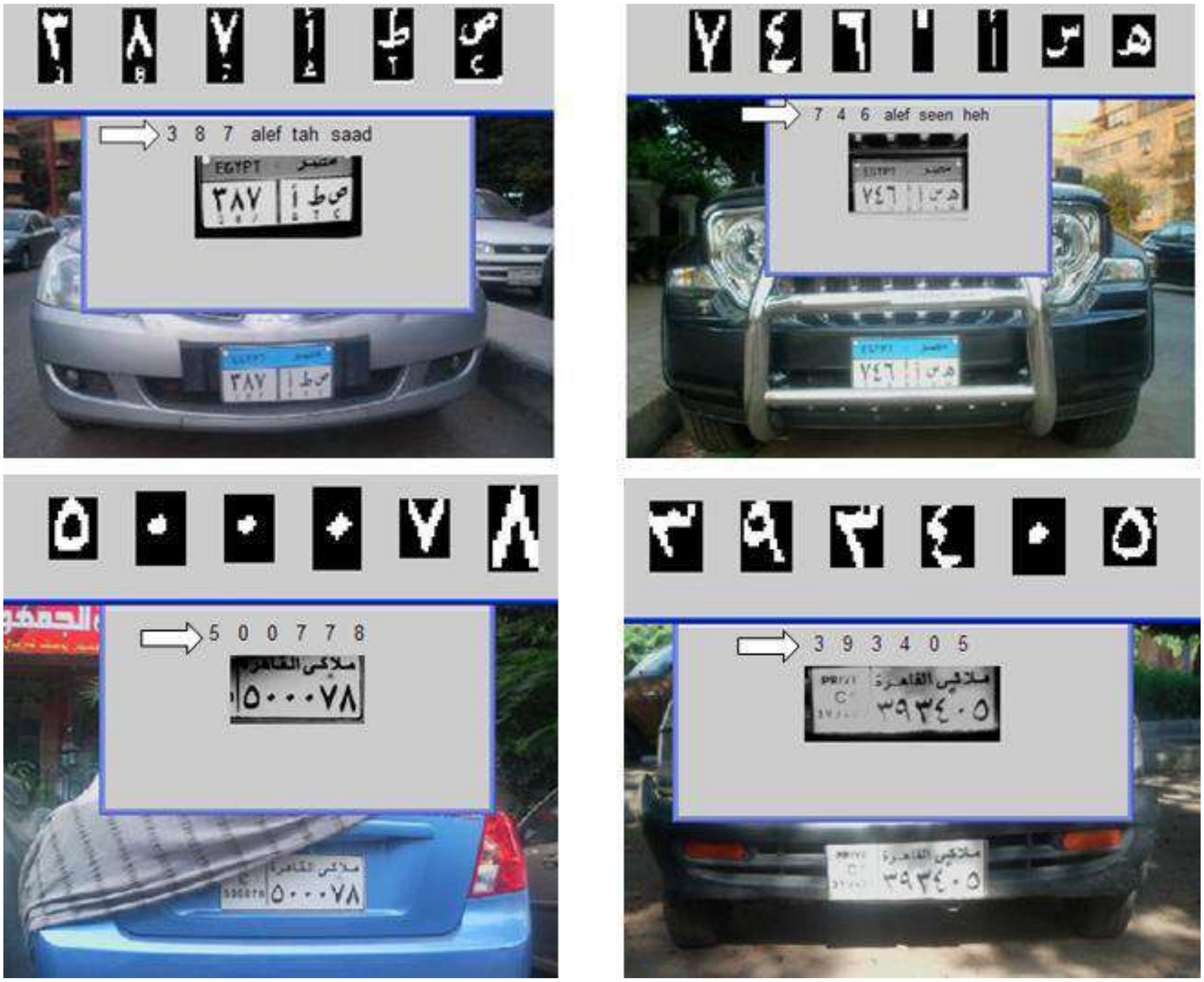


Figure 8. Difficult images due to inclined plates or car accessories

## 5 Conclusion

It seems clear from the results that the system was superior in the first stage, with also great results in the second stage, but the third stage still needs a more efficient recognition method especially for the old license plates to reduce the total error. The government is intending to replace all the license plates with the new ones, that is why more work was done on them. The problems in segmentation were due to sun reflection on the plate, scratched and dirty plates.

## 6 References

1. S. Youssef and S. AbdelRahman “A smart access control using an efficient license plate location and recognition approach” in ELSEVIER Expert Systems with Applications 34 (2008) 256–265.

2. C. Anagnostopoulos, I. Anagnostopoulos, I. Psoroulas, V. Loumos and E. Kayafas "License Plate Recognition From Still Images and Video Sequences: A Survey" in IEEE Transactions On Intelligent Transportation Systems, Vol. 9, No. 3, September 2008, pp. 377-391.
3. R. Gonzalez, R. Woods and S. Eddins, "**Digital Image Processing Using Matlab",** Prentice Hall, 2004.
4. R. Gonzalez and R. Woods, "**Digital Image Processing". Second edition**, Prentice Hall, 2002.
5. C. Hsieh, Y. Juan, K. Hung "Multiple License Plate Detection for Complex Background" in IEEE Proceedings of the 2005 19th International Conference on Advanced Information Networking and Applications (AINA'05) 1550-445X/05.

6. M. Sarfraz, M. Jameel and S. Ghazi. "Saudi Arabian license plate recognition system". In IEEE Proceedings of the 2003 International conference on Geometric Modeling and Graphics (GMAG03) 0-7695-1985-7/03.

# Fractal Image Compression Applied on Document Images

**Salwa H. El-Ramly**

Electronics and Communication
Engineering Department, Faculty of
Engineering, Ain-Shams University

sramlye@netscape.net

**Ramy F. Taki El-Din**

Engineering Physics and Mathematics
Department, Faculty of Engineering,
Ain-Shams University.

ramyfarouk@hotmail.com

## ABSTRACT

This paper presents a lossy compression technique for scanned document images. We introduce a two-step Quadtree partitioning fractal image compression method used for images containing Arabic and English texts. Fractal image compression makes use of the self-similarity in text images in a hierarchical way. A first step partitioning scheme partitions the image support to range blocks of size 16x16. A control parameter for each range block is computed, which provides the decision for a second step partitioning, 8x8 pixels, if required. Local neighboring searching is used for each range block for finding its best matching domain block in our results. Influences of fractal image compression on a group of document images are discussed. We attempt a significant improvement in the compression ratio with no visible artifacts.

## I. Introduction

Documents images, or textual images, appear in our daily life. They are pixels carrying intensity values representing their color. Compression is desirable for document image transmission and storage. For example, a letter size document if sampled at 512x512 pixels would require about 32 KB of data without compression. It would take almost 9 seconds to send this image over a telephone line at 28,800 bits/sec. Transmission time will matter when many document images are required for transmission. Power consumption is directly proportional to the transmission time; this is an important aspect in satellite systems.

Document images contain mostly aligned text. Self-similarity in document images appears in characters and lines, which may be repeated many times in the image. Fractal compression is based on the self-similarity in the document images, thus achieving compression results in reducing the transmission time and saving energy required for the transmission. Even though most general images are not ideal fractal images, they could be compressed by fractal image compression with a certain compression ratio.

The fractal image compression algorithm has the ability to detect the existence of local self-similarity in images. To find self-similar portions in an image, the image has to be partitioned into range blocks. For each range block, the encoder searches the image for a domain block that is mostly similar to the range block when applied to some transformation. Fractal image compression encoder gains the compression through the presentation of the image by its transformations. Quadtree partitioning is a way to partition the image automatically. The original image, document image in our case, is partitioned into range blocks. Since, the objective of the fractal image encoder is to find a domain block that can match the appointed range block after been subjected to some transformation; the range block is further partitioned into four sub-blocks if such domain is not found. The encoding process is finished when all the range blocks find their suitable domain matching block with optimal transformation parameters. In Fractal Image Compression, image is approximated as the attractor of a contractive operator called the fractal transform $W$ on the space of images. Representing an image as the unique fixed point of a fractal transform was first introduced by Barnsley and Sloan. Jacquin then came to devise the first practical fractal image encoder. Fractally encoding image yields pleasing results; although it suffers from long encoding time, it has advantages of fast decoding process, high compression ratio relative to other compression techniques at a certain peak signal to noise ratio and the resolution independent property. These advantages make it a very attractive method in the applications of the multimedia.

This paper is organized as follow. Section II gives an overview of the fractal image compression technique using Quadtree. In section III, fundamentals of fractal image compression are briefly introduced, followed by our experimental results and discussion in section IV. Finally, we summarized out our findings in the conclusion section V.

## II. Fractal Image Compression Using Quadtree Partitioning

The basic idea beyond fractal image compression is to partition the original image into non-overlapping image blocks of size kxk pixels, called range blocks. Each range is encoded by finding its best matching image block of size 2kx2k, called domain block, and the best affine transformation. The best matching domain block, for a given range block, is chosen such that under the affine transformation, the domain block is similar as much as possible to the range block. Specifically, for each range block $R$, we search the domain pool to find the domain block $D$ and a transformation $w$ such that $w\left(D\right)$provides the best matching for $R$. The distortion between the transformed domain block and the range block may be computed using the Haussdorff metric [1, 2] for black and white images (binary image) and the root mean square metric (RMSE) metric, equation (1), for grayscale images [3]. Thus, the key point in fractal image compression is to partition the image into small number of blocks that are similar to other image parts under certain transformations. Suitable domain block for a certain range block is found through searching the whole domain pool, which are all the possible overlapped blocks of size 2kx2k that can be extracted from the image with all 8 isometrics from reflection and rotation. A fractal compressed code for a certain range block consists of the parameters describing its affine transformation, the suitable domain address and its isometry. These codes are stored for each range block, and hence compression is achieved.

Many partitioning schemes have been proposed for fractal image compression [3]; among the most widely known partitioning methods are the fixed size range blocks, Quadtree partitioning [4, 5], the horizontal-vertical partitioning and region merging partitioning scheme [6]. In order to obtain high compression ratio, small number of blocks are required, as the size of the fractal code is proportional to the number of blocks of the partition.

Fixed size blocks partitioning scheme provides a compression ratio independent of the image contents, which is inversely proportional to the number of range blocks used to partition the image. Other techniques are hierarchical and built in a top-down fashion. In this paper, we use

Quadtree partitioning scheme [4] using two levels of range block sizes. First the image is partitioned into fixed size 16x16 pixels range blocks. For each range block a control parameter is computed which is proportional to the pixels intensity values variation, i.e. the variance of the intensity values of the range block. A second level partitioning is introduced to those ranges whose control parameter exceeds a certain threshold $\tau$. The second level partitioning divides the 16x16 range block into four 8x8 sub-blocks, for which the searching process is repeated.

According to the mathematical model used to model our image, we determine the transformation shape applied to the domain and the metric used to compare the similarity between the transformed domain block and the range block. For examples:

- If the document image to be compressed is a black and white image, i.e. binary image. We model the image as a compact set in the metric space $H\left(R^2\right)$, which is the space formed of all the nonempty compact subsets of the space $R^2$. In such a model, similarity between image blocks is measured using the Haussdorff measure [1].
- Document images may be modeled by image function $f\left(x,y\right):\square \rightarrow I$ defined over the image support□, where $I$ is an interval containing the pixels intensity values, possibly $I=[0,1]$ for document images containing shading or $I=\{0,1\}$ for black and white document images. A suitable metric for measuring the distance, dissimilarity, between two images models is the root mean square error metric, RMSE, defined as

$$d_{rms}\left(f,g\right)=\sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{M}\left(f\left(i,j\right)-g\left(i,j\right)\right)^2}{M^2}} \tag{1}$$

  where $M^2$ is the image size.

In this paper we investigate comparisons between different document images undergoing fractal image compression. We used document images containing Arabic (computer written and handwritten) and

English texts with different font sizes. Original images used are of size 512x512 pixels, these are shown in figure 3.

Searching the whole domain pool for the domain giving the minimum dissimilarity between $R$ and $w\left(D\right)$is the so called full search problem and is highly computationally intensive. Many researches suggest different searching techniques to reduce time complexity reduction [7]. We used a local neighboring search for each range block [8]. Local neighboring search not only reduces the encoding time, but also increase the compression ratio since fewer bits are required to address the domain block that best match a certain range block.

## III. Fundamentals of Fractal Image Compression

We briefly review the relevant concepts of Iterated Function Systems IFS encoding before explaining the results obtained for document images.

The image to be compressed is partitioned into a set of disjoint blocks, called range blocks. Another set of image blocks, possibly overlapping, called domains is used to approximate each range by means of an affine transformation. In Quadtree partitioning, both range and domains are typically square blocks. Domain blocks, usually with sides twice as long as the ranges, are shrunk by a spatial transformation to fit the range block size. For each range block $R$ , we have to find the domain block $D$ and the affine transformation $w$ parameters that give

$$\min_{\text{All D}}\left\{\min_{w}\left\{d\left(R,w\left(D\right)\right)\right\}\right\} \tag{2}$$

where $d$ is the metric distance used to measure dissimilarity between blocks. The effect of the transformation $w$ depends on the nature of the image model used to represent the image. Images modeled as compact sets, the transformation $w$ is a contraction, reflection and/or flipping followed by shifting. Images modeled as an image function defined over the image support, the transformation $w$ consists of a spatial contraction mapping followed by a pixel intensity transformation. The intensity level transformation is a composition of a contrast scaling and a luminance shift. The minimization problem, equation (2), is solved as a least square problem when the RMSE metric is used [9].

A first step partitioning divides the image into range blocks of size 16x16 pixels. Then a control parameter is computed to each range block, so that if its value is below a given threshold $\tau$ then no further block partitioning is required and this range block is left as 16x16 pixels. When the control parameters exceed the threshold value $\tau$, the range block is further divided into four 8x8 pixels blocks. Experiments and results are done on 512x512 8bits grayscale images.

Once each range in the image partition has been approximated by a domain from the domain pool, the image can be encoded by the IFS code of the affine transformations from domains to ranges. The construction of the original image in the decoder can be accomplished by the iterations of these transformations starting from any arbitrary image $X_o$. The Fixed Point Theorem guarantees the convergence of such iterations if the affine transformations are contractive [1, 10, 3]. The Collage Theorem ensures that the system's fixed point of convergence $W^{\infty}(X_o)$ can be made very close to the original image [1, 10, 3]. Iterations of the fractal IFS obtained from coding a document image is shown in figure 1, starting from an arbitrary image (Lenna image).

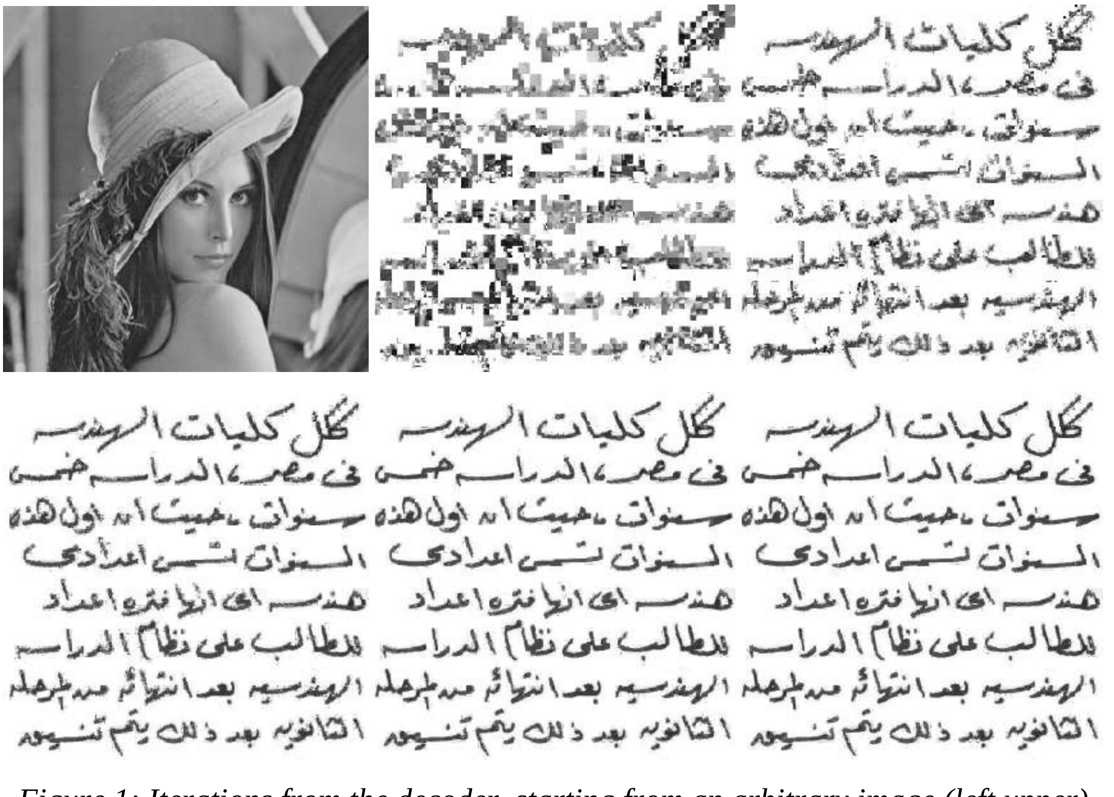


*Figure 1: Iterations from the decoder, starting from an arbitrary image (left upper), converging to our document image (right down). Control parameter threshold =2*

The PSNR versus the iterations for three different initial images (blank, text and Lenna images) shows that the convergence to the fixed point is independent on the initial image, see figure 2.

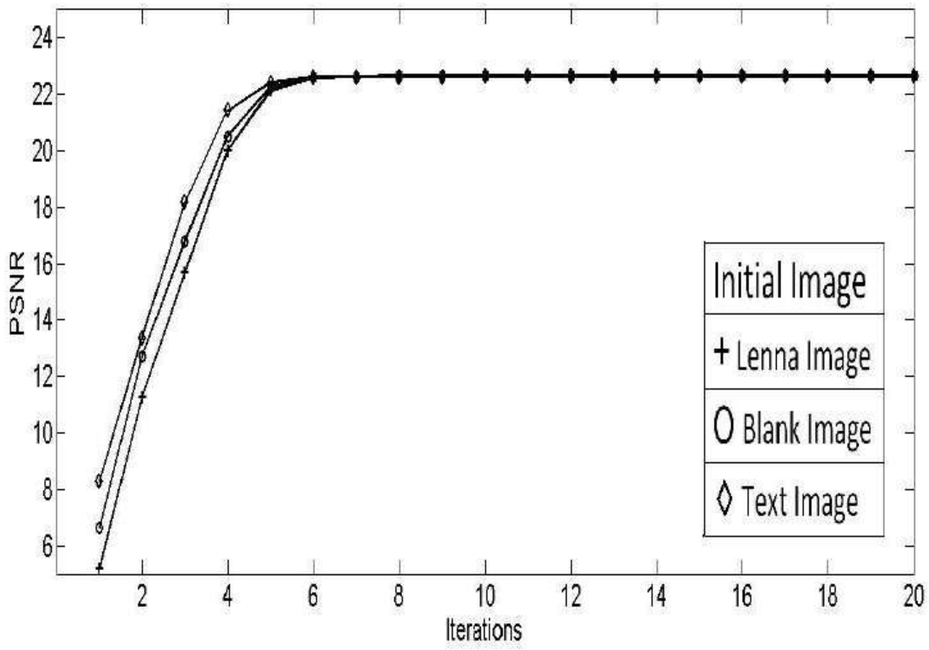
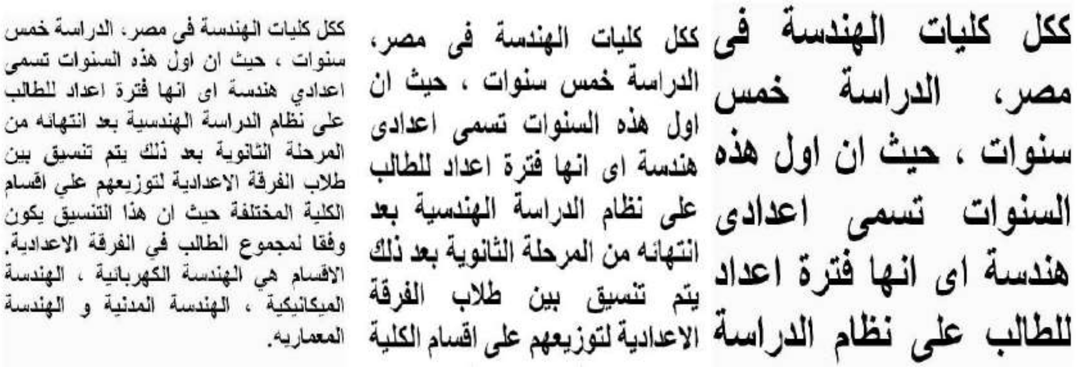


*Figure 2: PSNR versus iterations for 3 different initial images. The original image is document image #5, with a control parameter threshold equals to 2*

## IV. Experimental Results and Discussion

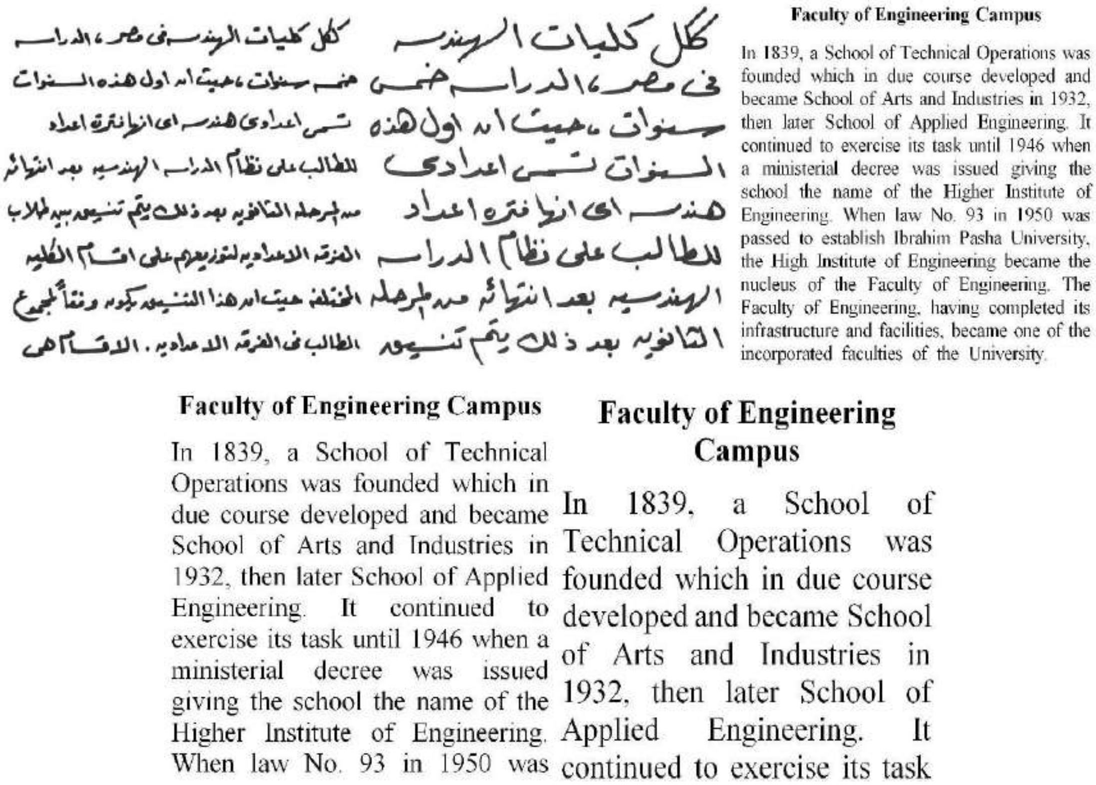
Fractal image compression is applied to some document images shown in figure 3.



ككل كليات الهندسة فى مصر، الدراسة خمس
سنوات ، حيث ان اول هذه السنوات تسمى
اعدادي هندسة اى انها فترة اعداد للطالب
على نظام الدراسة الهندسية بعد انتهائه من
المرحلة الثانوية بعد ذلك يتم تنسيق بين
طلاب الفرقة الاعدادية لتوزيعهم علي اقسام
الكلية المختلفة حيث ان هذا التنسيق يكون
وفقا لمجموع الطالب في الفرقة الاعدادية.
الاقسام هي الهندسة الكهربائية ، الهندسة
الميكانيكية ، الهندسة المدنية و الهندسة
المعماريه.

ككل كليات الهندسة فى مصر،
الدراسة خمس سنوات ، حيث ان
اول هذه السنوات تسمى اعدادى
هندسة اى انها فترة اعداد للطالب
على نظام الدراسة الهندسية بعد
انتهائه من المرحلة الثانوية بعد ذلك
يتم تنسيق بين طلاب الفرقة
الاعدادية لتوزيعهم على اقسام الكلية

ككل كليات الهندسة فى
مصر، الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادى
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة



ككل كليات الهندسه في مصر، الدراسه
خمس سنوات، حيث ان اول هذه السنوات
تسمى اعدادي هندسه اي انها فتره اعداد
للطالب على نظام الدراسه الهندسيه بعد انتهائه
من مرحلة الثانويه بعد ذلك يتم تشعيبه بين طلاب
الفرقه الاعداديه لتوزيعهم على اقسام الكليه
المختلفه حيث ان هذا التشعيب يكون وفقاً لمجموع
الطالب في الفرقه الاعداديه. الاقسام هي

ككل كليات الهندسه
في مصر، الدراسه خمس
سنوات، حيث ان اول هذه
السنوات تسمى اعدادي
هندسه اي انها فتره اعداد
للطالب على نظام الدراسه
الهندسيه بعد انتهائه من مرحلة
الثانويه بعد ذلك يتم تشعيبه

**Faculty of Engineering Campus**

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was passed to establish Ibrahim Pasha University, the High Institute of Engineering became the nucleus of the Faculty of Engineering. The Faculty of Engineering, having completed its infrastructure and facilities, became one of the incorporated faculties of the University.

**Faculty of Engineering Campus**

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was

**Faculty of Engineering Campus**

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task

*Figure 3: Document images numbered from 1 to 8. 1, 2 and 3 are Arabic text computer written, 4and 5 are Arabic handwritten text. Finally, 7 and 8 are English computer written text*

Different results are obtained when the control parameter is changed. If the range block control parameter is less than a certain predetermined threshold $\tau$, the block needs no further partitioning and is encoded as a 16x16 pixels block. When the control parameter exceeds its threshold $\tau$, the range block is divided into four sub-blocks of sizes 8x8 pixels and each is encoded separately. The higher the control parameter threshold $\tau$ is, the less is the probability that the range block is subjected to further partitioning, the higher the compression ratio is (less range blocks will partition the image support) and the lower is the PSNR.

Figure 4 represents the decoder output for the document images when a control parameter is infinity, i.e. fixed size partitioning scheme with size 16x16 pixels. Figures 5, 6 and 7 show results when the control parameter threshold $\tau$ is set to 20, 10 and 2 respectively. Figure 8 shows results obtained with a threshold $\tau$ equals to zero, i.e. fixed size partitioning of size 8x8 pixels.

Compression ratio and PSNR for different document images and different control parameter thresholds are shown in tables 1 and 2.

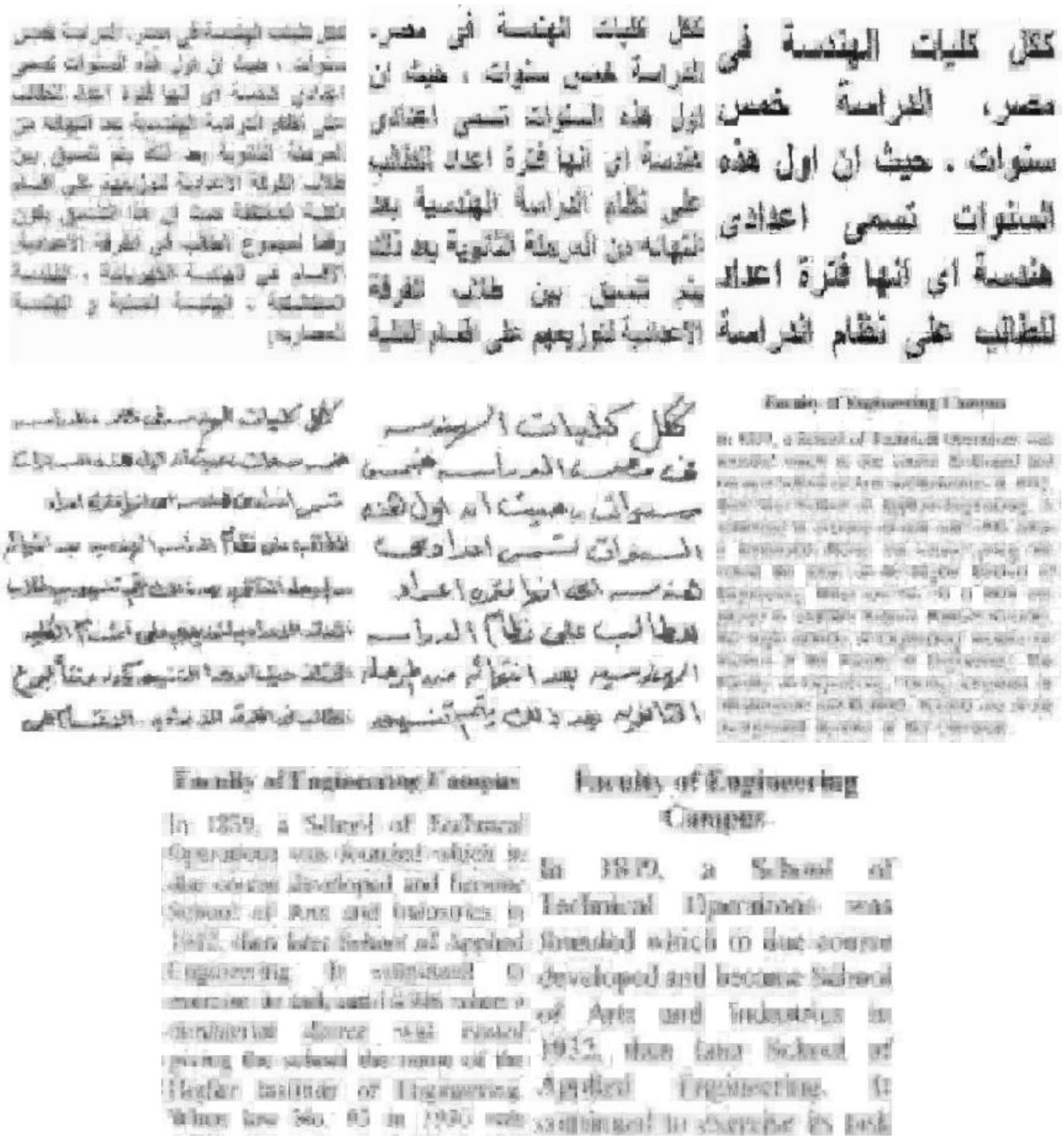
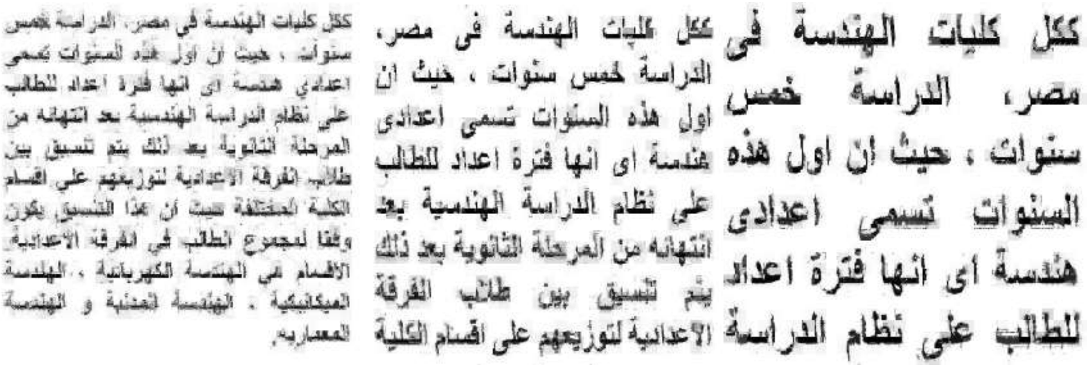
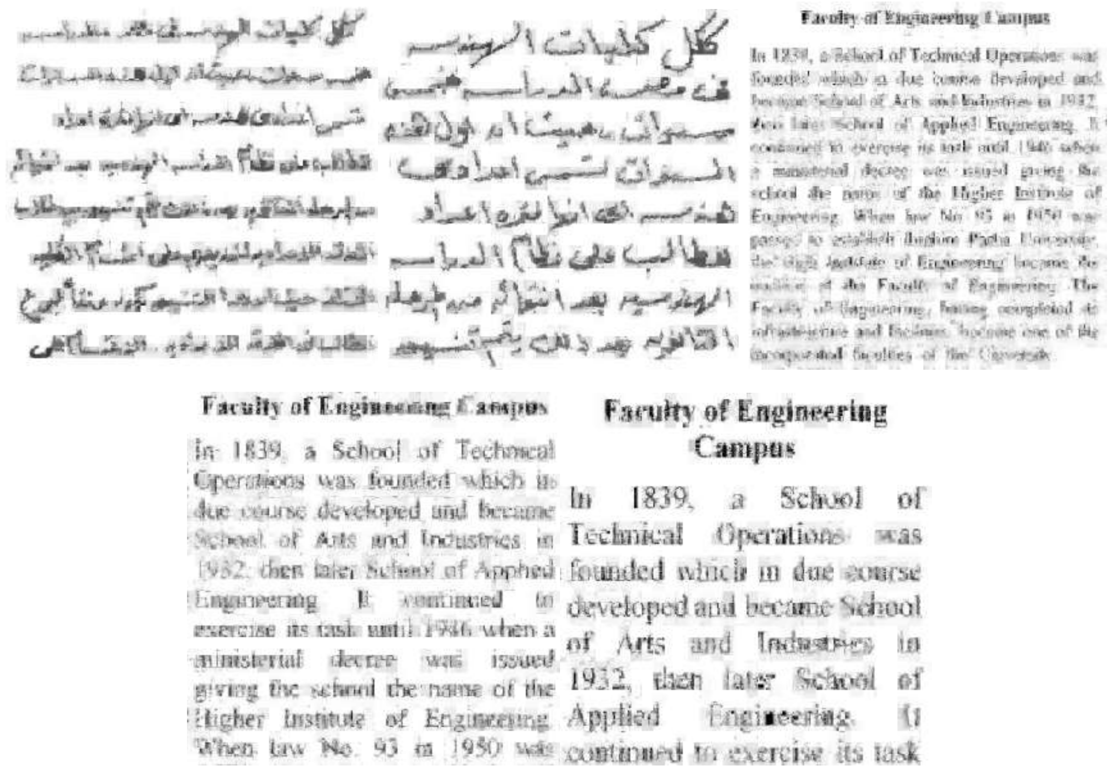


*Figure 4: Decoder output when the control parameter threshold is set to infinity*





ككل كليات الهندسة في مصر الدراسة
خمس سنوات ، حيث ان اول هذه السنوات
تسمى اعدادي هندسة اى انها فترة اعداد
للطالب على نظام الدراسة الهندسية بعد انتهائه
من المرحلة الثانوية بعد ذلك يتم تنسيق بين طلاب
الفرقة الاعدادية لتوزيعهم على اقسام الكلية
المختلفة حيث ان هذا التنسيق يكون وفقا لمجموع
الطلاب في الفرقة الاعدادية . الاقسام هي

ككل كليات الهندسة
في مصر الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادي
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة
الهندسية بعد انتهائه من المرحلة
الثانوية بعد ذلك يتم تنسيق

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was passed to establish Ibrahim Pasha University, the High Institute of Engineering became the nucleus of the Faculty of Engineering. The Faculty of Engineering, having completed its infrastructure and facilities, became one of the incorporated faculties of the University.

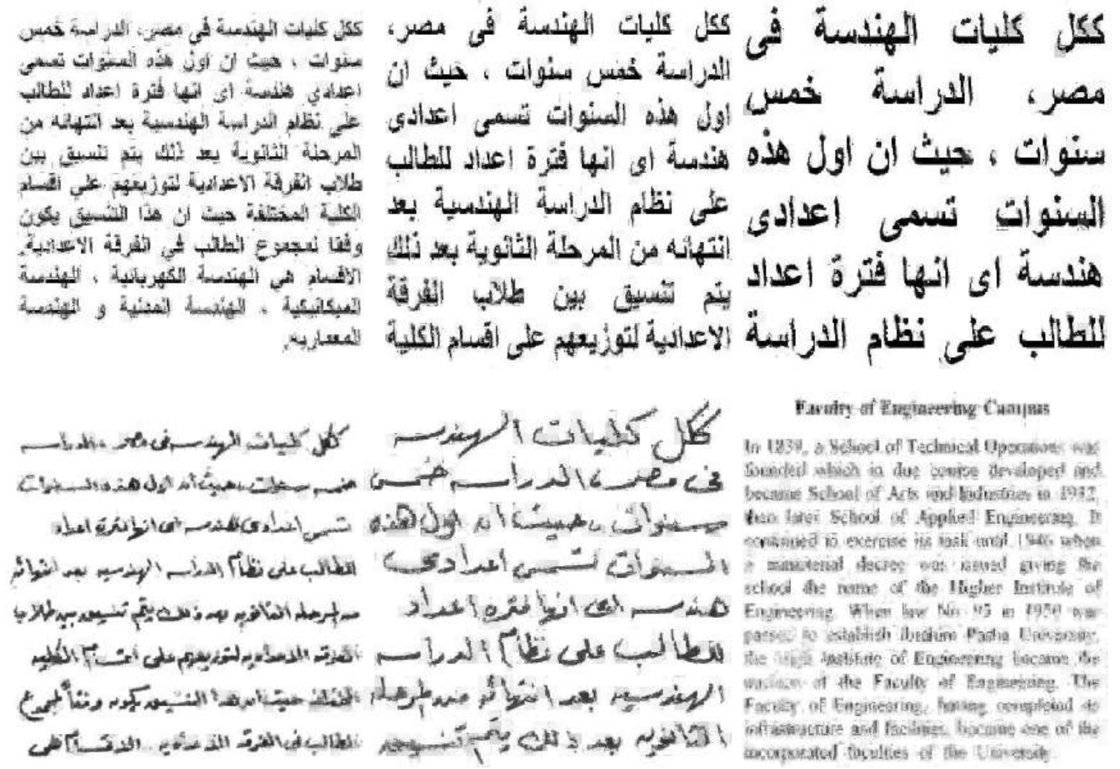
Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task

*Figure 5: Decoder output when the control parameter threshold is set to 20*



ككل كليات الهندسة في مصر، الدراسة خمس
سنوات ، حيث ان اول هذه السنوات تسمى
اعدادي هندسة اى انها فترة اعداد للطالب
على نظام الدراسة الهندسية بعد انتهائه من
المرحلة الثانوية بعد ذلك يتم تنسيق بين
طلاب الفرقة الاعدادية لتوزيعهم على اقسام
الكلية المختلفة حيث ان هذا التنسيق يكون
وفقا لمجموع الطالب في الفرقة الاعدادية.
الاقسام هي الهندسة الكهربائية ، الهندسة
الميكانيكية ، الهندسة المدنية و الهندسة
المعمارية.

ككل كليات الهندسة فى مصر،
الدراسة خمس سنوات ، حيث ان
اول هذه السنوات تسمى اعدادى
هندسة اى انها فترة اعداد للطالب
على نظام الدراسة الهندسية بعد
انتهائه من المرحلة الثانوية بعد ذلك
يتم تنسيق بين طلاب الفرقة
الاعدادية لتوزيعهم على اقسام الكلية

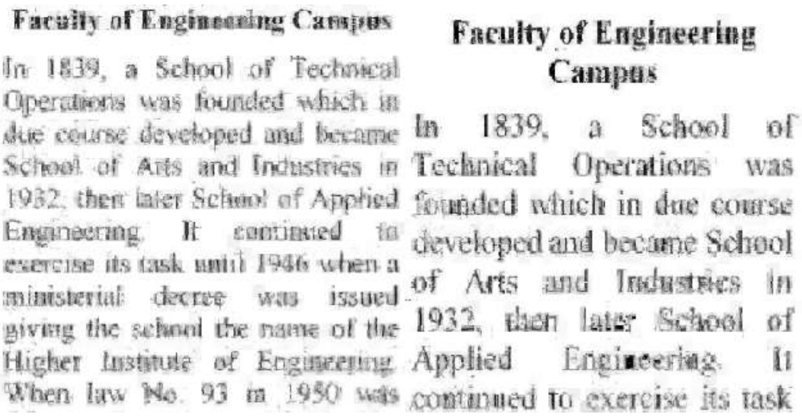
ككل كليات الهندسة فى
مصر، الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادى
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة

ككل كليات الهندسة في مصر الدراسة
خمس سنوات ، حيث ان اول هذه السنوات
تسمى اعدادي هندسة اى انها فترة اعداد
للطالب على نظام الدراسة الهندسية بعد انتهائه
من المرحلة الثانوية بعد ذلك يتم تنسيق بين طلاب
الفرقة الاعدادية لتوزيعهم على اقسام الكلية
المختلفة حيث ان هذا التنسيق يكون وفقا لمجموع
الطلاب في الفرقة الاعدادية . الاقسام هي

ككل كليات الهندسة
في مصر الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادي
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة
الهندسية بعد انتهائه من المرحلة
الثانوية بعد ذلك يتم تنسيق

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was passed to establish Ibrahim Pasha University, the High Institute of Engineering became the nucleus of the Faculty of Engineering. The Faculty of Engineering, having completed its infrastructure and facilities, became one of the incorporated faculties of the University.



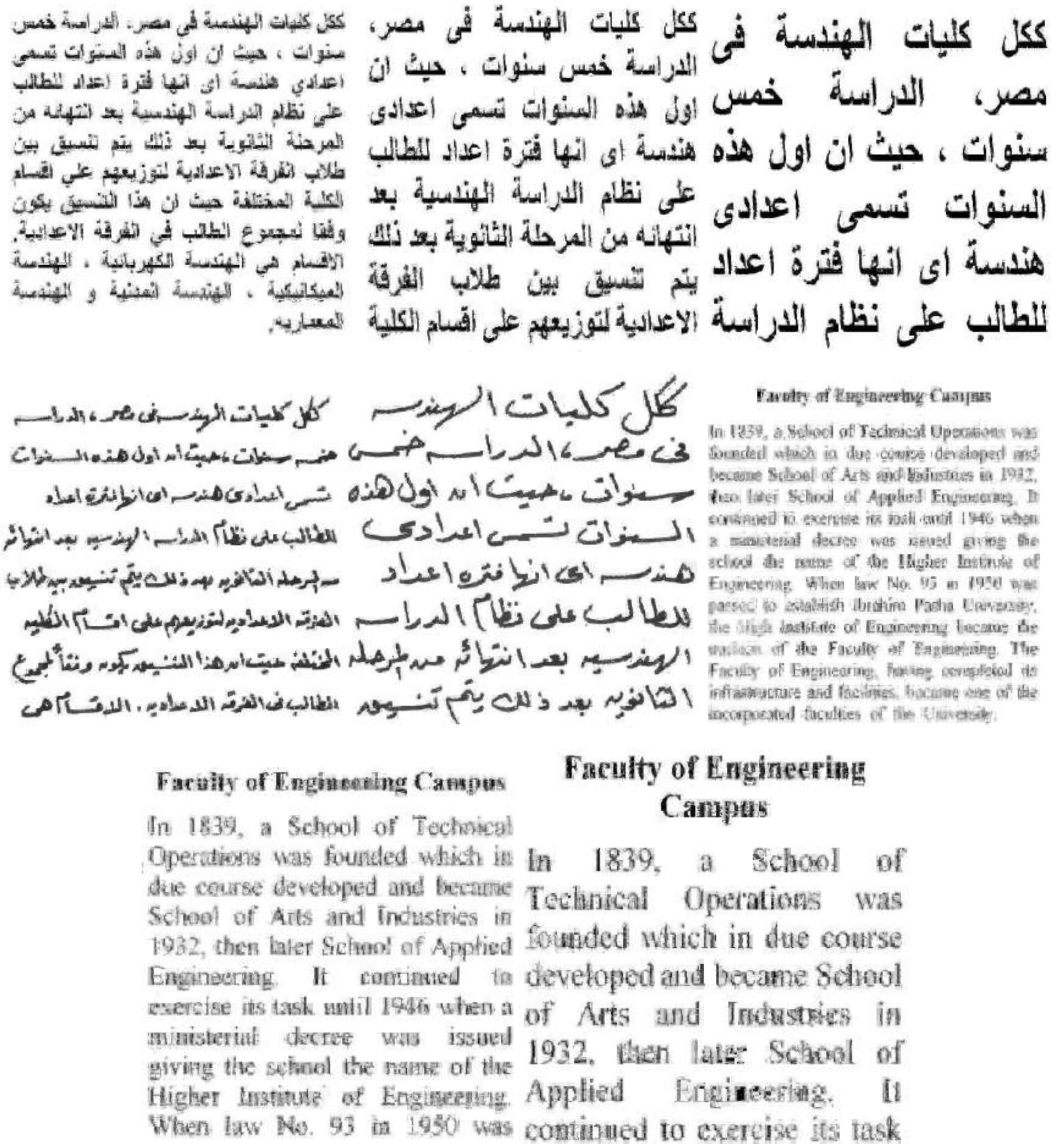
Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task

*Figure 6: Decoder output when the control parameter threshold is set to 10*



ككل كليات الهندسة فى
مصر، الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادى
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة

ككل كليات الهندسة فى مصر،
الدراسة خمس سنوات ، حيث ان
اول هذه السنوات تسمى اعدادى
هندسة اى انها فترة اعداد للطالب
على نظام الدراسة الهندسية بعد
انتهائه من المرحلة الثانوية بعد ذلك
يتم تنسيق بين طلاب الفرقة
الاعدادية لتوزيعهم على اقسام الكلية

ككل كليات الهندسة فى مصر، الدراسة خمس
سنوات ، حيث ان اول هذه السنوات تسمى
اعدادي هندسة اى انها فترة اعداد للطالب
علي نظام الدراسة الهندسية بعد انتهائه من
المرحلة الثانوية بعد ذلك يتم تنسيق بين
طلاب الفرقة الاعدادية لتوزيعهم علي اقسام
الكلية المختلفة حيث ان هذا التنسيق يكون
وفقا لمجموع الطالب في الفرقة الاعدادية.
الاقسام هي الهندسة الكهربائية ، الهندسة
الميكانيكية ، الهندسة المدنية و الهندسة
المعماريه.

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 95 in 1950 was passed to establish Ibrahim Pasha University, the High Institute of Engineering became the nucleus of the Faculty of Engineering. The Faculty of Engineering, having completed its infrastructure and facilities, became one of the incorporated faculties of the University.

ككل كليات الهندسة
فى مصر، الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادى
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة
الهندسية بعد انتهائه من المرحلة
الثانوية بعد ذلك يتم تنسيق

ككل كليات الهندسة فى مصر ، الدراسة
خمس سنوات ، حيث ان اول هذه السنوات
تسمى اعدادى هندسة اى انها فترة اعداد
للطالب على نظام الدراسة الهندسية بعد انتهائه
من المرحلة الثانوية بعد ذلك يتم تنسيق بين طلاب
الفرقة الاعدادية لتوزيعهم على اقسام الكلية
المختلفة حيث ان هذا التنسيق يكون وفقا لمجموع
الطالب فى الفرقة الاعدادية . الاقسام هى

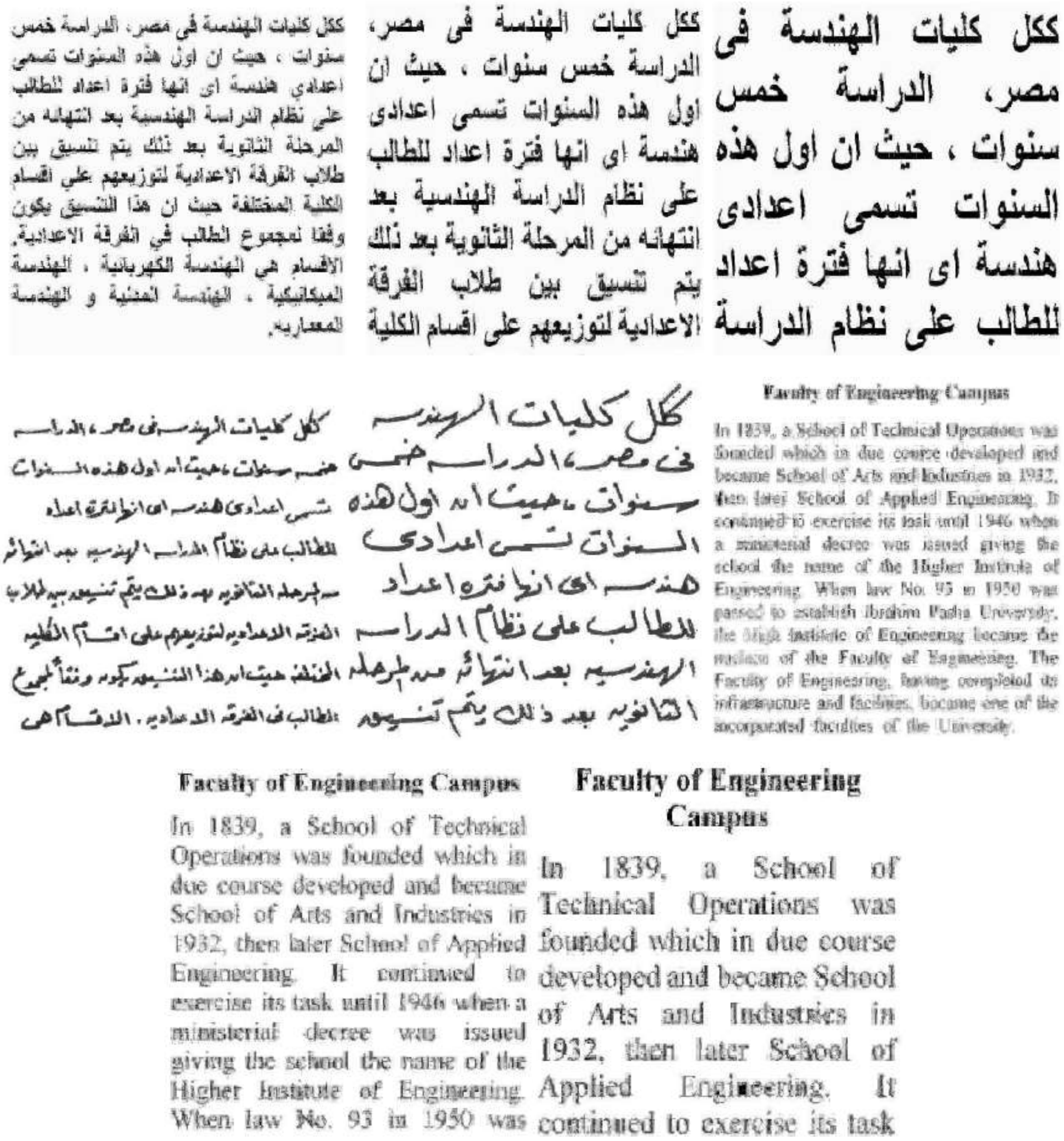
Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task

*Figure 7: Decoded outputs when the control parameter threshold is set to 2*



ككل كليات الهندسة فى مصر، الدراسة خمس
سنوات ، حيث ان اول هذه السنوات تسمى
اعدادي هندسة اى انها فترة اعداد للطالب
علي نظام الدراسة الهندسية بعد انتهائه من
المرحلة الثانوية بعد ذلك يتم تنسيق بين
طلاب الفرقة الاعدادية لتوزيعهم علي اقسام
الكلية المختلفة حيث ان هذا التنسيق يكون
وفقا لمجموع الطالب في الفرقة الاعدادية.
الاقسام هي الهندسة الكهربائية ، الهندسة
الميكانيكية ، الهندسة المدنية و الهندسة
المعماريه.

ككل كليات الهندسة فى مصر،
الدراسة خمس سنوات ، حيث ان
اول هذه السنوات تسمى اعدادى
هندسة اى انها فترة اعداد للطالب
على نظام الدراسة الهندسية بعد
انتهائه من المرحلة الثانوية بعد ذلك
يتم تنسيق بين طلاب الفرقة
الاعدادية لتوزيعهم على اقسام الكلية

ككل كليات الهندسة فى
مصر، الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادى
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة

ككل كليات الهندسة فى مصر ، الدراسة
خمس سنوات ، حيث ان اول هذه السنوات
تسمى اعدادى هندسة اى انها فترة اعداد
للطالب على نظام الدراسة الهندسية بعد انتهائه
من المرحلة الثانوية بعد ذلك يتم تنسيق بين طلاب
الفرقة الاعدادية لتوزيعهم على اقسام الكلية
المختلفة حيث ان هذا التنسيق يكون وفقا لمجموع
الطالب فى الفرقة الاعدادية . الاقسام هى

ككل كليات الهندسة
فى مصر ، الدراسة خمس
سنوات ، حيث ان اول هذه
السنوات تسمى اعدادى
هندسة اى انها فترة اعداد
للطالب على نظام الدراسة
الهندسية بعد انتهائه من المرحلة
الثانوية بعد ذلك يتم تنسيق

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was passed to establish Ibrahim Pasha University, the High Institute of Engineering became the nucleus of the Faculty of Engineering. The Faculty of Engineering, having completed its infrastructure and facilities, became one of the incorporated faculties of the University.

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task until 1946 when a ministerial decree was issued giving the school the name of the Higher Institute of Engineering. When law No. 93 in 1950 was

Faculty of Engineering Campus

In 1839, a School of Technical Operations was founded which in due course developed and became School of Arts and Industries in 1932, then later School of Applied Engineering. It continued to exercise its task

*Figure 8: Decoder output when the control parameter is set to zero, i.e. fixed size partitioning with 8x8 pixels range blocks*

*Table 1: Compression ratio for our document images for different $\tau$*

| | | Document Image # | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Control Parameter Threshold | ∞ | 62 | 62 | 62 | 62 | 62 | 62 | 62 | 62 |
| | 20 | 31.34 | 32.725 | 39.291 | 55.429 | 56.687 | 25.86 | 26.726 | 29.444 |
| | 10 | 24.77 | 25.86 | 28.932 | 33.776 | 37.101 | 18.854 | 18.854 | 24.407 |
| | 2 | 20.188 | 22.408 | 23.627 | 23.91 | 22.662 | 16.681 | 18.437 | 21.106 |
| | 0 | 14.63 | 14.63 | 14.63 | 14.63 | 14.63 | 14.63 | 14.63 | 14.63 |

*Table 2: PSNR for our document images for different control parameter threshold*

| | | Document Image # | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Control Parameter Threshold | $\infty$ | 13.322 | 13,665 | 15.05 | 15.736 | 16.633 | 11.537 | 11.99 | 12.617 |
| | 20 | 15.589 | 16.304 | 17.152 | 15.761 | 16.632 | 12.505 | 13.224 | 13.989 |
| | 10 | 16.378 | 17.896 | 19.982 | 18.58 | 18.737 | 12.903 | 13.701 | 14.65 |
| | 2 | 16.716 | 18.173 | 21.05 | 21.121 | 22.642 | 13.089 | 13.856 | 14.814 |
| | 0 | 16.894 | 18.369 | 21.212 | 21.412 | 22.964 | 13.205 | 14.005 | 14.922 |

## V. Conclusion

In this paper, a comparison is carried out between different document images subjected to fractal image compression. Experimental results on these document images show that fractal image compression can be useful for document images compression for specific threshold control parameter $\tau$ that decide the partitioning block sizes. Such threshold might be well chosen according to the font of the text in the document images. As shown in our experimental results, large fonts document images can be compressed at a relatively high compression ration as they only require high control parameter threshold without a great loss in the readability of the text. On the other hand, small font document images require smaller threshold control parameter, thus smaller compression ratio, to preserve the readability of their texts. Thus, Fractal image compression can be used with document images to reduce their transmission time, their transmission power consumption and the cost of the connection time. Although a significant loss in the image quality is observed, it can be compensated at the receiver through using OCR and text recognition algorithms.

## VI. References

[1] M.F. Barnsley, *Fractals Everywhere,* Academic Press, New York, 1992.

[2] J.Hutchinson, *Introduction to Mathematical Analysis,* 1995.

[3] Y. Fisher, *Fractal Image Compression: Theory and Application,* Springer-Verlag , New York, 1995.

[4] Farhadi G., "*An enhanced fractal image compression based on Quadtree partition* ", the 3rd IEEE International Symposium on Image and Signal Processing and Analysis Volume 1, 18-20 Sept. 2003 Pages: 213-216 vol.1.

[5] Cangju Xing, Yuan Ren and Xuebin Li "*A Hierarchical Classification Matching Scheme for Fractal Image Compression* ", Congress on Image and Signal Processing, 2008 (CISP '08), Volume 1, 27-30 May 2008, Pages:283-286.

[6] Hartenstein H., Ruhl M. and Saupe D., "*Region-based fractal image compression* ", IEEE Transactions on Image Processing Volume 9, Issue 7, July 2000 Pages: 1171-1184.

[7] Polvere M. and Nappi M., "*Speed-up in fractal image coding: comparison of methods* ", IEEE Transactions on Image Processing, Volume 9, Issue 6, June 2000 Pages: 1002-1009.

[8] Thao, N.T., "*Local search fractal image compression for fast integrated implementation* ", IEEE International Symposium on Circuits and Systems Volume 2, 9-12 June 1997 Pages: 1333-1336.

[9] Prasad V.R., Vaddella, Babu R. and Inampudi, "*Adaptive Gray Level Difference to Speed Up Fractal Image Compression*", International Conference on Signal Processing, Communications and Networking 2007 (ICSCN '07), 22-24 Feb. 2007, Pages: 253-258.

[10] M.F. Barnsley and L.P. Hurd, *Fractal Image Compression,* A.K. Peters, Wellesley, MA, 1993.

[11] Rikus Muller, "*A Study of Image Compression Techniques, with Specific Focus on Weighted Finite Automata",* M.Sc thesis, Stellenbosch University, South Africa, December 2005.