الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

الجمعية المصرية لهندسة اللغة

المؤتمر الرابع عشر لهندسة اللغة

مجلد الأوراق البحثية

3-4 ديسمبر 2014

القاهرة - جمهورية مصر العربية

الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

The Egyptian Society of Language Engineering

The Fourteenth Conference on Language Engineering
ESOLEC' 2014

PROCEEDINGS

December 3, 4 -2014

Cairo, Egypt

Designing a Hybrid Approach for Answer Extraction for Factoid Questions

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

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

[1]mahmoud.a.wahdan@gmail.com

[1]safia_abbas@yahoo.com

[3]mostafa.m.aref@gmail.com

***Abstract*—Answer Extraction is very important component in Statistical Question Answering systems. The goal of Answer Extraction is to generate a list of answer candidates to be ranked in the Answer Selection component. We designed hybrid approach architecture to increase both precision and binary recall of the answer candidates list without hurting the end-to-end system accuracy. The proposed approach is the top of the fruits of three precision-oriented approaches and one recall-oriented approach. We delivered a robust design that can be applied to other languages by replacing the current language components with targeted language components.**

1 INTRODUCTION

A common architecture of Question Answering (QA) system consists of three main components. These main components are: 1) Question Processing Component whose heart is the question analysisand query generation components, 2) the Information Retrieval also called Document Processing or Search component, and 3) the Answer Processing Component whose heart is the Answer Extraction and Selection. Figure 1 shows the pipelined architecture and the interaction of these components.

The Question Processing identifies the focus of the question, classifies the question, gets the set of keywords and terms, derives the expected answer type, and generates semantically equivalent formulations of the question. The formulation set beside a set of keywords, terms are used to generate both expanded and non-expanded queries against the targeted information retrieval system.

The Search component uses the generated queries from the previous step to retrieve collections of related documents and passages to the queries. QA uses corpus of documents, web corpus, web search engines, knowledge-bases and/or triple storages in the Search phase.

The Answer processing is the final component in question answering system. Both answer extraction and selection is the effective factor on question answering system for the final results. It is also the most important component because it is responsible for delivering the right precise answer to the user.

Answer Extraction is a key component in any Question Answering system and it is the focus of this paper. Answer Extraction is the process of extracting candidate answers from documents and passages retrieved by the search component. These passages and documents may be web documents, Wikipedia documents, Newswire documents, knowledge peace and/or triples. In other words, Answer Extraction is the process of candidate generation [4]. The result of Answer Extraction component will be a set of initially scored candidate answers that will be ranked or classified using Answer Selection component.

2 ANSWER EXTRACTION APPROACHES

There are two types of approaches based on answer type ontology, in another way – the targeted set of questions to be answered.

A. Answer Extraction independent on type ontology

This approach is very useful in domains where type-based answer extraction and reliable type identification is impossible. The scenario where type-independent answer extraction fits is introduced by [6] in developing IBM Watson’s QA system that compete against humans in Jeopardy! quiz. [3, 4] proposed a type-independent approach for search and candidate generation used in Watson’s QA. They exploited the unique nature of Wikipedia documents and its metadata. The goal of this approach is to increase the recall of answer extraction and eliminate the usage of answer type ontology. To apply this idea, they began with the search and applied a three-pronged search strategy (Document Search, Title in Clue (TIC) and Passage Search).

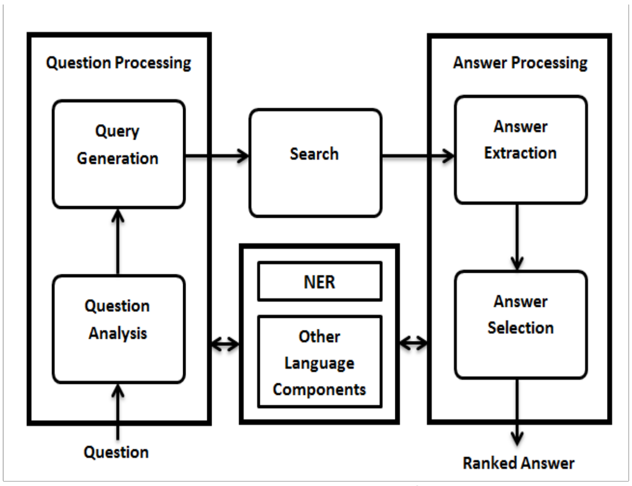


Figure 1: Common Architecture of QA systems

The next step was Candidate Generation from the retrieved content using Wikipedia metadata. In [4] they proposed dual strategies (Title of Document candidate generation and Anchor Text candidate generation). The Title of Document candidate generation strategy is effective for retrieved document with title that may be or contain an answer. Anchor Text candidate generation extracts the candidate answers using anchor text and redirects metadata from Wikipedia documents. The main concept they depend on here is salient concepts are often hyperlinked only in its first occurrence in a Wikipedia document. All salient concepts connected to the title of the document contains the retrieved text segment are retrieved. All salient concepts present in the text segment are extracted as candidate answers. This approach increases the binary recall of candidate generation in both Jeopardy! And TREC data sets.

The same authors of [4] completed their work in Watson's QA [3] by using extra approach beside the previous ones. They used the famous Answer Lookup approach. This approach tries to translate natural language into formal machine language (i.e. first order logic or Bayesian logic) then either looked up the answer from a structured knowledge base or use it to help in driving other answer using reasoning for known facts. Watson tries to exploit the capability to extract some useful relations in the question and convert it into a query against structured sources such as DBpedia [7] and the Internet Movie Database (IMDB) [8] and the retrieved results will be candidate answers. This approach increased both binary recall and end-to-end system accuracy.

B. *Answer Extraction dependent on type ontology*

This approach is widely used in QA systems due to the fact that if the system knows the expected answer type, this will help in getting a list of answer candidates with high precision. High precision will make the job easier for the next step, Answer Selection. There are a variety of techniques that depends on type ontology as follows:

1) *Pattern-based Answer Extraction*: Early QA systems used surface text information such as hand-crafting patterns and automatically acquiring surface text patterns [10, 11]. Gonzalez et al. [12] used regular expressions to extract answer candidate from the retrieved passages. It applies on Spanish language and they retrieve answer candidates depending on answer type. CINDI QA [13] extracts candidate answers from pre-defined lexical patterns and they call it templates. It's almost the same approach. There are a few shortcomings of pattern-based approach. Firstly, manually constructed surface patterns usually return a list of candidate answers with good precision and poor recall [9]. Ravichandran et al.[11] tries to address this issue by introducing an automatic learning approach for these patterns, but still got low recall.

These text patterns are collected automatically in an unsupervised fashion using a collection of trivia question and answer pairs as seeds. A seed is a sample containing the question term and the correct answer for a given question category. Consider the same example of the authors: the term is 'Mozart' and the correct answer is '1756', for the question category 'year'. The pair is submitted to a search engine and the top N results containing both terms are used to extract a pattern able to match the question term and answer. Those patterns are then generalized, by swapping the question term and answer by the <NAME> and <ANSWER> tags, respectively. In this example, patterns like:

a. born in <ANSWER> , <NAME>
b. <NAME> was born on <ANSWER> ,

c. <NAME>(<ANSWER> -
d. <NAME>(<ANSWER -)

could be extracted. The process is repeated with other seed pairs, learning thus more patterns and reining the existing ones. After this, new queries are created from the seeds with only the question terms (without the answer this time) and the obtained patterns are used to extract the answers. The answers retrieved with the patterns are then compared with the expected answer. The ratio of correctly extracted answers becomes the precision of the pattern that originated such answer. The values found represent a probability of each pattern to find the correct answer and are used later to decide what candidate answers are returned as the correct ones.

The problem of low recall is that it is the main cause of bad performance in many pattern-based approaches [14]. To solve this problem, [15] combine patterns with other statistical methods. They followed the approach described in [11] to extract a set of 22,353 patterns, and then they used a maximum-entropy classifier on a training set that has 4,900 questions to learn the appropriate weights of these patterns.

Another problem discussed by [11] is that surface patterns cannot capture long-distance dependencies. This problem can be solved by approach proposed by [16]. The approach aims to recover syntactic structures in the answer sentences, and enhance the patterns with such linguistic constructs.

2) *Named Entity-based Answer Extraction:*Almost all question answering systems uses named entity (NE) to extract candidate answers. The idea is that factoid questions can be classified into several distinctive types, such as “location”, “date”, “person”, etc. [9]. Let’s assume that we can recognize the question type correctly, then the potential answer candidates can be limited down to a few NE types that correspond to the question type. This will increase the precision of answer candidates list and make it easier for answer selection phase.

For example, if the question is asking for a date, then an answer string that is identified to be a location type named-entity is not likely to be the correct answer and will be discarded from the very beginning. However, it is important to bear in mind that neither question type classification nor NE recognition is perfect in the real world. They still have a lot of work because NE classifiers still work on three to seven classes, although question type ontology at least consists of 50 classes. Therefore, although systems can benefit from having fewer answer candidates to consider, using question type and named-entity to rule out answer candidates deterministically [17 - 18] can be harmful when classification and recognition errors occur.

3) *N-grams-based Answer Extraction:* This approach is to get N-grams from retrieved sentences from IR as a candidate answers. Deepak et al. 2003 [5] parse each of these sentences and get a set of chunks, where each chunk is a node of the parse tree. Each chunk is viewed as a candidate answer. They restrict the number of potential answers to be at most 5000.

Aranea [1, 2] generates all n-grams of terms, from unigrams to tetra-grams, from retrieved passages. These n-grams have an initial score, depending from which query they are from, and are considered the candidate answers.

The aforementioned type ontology-dependent approaches – pattern-based answer extraction and named entity-based answer extraction – both come from the answer sentence side. The only information we have extracted from the question side is the question type, which is used for selecting patterns and NE types for matching.

3 SYSTEM ARCHITECTURE

We are motivated to get the advantages of both main approaches form the literature to maximize the binary recall and don’t hurt the precision since we are focusing on factoid questions. It’s very important for Answer Extraction and the whole QA system to include the correct answer among answer candidates without add more noise that will affect the Answer Selection badly. Being large, candidate answers list will make Answer Selection fail to identify the correct final answer. So, we proposed a new hybrid Answer Extraction component as shown in Figure 2.

A. Knowledge-based Answer Extraction component

As we mentioned later, early QA systems tries to parse a natural language question into useful semantic representation that the machine can understand. Although this task is very old, it stills an open problem and no one provide a generic solution. This approach has high precision and very low recall but still can find right answers for some question while other approaches can’t. The parsed question will be translated into structured query representation against some knowledge-bases like DBpedia and IMDB. To construct such structured queries, a semantic relation like first-order logic should be extracted as a part of Question Analysis component. In our architecture, we used a pattern-based approach to detect some relations for frequent question types with the aid of our hybrid Named Entity Recognition (NER) component. This logic will be translated into a query written in a query language supported by the knowledge-base. (i.e. SPARQL). For example: consider the question “*What is the capital of Syria?*” (TREC 11, question number 1447) the question could be interpreted into the form “<Syria><capital> ?result” where “<Syria>” and “<capital>” are the resource link for Syria entity and capital property in the knowledgebase respectively and “?result” is the expected list of search results. The Answer Extraction will use this list and add it to candidate answers set. An example of a question that this component succeed to retrieve the correct answer among other candidate answers while other approaches fail is “*What lays blue* *eggs?*” (TREC 11, question number 1410). The question could be interpreted into the form “?resultEntity<eggcolor><blue>” where “<eggcolor>” is the property name, “<blue>” is a label represents the value of that property and “?resultEntity” is the expected list of search results which will contain entities that can lay blue eggs.

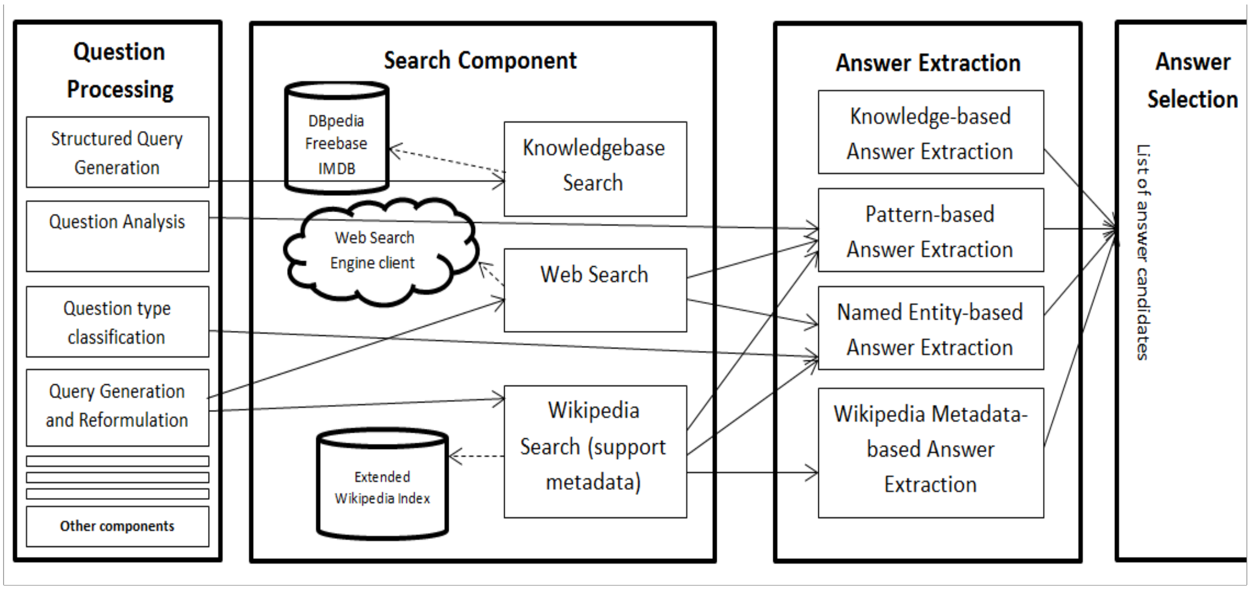

Question Processing
Structured Query Generation
Question Analysis
Question type classification
Query Generation and Reformulation
Other components
Search Component
DBpedia
Freebase
IMDB
Knowledgebase Search
Web Search Engine client
Web Search
Extended Wikipedia Index
Wikipedia Search (support metadata)
Answer Extraction
Knowledge-based Answer Extraction
Pattern-based Answer Extraction
Named Entity-based Answer Extraction
Wikipedia Metadata-based Answer Extraction
Answer Selection
List of answer candidates

Figure 2: Hybrid Answer Extraction and its dependencies system architecture

B. *Pattern-based Answer Extraction component*

Our pattern-based approach is a semi-supervised learning approach and flows from Question Analysis component to Answer Extraction component. The learning is done in offline step. We will discuss this approach via an example to make everything clear. Consider the question “*What is the capital city of New Zealand?*” (TREC 11, question number 1530) is our example. We used a set of questions and there answers from TREC to train the system. We expanded this set manually to provide more generalization for the form that the question can be asked in and provide some synonyms of properties and entities appeared in the original question.

For the above example, a valid set of expanded question could be {“*What is the capital of New Zealand?*”, “*What is New Zealand’s capital?*”, “*What city is the capital of New Zealand?*”}. We annotate these questions with the aid of our hybrid NER as follows: The Question Analysis component detects entity, property and zero or more context parameters. In our example entity will be “*New Zealand*”, property will be “*capital*” and there are no context parameters. It will interpreted to be “*What is the <PROPERTY> of <ENTITY>?*”. This analysis will be used in query generation. The generated queries will also contain the correct answer (ex: “*New Zealand*” “*capital*” “*Wellington*”) and will be submitted to a web search engine to get a set of relative sentences which contains both important question parts and the correct answer.

Consider a sentence “*Wellington, the capital city of New Zealand …*” will be analyzed and the result will be “*<ANSWER>, the <PROPERY> of <ENTITY> …*”. Another sentence could be “*Wellington is the capital city and second most populous urban area of New Zealand*” will be interpreted into“*<ANSWER> is the <PROPERTY> and second most populous urban area of <ENTITY>*”. A third sentence could be “*Wellington - New Zealand’s capital city - is also known ….*” will be interpreted into “*<ANSWER> - <ENTITY>’s <PROPERTY> - is also known ….*”.
We can then learn the pattern(s) from these interpreted forms by unifying them in one or more generic regular expression. It should be:

“<ANSWER>\s*(,|-|)\s*(is)?\s*the\s+<PROPERTY>\s+(.*?)\s*of<ENTITY>”
“<ANSWER>\s*(,|-|)\s*<ENTITY>’s\s*<PROPERTY>”

Then, in Answer Extraction we can extract <ANSWER> to be candidate answer.
Because this approach requires manual work and it’s too hard to figure out a lot of possible questions, it has a high precision and low binary recall.

C. *Named Entity-based Answer Extraction component*

Named Entity-based answer selection is used in majority of QA systems. It depends heavily on two components; Question type classifier and NER. Consider the question: “*When is Mexico's independence?*” (TREC 11, question number 1820). The question type classifier will classify this question to be “DATE” and this will help in formulating the appropriate query. Some search results could be “*Independence Day is a Mexican holiday to celebrate the cry of independence on September 16, 1810, which staked a revolt against the Spaniards.*”. Here, the NER will detect “*September 16, 1810*” as a DATE entity and the named entity-based component will count it as an answer candidate and will discard all other named entities like “*Independence Day*”, “*Mexican*” and “*Spaniards*”.

Another example is: “*Where are the British Crown jewels kept?*”. The question type classifier will classifies this question as of type “LOCATION”. A search result could be “*The Crown Jewels have been kept at the Tower of London since 1303 after they were stolen from Westminster Abbey.*”. A possible candidate answers could be “*Tower of London*” and “*Westminster Abbey*” because they are of type “LOCATION” and other named entities from other types like “*The Crown Jewels*” and “*1303*” will be discarded.

It’s now clear that this approach is highly precise but has low binary recall because question type classifier and NER are not perfect. Although, question type classifier accuracy is above 90%, NER accuracy still not that big especially in open domain. Beside if the question type classifier messes up in detecting the right question type, the problem will propagate in the whole system with a wider spectrum. For example: if the question originally asks for “PERSON” and misclassified to be of “LOCATION”type, then Named Entity-based component will mess up and the answer candidates will not contain the right answer.

It will be not valid to consider all named entities in the search result because this will increase the binary recall and harms the precision. We will not consider this as a solution because our hypothesis depends on getting advantages from other approaches to solve the current approach problems. To solve the problem, we consider a hybrid NER to increase the accuracy and both precision and recall of extracting candidate answers based on question type. This hybrid NER combines advantages of statistical classifier, extraction patterns and gazetteers. This will increase the binary recall of Answer Extraction phase with reasonable percentage.

Another solution to solve the decrease of binary recall in both Pattern-based and NamedEntity-based components which are precision oriented approaches is to use a recall oriented approach beside the precision oriented ones as in the next component.

D. Wikipedia Metadata-based Answer Extraction component

This component is responsible for offline indexing of Wikipedia text and its metadata (Extended Wikipedia Index in Figure 2) as in [3, 4]. The indexed metadata is document title, entities and its synonyms gathered by entities links redirects. This approach based on heuristics that the encyclopaedias’ title may contain the answer of some questions because these encyclopaedias are much organized than other documents. Another heuristic is most of questions’ answers are entities while Wikipedia have annotated entities (anchor text) in its documents body and also appears in the context of the question asked. The final heuristic is entities redirects are a natural source of how entities can be called in many contexts. All these heuristics are covered in Wikipedia. After submitting the query to Extended Wikipedia search, beside the retrieved passages all salient concepts (entities) connected to the title of the document contains the retrieved text segment are retrieved.All salient concepts present in the text segment are extracted as candidate answers. All the retrieved concepts are considered candidate answers. This component extracts candidate answers independent on the question type and so, it will enhance binary recall in the questions where question type classifier fails to get the correct type of. This approach proved an increase in binary recall on non-numbered factoid questions dataset from TREC 11 and 12.

4 CONCLUSIONS

We proposed a robust architecture for a hybrid Answer Extraction component of a QA system. The design is robust and not coupled to a single system. It can be applied to other language by replacing language components. It exploited the unique nature of each of the used approaches to build all-in-one approach that increase both precision and binary recall. It can be used in domains that the question types are easy to be extracted from the questions and in other domains that lack question type information. The hybrid approach depends heavily on other components like hybrid NER, search and question analysis components. The question interpretation can be enhanced in the future to convert more questions into structured queries. Also, Pattern-based Answer Extraction component could be enhanced in the future to cover more patterns without hurting precision.

REFERENCES

[1]Jimmy Lin, "An exploration of the principles underlying redundancy-based factoid question answering," *Journal of ACM Transactions on Information Systems (TOIS)*,vol. 25, no. 6, 2007.

[2]Jimmy Lin and Boris Katz, "Question answering from the web using knowledge annotation and knowledge mining techniques," in *Proc. of the twelfth international conference on Information and knowledge management*, pp.116-123, 2003.

[3]Jennifer Chu-Carroll, James Fan, BranimirBoguraev, David Carmel, DafnaSheinwald and Chris Welty, "Finding needles in the haystack: Search and candidate generation,"*IBM Journal of Research and Development*, vol. 56, no. 3, pp.300-311, IBM Corp. Riverton, NJ, USA, 2012.

[4] Jennifer Chu-Carroll and James Fan, " Leveraging Wikipedia Characteristics for Search and Candidate Generation in Question Answering," in *Proc. of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[5]Deepak Ravichandran, Eduard Hovy, and Franz Josef Och, "Statistical QA - Classifier vs. Re-ranker: What's the difference?, " *in Proc. of the ACL 2003 workshop on Multilingual summarization and question answering*, vol. 12, pp.69-75, 2003.

[6] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek,D., Kalyanpur, A., Adam, L., Murdock, J. W.; Nyberg, E; Prager, J.; and Schlaefer, N., "Building Watson: An overview of the DeepQA project," *AI Magazine*, vol. 31, no. 3, 2010.

[7] DBPedia Web Site: http://www.dbpedia.org.

[8] Internet Movie Database (IMDB) Web Site: http://www.imdb.com.

[9] Mengqiu Wang, "A Survey of Answer Extraction Techniques in Factoid Question Answering," *in Proc. of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2006.

[10] Soubbotin, Martin M. and Sergei M. Soubbotin, "Patterns for potential answer expressions as clues to the right answers," *in Proc. of the 10th Text REtrieval Conference (TREC-10)*, 2001.

[11] Ravichandran, Deepak and Eduard Hovy, "Learning surface text patterns for a question answering system," *in Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.41-47, Philadelphia, PA, USA, 2001.

[12] Antonio J. Gonzalez, Alberto T. Valero, Claudia D. Carral, Manuel, and Luis V. Pineda, "INAOE at CLEF 2006: Experiments in Spanish question answering," *In Alessandro Nardi, Carol Peters, and Joe L. Vicedo, editors, Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.

[13] Chedid Haddad and Bipin C. Desai, "Bilingual Question Answering Using CINDI QA at QA@CLEF 2007," *Advances in Multilingual and Multimodal Information RetrievalLecture Notes in Computer Science*, vol. 512, pp.308-315, 2008.

[14] Xu, Jinxi, Ana Licuanan, and Ralph Weischedel, "Trec2003 qa at bbn: Answering definitional questions," *in Proc. of the twelfth Text Retrieval Conferene (TREC 2003)* , 2003.

[15] Ravichandran, Deepak, AbharamIttycheriah, and SalimRoukos, "Automatic derivation of surface text patterns for a maximum entropy based question answering system," *in Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers*, vol. 2, pp.85-87, 2003.

[16] Peng, Fuchun, Ralph Weischedel, Ana Licuanan, and JinxiXu, "Combining deep linguistics analysis and surface pattern learning: A hybrid approach to chinese definitional question answering," *in Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp.307-314, 2005.

[17] Lee, Cheng-Wei, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, and Wen-Lian Hsu, "Asqa: Academia sinica question answering system for ntcir-5 clqa," in *Proc. Of the 5th NTCIR Workshop Meeting (NTCIR-5), 2005*.

[18] Yang, Hui, Hang Cui, Min-Yen Kan, MstislavMaslennikov, Long Qiu, and Tat-Seng Chua, "Qualifier in trec-12 qa main task," *in Proc. of the 12th Text REtrieval Conference (TREC-12)*, 2003.

BIOGRAPHY

Mahmoud A. Wahdan is a Software Engineer at Orange Labs Cairo and a Researcher in the field of NLP and Question Answering systems. He received B.Sc. degree in computer science from Ain Shams University in 2009 and currently a Master student in computer science at Ain Shams University. He worked for the leaders of Arabic NLP; Sakhr Software and RDI and then be the team leader of a small R&D team at Kngine.

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

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

تصميم طريقة مختلطة لإستخراج إجابات مقترحة لأسئلة الحقائق

[1]محمود عبدالرحمن وهدان – [2] صفية عباس – 3 مصطفى عارف

قسم علوم الحاسب – كلية الحاسبات والمعلومات – جامعة عين شمس

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

ملخص:

تعتبر عملية "استخراج الإجابات المقترحة"مهمة جدا فى أنظمة اجابات الأسئلة القائمة على الإحصاء. ويكون الهدف من هذه العملية هو توليد واستخراج قائمة من الاجابات المقترحة أو المرشحة لترتب وتصنف فى المكون التالى وهو "اختيار الإجابة". لقد قمنا بتصميم هيكلية معتمدة على طريقة مختلطة الهدف منها زيادة كلا من الدقة والانضباط (precision) ودقة استدعاء البيانات (binary recall)لتوليد قائمة الاجابات المقترحة بدون الايذاء بدقة نظام اجابات الأسئلة ككل. تعتبر الطريقة المقترحة خلاصة الفائدة من ثلاث طرق موجهة تجاه الدقة و طريقة أخرى موجهة تجاه دقة استدعاء البيانات. قدمنا تصميم قوى ومتين ومن السهل أن يطبق على لغات أخرى عن طريق استبدال المكونات اللغوية المستخدمة بأخرى تخص اللغة المستهدفة.

Solving Ambiguity in Requirements Specification to UML Conversion

Somaia Osama[*1], Safia Abbas[*2], Mostafa Aref[*3]

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

[1] somaia_mail@yahoo.com
[2] safia_abbas@yahoo.com
[3] aref_99@yahoo.com

Abstract-- **Requirements analysis phase is the first step of software building process. This phase formed a Software Requirements Specification document (SRS). SRS document will be the base for development team to build a software application. The document contains sentences and statement that state the functional and nonfunctional requirements. Sentences in SRS document are likely to contain ambiguous words. The ambiguity means that every word could have a different meaning for every person who reads it, in this case the development team. Therefore, it is necessary for a tool that can solve the ambiguity of SRS document. The tool helps the developers to extract a specific need in SRS document into UML diagrams. It can help system analyst in determining and removing the ambiguity of a requirement statement. In this paper, we describe an automated approach to identify and remove ambiguity, which occurs when text is interpreted differently by different readers.**

Key words: Software Requirements Specification, Natural Language Processing, Ambiguity

1 INTRODUCTION

In industrial practice, the requirements documents are written in natural language (NL), and so run the risk of being ambiguous. Ambiguity is a phenomenon essential in natural language. It means the capability of being understood in two or more possible senses or ways. The description of the functionality of the system has to be unambiguous, meaning that it is free of different interpretations. If a description has more possible interpretations, the software developer may interpret an ambiguous word in a way the customer did not mean. This may result in a system that does not meet the requirements of the company. Since ambiguity in the requirements can lead to specifications which do not accurately describe the desired behavior of the system to be developed. For example, if the customer's interpretation of the requirements is not the similar as that of the system's stakeholders, then the system might not be accepted after customer validation.

Stakeholders are often not even aware that there is an ambiguity in a requirement. Each stakeholder gets from reading the requirements an understanding that differs from that of others, without knowing this difference. So, the software developers design and implement a system that does not behave as intended by the users, but the developers honestly believe they have followed the requirements. Berry and his colleagues [1] illustrated the ambiguity phenomenon in requirements documents, and, following common practice in linguistics, classified them into four main categories, depending on whether the source of the ambiguity lies at the level of words (lexical ambiguity), syntax (syntactic ambiguity), semantic interpretation (semantic ambiguity), or the interaction between interpretation and context (pragmatic ambiguity). Previous work on ambiguity in RE tried to address the problem from at least two perspectives.

- Providing users with a restricted NL, tool, or handbook to assist with writing less ambiguously.
- Detecting ambiguity by investigative the text using lexical, syntactic, or semantic information, or with the help of quality metrics.

The remainder of the paper is structured into the following sections: First, Section 2 presents an overview of related work. Second, Section 3 states architecture of approach used to detect ambiguities and removing it. Finally, Sections 4 presents the conclusion of the paper and the future work.

2 RELATED WORK

A few scientists have proposed various approaches to identify and measure the typical ambiguities in NL based software requirements specifications. Different techniques exist to minimize the ambiguity in requirements. These techniques can be applied during the elicitation, specification and validation of requirements. However, decreasing the level of ambiguity in requirements is a labor-intensive activity, and it remains unclear whether investing effort is worthwhile, resolving requirements ambiguities that are likely problematic to requirements engineers. And it used as the basis for the presented tool [1]. Chantree et al. present an interesting approach with a focus on identifying ambiguities that are likely to lead to misunderstandings that deals with the coordination ambiguity [2]. Kiyavitskaya, Zeni, Mich, and Berry did some case studies with prototypes of a proposed tool for identifying ambiguities in NL RSs in an effort to identify requirements for such tools and find the reason of ambiguity. Their approach was to apply ambiguity measures to sentences identified by a parser based tool to try to increase the precision of the tool with respect to reporting genuine ambiguities. The measures are based on using lexical and syntactic proxies for semantic criteria and the WordNet thesaurus. The case studies found that many of what the tool thought was ambiguous were not problematic given the normal knowledge that the analyst user would have about the domain of the specification and that the tool failed to find many of what one analyst who was particularly attuned to finding ambiguities found manually [3].

Wilson et al. defined general quality criteria for RSs and developed an analysis tool ARM (Automated Requirements Management) to assess the structure of a given RS, the structure of the RS's RStats, the vocabulary used to write the RS, and thus to determine if the RS meets the quality criteria. It identifies potential problems, such as ambiguity, inaccuracy, and inconsistency, in natural language specification statements [4]. QuARS (Quality Analyzer of Requirements Specification) is a linguistic language tool based on a quality model for NL requirements specifications. It aims to detect lexical, syntactic, structural, and semantic defects including ambiguities. In QuARS, certain terms or syntactic structures are considered "dangerous" by themselves; for example, use of certain adverbs or syntactic structures are marked as potentially nocuous. The main obstacle to applying this approach in practice is the rather high number of false positives; in fact, there is no analysis of which among the potentially dangerous constructs are likely to really cause interpretation problems to the stakeholders [5].

Gervasi and Zowghi studied the nature of ambiguity in requirements specifications and provided deeper analysis on the causes and effects of different types of ambiguity in the system development process in order to help better understand the role of ambiguity in RE practices. In that work, Gervasi and Zowghi propose a role for the linguistic feature of *markedness* as a predictor of whether any ambiguity is intentional on the part of the writer, or not [6]. The definition of what constitute ambiguity is given *a priori* by describing the metrics apodictically [7]. Boyd et al. describe a controlled natural language to help reduce the degree of lexical ambiguity of requirements specifications. By substituting synonyms or hyponyms with corresponding terms, and thus obtaining a reduced vocabulary. This approach helps with pronominal anaphora in that it reduces the chances for multiple references [8]. Kamsties and his colleagues describe a pattern-driven inspection technique to detect ambiguities in NL requirements; the technique however is essentially human driven, and thus can draw on the knowledge of an expert inspector [9]. Goldin and Berry provide good examples of concept extraction techniques: they analyze occurrences of different terms, and basing on occurrence frequency, extract application specific terms from requirements documents [10]. RequirementsAssistant [11] is able to recognize lexical, syntactic, semantic, and pragmatic ambiguity. For all the ambiguity categories, there are rules defined to detect their instances.

3 Overall System Architecture

We will develop an automated system to detect and remove ambiguities from full text documents. The system architecture is shown in Figure 1. The initial input is a complete requirements document. The output is UML diagrams.

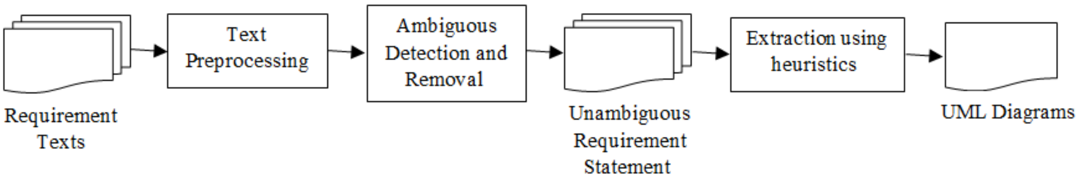

Requirement Texts
Text Preprocessing
Ambiguous Detection and Removal
Unambiguous Requirement Statement
Extraction using heuristics
UML Diagrams

Fig. 1 Overall system architecture

The system consists of three major functional process modules.

(a) **Text Preprocessing Module.** The input requirements document is split into separate sentences using an established sentence boundary detector. The individual sentences are then passed to Tokenizer, the Tagger which identifies the individual words' part of speech, and marks phrase boundaries and the finally syntactic parser.

(b) **Ambiguous Detection and removal Module.**
A tool would apply a set of ambiguity measures to a RS in order to identify potentially ambiguous sentences in the RS. The main goals for the tool for identifying and measuring ambiguities in NL RSs are: to identify which sentences in a NL RS are ambiguous and, for each ambiguous sentence, remove the ambiguity from the sentence, and thus improve the NL RS.

(c) **Extraction using heuristics module:** Finally, This section focuses in heuristics and their application to improve the generation of OO concepts from natural language texts. Usually, candidate classes can be extracted by considering the noun phrases in the requirements text. Candidate relationships can be found in the same way by considering verb phrases, with the UML diagrams being presented to the user as the final step.

4 CONCLUSIONS

Since many requirements documents continue to be written in natural language, we need ways to deal with the ambiguity essential in natural language which have a high risk of misunderstanding among different readers. Our overall research goal is to develop techniques to detect ambiguity in requirements in order to minimize their side effects at the early stages of the software development lifecycle, and removing it. And extract the object oriented information from software requirements specification such as classes, instances and their respective attributes, operations, associations, aggregations, and generalizations. In future work, we will develop prototype of proposed approach.

REFERENCES

[1] Berry, D.M., Kamsties, E., Krieger, M.M.: "From contract drafting to software specification: Linguistic sources of ambiguity," http://se.uwaterloo.ca/˜dberry/handbook/ambiguityHandbook.pdf, 2003.

[2] Chantree, F., Nuseibeh, B., de Roeck, A., Willis, A.: "Identifying nocuous ambiguities in natural language requirements," In: *Proceedings of the 14th IEEE International Requirements Engineering Conference*, Washington, DC, USA, pp. 56–65, 2006.

[3] Kiyavitskaya, N., Zeni, N., Mich, L., Berry, D.M.: "Requirements for tools for ambiguity identification and measurement in natural language requirements specifications," *Requirements Engineering Journal 13*, pp. 207–240, 2008.

[4] Wilson, W.M., Rosenberg, L.H., Hyatt, L.E.: "Automated analysis of requirement specifications," In: *Proceedings of the Nineteenth International Conference on Software Engineering (ICSE 1997)*, pp. 161–171, 1997.

[5] Fabbrini, F., Fusani, M., Gnesi, S., Lami, G.: "The linguistic approach to the natural language requirements, quality: Benefits of the use of an automatic tool," In: *Proceedings of the Twenty- Sixth Annual IEEE Computer Society – NASA GSFC Software Engineering Workshop*, pp. 97–105, 2001.

[6] Gervasi, V., Zowghi, D.: "On the role of ambiguity in RE. In: Requirements Engineering: Foundation for Software Quality," pp. 248–254, 2010.

[7] Mich L, Garigliano R "Ambiguity measures in requirement engineering," In: *Proceedings of international conference on software—theory and practice (ICS2000)*, pp. 39–48, 2000.

[8] Boyd S, Zowghi D, Farroukh A "Measuring the expressiveness of a constrained natural language: An empirical study," In: *Proceedings of the 13th IEEE International Conference on Requirements Engineering (RE'05)*, Washington, DC, pp. 339-352, 2005.

[9] Kamsties E, Berry D, Paech B "Detecting ambiguities in requirements documents using inspections," In: *Proceedings of the First Workshop on Inspection in Software Engineering (WISE'01)*, pp. 68-80, 2001.

[10] Goldin, L., Berry, D.M.: "AbstFinder, a prototype natural language text abstraction finder for use in requirements elicitation," *Automated Software Eng.* 375–4124, 1997.

[11] Driessen, H.: RequirementsAssistant. http://www.requirementsassistant.nl/ 2012.

BIOGRAPHY

Somaia O. Rashad graduated from faculty of the Computer Science in 2009 at Akhabr El Yom Academy, Cairo, Egypt. She started working as a teaching assistant in the Computer Science department at Akhabr El Yom Academy since Sept 2009 till now. Then she got a diploma in software architect from Information Technology Institute, Smart Village, Egypt in 2011.

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

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

حل الغموض في مواصفات المتطلبات للتحويل UML

سمية اسامة، صفية عباس، مصطفى عارف
قسم علوم الحاسب، كلية الحاسبات والمعلومات، جامعة عين شمس

والوثيقة أساسية لفريق التطوير لبناء متطلبات البرامج. تحليل المتطلبات هي الخطوة الأولى فى عملية بناء البرمجيات. شكلت هذه المرحلة وثيقة مواصفات
تطبيقات البرمجيات. تحتوي الوثيقة على المتطلبات الوظيفية وغير الوظيفية و من المحتمل أن تحتوي الوثيقة على كلمات غامضة المعنى. يعني الغموض أن
يكون للكلمة معانى مختلفه تختلف بأختلاف الاشخاص القارئين للوثيقة ، في هذه الحالة فريق التطوير. ولذلك فمن الضروري عمل نهج يمكنه حل الغموض
ويمكن أن تساعد محلل النظم في تحديد وإزالة الغموض. يصف هذا البحث النهج و يساعد المطورين لاستخراج الاحتياجات المحددة على هيئة مخططات
.الآلي العام لتحديد وإزالة الغموض الذي يحدث عندما يتم تفسير النص بشكل مختلف من قبل مختلف القراء

KEYS: A Knowledge Extraction System Based on UNL Knowledge Infrastructure

Sameh Alansary [*1], Magdy Nagi [**2]

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

***Computer and System Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt*
[2]magdy.nagi@bibalex.org

***Abstract*— With the revolution of information available on the internet pages, humans need to extract specific information. This paper presents KEYS (Knowledge Extraction sYStem); an information retrieval and extraction system. It searches for information inside documents represented in UNL, i.e., in semantic hyper-graphs. This allows for retrieval and extraction practices that are language-independent and semantically-oriented. It is expected to provide high-quality knowledge extraction through a shallow analysis of the source text into the Universal Networking Language (UNL) using a specific ontological relations and fully-automatic generation from the resulting UNL document into several different target languages. This is expected to present a novel approach to the topic of identifying the named entity; extracting names with all its types from a natural language form**.

1 INTRODUCTION

Information extraction (IE) is the process of scanning text for information relevant to some interest, including extracting entities, relations, and, most challenging, events{or who did what to whom when and where}. It requires deeper analysis than keyword searches, but its aims fall short of the very hard and long-term problem of text understanding. Information extraction technology arose in response to the need for efficient processing of texts in specialized domains. For example, an information extraction system designed for a terrorism domain might extract the names of perpetrators, victims, physical targets, weapons, dates, and locations of terrorist events. An information extraction system designed for a business domain might extract the names of companies, products, facilities, and financial figures associated with business activities. Full-sentence parsers expended a lot of effort in trying to arrive at parses of long sentences that were not relevant to the domain, or which contained much irrelevant material, thereby increasing the chances for error. Information extraction technology, by contrast, focuses on only the relevant parts of the text and ignores the rest [1].

Message Understanding Conferences (MUC) have described IE as consisting of different tasks. These various tasks differ mainly in their complexity degree and in the depth of the extracted information. For instance, the named entity (NE) task identifying within free text, person, location and organization names, and quantities, such as dates, monetary amounts, etc. Then a more complicated task which is the coreference task (CO) that involves the identification of coreferent entities in text. The template elements (TE) task is responsible for discovering specific attributes about these entities. Next, the relation extraction (RE) task which implies the detection of specific relations (such as employee of, author of, etc.) within the identified entities. Finally, the most complex task which is the scenario template (ST) task in which the system is required to identify instances of a specific predefined event in the text, and extract the information related to each instance of the found event. The system is expected to provide an event template containing various pieces of event information corresponding to each event detected within the given text. Thus, locating the various forms of interesting information embedded in free text is highly complicated [2].

Knowledge extraction has originated from people`s need to obtain and manage the vast amounts of information described in free text more accessibly. Free text contains a multitude of information such as (name of people, places, organizations, roles played by entities in events, relations between entities, etc) that if effectively extracted, can be of great use to many real-world text/web applications, for example, integration of product information from various websites, question answering, contact in formation search, finding the proteins mentioned in a biomedical journal article, and removal of the noisy data [2].

Lehnert and Cowie [3] have summarized some of the early work done in the field of information extraction, they have mentioned the work by DeJong [4], [5] who has analyzed news stories as one of the early attempts in the field of IE. The system is called FRUMP; it is a general purpose NLP system designed to analyze news stories and to generate summaries for users logged into the system. This system is very similar to the current IE systems, since the generated summaries are essentially event templates filled in by FRUMP and presented as single sentence summaries of the events. FRUMP uses hand-coded rules for 17 “prediction” and “substantiation” (the two components of the system) to identify role fillers of 48 different types of events. FRUMP uses a data structure called “sketchy script”, which is a variation of “scripts” that was previously used to represent events or real-world situations described in text [6], [7]. This era of the field of IE has included other approaches as the Prolog-based system by Silva and Dwiggins [8] for identifying information about satellite-flights from multiple text reports. Cowie [9] has also implemented a system, based on Prolog that uses “sketchy syntax” rules to extract information about plants. By segmenting the text into smaller parts, depending on pivotal points, like pronouns, conjunctions, punctuation marks, etc., the system can avoid the need for complex grammars to parse texts. Sager [10] has also developed a system which is applied to highly domain-specific medical diagnostic texts (patient discharge summaries) to extract information into a database for later processing. The system uses English grammar rules to map the text into a structured layout. Zarri`s work is also worth of noting whose goal was to identify information about relationships and meetings of French historical personalities and represent this information in a structured form in the “RESEDA semantic metalanguage” [11]. The system uses rules for semantic parsing and heuristic rules of identifying slot-fillers required by the RESEDA metalanguage. During this period of NLP research, IE has been a field of interest where a fair amount of efforts have been exerted. Much of this work focused on specific domains, used hand-crafted rules and did not have standard data sets or standard evaluation procedures. Defense Advanced Research Projects Agency (DARPA) has organized a series of Message Understanding Conferences (MUC)[12] as a competitive task with standard data and evaluation procedures in the late 1980s and early 1990s. DARPA has also introduced another program towards the end of the MUC era, it was called the TIPSTER program [13], [14], [15]. It was designed to advance the state of the art in text processing. Naturally speaking research in IE has continued to grow over the years since MUC and TIPSTER. Moreover, the definition of IE has also gradually broadened to include many different types of information and tasks that differ in their complexity [14]. Many systems, for example, GE [16], SRI [17], UMass [18], NYU [19], etc. have participated in MUC tasks, which considerably helped in the advancement of IE research.

However, the field of Information Extraction (IE) still includes vast potentials for large-scale knowledge acquisition, since the current systems are still unable to form a coherent theory from a textual corpus which involves representation and learning abilities, although, the current IE systems are able to uncover assertions about individual entities with an increasing level of sophistication and text understanding. Compared to individual relational assertions provided by IE systems, a theory includes coherent knowledge of abstract concepts and the relationships among them. Previous efforts in text-based knowledge acquisition can largely be attributed to the field of Information Extraction (IE), where the task is to recognize entities and relations mentioned within text corpora. Traditional IE systems focused on identifying instances of narrow, pre-specified relations, such as the time and place of events, from small homogeneous corpora. Furthermore, the current IE systems are typically designed for a single domain, there is a lot of interest in building systems that are easily applicable to new domains [20].

The KnowItAll system is considered as an advancement in the field of IE by capturing knowledge in a manner that scaled to the size and diversity of relationships that are present within millions of Web pages. It is a system that aims to automate the tedious process of extracting large collections of facts from the web in an autonomous, domain-independent, and scalable fashion. By learning to label its own training examples using only a set of domain-independent extraction patterns and a bootstrapping procedure, KnowItAll has managed to accomplish this task. KnowItAll is capable of self-supervising its training process; however, the extraction is not fully automatic. KnowItAll requires a user to determine the relation before each extraction cycle for every relation of interest. When acquiring knowledge from corpora as large and varied as the Web, the task of anticipating all relations of interest becomes extremely complicated [20].
Some of the previous information extraction tools can deal with Arabic, Such as Rocket AeroText and NetOwl Extracto. Both of them are capable of discovering entities (people, products, dates, places, and more) and the relationships between them, as well as sentiment analysis in multiple languages. However, both systems are not for free, they were developed as commercial products.

Huge amounts of information in natural language forms exist only in lists of documents and to search all these documents to find just a certain piece of information will be a waste of time. Implementing the Information Extraction techniques in a certain system will with no doubt save a lot of time and efforts while providing precise results. Information Extraction techniques can be used to search various types of documents like historical articles, medical researches and newspapers reports.
Since 1950’s, many research groups have recognized the vital role that the IE plays and started to create projects for tasks like the transformation of a whole encyclopedia to structured forms.
Although these projects have faced some natural language processing problems, modest extraction systems have appeared and have been used in extracting information from a relatively small number of forms.
IE technology still needs mature systems in order to match the human performance.

The IE systems usually support one of two approaches either knowledge engineer approach or automatic training approach. In the knowledge engineer approach, after analyzing huge number of natural language data, the designer identifies sets of common patterns for which he develops rules manually that get interpreted by the components of the IE system. However, using this kind of approach in building the system is considered to be highly time and effort consuming.

In the automatic training approach, there is no need to develop the rules manually, since it depends on implementing a machine learning algorithm in the system which is able to detect and create these rules.
The algorithm must get access to a large number of training texts, these texts have to be annotated manually in order to give the algorithm a sufficient amount of examples which it can learn from and provide the extraction rules [21], [22], [23].

This paper adopts the Universal Networking Language (UNL) framework in building a knowledge extraction system. The aim of UNL is to provide a large collection of semantically annotated texts belonging to different languages. We will present a Knowledge Extraction sYStem named KEYS. It searches for information inside documents that are represented in natural language or UNL expression, i.e., in semantic hyper-graphs. It allows for retrieval and extraction practices that are language-independent and semantically-oriented. With KEYS, we try to start a new fashion in IE by targeting the users aspirations from such application. It is based on a philosophy that is different from the mainstream in the field of IE, since it aims to serve the public which is similar to Google`s goal. Moreover, KEYS originality stems from the fact that it can identify and understand the object depending on its context; it is also able to provide all the suggestions related to this object. KEYS includes SEAN and EUGENE. The former is a shallow enhanced natural language analysis system, it represents natural language texts as semantic networks in the UNL format. While the latter is a natural language generation system, it generates natural language sentences out of semantic networks represented in the UNL format. KEYS is expected to synthesize and normalize the information available on the Web, and to provide summaries extracted out of several different input documents. KEYS has been developed by the Library of Alexandria.

In what follows, section 2 will present the different techniques of information extraction systems. Section 3 sheds light on the project`s history and current status. Section 4 illustrates the basic components of KEYS; the system’s open-source components. First, the language resources (dictionaries and grammars). Second, the software used in building and operating the system (analysis and generation engines). Each of these components is described and their current state is specified. Section 5 will describe KEYS`s interface and illustrate how this system is used. In section 6 KEYS`s output will be evaluated. Finally, section 7 will conclude the paper.

2 THE BASIC TECHNIQUES OF INFORMATION EXTRACTION

The basic techniques are pattern matching, lexical analysis, name recognition, syntactic structure, scenario pattern matching, coreference analysis and event merging. These techniques are divided into two main parts. First, all the individual facts are extracted from the documents, these individual facts are integrated together to form larger facts and translated into the required output format, this stage is called the integration phase. Second, coreference analysis is done and inferences are drawn from the explicitly stated facts in the document. The final output of the information extraction is called a template [24]. The first step consists of developing a set of patterns that matches the various linguistic realizations of the individual facts and these patterns are not just sequences of words, they are more complex than that. To develop such patterns many linguistic processes are required starting from lexical analysis and ending with name recognition. Most of the current systems use partial syntactic analysis just to identify the verbal or nominal constituents in the text. After using these general patterns, task specific patterns are used to identify the facts of interest, which are called scenarios according to the Message Understanding Conference (MUC). The second step includes conference analysis and drawing inferences from the explicitly stated facts in the document. At the end the final output from the information extraction is called template.

The Pattern matching is done through matching the text against a set of regular expressions, when a segment of a text (constituent) is matched with one of these regular expressions, the text segment become a label with one or more assigned features. When there is any semantic feature associated with the constituent, they are called events or entities.

In lexical Analysis phase, first, the text is split into sentences then into tokens. Each token is looked up in the dictionary to assign its features and part of speech.

In the name recognition phase, the different types of names and other special forms like currency and amounts are identified and classified. This simplifies the further processing.

Some systems do not have a separate phase for syntactic analysis, others attempt to build a complete parser of sentences. However, most of the systems fall in between by building a shallow parser. Identifying some of the syntactic structure simplifies the extraction of the information or the knowledge. The argument to be extracted often correspond to noun phrase [24]. After dividing the text into syntactic constituents, each constituent has to be associated with some features like the tense, voice and root of the verb in the verbal constituents and for nominal constituents information are associated to the head of the constituent like its number whether it is a proper name or not and so on.

Then, larger nominal phrases are built up by attaching their modifiers to them and in this case these patterns will have some semantic constraints.

In the scenario pattern matching phase, the main target is to extract the main events of the scenario. Then the coreference analysis phase comes next which includes the task of resolving the anaphoric references by searching for the most recent previously mentioned entry of the same type, for example, person if the anaphora was one of the personal pronouns.

In the event merging phase, all the information about an event is collected which may constitute a hard task, because the information may be spread over many sentences. Another problem that may face the extraction systems when collecting information about a certain event is that its information may be implicit and needs to be more explicit.

3 PROJECT HISTORY AND CURRENT STATUS

KEYS is a rule based Knowledge Extraction system; this system requires different linguistic resources and tools with certain features in its background in order to work efficiently. It requires a dictionary that is enhanced with certain features that encompass all the levels of linguistic information whether it is morphological, semantic or syntactic (will be described in details in section 4. It also requires a grammar that is capable of providing an adequate semantic and syntactic analysis. Moreover, it requires tools that exploit these resources. These tools are called SEAN and EUGNE which have been developed in Bibliotheca Alexandrina (will be described in details in section 4. The linguistic resources were developed using the universal networking language within the UNL framework.

The UNL project has been originally proposed in 1996. The responsible organization is the Universal Networking Digital Language (UNDL) Foundation[1] in Geneva, Switzerland [25], [26], [27], [28] and [29]. UNL is the interlingua employed here; it is capable of representing the meaning of the content of natural language texts in an abstract universal format that is not influenced by any language. UNL aims ultimately to allow people to generate, have access to, information and knowledge, in their own native language by breaking down the language barriers that exclude the majority of people from gaining access to information in their native language. The UNL also assumes that any information conveyed by natural language can be formally and usefully represented by semantic networks (sometimes called UNL expression). In UNL approach; the semantic network must be independent of any natural language in particular (i.e., it must be "universal"). This semantic network is made of three different types of discrete semantic entities: concepts, relations and attributes. Concepts are nodes in the network; relations are arcs linking nodes; and attributes are used to represent information conveyed by natural language grammatical categories (such as tense, mood, aspect, number, etc.) [25] which are a standard set of universally-accessible semantic entities. The semantic network is derived by passing through different stages; tokenization and disambiguation, morphological analysis, syntactic analysis and semantic analysis.

In the UNL framework, the different linguistic levels of analysis are achieved via three types of grammar: N-Grammar, or Normalization Grammar which is a set of rules used to segment the natural language text into sentences and to prepare the input for processing, T-Grammar, or Transformation Grammar which is a set of rules used to transform natural language into UNL or UNL into natural language.

The transformation should be carried out progressively, i.e., through a transitional data structure: the tree, which could be used as an interface between lists and networks. Accordingly, the UNL grammar states seven different types of rules which is divided into two types of grammar; analysis and generation. Three types of rules are common between the two grammars the other four depend on the type of grammar. The seven types of rules are (LL, TT, NN, LT, TL, TN, NT), specified as indicated below:

ANALYSIS (NL-UNL)

- LL - List Processing (list-to-list)
- LT - Surface-Structure Formation (list-to-tree)
- TT - Syntactic Processing (tree-to-tree)
- TN - Deep-Structure Formation (tree-to-network)

[1] The official website of the foundations is available at http://www.undl.org

- NN - Semantic Processing (network-to-network)

GENERATION (UNL-NL)

- NN - Semantic Processing (network-to-network)
- NT - Deep-Structure Formation (network-to-tree)
- TT - Syntactic Processing (tree-to-tree)
- TL - Surface-Structure Formation (tree-to-list)
- LL - List Processing (list-to-list)

Finally, D-Grammar, or Disambiguation Grammar which is a set of rules used to improve the performance of the transformation rules by constraining or forcing their applicability. Grammars are not bidirectional, although they share the same syntax. In the UNLization, the N-Grammar contains the normalization rules for natural analysis, the analysis T-Grammar contains the transformation rules used for natural language analysis and the analysis D-Grammar contains the disambiguation rules used for tokenization as well as for improving the results of the NL-UNL T-Grammar. While in the NLization process, the generation T-Grammar contains the transformation rules used for natural language generation and the generation D-Grammar contains the disambiguation rules used for improving the results of the UNL-NL T-Grammar.

KEYs takes advantage of the UNL approach along with the new trend in NLP applications, that is being an open-source application, because of its vast advantages, opportunities and potentials. A rule-based knowledge extraction system is open source only when the source code of its engines and tools are distributed along with the linguistic data of the extraction pairs. In addition, tools to maintain and develop the linguistic resources so that they can be used with the engines should also be distributed. KEYs fulfills all of the criteria and, hence, can be positively considered an open-source Knowledge extraction system. Moreover, not only its components are open-source, they are also free. The basic components of KEYs, its linguistic resources and tools will be described in details in section 4.

4 THE BASIC COMPONENTS OF KNOWLEDGE EXTRACTION SYSTEM (KEYS)

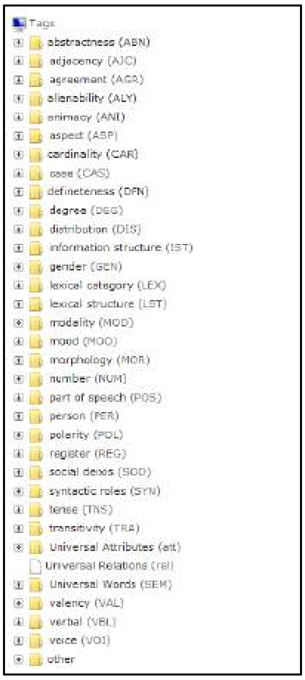
As mentioned before a knowledge extraction system depends on different linguistic resources and tools in order to be able to operate. KEYS depends on three linguistic resources which are dictionary, corpus and grammar. It also depends on two tools called SEAN and EUGENE. All these resources and tools are developed by Bibliotheca Alexandrina. In this section these resources and tools are going to be described in details.

A) Language Resources

1) *Dictionary:* it presents the linguistic information that constitutes the linguistic infrastructure of the dictionary (the UNL dictionary) used by KEYS application. The linguistic information that appears in the UNL dictionary has been assigned to all of the words of the dictionary through UNLarium[2], encompassing the different linguistic levels: morphological information, morpho-syntactic information, syntactic information and semantic information. UNL uses a standard and universal list of features (Tagset) to describe all types of the linguistic information concerning every natural language word. The words are described using a list of features extracted from the UNDL Foundation Tagset. The UNDL Foundation recommends adopting the following tags for some specific and pervasive grammatical phenomena to boost the standardization of the lexical resources used in the UNL framework. The Tagset’s features depending on the structure of the natural language. Several of those linguistic constants have been already proposed in the Data Category Registry (ISO 12620)3, see Fig. 1.

[2] http://www.unlweb.net/unlarium/

[3] http://media.dwds.de/clarin/userguide/text/concepts_ISOcat.xhtml



Tags
- abstractness (ABN)
- adjacency (AJC)
- agreement (AGR)
- alienability (ALY)
- animacy (ANI)
- aspect (ASP)
- cardinality (CAR)
- case (CAS)
- defineteness (DFN)
- degree (DEG)
- distribution (DIS)
- information structure (IST)
- gender (GEN)
- lexical category (LEX)
- lexical structure (LST)
- modality (MOD)
- mood (MOO)
- morphology (MOR)
- number (NUM)
- part of speech (POS)
- person (PER)
- polarity (POL)
- register (REG)
- social deixis (SOD)
- syntactic roles (SYN)
- tense (TNS)
- transitivity (TRA)
- Universal Attributes (att)
- Universal Relations (rel)
- Universal Words (SEM)
- valency (VAL)
- verbal (VBL)
- voice (VOI)
- other

Figure 1: List of tags in alphabetical order

The tagset is providing the technical means for describing any linguistic behavior which should be done in a highly standardized manner, so that others could easily understand and exploit the data for their own benefit. The main intention is to create a harmonized system in order to make language resources as easily understandable and exchangeable. The dictionary is enhanced by morphological information indicating the structure of words, some of this morphological information such as part of speech, lexical structure and the inflections of words.

Part of speech feature: It is used to classify words into main classes and each class may include subclasses. The classes are nouns, verbs, adjective, adposition, adverb, affix, classifier, conjunction, determiner, interjection, numeral, particle and pronoun. The system is designed as such in order to create much flexibility in describing the different types of words. Moreover, the classes are divided into subclasses. For example, the used features in the dictionary differentiate between two types of nouns, common noun such as "صندوق" 'box' - "باب" 'door' – "ورقة" 'paper' and proper noun as "نجيب محفوظ" 'Naguib Mahfouz' - "مصر" 'Egypt' - "اليونسكو" 'UNESCO'.

The dictionary used in the knowledge extraction system differentiates between common and proper nouns and is enhanced with information for the proper names such as the names of rivers, mountains, the names of humans which are considered as public figures (common Arab and non- Arab first and second names).

lexical structure: It is used to classify the words into simple words as the Arabic words "قرأ" 'read' - "مكتب" 'office' - "رائع" 'wonderful', and multiword expressions such as the word "سور الصين العظيم" 'the great wall of China'.

Inflectional paradigms: It is a stored feature that is responsible for generating the different word forms out of the stored lexemes. The dictionary also includes syntactic information that describes the principles and processes by which sentences are constructed. It deals with phrase and sentence formation out of words, such as valency, aspect and subcategorization information. Moreover, the dictionary also is enhanced by information that is concerned with the grammatical categories such as gender, number, person, transitivity, tense, case, voice and mood.

The most important feature concerning building any knowledge extraction system is the semantic classification of the words; the UNL dictionary utilizes a semantic ontology. This ontology classifies the entities existing in the natural world into a semantic hierarchy. This hierarchy points out the particular type of each concept and the kind of relation it indicates with other concepts in the ontology. Each entry in this hierarchy carries a set of features and attributes and all subclasses of this concept inherit the properties of that class. Ontologies are useful in NLP as they play a crucial role in the disambiguation of word senses as well as the understanding of a natural language text by determining the exact sense of a word via its position in the semantic hierarchy. The semantic ontology adopted in the UNL dictionary is the English WordNet 3.0. ontology. In WordNet, English nouns, verbs, adjectives and adverbs are organized into sets of synonymous words (called synsets), each synset representing one distinct concept. For example, the words "coast", "seacoast", "sea-coast" and "seashore" are all synonyms grouped together in a single synset that refers to a unique cognitive concept which is "the shore of a sea or ocean".

Nouns in the WordNet hierarchy are divided into several semantic fields each having a “unique beginner” as the starting node. A unique beginner is a semantic entity that probably has no hypernym and from which nouns that belong to this distinct semantic field can be pulled out. The WordNet employs a set of 25 unique beginners, 8 of which refer to tangible things or “entities”, 5 denote “abstractions” and 3 are “psychological features”. Verbs, modifiers and adverbs are also classified into distinct semantic hierarchies see Fig. 2. For more details about the dictionary and the stored features [30]. Moreover, it is important to mention that any lexical item that is not included in the dictionary will be labeled as “TEMP”.

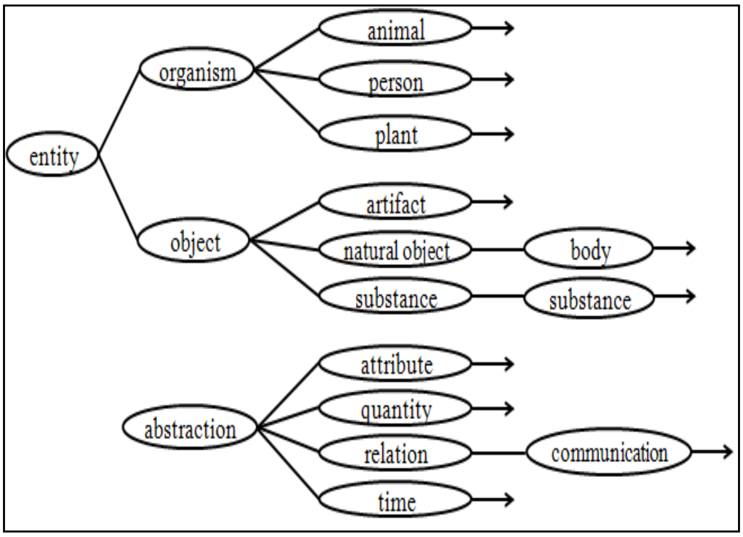
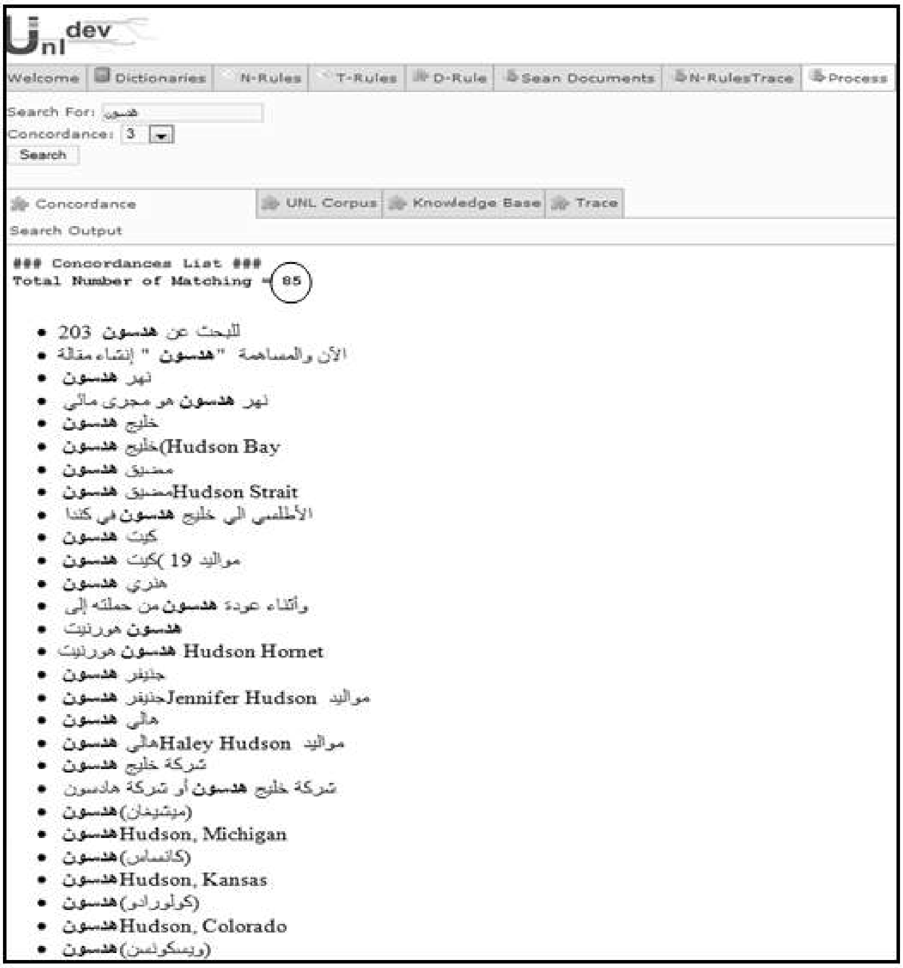


Figure 2: The semantic ontology used in the dictionary

2) *CORPUS:* In order to build an sufficient corpus for proper names, 1000 pages of proper names have been selected from the Wikipedia. These pages represent a rich material for the corpus, since these pages will include these proper names in real contexts. Furthermore, these proper names are from Wikipedia which means that the coverage rate will be high and the corpus will be considered robust. These pages are segmented into sentences using concordance. The total number of occurrences for these proper names is 22,000 with maximum 7 words length; 3 words before and after the proper name. The data is divided into training data which includes 17,000 occurrences and testing data which includes 5,000 occurrences. The testing data will be used later in the evaluation phase. Fig. 3 represents an example for the corpus of the searched word “هدسون” with its 85 instances.



Search For: هدسون
Concordance: 3
Search

Search Output

```
### Concordances List ###
Total Number of Matching = 85
```

- للبحث عن هدسون 203
- الآن والمساهمة "هدسون" إنشاء مقالة
- نهر هدسون
- نهر هدسون هو مجرى مائي
- خليج هدسون
- خليج هدسون(Hudson Bay
- مضيق هدسون
- مضيق هدسونHudson Strait
- الأطلسي الى خليج هدسون في كندا
- كيت هدسون
- مواليد 19)كيت هدسون
- هنري هدسون
- وأثناء عودة هدسون من حملته إلى
- هدسون هورنيت
- هدسون هورنيت Hudson Hornet
- جنيفر هدسون
- مواليد Jennifer Hudsonجنيفر هدسون
- هالي هدسون
- مواليد Haley Hudsonهالي هدسون
- شركة خليج هدسون
- شركة خليج هدسون أو شركة هادسون
- هدسون (ميشيغان)
- هدسونHudson, Michigan
- هدسون (كانساس)
- هدسونHudson, Kansas
- هدسون (كولورادو)
- هدسونHudson, Colorado
- هدسون (ويسكونسن)

Figure 3: Example for corpus

3) GRAMMAR: In KEYS application, the knowledge extraction process has to pass through different stages. Firstly, lexical analysis stage which split sentences into tokens. Each token is looked up in the dictionary to assign its features; part of speech and the other features that are stored in the dictionary. Then a pattern matching stage starts to build ontological relations between the tokens, but if the grammar fails to match the sentences with any pattern, the grammar will try to retrieve the ontological relations using the semantic features that are stored in the dictionary. Finally, if the previous stages fails to figure out the ontological relation between the words the context prediction will take place by predicting the identity of the proper name from the context. However, it is worth mentioning that not all of the grammar levels that are mentioned earlier are applied in this application, since that this application provides a shallow parsing only.

Lexical analysis module

This module is responsible for splitting the sentences into tokens, then matches these tokens with the dictionary in order to assign the different necessary features to each token. However, some words may be misrecognized, because of spelling mistakes or morpho-syntactic changes. For example, the most common mistake in the Arabic writings is /Hamza/ in the initial position as in "استقبل" 'receive'. The rules are able to solve this problem by investigating the morphological pattern of the wrong spelled word by the regular expression technique. For example, if a six-letters word begins with the sequence "/...است/" as in the pattern "استفعل" /?istifʕaal/, the wrong written /Hamza/ ("أ", "إ" or "آ") will be modified to "ا" according to the Arabic grammar as in the rule in (1).

```
(1)
({SHEAD|BLK|PUT|PFX},%e)(TEMP,"/(نا|تم|ن|ت|وا|و|ي|ا|)...ست(إ|آ|أ)/",%x,^Hamza_modified)(%y,{STAIL|BLK|PUT})
:=(%e)("1>"ا,%x,X,Hamza_modified)(%y);
```

The module also deals with cases of morpho-syntactic changes as in the nominative form “علماؤ” /ʕulamaa?u/ when it is attached to the pronoun “ه” ‘its’. Rules are able to extract the deep form “علماء” ‘scientists’ from the surface form “علماؤه” ‘its scientists’ as in rule in (2).

```
(2) ({SHEAD|BLK|PUT|PFX},%e)("/.+(ؤ|ئ)/",^Y,%x)(POD,%w):=(%e)(%x,"ء"<"ؤ","ء"<"ئ",Y)(%w);
```

After the completion of the task of spelling correction, if some words are still undefined, the feature ‘TEMP’ will be assigned to it. Then, it will be considered as a proper name and the ‘PPN’ feature will be assigned to it instead of the feature ‘TEMP’.

After the lexical analysis module, the semantic relations between the words of the sentences using the UNL ontological relations should be established. These relations are stated in table 1 below:

TABLE 1

ONTOLOGICAL RELATIONS IN THE UNL SYSTEM

Tag	**Relation**	**Definition**	**Example**
ant	opposition or concession	Used to indicate that two entities do not share the same meaning or reference. Also used to indicate concession.	John is not Peter = ant(Peter;John)
cnt	content or theme	The object of an stative or experiental verb, or the theme of an entity.	Book about linguistics = cnt(book;linguistics)
icl	hyponymy, is a kind of	Used to refer to a subclass of a class.	Dogs are mammals = icl(mammal;dogs)
iof	is an instance of	Used to refer to an instance or individual element of a class.	John is a human being = iof(human being;John)
nam	name	The name of an entity.	The city of New York = nam(city;New York)
pof	is part of	Used to refer to a part of a whole.	John is part of the family = pof(family;John)
fld	field	Used to indicate the semantic domain of an entity.	sentence (linguistics) = fld(sentence;linguistics)

The following sections discuss the followed techniques to build the ontological relations. Each of the following sub-sections represents an attempt to recognize the identity of the proper names that have occurred in the instances; if one attempt fails to reach the recognition, the following attempt will take place.

Pattern matching

Sometimes the ontological relations mentioned in table 1 would have fixed structures with keywords that are stated in them, these structures would represent the type of the relation as in table 2.

TABLE 2

RELATIONS KEYWORDS AND EXAMPLES FOR THE ONTOLOGICAL RELATIONS

Relation Tag	**Relation key words**	**Example**
ant	عكس – مقابل - يقابل	الحق مقابل الباطل
cnt	عن - حول	كتاب عن التاريخ
icl	نوع / صنف من – أحد أنواع / واحد من –أصناف	القطط نوع من الحيوانات
iof	مثال ل/على	الإسكندرية مثال لبلدان مصر
nam	تسمى ب – اسم - تحت اسم	محمد اسم انسان
pof	جزء من – أحد أجزاء	الإسكندرية جزء من مصر
fld	في مجال – في علم	المورفولوجيا في علم اللغويات

The pattern matching module is responsible for building the ontological relations for structures that have keywords such as those mentioned in table (2). For example, the rule in (3) states that if a noun is followed by “جزء” ‘part’ then “من” ‘of’ and is followed by another noun, then both nouns would be linked with a ‘pof’ relation. However, not all of the relation keywords that are stated in table (2) are found in the corpus, but they are taken into consideration in order to achieve grammar robustness.

```
(3)       (%a,N)("جزء",%b)("من",%b)(N,%d):=(pof(%d,with_rel;%a,rel=pof)) #L(%a,rel=pof,#CLONE;e);
```

Most sentences of the corpus did not contain keywords that represent the relations. Therefore, the grammar depends on the features assigned to the words that need to be related. It specially depends on the semantic classification in order to determine which ontological relation should be used. Table 3 lists some semantic features that have been observed:

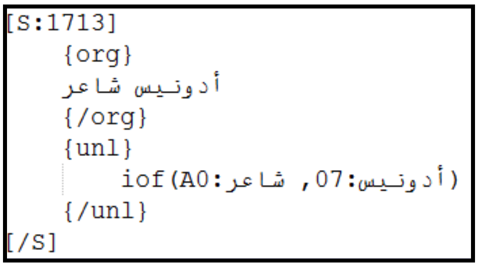
TABLE 3

SEMANTIC FEATURES OBSERVED FOR THE UW1 OF THE ONTOLOGICAL RELATIONS IN THE CORPUS

Semantic feature	Explanation	Example
HUM	person (Nouns denoting people.)	طبيب
GRO	group (Nouns denoting groupings of people or objects.)	جامعة
ARF	artifact (Nouns denoting man-made objects.)	شركة
NOB	natural object (Nouns denoting natural objects (not man-made).)	بحر
CGN	cognitive noun (Nouns denoting cognitive processes and contents.)	قانون
LCT	location (Nouns denoting spatial position.)	مدينة

For example, in the sentence “أدونيس شاعر” ‘Adonis is a poet’, the two words are defined in the dictionary as [أدونيس POS=PPN, GEN=MCL, SEM=HUM] and [شاعر POS=N, GEN=MCL, SEM=HUM]. A rule can link between “أدونيس” ‘Adonis’ and “شاعر” ‘poet’ with an ‘iof’ relation through depending on the ‘HUM’ (human) feature. The rule in (4) states that if a noun such as “شاعر” ‘poet’ with the semantic feature ‘HUM’ comes after a proper noun such as “أدونيس” ‘Adonis’, then both nouns would be linked with a ‘iof’ relation as in Fig. 4.

```
(4) (%x , PPN , HUM , ^rel = iof ) (%y , HUM , ^PPN , GEN = %x) ({^N | STAIL }, %q ) :=
    (iof(%x , +with_rel ; %y , +rel = iof ) , %01 ) #L(%y ,  #CLONE , +rel = iof ; %q ) ;
```



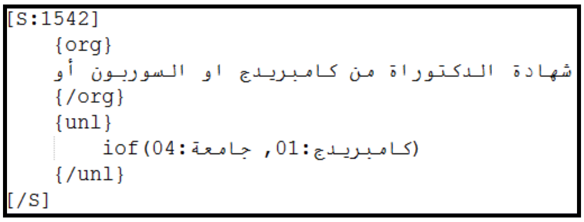
```
[S:1713]
    {org}
    أدونيس شاعر
    {/org}
    {unl}
        iof(A0:شاعر ,07:أدونيس)
    {/unl}
[/S]
```

Figure 4: The UNL ontological relation using semantic features technique

Context prediction

The identity of a proper name can be predicted from the context in which it occurs. For example, the proper noun “كامبردج” ‘Cambridge’ in “شهادة الدكتوراه من كامبردج أو السوربون أو” ‘PHD from Cambridge or Sorbonne or’ doesn’t have an adjacent noun that has one of the semantic features mentioned in table 3, but one of the adjacent nodes (words) can help in predicting the identity of the proper name “كامبردج” ‘Cambridge’ which is the noun “شهادة” ‘certificate’. If the rules find this list of words, then the noun “جامعة” ‘university’ will be inserted by the rule in (5) in order to be “شهادة الدكتوراه من جامعة كامبردج” ‘PHD certificate from Cambridge university’ that is linked by the ‘iof’ relation as in Fig. 5 .

```
(5) (%a,{“شهادة الماجستير”|“شهادة الدكتوراه|”التعليم العالي” },^ins):=(%a,ins)(?[جامعة],blk,INS);
```



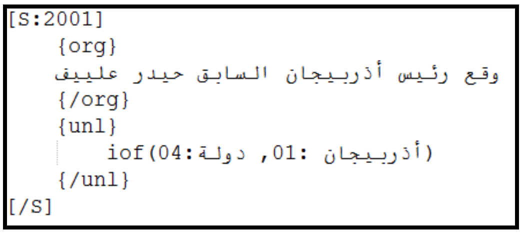
```
[S:1542]
    {org}
    شهادة الدكتوراة من كامبريدج او السوربون أو
    {/org}
    {unl}
        iof(04:جامعة ,01:كامبريدج)
    {/unl}
[/S]
```

Figure 5: The UNL ontological relation using context prediction technique

Knowledge base

In the sentence “وقع رئيس أذربيجان السابق حيدر علييف” ‘former President of Azerbaijan Heydar Aliyev has signed’, the two nouns “رئيس” ‘president’ and “أذربيجان” ‘Azerbaijan’ cannot be linked with a direct relation, given the fact that the dictionary includes [أذربيجان POS=PPN, GEN=MCL, SEM=LCT, CAR=ONE] and [رئيس POS=N, GEN=MCL, SEM=HUM], since “أذربيجان” ‘Azerbaijan’ is not an instance for “رئيس” ‘president’. However, such adjacent words with those semantic features should be linked with ‘mod’ relation, but not an ontological one as in rule (6); the mod relation is not displayed in the final output, but it is a method to block applying the ontological relation. All of the adjacent nodes; the context around the proper name “أذربيجان” ‘Azerbaijan’ fail to help in recognizing its identity. Therefore, only the dictionary features can help in predicting the identity of “أذربيجان” ‘Azerbaijan’, as it is a location ‘SEM=LCT’ and it is the only one in the world ‘CAR=ONE’, so it could be concluded that it is an instance of a country. In the case of prediction of a proper name identity, the rule inserts the identity noun “دولة” ‘country’ before the proper name by the insertion rule described in (7). The ‘iof’ relation will link “أذربيجان” and “دولة” as in Fig. 6.

```
(6)
(SHEAD,%c)(N,HUM,^with_rel,^INS,%b)(%a,N,PPN,^GEN=%b):=(mod(%b;%a,rel=mod))
#L(%c;%a,rel=mod,#CLONE);
(7)
({^GRO,^ARF,^HUM,^NOB,^LCT,^CGN|SHEAD
|LCT,PPN|PPN,GRO|CGN,DEF|HUM,PLR|"سكان"|"عاصمة"},%a)(%b,LCT,PPN,ONE,{^ins|ins,SPLIT},^with_add,^rel=
iof)({^"دولة",^NOB,^LCT|STAIL|node_left_del|LCT,PPN|COO|PPN,GRO|DEF},%c):=(%a)(?[دولة],blk,INS)(%b,ins)(%c
).
```



```
[S:2001]
    {org}
    وقع رئيس أذربيجان السابق حيدر علييف
    {/org}
    {unl}
        iof(04:دولة ,01: أذربيجان)
    {/unl}
[/S]
```

Figure 1: The UNL ontological relation using knowledge base technique

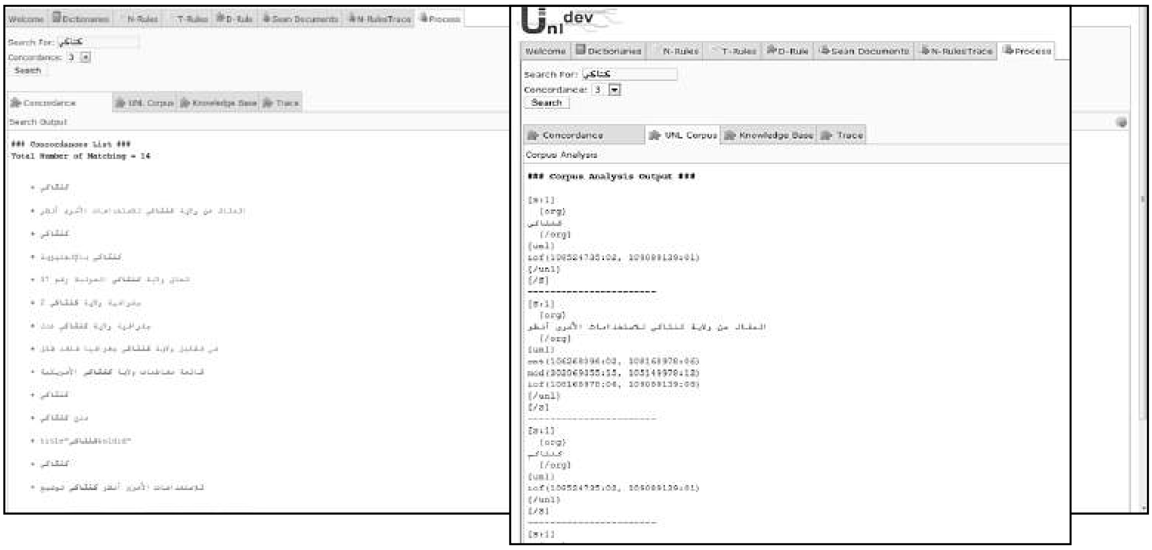
B) Tools and Engines

1) SEAN: is the acronym for Shallow Enhanced ANalyser. It is fully automatic; it does not allow for any human intervention. It is a multi-document analyzer . Moreover, it is a word-driven analyzer: the unit of analysis is a word that is provided by the user. It is also a shallow analyzer: the analysis targets the surface structure of natural language sentences.

SEAN is appropriate for information retrieval and extraction task, because it provides a rather rough and partial analysis of the natural language input. SEAN has been developed by the engineering team in the Library of Alexandria.

Dictionaries, N-rules, T-rules and D-rules tabs in SEAN are provided the dictionary, normalization rules, transformation and disambiguation rules. In the Sean Documents tab, the NL documents can be uploaded either as web links or a text file in the UTF8 format.

Moreover, the Process tab, allows the user to search for a word in the uploaded text. The number of words around the searched word can be specified from the combo box ‘Concordance’. The process tab consists of 4 sub-tabs; concordance, UNL corpus, Knowledge base and trace tabs. By clicking the ‘search’ button under the comb box, search results will be shown in the concordance tab in the left pane as shown in Fig. 7.



Search For: كتاكي
Concordance: 3

```
### Concordances List ###
Total Number of Matching = 14

    • كتاكي
    • [illegible] من رواية كتاكي [illegible] الأخرى أنظر
    • كتاكي
    • كتاكي [illegible]
    • [illegible] رواية كتاكي [illegible] رقم 17
    • [illegible] رواية كتاكي 2
    • [illegible] رواية كتاكي [illegible]
    • في [illegible] رواية كتاكي [illegible]
    • [illegible] رواية كتاكي الأمريكية
    • كتاكي
    • [illegible] كتاكي
    • title="كتاكي[illegible]"
    • كتاكي
    • [illegible] الأخرى أنظر كتاكي [illegible]
```

```
### Corpus Analysis Output ###

[S:1]
  [org]
كتاكي
  [/org]
[unl]
icf(106524735:02, 108088139:01)
[/unl]
[/S]
------------------------
[S:1]
  [org]
[illegible] من رواية كتاكي [illegible] الأخرى أنظر
  [/org]
[unl]
mos(106268396:03, 108168978:06)
mod(302069355:15, 105149978:12)
icf(108168978:06, 108088139:08)
[/unl]
[/S]
------------------------
[S:1]
  [org]
كتاكي
  [/org]
[unl]
icf(106524735:02, 108088139:01)
[/unl]
[/S]
------------------------
[S:1]
```

Figure 7: The “Process” tab

The selected natural language text from the concordance result is processed by the selected dictionary and the selected rules files, the UNL expressions of all search results are shown in the sub-tab ‘UNL corpus’. The behavior of the applied rules can be viewed in the sub-tab ‘trace’.

2) EUGENE: This tool is responsible for generating the natural language sentences out of semantic networks represented in the UNL format. In its current release, it is a web application developed in Java and available at the UNLdev4. EUGENE is an acronym for dEp-to-sUrface GENErator. As a multilingual engine, EUGENE must be parameterized to the target natural languages with the following files that are provided through EUGENE's interface: The input document in the UNL document structure, i.e., the universal semantic network to be generated in natural language, the UNL-NL (generation) dictionary, i.e., a lexical database where UWs are mapped into natural language entries, along with the corresponding features, the UNL-NL (generation) transformation grammar, i.e., a set of transformation rules used to convert the UNL graphs into natural language sentences and the UNL-NL (generation) disambiguation grammar, i.e, a set of disambiguation rules used to improve the results of the tokenization and of the transformation[5][31].

5 KEYS (KEY’S INTERFACE)

KEYS is an automatic language-independent knowledge extraction system, it automatically extracts structured information, from unstructured machine-readable natural language documents. The system is able to work with any language as long as it contains the required resources of this given language. The following sub-sections will describe how KEYS works starting from the point of files uploading to the point of obtaining the results.

A. Uploading documents

The user has to select the language of the file that will be uploaded from a dropdown list. The user can also select the desired text file or zip folder from his file system by clicking on the browse button then clicking on the upload button. The user can add a URL file, by inserting the URL.

B. Search for a query

The user has to select the desired language, concordance size and documents in order to search in these documents, then enter the search query and click the search button. The results will be viewed in tabs (Visualization, Simplified and Eugene).

[4] http://dev.undlfoundation.org/index.jsp

[5] http://www.unlweb.net/wiki/EUGENE

1) Visualization: The output will be presented in the form of a graph by clicking on the “Visualization” tab. The output of this option depends on the results provided through SEAN. In this view, the user can click on each node in the graph to display its relations; the thick arrow refers to the most frequent instance of the word as shown in Fig. 8.

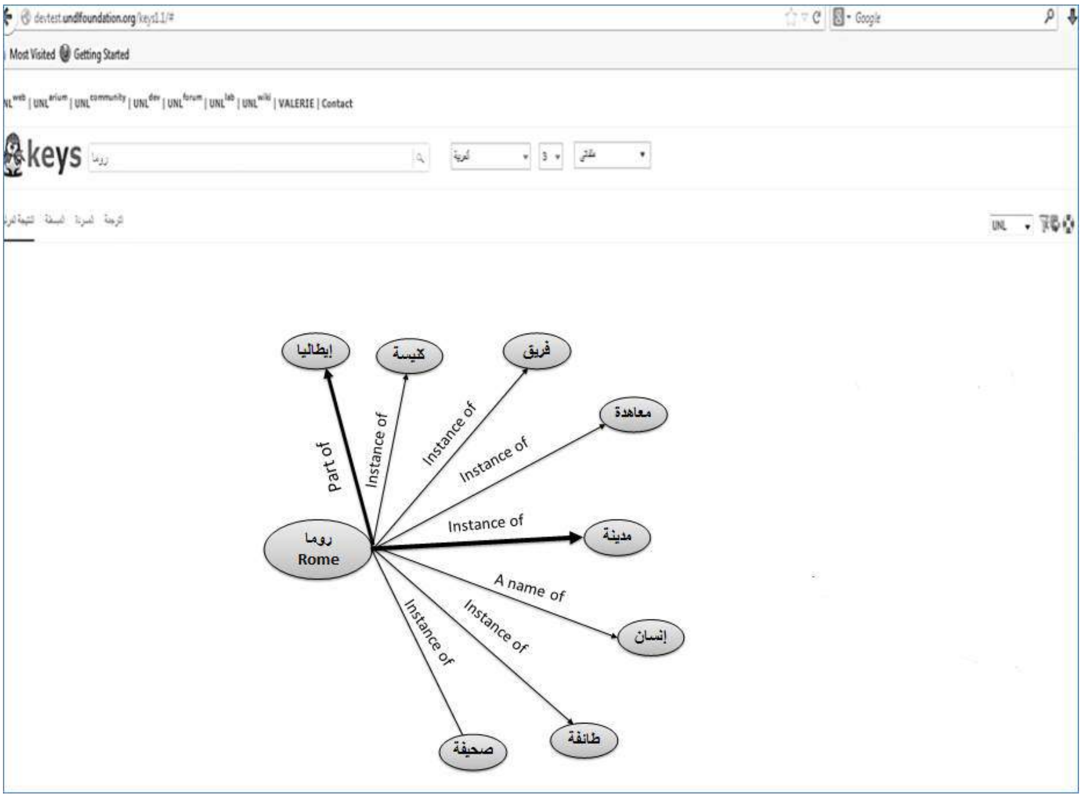
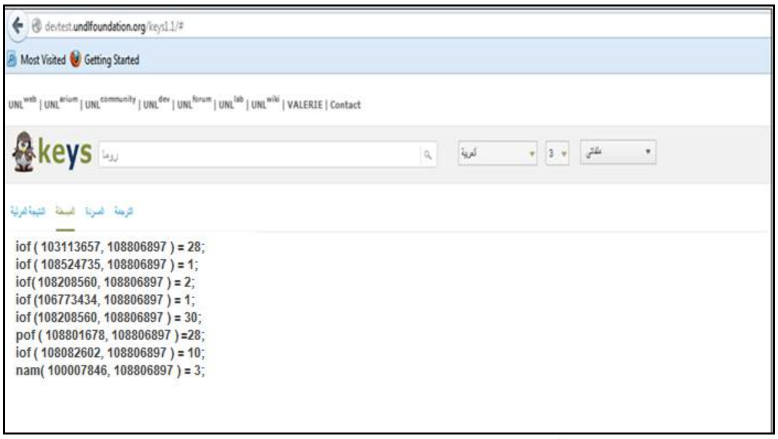


Figure 8: KEYS output in the UNL view (visualization)

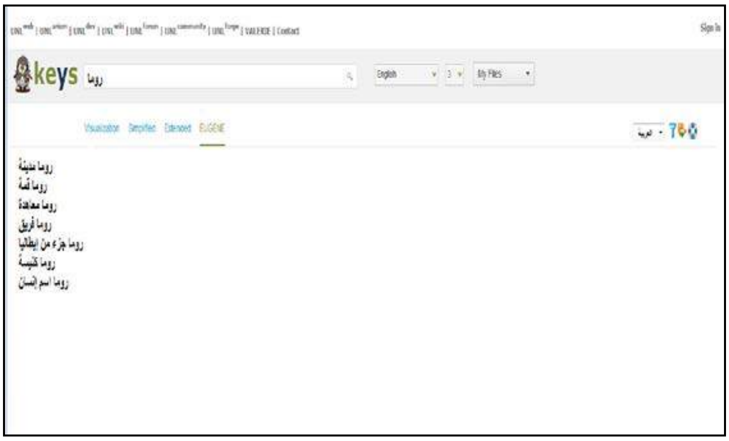
2) *Simplified KB:* The output will be presented in the form of UNL expressions by clicking on the “Simplified” tab. The output of this option also depends on the results provided through SEAN. The output will appear as shown in Fig. 9.


devtest.undlfoundation.org/keys1.1/#
Most Visited
Getting Started
UNLweb | UNLarium | UNLcommunity | UNLdev | UNLforum | UNLlab | UNLwiki | VALERIE | Contact
keys
روما
3

```
iof ( 103113657, 108806897 ) = 28;
iof ( 108524735, 108806897 ) = 1;
iof( 108208560, 108806897 ) = 2;
iof (106773434, 108806897 ) = 1;
iof (108208560, 108806897 ) = 30;
pof ( 108801678, 108806897 )=28;
iof ( 108082602, 108806897 ) = 10;
nam( 100007846, 108806897 ) = 3;
```

Figure 9: KEYS output in the UNL view (simplified)

3) EUGENE: The output will be presented in the form of natural language sentences by clicking on the “EUGENE” tab. The output of this option depends on the results of the generation tool EUGENE. The output will appear as shown in Fig. 10.



روما مدينة
روما قمة
روما معاهدة
روما فريق
روما جزء من ايطاليا
روما كنيسة
روما اسم إنسان

Figure 10: KEYS output (Eugene)

6 EVALUATING THE RESULTS

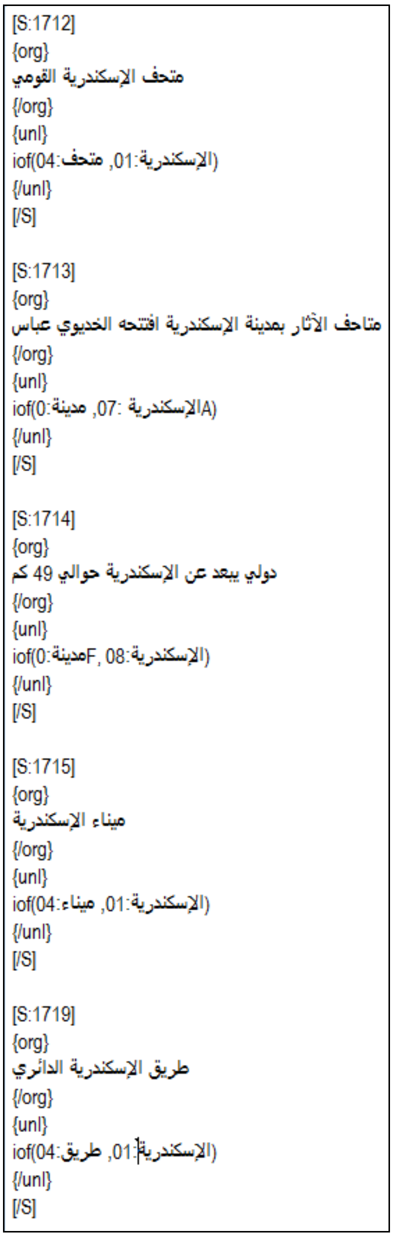
Evaluation has been performed in order to investigate the accuracy and robustness of the grammar. The used data consists 1000 proper nouns as keys for search with total number of occurrences being 22,000. The instances are divided into a training set which includes 17,000 instances and a testing set which includes 5,000 instances. The same proper name used in the training data has been tested in different contexts, different from the trained instances. For example, the output of the proper name "هدسون" 'Hudson' in the trained data was (14 instances), while the tested contexts represent 4 instances. The primary scores are precision and recall. Let N_{key} be the total number of filled slots in the answer key, $N_{response}$ be the total number of filled slots in the system response, and $N_{correct}$ be the number of correctly filled slots in the system response (i.e., the number which match the answer key). Then

$$\textbf{Precision} = \frac{N\ correct}{N\ response} = 19.500/21.000 \quad = 0.92$$

$$\textbf{Recall} \quad = \frac{N\ correct}{N\ \ keys} = 19.500/22.000 \quad = 0.886$$

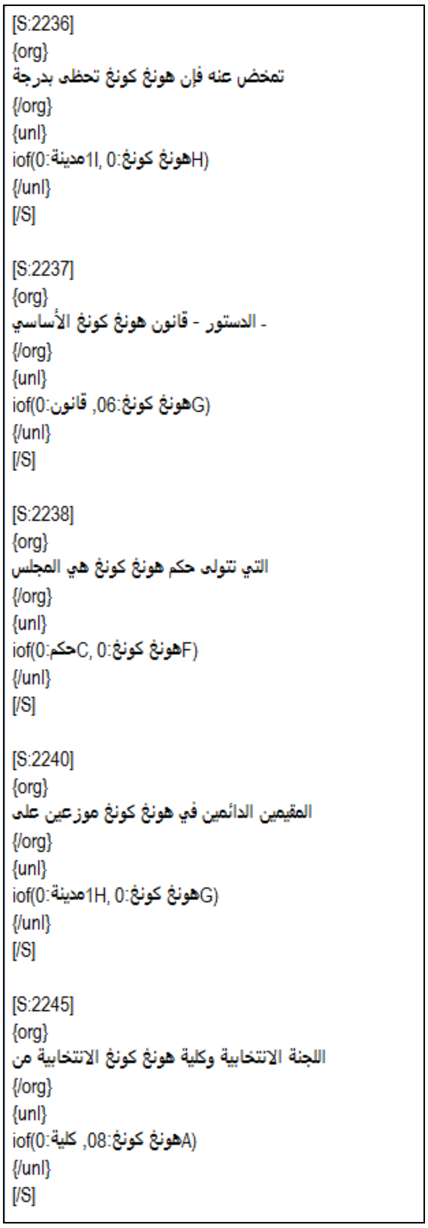
The F measure was calculated with the equation: F= (2 × precision × recall)/(precision + recall) and the accuracy was 90.2 %. These equations were calculated for each answer key in the corpus, then all of their results were added to provide the total accuracy of the corpus.

In addition, a different set of proper names other than those used in the training and testing sets have been used to see whether the system has enough knowledge to search for any other entities in other contexts. For instance, Fig. 11, 12, and 13 represent the output samples of the proper names "الإسكندرية" 'Alexandria', "هونج كونج" 'Hong Kong' and "مونتريال" 'Montreal' respectively which were not included in the 1000 proper names we worked with. Fig. 11,12 and 13 reflect that the system has learned abstracted knowledge that made it able to deal with both new keys and new contexts.



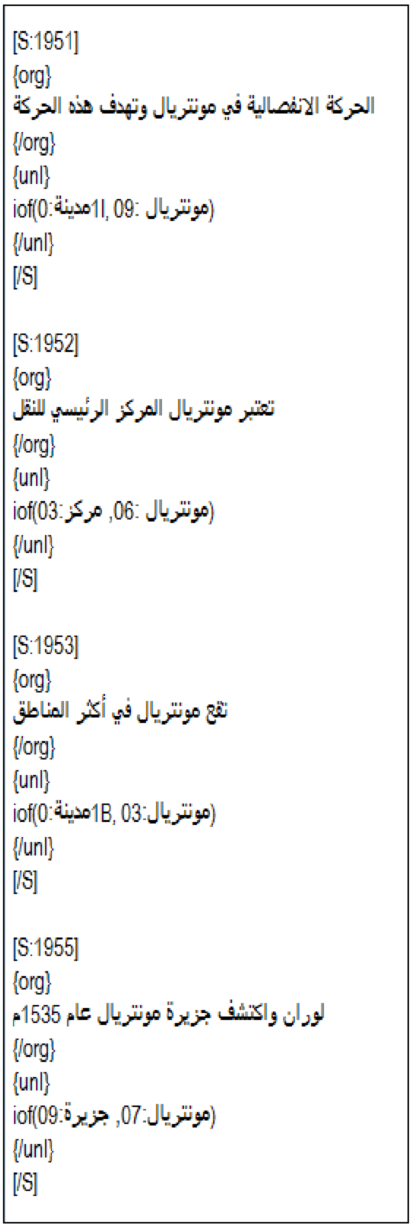
```
[S:1712]
{org}
متحف الإسكندرية القومي
{/org}
{unl}
iof(04:متحف ,01:الإسكندرية)
{/unl}
[/S]

[S:1713]
{org}
متاحف الآثار بمدينة الإسكندرية افتتحه الخديوي عباس
{/org}
{unl}
iof(0:مدينة ,07: الإسكندريةA)
{/unl}
[/S]

[S:1714]
{org}
دولي يبعد عن الإسكندرية حوالي 49 كم
{/org}
{unl}
iof(0:مدينةF, 08:الإسكندرية)
{/unl}
[/S]

[S:1715]
{org}
ميناء الإسكندرية
{/org}
{unl}
iof(04:ميناء ,01:الإسكندرية)
{/unl}
[/S]

[S:1719]
{org}
طريق الإسكندرية الدائري
{/org}
{unl}
iof(04:طريق ,01:الإسكندرية)
{/unl}
[/S]
```

Figure 11: Sample of "الإسكندرية" output



```
[S:2236]
{org}
تمخض عنه فإن هونغ كونغ تحظى بدرجة
{/org}
{unl}
iof(0:مدينة1I, 0:هونغ كونغH)
{/unl}
[/S]

[S:2237]
{org}
- الدستور - قانون هونغ كونغ الأساسي
{/org}
{unl}
iof(0:قانون ,06:هونغ كونغG)
{/unl}
[/S]

[S:2238]
{org}
التي تتولى حكم هونغ كونغ هي المجلس
{/org}
{unl}
iof(0:حكمC, 0:هونغ كونغF)
{/unl}
[/S]

[S:2240]
{org}
المقيمين الدائمين في هونغ كونغ موزعين على
{/org}
{unl}
iof(0:مدينة1H, 0:هونغ كونغG)
{/unl}
[/S]

[S:2245]
{org}
اللجنة الانتخابية وكلية هونغ كونغ الانتخابية من
{/org}
{unl}
iof(0:كلية ,08:هونغ كونغA)
{/unl}
[/S]
```

Figure 12: Sample of "هونغ كونج" output



```
[S:1951]
{org}
الحركة الانفصالية في مونتريال وتهدف هذه الحركة
{/org}
{unl}
iof(0:مدينة1I, 09: مونتريال)
{/unl}
[/S]

[S:1952]
{org}
تعتبر مونتريال المركز الرئيسي للنقل
{/org}
{unl}
iof(03:مركز ,06: مونتريال)
{/unl}
[/S]

[S:1953]
{org}
تقع مونتريال في أكثر المناطق
{/org}
{unl}
iof(0:مدينة1B, 03:مونتريال)
{/unl}
[/S]

[S:1955]
{org}
لوران واكتشف جزيرة مونتريال عام 1535م
{/org}
{unl}
iof(09:جزيرة ,07:مونتريال)
{/unl}
[/S]
```

Figure 13: Sample of "مونتريال" output

7 Conclusion

Many applications depend on the automatic extraction of structure data from unstructured data for better means of querying, organizing, and analyzing data. KEYS is a knowledge extraction system that promises to fulfil the human needs in providing an easy access to the vast amount of information that is readily available on the internet. The amount of information on the internet is rapidly increasing, it is increasing every second, which makes benefiting from this amount of information difficult. Hence, the importance of knowledge extraction systems is manifested in providing an easy method to obtain the needed information. Knowledge extraction systems maximize the magnitude of utilizing the available information. In this article, the infrastructure of KEYS system is discussed. The linguistic resources and the tools involved in KEYS are presented, they are all provided in an open-source form for free at www.unlweb.net. The precision measurement of the Arabic grammar was 0.92 while recall measurement was 0.886.

REFERENCES

[1] J. Hobbs, and E. Riloff, E. "Information Extraction",_*Handbook of Natural Language Processing, 2nd Edition*, Editors: Nitin Indurkhya and Fred J. Damerau, Chapman & Hall/CRC Press, Taylor & Francis Group, 2010.

[2] S. Patwardhan ,"widening the field of view of information extraction through sentential event recognition", A dissertation submitted to the faculty of The University of Utah in partial fulfillment of the requirements for the degree of Doctor of Philosophy 2010.

[3] W. Lehnert, and J. Cowie, " Information Extraction". Communications of the ACM 39, 80–91,1996.

[4] G. DeJong, "Prediction and Substantiation: A New Approach to Natural Language Processing", *A Multidisciplinary Journal 3*, 251–271. 1, 1996.

[5] G. DeJong, "An Overview of the FRUMP System". *In Strategies for Natural Language Processing*, W. Lehnert and M. Ringle, Eds. Erlbaum, Hillsdale, NJ, 1982, pp. 149–176.

[6] R.Schank, and R.Abelson, "Scripts, Plans, and Understanding". Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

[7] R. Cullingford, "Computer Understanding of Newspaper Stories", PhD thesis, Yale University, 1978.

[8] G. Silva, , and D. Dwiggins, , "Towards a Prolog Text Grammar", ACM SIGART Bulletin 73, 20–25, 1980.

[9] J. Cowie, " Automatic Analysis of Descriptive Texts". *In Proceedings of the First Conference on Applied Natural Language Processing*, Santa Monica, CA, pp. 117–123, 1983.

[10]N. Sager, "Natural Language Information Processing: A Computer Grammar of English and Its Applications", Addison-Wesley, Boston, MA, 1981.

[11] G. Zarri, , "Automatic Representation of the Semantic Relationships Corresponding to a French Surface Expression", *In Proceedings of the First Conference on Applied Natural Language Processing* ,Santa Monica, CA, pp. 143–147, 1983.

[12] R. Grishman, , and B. Sundheim, "Message Understanding Conference - 6: A Brief History". *In Proceedings of the 16th International Conference on Computational Linguistics* ,Copenhagen, Denmark, pp. 466–471, 1996.

[13] R. Merchant, "TIPSTER Program Overview", *In TIPSTER Text Program Phase I*: Proceedings of the Workshop (Fredricksburg, VA), pp. 1–2, 1993.

[14] P. Altomari, and P. Currier, "Focus of TIPSTER Phases I and II". In TIPSTER Text Program Phase II: Proceedings of the Workshop ,Vienna, VA, pp. 9–11, 1996.

[15] F. Ruth Gee, "The TIPSTER Text Program Overview", In Proceedings of the TIPSTER Text Program Phase III ,Baltimore, MD, pp. 3–5, 1998.

[16] G. Krupka, L. Iwariska, P. Jacobs, and L. Rau, "GE NLToolset: MUC-3 Test Results and Analysis", *In Proceedings of the Third Message Understanding Conference (MUC-3)* ,San Diego, CA, pp. 60–68, 1991.

[17] J. Hobbs, "SRI International's TACITUS System: MUC-3 Test Results and Analysis", *In Proceedings of the Third Message Understanding Conference (MUC-3)* ,San Diego,CA , May 1991, pp. 105–107.

[18] W. Lehnert, , C. Cardie, D.Fisher, E. Riloff, , and R. Williams, "MUC-3 Test Results and Analysis", *In Proceedings of the Third Message Understanding Conference (MUC-3)* ,San Diego, CA, May 1991 ,pp. 116–119.

[19] R. Grishman, J. Sterling, , and C. MacLeod, "MUC-3 Test Results and Analysis" ,*In Proceedings of the Third Message Understanding Conference (MUC-3)* ,San Diego, CA, pp. 95–98, 1991.

[20] M.Banko ,O. Etzioni, "Strategies for Lifelong Knowledge Extraction from the Web", Turing Center, University of Washington Computer Science and Engineering, 2007.

[21] L. Elikuil, "Information Extraction from the World Wide Web: A Survey". Norwegian Computer Center, Report no. 945, 1999.

[22] N. Naw and E. Hlaing, "Relevant Words Mining with Compiling Technique" , *International Journal of Emerging Technology and Advanced Engineering* ,Volume 4, Issue 4, 2014.

[23] N. Naw and E. Hlaing, "Relevant Words Extraction Method for Web Recommender System" , *International Conference on Advances in Engineering and Technology (ICAET'2014)* March 29-30, 2014.

[24] R. Grishman, "Information Extraction: Techniques and Challenges". *International Summer School on Information Extraction(SCIE'97)* Frascati, Italy(=Lecture Notes in Computer Science,1299),10-27.Berlin:Springer. 1997.

[25] S. Alansary, M. Nagi, N. Adly, "UNL+3: The Gateway to a Fully Operational UNL System", *in Proceeding of 10th Conference on Language Engineering*, Cairo, Egypt, 2010.

[26] H. Uchida, "UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration". UNU/IAS/UNL Center. Tokyo, Japan,1996.

[27] H. Uchida, M. Zhu, "The Universal Networking Language beyond Machine Translation", UNL Foundation, 2001.

[28] C.Jesús, A. Gelbukh, E. Tovar (eds.), "*Universal Networking Language: advances in theory and applications*". Mexico City: National Polytechnic Institute, 2005.

[29] H. Uchida, M. Zhu, "UNL2005 for Providing Knowledge Infrastructure" , *in Proceeding of the Semantic Computing Workshop (SeC2005)*, Chiba, Japan, 2005.

[30] S. Alansary, "MUHIT: A Multilingual Harmonized Dictionary", The 9th edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland, 26-31 May. , 2014.

[31] S. Alansary, M. Nagi and N. Adly, "Generating Arabic Text: the Decoding Component in an Interlingual System for Man-Machine Communication in Natural Language", 6th International Conference on Language Engineering, Cairo, Egypt, December 2006.

Biographies

Dr. Sameh Alansary*: Director of Arabic Computational Linguistics Center* Bibliotheca Alexandrina

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

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

Dr. Magdy H. Nagi: Senior Consultant, ICT Sector Bibliotheca Alexandrina.

Dr. Nagi is a Professor in the Computer and Systems Engineering department, Faculty of Engineering, Alexandria University. He obtained his Ph.D. from the University of Karlsruhe, in 1974, where he served as Lecturer for two years and as a Consultant to its Computer Center from 1974-1990. During this period he also served as Consultant to many companies in Germany such as Dr. Otker, Bayer, SYDAT AG, and BEC. He served, since 1995, as Consultant to the Bibliotheca Alexandrina. Among his activities were the design and installation of Bibliotheca Alexandrina's network and information system, namely a trilingual information system that offers full library automation. In 2001, he got appointed as the Head of the Information and Communication Technology (ICT) Sector of the Bibliotheca Alexandrina and occupied that post till 2012. He currently serves as a senior Consultant to the ICT Sector and continues to oversee the various projects and partnerships established between the ICT Sector and many international institutions. Dr. Nagi is a member of the ACM and the IEEE Computer Society as well as several other scientific organizations. His main research interests are in operating systems and database systems. He is author/co-author of more than 100 papers.

ملخص:

KEYS: نظام استخلاص معرفي اعتمادا على البنية التحتية المعرفية للغة الشبكات العالمية

سامح الأنصاري[1*] ، مجدي ناجي[2**]

مكتبة الإسكندرية، الشاطبي، الإسكندرية، مصر
*قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر
[1]sameh.alansary@bibalex.org
** قسم هندسة النظم والحاسب، كلية الهندسة، جامعة الإسكندرية، الإسكندرية، مصر
[2]magdy.nagi@bibalex.org

ملخص — في وجود ثورة المعلومات واتاحة كم كبير منها على صفحات الانترنت زاد احتياج الانسان لاستخلاص معلومات محددة من هذه الصفحات ، لذلك فإن هدف هذه الورقة البحثية تقديم النظام (KEYS) كنظام استخلاص معرفي يهدف إلى استخلاص واسترجاع المعلومات إذ يقوم بالبحث عن المعلومات داخل نصوص ممثلة دلاليا باستخدام لغة الشبكات العالمية (UNL) مما يجعل مهمة استخلاص واسترجاع المعلومات أكثر دقة وعندها يكون من المتوقع أن تكون النتائج ذات جودة عالية ؛ إذ يتم تحليل النص المصدر تحليلا سطحيا وتحويله إلى شبكة دلالية تحتوي على علاقات انطولوجية محددة ممثلة باستخدام لغة الشبكات العالمية، بعد ذلك يتم توليد آلي لأي لغة طبيعية من هذه الشبكة الدلالية. وبهذا يكون من المتوقع أن يقدم هذا النظام نهجا جديدا لتحديد الكيان الاسمي وهو استخراج الاسماء بجميع تصنيفاتها الانطولوجية من اللغة الطبيعية أيا كانت هذه اللغة.

قضايا صوتية خلافية في ضوء التحليل الصوتي الحاسوبي

دكتور أحمد راغب أحمد*
*أستاذ علم اللغة المشارك
الجامعة الإسلامية العالمية بماليزيا
ragheb@iium.edu.my

المستخلص:
نشأ علم الأصوات العربي نتيجة طبيعية لاهتمام العلماء بالقرآن الكريم، ومحاولتهم ضبط طرق نطقه، وبيان الأسس والأساليب التي تأسست عليها لغة القرآن. لقد كان نزول القرآن الدافع الأساس لظهور العلوم العربية عامة وعلم الأصوات بصفة خاصة، ولقد أدت عناية القراء بضبط قراءة النص القرآني وتلاوته تلاوة صحيحة إلى نشأة الدرس الصوتي العربي.

ولا شكَّ أن علم التجويد يعتبر مصدرا أصيلا من مصادر الدراسة الصوتية العربية، وهذه نتيجة مبنية على أساس الإنجازات القيِّمة التي حقَّقها علماء التجويد في مجال الدراسة الصوتية، لا على أساس وفرة المصنفات في هذا العلم. بل لعلي لا أتجاوز الحدَّ إن زعمت أن ما وصل إليه علماء الأصوات حاليًا إنما هو تَتِمَّة لما وصل إليه علماء العرب قديما ومن قبلهم علماء الهنود. لقد كانت جهود علماء العربية في دراسة الأصوات اللُّغوية من الإنجازات المتميزة في الدرس اللغوي، وقامت حولها دراسات ليست قليلة، ولكن أحدًا من المشتغلين بدراسة الأصوات العربية المحْدثين لم يلتفتْ إلى كتب التجويد التي تتضمن دراسة للأصوات اللغوية لا تقل أهميتها عن جهود علماء العربية.

وتتناول هذه الدراسة مجموعة من القضايا الصوتية الخلافية بين علماء اللغة المتقدمين واللغويين المعاصرين، فهي دراسة تجمع بين النظرية والتطبيق، وذلك من خلال الاتكاء على نتائج التحليل التقني لأصوات اللغة، إلا أن تلك الدراسة لم تكن منبتة الصلة عن جهود علماء العربية الذين قدموا وصفًا تفصيلا لأصوات اللغة العربية بغية الحفاظ على النطق العربي من اللحن أو التبديل وحفاظًا على نطق القرآن الكريم بصورة سليمة معيارية.

الكلمات المفتاحية: علم التشكيل الصوتي، الموجة الصوتية، منحنى التنغيم الأساسي، الصورة الطيفية.

***Keywords:* Arabic Phonology, wave form, Fundamental Frequency, spectrogram.**

قضايا صوتية خلافية في ضوء التحليل الصوتي الحاسوبي

ذهب سيبويه إلى أنَّ الهمزة والقاف والطاء أصوات مجهورة، وذلك حين حصر الحروف المجهورة في تسعة عشر حرفًا هي "الهمزة والألف، والعين، والغين، والقاف، والجيم، والياء والضاد، واللام، والنون، والراء، والطاء، والدال، والزاي، والظاء، والذال، والباء، والميم، والواو. فذلك تسعة عشر حرفًا"[1].

وقد تبعه فيما ذهب إليه كل علماء اللغة الأقدمين الذين أتوا بعده، ولم يضيفوا جديداً إلى هذه المسألة سوى مزيد من الشرح والتحليل والاستدلال.

غير أن هذه الرؤية قد انقلبت رأساً على عقب فيما يخص وصف هذه الأصوات لدى علماء اللغة المحدثين، الأمر الذي أدى إلى سؤالين مقتضاهما:

أولاً: إذا كان هذا الخلاف على هذا القدر من الوضوح فأي الفريقين قد أصاب الحقيقة في وصف هذه الأصوات وأيهما قد جانبه الصواب؟

ثانياً: ما الأسباب الحقيقية التي أدت إلى هذا الاختلاف؟

وسأشرع في محاولة الإجابة عن هذين السؤالين على النحو الآتي:

أولًا: الطاء بين الجهر والهمس:

الطاء صوت مهموس في العربية المعاصرة، واعتبرها سيبويه مفخم الدال ورأى أنه "لولا الإطباق لصارت الطاء دالًا"[2].

وصوت الطاء، كما ينطق بها اليوم، يقابل صوت التاء في التفخيم والترقيق، وكلاهما صوت شديد مهموس، "ولا فرق بينهما إلا في أن مؤخرة اللسان ترتفع تجاه الطبق عند نطق الطاء، ولا ترتفع نحوه في نطق التاء"[3].

وهكذا ينحصر الخلاف بين سيبويه والعلماء المحدثين في وصف الطاء بين الهمس والجهر، فقد عدها سيبويه من الأصوات المجهورة بينما أكد البحث الحديث على عدم اهتزاز الوترين الصوتيين أثناء النطق بها، وقد ذهب الكثير من المحدثين إلى اعتبار الطاء صوتًا مجهورًا في القديم وقد تحول إلى الهمس بفعل عامل التطوير، فالطاء "مهموسة اليوم، مجهورة عند القدماء، ونطق الطاء العتيق قد انمحى وتلاشى تمامًا"[4].

بينما يحاول "شاده" رد هذا الاختلاف إلى الظواهر اللهجية أو الجغرافية فيقول: "سيبويه يعد من المجهورة الطاء والقاف. وفي لفظ عصرنا لا نصيب للأوتار الصوتية في إنتاجهما، ولكن ذلك لا يصح إلا عن لفظ المدارس [يقصد الفصحى الحالية]، وأما اللهجات فتخالفها مخالفة شديدة......"[5]، وهذا الرأي ربما لا يجد ما يدعمه إلا محاولة الإشادة بجهود علماء العرب القدامى.

والذي أميل إليه في هذه المسألة أنه إذا جاز لنا تقديم الأعذار لعلماء العربية السابقين الذين بذلوا وسعهم ولم يألوا جهدًا في حدود الإمكانات التي أتيحت لهم، إننا إذا كنا نقبل منهم شاكرين ما وصلوا إليه في هذه المسألة إلا أننا لا نستطيع أن نقبل بحال من الأحوال أن تبقى هذه

[1] سيبويه،الكتاب، تحقيق/ عبد السلام محمد هارون، دار الجيل، بيروت 434/4.

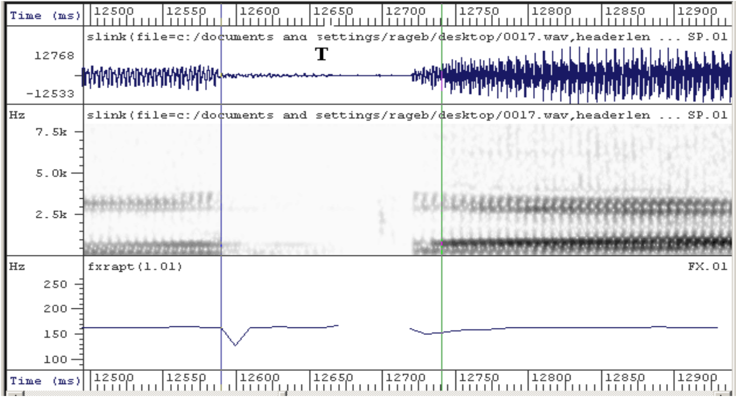
[2] سيبويه، الكتاب، 436/4.

[3] عبد التواب، رمضان، المدخل إلى علم اللغة، ص: 75.

[4] عبد التواب، رمضان، التطور النحوي، ص: 9.

[5] عبد التواب، رمضان، المدخل إلى علم اللغة، ص75.

المسألة مسألة خلافية في وقتنا الحالي، فيكفينا عرض صورة طيفية واحدة لتثبت بما لا يدع مجالًا للشك أن الطاء صوت انفجاري مهموس لا عمل للوترين الصوتيين في إنتاجه، غير أن هذا الصوت كباقي الأصوات الانفجارية يتكون من عمليتين صوتيتين متتاليتين؛ حيث يندفع الهواء من الصدر فتقف برهة أمام الوترين الصوتيين ثم ما يلبثان حتى ينفرجا ليحدث الصوت الانفجاري، شأنها في ذلك شأن القاف والكاف والتاء والدال والباء. وللتدليل على ذلك أعرض في ما يلي صورة طيفية للطاء ثم أعقب بعدها بقراءتي لهذا الصوت على النحو التالي:



الشكل رقم [1]

صورة طيفية لجملة: [ويقطعون ما أمر الله به أن يوصل]، مع التركيز على صوت الطاء الانفجاري المهموس.

وهذه الصورة تعرض ثلاثة من مستويات التحليل الصوتي لصوت الطاء، ويظهر في المستوى الأعلى شكل الموجة [wave form] ويبدو جليًا أنها لصوت مهموس؛ حيث لا يوجد أثر للذبذبات التي تقترن دائمًا بالأصوات المجهورة، مثل الصوت التالي لها وهو صوت العين. أما المستوى الثاني فيعرض الصورة الاسبكتروجرام ويتضح فيه أيضًا خلو هذا الصوت من الذبذبات المجهورة، أما المستوى الأخير فهو المستوى الذي يحدد النغمة الأساسية [Fundamental Frequency] ونلاحظ انقطاع الخط القاعدي لها وهو أمر ملازم للأصوات المهموسة فقط. وعليه فإن صوت الطاء صوت مهموس لا تظهر فيه من معالم الجهر شيء.

ثانيًا: القاف مهموسة أم مجهورة؟:

هو صوت لهوي شديد مهموس في العربية المعاصرة، أما سيبويه ومن تبعه من النحاة والقراء فقد ذهبوا إلى أنه صوت مجهور، ويستنتج الدكتور إبراهيم أنيس "من وصف القدامى لهذا الصوت أنه كان يشبه إلى حد كبير تلك القاف المجهورة التي نسمعها الآن بين القبائل العربية في السودان، وجنوب العراق، فهم ينطقون بها نطقًا يخالف نطقها في معظم اللهجات العربية الحديثة؛ إذ نسمعها منهم نوعا من الغين"[6].

بينما ينحو الدكتور رمضان عبد التواب في هذه المسألة منحى التعدد اللهجي فيذكر أن "القبائل العربية لم تكن تنطق القاف بصورة موحدة، فها هو ابن دريد اللغوي يقول: "فأما بنو تميم، فإنهم يلحقون القاف بالكاف؛ فيقول: الكوم، يريد: القوم؛ فتكون القاف بين الكاف والقاف. وهذه لغة معروفة في بني تميم، قال الشاعر:

ولا أكول لكدر الكوم كد نضجت ** ولا أكول لباب الدار مكفول"[7][8]

ولنا أن نسأل:هل أخطأ سيبويه في وصف هذه الأصوات أم أن التطور اللغوي قد ألقى وقال كلمته في هذه المسألة باعتبار أن هذه الأصوات كانت تنطق مجهورة ثم حدث لها نوع من التطور اللغوي فتحولت إلى نظائرها المهموسة؟

والحق أنني لا أميل إلى هذا الرأي الأخير بأي حال من الأحوال، ودليلي على ذلك أن قراء القرآن وأئمة الأداء ما زالوا يقرأون القرآن بهذه الصورة التي لا تختلف عن العربية القديمة، وقد ورثوا هذا الأداء وتعلموه من مشايخهم عن طريق المشافهة والسماع.

ومن باب آخر فإن الزعم بالتطور اللغوي يؤدي إلى نتيجة مفادها أننا نقرأ القرآن الآن بطريقة متباينة في بعض الوجوه عن تلك الطريقة التي قرأه بها النبي صلى الله عليه وسلم وعلمها لأصحابه رضوان الله عليهم. وهذا الزعم محال؛ لأنه يخالف أصلًا إسلاميًا وهو حفظ الله تعالى للقرآن الكريم من اللحن أو التحريف.

أما الرأي الأول فإنه ينطوي هو الآخر على شيء من المجازفة، وبتأمل مفهوم سيبويه للجهر والهمس يتبين لنا أنه يختلف عن مفهومهما عند علماء العربية في العصر الحديث.

حيث يرى سيبويه أن "المجهور: حرف أشبع الاعتماد عليه ، ويجري الصوت، فهذه حال المجهورة في الحلق والفم، إلا أن النون والميم قد يعتمد لهما في الفم والخياشيم فتصير فيهما غنة، والدليل على ذلك أنك لو أمسكت بأنفك ثم تكلمت بهما لرأيت ذلك قد أخل بهما"[9].

فالجهر عند سيبويه صفة صوتية ترتبط بإشباع الاعتماد في موضعه، ومنع النفس أن يجري مع أداء الصوت المتصف بتلك الصفة حتى ينقضي الاعتماد عليه ويجري الصوت"[10].

بينما الجهر عند علماء العربية المحدثين هو "صفة صوتية ترتبط بتذبذب الأوتار الصوتية حين النطق"[11].

أما مفهوم سيبويه عن الهمس فيتمثل في كونه"صفة صوتية تتعلق بإضعاف الاعتماد في موضعه بصورة تسمح بأن يجري النفس مع أداء الصوت المتصف به "[12].

[6] أنيس، إبراهيم، الأصوات اللغوية، مكتبة الأنجلو المصرية، ط4، 1971م، ص: 77.

[7] ابن دريد، جمهرة اللغة، 5/1، وعلق كرنكو في الهامش بقوله: "معنى تغليظ القاف التلفظ بالكاف الفارسي... وهذا الشعر لأبي الأسود الدؤلي، ويروى لحاتم الطائي ولغيره"، وانظر النص كذلك في الصاحبي لابن فارس، ص: 36.

[8] انظر: عبد التواب، رمضان، المدخل إلى علم اللغة، ص: 79.

[9] سيبويه، الكتاب، 434/4.

[10] فتيح، محمد، الأصوات العامة والأصوات العربية، دار الثقافة العربية، القاهرة، ص:199.

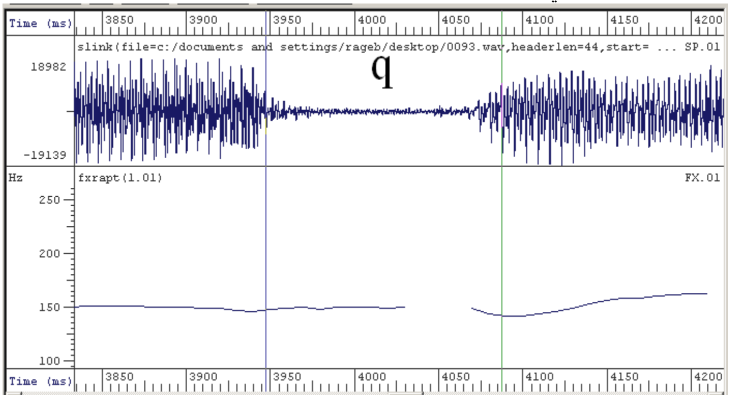
[11] فتيح، محمد، الأصوات العامة والأصوات العربية، ص: 198.

[12] فتيح، محمد، الأصوات العامة والأصوات العربية، ص:199.

بينما يرتبط هذا المفهوم عند اللغويين المحدثين بعدم تذبذب الأوتار الصوتية أثناء القيام بالعملية النطقية .

وربما يرجع الاعتماد المذكور في عبارات سيبويه إلى شدة الضغط في الحجاب الحاجز"[13].

وخلاصة الأمر أن اختلاف مفهوم الجهر والهمس بين سيبويه من ناحية وبين علماء الأصوات المحدثين من الناحية الأخرى هو الذي أدى إلى الاختلاف في تصنيف بعض الأصوات العربية وتوزيعها بين القسمين. ولكن كما سبق أن قلت أثناء تناولي لصوت الطاء فإنه لم يعد يحق لنا الاكتفاء بعرض هذا الخلاف أو ربما مجرد قبوله مع ما أتيح لنا من أجهزة تمكننا من إعطاء صورة حقيقية للجهر والهمس، وعليه فسوف أعرض الآن صورة طيفية للقاف متبعًا ذلك بقراءتي لتلك الصورة على هذا النحو:



الشكل رقم [2]

صورة طيفية لجملة: [إذا تتلى عليه آياتنا قال أساطير الأولين]، مع التركيز على صوت القاف الانفجاري المهموس.

وتعرض هذه الصورة أيضًا ثلاثة من مستويات التحليل الصوتي لصوت القاف، ويظهر في المستوى الأعلى ال [wave form] بوضوح أنها لصوت مهموس؛ حيث لا يوجد أثر للذبذبات التي تقترن دائمًا بالأصوات المجهورة، مثل الصوت التالي أو السابق لها وهو صوت الفتحة الطويلة. أما المستوى الثاني فيعرض الصورة الاسبكتروجرام، ويتضح فيه أيضًا خلو هذا الصوت من الذبذبات المجهورة، أما المستوى الأخير فهو النغمة الأساسية [Fundamental Frequency] ونلاحظ انقطاع الخط القاعدي لها، وهو أمر ملازم للأصوات المهموسة فقط. وعليه فإن صوت القاف صوت مهموس لا تظهر فيه معالم الجهر.

ثالثًا: بينيَّة الهمزة:

لقد اعتبرها سيبويه أولى الحروف المجهورة، وتبعه في ذلك علماء العربية القدامى على حين ذهب بعض المحدثين إلى أن الهمزة العربية صوت مهموس، وذهب فريق ثالث على رأسهم إبراهيم أنيس، و أحمد مختار عمر ، وكما أكد كمال محمد بشر أن " نأخذ بالرأي الذي تبيناه وهو كونها صوتًا لا بالمجهور ولا بالمهموس"[14]؛ ذلك أنه في حال نطق هذا الصوت "ينطبق الوتران انطباقا تاما فلا يسمح بمرور الهواء إلى الحلق مده هذا الانطباق، ومن ثم ينقطع النفس، ثم يحدث أن ينفرج هذان الوتران، فيخرج صوت انفجاري نتيجة لاندفاع الهواء الذي كان محبوسًا حال الانطباق التام، هذا الصوت هو همزة القطع. فهمزة القطع العربية إذن صوت صامت لا هو بالمهموس ولا بالمجهور"[15].

وهذا هو ما أكده أحمد مختار عمر حين ذهب إلى أنه "قد يوضع الوتران في حالة غلق تام محكم يمنع تيار الهواء من تفريقهما، وهو وضع ينتج أصواتًا كثيرة غير لغوية، كما أنه وضع إنتاج "الوقفة الحنجرية" [الهمزة]"[16]. وذهب إلى تعليل ذلك بأنه "لا توجد أعضاء نطق مستعملة في إنتاج هذا الصوت، ولكن الأوتار الصوتية تقوم بدور هذه الأعضاء، لتنتج غلقا تامًا – وإن كان قصيرا – في مجرى تيار الهواء. وحيث إن الأوتار الصوتية نفسها هي المنتجة لهذا الصوت فلا معنى لوصفه بأنه مجهور أو مهموس أو موشوش"[17].

رابعاً: إشكالية الضاد

صوت الضاد في اللغة العربية صوت دار حوله جدل طويل، ولا تكاد تجد في كتب التجويد ولا في كتب الصوتيات العربية أكثر إثارة للجدل من حرف الضاد، ومن أجل ذلك سميت اللغة العربية بلغة الضاد، ولا تكاد تجد بين علماء التجويد خلافًا في غيره.

وهذه القضية قد أخذت أبعادًا كثيرة في وسط القراء على وجه الخصوص، فكل يذهب إلى سداد رأيه وتخطئة رأي مخالفه، مع احتجاج كلٍّ بتلقيه ذلك بالإسناد، فكلٌّ فريق يكتب تأييدًا لرأيه، وتسفيهًا لرأي مخالفه، وهذا في كتب المتأخرين على وجه الخصوص. وقد صُنف العديد من الكتب في هذا الموضوع، وهي تتبنى رأيًا واحدًا هو القول بأن النطق الصحيح للضاد هو كما يقرأه القراء المصريون الآن، كالشيخ عبد الباسط عبد الصمد، والشيخ محمود خليل الحصري، والشيخ محمد رفعت، رحمهم الله جميعًا، وكما يقرأ أئمة الحرمين في هذا الزمان -الشيخان سعود الشريم وعبد الرحمن السديس، والشيخ علي عبد الرحمن الحذيفي.

وإلى عهد قريب كانت المصادر التي يمكن الاعتماد عليها في تتبع صوت الضاد لا تزال مخطوطة، ثم حقق غانم قدوري عددا كبيرا من هذه المصادر؛ وحمل على عاتقه عبء الكشف والتنقيب عن مكنونات الدراسات الصوتية عند علماء التجويد، فانكشف للباحثين كثير من النصوص التي توضح رحلة صوت الضاد الحقيقية عبر الأجيال حتى وقتنا القريب.

وقد ذهب غانم قدوري الحمد إلى أن دراسة هذا الصوت لابد أن تنطلق من قاعدتين:

الأولى: أن الأصل في القراءة الاتِّباع، فهو سنة متبعة يأخذها اللاحق عن السابق؛ لقول النبي صلى الله عليه وسلم :[اقرأوا كما عُلِّمتُم].

الثانية: أن أقدم وصف مكتوب للضاد وصل إلينا هو وصف إمام النحاة سيبويه في كتابه العظيم [الكتاب]، وكل من جاء بعده ينقل

[13] أنيس، إبراهيم، الأصوات اللغوية، ص: 68.

[14] بشر، كمال محمد، علم الأصوات، دار غريب، القاهرة، 2000م، ص: 175.

[15] السابق، ص175.

[16] السابق، ص128.

[17] السابق، ص 129.

عنه[18].

ويمكن تلخيص ما ذكره سيبويه عن الضاد في النقطتين التاليتين:

1- مخرج الضاد، ذكر سيبويه أنها تخرج "من بين أول حافة اللسان وما يليها من الأضراس"[19].

فالضاد تميزت بمخرجها، فهي من حافة اللسان من أقصاها، مع ما يقابلها من الأضراس، وكان سيبويه قد ذكر الضاد قبل الجيم حين رتب الحروف، لكنه جعل مخرج الضاد بعد مخرج حروف وسط اللسان [ج ش ي] باتجاه طرف اللسان[20].

2- صفات الضاد. وقد ذكر سيبويه أنها تتصف بالجهر، والرخاوة، والإطباق، والاستعلاء، والاستطالة[21].

فالضاد التي وصفها سيبويه صوت رخو لا ينحبس النفس في مخرجه، مجهور يتذبذب الوتران الصوتيان عند النطق به، مطبق، مستعل، يتميز بالاستطالة.

وبعد ذكر سيبويه لمخرج الضاد وصفاتها ذهب إلى أن "كل حرف فيه زيادة صوت لا يدغم في ما هو أنقص صوتًا منه. وفي الضاد استطالة ليست لشيء من الحروف فلم يدغموها في شيء من الحروف المقاربة لها، إلا ما روي من إدغامها في الشين في قوله تعالى :[لبعض شأنهم] وسوغ ذلك ما في الشين من تفش يشبه الاستطالة يقربها من الضاد. ومن ثم أدغمت اللام والتاء والدال والطاء والثاء والذال والظاء في الضاد، ولم تدغم هي فيها"[22].

وعليه فإن الضاد بهذه الصفات التي ذكرها سيبويه صوت متفرد، ولهذا قال سيبويه : "لولا الإطباق لصارت الطاء دالًا، والصاد سينًا، والظاء ذالًا، ولخرجت الضاد من الكلام؛ لأنه ليس من موضعها شيء غيرها"[23].

وقد بقي ما كتبه سيبويه دستورا للعلماء الذين جاءوا بعده، وفي القرن الرابع الهجري بدأ الأمر يأخذ منحى آخر، حين بدأ الانحراف يظهر في النطق بالضاد وخاصة التباسها بصوت الظاء؛ مما جعل العلماء يكتبون الكتب في التفريق بين الضاد والظاء، وذلك بجمع الألفاظ التي تكتب بالضاد والتي تكتب بالظاء.

وقد قام رمضان عبد التواب بإحصاء المصنفات التي ألفت في التفريق بين الضاد والظاء في تحقيقه لكتاب [زينة الفضلاء في الفرق بين الضاد والظاء] لأبي البركات الأنباري.

وهناك مؤلفات أخرى لعدد من العلماء تناولت الجانب الصوتي، فتحدثت عن خصائص صوت الضاد النطقية، والانحرافات التي تلحقه على ألسنة الناطقين، والأصوات التي يختلط بها أو يقترب منها، وكان لعلماء التجويد مشاركة واضحة فعالة في هذا الأمر، ومن أهم هذه المصنفات:

□ رسالة [غاية المراد في إخراج الضاد]، لابن النجار [870هـ]، التي حققها طه محسن في مجلة المجمع العلمي العراقي في عدد ذي القعدة عام 1408هـ.

□ ورسالة [بغية المرتاد لتصحيح الضاد]، لابن غانم المقدسي [1004هـ]، التي حققها محمد عبد الجبار المعيبد، ونشرها في مجلة المورد العراقية.

□ ورسالة في كيفية الضاد لساجقلي زاده [1150هـ]. وقد طبعت بتحقيق حاتم الضامن.

وقد أكدت هذه الرسائل والمؤلفات حقيقتين:

الأولى: أن هناك تغيرًا صوتيًا يحدث في نطق الضاد.

الثانية: أن علماء التجويد كانوا مشغولين بتحديد ملامح ذلك التغير، وأنهم كانوا حريصين على التمسك بالصورة الأولى لنطق الضاد؛ مراعاة لهدف مصنفاتهم الأول وهو البعد عن اللحنين الجلي والخفي.

وقد أشار سيبويه إلى صوت وليد أسماه [الضاد الضعيفة]، وهي أحد الحروف الفرعية غير المستحسنة لا في قراءة القرآن ولا في الشعر، [إلا أن الضاد الضعيفة تُتَكَلَّف من الجانب الأيمن، وإن شئت تكلفتها من الجانب الأيسر، وهو أخف؛ لأنها من حافة اللسان مطبقة؛ لأنك جمعت في الضاد تكلف الإطباق مع إزالته عن موضعه، وإنما جاز هذا فيها لأنك تحولها من اليسار إلى الموضع الذي في اليمين وهي أخف؛ لأنها من حافة اللسان وأنها تخالط مخرج غيرها بعد خروجها، فتستطيل حين تخالط حروف اللسان، فسهل تحويلها إلى الأيسر؛ لأنها تصير في حافة اللسان في الأيسر إلى مثل ما كانت في الأيمن، ثم تنسلُّ من الأيسر حتى تتصل بحروف اللسان، كما كانت كذلك في الأيمن][24].

وقد كان علماء اللغة والتجويد والتفسير على وعي تام بهذا التغير الطارئ على صوت الضاد، مع وعيهم بالخلط المفتعل الذي قد يحدث بين صوتي الضاد والظاء، وسوف أسرد بعضًا من كلامهم في الضاد لنعرف أن الانحراف في نطق هذا الحرف قديم.

قال مكي بن أبي طالب القيسي [ت 437هـ] :[ولا بد للقارئ من التحفظ بلفظ الضاد حيث وقعت، فهو أمر يُقَصِّرُ فيه أكثر من رأيت من القراء والأئمة ... ومتى فَرَّطَ في ذلك أتى بلفظ الظاء أو بلفظ الذال فيكون مُبَدِّلًا وَمُغَيِّرًا، والضاد أصعب الحروف تكلفًا في المخرج، وأشدها صعوبة على اللافظ، فمتى لم يتكلف القارئ إخراجها على حقها أتى بغير لفظها، وأخل بقراءته][25].

وقال أبو عمرو الداني [ت 444هـ] عن نطق الضاد :[ومن آكد ما على القراء أن يخلصوه من حرف الظاء بإخراجه من موضعه،

[18] انظر: الحمد، غانم قدوري، أبحاث في علم التجويد، ص 146-159، دار عمار للنشر والتوزيع، الأردن.

[19] سيبويه، الكتاب، 433/4.

[20] انظر: سيبويه، الكتاب، 433/4.

[21] السابق، نفس الصفحة.

[22] لنظر: سيبويه، الكتاب 434/4، بتصرف.

[23] سيبويه، الكتاب، 436/4.

[24] سيبويه، الكتاب، 432/4.

[25] القيسي، مكي بن أبي طالب، الرعاية في تجويد القراء وتحقيق لفظ التلاوة، ص: 158-159.

وإيفائه حقه من الاستطالة][26].

وقال ابن كثير[ت 774ه]: "والصحيح من مذاهب العلماء أنه يُغتفر الإخلال بتحرير ما بين الضاد والظاء؛ لقرب مخرجهما وذلك لأن الضاد مخرجها ... فلهذا اغتفر استعمال أحدهما مكان الآخر لمن لا يميز ذلك، وأما حديث : أنا أفصح مَنْ نطق بالضاد فلا أصل له"[27].

وقال عبد الوهاب القرطبي [461هـ] :[وأكثر القراء اليوم على إخراج الضاد من مخرج الظاء، ويجب أن تكون العناية بتحقيقها تامة؛ لأن إخراجها ظاءً تبديل][28].

وكان ابن الجزري [833هـ] قد حدد الأصوات التي يتحول إليها الضاد على ألسنة المعاصرين له فقال في النشر :[والضاد انفرد بالاستطالة، وليس في الحروف ما يعسر على اللسان مثله، فإن ألسنة الناس فيه مختلفة، وقلَّ من يحسنه، فمنهم من يخرجه ظاءً. ومنهم من يمزجه بالذال. ومنهم من يجعله لامًا مفخمة. ومنهم من يُشِمُّه بالزاي. وكل ذلك لا يجوز][29].

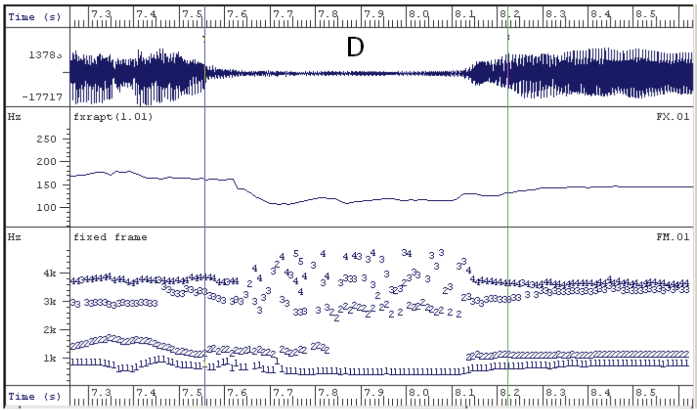
وقال ابن الجزري في التمهيد: [واعلم أن هذا الحرف ليس من الحروف حرف يعسر على اللسان غيره ، والناس يتفاضلون في النطق به: فمنهم من يجعله ظاء مطلقًا... وهم أكثر الشاميين وبعض أهل المشرق. ومنهم من لا يوصلها إلى مخرجها، بل يخرجها دونه ممزوجة بالطاء المهملة، لا يقدرون على غير ذلك، وهم أكثر المصريين وبعض أهل المغرب. ومنهم من يخرجها لامًا مفخمة، وهم الزيالع ومن ضاهاهم][30]. وفي بلاد الحمران جنوب مكة المكرمة يقلبونه لاماً غير مفخمة إلى اليوم فيقولون في البيض والقاضي [بيل والقالي].

وقال الألوسي[ت 1270 ه]: "والفرق بين الضاد والظاء مخرجا أن الضاد مخرجها من أصل حافة اللسان وما يليها من الأضراس من يمين اللسان أو يساره ومنهم من يتمكن من إخراجها منهما والظاء مخرجها من طرف اللسان وأصول الثنايا العليا واختلفوا في إبدال أحدهما بالأخرى هل يمتنع وتفسد به الصلاة أم لا فقيل تفسد قياسا ونقله في المحيط البرهاني عن عامة المشايخ ونقله في الخلاصة عن أبي حنيفة ومحمد، وقيل: لا استحسانا ونقله فيها عن عامة المشايخ كأبي مطيع البلخي ومحمد بن سلمة وقال جمع أنه إذا أمكن الفرق بينهما فتعمد ذلك وكان مما لم يقرأ به كما هنا وغير المعنى فسدت صلاته وإلا فلا لعسر التمييز بينهما خصوصا على العجم وقد أسلم كثير منهم في الصدر الأول ولم ينقل حثهم على الفرق وتعليمه من الصحابة ولو كان لازماً لفعلوه ونُقل، وهذا هو الذي ينبغي أن يعوَّل عليه"[31].

وأرخ سالم السحيمي من المحدثين لهذا الخلط بين الصوتين فذكر أن "العرب كانت تفرق بين هذين الصوتين تفريقا واضحا في الرسم والنطق، وقد ظهر الفرق بينهما جليا في النقوش اليمنية التي كتبت بالخط المسند، وإنما سبب الخلط بينهما فساد اللغة، ولعل ذلك كانت نتيجة لاختلاط العرب بغيرهم من الأمم الأخرى، وقد وضح القاضي محمد بن نشوان في مختصره الذي ألفه في الفرق بين الضاد والظاء أن العرب كانت تميز بين هذين الصوتين تمييزا واضحا"[32].

ثم أكد أن "بين الضاد والظاء فرق واضحا في اللفظ والمخرج والخط، فأما اللفظ فصميم العرب لا يخلطون بعضهما ببعض ويميزون إحداهما عن الآخر، فلا يقع عندهم بينهما اشتباه، كما لا يشتبه سائر الحروف، حتى إن بعضهم يميل في نطق الضاد إلى شين لقرب مخرج الشين من مخرج الضاد، وبعضهم يميل في نطق الظاء إلى الثاء لقرب مخرجها منها"[33].

وخلاصة الأمر أن علماء العربية والتجويديين وأهل التفسير كانوا على وعي تام بالمفارقة النطقية بين صوتي الضاد والظاء، وكلاهما له موقعه الخاص على خارطة الأصوات اللغوية العربية، غير أنني أرى أن الخلط قد نشأ عند بعض أهل الأداء نظرًا لمحاولتهم تطبيق صفة الاحتكاكية التي أكد عليها علماء العربية والتجويد والتفسير جميعًا جريًا وراء حديث سيبويه في وصفه لهذا الصوت، وإذا كنت قد ذكرت من قبل عدم قبولي لفكرة التطور اللغوي في الصوت القرآني فإنني سأعود وأؤكد أن صوت الضاد الذي نسمعه من أئمة القرآن في هذا العصر هو هو بنفس مخرجه وصفاته كما نطقه النبي صلى الله عليه وسلم وصحابته الكرام منذ بداية نزول القرآن وحتى عصرنا هذا، إنه صوت شديد انفجاري غير لين أو احتكاكي، وسوف أختم حديثي حول هذه الإشكالية بعرض صورة طيفية لهذا الصوت المميز كما نطقه فضيلة الشيخ محمود خليل الحصري على هذا النحو:



الشكل رقم [3]

صورة طيفية لصوت الضاد من كلمة [ولا الضالين] من قوله تعالى: [صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين]، سورة الفاتحة، الآية [7].

[26] الداني، أبو عمرو بن سعيد، التحديد في الإتقان والتجويد، تحقيق غانم قدوري الحمد، مطبوعات جامعة بغداد، 1998م ص: 164.

[27] ابن كثير، تفسير القرآن العظيم، مكتبة السنة، 54/1.

[28] القرطبي، عبد الوهاب، الموضح، 114.

[29] ابن الجزري، النشر في القراءات العشر، 219/1.

[30] ابن الجزري، التمهيد في علم التجويد، 140-141.

[31] الألوسي، روح المعاني 30 /61.

[32] سالم السحيمي، إبدال اللهجات العربية، ص 428.

[33] السابق، ص: 429.

ويضع لنا الشكل السابق تحليلًا لصوت الضاد على ثلاثة مستويات، ففي المستوى الأول [wav form] لا نجد أثراً للذبذبات التي تصاحب الأصوات المجهورة الاحتكاكية، وفي المستوى الثاني نلاحظ انخفاض خط منحنى التنغيم الأساسي [Fundamental Frequency]، وظهوره على هيئة مقعرة، الأمر الذي يؤكد عدم احتكاكية هذا الصوت، بينما نجد في المستوى الثالث [formants] توزيعا عشوائيًا غير متتابع لقيم المعالم الأولى والثانية والثالثة [f1.f2.f3] مما يدل على انتماء هذا الصوت إلى مجموعة الأصوات الشديدة الانفجارية وليست مجموعة الأصوات الرخوة. وعليه فإن صوت الضاد صوت شديد انفجاري وليس لينًا رخوا.

نتائج الدراسة:

1. أظهر التحليل الصوتي الحاسوبي لصوتي الطاء والقاف، عدم وجود ذبذبات وترية في المستوى الأعلى الذي يمثل شكل الموجة [wave form]، في حين يوضح المستوى الثاني المتعلق بالصورة الطيفية [Spectrogram] خلو هذا الصوت من الذبذبات المجهورة، أما المستوى الأخير الذي يحدد النغمة الأساسية [Fundamental Frequency] فنلاحظ فيه انقطاع الخط القاعدي، وهو أمر ملازم للأصوات المهموسة. وعليه فإنهما صوتان مهموسان لا تظهر فيهما أية معالم من معالم الجهر.
2. إن اختلاف مفهوم الجهر والهمس بين سيبويه ومن تبعه من ناحية؛ وبين علماء الأصوات المحدثين من الناحية الأخرى هو الذي أدى إلى الاختلاف في تصنيف بعض الأصوات العربية وتوزيعها إلى أحد القسمين.
3. الهمزة صوت لا بالمجهور ولا بالمهموس؛ ذلك أنه في حال نطق هذا الصوت ينطبق الوتران انطباقاً تاماً؛ فلا يسمح بمرور الهواء إلى الحلق، ومن ثم ينقطع النفس، ثم يحدث أن ينفرج هذان الوتران، فيخرج صوت انفجاري نتيجة لاندفاع الهواء الذي كان محبوسًا حال الانطباق التام، فهمزة القطع العربية إذن صوت صامت لا هو بالمهموس ولا بالمجهور. حيث إنه لا توجد أعضاء نطق مستعملة في إنتاج هذا الصوت، ولكن الأوتار الصوتية تقوم بدور هذه الأعضاء، لتنتج غلقا تامًا – وإن كان قصيرا – في مجرى تيار الهواء. وحيث إن الأوتار الصوتية نفسها هي المنتجة لهذا الصوت فلا معنى لوصفه بأنه مجهور أو مهموس أو موشوش.
4. أما بالنسبة فيما يتعلق بالضاد فإنَّ الخلط بينه وبين الظاء قد نشأ عند بعض أهل الأداء نظرًا لمحاولتهم تطبيق صفة الاحتكاكية التي أكد عليها علماء العربية والتجويد والتفسير جميعًا؛ جريًا وراء حديث سيبويه في وصفه لهذا الصوت، غير أنني لا يمكنني قبول فكرة التطور اللغوي لهذا الصوت القرآني، وعليه فإنَّ صوت الضاد الذي نسمعه من أئمة القرآن في هذا العصر هو هو بنفس مخرجه وصفاته كما نطقه النبي صلى الله عليه وسلم وصحابته الكرام، فهو صوت شديد انفجاري غير لين أو احتكاكي.

ثبت المراجع

[1] أنيس، إبراهيم: الأصوات اللغوية، ط4، مكتبة الأنجلو المصرية، (1971م).
[2] أيوب، عبد الرحمن: الكلام إنتاجه وتحليله، مطبوعات جامعة الكويت، (1994م).
[3] بشر، كمال محمد: علم اللغة العام (الأصوات)، ط2، دار غريب، (1971م).
[4] بعبولة، سيد: البرهان في تجويد القرآن، ط2، مطبعة الإيمان (2002م).
[5] التوني، مصطفى زكي: النون في اللغة العربية "دراسة لغوية في ضوء القرآن الكريم"، حوليات كلية الآداب جامعة الكويت، الحولية السابعة عشرة، (1416-1417هـ)، (1996- 1997م).
[6] ابن الجزري، محمد بن محمد بن محمد: التمهيد في علم التجويد، ط1، القاهرة، (1908م).
- متن الجزرية في معرفة تجويد الآيات القرآنية، مكتبة صبيح بالأزهر، (1956م).
[7] ابن جني، أبو الفتح عثمان: سر صناعة الإعراب، تحقيق مصطفى السقا وآخرين، ط1، مطبعة مصطفى البابي الحلبي وأولاده بمصر، (1374هـ).
[8] حسان، تمام: مناهج البحث في اللغة، ، ط2، دار الثقافة، الدار البيضاء، (1394م، 1974م).
[9] الحصري، محمود خليل: أحكام قراءة القرآن الكريم، ط1، مكتبة السنة، (2000م).
[10] الحمد، غانم قدوري: أبحاث في علم التجويد، دار عمار للنشر والتوزيع، الأردن.
[11] الداني، أبو عمرو بن سعيد: التحديد في الإتقان والتجويد، تحقيق غانم دوري الحمد، ط1، مطبوعات جامعة بغداد،(1998م).
[12] ابن دريد، أبو بكر محمد بن الحسن: جمهرة اللغة، دار صادر بيروت، (طبعة بالأوفست).
[13] الزجاج، أبو إسحاق إبراهيم بن السري: إعراب القرآن، تحقيق عبد الجليل عبده شلبي، مطبوعات الهيئة العامة لشئون المطابع الأميرية، القاهرة، (1973م).
[14] السعران، محمود: علم اللغة مقدمة للقارئ العربي، دار المعارف بمصر، 1962م.
[15] أبو سكين، عبد الحميد محمد: دراسات في التجويد والأصوات اللغوية، مطبعة الأمانة، القاهرة، 1404هـ/ 1983م.
[16] سيبويه، أبو بشر عمرو بن عثمان بن قنبر: الكتاب، تحقيق: عبد السلام محمد هارون، ط1، دار الجيل، بيروت.
[17] السيرافي، أبو سعيد الحسن بن عبد الله: شرح كتاب سيبويه، مخطوط بدار الكتب المصرية، رقم (528 نحو تيمور).
[18] الضالع، محمد صالح: التجويد القرآني "دراسة صوتية فيزيائية"، دار غريب 2002م.
[19] ضوة، إبراهيم: محاضرات في اللغة العربية والحاسب، ط1، دار الثقافة العربية، (2000م).
[20] العاني، سلمان حسن: فونولوجيا العربية، ترجمة ياسر الملاح، ط1، مطبوعات النادي الأدبي الثقافي بجدة، (1403هـ - 1983م).
[21] عبد البَاقي، نعيم: قواعد تشكل النغَم في مُوسيقى القرآن، مجلة التراث العربي، النسخة الإلكترونية، العدد25، "أكتوبر"(1986م).
[22] ابن الطحان، أبو الإصبع السماتي الأشبيلي: مخارج الحروف وصفاتها، تحقيق محمد يعقوب تركستاني.
[23] عمر، أحمد مختار: دراسة الصوت اللغوي، عالم الكتب، (2000م).
[24] الغامدي، منصور بن محمد: الصوتيات العربية، ط1، مكتبة التوبة، (2000م).
[25] فتيح، محمد: الأصوات العامة والأصوات العربية، دار الثقافة العربية، القاهرة.
[26] قدور، أحمد محمد: أصالة علم الأصوات عند الخليل من خلال مقدمة كتاب العين، ط1، دار الفكر المعاصر، بيروت- لبنان، (1419هـ/1998م).
[27] القيسي، مكي بن أبي طالب: الرعاية لتجويد القراءة وتحقيق لفظ التلاوة، تحقيق أحمد حسن فرحات، دمشق، (1973م).
[28] ليونز، جون: اللغة وعلم اللغة، ترجمة مصطفى زكي التوني، دار النهضة العربية، (1988م).
[29] المقدسي، عبد الرحمن بن إسماعيل أبو شامة: إبراز المعاني من حرز الأماني، تحقيق غانم قدوري الحمد.
[30] ابن منظور، أبو الفضل محمد بن مكرم: لسان العرب، ، ط1، مطبعة بولاق.
[31] الموسوي، مناف: علم الأصوات اللغوية، ط1، عالم الكتب، بيروت، (1998م).
[32] نصر، عطية قابل: غاية المريد في علم التجويد، ط 6، دار الحرمين للطباعة، القاهرة.
[33] نصر، محمد مكي: نهاية القول المفيد في علم التجويد، مكتبة الحلبي، (1349هـ).
[34] هلال، عبد الغفار حامد: أصوات اللغة العربية، ط2، مكتبة الأنجلو المصرية، (1988م).
[35] الوعر، مازن: صلة التراث اللغوي العربي باللسانيات، مجلة التراث العربي، اتحاد الكتاب العرب، دمشق، النسخة الإلكترونية.

السيرة الذاتية:

أحمد راغب أحمد

– أستاذ مشارك بقسم اللغة العربية بكلية معارف الوحي والعلوم الإنسانية، بالجامعة الإسلامية العالمية بماليزيا. حصل على الدكتوراه في اللغويات الحاسوبية مع مرتبة الشرف الأولى، من قسم علم اللغة بكلية دار العلوم، جامعة القاهرة،، وله منشورات عن التحليل الصوتي الحاسوبي وسبل معالجة الإنحراف اللغوي في الترجمة الآلية وتطوير آليات التعليم الإلكتروني، كما شارك في تطوير مناهج الحوسبة اللغوية وتكنولوجيا التعليم بالجامعة الإسلامية العالمية بماليزيا.

Controversial Voices Issues In the Light of Computational Speech Analysis

Ahmed Ragheb Ahmed
Kulliyyah of Islamic Revealed Knowledge and Human Sciences - International Islamic University Malaysia
ragheb@iium.edu.my

***Abstract*- Arabic Phonology has a standard pronunciation, language used to recite the Islamic sacred book Qu'ran. The first codification of the Arabic language was undertaken by early Arab linguistician who regarded the language of the Quran as the model of correctness. This was the first time the Arabic language methods was adjusted and standardized with an explicit recitation grammar defining correct usage. The codification included all its linguistic levels such phonetics, phonology, morphology, syntax and semantics. There is no doubt Qu'ran is a best source, for over decade and to the present, this grammar of classical Arabic is still taught to all Arabic speakers in their general education courses/studies in Arabic Speech. Thus, the function of Arabic language extended beyond the communicative needs of its native speaker as it served as the standard that all Arabic speakers aspire to master . This study examines range of contention issues between traditional and modern linguists. It is a study to compare both theory and practical based on technical analysis of language phonetic sound. It is not to adjust the efforts made by grammarians but to perceive and comprehend linguists and preserve the proper standard pronunciation of the Quran**.

Topic Clustering of Stemmed Transcribed Arabic Broadcast News

Ahmed Abdelaziz Jafar*[1], Mohamed Waleed Fakhr*[2], Mohamed Hesham Farouk **[3]

**Department of Computer Science, College of Computing and Information Technology, Arab Academy for Science and Technology (AAST), Cairo, Egypt*

[1]a.jaf84@gmail.com

[2]waleedf@aast.edu

***Department of Engineering Math & Physics, Faculty of Engineering, Cairo University, 12613 Egypt*

[3]mhesham@eng.cu.edu.eg

***Abstract*— in this research different clustering techniques are applied for grouping transcribed textual documents obtained out of audio streams. Since audio transcripts are normally highly erroneous, it is essential to reduce the negative impact of errors gained at the speech recognition stage. In attempt to overcome some of these errors, different stemming techniques are applied on the transcribed text. To further improve the clustering accuracy, documents causing topic confusion are detected by using fuzzy and possibilistic techniques and then excluded from the dataset. The goal of this research is to achieve automatic topic clustering of transcribed speech documents, and investigate the impact of applying stemming techniques in combination with a Chi-square similarity measure on the accuracy of the selected clustering algorithms. The evaluation has showed that using rule-based light stemming in combination with spectral clustering technique achieved the highest accuracy, and this accuracy is further increased after excluding the confusing documents by using the possibilistic GK algorithm.**

1 INTRODUCTION

The growing amount of audible news broadcasted on TV channels, radio stations and on the Internet demands reliable and fast techniques to organize and store those vast amounts of news in order to facilitate future search and retrieval.

In this work, Automatic Speech Recognition (ASR) – a technology that converts spoken words to written text – is applied to audible Arabic news documents. Then a set of pre-processing and clustering techniques are applied on the transcribed documents in order to categorize them into a set of predefined topics.

Since the transcription process is normally highly erroneous [11] and as an attempt to overcome some of these errors two pre-processing steps: words formatting and stemming are considered to evaluate their impact on limiting the negative impact of such errors. Three stemming techniques are selected: light stemming [17], root-based stemming [13], and rule-based light stemming [12].

At the clustering stage, two similarity measures are used: Chi-square and the traditional cosine similarity measure. Chi-square similarity measure is based on the Chi-square method [11]. This similarity measure is designed to eliminate non informative words (usually erroneous words when applied on transcribed documents). Two clustering techniques are utilized to achieve topic clustering: *k*-means, [24] and spectral clustering, [18]. *K*-means is selected as a simple and fast traditional clustering algorithm. Spectral clustering is selected as it is one of popular, effective, and simple to implement modern clustering techniques.

To further enhance the topic-clustering accuracy, the fuzzy *c*-means [8] and a possibilistic [15] version of the fuzzy clustering algorithm "Gustafson–Kessel (GK)" [10] is used to measure the degree of membership of each document to every topic cluster, hence all documents that don't belong vividly to one topic can be identified and scheduled for manual topic assignment.

This research is organized as follows: in section 2, speech transcription challenges are overviewed. In Section 3, the used ASR system's accuracy is evaluated. In Section 4, data set pre-processing steps are discussed. In Section 5, topic clustering is discussed in details. In Section 6, experimental results are evaluated and discussed in section 7. The last section concludes the research.

2 SPEECH TRANSCRIPTION CHALLENGES

The process of transcribing audible media to textual form using ASR system confronts many challenges that are typically not present in normal textual documents [11]. The main challenges include: transcription errors, grammatical errors, and out-of-vocabulary problem (OOV), or combination of the previously mentioned problems.

The occurrence of such problems can seriously restrict the transcription process efficiency and hence restricts any further analysis applied on the transcripts. This work targets the overcoming the problem of transcription errors as it is the most common problem when dealing with news transcripts.

The transcription errors occur due to limitation in the ASR system. The correction or elimination of such errors is a challenging task and requires understanding the nature of these errors. According to authors' observation, the transcription errors regarding Arabic language can be categorized into four sets:

1) *Omission errors*: happen when the ASR fails completely to recognize a word or a series of consecutive words. In this case, the words are dropped out from the transcribed text and recognition process is continued. Omission errors are irrecoverable.
2) *Word insertion errors*: occur when the ASR confuses word syllables with a separate word or multiple words. In this kind of errors, the original word is irrecoverable.
3) *Misidentification errors*: identifying a pronounced word as a different word similar in pronunciation. The transcribed word may or may not belong to the valid Arabic vocabulary set.
4) *Minor spelling errors*: a spoken word is identified correctly, but spelled wrong in transcription. These errors usually affect the way a word should be pronounced and it may affect its meaning as well. Common minor spelling errors generated by ASR are replacing the letter 'ه' with 'ة' at the end of the word and vice versa, replacing one of the following letters with one another 'ا', 'أ', 'إ' , and 'آ', and diacritics related errors.

3 ASR System's Accuracy Evaluation

The performance of speech recognition systems is usually evaluated in terms of accuracy and speed. Accuracy is usually rated with Word Error Rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

The WER [53] is commonly used to measure speech recognition performance. It is based on the frequency of occurrences of three types of errors: substitutions–a reference word is replaced by another word, insertions –a word is hypothesized that was not in the reference and deletions–a word in the reference is missed. WER is calculated as in (4.1).

$$WER = \frac{\#Substitutions + \#Insertions + \#Deletions}{\#Reference\ Words} \quad (1)$$

WER is typically calculated by matching the reference and the corresponding transcriptions and it can be over 100%. Table I and Table II show the performance evaluation of the Dragon Dictation recognition system before and after removing stop words. It is notable that removing stop words in the preprocessing phase has reduced WER to ≈20.65% instead of ≈29.11%; it also reduced the vocabulary size by ≈56.29%, which is great for reducing storage size and any further processing time.

TABLE

OF THE DRAGON DICTATION RECOGNITION SYSTEM BEFORE REMOVING STOP WORDS

Reference Words	Substitutions%	Insertion%	Deletion%
68720	17327	2105	574
WER %	29.1123399		

TABLE I

WER OF THE DRAGON DICTATION RECOGNITION SYSTEM AFTER REMOVING STOP WORDS

Reference Words	Substitutions%	Insertion%	Deletion%
30040	3607	1831	766
WER %	20.6524634		

The Substitution errors have the highest effect on the accuracy of this system as it form ≈86.6% of all the errors. Fortunately, many of those errors occur among stop words, and it occurs in a way that can be fixed during the stop words removal phase of the preprocessing, which reduce this percentage to ≈58.14%. An example of a common substitution error related to stop words is substituting the article "ان" with either "انه" or "انا" according to the pronunciation. In both cases they are both valid stop words and hence would be removed. Substitutions that occur among stop words represent ≈79.18% of all the substitutions.

Insertion and deletion errors are much less common than substitution errors as they represent ≈3.1% and ≈0.84% respectively of the whole reference words. Compared to substitution errors that represents ≈25.21% of the reference words, Insertion and deletion errors are acceptable. Insertion errors count is slightly decreased by removing stop words whereas deletion errors count is increased (Table II). Handling insertion and deletion errors is beyond the scope of this research.

The decrement in the insertion errors count can be explained as some of the inserted words are stop words, and thus can be removed safely. For Example the word “الذهاب” could be transcribed into “لا ذهاب” confusing the definite article “ال” with the stop word “لا”. This counts as insertion error until fixed by removing stop words. The conclusion is that stop words removal didn’t actually help much with this type of errors.

The increment in deletion errors count is due to the transformation from some of substitution errors to deletion errors after removing stop words. For example, in one of the transcriptions, the phrase “بحلول الخامس من فبرايرالجاري”, is transcribed to “بعد الخامس من فبراير الجاري”. By comparing the two phrases, it is clearly obvious that the word “بحلول” is substituted by “بعد”. Now after removing stop words from both phrases they become: “بحلول الخامس الجاري” and “ الخامس الجاري” respectively. Comparing the two phrases after removing stop words would indicate that the word “بحلول” is now missing as its substituted word “بعد” is removed as being one of the stop words.

4 DATASET PRE-PROCESSING

As any written text documents, the transcribed documents produced by the ASR system are of unstructured format. Thus it is required to transform these unstructured documents to structured format using pre-processing in order to facilitate any further analysis applied on them. The following are the steps involved in the pre-processing applied in this work.

1) *Tokenization:* the process of mapping sentences from character strings into strings of words. For example, the sentence “اللغة العربية تعد من أشهر اللغات” would be tokenized into “اللغة/”, “العربية/”, “تعد/”, “من/”, “أشهر/”, “اللغات/”.
2) *Stop words removal:* Stop words are typical frequently occurring words that have little or no discriminating power, or other domain-independent words. Stop words removal can increase the effectiveness of the information retrieval process [2], [14], especially when dealing with large volume of text [21]. Stop words identified in this work include numbers, days and months names, prepositions, pronouns, and conjunctions.
3) *Words Formatting:* An extra step applied in this work to unify all different shapes of the same letter to one form and also to remove some unwanted suffixes.
4) *Stemming:* Removes the affixes in the words and produces the root word known as the stem. Typically, the stemming process will be performed so that the words are transformed into their root form. Automatic Arabic stemming is effective technique for text processing for small collections as in [4], [5] and large collections of documents as in [16], [17]. It also can enhance clustering as in [5]. Arabic stemmers are categorized as either root-based as in [6], [13] or stem-based (light stemmers) as in [16], [17]. Also the research for hybrid techniques, like the rule-based light stemming technique [12], has evolved to minimize the drawbacks associated with standard stemming techniques.
5) *Weighted matrix construction:* the process of representing the text document into a machine readable form [12].

Besides transforming the unstructured transcribed text to structured form, the pre-processing is also used as the first phase to reduce the transcription errors by either correcting or help overcoming some of these errors. This happens during the pre-processing steps: words formatting and stemming. The following discuss how applying pre-processing can correct or help overcoming some of the transcription errors.

In Fig. 1, and Fig. 2 erroneous words like “بدات”, “امس”, “اذا”, “الفتره”, “التجريبيه” and “ملاحقه” are examples of minor spelling errors. According to Arabic syntactic rules [3], the correct spelling for these words should be: “بدأت”, “أمس”, “إذا”, “الفترة”, “التجريبية” and “ملاحقة”. Such a problem is common in Arabic ASR systems and leaving it without handling would cause problems in any further analysis as in any computer system words like “بدات” and “بدأت” aren’t the same and actually will process them as two separate words. In this kind of errors, the correct spelling is determined on syntactic rules that depend in most cases on pronunciation. The correct pronunciations depends on the meaning of the word which is determined according to its context, thus the only way to detect and correct such errors is searching massive dictionary of correctly spelled Arabic words and searching for the correct syntax if the word is misspelled. It is possible for a word to take many correct forms in different contexts, hence automatically selecting the correct form needs understanding the word context. Such process is inefficient in terms of the processing time and power required as Arabic is very complex language, moreover, many existing machine learning techniques don’t require understanding the language structure to operate on text, so it is not efficient to do such thing as a pre-processing step just to correct some errors. The most suitable solution is not to correct these errors but to work around them by unifying all different formats of a letter into one form. The unification process is performed at the words formatting step. Some suffixes are also removed at the word formatting step to fine-tune the input text for the stemming step.

Three stemming techniques are utilized in this work: light stemming represented in Larkey's light10 stemmer, root-based stemming represented in Khoja's root-based stemmer, and rule-based light stemmer. The stemming techniques are applied in this work to unify vocabulary and also to overcome some transcription errors.

The words "تتويجا", "اللعب", "أقدم", "داخلة", "لحس", "لكونى", "قبلة", and "انتهى" in Fig. 1 are examples of misidentification errors. The original words are "توج", "تغلب", "قد", "دخل", "لحسم", "لكونه", "قبل", and "انتهاء". If a word is misidentified into one of its relative forms, so there is a good chance that this mistake would be overcame when root-based stemming is applied. The light and rule-based stemmers would either transform the word to another form or leave it without any transformation in some cases.

Both light and rule-based stemmers actually tend to correct the error in case of occurrence of inserted letters in the start and the end of the word as long as the inserted letters exist on their prefix/suffix removal list. The root-based stemmer can also correct an erroneous word if by chance the original word is the same as the root of erroneous word. The word "تتويجا" (Fig. 1) is a good example, after removing the suffix 'ا' at the word formatting step the result is "تتويج" which would be transformed by the root-based stemmer to "توج" (Fig. 3) which is also the same as the original word in spelling.

تتويجا يوفنتوس رسميا بلقب الدوري الإيطالي لكرة القدم للمرة الثانية على التوالي وذلك بعدما اللعب على ضيفه باليرمو
وأقدم داخلة فريق المدرب انطونيو كونتي المباراة وهو بحاجة الي نقطة 1 فقط لحس اللقب للمرة ثانية على التوالي والتاسعة والعشرون في تاريخه لكوني
يتصدر الترتيب بفارق ١١ نقطة عن ملاحقة نابولي قبلة أربعة مراحل على انتهى الموسم

Figure 1: Sample of transcribed text with various transcription errors

تتويج يوفنتوس رسم بلقب دور ايطال لكر قدم مر ثان توال لعب ضيف باليرم
اقدم داخل فريق مدرب انطوني كونت مبارا بحاج نقط فقط لحس لقب مر ثاني توال والتاسع عشر في تاريخ لكون يتصدر ترتيب بفارق نقط ملاحق
نابول قبل اربع مراحل انته موسم

Figure 2: Sample of transcribed text after applying pre-processing with light stemming

توج يوفنتوس رسم لقب دور ايطالي كور قدم مرر ثاني ولي لعب ضيف باليرم
قدم دخل فرق درب انطوني كونتي برا حوج نقط فقط لحس لقب مرر ثاني ولي تاسع عشرون ارخ كون صدر رتب فرق نقط لحق نابولي قبل اربعه رحل
نهي وسم

Figure 3: Sample of transcribed text after applying pre-processing with root-based stemming

تتويج يوفنتوس رسم لقب دور إيطال كرة قدم مره ثان وال لعب ضيف اليرم
أقدم داخل فريق مدرب انطوني كونت مباراه حاج نقط فقط لحس لقب مره ثان توال التاسع عشر تاريخ كون يتصدر الترتيب بفارق نقطه عن ملاحق نابول
قبل اربعه مراحل على انته موسم

Figure 4: Sample of transcribed text after applying pre-processing with root-based stemming

All three stemming techniques fail when the erroneous word is substituted by completely another word of a different spelling and meaning like "اللعب", "أقدم", and "لحس" as in Figures 2-4 or if by chance a letter is inserted in the middle of the word. If such erroneous words are not repeated frequently along the whole set of documents, they would probably have poor information contribution, and hence they would be eliminated by the chi-square based similarity measure if their information contribution assessment doesn't comply with a certain threshold, otherwise they would be retained.

After applying stop words removal, words formatting and stemming, and in order to process the transcribed documents for topic clustering, they must be represented in a machine readable form. Vector-Space Model (VSM) the model applied in this work because of its effectiveness in proximity estimation between text documents in addition to its conceptual simplicity [12].

In VSM all documents are represented as vectors of weights in an n-dimensional space of terms. At the recent time, there are a number of well-known methods that have been developed to evaluate term weight [23], in this work, Okapi method

[19] is applied, which is a modification of a classic TFIDF (Term Frequency × Inverse Document Frequency) weighting scheme and proved to be efficient in a number of applications [1], [7].

According to the Okapi method Combined Weight of the word (CW) is calculated as in (1).

$$CW(w_i|D_j) = \frac{(K+1) \times CFW(w_i) \times TF(w_i, D_j)}{K \times ((1-b) \times +b \times NDL(D_j) + TF(w_i, D_j)} \tag{1}$$

The quantity $CFW(w_i) = \log \frac{N}{n(w_i)}$ is the data set collection frequency weight; *N* is the total number of documents in the collection and $n(w_i)$ is the number of documents containing the word w_i. The quantity $TF(w_i, D_j)$ is the frequency of occurrences of word w_i in the document D_j and $NDL(D_j)$ is the length of the document D_j normalized by the mean document length. The constant *b* controls the influence of document length and is empirically determined to the value 0.75. The other constant *K* acts as a discounting parameter on the word frequency: when *K* is 0, the combined weight reduces to the collection frequency weight; as *K* increases the combined weight approaches $tf \times itf$. *K* is set to 1.25 in this work.

Once *CW* is calculated for all words, it is easy to calculate the Weight of a Document (*DW*) the same way. *DW* can be calculated for the whole document or any of its parts via applying (2).

$$DW(D_i) = \sum\nolimits_{w_k \in X_i} CW(w_k) \tag{2}$$

5 Topic Clustering

Topic clustering is the process of assigning one or more labels to text documents chosen from a pre-defined list of topics using a similarity measure. A Chi-square based similarity measure and cosine similarity is used along with *k*-means and spectral clustering algorithms.

The Chi-square similarity measure determine the word co-occurrences between matching transcripts by sorting all words in transcripts by their weights and retain only those whose weights are greater than some empirically preset threshold. Thus non-informative words including low frequently repeated erroneous words should appear at the bottom of the sorted list and hence eliminated according to the empirically determined threshold. The Chi-square similarity is calculated as in (3).

$$sim(Inter(D_i, D_j) = \sigma \times Inter(D_i, D_j), \tag{3}$$

where σ is given by evaluating the Chi-square test in (4) and $Inter(D_i, D_j)$ is given by (5). The calculated similarity will range between 0 and 1 and it will be equal to 1 if and only if $D_i = D_j$.

$$X^2 = \sum\nolimits_{w_k \in D_i \cap Dj} \frac{(CW(w_k \mid w_k \in D_i) - CW(w_k \mid w_k \in D_j))^2}{CW(w_k \mid w_k \in D_j)} \tag{4}$$

$$Inter(D_i, D_j) = \frac{DW(D_i \cap D_j)}{DW(D_i)}, \tag{5}$$

such that in case of $D_i \neq D_j$ is true, then the inequality in (6) is also true.

$$Inter(D_i, D_j) \neq Inter(D_j, D_i) \tag{6}$$

The widely used cosine similarity (7) measure is also used. The clustering accuracy is, then, compared to the accuracy achieved via using the Chi-square measure.

$$S_c(D_i, D_j) = \frac{\sum_{k=1}^{N} (w_{ki} \times w_{kj})}{\sqrt{\sum_{k=1}^{N} w_{ki}^2 \times \sum_{k=1}^{N} w_{kj}^2}} \tag{7}$$

For Topic clustering process, three clustering approaches are used: hard clustering, fuzzy clustering, and possibilistic clustering.

A. *Hard Clustering*

Hard clustering means partitioning the data into a number of subsets (clusters) such that an object either belong or doesn't belong to a cluster. Two hard clustering techniques are utilized in this work: k-Means, spectral clustering.

K-means is based on the idea that a center point (centroid) can represent a cluster. It is one of the most popular traditional data clustering algorithms because of its simplicity and computational efficiency. The main problem with this clustering method is its tendency to converge at a local minimum and the final results highly depends on the initial choices of centroids.

Spectral clustering reformulation of the clustering process takes place using a similarity graph *G=(V, E)* where the goal is to find a partition of the graph such that the edges between different groups have very low weights, and the edges within a group have high weights. The similarity graph used in this work is the fully connected graph because the Chi-square similarity measure itself models local neighborhoods, so it best suite this kind of graphs as described in [22].

B. *Fuzzy Clustering*

Besides the fact that the transcribed data are being erroneous, there is also a possible chance of the existence of high percentage of topic overlaps and/or noisy documents that don't belong to any predefined topic, which limits any effort to correctly cluster such data into topics using hard clustering techniques. Thus the need emerged for a clustering method that allows a document to belong to more than one cluster simultaneously with different membership degrees. Fuzzy clustering method [25] allows object memberships satisfying the following constraints:

$$\mu_{ij} \in [0,1], \forall_i, \forall_j \tag{8}$$

$$0 < \sum_{j=1}^{N} \mu_{ij} < N, \forall_i, and \tag{9}$$

$$\sum_{i=1}^{C} \mu_{ij} = 1, \forall j. \tag{10}$$

The parameter μ_{ij} is the degree of membership of the feature point *x* in cluster β_i. *C* denotes the number of classes, and *N* denotes the total number of feature points. The fuzzy clustering technique utilized in this work is the fuzzy *c*-means algorithm.

The fuzzy *c*-means algorithm is based on the optimization of the following basic objective function:

$$J(L,U) = \sum_{i=1}^{C} \sum_{j=1}^{N} (\mu_{ij})^m d_{ij}^2 \tag{11}$$

The parameter *L*=($\beta_1, \ldots, \beta_C$) is the collection of cluster prototypes, and *U*=[μ_{ij}] is a *C*×*N* fuzzy *c*-partition matrix (membership matrix) satisfying the conditions (8), (9), and (10). d_{ij}^2 is the distance of feature point x_j to cluster center c_i (12).

$$d_{ij}^2 = \| x_j - c_i \|^2 \ A = (x_j - c_i)^T A(x_j - c_i), \tag{12}$$

The fuzzy c-means algorithm is explained in [8].

C. *Possibilistic Clustering*

Because of the restriction in (10), the generated memberships degree don't always correspond well to the actual degree of belonging of the data, thus it is difficult to detect outliers in a noisy environment using fuzzy clustering. As a solution to this problem, possiblistic clustering is used, in which the restriction in (10) is relaxed [6] as in (13). A possibilistic version of the fuzzy GK algorithm is utilized in this work.

$$\max_i \mu_{ij} > 0, \forall_j \tag{13}$$

GK extended the standard fuzzy c-means algorithm by employing an adaptive distance norm, in which each cluster has its own norm-inducing matrix A_i, which yields the inner-product norm in (14). The choice of the norm-inducing matrix A determines the cluster shape; hence the employing of adaptive distance norm adds the capability of detecting clusters of different geometrical shapes.

$$d_{ij}^2 = \| x_j - c_i \|^2 \; A_i = (x_j - c_i)^T A_i (x_j - c_i), \tag{14}$$

The GK algorithm is based on the minimization the following objective function:

$$J(L,U) = \sum_{i=1}^{C} \sum_{j=1}^{N} (\mu_{ij})^m d_{ij}^2 \tag{15}$$

This objective function cannot be directly minimized with respect to A_i, since it is linear in A_i, thus the determinant of A_i must be constrained as in (16).

$$| A_i | = \rho_i, \rho_i > 0, \forall_i. \tag{16}$$

Fixing the determinant of the matrix A_i while allowing it to vary leads to optimizing the cluster's shape while its volume remains constant. The expression for A_i is obtained by using the Lagrange multiplier method as follows:

$$A_i = [\rho_i \det(F_i)]^{1/n} F_i^{-1}, \tag{17}$$

where F_i is the fuzzy covariance matrix of the *i*th cluster given by (18).

$$F_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m (x_j - c_i)(x_j - c_i)^T}{\sum_{j=1}^{N} (\mu_{ij})^m} \tag{18}$$

A generalized squared Mahalanobis distance norm is given by the substitution of equations (17) and (18) into (14), where the covariance is weighted by the membership degrees in *U*.

The possibilistic version of the GK algorithm is derived from the general form of possiblistic algorithms introduced in [6]. The objective function of the possibilistic GK is obtained by (19).

$$J(L,U) = \sum_{i=1}^{C} \sum_{j=1}^{N} (\mu_{ij})^m d_{ij}^2 + \sum_{i=1}^{C} \eta_i \sum_{j=1}^{N} (1 - \mu_{ij})^m \tag{19}$$

where μ_{ij} is updated using (20) and η_i are suitable positive numbers. The first term makes the distances from the feature vectors to the prototypes (cluster centers) as low as possible, whereas the second term leads to avoiding the trivial solution by forcing μ_{ij} to be as large as possible. The parameter η_i is calculated as in (21) and (22).

$$\mu_{ij} = \frac{1}{1 + (\frac{d_{ij}^2}{\eta_i})^{1/m-1}} \tag{20}$$

$$\eta_i = K \frac{\sum_{j=1}^{N} \mu_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} \mu_{ij}^m}, \tag{21}$$

where *K* is typically chosen to be 1. This choice of η_i makes it proportional to the average fuzzy intra-cluster distance of cluster β_i.

$$\eta_i = \frac{\sum_{x_j \in (\Pi_i)_\alpha} d_{ij}^2}{|(\Pi_i)_\alpha|}, \tag{22}$$

where $(\prod_i)_\alpha$ is an appropriate α -cut of $\prod_i$. In this case, η_i is the average intracluster distance for all of the "good" feature vectors (vectors with memberships greater than or equal to α). The parameter α is typically set to a value between 0.1 and 0.4 for consistent results. The possibilistic GK algorithm is explained in [15].

6 Experimental Results

The dataset used in this research consists of audio news stories collected and recorded manually from various Arabic news broadcast networks: Al-Jazeera, Al-Arabiya, and BBC Arabic. The dataset size is about 30 hours of recorded Arabic news stories. The average length of the news story is two minutes. The news stories are transcribed generating 1000 text files divided into five topics: culture and arts, economics, politics, science, and sports.

The reason behind the manual selection of the news stories is to minimize speaker related problems, like unclear pronunciation and grammatical errors. The collected news are then transcribed into text documents using ASR system "Dragon Dictation" [9], and then pre-processed for topic-clustering.

After applying pre-processing steps on the documents and performing clustering techniques, the accuracy of clustering is evaluated using F-Measure (21), a measure that combines the recall and precision ideas from information retrieval [19].

$$F = \sum_i \frac{n_i}{n} Max\{F(i,j)\} \tag{21}$$

The *max* is taken over all clusters at all levels, and *n* is the number of documents and *F(i,j)* is defined by:

$$F(i,j) = \frac{2 \times \mathrm{Re}call(i,j) \times \mathrm{Pr}ecesion(i,j)}{\mathrm{Re}call(i,j) + \mathrm{Pr}ecesion(i,j)} \tag{22}$$

$$\mathrm{Re}call(i,j) = n_{ij} / n_i \tag{23}$$

$$\mathrm{Pr}ecision(i,j) = n_{ij} / n_j \tag{24}$$

The quantities *Recall* and *Precision* are calculated as in (23) and (24), where n_{ij} is the number of members of class *i* in cluster *j*, n_i is the number of members of class *i*, and n_j is the number of members of cluster *j*.

The dataset is divided into subsets of sizes ranged from 50 to 200 documents per category. Experiments are carried out on each of these subsets four times for each clustering algorithm: when no stemming is applied, when light-stemming is applied, when root-based stemming is applied, and finally when rule-based light stemming is applied. Each clustering algorithm is run twice: one time with the use of the Chi-square similarity measure, and the other time with the use of the popular cosine measure. The accuracy of the clustering is evaluated for each subset, and then the average accuracy is calculated among all the subsets (Table III).

The dataset is divided into subsets to consider the effect of the change in dataset size and hence the change in the amount of information contained in each subset on the average clustering accuracy. The reason why clustering is first applied on non stemmed data is to measure the impact of applying stemming techniques on improving the accuracy of the clustering algorithms operating on such erroneous data. The use of two similarity measures is to ensure that the Chi-square measure makes a positive effect on clustering the erroneous data into topics.

TABLE II

ACCURACY EVALUATION OF TOPIC CLUSTERING OF THE TRANSCRIBED DOCUMENTS USING HARD CLUSTERING METHODS

Clustering Approach/Similarity Measure	**Average Accuracy**			
	Non-Stemmed	*Light-Stemmed*	*Root-Stemmed*	*Rule-Stemmed*
k-Means /Cosine	39.42%	44.61%	54.41%	60.04%
k-Means/Chi-square	44.3%	47.6%	56.5%	63.35%

Spectral Clustering/Cosine	45.62%	50.96%	65.57%	71.33%
Spectral Clustering/Chi-square	46.5%	53.8%	68.9%	76.11%

The experimental scenarios applied on the transcribed news documents are also applied on the error-free version of those transcribed documents, and then the average accuracy is calculated (Table IV). This step is carried on to gain knowledge about the sensitivity of the original data to clustering, thus an assessment to what extent the techniques proposed in this work have managed to overcome transcription errors can be performed.

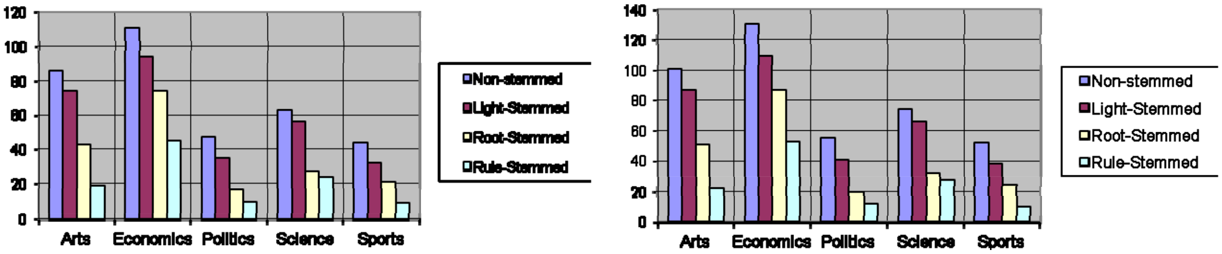
TABLE IV

ACCURACY EVALUATION OF TOPIC CLUSTERING OF ERROR-FREE VERSION OF THE TRANSCRIBED DOCUMENTS USING HARD CLUSTERING METHODS

Clustering Approach/Similarity Measure	**Average Accuracy**			
	Non-Stemmed	*Light-Stemmed*	*Root-Stemmed*	*Rule-Stemmed*
k-Means /Cosine	62.2%	64.63%	68.06%	76.84%
k-Means/Chi-square	65.9%	67.97%	72.84%	79.05%
Spectral Clustering/Cosine	72.2%	74.97%	80.77%	85.15%
Spectral Clustering/Chi-square	74.87%	76.85%	82.74%	87.21%

By comparing the accuracy results in Table III and Table IV, and by observing the clustering confusion matrix for each clustering scenario for both original and transcribed date, it is concluded that in both sets of data, there are documents causing clustering confusion. The existence of topic overlaps in the original data is the main cause of such confusion. The information loss due to the transcription errors is increasing the confusion even more in the transcribed data.

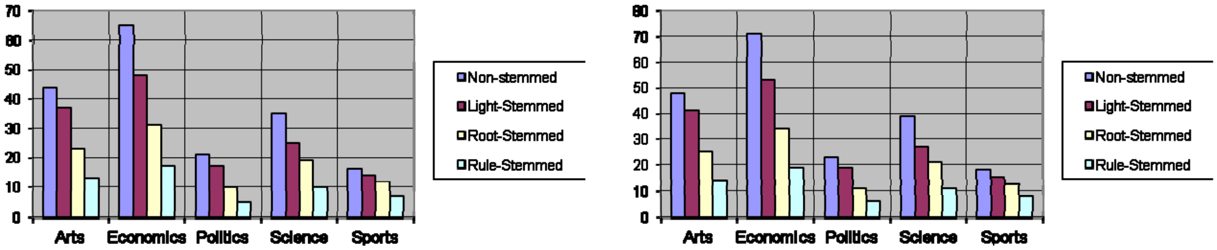
In the next phase of experiments fuzzy *c*-means and possibilistic GK algorithms are applied on both the transcribed and the original data, and the membership matrix is analyzed to evaluate the amount of confusing documents in each topic as in Fig. 5 and Fig. 6. A document is considered confusing to the clustering process if its membership degrees to all clusters are under a certain empirical threshold, or if its membership degrees to all clusters are convergent. By determining which documents are affecting the clustering accuracy, they can be excluded and the rest of the documents are maintained. Doing such exclusion, would improve the clustering accuracy for the rest of the documents. After re-applying the experiential scenarios on the remaining data on both transcribed and original data, the average clustering accuracy improved to an a maximum 79.34% and 90.52% respectively for the remaining data after using fuzzy *c*-means and maximum of 85.62% and 92.26% respectively for the remaining data after using possibilistic GK algorithm. In both cases, the maximum average accuracy is obtained when spectral clustering is used on rule-based stemmed data. Manual categorization can be considered a solution to categorize the excluded documents.



a) Confusing documents detected by fuzzy *c*-means

b) Confusing documents detected by possibilistic GK algorithm

Figure 5: Confusion documents detected in the transcribed data



a) Confusing documents detected by fuzzy *c*-means

b) Confusing documents detected by possibilistic GK algorithm

Figure 6: Confusion documents detected in the original data

7 DISCUSSION

The results have showed that stemming techniques have improved the accuracy of all clustering algorithms applied on the transcribed Arabic documents in all scenarios by an average of 19.7%. Rule-based light stemming has improved the accuracy of the clustering process by an average of 23.75%, which is better than the rule-based and light stemming techniques which improved the accuracy of the clustering process by an average of 17.39% and 5.28% respectively. The reason behind that is the nature of the Arabic language and its related transcription errors. Light-stemming techniques only removes certain prefixes and suffixes which will not truly and effectively transform all similar words to one root, hence limit its ability to overcome misidentification errors. Light stemming is also known for causing high mis-stemming and under-stemming errors. Mis-stemming occurs when an original part of the word is confused with an affix, and hence is removed. Under-stemming occurs when stemming two word with the same root, but instead, the stemmer produce two different roots. In contrast to light stemming, root-based stemming transforms all similar words to one root, except for some limitations that may exist in the algorithm performing the transformation, which makes it more efficient than light-stemming in overcoming errors. Root-based stemming causes high over-stemming errors where two words with different roots are transformed to one root. Rule-based light stemming has more ability to overcome transcription errors than light stemming, but less ability than root-based stemming. It also has the benefit of balancing between the stemming errors caused by the light and root-based stemming techniques; hence it leads to the best performance in the clustering phase.
The spectral clustering algorithm achieved more accuracy than the *k*-means algorithm in all cases which may be explained due to the nature of the data set. In contrast to spectral clustering, *k*-means tends to perform best in linearly separable data. Since the topics chosen are general and limited in number, thus the chance of existence of cross topic documents is increased and therefore the process of linearly separating data becomes more difficult.

The results also have showed that Chi-square similarity method has showed superiority over the popular and traditional cosine similarity and it is best utilized by the spectral clustering algorithm.

Applying fuzzy *c*-means and possibilistic GK algorithms on both the transcribed and original data has revealed some of the characteristics of the data. By analyzing the membership matrix and manual observation of the detected confusing documents, it is notable that the Economics topic has the biggest number of confusing documents; this may be explained due to the numerical-based nature of the content involved in this topic. Thus considering that the news stories are relatively short, it is hard to extract unique features that can qualify such document to be assigned vividly to a topic. Arts and Science have the second and third places in the number of occurrences of confusing documents. This may be explained to the possibility of existence of sub-topics within them, in addition to the relatively small size of the dataset that doesn't cover all of those sub-topics well. Although Politics topic has the least confusing documents, it is the most topic that received wrong-clustered documents from all other categories. This may be explained due to the interference of politics in many aspects of life, so it is possible for an economical decision to be based on some political background and both are melded into one news story. The possibilistic GK algorithm showed better ability to detect confusing documents than the fuzzy *c*-means especially when applied on the transcribed data. The reason is that GK algorithm is better at dealing with clusters of different geometrical shapes – caused by topic overlaps in the original documents – in addition to the advantage of using its possibilistic version that makes it superior in dealing with outliers mostly existing in noisy data (transcribed documents).

The number of confusing documents is increased in the transcribed documents due to transcription errors, especially when such errors occur frequently in named entities (names of persons, organizations, places, etc) which usually represent important features for guiding the clustering process. It is also notable that stemming played an important role in reducing the amount of confusing documents in all topics of the transcribed and original data.

8 Conclusions

In this research a set of transcribed textual documents obtained from a set of spoken documents are clustered into topics, and the impact of applying three stemming techniques along with the Chi-square and cosine similarity measures on the accuracy of the topic-clustering process is measured. The confusion nature of the transcribed data is investigated and compared to its original correct form by the use of fuzzy *c*-means and a possibilistic version of the GK fuzzy algorithm which showed that possibilistic clustering is best suitable for this kind of erroneous confusing data, compared to fuzzy clustering, as it has more ability detect those confusing members and hence give the option for excluding them.

The clustering accuracy evaluation showed that stemming has improved the clustering accuracy in all test scenarios. It also showed that Chi-square similarity measure has improved the clustering accuracy better than the traditional cosine similarity. The best topic clustering results are achieved by applying the spectral clustering algorithm with the aid of the Chi-square similarity measure on rule-based stemmed data scoring average accuracy of 76.11%. This accuracy is best improved to 85.62% by excluding the confusing document using the possibilistic GK algorithm.

REFERENCES

[1] Abberley D., Renals S., and Cook G., "Retrieval of broadcast news documents with the THISL system," *Proc. IEEE International Conference on Acoustic*, *Speech, and Signal Processing*, pp. 3781-3784, 1998.

[2] Abu El-Khair I., "Effects of stop words elimination for Arabic information retrieval: a comparative study," *International Journal of Computing & Information Sciences*, vol. 4, no. 3, pp. 119-133, July 2006.

[3] Al-Fares W., *Arabic root-based clustering: an algorithm for identifying roots based on n-grams and morphological similarity*, University of Essex, UK, 2002.

[4] Al-Kharashi I. and Evens M., "Comparing words, stems, and roots as index terms in an Arabic information retrieval system," *Journal of the American Society for Information Science*, vol. 45, no. 8, pp. 548-560, 1994.

[5] Al-Shammari E. and Lin J., "A novel Arabic lemmatization algorithm," *Proc. second workshop on Analytics for noisy unstructured text data*, pp. 113-118, 2008.

[6] Awde N. and Samano P., *The Arabic Alphabet: How to Read & Write It*, Lyle Stuart, 2000.

[7] Coden A., and Brown E., "Speech transcript analysis for automatic search," *Proc. 34th Annual Hawaii International Conference*, Jan. 2001.

[8] Dave R., "Boundary detection through fuzzy clustering," *Proc. IEEE International Conference on Fuzzy Systems*, San Diego, USA, pp. 127–134, 1992.

[9] Dragon Dictation App home page on iTunes store, https://itunes.apple.com/us/app/dragon-dictation/id341446764?mt=8 , (accessed August 2014)

[10] Gustafson D. and Kessel W. "Fuzzy clustering with a fuzzy covariance matrix," *Proc. IEEE CDC*, pp. 761–766, San Diego, CA, USA, 1979.

[11] Ibrahimov O., Sethi I., and Dimitrova N, "A novel similarity based clustering algorithm for grouping broadcast news," *Proc. SPIE Conf. Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*, 2002.

[12] Kanaan, G., Al-Shalabi, R., Ababneh, M., Al-Nobani, A., "Building an effective rule-based light stemmer for Arabic language to improve search effectiveness," *The International Arab Journal of Information Technology*, vol. 9, no. 4, pp. 368-372, July 2012.

[13] Khoja S. and Garside R., *Stemming Arabic text*, Lancaster, UK, Computing Department, Lancaster University, 1999.

[14] Korfhage R., *Information storage and retrieval*, John Wiley, 1997.

[15] Krishnapuram R., Keller J., "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*. vol. 1, no. 2, pp. 98–110, 1993.

[16] Larkey L., Ballesteros L., and Connell M., "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," *Proc. 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, pp. 275-282, 2002.

[17] Larkey L. and Connell M., "Arabic information retrieval at UMass in TREC-10," *Proc. Tenth Text REtrieval Conference (TREC-10)*, EM Voorhees and DK Harman ed, pp. 562-570, 2001.

[18] Luxburg U., "A tutorial on spectral clustering," *Springer Statistics and Computing*, vol. 17, no. 4, pp. 395-416, December 2007.

[19] Robertson S., Walker S., Jones S., Hancock-Beaulieu M., and Gatford M. "Okapi at TREC-3," *Proc. Third Text REtrieval Conference (TREC 1994)*. Gaithersburg, USA, NIST SP 500-225, pp. 109-126, November 1994.

[20] Salton G., Automatic text processing: the transformation, analysis, and retrieval of information by computer, Addison-Wesley, 1989.

[21] Schauble P., Multimedia information retrieval: Content-based information retrieval from Large Text and Audio Databases, Kluwer Academic Publishers, 1997.

[22] Shi J. and Malik J., “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, August 2000.

[23] Singler M., Jin R., and Hauptmann A., “CMU spoken document retrieval in Trec-8: analysis of the role of term frequency TF,” Proc. *8th Text REtrieval Conference*, NIST, Gaithersburg, MD, 1999.

[24] Steinbach M., Karypis G., and Kumar V., “A comparison of document clustering techniques,” *KDD Workshop on Text Mining*, University of Minnesota, 2000.

[25] Yang M., “A survey of fuzzy clustering,” *Mathematical and Computer Modelling journal*. vol. 18, no. 11, pp. 1–16, 1993.

BIOGRAPHY

Ahmed Abdelaziz Jafar obtained his B.Sc in Computer Science from Faculty of Information systems and Computer Science, October Six University, Egypt in July 2006. He received his MS.c. Degree in Computer Science from College of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt in September 2014. He is a demonstrator at the Faculty of Information systems and Computer Science, October Six University, starting from March 2008 – Present.

Mohamed Waleed Fakhr received his Ph.D. in Electrical Engineering from Electrical and Computer Engineering department, University of Waterloo, Waterloo, Canada in May 1994. He is currently a full professor at College of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt starting from September 2013 – Present. He teaches the following courses: Digital Signal Processing, Multimedia Systems, Computer Networks, Pattern Recognition, Neural Networks, Digital Logic design. His fields of interest are Image Processing, Audio Processing, Pattern recognition, Sparse Coding, Sparse Recovery, and Machine Learning.

Mohamed Hesham Farouk received his Ph.D. in Engineering Physics from Faculty of Engineering, Cairo University, Egypt in January 1994. He is a Full Professor of Engineering Math. & Physics Dept., Faculty of Engineering, Cairo University starting from 2007-Till Now. His publications include: 11 papers on numerical analysis of acoustic scattering integral equations and wavelets, 10 Papers on speech processing and Vocal Tract Modeling, 4 Papers on Genetic Algorithms & Ultrasonic Parameters estimation, 1 paper on digital signal processing of x-ray, and 1 Paper on the development of Real-time SCADA systems using state machines.

ملخص

عنقدة الموضوع للأخبار العربية المجذرة التي يتم بثه

احمد عبد العزيز جعفر[*]، محمد وليد فخرو[*]، محمد هشام فاروق[**]
[*]قسم علوم الحاسب-كلية الحاسبات و تكنولوجيا المعلومات-الأكاديمية العربية للهندسة و العلوم و التكنولوجيا والنقل البحرى
[**]قسم الفيزيقا و الرياضيات الهندسية- كلية الهندسة – جامعة القاهرة

في هذا البحث، تم تطبيق أساليب عنقدة مختلفة لتجميع ملفات نصية تم إملاؤها من مصادر صوتية عن طريق نظام للتعرف على الحديث. وحيث أن هذه الملفات المُملاة عادة ما تحتوي على العديد من الأخطاء، كان من اللازم الحد من التأثير السلبي لهذه الأخطاء. لذلك، وفي محاولة للسعي إلى التغلب على بعض هذه الأخطاء، تم إستخدام أساليب مختلفة لتجذير الكلمات الموجودة في هذه النصوص المُملاة. كما تم ايضا تحديد واستبعاد الملفات التي تؤدي إلى تشويش عملية العنقدة بمساعدة أساليب عنقدة ترجيجية وإحتمالية.

إن الهدف من هذا البحث هو تقسيم الملفات النصية المُملاة طبقا للموضوعات التي تمثلها واستكشاف مدى تأثير عملية تجذير النصوص، مصحوبة بتطبيق معيار تشابه معتمد على توزيع "مربع كاي" الإحصائي، على دقة عملية التقسيم طبقا للموضوع التي تتم بواسطة الخوارزميات المختارة لهذا الغرض.

لقد أظهر تقييم نتائج التجارب أن استخدام التجذير الخفيف المعتمد على القواعد مع العنقدة الطيفية قد حققا معاً أعلى دقة لتقسيم الملفات وهذه الدقة تمت زيادتها بدرجة أكبر بعد استبعاد الملفات المُشوِشة بإستخدام خوارزمية "جيستافسون كيسيل" الإحتمالية.

Band Mapped and Distributed Energy Best Tree Encoding Features for Tri-Phone-Based Automatic Speech Recognition

Amr M. Gody [*1], Dr. Tamer M. Barakat[*2], Sayed A. Zaky[**3]

** Department of Electrical Engineering, Fayoum University*
Fayoum- Egypt
[1]amg00@ fayoum.edu.eg
[2] tmb00 @ fayoum.edu.eg
[**] Department of Engineering Affairs, Fayoum University
Fayoum- Egypt
[3]sa134 @ fayoum.edu.eg

Abstract — **Best Tree Encoding is a promising feature extraction technique used in automatic speech recognition it based on wavelet packet decomposition. This research provides comparison results between the standard BTE and the new proposed model BMDE-BTE for solving the Tri-Phone Automatic speech Recognition. In Addition; comparison to mono phone Recognition problem is introduced. For mono phone the encoding algorithm with the new entropy give better results for correctness (24.60) and accuracy (-52.00) when adding Delta and Acceleration coefficients, but in case of tri phone correctness get worst (19.56) and accuracy (-52.44).**

Keywords: Arabic tri phone recognition, Best Tree Encoding, BTE

1 INTRODUCTION

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. It has a wide area of applications: Command recognition (Voice user interface with the computer), Dictation, Interactive Voice Response, it can be used to learn a foreign language. ASR can help also, handicapped people to interact with society. It is a technology which makes life easier and very promising.

View the importance of ASR too many systems are developed, the most popular are: Dragon Naturally Speaking, IBM Via voice, Microsoft SAPI. Open source speech recognition systems are available too, such as HTK, ISIP, AVCSR and CMU Sphinx-4. The interested tool is Hidden Markov toolkit (HTK), which is based on Hidden Markov Models (HMMs).

A Hidden Markov Model (HMM) is a statistical model where the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters, from the observable parameters, based on this assumption. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications.

Speech recognition systems commonly carry out some kind of classification / recognition based on speech features which are usually obtained via Fourier Transforms (FTs), Short Time Fourier Transforms (STFTs), or Linear Predictive Coding techniques. However, these methods have some disadvantages. These methods accept signal stationary within a given time frame and may therefore lack the ability to analyze localized events correctly. Moreover, the LPC method accepts a particular linear (all-pole) model of speech production which strictly speaking is not the case. The wavelet transform copes with some of these problems. The wavelet transform decompose the signal into a set of basis functions that are scale variant with frequency. Wavelets are shifted scaled versions of original or mother wavelets. There are many types of mother wavelets but the most suitable one for speech recognition is Duabiechi wavelet. The wavelet families are commonly orthogonal to one another, since this situation gives computational efficiency and ease of numerical implementation. Other factors influencing the selection of Wavelet Transforms (WT) over conventional methods include their ability to determine localized features.

Even though many speech recognition systems have obtained satisfactory performance in clean environments, recognition accuracy significantly degrades if the test environment is different from the training environment. These environmental differences might be due to additive noise, channel distortion, acoustical differences between different speakers, and so on. Many algorithms have been developed to enhance the environmental robustness of speech recognition systems. The most frequently used approach is based on HMM phone models, where each speech waveform is initially decomposed into a sequence of feature vectors. Then, a set of HMM phone models (phone recognizer) is utilized to extract the corresponding phonetic sequence.

Various speech recognition techniques have been utilized in the phonetic recognition task, with Mel Frequency Cepstral Coefficients (MFCC) among the most widely used, especially in the HMM-based approach. Other speech features such as Perceptual Linear Prediction (PLP), Line Spectral Frequencies (LSF), Linear Predictive Coding (LPC), short-time energy, formants and wavelet-based have also been used. Due to the requirements of the large vocabulary and need for speed in real time applications, the research affords has been dedicated to memory reduction and efficient search algorithms. Terms such as tri phones [2-7] and syllables [8-12] are applied in ASR systems.

BTE algorithm introduced and developed to improve the speech recognition systems performance.BTE was first introduced by Amr M.Gody [1], first generation BTE4.And where BTE based on optimal adjacency in frequency domain which made manual, and where manual annotation for time-aligning speech waveform against the corresponding phonetic sequence is a tedious and time consuming task a completely automated Arabic phone recognition system based on enhanced Wavelet Packets Best Tree Encoding (EWPBTE) was introduced by Mohamed Hassan [13]. BTE second generation BTE5 was developed by Maha Adham [14], the third generation BTE7 was developed by Eslam El Maghraby [15], by increasing the decomposing levels to 7 levels and uses the Log energy entropy instead of Shannon's. The BTE second and third generations show improvement over BTE4, but still not the expected success rate of BTE specially for Arabic language, where maximum success reach 26.0%.

As a development to BTE, this paper introduces a new encoding algorithm which use down sampling to the human auditory perception frequency (10000 Hz), then the best tree leave node encoded based on its location and energy to increase the discrimination of the feature vector, all of these done by using Band Mapped and Distributed Energy Best Tree Encoding Features (BMDE-BTE), which applied for tri-phone-based automatic speech recognition. To evaluate this development many experiments were performed, and their results reported and discussed.

Section 2 is a BTE overview. Next, an overview for Band Mapped is addressed in section 3. Section 4 is a detailed description of the proposed model Band Mapped and distributed Energy BTE. Section 5 explains HMM design for tri-phone recognition. Section 6 presents the experiments and results, and section 7 contains summary and conclusions.

2 BTE OVERVIEW

A. Wavelet transform and best tree

The Wavelet Transform (WT) is used in a variety of signal processing applications, such as video compression, speech recognition, and numerical analysis. It can efficiently represent some signals, especially ones that have localized changes.

Its representation basically involves the decomposition of the signals in terms of small wave components called wavelets.

The wavelet transform has proven to be very efficient and effective in analyzing a very wide class of signals and phenomena. It has the ability to compact the signal energy into few large coefficients. The original signal can be reconstructed perfectly from these few coefficients while suppressing the other coefficients without losing most of the features of the signal. Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulator and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. An important problem in speech recognition systems is to determine a representation that is well adapted for extracting information content of speech signals. In general, transforming a signal to a different domain is done to get a better representation of the signal. For recognition, better means having more ability to separate signals which belong to separate classes or categories in the new domain than in the original domain. Also, speech signal is not a constant frequency, where frequency vary with time, see figure 1, so Fourier transform which applied to a constant frequency and periodic signal cannot used for speech signal, also, Short Fourier transform, where a fixed size window applied to the speech signal to deal with the local changes in the signal frequency, cannot be used for speech signal where too much information due to the window size not suitable to the signal.

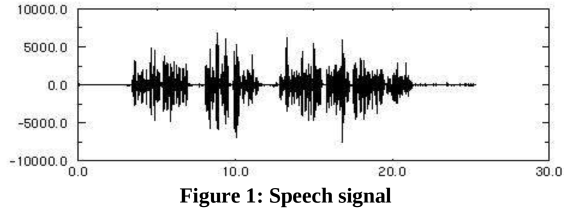


Figure 1: Speech signal

Wavelets have the ability to analyze different parts of a signal at different scales. The wavelet transform (WT) is a transformation that provides time-frequency representation of the signal. The continuous one dimensional WT is a decomposition of f (t) into a set of basis function Ψa,b(t) called wavelets. The wavelets are generated from a single mother wavelet Ψ (t) by dilation and translation

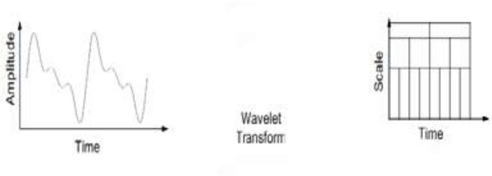


Figure 2: Wavelet transform

$$W(\mathrm{a,b}) = \int f(\mathrm{t})\Psi^{*}{}_{\mathrm{a,\,b}}(\mathrm{t})\,\mathrm{dt} \qquad (1)$$

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right) \qquad (2)$$

where: f (t) is the signal to be analyzed, a is the scale, and b is the translation factor. W (t) is the transforming function and is called the mother wavelet. Filters of different cutoff frequencies are used to analyze the signal. A continuous wavelet transform can operate at every scale. Also the analyzing wavelet is shifted smoothly over the full domain of the analyzed function. In discrete wavelet transform (DWT), scales and positions of powers of two are chosen. Given a signal S of length N, the DWT consists of log 2 N stages at most. The first step produces, starting from S, two sets of coefficients: approximation coefficients A1 and the detail coefficients D1 see figure 3. These vectors are obtained by convolving S with a low pass filter for approximations, and with a high pass filter for details, followed by dyadic decimation. The next step splits the approximation coefficients A1 into two parts using the same scheme, replacing S by A1 and producing A2 and D2, and so on.

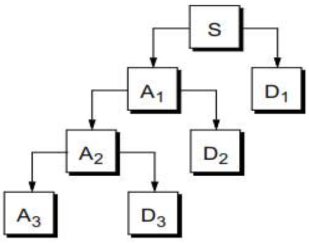


Figure 3: Wavelet decomposition

This technique is most effective when it is applied to the detection of short-time phenomena, discontinuities, or abrupt changes in the signal.BTE based on wavelet packet decomposition, the difference between the wavelet transform and wavelet packet decomposition in that, in wavelet packet analysis, the details as well as the approximations can be split, see figure 4.

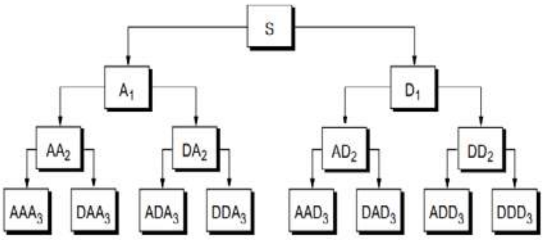


Figure 4: Wavelet packet decomposition

For speech signal not all the wavelet tree leaves have useful information so best tree is used to reduce the analysis time and complexity. Best tree obtained by applying entropy function, see figure 5.There are many type of entropy functions, here Shannon entropy is used.

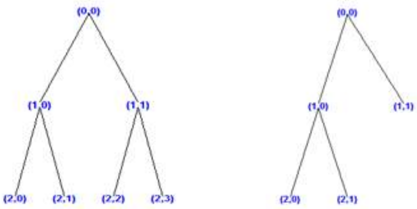


Figure 5: Two-levels wavelet packet and best tree

B. Best Tree Encoding

BTE was first introduced by Amr M.Gody [1], the BTE first generation, as a speech recognition algorithm to solve the problems in the existing speech recognition algorithms, bad performance in case of noisy environments. Figure 6 shows a block diagram of BTE, it works as follow, for detailed description see [1], the input speech signal is divided and windowed, using Hamming window, into frames of length 20 ms, then the frame is decomposed into wave packet tree which consists of a number of levels and each level contain many nodes, not all nodes have a useful information about the speech signal so, an entropy is applied to obtain which called best tree, Shannon entropy, this best tree is encoded into a feature vector of four components.

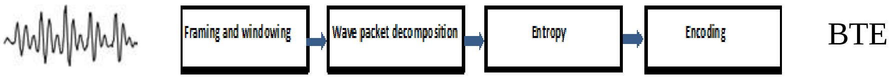


Figure 6: BTE block diagram

The first generation of BTE encodes the best tree based on the terminal nodes location as follow: [1]
After obtaining the best tree from the wave packet tree, figure 7.

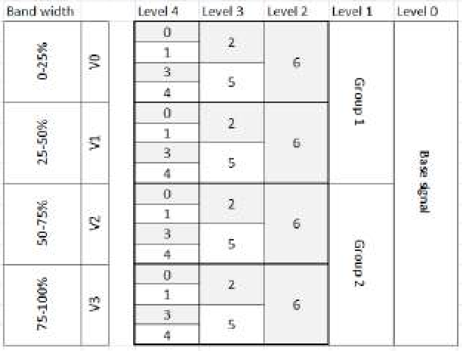


Figure 7: Wave packet tree for 4 points encoding [1]

The leave nodes encoded based on its location to a 7 bit value figure 8 shows an example for encoding a frame of speech signal.

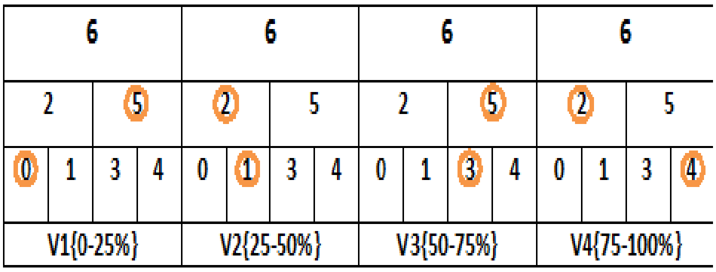


Figure 8: Best tree 4 point encoding example

Circled numbers in figure 8 represent leave nodes in the best tree decomposition, which will be encoded into features vector of 4 elements as shown in table 1.

Table 1
Best tree 4 point encoding evaluation.

Element	Binary value	Decimal value	Frequency band
V1	0100001	33	0-25%
V2	0000110	6	25-50%
V3	0101000	40	50-75%
V4	0010100	20	75-100%

Features vector for this example speech frame will be:

$$F= [12; 64; 0; 4] \qquad (3)$$

The second generation of BTE5was developed by Maha Adham [14], where the decomposing level increased to 5 levels to increase the discrimination of the feature vector to give 25% success rate, the third generation BTE7 was developed by

Eslam El Maghraby [15], by increasing the decomposing levels to 7 levels and uses the Log energy entropy instead of Shannon's. The last generation BTE7 gives 22% enhancements over BTE4 and BTE5.

3 BAND MAPPED OVERVIEW

Before being processed linguistically, speech sounds must pass through the auditory system where the perceptually-salient cues or features present in the acoustic signal are transformed in various, mostly non-linear, ways. When sound enters the outer ear it is affected by the resonances of the pinna (ear lobe), concha (funnel-like opening to the outer ear canal), and external auditory meatus (outer ear canal). The main effect of these resonances is to produce a broad peak of 15-20 dB at 2500 Hz and spreading relatively uniformly from 2000-7000 Hz (Pickles, 1988). What is of particular relevance to the present discussion of the non-linear transduction of the ear is the finding that the pressure gain transfer function of the middle ear is not uniform but shows a peak at 1000 Hz and gradually drops off to about 20 dB below peak level at 100 Hz and 10,000 Hz (Nedzelnitsky, 1980). This peak is relatively flat, however, over most of the frequency range containing speech cues (<10 dB variation from 300 Hz to 7000 Hz).

A. Frequency

There are three major types of auditory behavior that are of interest when examining auditory processing of the frequency dimension. They are frequency discrimination, frequency selectivity (or resolution) and judgments of relative pitch. Here, the more interested is the frequency discrimination.

1) Frequency Discrimination

Frequency discrimination refers to the ability to detect differences in the frequencies of sounds which are presented successively. They are also referred to as frequency difference thresholds (df), frequency difference limen (DLF) or just noticeable differences in frequency (or frequency jnd). Many studies showed that from 125-2000 Hz df is constant at about 3 Hz. It rises to about 12 Hz by 5000 Hz, 30 Hz by 10000 Hz, and 187 Hz by 15000 Hz. So the most useful information exist in the frequency range below 5000 Hz or bandwidth 10000 Hz.

In the previous BTE generations the input audio signal has sampling rate 32000 Hz, which contains a lot of non useful information, also this consume processing time and increase complexity. To overcome these problems and improve the discrimination rate, down sampling must be applied.

2) Down sampling

Down sampling is decreasing the sampling rate of a signal. Let's consider a simple case of down sampling a signal to half of its original sampling rate. Simplest way to do this is to forget every other sample and we'll have the desired sampling rate. See figure 9.

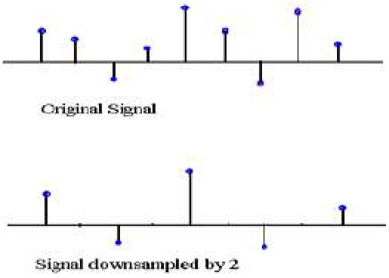


Figure: 9 down sampling by 2

But if we reduce the sampling rate just by selecting every other sample of x (n), the resulting with folding frequency Fs/2. The frequency spectrum tries to spread up but it cannot do so. Hence it winds up on itself. See figure 10.

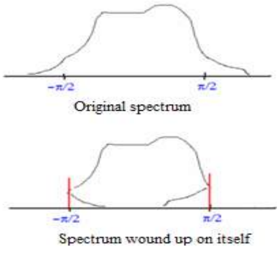


Figure 10: spectrum wound up itself

To avoid aliasing, we must first reduce the bandwidth of x (n) to w $_{max}$= π/2. So we should cut of high frequency contents to avoid aliasing.

4 PROPOSED MODEL BAND MAPPED AND DISTRIBUTED ENERGY BTE (BMDE-BTE)

As mentioned earlier the previous BTE generations encode the best tree leaves based on its location, to increase the discrimination of the feature vector and so that the success rate increased the best tree encoded based on the leaves location and its energy, so eight components feature vector obtained. Also, to increase the success rate the proposed algorithm BMDE-BTE was applied in Arabic tri phone recognition system. The proposed algorithm BMDE-BTE based on the same idea of the previous algorithms but it differ in the sampling rate used where the sample frequency is down sampled by a sampling rate to be 10000 Hz, also the output feature vector consists of 8 components which represent the location and energy of the terminal nodes .The proposed algorithm BMDE-BTE not like the previous BTE algorithms which recognize continuous words using mono phone recognition but it use tri phone recognition.

5 HMM DESIGN FOR TRI-PHONE RECOGNITION

A Hidden Markov Model is a statistical model of a sequence of feature vector observations. In building a recognizer with HMMs we need to decide what sequences will correspond to what models. In the very simplest case, each utterance could be assigned an HMM: for example, one HMM for each digit in a digit recognition task. To recognize an utterance, the probability metric according to each model is computed and the model with the best fit to the utterance is chosen. However, this approach is very inflexible and requires that new models be trained if new words are to be added to the recognizer. A more general approach is to assign some kind of sub-word unit to each model and construct word and phrase models from these.

The most obvious sub-word unit is the phoneme. If we assign each phoneme to an HMM we would need around 45 models for English; an additional model is also created for silence and background noise. Using this approach, a model for any word can be constructed by chaining together models for the component phonemes. Each phoneme model will be made up of a number of states; the number of states per model is another design decision which needs to be made by the system designer. Each state in the model corresponds to some part of the input speech signal; we would like the feature vectors assigned to each state to be as uniform as possible so that the Gaussian model can be accurate. A very common approach is to use three states for each phoneme model; intuitively this corresponds to one state for the transition into the phoneme, one for the middle part and one for the transition out of the phoneme. Similarly the topology of the model must be decided. The three states might be linked in a chain where transitions are only allowed to higher numbered states or to themselves. Alternatively each state might be all linked to all others, the so called ergodic model. These two structures are common but many other combinations are clearly possible.[17]

When phoneme based HMMs are being used, they must be concatenated to construct word or phrase HMMs. For example, an HMM for `cat' can be constructed from the phoneme HMMs for /k/ /a/ and /t/. If each phoneme HMM has three states the `cat' HMM will have nine states. While phoneme based models can be used to construct word models for any word they do not take into account any contextual variation in phoneme production. One way around this is to use units larger than phonemes or to use context dependant models. The most common solution is to use tri-phone models where there is one distinct phoneme model for every different left and right phoneme context. Thus there are different models for the /ai/ in /k-ai-t/ and in /h-ai-t/. Now, a word model is made up from the appropriate context dependant tri-phone models: 'cat' would be made up from the three models [/sil-k-a/ /k-a-t/ /a-t-sil/]. See figure 11.

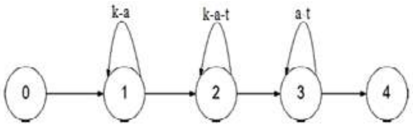


Figure 11: Tri-phone model for the English word Cat

While the use of tri-phones solves the problem of context sensitivity it presents another problem. With around 45 phonemes in English there are 453 = 91125 possible tri-phone models to train (although not all of these occur in speech due to phonotactics constraints). The problem of not having enough data to effectively train these models becomes very important. One technique (state tying) but another is to use only word internal tri-phones instead of the more general cross word tri-phones. Cross word tri-phones capture co articulation effects across word boundaries which can be very important for continuous speech production. A word internal tri-phone model uses tri-phones only for word internal triples and di-phones for word final phonemes; `cat' would become: [sil /k-a/ /k-a-t/ /a-t/ sil]. This will clearly be less accurate for continuous speech modeling but the number of models required is smaller (none involve silence as a context) and so they can be more accurately trained on a given set of data.

Here, HTK tool kit was used in HMM training. HTK first is driven data and uses a similarity measure between states. The second uses decision trees and is based on asking questions about the left and right contexts of each tri-phone. The decision tree attempts to find those contexts which make the largest difference to the acoustics and which should therefore distinguish clusters. This tree is driven based on good pronunciation knowledge for English language.

For Arabic language there are no enough pronunciation researches, so the decision tree can't be driven and HTK based only on the tied list obtained from merging mono phone, bi phone and tri phone in one file (fullist) and use it for training the HMM model.

6 EXPERIMENTS AND RESULTS

A. Experiments procedures

In the experiments reported in this paper, continuous word recognition experiments were performed using an Arabic database in an Egyptian colloquial language. This database consists of 2977 speech files in the wav format with 32000 Hz sampling rate, 16 bit, before any processing the wav file was down sampled, as a preprocessing step, to a sampling frequency 10000 Hz where the speech signal band width is 8000 Hz and assume 2000 Hz as guard band where some voices increase more than 4000 Hz so 10000 Hz sampling frequency is efficient for speech processing and provide only the useful information. The database divided into two sets one for training the Hidden Markov Model (HMM) models using HTK toolkit and the other for testing and evaluation of the recognition process. The new feature extraction algorithm used in Arabic mono phone and tri phone recognition, then a new entropy which introduced by Mai Ezz [16] used in the recognition process for mono and tri phone. The new entropy based on Mel frequency where not all speech frequencies detected by the human auditory system so not all leave nodes have useful information, then using this entropy gives a new best tree contain the most important information in the speech signal.

Experiments were performed to obtain comparative results for the new BTE algorithm with the previous BTE algorithms and then evaluate the new algorithm. The new algorithm was applied for mono phone recognition, in all BTE generations, and then the new entropy used in the new algorithm and applied for mono phone recognition, the above experiments repeated for tri phone recognition. Finally, the above experiments repeated with adding Dleta and Acceleration coefficients to the feature vector. In each experiment Gaussian Mixture Model (GMM) take the value 2. The next section shows the detailed results of each experiment.

B. Results

This section report the results of all experiment, see Tables 2, 3 and 4, and discus these results to evaluate the new BTE algorithm.

The HTK tools used for evaluating the recognition results, HResults tool compares the transcriptions output by HVite tool with the original reference transcriptions and then outputs various statistics. HResults matches each of the recognized and reference label sequences by performing an optimal string match using dynamic programming. After the optimal alignment founded, the number of substitution errors (S), deletion errors (D) and insertion errors (I) calculated. Then the percentage correctness and accuracy calculated as follow:

$$\text{Percent Correct}= ((\mathbf{N\text{-}D\text{-}S)/N}) \text{ X}100\% \qquad (3)$$

where ***N*** is the total number of labels in the reference transcriptions ignores insertion errors. For many purposes, the percentage accuracy is defined as

$$\text{Percent Accuracy}= ((\mathbf{N\text{-}D\text{-}S\text{-}I)/N}) \text{ X}100\% \qquad (4)$$

Table 2
BTE-4 Results

	Entropy	**Qualifiers**	**% Correctness**	**Accuracy**
Mono phone	Shannon	---	19.35	-136.83
		A_D	22.26	-64.56
	New entropy	---	19.56	-52.02
		A_D	24.60	-52.40
Tri phone	Shannon	---	19.37	-136.92
		A_D	22.19	-64.25
	New entropy	---	19.56	-56.00
		A_D	24.60	-52.44

Table 3
BTE-5 Results

	Entropy	**Qualifiers**	**% Correctness**	**Accuracy**
Mono phone	Shannon	---	15.02	-141.90
		A_D	21.32	-82.75
	New entropy	---	17.89	-74.31
		A_D	23.11	-48.44
Tri phone	Shannon	---	15.01	-123.82
		A_D	25.33	-36.20
	New entropy	---	15.05	-120.00
		A_D	24.91	-41.24

Table 4
BTE-7 Results

	Entropy	Qualifiers	% Correctness	Accuracy
Mono phone	Shannon	---	13.44	-253.08
		A_D	19.60	-225.00
	New entropy	---	16.12	-181.20
		A_D	20.45	-102.11
Tri phone	Shannon	---	13.44	-249.89
		A_D	18.78	-202.04
	New entropy	---	16.00	-177.49
		A_D	19.80	-92.46

The negative value of Accuracy is a reflection of low stability of the recognizer, where the insertion errors are too much. The optimal situation or the most stable situation exists when the total numbers of recognized phones equal the total number of expected phones. The expected phones are pre-evaluated before the recognition process, this process is called transcription.

With reference to MFCC technique with the same database and using three qualifiers Delta, Acceleration and Energy the features vector becomes 39 components vector the output correctness equal 37%, by using the proposed model for two qualifiers only Delta and Acceleration the feature vector becomes 16 components vector and give correctness about 25% or about 68% with reference to the common technique MFCC, so the proposed is a promising technique if considering the not 100% database accuracy and also if the transcription done manually. Also, the proposed model is faster than the previous BTE generations and MFCC because it based on a reduced feature vector so it save the processing time and power. As a future work a new and more accurate database and more accurate transcription method will be used to increase the accuracy and make it more stable, positive correctness values.

7 Summary and Conclusion

The above results show that BTE-4 and using the new entropy for mono phone recognition give the best results, and using tri phone not give enhancement in the success rate, also increasing the decomposition results not improve the performance. Finding efficient automatic speech recognition techniques for Arabic words is of great interest since the research efforts remain limited.BTE is a promising technique, mono phone recognition shows better results than tri phone results. Also, when using the new entropy and adding Delta and Acceleration coefficients more improvement obtained.

REFERENCES

[1] Amr M. Gody, "Wavelet Packets Best Tree 4 Points Encoded (BTE) Features", The Eighth Conference on Language Engineering, Ain-Shams University, Cairo, Egypt, 17-18 December 2008.

[2] R. Thangarajan and M. Selvam, “Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language,” J. Res. Dev. VOL. 48, NUMBER 5/6, vol. 4, no. 3, 2008.

[3] G. A. O. Sheng, X. U. Bo, and H. Tai-yi, “CLASS-TRIPHONE ACOUSTIC MODELING BASED ON DECISION TREE FOR MANDARIN CONTINUOUS SPEECH RECOGNITION,” Int. Symp. CHINESE Spoken Language Processing. , ISCSLP, SINGAPORE. 1998.

[4] S. Darjaa and M. Rusko, “Rule-Based Triphone Mapping for Acoustic Modeling in Automatic Speech Recognition,” INT J SPEECH TECHNOL, 2004.

[5] F. Hassan, M. R. A. Kotwal, G. Muhammad, and M. N. Huda, “MLN-based Bangla ASR using context sensitive triphone HMM,” Int. J. Speech Technol., vol. 14, no. 3, pp. 183–191, Jun. 2011.

[6] P. Banerjee, G. Garg, P. Mitra, and A. Basu, “Application of Triphone Clustering in Acoustic Modeling for Continuous Speech Recognition in Bengali,” WSEAS Trans. SIGNAL Process., pp. 2–5, 2010.

[7] S. Darjaa and M. Rusko, “Effective Triphone Mapping for Acoustic Modeling in Speech Recognition,” INTL. Conf. AEI, VENICE, ITALY, 2010.

[8] M. Y. Tachbelie, “Morphology-Based Language Modeling for Amharic,” PH.D. THESIS, Univ. HAMBURG, Ger., no. August, 2010.

[9] S. T. Abate and W. Menzel, “Syllable-Based Speech Recognition for Amharic,” *INTERSPEECH 2006 – ICSLP*, no. June, pp. 33–40, 2007.

[10] A. Ganapathiraju, J. Hamaker, J. Picone, S. Member, M. Ordowski, and G. R. Doddington, “Syllable-Based Large Vocabulary Continuous Speech Recognition,” *SPECOM 2005, 2005, PP. 499– 502.*, vol. 9, no. 4, pp. 358–366, 2001.

[11] A. Lakshmi and H. A. Murthy, “A SYLLABLE BASED CONTINUOUS SPEECH RECOGNIZER for TAMIL,” *IEEE SIGNAL Process. Mag. VOL. 26, NO. 4, PP. 78–85*, 2010.

[12] . Martha Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier, “Part-of-speech tagging for underresourced and morphologically rich languages - the case of amharic,” in Proceedings of the HLTD 2011, 2011, pp. 50–55.

[13] Amr M. Gody, Rania Ahmed AbulSeoud, and Mohamed Hassan, “Automatic Speech Annotation Using HMM based on Enhanced Wavelet Packets Best Tree Encoding (EWPBTE) Feature.” the th Conference on Language Engineering. 2008, Cairo, Egypt.

[14] M. Adham, "Phone level speech segmentation using wavelet packets," Master thesis, Department of electrical engineering, Fayoum university , Fayoum, Egypt, 2013.

[15] Eslam El Maghraby, "ENHANCEMENT SPEED OF LARGE VOCABULARY SPEECH RECOGNITION SYSTEM," Master thesis, Department of electrical engineering, Fayoum university , Fayoum, Egypt, 2013.

[16] Amr M. Gody, Rania AbulSeoud, and Mai Ezz, ″ Enhanced Best Tree Encoding (BTE) Model using adapted wavelet filter ″, the 14th Conference on Language Engineering. 2014, Cairo, Egypt

[17] GHAI, W.; SINGH, N. "Phone based acoustic modeling for automatic speech recognition for punjabi language", Journal of Speech Sciences 1(3):69-83.2013.

Bibliography

Amr M. Gody received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University. Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is author and co-author of about 40 papers in national and international conference proceedings and journals. He is the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012, 2013 and 2014. His current research areas of interest include speech processing, speech recognition and speech compression.

Tamer M. Barakat received his BSc in communications and computers engineering from Helwan University, Cairo; Egypt in 2000. Received his MSc in Cryptography and Network security systems from Helwan University in 2004 and received his PhD in Cryptography and Network security systems from Cairo University in 2008. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt, in 2009. His main interests are: Cryptography and network security, Digital Image, and Digital Signal Processing. More specially, he is working on the design of efficient and secure cryptographic algorithms, in particular, security in the wireless sensor networks. Other things that interest him are number theory and the investigation of mathematics for designing secure and efficient cryptographic schemes.

Sayed A. zaky received the B.Sc. from the Faculty of Engineering, Cairo University. Egypt, in 2003. He joined the Engineering Department, Fayoum University, Egypt in 2012. His main interests are: Digital Signal Processing and speech processing.

نموذج مُحسن لتشفير الشجرة المُثلى (BTE) باستخدام تشفير بيانات عُقد شجرة حزم المويجات

عمرو محمد رفعت جودي[1] ، تامر محمد عبدالرحمن بركات[2] ،سيد عبدالباسط زكي[3]
[1،2] قسم الهندسة الكهربية،جامعة الفيوم
الفيوم- جمهورية مصر العربية
[1]*amg00@ fayoum.edu.eg*
[2] *tmb00 @ fayoum.edu.eg*
[3] ادارة الشئون الهندسية،جامعة الفيوم
الفيوم –جمهورية مصر العربية
[3]*sa134 @ fayoum.edu.eg*

ملخص

تقنية تشفير الشجرة المُثلى هي تقنية للحصول على سمات الاشارة الصوتية لتطبيقها في أنظمة التعرف التلقائي على الكلام،تعتمد تقنية تشفير الشجرة المُثلى على تحلل المويجات (WPD). يُقدم هذا البحث نتائج مقارنة لتقنية

تشفير الشجرة المُثلى (BTE) القياسي و النموذج المقترح "تشفير الشجرة المُثلى اعتماداً على محتوى الطاقة بالعقد الشجرية في الشجرة المُثلى لتقنية التعرف التلقائي على الكلام المعتمدة على جزء من الكلمة مكون من ثلاثة احرف صوتية" (BMDE-BTE).بالاضافة الى ذلك، يُقدم نتائج مقارنة لتقنية التعرف التلقائي على الكلام المعتمدة على جزء من الكلمة مكون من حرف صوتي واحد و التي تعطي درجة صحة [correctness= 24.60%] و درجة دقة [accuracy= -52.00%] و ذلك باستخدام النموذج المقترح لتشفير العقد الشجرية اعتماداً على موضعها الشجري و ايضاً محتواها من الطاقة بالاضافة الى تطبيقEntropy جديد معتمد على معادلة تأخذ في اعتبارها أن تردد اشارة الكلام غير ثابت مع الوقت،بينما في حالة تطبيق النموذج المقترح على تقنية التعرف التلقائي على الكلام المعتمدة على جزء من الكلمة مكون من ثلاثة حروف صوتية تصبح النتائج سيئة.

Sentiment Analysis of Arabic Conversations on Social Media A Study of Linguistic Mechanisms

Amr Gomaa Abdel-Rasool[*1], Muhammad Allam Farghaly[**2]

[*]*Faculty of Dar Al Olum, Cairo University*
Giza, Cairo, Egypt
[1]amr1979go@yahoo.com

[**]*Faculty of Commerce, Cairo University*
Giza, Cairo, Egypt
[2]muhamad.allam@gmail.com

***Abstract*—Social media has recently witnessed major developments, as its applications and platforms are now being used by many people all around the world. As its importance kept rising, the need for applications that can manage this enormous amount of digital data has increased as well. One of the most important applications needed is Sentiment Analysis application, which is to be used to identify the sentiment of those who engage in online conversations on social media towards a certain issue. This research is based upon building a linguistic mechanism to analyze opinions and break them down to positive, neutral, and negative. This mechanism can also identify the sentiment of the user based on his tone of voice expressed in writing on social media, which can be broken-down into negative sentiments; e.g. hate, anger and grief, positive sentiments; e.g. love, happiness, and optimism, or neutral sentiments (objective). This would be done through building a group of lexical databases and linguistic rules in order to analyze such opinions and conversations automatically. It is necessary to point out the possibility to update these databases based on the semantic field or the domain we want to analyze its content on social networking websites.**

1 INTRODUCTION

A. Social Media

There are so many definitions of social media; each of which focuses on certain group of aspects or characteristics. We can define it as a developed online interactive tool by which people can share knowledge and experiences.Social media has a special importance in today's world as a free and easy way of human communication. It is applied not only in social life as its name indicates, but also in fields such as industry, health, education, public sector…etc.

Worth mentioning that Facebook is the largest social media platforms in terms of active monthly users, as it recorded 1.32 billion monthly active users as of June 30, 2014.The total number of Facebook users in the Arab world as of beginning of May 2014 is 81,302,064 up from 54,552,875 in May 2013. Total number of active Twitter users in the Arab world reached 5,797,500 users as of March 2014. The estimated number of tweets produced by twitter users in the Arab world in March 2014 was 533,165,900tweets or 17,198,900tweets per day.

1) *Social Media Types:* Social media has many types that can be grouped according to many features; here we elaborate some of the most common types according to purpose of the platform. (Social Networks, Social Q&A, Microblogs, Video Sharing, Photo Sharing, Forums, Wikis…etc.)

2) *Social Media Analytics:* Analyzing user generated content shared on social media is becoming very important in a world that is increasingly using social media in a variety of fields. Analysis of social media conversations is especially concerned with quantitative and qualitative metrics such as the sentiment of users' opinions, influence of users, users' demographics, language and accents used, topics discussed…etc.
There are many automatic social media analytics tools, paid and free, but only few give accurate comprehensive analysis results. These tools can work as monitoring dashboards (e.g., HootSuite), social search engines (e.g., Topsy), comprehensive analysis tools (e.g., Radian6).

B. Sentiment Analysis

Sentiment analysis is analyzing a text in order to define the dominating sentiment of the author when he wrote the text. In social media it is concerned with analyzing posts and tweets in order to define the sentiment of the user who wrote them.

The sentiment can be one of the following; positive, negative, or neutral. One tweet/post can have more than one sentiment, but there's always a dominating sentiment.

Sentiment analysis is very challenging in terms of automation, especially in Arabic, because of many factors; one of them is the lack of accurate text analysis mechanisms, and another factor is that users on social media do not write in a standard unified language, but rater use a developed form of expression, in which we can find a mixture of standard language, colloquial language, and a developed way of using punctuation marks and emotion icons.

Importance of Arabic conversations sentiment analysis comes from the fact that social media users in the Arab world are increasing every day, hence the need for an automated tool to analyze a large set of data (text) produced by Arab users.

Sentiment analysis and opinion mining results can help in many ways, for instance it can help companies understand how they and their products or services are perceived by the public. This can be achieved through collecting all relevant mentions of the company and its products or services from all public social media platforms, then analyzing the dominating sentiment in these opinions.

C. Application of Sentiment Analysis in Business Research

Suggested mechanism can be used in various ways; for instance companies and organizations can breakdown users' opinions/reviews by platform, because each social platform has its own audience, hence analyzing the sentiment of every platform separately, and compare results not only to identify how satisfied audience are, but also to measure how successful the company was in building its image and reputation on these platforms. If for instance analysis results have proven a general dominant negativity in Twitter users' opinions, on the other hand opinions on Facebook were mostly positive. Considering the previous results, the company needs to review its communication/marketing strategy used on Twitter.

Sentiment analysis results could also be used in chronological comparison between overall users' opinion in a period (e.g., month) and a following or any other period. This can help companies identify the trend of users' opinions towards them, which is an indicator of sales and customer retention.

Another useful application is competition analysis. A company can analyze users' sentiments towards its competitors, and compare analysis results with its own, in order to know how people perceive all companies in the field, hence they can take data informed decisions, based on real-time data.

D. Related Work

We tried some free-to-use and paid social media analytics tools that provide sentiment analysis as part of their analytics products. Their Arabic sentiment analysis results were mostly not accurate due to - as we believe - their lack of a comprehensive linguistic analysis mechanism. Arabic is different than English in some ways, so a sentiment analysis mechanism should be specifically made and used for Arabic conversations.

Some of the common problems that we found in these tools[1] are: they rely on a very small corpus, don’t use both linguistic and lexical databases, neglect the common typing mistakes, process only one dialect, or they rely only on one level analysis without further classification of positive for instance into gratitude, content and so on.

As it's a relatively new field of interest for researchers, only a few research papers provide mechanisms of sentiment analysis of Arabic conversations. We used internet research to explore published research papers in the same field of Sentiment Analysis of Arabic Conversations. We found some interesting work that tackled this area and some even built linguistic datasets, but they mostly relied on book reviews, Ref [1] or movie reviews, Ref. [3], but we could not find any who built linguistic rules and corpus from social media public conversations, specifically tweets.

2 RESEARCH METHODOLOGY AND SAMPLE (SUGGESTED MECHANISM COMPONENTS)

A. Corpus

Constructing lexical and linguistic rules for any linguistic mechanism is known to be done through large linguistic repertoire (corpus); therefore the researchers have chosen Arabic conversations and discussions on the micro-blogging

[1] http://www.sdl.com/products/SM2/ | http://www.trackur.com/ | https://www.repustate.com/| http://www.brandwatch.com/

website Twitter to act as a kernel to a large-scale corpus. Researchers have chosen - as a sample - tweets mentioning a telecommunications company in Egypt called Vodafone, to be the kernel of a communications field corpus, and a start point to construct lexical and linguistic databases in the field of communications. The outcome includes about 300,000 Arabic words of comments, opinions, and conversations mentioning the company, its competitors, and/or the telecommunications field in Egypt.The linguistic mechanism relies on the following two types of databases.

B. Lexical Databases

Building a lexical database requires the following steps:

1) *Step 1:* to extrapolate the linguistic repertoire, in order to extract sentiment-bearing words and compositions that can help in judging the opinions and discussions (opinion mining) in terms of being positive or negative towards the entity subject to the analysis, which is Vodafone in our research sample. Table 1 presents some of these opinions "tweets":

TABLE 1.
SAMPLE OF TWEETS TO BE USED IN BUILDING THE CORPUS

ينعل ابو فودافون لابو شبكتها في
يعني هو أختك وأمك وقرايبك البنات خط أحمر وبنات الناس خط فودافون!!
يعني انا مشترك في خدمة "اعلن" ليا اكتر من شهرين ومش عارف اشغلها ومحدش في ام شركة فودافون كلها عارف يشغلهالي ؟ ينفع كدة؟
يـعـنـىآيـه !! SingLe يـعـنـىيـجـيـلـكمـسـجعـلـىآلـمـوبـآيـلومـتـفـتـحـآشعـلـشـآنعـآرفآنـهـآمـن فودافون
يا شركه وسخه #فودافون
يا جماعه اللي معاه فودافون يمسحه خلاص‌لان انا لغيت الخط
يا اوسخ شركه فى الوجود #فودافون
ولما تبقى مستنى رسالة من حد معين #فودافون مش هتنساك!

2) *Step 2:* to assign polarity (orientation) to words and compositions that were extracted in step 1. To guarantee the accuracy of this polarity assignment and hence the analysis results; and because one opinion can bear negative and positive words, we suggest that classification take the following form: Very Negative, Negative, Somewhat Negative, Somewhat Positive, Positive, and Very Positive. Table 2 presents examples, and the weight of each class:

TABLE 2.
SENTIMENT CLASSES, THEIR WEIGHTS, AND EXAMPLES

Class	Weight	Sample Tweet	Word Extracted
Very Negative	-100	يا شركه وسخه #فودافون	وسخة
Negative	-50	فيه حد هنا عنده مشكلة في نت فودافون G3 ؟؟ تقريبا قاطع عند كذا واحد	مشكلة
Somewhat Negative	-25	يا جماعه اللي معاه فودافون يمسحه خلاص‌لان انا لغيت الخط	لغيت
Somewhat Positive	+25	معظم المصريين اتصالات بس بتعجبهم اعلانات فودافون و بيحبوا اغاني موبينل.	بتعجبهم
Positive	+50	فودافون ريد .. حاجة حلوة (بسبوسه)	حلوة
Very Positive	+100	مع ان فودافون بردوا طلعت الشركة الوحيدة المحترمة واعلنت ان القطع تم بأمر الحكومة	المحترمة

No doubt that having one of these words in a sentence is a major indication of user's sentiment. With the class weight mechanism, we can identify the sentiment by identifying the dominant sentiment. Strength (here we refer to as weight) is defined as the degree to which the word, phrase, sentence, or document in question is positive or negative, Ref. [4].
Example: if an opinion has a word with weight -25 and another with +50 then the dominant sentiment here is positive.

3) *Step 3:* to identify the synonyms of extracted words, whether in Modern Standard Arabic (MSA) or Colloquial Arabic (we suggest using different dialects like Egyptian, Jordanian, and Kuwaiti…etc.). This step aims to enrich the mechanism databases with more sentiment-bearing words. Table 3 presents two words extracted from the sample corpus, and the synonyms derived from them.

TABLE 3.
SAMPLE OF TWO WORDS AND THEIR SYNONYMS

Sentiment	Positive	Negative
Sample Tweet	فودافون أجمل من يورك مفيش كلامD=	فودافون اصبحت أسوأ شبكة فى مصر من كافة النواحى وبصراحه اتصالات هى افضلهم

Extracted Word	أجمل	أسوأ
Synonyms	أفضل، أحسن، أحلى، أجود، أعظم	أردأ، أوسخ، أقبح، أحقر، أدنى

4) *Step 4:* to identify the different orthographic variations of all the sentiment-bearing words extracted from the corpus. This step aims at further enrich the lexical databases with different forms of spelling for the same word, considering not only different dialects, but also common spelling mistakes, and the so-called Franco-Arab ; which is Arabic words written using Latin Alphabet.
Example 1: the word سيئ can be written: سيء, سئ, سىء...etc.
Example 2: the word أحسن can be written: ahsan, a7sn, 27san…etc.

5) *Step 5:* the final step in constructing the lexical database; is deepen the level of analysis, by classifying words extracted from the corpus into a sub-level of tone and sentiment. Positive words can be grouped according to its sentiment to words that express/reflect content, gratitude, satisfaction…etc. Negative words can be grouped to words of anger, sadness, discontent…etc.
This second level of analysis aims at deepens the judgment on the opinions, by classifying it into sub-fields, each of which express a certain emotion or sentiment towards the entity subject to the analysis.

C. Linguistic Databases (morphological, syntactic, and semantic rules)

In our suggested analysis mechanism; we do not only rely on lexical databases to be used in Sentiment Analysis of Arabic Conversations on Social Media, but we also rely on linguistic rules that should be extracted from samples of tweets and posts mentioning the analysis subject, in our case here the telecom company Vodafone.These rules consist of the following 3 types.

1) *Morphological Rules:* When dealing with the Arabic language, we have to consider morphological rules, especially etymology; as Arabic words usually have a prefixes and/or suffixes, which results in many morphological variations of the same word. In the suggested mechanism we try to accommodate as many variations as possible for the sentiment-bearing words. For doing this we suggest to use the "stem" of the word, in order to guarantee the inclusion of all variations of the word, whether alone or with prefixes and/or suffixes.
Example: the word أجمل can have the following forms: أجمله, أجملهم, أجملهن, أجملكم, بأجمله. بأجملهم, بأجملهن, بأجملكم...etc.

2) *Syntactic Rules:* One of the most important syntactic roles in Arabic sentiment analysis is affirmation and negation, which plays an important role in opinion mining. By extrapolating the mechanism corpus; we can identify some of the negation particles in MSA and colloquial, e.g., لا – ليس – ليست - مش - ما – غير – مافي – مو – مب – موب –.
We can also identify questions by Interrogative Particles such as:إيه – فين – أين – وين.
Considering these particles, we came up with the following linguistic rules that help in identifying the conversation tone and users' sentiments:
1. Negation Particle + Negative Word = Positive Sentiment. E.g. مش غالي
2. Negation Particle + Positive Word = Negative Sentiment. E.g. مش رخيص
3. Interrogative Particle + Negative Word = Positive Sentiment. E.g. وين السيء؟
4. Interrogative Particles + Positive Word = Negative Sentiment. E.g. وين الجديد؟

3) *Semantic Rules:* Semantic analysis of punctuation marks (traditional or developed) is important in sentiment analysis. Also it is important to analyze the emotion icons (emoticons) e.g. :).

At the level of semantic analysis of the punctuation marks used in social media websites, we can conclude the following:
1. One question mark (?) indicates a Neutral sentiment, as the opinion is just a question waiting an answer.
Example: ازاي ممكن ابعت رسالة كلمني شكرا ببلاش من فودافون؟
2. Repeated question marks (???) or one or more exclamation marks (!!) indicates a negative sentiment towards the entity subject to the analysis.
Example: ؟ هو فودافون بظ ده الغي الاشتراك بتاعه ازاي!!!!
3. A good indication is the emoticons used in the opinion, as :), :D, ^^ indicate a positive sentiment, but :(, :/, :s indicate a negative sentiment. A database of these emoticons and their meanings can be collected from the internet, as a reference for the analysis mechanism.
At the level of the developed methods of written language on social media, we can conclude the following:

- Repeating a letter in a sentiment bearing word means a meaning emphasis, which indicates a very negativeor very positive sentiment.
  Example: أحلى وامتع دعايات حقت فودافون كوكاكولا مع أني مااحب اشربو بس اعلاناتو رااائعة

3 CONCLUSION

This research is a presentation of how to build a linguistic mechanism to analyze Arabic opinions and conversations, and provide judgment and classification to the tone of conversation and sentiment of users, in terms of being positive (e.g., hate, sadness, anger), negative (e.g., happiness, love, optimism) or neutral (objective). In this regard we presented how to build a group of lexical and linguistic rules in order to automatically analyze opinions' sentiments.

As for lexical databases, it contains a semantic classification for commonly used sentiment-bearing words (nouns and adjectives). The researchers chose to perform the research in the Telecommunications field, especially the telecom company Vodafone.

Linguistic databases consist of morphological, syntactic, and semantic rules concerning affirmation and negation and standard and developed punctuation marks. This research focuses on the linguistic side of the suggested mechanism and left the technical computational part to another research.

REFERENCES

[1] Aly, M., & Atiya, A., *LABR: Large Scale Arabic Book Reviews Dataset*, Meetings of the Association of Computational Linguistics (ACL), Sofia, Bulgaria, 2013.

[2] Mohammed Korayem, D. C.-M., *Advanced Machine Learning Technologies and Applications*, 2012.

[3] Rushdi-Saleh, M., Martín-Valdiv, M., Ureña-López, L., & Perea-Ortega, J., *Bilingual Experiments with an Arabic-English Corpus for Opinion Mining*, 2011.

[4] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M., *Lexicon-based methods for sentiment analysis,* computational linguistics 37, no. 2, 267-307, 2011.

[5] Abdul-mageed, M., & Diab, M. *AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis*, 2012.

[6] Abdul-Mageed, M., Kübler, S., & Diab, M., *SAMAR: A system for subjectivity and sentiment analysis of Arabic social media*, in Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (pp. 19-28), Association for Computational Linguistics, 2012.

[7] Al-Subaihin, A. A., Al-Khalifa, H. S., & Al-Salman, A. S.,*A proposed sentiment analysis tool for modern arabic using human-based computing,* in Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services (pp. 543-546), ACM, 2011.

[8] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S., *A system for real-time twitter sentiment analysis of 2012 us presidential election cycle*, in Proceedings of the ACL 2012 System Demonstrations (pp. 115-120), Association for Computational Linguistics, 2012.

[9] Mohammad, S. M., *# Emotional tweets,* in Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 246-255), Association for Computational Linguistics, 2012.

تحليل المشاعر فى المحادثات بالعربيةعلى شبكات التواصل الاجتماعى
دراسة للآليات اللغوية

عمرو جمعة عبدالرسول , محمدعلام فرغلى*
*كلية دار العلوم – جامعة القاهرة
**كلية التجارة – جامعة القاهرة
القاهرة – جمهورية مصر العربية

تطورت شبكات التواصل والإعلام الاجتماعي في الأعوام الأخيرة تطورا كبيرا، وغزت منصاتها الكثير من المجتمعات، حتى غدت وسائل إعلام جديدة تقوم على التواصل والتبادل وليس على التلقي فقط. ومع ازدياد هذه الأهمية زادت أهمية التطبيقات التي

تتعامل مع المحتوى الرقمي على هذه الشبكات، ومن أهم هذه التطبيقات تطبيقات تحليل المشاعرSentiment analysis ؛ وتحليل العواطف المسيطرة على مستخدم هذه الشبكات تجاه موضوع معين.

وتقوم فكرة البحث على بناء آلية لغوية تعمل على تحليل الآراء والحكم عليها من حيث كونها سلبية أو إيحابية أو محايدة، وكذلك الحكم على عواطف أصحابها من كونها عواطف سلبية مثل (الكره، والغضب، والحزن، ...) أو عواطف إيجابية مثل (الحب، والسعادة، والتفاؤل، ...) أو عواطف حيادية (موضوعية).

وتشمل المكونات اللغوية لآلية تحليل المشاعر وتحليل الآراء:

1- قاعدة البيانات (وتحوي تصنيفا دلاليا لكلمات (صفات وأسماء) تحمل معاني سلبية أو إيجابية أو يمكن أن توحي بعواطف سلبية أو إيجابية، وذلك حسب حقولها الدلالية ومجالاتها المختلفة، وتراعي الفصحى والعامية، فمحتوى هذه المنصات الاجتماعية يشتمل على الفصحى والعامية في الوقت نفسه.

2- القواعد اللغوية (الصرفية والدلالية) وتحوي قواعد لغوية أو قواعد للرسم الإملائي تساعد في تحليل مضمون منصات التواصل الاجتماعي والتعرف على العاطفة السائدة في نصوصها موضع التحليل.

الكلمات المفتاحية

تحليل المشاعر – تحليل الآراء - وسائل الإعلام الاجتماعي – الشبكات الاجتماعية - المنتديات – المدونات – المدونات المصغرة

BIOGRAPHY

Dr. Amr Gomaa obtained his doctorate in Arabic language and literature from the University of Cairo in 2014, with first-class distinction. He works as linguistic researcher at a natural-language processing firm, and contributed to the developments of several linguistic technologies. He has participated in several scientific conferences, including the ‘Arabic Content Online Challenges and Aspirations’ conference at Muhammad bin Saud Islamic University in Saudi Arabia in 2011 With a paper entitled ‘Linguistic criteria for evaluating Arabic-language internet search engines’, as well as the 8th and 10th Conferences of The Egyptian Society of Language Engineering (ESOLE) with papers entitled ‘Problems in the analysis of Arabic textual content on the Internet’ and ‘Linguistic and technical criteria for evaluating Arabic optical character recognition programs’.

Muhammad Allam received the bachelor degree in Commerce from the Faculty of Commerce, Cairo University in 2008. He has been working since graduation in the field of social media marketing, research and analysis. This is Muhammad’s first contribution to a scientific research paper as an independent researcher. His research interests are social media usages and applications, social media analytics and sentiment analysis.

Survey of Sentiment Analysis for Egyptian Dialect

Manal Mustafa*1, A.Shakour Al-Samahy*2, A.Fattah Elsharkawi*3, M. Gamal*4, Alaa Hamouda*5

**Systems and Computer Department, Faculty of Engineering*
Al-Azhar University, Cairo – Egypt
1manalmustafa10@yahoo.com
2amsamahy@windowslive.com
3sharkawi_eg@yahoo.com
4mgamal@4s-systems.com
5alaa_ham@gega.net

Abstract— **Using of Egyptian dialect in a text in order to express sentiments is posing a challenge to the automated sentiment analysis tools to correctly account for such colloquial words for sentiment. This paper surveys the sentiment analysis for Egyptian dialect and tackles an overview of the last update in this field. The techniques and methods in sentiment analysis and challenges appear in this field are outlined. We mentioned a popular approach and classification techniquesto perform a sentiment analysis on text-based status updates & comments to detect both positive and negative sentiments.**

1 INTRODUCTION

As social media gains popularity, it becomes more useful to analyze trends and sentiments of its users towards various topics. Sentiment analysis ([7], [11]-[18]) is the task dealing with the automatic detection and classification of opinions expressed in text written in natural language [20]. According to Simm et al [17] it is "the task of identifying positive and negative opinions, emotions, and evaluations". Subjectivity, it is the way that emotions and opinion can be expressed in the language while objectivity refers to the factual phrases [10]. Sentiment analysis is considered as a subsequent task to subjectivity detection, which should ideally be performed to extract content that is not factual in nature [20]. One goal of a sentiment analysis system is"given a text document; infer its polarity toward entities and events mentioned in the text" [8]. Sentiment analysis solutions can be divided according to the scope of the input such as document level, sentence level and word level sentiment analysis [14]. In addition, it can be divided according to the applied approaches to solve the problem of sentiment classification into three categoriesmachine learning (ML) approach, lexicon based approach and hybrid approach [19].

Mining opinions and sentiments from natural language is challenging, because it requires a deep understanding of the explicit, implicit, regular and irregular syntactical and semantic language rules [4]. Most of the current studies related to this topic focus mainly on English texts [7] with very limited resources available for other languages such as Arabic, especially the colloquial Arabic. Consequently, most of the resources developed (such as sentiment lexicons and corpora) are in English. Applying this research to other languages is a domain adaptation problem [4]. Even though Arabic ranks as the fifth largest natural language among the top 100 used natural languages worldwide [12] and is considered to be among the top ten languages mostly used on the Internet. Only few researches have been conducted on sentiment classification for Arabic languages due to the lack of resources for managing sentiments or opinions such as senti-lexicons and opinions [14]. A possible reason for that is the complex morphological, structural and grammatical nature of Arabic [16].

Arabic can be classified with respect to its morphology, syntax, and lexical combinations into three different categories: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectical Arabic (DA). Users on social networks typically use the latter, i.e. varieties of Arabic such as Egyptian Arabic, Iraqi,and Gulf Arabic. Moreover, dealing with Egyptian dialect (ED) creates additional challenges for researchers working on NLP: Being mainly spoken dialects, they lack standardization, are written in free-text and show significant variation from MSA. Actually, grammar and rules govern the use of MSA, but colloquial Arabic lacks grammar rules showing how to use it.

Based on our conducted survey, the main problems in the context of Arabic sentiment analysis can be concluded as: first, it is a very important task that contains natural language processing, web mining and machine learning [14]. Second, most research on sentiment analysis focus on text written in English and, consequently, most of the resources developed (such as sentiment lexicons and corpora) are in English. Third, Arabic language is both challenging and interesting because of its history, the strategic importance of the Arabic region, and its cultural and literary heritage [16].
The majority of the text produced by the social websites is considered to have an unstructured or noisy nature. This is due to the lack of standardization, spelling mistakes, missing punctuation, nonstandard words, repetitions, etc…That is why

the importance of pre-processing this kind of text is attracting the attention these days especially with the presence of several websites producing noisy text. We can conclude that the main problems in the context of ED sentiment analysis:

- There's no exact or specific pattern the user have to follow to write his review.
- The lack of suggested feature sets.
- Lack of classification algorithms to be used in the classification of the Arabic slang text.
- The majority of the available pre-processing tools like stemmers, stop-words lists, etc. are mainly built for the modern standard Arabic (MSA) lacking the dialect specific rules.
- The absence of slang specific lexicon with weights for each sentiment word makes this approach to be less investigated in the field of sentiment analysis for Arabic slang.

The reminder of this paper is organized as follows: Section 2surveyssentiment analysis for Arabic and dialectical Arabic, while section 3 gives the popular sentiment analysis methodology. Section 4 outlines sentiment classification techniques, while section 5 discusses training the classifier. Section 6 gives related fields to sentiment classification and section 7 defines the evaluation metrics. Finally, section 8 concludes and gives possible directions for future work.

2 SENTIMENT ANALYSIS FOR ARABIC AND DIALECTICAL ARABIC

Related studies used different tools and techniques to extract Arabic sentiments to automatically determine their polarities. Examples of these studies are summarized in this survey. Refaee and Rieser [15] presented a newly collected corpus of annotated twitter feeds annotated for subjectivity and sentiment analysis. They employ morphological features, simple syntactic features, such as n-grams, as well as semantic features. Khalifa and Omar [10] addressed the difficulties of Arabic opinion question answering(QA) by proposing a hybrid method of lexicon-based approach and classification using Naïve Bayes classifier. Their experimental results achieve 91% accuracy. El-Halees proposed a combination of lexicon- based, maximum entropy and k-nearest neighbor with using a corpus of Arabic words in order to classify opinionated Arabic documents [6].

Moreover, Al-Subaihin et al [1] designed and implemented a lexicon-based sentiment analysis tool dedicated to colloquial Arabic text used in some Arabic social media Websites. Those researchers proposed that, their tool should rely partially on human judgment to overcome the problem arose from using non-standardized colloquial Arabic text. Bautin et al [2] utilized machine translation to translate source foreign text (non-English) to English, and then to conduct sentiment analysis on the machine translated English text. This study includes conducting sentiment analysis on textual news and blogs, which use one of the following eight natural languages: Arabic, Chinese, French, German, Italian, Japanese, Korean, and Spanish. Non-English text first has to be translated automatically to English, in order to be analyzed by a system designed only to analyze English text.

Al-Kabiet et al [12] used sentiment analysis to identify sentiments with their subjectivity from the huge volume of reviews. In order to conduct their study a small dataset consisting of 4,050 Arabic and English reviews were collected. Three polarity dictionaries were also created (Arabic, English, and Emoticons). The collected dataset and those dictionaries were used to conduct a comparison between two free online sentiment analysis tools (SocialMention and Twendz). Finally, Omar et al [14] proposed an ensemble of machine learning classifier framework for handling the problem of subjectivity and sentiment analysis of Arabic customer reviews.

3 SENTIMENT ANALYSIS METHODOLOGY

Khalifa and Omar [10] build a framework to determine the flow of the phases which include transformation, normalization, tokenization, feature extraction and classification. Also, Shoukry and Rafea [16] summaries the ML process of the sentence's sentiment analysis in the Arabic language using Arabic tweets from the social network website twitter. Fig. 1 outlines these different phases.

A. Dataset Characteristics

Sentiment analysis can be a bit challenging as people generally don't pay any heed to the spellings and deliberately modify the spellings of the words. For example, many people write the word “فى التليفون” as “ف التليفون”, “على الناصية” as “ ع الناصية”. Also, they may use short forms whenever required. For example, “صلى الله عليه وسلم” as “ص”, and the list goes on. This poses a challenge to correctly process the language and find out the polarity of the sentence or the paragraph. In addition, unstructured or noisy data, which may due to:

- Lack of standardization and ambiguity.
- Very complex morphology (lack of standard Arabic morphological analysis tools) [10].
- Arabic opinions are highly subjective to context domains, where you may face words that have different polarity categories in different contexts (عملية قلب مفتوح،فى قلب الاحداث، فى قلب الملعب).
- Mixture of English and Arabic text (fantastic – شكله رائع).
- Other reviews such as:” 7ilwi awi” which means in English:”very sweet”.

- Latin letters and English phonetics (transliteration) are used to express Arabic phrases such as “jamiljidannnnn”.
- Emoticons as noisy labels (e.g. Happy emoticons: “:-)”, “:)”, “=)”, “:D” and Sad emoticons: “:-(”, “:(”, “=(”, “;(”) and acronyms. Vashisht and Thakur [18] analyzed the role that emoticons play in delivering the overall sentiment of the text. They identified the commonly used emoticons and exploited them to devise a finite state machine that takes these typographical symbols as an input and conveys the associated sentiment as an output.
- Frequency of special characters such as (!) and (?) which have significant effect on sentiment analysis.

The extracted data was cleaned in a pre-processing phase, e.g. by normalizing user-names, digits, and eliminating Latin characters (i.e. URLs, emails).

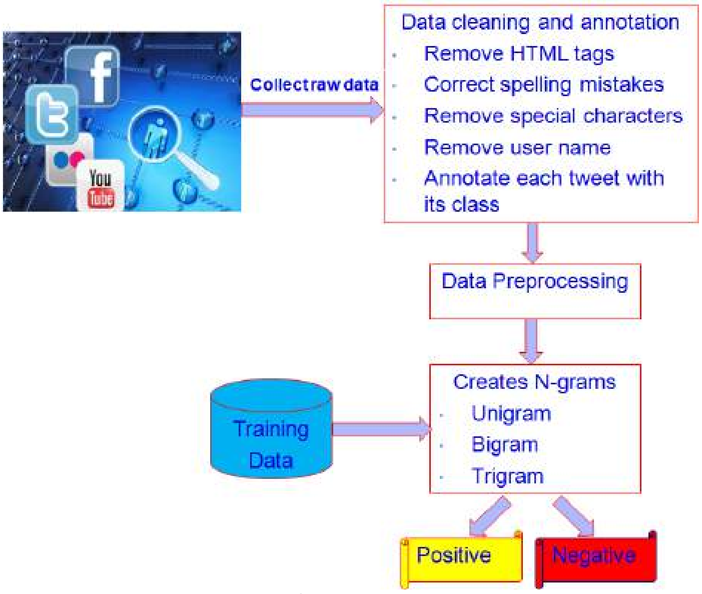


Figure 1: An Outline of Sentiment Analysis Methodology

B. *Data pre-processing*

After the data has been collected and manually classified, it is passed through a series of pre-processing steps as shown in Fig. 2. Then, each message is decomposed into a set of features, which in the model used are represented mostly by words that can be taken as input for the classifier.

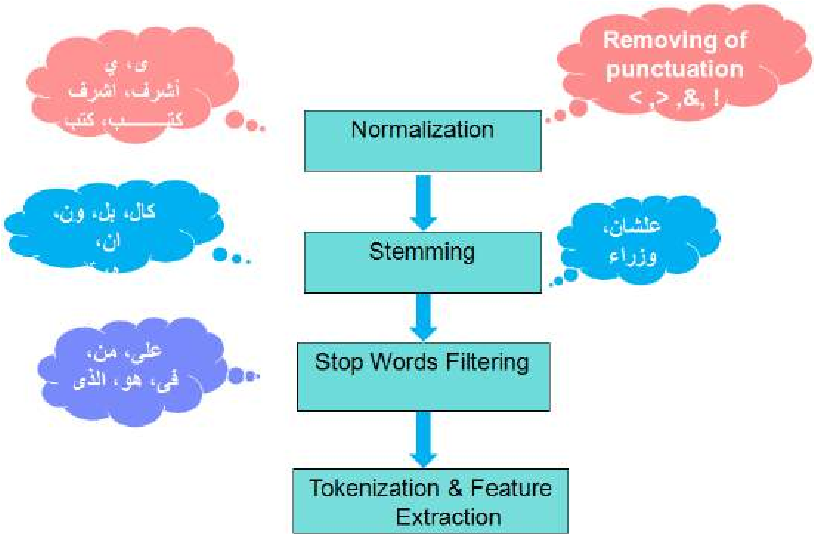


Figure 2: Pre-processing Stages

1) *Normalization:* Before normalization process the text need to be sanitized. That is, HTML tags and non-textual contents were striped out. After that, the text is ready to be normalized. The normalizing process puts the Arabic text in a consistent form, thus converting all the various forms of a word to a common form. Normalization step executes a set of text processing techniques that sanitize and normalize the raw data. Because of the informal and non-grammatical *nature*

of the colloquial language used in social media, these steps gain significant importance. These techniques have been proven to improve the quality of the features extracted from such data and therein improve the performance of the classifier used afterwards.

- *Punctuation.* Irrelevant punctuations were removed from the messages for consistency. However, in social media it is common to use excessive punctuation in order to convey emotions. Soothe series of exclamation marks or combinations of exclamation and question marks were replaced with a keyword before removing all punctuation.
- *Detecting repeated alphabets.* For Arabic scripts, some alphabets have to be normalized, this appears in different forms.

    - Some repeated letters have been cancelled (that happens in discussion when the user wants to insist on some words. i.e., is a common method for indicating powerful emotions), and therefore can relate to the message sentiment. E.g., (رائع جداااااااااااا).
    - The letters which have more than one form, e.g., (أنت، إنت ، انت).
    - Some of the wrong spelling words are corrected. E.g., (جميل– جمي).
    - Diacritics extraction. All diacritics have to be removed (بُ بْ بّ بِ ب).
- *Replacing emoticons.* Many social media messages make use of emoticons in order to transmit emotion, making them very useful for sentiment analysis. In addition, variations of laughter such as "ههههههه" were all replaced with a particular keyword.
- *Replacing URLs.* Many reviews contain URLs in order to share more content than can be given in the limited messages. Since URLs are unique, they were removed from the messages to avoid including them as possible features.
- *Replacing platform specific characters.* Twitter messages use the '@' character in front of a username to address other users inside the platform. As done before, these pointers were replaced with a special keyword.
- *Removing repeated characters.* For emphasizing messages, some words might include repeated characters. These occurrences were compressed to their original form by using regular expressions techniques.
- *Special Character Extraction:* Every character seems to be non-Arabic letter has to be removed (e.g.,><& ; ""). These steps are important for improving consistency.

2) *Stemming:* As mentioned before, Arabic language has numerous forms of verbs; those derivations have to be stemmed to its own roots. Stemming is the process for removing and replacing common suffixes of Arabic words.E.g.,(المسافر - المسافرون - المسافرين).The purpose of this step is to reduce the size of the feature set presented to the classifier. Stemming Arabic terms has proven in several researches that it is not an easy task because of its highly inflected and derivational nature. Shoukry and Rafea [16] discuss a stemmer in more details.

3) *Stop Words Filtering:* It is also useful to ignore very common words of the messages that do not provide any useful information in the classification. In sentiment analysis, these words all called stop words and mainly consists of pronoun, articles, and prepositions (e.g., ثم - عند–من – هو –الذى–فى) and so on. For this step, the Arabic stop word included in NTLK corpus was used.

After applying Arabic stemmer and removing stop words, the sentences are tokenized. Then, the obtained vector representations for the terms from their textual representations by performing TFIDF (Term Frequency–Inverse Document Frequency) weight which is a well known weight presentation of terms often used in text mining.

4)Tokenization and Feature Extraction: The first step in the sentiment classification problem is to extract and select text features. Walaa, Ahmed and Hoda [19] refer to these current features such as:

- *Terms presence and frequency:* These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears or one if otherwise) or uses term frequency weights to indicate the relative importance of features.

- *TFIDF technique:* This method evaluates words in each sentence to minimize the influence of those that are very common across the document and do not carry much meaning. The importance of a word is high if it is frequent in a particular sentence, but less frequent in other.
- *Parts of speech (POS):* finding adjectives, as they are important indicators of opinions.
- *Opinion words and phrases:* these are words commonly used to express opinions including good or bad, like or hate. On the other hand, some phrases express opinions without using opinion words.
- *Negations:* the appearance of negative words may change the opinion orientation like not good is equivalent to bad.

Then, a Bag of words (BOW) model was used to transform the list into a feature set that is consumable by the classifier. This simplifying model takes individual words as features, assuming their conditional independence and equality. So, the messages are represented by an unordered collection of words, disregarding grammar and even word order. Each feature represents the existence of one word. For improving accuracy, Refaee and Rieser [15] incorporate some additional features to the classifier. Those features are summarized as follows:

1) *Morphological features:* Considering the morphologically rich nature of Arabic, the following features may be annotated: aspect, gender, mood, number, person, and voice (e.g. active). Using automatic morphological analyzer for Arabic text to obtain these features. In particular, current version of MADA+TOKAN can be incorporated which performs tokenization, morphological disambiguation, Part-of-Speech (POS) tagging, stemming and lemmatization for Arabic [9]. It is important to note that MADA is developed for Modern Standard Arabic (MSA) only.
2) *Semantic features:* It includes a number of binary features that check the presence of sentiment bearing words of a polarity lexicon in each given tweet.
3) *Stylistic features:* This feature-set includes two binary features that check the presence of positive/negative emoticons [15].
4) *Syntactic features:* Refers to the annotated part-of-speech for each word (adjectives, adverb, nouns and verbs). The common syntactic that usually used is the adjectives.

4 SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment Classification (SC) techniques can be divided into machine learning approach, lexicon based approach and hybrid approach [19]. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. Khalifa and Omar [10] proposed a hybrid classification method consists of Naïve Bayes classifier and lexicon-based approach in order to build Arabic Opinion question answering. Three classifiers have been carried out with the lexicon approach. Those classifiers are NB, SVM and KNN. Their experiment shows that, NB has demonstrated best results of macro-average F-measure of 91%.Therefore, NB was selected to be combined with the lexicon-based approach in order to solve the problem of sentence-level sentiment analysis. The various approaches and the most popular algorithms of SC are illustrated in Fig 3.

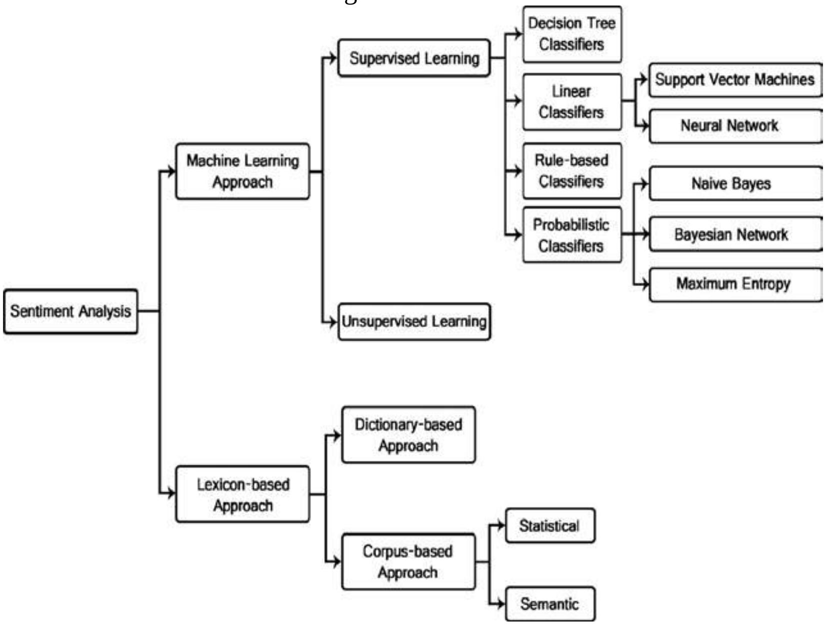


Figure 3:Sentiment Classification Techniques [19]

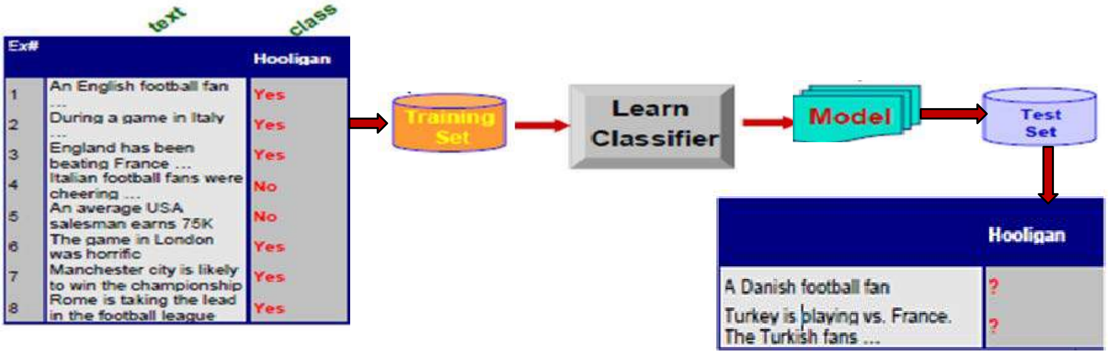
The text classification methods using ML approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents as indicated inFig. 4. N. Anitha et al [13] covered a number of categorization algorithms and semantic orientation approaches which are widely used in sentiment classification. The authors explained the importance and usage of several techniques and look insight into the various machine learning techniques for sentiment classification such as 1) Machine Learning Approaches, 2) Semantic Orientation Approaches and 3) Novel Machine Learning Approaches. Another study introduced by Vinodhini and Chandrasekaran [7] surveyed a number of ML algorithms that can be used in sentiment classification.

The unsupervised methods are used when it is difficult to find labeled training data. The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approachwhich depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms [19]. The corpus based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

5 TRAINING THE CLASSIFIER

The ML approach is typically a supervised approach [7] in which a set of data labeled with its class such as “positive” or “negative” are represented by feature vectors. Then, these vectors are used by the classifier as a training data inferring that a combination of specific features yields a specific class employing one of the supervised categorization algorithms as depicted inFig 4.The general process of sentiment analysis is to induce a classifier by using a set of training data with manually assigned category labels (positive, negative or neutral) and then apply it to predict labels for uncategorized reviews.

- Given: a collection of labeled records (training set). Each record contains a set of features (attributes), and the true class (label).
- Find: a model for the class as a function of the values of the features.
- Goal: previously unseen records should be assigned a class as accurately as possible.



Ex#		Hooligan
1	An English football fan …	Yes
2	During a game in Italy …	Yes
3	England has been beating France …	Yes
4	Italian football fans were cheering …	No
5	An average USA salesman earns 75K	No
6	The game in London was horrific	Yes
7	Manchester city is likely to win the championship	Yes
8	Rome is taking the lead in the football league	Yes

	Hooligan
A Danish football fan	?
Turkey is playing vs. France. The Turkish fans …	?

Figure 4: Classification Process

6 RELATED FIELDS TO SENTIMENT CLASSIFICATION

There are some topics that work under the umbrella of SC and should be taken in consideration. In the next subsection, some of these topics are presented with related articles.

A. *Negation*

Also, negations [3] will be considered as a feature in ML approach because their presence in the sentence can result in changing the sentiment of the whole tweet.It is a very common linguistic construction that affects polarity and, therefore, needs to be taken into consideration in sentiment analysis [20]. Negation is not only conveyed by common negation words (ليس- غير-لا – لم- مش) but also by other lexical units. Research in the field has shown that there are many other words that invert the polarity of an opinion expressed, such as valence shifters, connectives or modals “ بصراحة أداء المنتخب اقل من المتوقع”. The scope size of a negation expression determines which sequence of words in the sentence is affected by negation words.

Roth et al [20] presented a survey on the role of negation in sentiment analysis and discussed various computational approaches modeling negation in sentiment analysis. Actually, negation terms affect the contextual polarity of words but the presence of a negation word in a sentence does not mean that all of the words conveying sentiments will be inverted. That is why we also have to determine the scope of negation in each sentence. Blanco and Moldovan [5] explore the importance of both scope and focus to capture the meaning of negated statements and outline some issues on detecting negation from text. The authors also depict the forms in which negation occurs and heuristics to detect its scope and focus.

B. *Emotion detection*

SA is concerned mainly in specifying positive or negative opinions, but emotion detection (ED) are concerned with detecting various emotions from text. As a Sentiment Analysis task; ED can be implemented using ML approach or Lexicon-based approach, but Lexicon-based approach is more frequently used [19].There are eight basic and prototypical emotions which are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. Vashisht and Thakur [18] have analyzed the role that emoticons play in delivering the overall sentiment of the text. They identified the commonly used emoticons and exploited them to devise a finite state machine that takes these typographical symbols as an input and conveys the associated sentiment as an output.

C. *Building resources*

Building resources (BR) aims at creating lexica, dictionaries and corpora in which opinion expressions are annotatedaccording to their polarity. Building resources is not a SA task, but it could help to improve SA and ED as well.

The mainchallenges that confronted the work in this category are ambiguity of words, granularity and thedifferences in opinion expression among textual genres [19]. One of the problems in the area of sentiment analysis of the Arabic text is the unavailability of free corpora specified for subjectivity and sentiment analysis. Omar et al [14] decide to create their own subjectivity and sentiment analysis annotated Arabic data.Also, the study proposed by Al-Kabi et al [12]started by collecting 4,050 Arabic/English reviews to construct a dataset, then the reviews in this dataset are tokenized to construct manually three polarity dictionaries to determine the polarity of each Arabic/English review. One of these three dictionaries is dedicated to Arabic reviews, while the second is dedicated to English reviews, and the third one is dedicated to emoticons.

Moreover, Arabic language has many slangs. Most of times, reviewers write their reviews in their own dialects. Different dialects may use different words to express the same opinion, for example, (جامدة جدا، روش طحن، كويس، جميل، حلو، لذيذ). To handle this problem, researches create a lexicon containing dialectical words and their slandered Arabic equivalents.Sentiment lexicon or so called senti-lexicons is the process where each word is associated with a polarity score which indicates the orientation of the word (positive or negative). The sentiment lexicon is the most sensitive resource for most sentiment analysis algorithms. Arabic sentiment lexicon may contain a syntactic refers to the annotated part-of-speech for each words (adjectives, adverb, nouns and verbs). The common syntactic that usually used is the adjectives. In addition, it contains equivalent dialectical synonyms and a score refers to the degree of polarity from most bad to most good which has been ranged between -5 to 5. Eventually, inflections forms refer to the forms that the word can be formulated whether for singular, female, dual, or plural [10].

7 EVALUATION METRICS

The performance of different methods used for sentiment analysis is evaluated by calculating various metrics like precision, recall and F-measure. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. The two measures are sometimes used together in the F- score (also F-score or F-measure) [7].

8 CONCLUSIONS AND FUTURE WORK

Due to the sheer volume of opinion rich web resources such as discussion forum, review sites, blogs and news corpora, social media; much of the current research is focusing on the area of sentiment analysis. We briefly surveyed sentiment analysis problem and mentioned different methods for building subjectivity and sentiment analysis systems for Arabic text. Actually, usage of informal language, short cuts & emoticons is increasing rapidly. Technical challenges of the Egyptian dialect and some of its solutions are exposed. Different types of features and classification algorithms are incorporated in an efficient way in order to overcome their individual drawbacks which has significant effect on classification accuracy.

One possible direction for future work is to extend the proposed feature set. More research may be conducted on negation expressions to produce a summary of sentiments based on product features/attributes. Complexity of sentence and handling of implicit product features is also a versatile area of research in this topic.

REFERENCES

[1] Al-Subaihin, A.; Al-Khalifa, H.; Al-Salman, A.;“A proposed sentiment analysis tool for modern Arabic using human based computing”, *in Proc. of the 13th International Conference on Information Integration and Web-based Applications and Services (iiWAS '11)*, pp. 543-546, ACM. New York, NY, USA, 2011.

[2] Bautin, M.; Vijayarenu, L.; Skiena, S.*;“International Sentiment Analysis for News and Blogs”, In 2nd International Conference on Weblogs and Social Media (ICWSM)*, pp.19–26, 2008.

[3] Blanco, E.; Moldovan, D.; “Some Issues on Detecting Negation from Text”, *in Proc. of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, pp. 228- 233, 2011.

[4] Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C.; “New Avenues in Opinion Mining and Sentiment Analysis”, pp. 15-21, March /April 2013 IEEE.

[5] Dadvar, M.; Hauff, C.; Jong, F.; “Scope of Negation Detection in Sentiment Analysis”, 2011.

[6] El-Halees, A.; “Arabic opinion mining using combined classification approach”, *in Proc. of the International Arab Conference on Information Technology, (CIT' 11)*, Azrqa, Jordan, pp. 1-8, 2011.

[7] G. Vinodhini; RM. Chandrasekaran; “Sentiment Analysis and Opinion Mining: A Survey”, *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 6, pp.282-292, June 2012.

[8] Guerra, P. C.; Meira, Jr.; Cardie, C.; “Sentiment Analysis on Evolving Social Streams: How Self-Report Imbalances Can Help”, ACM, New York, USA, Feb. 2014.

[9] Habash, N.; Rambow, O.; Roth, R.; “MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization”, In Choukri, K. and Maegaard, B., editors, *Proc. of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium, pp. 102 – 109, 2009.

[10] Khalifa, K.; Omar, N.; “A Hybrid Method Using Lexicon-Based Approach and Naïve Bayes Classifier for Arabic Opinion Question Answering ”, *Journal of Computer Science (JCS)*, vol. 10, no.10, pp. 1961-1968, 2014.

[11] Mageed, A.; M.; Diab, M.,T.; Korayem, M.; “Subjectivity and Sentiment Analysis of Modern Standard Arabic”, *in Proc. of the 49th Annual Meeting of the Association for Computational Linguistics:* Portland, Oregon, Association for Computational Linguistics. Shortpapers, pp. 587–591, June 2011.

[12] M. Al-Kabi; N. M. Al-Qudah; I. Alsmadi “Arabic / English Sentiment Analysis: An Empirical Study”, Irbid – Jordan, ACM, April 2013.

[13] N. Anitha; B. Anitha; S. Pradeepa; “Sentiment Classification Approaches – A Review”, *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 3,no 1, Oct. 2013.

[14] Omar, N.; Albared, M.; Al-Shabi, A.; Al-Moslmi, T.; “Ensemble of classification algorithms for subjectivity and sentiment analysis of Arabic customer's reviews”, *International Journal of Advancements in Computing Technology (IJACT)* vol. 5, no. 14, pp. 77 – 85, Oct. 2013.

[15] Refaee, E.; Rieser, V.; “An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis”, pp. 2268 – 2273, 2014.

[16] Shoukry, A.; Rafea. A. ; “Sentence level Arabic sentiment analysis”, *International Conference of Collaboration Technologies and Systems (CTS)*,pp.546-550, May 2012.

[17] Simm, W.; Ferrario, M. A.; Piao,S.; Whittle, J.; Rayson, P.; “Classification of Short Text Comments by Sentiment and Actionability for VoiceYourView”, pp.552 – 557,2010 IEEE.

[18] Vashisht, G.; Thakur, S.; “Facebook as a Corpus for Emoticons-Based Sentiment Analysis”, *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 5, pp. 904 – 908, May 2014.

[19] Walaa M.; Ahmed H.; Hoda K.; “Sentiment analysis algorithms and applications: A survey”, *Ain Shams Engineering Journal*, pp. 1 – 21, April 2014.

[20]Wiegand, M. ; Balahur, A.; Roth, B.; and Klakow, D.; Montoyo, A.; “A Survey on the Role of Negation in Sentiment *Analysis”, 2011.*

BIOGRAPHY

Manal Mustafa, is an Assistant Lecturer at the Engineering faculty of Azhar University. She worked at National Authority for Remote Sensing and Space Sciences (NARSS) form 2009 to 2011. She got her M.Sc. and B.Sc. from Computers and Systems Engineering Department, Azhar University in 2014, 2008 respectively. Fields of interest: Software Engineering, Text Mining, Database System, and Artificial Intelligence.

Abdelfatth Elsharkawi, Born 1957 Egypt, BS 1981 Systems and Computer Eng, AlAzhar University Egypt. He got MS.C 1986 Systems and Computer Eng, AlAzhar University Egypt, PH.D 1993 Systems and Computer Eng, AlAzhar University Egypt. Current occupation Associate Professor Software Engineering Systems and Computer Eng, AlAzhar University Egypt.

Mohamed Gamal, is a Teacher at the Engineering faculty of Azhar University, and CEO of 4S Technology Company. He publishes five researches papers in International conferences. In practical field, he has more than 30 years of experiencefor playing all roles of the software development cycle starting as a programmer to a system analyst and designer to project manager, division manager, director of professional services and managing director.

Alaa Hamouda is an Associate Professor at the Engineering faculty of Azhar University. He has a lot of published papers in the international conferences and referred journals. In practical field, he has more than 15 years of experience in software project management, software quality models of Capability Maturity Models Integration (CMMI), and Agile. As a consultant, he participated in applying CMMI quality models in Egyptian companies. He participated in establishing reliable standard work procedures in software houses to apply standard project management processes (PMI). He has played the role of Software Board member and Engineering Process Group (EPG) member as well as Project Management Office Director (PMO Director) in several Egyptian Companies.

دراسة تحليل الآرء في اللهجة العامية المصرية

*منال مصطفى ،*عبدالشكور السماحى،*عبدالفتاح الشرقاوى،*محمد جمال،*علاء حمودة
* قسم النظم والحاسبات، كلية الهندسة، جامعة الأزهر

استخدام العامية المصرية على مواقع التواصل الاجتماعى يمثل تحديا كبيرا بالنسبة لتحليل الرؤى والاتجاهات المختلفة إزاء موضع معين. هذاالبحث يقوم بعمل مسح للتحديات والصعوبات التى تواجه تحليل الآراء والمشاعر المعبر عنها بالعامية المصرية. كما تم عرض طرق التصنيف والتقنيات المختلفة المستخدمة فى التعلم الآلى، والاساليب العلمية التى يمكن توظيفها لتحديد الانفعالات والاتجاهات المختلفة. وبناءا على ذلك يتم تصنيف ما اذا كان هذا الرأى إيجابيا او سلبيا.

Corpora Preparation and Stopword List Generation for Arabic Data in Social Network

Walaa Medhat[*1], Ahmed H. Yousef[**2], Hoda Korashy[**3]

[*] *School of Electronic Engineering, Canadian International College, Cairo campus of CBU*
Land 6, south of Police Academy, Eltagamo Elkhames, New Cairo, Egypt

[1]walaamedhat@gmail.com

[**] *Computers & systems Department, Faculty of Engineering, Ain Shams University*
1 El-Sarayat st., Al-Abbasiyah, Cairo, Egypt

[2]ahassan@eng.asu.edu.eg
[3]hoda.korashy@eng.asu.edu.eg

Abstract—**This paper proposes a methodology to prepare corpora in Arabic language from online social network (OSN) and review site for Sentiment Analysis (SA) task. The paper also proposes a methodology for generating a stopword list from the prepared corpora. The aim of the paper is to investigate the effect of removing stopwords on the SA task. The problem is that the stopwords lists generated before were on Modern Standard Arabic (MSA) which is not the common language used in OSN. We have generated a stopword list of Egyptian dialect and a corpus-based list to be used with the OSN corpora. We compare the efficiency of text classification when using the generated lists along with previously generated lists of MSA and combining the Egyptian dialect list with the MSA list. The text classification was performed using *Naïve Bayes* and *Decision Tree* classifiers and two feature selection approaches, *unigram* and *bigram*. The experiments show that the general lists containing the Egyptian dialects stopwords give better performance than using lists of MSA stopwords only.**

1 INTRODUCTION

Sentiment Analysis is the computational study of people's opinions, attitudes, and emotions towards topics covered by reviews or news as in Ref. [1]. SA is considered also a classification process which is the task of classifying text to represent a positive or negative sentiment as in Ref. [2] – [4]. The classification process is usually formulated as a two-class classification problem; positive and negative. Since it is a text classification problem, any existing supervised learning method can be applied, e.g., Naïve Bayes (NB) classifier.

The web has become a very important source of information recently as it becomes a read-write platform. The dramatic increase of OSN, video sharing sites, online news, online reviews sites, online forums and blogs has made the user-generated content, in the form of unstructured free text gains a considerable attention due to its importance for many businesses. The web is used by many languages' speakers. It is no longer used by English speakers only. The need of SA systems that can analyze OSN in other languages than English is compulsory.

Arabic is spoken by more than 300 million people, and is the fastest-growing language on the web (with an annual growth rate of 2,501.2% in the number of Internet users as of 2010, compared to 1,825.8% for Russian, 1,478.7% for Chinese and 301.4% for English) (http://www.internetworldstats.com/stats7.htm). Arabic is a Semitic language as in Ref. [5] and consists of many different regional dialects. However, these dialects are true native language forms which are used in informal daily communication and are not standardized or taught in schools as in Ref. [6]. Despite this fact but in reality the internet users especially on OSN sites and some of the blogs and reviews site as well, use their own dialect to express their feelings. The only formal written standard for Arabic is the MSA. It is commonly used in written media and education. There is a large degree of difference between MSA and most Arabic dialects as MSA is not actually the native language of any Arabic country as in Ref. [7].

There is lack of language resources of Arabic language as most of them are under development. In order to use Arabic language in SA, there are some text processing techniques are needed like removing stopwords or Part-of-Speech (POS) tagging. There are some sources of stopword lists and POS taggers are publicly available but they work on MSA not on Dialect Arabic (DA). This paper tackles the first problem of removing stopwords. Stopwords are more typical words used in many sentences and have no significant semantic relation to the context in which they exist.

In the literature, there are some research works have generated stopword lists but as far as our knowledge no one has generated a stopword list for DA. Reference [8] has proposed an algorithm for removing stopwords based on a finite state machine. They have used a previously generated stopword list on MSA. Reference [9] has created a corpus-based list from newswire, query sets and a general list using the same corpus. Then compares the effectiveness of these lists on the information retrieval systems. The lists are on MSA too. Reference [10] has generated a stopword list of MSA from the highest frequent meaningless words that appear in their corpus.

The aim of this paper is to investigate the effect of removing stopwords on SA for OSN Arabic data. Since the OSN sites and the reviews sites use the simple Egyptian dialect. The creation of a stopword list of Egyptian dialect is mandatory. The data are collected from OSN sites Facebook and Twitter as in Ref. [11] – [20] on Egyptian movies. We used an Arabic review site as well that allow users to write critics about the movies (https://www.elcinema.com). The used language by the users in the review is syntactically simple with many words of Egyptian dialects included. The data from OSN is characterized by being noisy and unstructured. Abbreviations and smiley faces are frequently used in OSN and sometimes in review site too. There is a need for many preprocessing and cleaning steps for this data to be prepared for SA. The Arabic users either write with Arabic or with Franco-arab (writing Arabic words in English letters) e.g. the word “maloosh” which stands for “مالوش” which means “doesn’t have”. This is an Egyptian dialect word which is written in MSA as “ليس له”. Sometimes they use English word in the middle of an Arabic sentence which must be translated.

We are tackling the problem of classifying reviews and OSN data about movies into two classes, positive and negative as was first presented in Ref. [2]; but on Arabic language. In their work they used unigram and bigram as Feature Selection (FS) techniques. It was shown in Ref. [2] that using unigrams as features in classification gives the highest accuracy with NB. We have used the same feature selection techniques, unigram and bigram along with NB and Decision Tree (DT) as classifiers.

We have proposed a methodology for preparing corpora from OSN which consists of many steps of cleaning, converting Franc-arab to Arabic words and translation of English words that appear in the middle of Arabic sentences to Arabic. We have also proposed a methodology of generating stopword lists from the corpora. The methodology consists of three phases which are: calculating the words’ frequency of occurrence, check the validity of a word to be a stopword, and adding possible prefixes and suffixes to the words generated.

The contribution of this paper is as follows. First, we propose a methodology for preparing corpora from OSN sites in Arabic language. Second, we propose a methodology for creating a stopword list for Egyptian dialect to be suitable for OSN corpora. Third, we prepare corpus from Facebook which was not tackled in the literature before for Arabic language. Fourth, tackling OSN data in Arabic language is new as it wasn’t investigated much. Fourth, tackling DT classifier with these kinds of corpora is new as it wasn’t investigated much in the literature. Finally, the measure of classifiers’ training time and considering it in the evaluation is new in this field.

The paper is organized as follows; section 2 presents the methodology. The stopword list generation is tackled in section 3. The Experimental setup and results are presented in section 4. A discussion of the results and analysis of corpora is presented in section 5. Section 6 presents the conclusion and future work.

2 METHODOLOGY

The aim of our study is to prepare data from Twitter, Facebook, and a review site on the same topic in Arabic language for SA. We have chosen a hot topic on the recently shown movies in the theatres for the last festival in first of August 2014. The movies were:”*الفيل الأزرق*” means “*The blue elephant*”; “*صنع فى مصر*” means “*Made in Egypt*”; “*الحرب العالمية التالتة*” means “*The third world war*”; and “*جوازة ميري*” means “*official marriage*”. We have downloaded related tweets from twitter, comments from some movies’ Facebook pages, and users’ reviews from the review site elcinema.com.

Tweets were downloaded about the movies using the regular search of Twitter as many of the sites that retrieve tweets are closed like (http://searchhash.com/, http://topsy.com/). We have searched using the movies' names and downloaded all the tweets that appear at the time of search. There were many unrelated tweets downloaded as some of the movies like "صنع فى مصر" and "الحرب العالمية الثالثة" can hold other meanings than the movies' title. The retrieved tweets are tweets that contain the whole words or any word either in the text or with hashtag.

The methodology we have used is very close to what was proposed in Ref. [21]. But there are some discrepancies related to the nature of the Arabic language. We have also used the removing stopwords only as a text processing technique due to lack of sources especially for DA.

A. Corpora Preparation

The data downloaded are prepared to be able to be fed to the classifier as shown in Fig. 1.

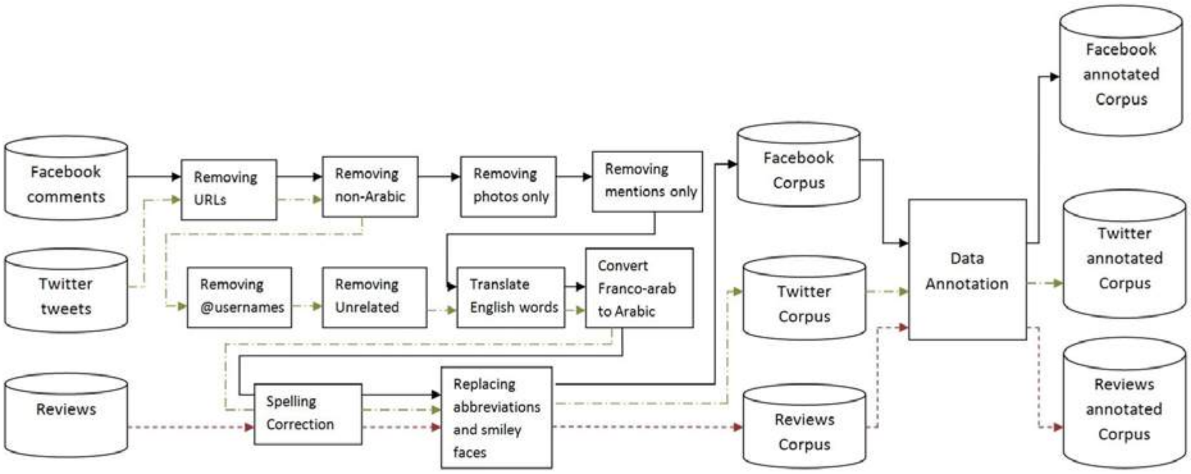


Figure 1: Arabic Corpora Preparation from Reviews, Facebook, and Twitter

The number of comments after removing the comments that contain URLs only or advertising links from Facebook was 1459. Removing comments expressed by photos only reduced them to 1415. Removing comments that contain mentions to friends with no other words reduced them to 1296. Then, after removing non-Arabic comments, they were reduced to 1261.

The final number of tweets downloaded was 1787 tweets. After removing the tweets that contain URLs only or advertising links or some who put links to watch the movie only, they were reduced to 1069. Some were links to certain scenes or related videos on Youtube. After removing unrelated tweets as the search on twitter was just by the movies' names which can imply other meanings, they were reduced to 862. Removing non-Arabic tweets reduced them to 781.

The number of reviews downloaded from the review sites was 32. The reviews needed only two steps of preparation as shown in Fig. 1.

After the preprocessing, cleaning and filtering of the data, they must be annotated to be fed to the supervised classifiers. The first Experiment shows the method of annotation and the number of positive and negative data.

B. Text processing and Classification

After annotation, we have applied removing stopwords text processing technique on the three corpora with different alternatives of stopwords list which are:

-A general MSA list: this list contains a combination of three published lists. The first one is a project that generated stopwords with all possible suffixes and prefixes. The other two were published in

(https://code.google.com/p/stop-words/source/browse/trunk/stop-words/stop-words/stop-words-arabic.txt) and (http://www.ranks.nl/stopwords/arabic) respectively.
-A generated corpus-based list: this list is generated from the most frequent words from the corpora regardless of their nature.
-A generated Egyptian-dialect list: this list is generated from the most frequent words in the corpora that can be a stopword in addition to the Egyptian dialect stopwords that appeared in the corpora.
-A combination of the Egyptian dialect list and the MSA list.

Text classification is applied on the three corpora using two feature selection techniques and two classifiers as shown in Fig. 2. We have used two well known supervised learning classifiers; Naïve Bayes (NB) in Ref. [22] and Decision tree (DT) in Ref. [23] in testing. There are many other kinds of supervised classifiers in the literature as in Ref. [24]. The two chosen classifiers represent two different families of classifiers. NB is one of the probabilistic classifiers which are the simplest and most commonly used classifier. DT on the other hand is a hierarchical decomposition of data space and doesn’t depend on calculating probability. The test used two different feature selection (FS) techniques. These are; unigrams which depend on word presence; and bigrams as in Ref. [2].

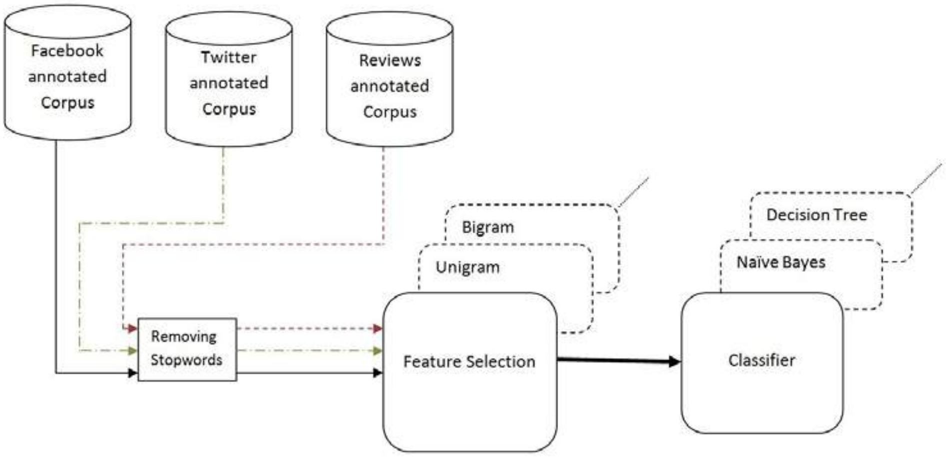


Figure 2: Text Processing and Text classification of the prepared corpora

3 STOPWORD LIST GENERATION

Stopwords are common words that generally do not contribute to the meaning of a sentence, specifically for the purposes of information retrieval and natural language processing. The common English words that don’t affect the meaning of a sentence are like “a”, “the”, “of”…. Removing stopwords will reduce the corpus size without losing important information. In some corpora, specific words could not contribute to the meaning like the word “movie” in a movie reviews corpus but means something in news corpus. This word could be considered a stopword when analyzing the movie reviews corpus.

The common strategy for determining a stopword list is to calculate the frequency of appearance of each word in the document collection then to take the most frequent words. The selected terms are often hand-filtered for their semantic content relative to the domain of the documents being indexed, and marked as a stopword list.

The English stopword list is general and contains 127 words like (all, just, being…). In order to generate the stopword list for Arabic which is a very rich lexical language; we have done this through many steps. First, we should specify some general conditions for the word to be a stopword:
-They give no meaning if they are used alone.
-They appear frequently in the text.
-They are general words and not used specifically in a certain field.

The methodology of generating the stopword lists are shown in Fig. 3. The methodology consists of three phases as illustrated in the following subsections.

A. Calculating words frequency

The three corpora are tokenized to words. This phase was done totally automatic using python code and the nltk 2.0 toolkit. The results are not totally meaningful as the tokenization could consider the “comma” as a word if it is not correctly used. There is some manual filtering after tokenization.

The reviews corpus give 3781 unique words, the Facebook corpus give 1451 unique words, and the Twitter corpus give 1160 unique words. This shows that despite the number of reviews are much less than the OSN corpora but they are lexically rich. After combining them together and removing the duplicates, the list of all words are 4818 words. Then we have calculated the frequency of occurrence of each word from the list of all words in the three corpora combined together.

B. The validity of words to be a stopword

To generate the corpus based list, we have taken the most frequent 200 words. These words are not all general and they are domain specific like the words “المشاهد” or “الفيلم” which means (the spectator, the movie) respectively. This list contains words in MSA and Egyptian dialect as well.

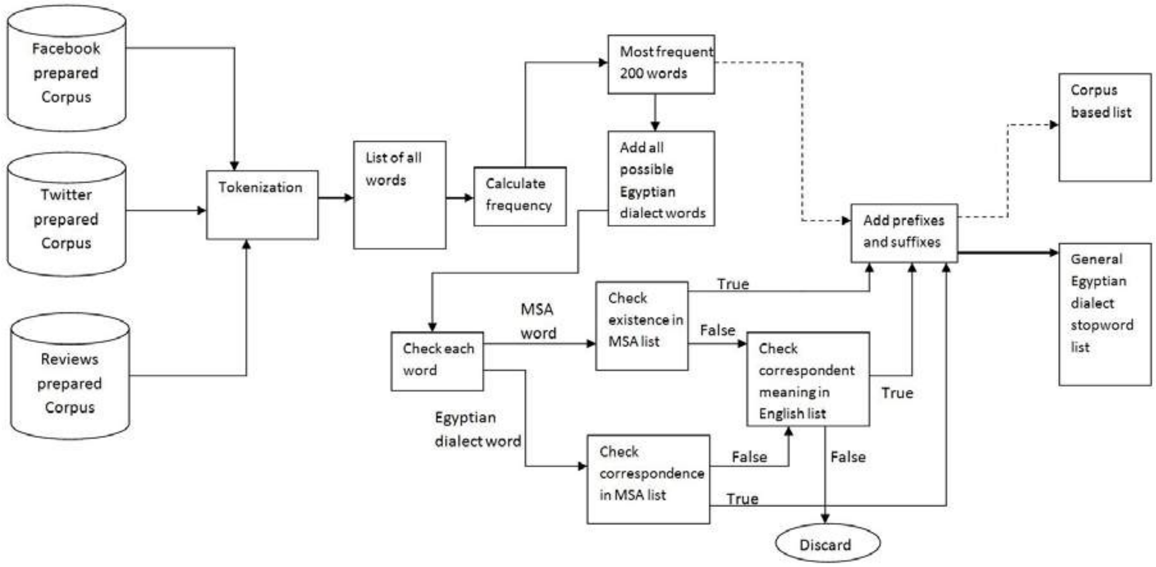


Figure 3: A methodology of generating the stopword lists

Diacritics could change the meaning of a word i.e. the word “المشاهد” could mean (the spectator or the scenes). The difference could be told through the meaning of the sentence. The OSN users use simple language without diacritics. Since the word is in the context of the corpora, it is more likely to appear frequently expressing both meanings. The problem will occur if a word appeared as a frequent word but outside the context of the corpora. This case didn’t happen here.

To generate a general list of Egyptian dialect stopwords, we have taken the most frequent 200 words and remove the semantically recognized words which are likely to be nouns and verbs. Then, to generate a general list of Egyptian dialect, we have added every word in the corpora in Egyptian dialect to the most frequent words that are semantically meaningless. To validate if the word is a stopword or not; if the word is a MSA word we check its existence in the MSA stopword lists. If it doesn’t exist, we check its corresponding meaning in the English stopword list. If the word is in Egyptian dialect, we see its correspondence in the MSA list and if doesn’t exist we check its correspondent meaning in the English stopword list. For example the word “بس”, its correspondence in MSA is “فقط” and it has a corresponding meaning in the English stopword list too which is “only”. On the contrary, the word “لازم” has no correspondence in the MSA list which should be “لابد” but it has a correspondent meaning in the

English list which is the word “should”. Therefore, it is considered a stopword. The final list of valid unique words contains 100 words. This phase was done in a semi automatic way that includes manual check.

C. Adding possible prefixes to the words

Arabic is a very rich lexical language which has a large number of prefixes and suffixes that could be added to a word to change its meaning. For example the prefix “ال” which means “the” change the word from indefinite to definite. The suffix “هم” gives the meaning of pronoun “them”. We have added some frequent used prefixes to the words generated in both lists which are (ال، و ، ب، ف، ل). If necessary we give pronoun suffixes which are (نا، ى، ه، ها، هم). We have added these suffixes to possession words in Egyptian dialect like the word “بتاعى” which means (mine).

There is also some letters are written in different forms so we write any word that contains these letters’ possible forms such as (ي، ى), (ه، ة), (ا، أ، إ). The last one is according to the word itself. The lists are manually revised for improper words or meaningless words.

After adding the prefixes and suffixes, the final corpus-based list contains 1061 words and can be found in (http://goo.gl/JW0jKP). The final general Egyptian dialect list contains 620 words and can be found in (http://goo.gl/263J5L).

4 EXPERIMENTAL SETUP AND RESULTS

We used a HP pavilion desktop computer of model: p6714me-m. The processor is Intel(R) core (TM) i5-2300 CPU @ 2.80 GHZ; RAM is 4GB; and 64-bit operating system. We have calculated the training time using a build-in function written with python code which calculates the processing time in terms of seconds. These tests were all performed using the Natural Language Toolkit (nltk 2.0) which is implemented inside python 3.1 as in Ref. [25].

A. Data Annotation

The reviews from the review site were previously rated from the site. They were given a degree from 1 to 10. The ratings bigger than 5 are considered positive and less than 5 are considered negative. The ratings equal to 5 are neutral. We have annotated the reviews according to the site rating.

For the OSN data, we have manually annotated the corpora. The manual annotation was more reliable as the human analyzing of data is better than the machine so far. Table I shows the number of positive, negative and neutral reviews, comments, and tweets resulted from annotation.

Table I

NUMBER OF POSITIVE, NEGATIVE AND NEUTRAL REVIEWS, COMMENTS, AND TWEETS FROM REVIEW SITE, FACEBOOK AND TWITTER

	Reviews	**Facebook**	**Twitter**
No. of positive	25	369	160
No. of negative	6	33	77
No. of neutral	1	859	544

B. Classifiers Preparation

We trained Naive Bayes, and Decision Tree classifiers. The classifiers were conducted with the nltk 2.0 toolkit. There are some parameters passed in to the DT classifier can be tweaked to improve accuracy or decrease training time as in Ref. [25].

The parameters are:
-*Entropy cutoff*: used during the tree refinement process. If the entropy of the probability distribution of label choices in the tree is greater than the entropy_cutoff, then the tree is refined further. But if the entropy is lower than the entropy_cutoff, then tree refinement is halted. Entropy is the uncertainty of the outcome. As entropy approaches

1.0, uncertainty increases and vice versa. Higher values of entropy_cutoff will decrease both accuracy and training time. It was set to ‘0.8’.
-*Depth cutoff*: used during refinement to control the depth of the tree. The final decision tree will never be deeper than the depth_cutoff. Decreasing the depth_cutoff will decrease the training time and most likely decrease the accuracy as well. It was set to ‘5’.
-*Support cutoff*: controls how many labeled feature sets are required to refine the tree. When the number of labeled feature sets is less than or equal to support_cutoff, refinement stops, at least for that section of the tree. Support_cutoff specifies the minimum number of instances that are required to make a decision about a feature. It was set to ‘30’.

C. Feature Selection

There are two Features selection (FS) techniques used in the test:
-*Unigram*: treats the documents as group of words (Bag of Words (BOWs)) which constructs a word presence feature set from all the words of an instance.
-*Bigram*: is the same as unigram but finds pair of words.

D. Results

We have made many experiments to test the effect of removing stopwords from different lists with the combination of two FS techniques and two classifiers with the three different corpora. We have made the tests on splitting 75% of the total number of the data in each corpus for training and 25% for testing data.

The standard Accuracy was used to evaluate the performance for each test. The accuracy is defined as: the ratio of number of correctly classified reviews, comment, and tweets to the total number of data.

Table II contains the results of the various tests we have made. The accuracy of the reviews is relatively high as the number of reviews is small and the data is highly unbalanced. The accuracy decreases when using corpus-based list on the lexically rich reviews and the general lists including Egyptian dialects give better results than the others. The timing is not changed a lot but in general it decreases when removing stopwords. The DT gives better results with Facebook data than NB as it is extremely unbalanced.

Table II

ACCURACY AND TRAINING TIME OF SENTIMENT ANALYSIS ON REVIEWS, FACEBOOK AND TWITTER CORPORA USING NB AND DT CLASSIFIERS WITH UNIGRAM AND BIGRAM AS FS AFTER REMOVING STOPWORDS FROM DIFFERENT LISTS

Classifier	Feature selection	Removing Stopwords	Accuracy			Time (sec)		
			Reviews	Facebook	Twitter	Reviews	Facebook	Twitter
Naïve Bayes	Unigram	Without	100%	48.04%	78.33%	0.050	0.018	0.015
		Other lists	100%	47.06%	68.33%	0.042	0.017	0.015
		Corpus-based	44.44%	53.92%	68.33%	0.038	0.014	0.015
		General	100%	50%	68.33%	0.043	0.016	0.014
		All lists	100%	48.03%	70%	0.041	0.017	0.014
	Bigram	Without	77.77%	40.20%	80%	0.117	0.060	0.052
		Other lists	88.88%	29.41%	68.33%	0.090	0.063	0.053
		Corpus-based	22.22%	43.13%	65%	0.098	0.047	0.035
		General	88.88%	31.37%	63.33%	0.094	0.059	0.053
		All lists	100%	29.41%	65%	0.097	0.061	0.052
Decision Tree	Unigram	Without	100%	90.20%	70%	0.217	0.620	0.589
		Other lists	100%	91.17%	68.33%	0.195	0.594	0.560
		Corpus-based	77.77%	90.20%	70%	0.187	0.503	1.170
		General	100%	90.20%	68.33%	0.196	0.559	0.530
		All lists	100%	91.17%	68.33%	0.192	0.560	0.521
	Bigram	Without	100%	90.20%	73.33%	0.510	1.849	2.471

		Other lists	100%	90.20%	68.33%	0.436	1.920	1.737
		Corpus-based	77.77%	90.20%	68.33%	0.414	1.561	3.011
		General	100%	90.20%	68.33%	0.416	1.787	0.530
		All lists	100%	90.20%	68.33%	0.410	1.795	1.647

The following figures show the accuracy and logarithmic graphs of training time for each corpus. The logarithmic graphs are used to clarify the difference in timing.

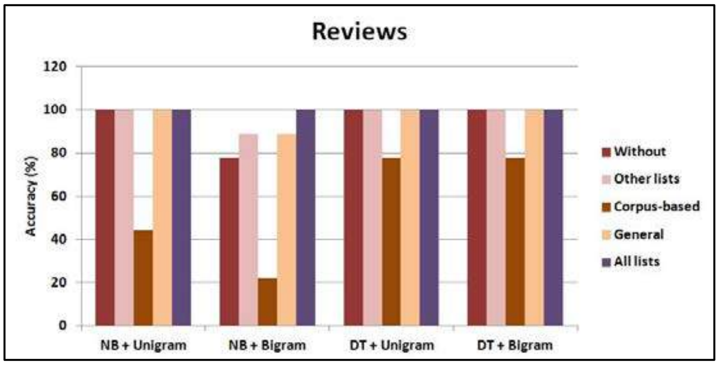


Figure 4: Classification accuracy of Reviews corpus

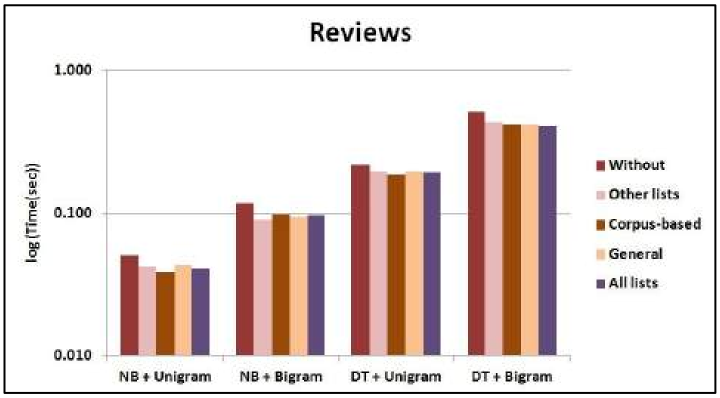


Figure 5: Classification training time of Reviews corpus

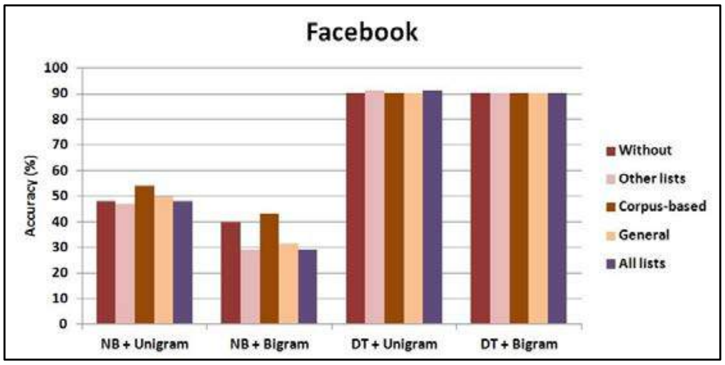


Figure 6: Classification accuracy of Facebook corpus

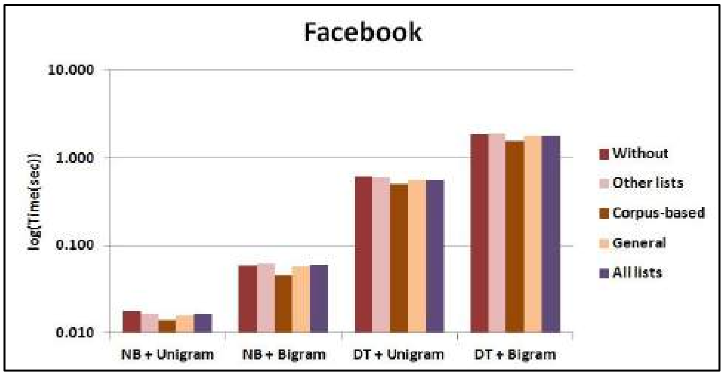


Figure 7: Classification training time of Facebook corpus

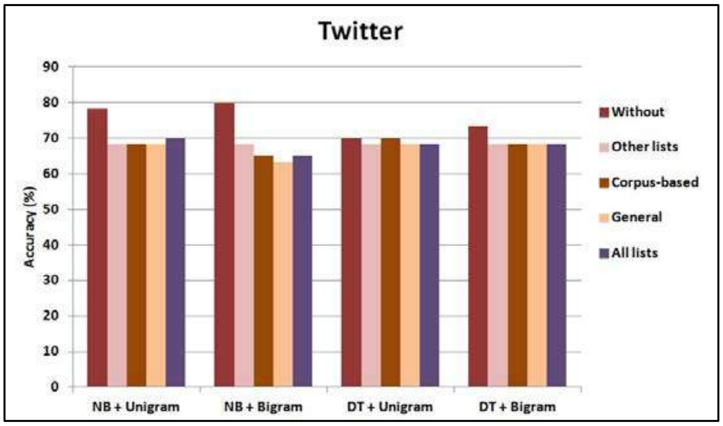


Figure 8: Classification accuracy of Twitter corpus

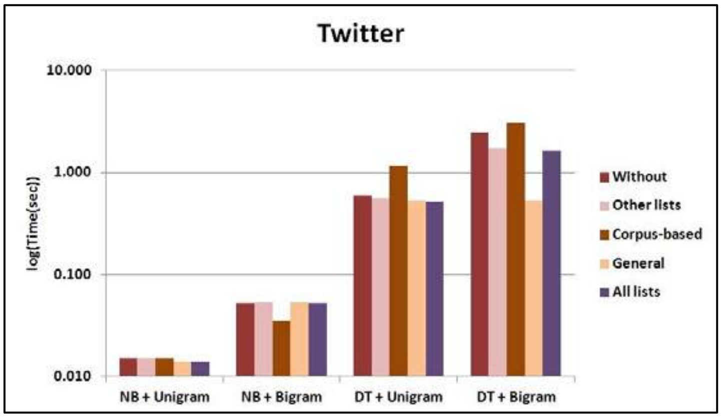


Figure 9: Classification training time of Twitter corpus

5 DISCUSSION

A. Corpora Analysis

The number of neutral reviews from the review site represents 3% of the whole data. This is not a big number. We believe that people who write whole reviews on reviews sites are mainly having a complete opinion about the movie and they want to show it. They don't lean to be neutral. The number of positive reviews represents 78% of the whole data while the number of negative reviews represents 18% of the entire data. The data are obviously unbalanced since the movies were successful in this season, not many users' reviews were negative.

The number of neutral comments on Facebook represents 68% of the whole data. These are not neutral opinions on the movie. People who write in OSN are not neutral at all. The neutral comments are mainly objective sentences that don't contain any sentiments. Many comments were just debates between users. Some were expressing their personal feelings and some were using adjectives without specifying on whom or what. The number of positive comments represents 29% of the entire data and the number of negative comments represents 2% of the whole data which is an extremely small percentage. This is also an unbalanced data. We believe that people who access a movie page they do like it.

The number of neutral tweets represents 69% of the whole data. These are not neutral opinions on the movie too. The neutral tweets are mainly objective sentences that don't contain any sentiments. Many of the tweets were repetition of a dialogue from a movie without expressing any feelings. Others were tweets expressing the users' personal feelings like feeling excited to see the movie. The number of positive comments represents 20% of the entire data and the number of negative comments represents 9% of the whole data which is a small percentage. We believe that people who mention the movie in their tweets; do like it.

Using abbreviations and smiley faces in OSN are very frequent. There are some abbreviations were used also in Reviews. The meaning of these abbreviations and smiley faces were found from different sources on the web (Yahoo answers, Facebook emoticons sites) and translated to Arabic. For the Arabic abbreviations they were manually translated. Table III contains sample of Abbreviations and smiley faces found in the three corpora.

Table III

SAMPLE OF ABBREVIATIONS AND SMILEY FACES FOUND IN FACEBOOK, TWITTER, AND REVIEWS

Abbreviations	**Facebook**	**Twitter**	**IMDB**
ههه → ضحك	Found	Found	
:D → ابتسامة كبيرة	Found	Found	Found
3> → قلب		Found	
^_^ → مبسوط	Found	Found	

B. Specializations of Arabic Language

Words with the same meaning could be written in different correct ways like the words "هنروح، حنروح". They both give the future tense of the verb "نروح" which means "we will go". As we can notice three words in English are just written in one word in Arabic and give the same meaning. The pronouns in English are expressed in Arabic by adding a prefix letter that modify the verb especially when it is used in the middle of the sentence like "اروح، نروح" which means (I go, we go) respectively. Some prepositions and causal words are expressed in Arabic with one letter like the words "انى، لانى" which means (I am, because I am) respectively.

The many forms that the Arabic words could take are very common characteristics of MSA which make the dealing with the language is complicated. For DA, it is a tragedy. We have a special dialect for each Arab country and different dialects in the same country. For Egyptian dialect, there are many words that have no resemblance in MSA like the word "مفيش" which means (there is not). It has only a correspondent in MSA which is "لا يوجد" which are complete different words. In the OSN corpora some other dialects appear like the Moroccan word "بزاف" which means (too much) and the Syrian word "مليح" which means (good). The number of other dialects in Facebook corpus represents 1% of the whole corpus which is very small percentage. The number of other dialects in Twitter corpus represents 0.5% of the whole corpus which is extremely small percentage. There were no other dialects in reviews corpus. They used a mix between MSA words and Egyptian dialect words as they are user reviews not formal reviews from critics.

The other phenomenon of Arab users is using the Franco-arab. This means that people use English letters for writing Arabic words like the word “de7k” which stands for “ضحك” which means (laugh). The number of Franco-arab comments in Facebook corpus represents 18% of the whole corpus which is not a big percentage. The number of Franco-arab tweets in Twitter corpus represents 3% of the whole corpus which is a small percentage. However, we have to unify the language used for the classifier to perform well. These are not even English words that have meanings so; they must be rewritten in Arabic letter. We have used the website (www.yamli.com). They give variations for each word that have to be chosen from. Sometimes the users don’t even write correct words in Franco-arab. In this case the site translates the letters only which give funny Arabic words. This transformation was manually revised.

C. Results Analysis

Fig. 4 shows that removing stopwords from reviews didn’t change the accuracy when using general lists but decrease the accuracy when using the corpus-based list. It also shows that that unigrams are better FS than bigrams with NB. The training time decreases after removing stopwords and the training time of DT is higher than NB as shown in Fig. 5.

Fig. 6 shows that the accuracy of DT is much bigger than NB because the data is extremely unbalanced. There is no significant difference between unigrams and bigrams in DT but unigrams is better than bigrams with NB. Using corpus-based list increase the accuracy than using the general lists with NB but the accuracies are almost the same with DT. The training time decreases after removing stopwords and the training time of DT is higher than NB as shown in Fig. 7.

Fig. 8 shows that NB and DT give very close accuracies as the data is not very unbalanced. Unigrams are better than bigrams in case of NB. Using different lists didn’t change the accuracy much but the general lists give good performance too. The training time decreases after removing stopwords and the training time of DT is higher than NB as shown in Fig. 9.

In case of lexically rich corpus like the reviews, using the corpus-based list decrease the accuracy of classification which is similar to what Ref. [9] has found. But in case of OSN where they were not lexically rich the three lists wasn’t varying the accuracies much but still the general lists containing Egyptian dialect stopwords give better results than using MSA stopwords only. The difference in performance between Facebook and Twitter data is due to the degree of imbalance. The nature of the data is the same but Facebook corpus is much more unbalanced than Twitter corpus.

Decision Tree is a hierarchical decomposition of data space and doesn’t depend on calculating probability but Naïve Bayes depends on calculating probability for the whole data. Although NB usually gives higher accuracy than DT, but this was not the case when testing these corpora. This is due to the unbalance of the data as the positive class in these cases where much bigger than the negative class. NB calculates the probability on the whole data but DT is more specifically build hierarchy decomposition of data. That is why DT is better for unbalance data as it is more specific than NB. But still DT has longer processing time than NB because it builds the hierarchical decomposition on the whole data but the difference in time is not big as the data size was not so big. In NB tests, the accuracy is better when using unigram which is similar to what Ref. [2] has found. In DT tests, unigram and bigrams give nearly similar results.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a methodology for generating a stopword list from online social network (OSN) corpora. The methodology consists of three phases: calculating the words’ frequency of occurrence, check the validity of a word to be a stopword, and adding possible prefixes and suffixes to the words generated. We have generated a stopword list of Egyptian dialect and a corpus-based list to be used with the OSN corpora. We compared them with other lists. The lists used in the comparison were: previously generated lists of MSA, the corpus-based generated list, the general generated list of Egyptian dialect, and a combination of the Egyptian dialect list with the MSA list.

We have also proposed a methodology to prepare corpora in Arabic language from OSN and review site for Sentiment Analysis (SA) task. It includes the translation of English words that appear in text and the transformation of Franco-arab to Arabic words. The text classification was performed using Naïve Bayes and Decision Tree classifiers and two feature selection approaches, unigram and bigram.

We have selected the movie reviews topic to download data about movies from three different sources (Review site, Facebook, and Twitter). The data are extremely unbalanced as the movies were successful and most of the OSN users like it and the reviewers as well. The data contain many spams like advertising URLs, debates, and using of abbreviations and smiley faces. It needed many preprocessing and cleaning steps to be prepared for classification.

Applying removing stopwords with multiple lists show that the corpus-based list negatively affects the accuracy of classification incase of reviews. Reviews are more lexically rich than OSN corpora. It also shows that the general lists containing the Egyptian dialects words give better performance than using lists of MSA stopwords only. The results of Decision tree classifier are better than Naïve Bayes classifier for these kinds of corpora. Using unigrams give better results than bigrams.

In the future we plan to try more text processing techniques on Arabic OSN data like POS tagging and try to fulfill the gap of using the Arabic dialect in the OSN data as all resources are designed for MSA. We could tackle other dialects other than Egyptian.

References

[1] W. Medhat, A. Hassan and H. Korashy, *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal, 2014.

[2] B. Pang, L. Lee, and S. Vaithyanathan, *Thumbs up?: sentiment classification using machine learning techniques*, in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), 2002.

[3] X. Bai, Padman, R. & Airoldi, *Sentiment Extraction from Unstructured Text Using Tabu Search -Enhanced Markov Blanket*, Technical Report CMU-ISRI-04-127, 2004.

[4] M. Gamon, *Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis*, in Proceedings of COLING'04, Geneva, Switzerland, 2004, pp. 841-847.

[5] K. Versteegh, C. Versteegh, *The Arabic Language*, Columbia University Press, 1997.

[6] N. Habash, *Introduction to Arabic natural language processing*, Synthesis Lectures on Human Language Technologies, vol. 3, 2010.

[7] M. Korayem, D. Crandall, and M. Abdul-Mageed, *Subjectivity and Sentiment Analysis of Arabic: A Survey*, AMLTA 2012, CCIS 322, pp. 128–139, 2012.

[8] R. Al-Shalabi, G. Kanaan, Jihad M. Jaam, A. Hasnah, and E. Hilat, *Stop-word removal algorithm for Arabic language*, in 1st International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, 2004, pp. 545-550.

[9] I. Abu El-Khair, *Effects of stop words elimination for Arabic information retrieval: a comparative study*, International Journal of Computing & Information Sciences, vol. 4, pp. 119-133, 2006.

[10] A. Alajmi, E. M. Saad, and R. R. Darwish, *Toward an ARABIC Stop-Words List Generation*, International Journal of Computer Applications, vol. 46, 2012.

[11] L. Hong and Brian D. Davison, *Empirical Study of Topic Modeling in Twitter*, presented at the 1st Workshop on Social Media Analytics (SOMA ’10), Washington, DC, USA., 2010.

[12] A. Bifet and E. Frank, *Sentiment Knowledge Discovery in Twitter Streaming Data*, 2010.

[13] A. Pak and P. Paroubek, *Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives*, in Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, Uppsala, Sweden, 2010, pp. 436–439.

[14] X. Liu, S. Zhang, F. Wei and M. Zhou, *Recognizing Named Entities in Tweets*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 2011, pp. 359–367.
[15] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, *Target-dependent Twitter Sentiment Classification*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 2011, pp. 151–160.
[16] Maynard, Diana, and A. Funk, *Automatic detection of political opinions in tweets*, The semantic web: ESWC 2011 workshops. Springer Berlin Heidelberg, 2012.
[17] Meador, Catie, and J. Gluck, *Analyzing the Relationship Between Tweets, Box-Office Performance and Stocks*, Methods (2009).
[18] A. Go, R. Bhayani, and L. Huang, *Exploiting the unique characteristics of tweets for sentiment analysis*, Technical report, Technical Report, Stanford University, 2010.
[19] A. Go, R. Bhayani, and L. Huang, *Twitter sentiment classification using distant supervision*, CS224N Project Report, Stanford (2009): 1-12.
[20] M. Cohen, P. Damiani, S. Durandeu, R. Navas, H. Merlino And E. Fernández, *Sentiment Analysis in Microblogging: A Practical Implementation*, presented at the CACIC 2011 - XVII CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN, Buenos Aires, Argentina, 2011.
[21] W. Medhat, A. Hassan and H. Korashy, *A Framework of preparing corpora from Social Network sites for Sentiment Analysis*, accepted in the International Conference on Information Society (i-Society 2014), London, UK, 2014.
[22] A. McCallum and K. Nigam, *A Comparison of Event Models for Naive Bayes Text Classification*, presented at the AAAI Workshop on Learning for Text Categorization, 1998.
[23] J. R. Quinlan, *Induction of Decision Trees*, Machine Learning, vol. 1, pp. 81–106, 1986.
[24] C. Aggarwal and C. Zhai, *Mining Text Data*, Springer New York Dordrecht Heidelberg London: © Springer Science+Business Media, LLC 2012, 2012, p.^pp. Pages.
[25] J. Perkins, *Python Text Processing with NLTK 2.0 Cookbook*, Birmingham: Packt Publishing Ltd., 2010.

BIOGRAPHY

Walaa Medhat, is an Engineering Lecturer in School of Electronic Engineering, Canadian International College, Cairo campus of CBU. She got her M.Sc. and B.Sc. from Computers and Systems Engineering Department, Ain Shams University in 2008, 2002 respectively. Fields of interest: Text mining, Natural Language Processing, Data mining, Software engineering, Programming Languages and Artificial Intelligence.

Ahmed Hassan Yousef is an associate professor in the Computers and Systems Engineering Department, Ain Shams University since 2009. He is the vice director of the Knowledge and Electronic Service Center (EKSC), Supreme Council of Universities, Egypt. He got his Ph.D., M.Sc. and B.Sc. from Ain Shams University in 2004, 2000, 1995 respectively. He works also as the secretary of the IEEE, Egypt section since 2012. His research interests include Data Mining, Software Engineering, Programming Languages, Artificial Intelligence and Automatic Control.

Hoda Korashy, is a Prof. at Department of Computers & Systems, Faculty of Engineering, Ain Shams University, Cairo, Egypt. Major interests are in database systems, data mining, web mining, semantic web and intelligent systems.

إعداد المكانز و إنشاء قائمة الكلمات الهامشية لبيانات باللغة العربية في شبكات التواصل الاجتماعي

***ولاء مدحت، **احمد حسن يوسف، **هدى قرشي**
* قسم الالكترونيات، المعهد الكندي العالي لتكنولوجيا الهندسة و الإدارة
** قسم الحاسبات و النظم، كلية الهندسة، جامعة عين شمس

تقترح هذه الورقة منهجية لإعداد المكانز باللغة العربية من مواقع شبكات التواصل الاجتماعي ومواقع المشاركات واستخدامها في عملية تحليل المشاعر. كما تقترح هذه الورقة أيضا منهجية لإنشاء قائمة للكلمات الهامشية من المكانز المعدة. تهدف هذه الورقة لتدارس تأثير إزالة الكلمات الهامشية على عملية تحليل المشاعر. تكمن المشكلة في أن قوائم الكلمات الهامشية المعدة مسبقا كانت على اللغة العربية الفصحى و هي ليست اللغة المعتادة المستخدمة في شبكات التواصل الاجتماعي. لقد قمنا بإنشاء قائمة الكلمات الهامشية باللهجة المصرية و أخرى معتمدة على المكنز لاستخدامها مع مكانز شبكات التواصل الاجتماعي. و قمنا بمقارنة كفاءة تصنيف النص عند استخدام القوائم المعدة مسبقا على اللغة العربية و بعد إدماج القائمة باللهجة المصرية مع القائمة باللغة العربية الفصحى. طبق تصنيف النص باستخدام أسلوب تصنيف بايز الساذج وشجرة القرارات كما يتم استخدام كلمة واحدة أو كلمتين كأسلوب للتعرف على مميزات الجملة. بينت التجارب أن قوائم الكلمات الهامشية التي تحتوى على اللهجة المصرية تعطى أداء أفضل من استخدام قوائم باللغة العربية الفصحى فقط.

معجم تكراري للغة العربية المعاصرة

"معالجة لغوية حاسوبية"

محمد مجدي لبيب حامد

قسم علم اللغة والدراسات السامية والشرقية بكلية دار العلوم

Medo_002@yahoo.com

ملخص

يهدفالبحثُ إلى وضع نموذج للمعجم التكراري للغة العربية المعاصرة يكون نواةً لمعجم يخدم أربابَ العربيـة من أبنائها والواردين عليها، عن طريق الاعتماد على معلومـات التكرار التـي تساعد في ترتيب مداخلـه وموادِه حسب عدد تكرارها وشيوعها في المادة المقدمة، مما يتيح لمستخدم اللغة من معلم ومتعلم الإلمامَ بأشـهر كلمـات اللغة التي تمثل ثلثي اللغة في وقت بسيط وبسرعة.

ويتخذ البحثُ من لسانيات المدونة وسيلةً للوصل إلى المنتج النـهائي المتمثل فـي المعجم المنشود، عن طريق جمع عينة عشوائية تمثل المستوى اللغوي المدروس وتنسيقها وتـهيئتها داخل مدونة لغوية تتيح التعامل آليًّا مـع المادة المجموعة بالحذف والتعديل والإضافة؛ مما ييسّر آليات البحث ويعين على إنجازه بصورة دقيقة.

كما يساهم علمُ الإحصاء في التأكيد على مدى دقة مخرجات المعجم وأدوات التحليل عن طريق إجراء معادلات الإحصاء للمخرجات وحساب عدد مرات تكرار مداخل المعجم، وترتيبها من الأكثر إلى الأقل شيوعًا.

كل ذلك في سبيل الإسهام في العمليـة التعليمية بمعجم يحـاول تسهيلها وإفادة طرفيها المعلم والمـتعلم بصورة سهلة وفي وقت يسير.

الكلمات المفتاحية

المعجم التكراري Frequency Dictionary – معجم تعليمي Educational Dictionary – المدونة اللغوية linguistic corpus – الصناعة المعجمية Lexicography

Abstract:

The research aims to develop a model of a frequency dictionary for Arabic language contemporary to be the nucleus of a lexicon serve native- speakers of Arabic and contained them, counting on frequency information, that helps arranging entrances lexicon and his headwords depending on the number of frequency and commonness in the material presented, by allow to language teacher and learner at knowledge the most popular words of the language, which represents two-thirds of the language quickly and simply.

the research use the corpus linguistics as mean to reaching to the final product represented in the lexicon desired, by collecting a random sample representing the linguistic level studied, coordination and configured within the language corpus allows to deal automatically with material which collected by deletion, modification and addition ; thus support the search mechanisms and accomplishing it accurately.

Statistics also contributes in the Confirmation on a accuracy of the outputs of the lexicon and analysis tools by conducting statistical equations and calculating the number of frequency of dictionary entries, arranged from most to least common**.**

All this in aims to contribute to the educational process by using a lexicon facilitating and helping both teacher and the learner quickly as well as easily.

أولا: مُقدِّمة

هذا البحث مستخلص من أطروحة معدة لنيل درجة الماجستير بقسم علم اللغة والدراسات السامية والشرقية بكلية دار العلوم – جامعة القاهرة، تحمل نفس العنوان "معجم تكراري للغة العربية المعاصرة.. دراسة لغوية حاسوبية"، تحت إشراف أ.د/ إبراهيم الدسوقي، ود.المعتز بالله السعيد.

شغلَ مضمارُ تعليم اللغات فِكْرَ اللغويين والتربويين حتى علماء النفس، فقد فكَّر عالمُ النفس الأمريكي إدوارد ثورنديك Edward Thorndike (1874–1949) في وسيلة تعليميَّة ناجعة تساعد المعلِّمين والمتعلمين في تعلُّم اللغة الإنجليزية بسهولة ويسر، فقام بوضع قائمة مفردات للغة الإنجليزية بعنوان قائمة المفردات للمعلمين "Teacher’s Wordbook" عام 1921، جمعَ فيها أشهرَ مفردات اللغة الإنجليزية معتمدًا في جمع مادته على مدونة لغوية مكونة من 41 مصدرًا مختلفًا للغة الإنجليزية، ثم عمل على تطوير قائمته تلك بمساعدة إرفنج لورج Irving Lorge، وقاما بإخراج الكتاب بعنوان The Teacher’s Wordbook of 30000 words عام 1944، وقد احتوت مدونة المعجم على ثماني عشرة مليون كلمة. ومنذ ذلك الوقت أخذت فكرةُ صناعة المعاجم التعليمية طريقها إلى اهتمام اللغويين والعاملين في مجال التربية.

وقد كانت محاولةُ ثورنديك نواةَ أعمال كثيرة انتهجت نفسَ الطريق، تهدف إلى المساهمة في تيسير العملية التعليمية على المعلمين والمتعلمين، خاصة المتعلمين المبتدئين.

1. المعجمُ التكراريُّ

المعجم التكراريFrequency Dictionary نوعٌ من المعاجم اللفظية المتخصصة، تُرَتَّب فيه الأبوابُ والفصولُ المعجمية بحسب شيوعها، متدرجةً من الأكثر شيوعًا إلى الأقل شيوعًا، ويعتمد المعجمُ في تحديد نسبة شيوع المواد المعجمية على مدونة لغويةlinguistic corpus تعكس واقعَ اللغة، وتمثل المستوى اللغوي الذي يُعنَى به المعجم[1].

تقوم فكرةُ المعجم التكراري على حصر ألفاظ اللغة بما يمثل الواقعَ الحقيقي للغة، وترتيبِ تلك الألفاظ حسب نسبة تكرارها وترددها، في مجموعاتٍ تدرَّج في الأهمية؛ حيث تضم المجموعةُ الأولى الألفاظَ الأكثر شيوعًا وانتشارًا، ثم تأتي المجوعةُ الثانية التي تضم الألفاظَ والكلمات الأقل في الشيوع، ثم الأقل فالأقل، إلى أن نصل إلى المجموعة الأخيرة التي تضمُّ الألفاظَ النادرَ شيوعُها واستخدامُها؛ فكان مقياسُ ورود الكلمات في المعجم معتمِدًا على عدد تكرارها وترددها في الاستخدام "تم تحديد مقياس لورود الكلمات عن طريق عدد معطى لتكرارها، ويُجِيب هذا الكم عن السؤال: كم عدد المصادر المختلفة التي تستعمل هذه الكلمة؟ أو –بصيغة أخرى– كيف تستعملها على نحو واسع؟ ويجيب عدد ورودها على كم مرة استعملت الكلمة؟"[2].

فمادةُ المعجم تُجْمَعُ من مصادر مختلفة ومجالات متعددة في اللغة؛ لكي تغطي جميعَ استخدامات الكلمة الواحدة واحتمالات وقوعها وتكرارها في اللغة بما يعكس واقعَ اللغة، ثم تُصنَّف تلك الكلمات حسب عدد مرات تكرارها الذي يعد حدَّ ائتمان يُعْتَمد عليه في تصنيف الكلمات من حيث تكرارها، ومن ثمّ تُقسَّم الكلمات المجموعة على مجموعات تحتوي المجموعةُ الأولى على المفردات الأكثر انتشارًا في اللغة والأكثر أهمية للطالب المتعلم، ثم مجموعة الكلمات الأقل فالأقل. والمحكُّ المعتمَدُ لتصنيف الكلمات داخل المجموعات هو معلوماتُ التكرارFrequency informationوهي تلك المعلوماتُ الإحصائيةُ التي يمكن من خلالها معرفةُ النسبة التقريبية لشيوع الكلمة (في اللغات اللصقية)، أو الجذر اللغوي في (اللغات الاشتقاقية)؛ والغرضُ من هذه المعلومات تعليميٌّ صرف، إذ إن معرفةَ نسبة شيوعالكلمة أو الجذر اللغوي سيوفر على متعلم اللغة الكثير من الوقت والجهد، إذ سيتجه أولا إلى معرفة أكثر الجذور اللغوية شيوعا، فالتي تليها، وهكذا[3].

وتلعبُ معلوماتُ التكرار دورًا محوريًّا في تعلُّمِ اللغة؛ إذ إنها المحكُّ الذي يَعتمدُ عليه المعجمُ التعليميُّ في نظم مادته وتوجيه المتعلم ومستخدم المعجم إلى أكثر كلمات اللغة المدروسة استخدامًا ودورانًا في اللغة، مما يوجِّه اهتمامَه إلى أشهر ألفاظ اللغة، وهو ما يوفر عليه الجهد والوقت في تعلُّم ألفاظ ومفردات قد تكون أُهملت، أو يكون استخدامُها قليلا بما لا يستدعي ورودها في الكلام اليومي. بيد أنها ليست مصدر المعلومات الوحيد الذي يتم بناءً عليه توجيه المتعلمين، بل إنها

[1](السعيد) المعتز بالله: "المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)"، ندوة القرآن الكريم والتقنيات المعاصرة (تقنية المعلومات)، المملكة العربية السعودية، مجمع الملك فهد لطباعة المصحف الشريف، المدينة المنورة، 2009 ص1.

[2]**EDWARD L. THORNDIKE "The Teacher's Word Book" NEW YORK, 1921, TEACHER'S COLLEGE, COLUMBIA UNIVERSITY.piii**.

[3](السعيد) المعتز بالله "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، أطروحة لنيل درجة الماجستير بكلية دار العلوم، القاهرة، 2008، ص3.

تُعَدّ نقطة انطلاقة جيدة للمتعلمين، كما أنها تعطي فائدة سريعة؛ لذلك فهي تمثل وسيلةً مهمةً لتعليم الكلمات[4]، ويتم استخلاصُها بصورة منهجية عن طريق المعالجة الحاسوبية والإحصائية لنصوص المدونة المستخدمة في صناعة المعجم المنشود[5].

2. أهمية المعاجم التكرارية

يضم المعجم التكراري ألفاظ اللغة مرتَّبًا إياها حسب الأكثر انتشارًا واستخدامًا، وعلى هذا تَكْمُن أهميةُ المعجم في أنه مرجعٌ للمعلمين والمتعلمين رافدٌ لاستخدامهم للكلمات وتوظيف ألفاظ اللغة ومعالجتها تربويًّا، فتمثلت أهميةُ المعجم التكراري في:

1) مساعدة المعلمين والمتعلمين في معرفة الكلمات الشائعة والمهمة؛ الأمر الذي يوفر الوقت والجهد، فقد تُمَثِّل أهمُّ كلمات اللغة 70% من اللغة الأم، فيهتم المتعلمُ بمعرفة تلك الكلمات، ومن ثَمَّ يأتي الجزءُ المتبقي بالممارسة "على أية حال، تبقى هذه القائمة أفضل من غيرها الموجود حتى الآن، وستساعد كل الأساتذة في معرفة الكلمات الشائعة والمهمة"[6].

2) يساعد المعلمَ في تقرير المعالجة التربوية للكلمة بتزويده بمدى أهمية تلك الكلمة في مجالها المعرفي، "يساعد هذا الكتاب –كتاب الكلمة– المعلمَ في التقرير السريع للمعالجة الملائمة بالإخبار مباشرة بمدى أهمية أي كلمة"[7].

3) يساعد المعلمين المبتدئين في مجال تعليم اللغات على الإلمام بأهم الكلمات وأكثرها انتشارًا في اللغة، وهو ما يجعلهم على مستوى المعلم الذي اكتسب تلك الخبرة من سنوات التعامل مع الطلاب والكتب، "الأهمية العملية الثانية لـ (كتاب الكلمة) هي تزويد المعلم الأقل خبرة بمعرفة أهمية الكلمات وصعوبتها على حد سواء بما يضاهي الخبرة التي اكتسبها المعلم الخبير أثناء سنوات التجريب مع الطلاب والكتب"[8].

4) كما يساعد العاملين في حقل تدريس العربية (لأبنائها أو لغير الناطقين بها) على اختيار الكلمات المناسبة في إعداد المواد التعليمية.

5) يحتوي على الكلمات في صورتها الأولية مع توضيح كيفية الاشتقاق منها بالطرق المختلفة.

[4] A Frequency Dictionary of Arabic: Core Vocabulary for Learners ،Op.cit. p(vii). Adapted

[5] انظر، نحو معجم تعليمي للغة العربية لغير الناطقين بغيرها: معالجة لغوية حاسوبية "مرجع سابق"، ص13.

[6] "The Teacher's Word Book", p(iv).Op.cit.

[7] "The Teacher's Word Book".P(iv) Op.cit.

[8] "The Teacher's Word Book".P(iv) Op.cit.

6) يساعد في وضع "المناهج الدراسية" حيث إن الكتب المدرسية تعمل على تطوير المفردات المقدمة للطلاب إلا أنها لا تنص على أيها الأكثر انتشارًا ودورانًا في اللغة، مما يجعل الطلاب يشتتون أوقاتهم بين الكتاب المدرسي والبحث في المعجم[9].

7) المعاجم التكرارية أداة مساعدة ومهمة لكل من معلمي اللغة العربية ومطوري المناهج؛ حيث تساعد المعلم في معالجة قصور حصيلة الطلاب اللغوية عن طريق إيجاد رؤية منهجية لتحديد قائمة من الكلمات التي تساعدهم في معالجة نقص مفرداتهم واستخدامها بطريقة أكثر فاعلية[10].

3. المعاجمُ التكرارية في اللغة العربية

لم تحظَ اللغةُ العربية بمحاولات منهجية لوضع معجم تكراري، فقد كانت أغلب تلك المحاولات عبارة عن وضع قائمة لأكثر الكلمات شيوعًا في اللغة المدروسة وهو ما عُرف بـ "قوائمُ الشيوع Common Words Lists" وهي: قوائمٌ تَضُمُّ أكثرَ الألفاظ شيوعًا في لغة ما، ومن ثَمَّ ترتيبُ تلك الألفاظ المجموعة على إحدى طُرُقٍ ثلاث، إما ترتيبٌ ألفبائيٌّ أو الترتيبُ حسب الأكثر ورودًا في الاستخدام اللغوي للألفاظ، أو الطريقةُ المعجميةُ التي تَعْتَمِدُ شكلَ الحرفِ لا لَفْظَهُ.

وقد بدأ المستشرقون العمل عليها فكانت البداية بقائمة بريل الموسومة بـ"قائمة الكلمات الأساسية للغة الصحافة العربية اليومية The Basic Word List of the Arabic Daily Newspaper" التي عرفت بـ"قائمة بريل" أو "قاموس الصحافة العربية"[11] ثم التربويون العرب كما في قائمةُ الدكتور داود عطية عبده التي ضمت ثلاثة آلاف كلمة ونيف من المفردات الأكثر شيوعًا الواردة في قوائم عاقل ولانداو وبريل وعبده –قائمة من وضع الدكتور داود عبده لم تنشر– والمجموعة من المواد الصحفية وكتب القراءة والأدب –نثرًا وشعرًا– وموضوعات أخرى متفرقة كالتاريخ والاقتصاد والتربية والاجتماع[12].

وقد انبثقت فكرةُ المعاجم التكرارية عن قوائم الشيوع حيث إن كليهما يهدف إلى حصر كلمات اللغة محتكمين إلى ما سُمي "معلومات التكرار" إلا أن الفارق بينهما كبير في نقطة الانطلاق، الهدف منها، الآلية المستخدمة، المخرجات المنتجة، الأهمية اللغوية، ومجالات الاستخدام.

[9] **Buckwalter , T. & Parkinson, D. (2011). A Frequency Dictionary of Arabic: Core Vocabulary for Learners. Routledgep(series preface-vii). Adapted.**

[10] **Look, A Frequency Dictionary of Arabic: Core Vocabulary for Learners, Op.cit. p1, Adapted.**

[11] **موسى بريل "قاموس الصحافة العربية"، القدس، الجامعة العبرية 1940.**

نقلا عن: (عبده) داود عطية: "المفردات الشائعة في اللغة العربية: دراسة في قوائم المفردات الشائعة في اللغة العربية" 1979م.

[12] **"المفردات الشائعة في اللغة العربية: دراسة في قوائم المفردات الشائعة في اللغة العربية" مرجع سابق، ص(ز،ح). بتصرف.**

وقد توافرت ثلاث دراسات قائمة على دراسة معلومات التكرار في اللغة العربية، وهي لا تنتمي إلى المعاجم، سوى المحاولة الثالثة وهي النواة الأولى لوضع المعجم التكراري محل الدراسة، إلى جانب دراستان قائمتان باللغة الإنجليزية:

1.أطروحة دكتوراه بجامعة برمنجهام the University of Birmingham بعنوان

"MODERN MEDIA ARABIC: A STUDY OF WORD FREQUENCY IN WORLD AFFAIRS AND SPORTS SECTIONS IN ARABIC NEWSPAPERS"

للباحث ZAINUR RIJAL ABDUL RAZAK

وفيها تناول لغةَ الصحافة العربية في ميداني "الشئون العالمية والرياضة" وينتهج فيها تحليل "معلومات التكرار" لمادة المدونة المجموعة.

2.معجم تيم بكوالتر Tim Buckwalter الموسوم بـ"المعجم التكراري للغة العربية A Frequency Dictionary of Arabic"، وقد ساهم في صناعته أيضًا ديلوورث باركنسون Dilworth Parkinson.

وهو يحوي خمسة آلاف كلمة مستخدمة في اللغة العربية الفصحى الحديثة Modern Standard Arabic (MSA)، مرتبة باعتبار درجة شيوعها، الأمر الذي يُخرجُهُ عن دائرة المعجمات ليصبح مماثلا لقوائم الشيوع في مناهجها، لولا أن هذا المعجم يهتم بإيراد الأمثلة والشواهد على استخدام الكلمات؛ حيث إنه لم يعتمد معايير الصناعة المعجمية في ترتيب مداخله فهو لم يرتب وحداته جذريًّا أو جذعيًّا أو ألفبائيًّا وإنما اعتمد معيار الشيوع لترتيب هذه الوحدات.

3.بحث للباحث المعتز بالله السعيد بعنوان "المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)"، المنشور في ندوة القرآن الكريم والتقنيات المعاصرة بالمملكة العربية السعودية عام 2009م.

وتقوم الدراسة فيه على إحصاء معلومات التكرار لألفاظ القرآن الكريم.

4. المدونات اللغوية

لقد تطورت لسانيات المدونات(Corpus Linguistics)في برامجها وأصبحت مادة مهمة تدرس في أقسام اللسانيات في عدد من الجامعات الغربية والعربية، ويمكن تعريفُ المدونات بأنها "مجموعةٌ ضخمةٌ من النصوص اللغوية (منطوقة أو مكتوبة) مودعةٌ في مخازن حاسوبية"[13].

وتكمن أهمية المدونة في الواقع والتمثيل الحقيقي للغة، والشمول من حيث المصادر والتنوعات والاستعمالات اللغوية والأساليب والأجناس الأدبية والتخصصات العلمية والتقنية، وذلك بشرط مراعاة ذلك عند إعداد المدونة.. كما تتميز بإمكانية إخضاعها للتحليل الإحصائي من جوانب مختلفة ولأغراض متباينة، والتعرف على شيوع الكلمة وشيوع معانيها المختلفة

[13] **تطبيقات استعمال لسانيات المدونات في إصدار معجم تأريخي للغة العربية، د.جمعان عبدالكريم.**

ونسبة شيوع الكلمة مقارنة بمجموع الكلمات في المدونة، إضافة إلى شيوعها من عدمه في أنواع النصوص المختلفة، وهو ما يفيد في استخلاص المصطلحات الشائعة في كل تخصص من التخصصات العلمية والتقنية. هذا فضلا عن إمكانية التعرف على شيوع الأوزان والصيغ الصرفية المختلفة، وإمكانية إجراء أنواع من التحليل النحوي والتركيبي، مع توافر بعض المتطلبات اللازمة. وأخيرًا إمكانية إجراء التحليل الصوتي (بوصف الحروف تمثيلا للأصوات العربية من حيث شيوعها ومواقعها في الألفاظ إلى غير ذلك)[14].

وتعتبر المدونات اللغوية كنزًا لغويًّا مهمًّا، إذ إنها قادرة على المساهمة في حل أغلب المشكلات اللغوية المعاصرة. وقد استفاد منها الباحثون في عمل العديد من التطبيقات كالمعاجم اللغوية، وتعليم اللغات [15]. وقد صارت مدونات اللغة العربية المعاصرة تشغل موقعًا مهمًّا ضمن المعالجة الحاسوبية للغة العربية، فوَضْعُ معاجم آلية حديثة أو ورقية يستوجب الأخذ بعين الاعتبار خصائص الكلمات بغية تجاوز ثغرات المعاجم التقليدية، ومن ثم إيجاد حل ناجح للكثير من القضايا اللغوية التي تواجه المعالجة الحاسوبية للغات الطبيعية، من قبيل اللبس الدلالي للمفردات والعبارات والنصوص، والترجمة الآلية، وبنوك المعطيات وغيرها من التطبيقات الآلية.

أما فيما يخص الصناعة المعجمية فقد كانت البداية للأديب الإنجليزي صموئيل جونسون Samuel Johnson (1707-1784) في استخدام المدونات اللغوية في الصناعة المعجمية Lexicography؛ حيث أنجز "معجم اللغة الإنجليزية A Dictionary of the English language"، الذي نُشِرَ عام 1755م، وقد احتوى على أكثر من أربعين (40) ألف مدخل معجمي معتمدًا على مدونة لغوية، جُمعتمن الأعمال الأدبية لشكسبير Shakespeareوملتون miltun ودريدن Dryden وغيرهم من أعلام الأدب الإنجليزي في ذلك الوقت[16].

وبعد محاولة صموئيل شقت آليةُ المدونات اللغوية طريقها إلى صناعة المعاجم خاصة التعليمية منها، وذلك لدقة التمثيل اللغوي لواقع اللغة المدروسة، إلى جانب تطورها ودخولها طور الحوسبة مما أدى إلى سهولة التعامل معها جمعًا وتحليلا ونشرًا.

كما عرفت المدوناتُ اللغوية طريقها إلى التوظيف في مجال "تعليم اللغات" عندما وضع ثورنديك Thorndike في بداية القرن العشرين قائمته teacher's word book، معتمدًا على مدونة لغوية قام بجمعها من 81 مصدرًا مختلفًا.

وبداية بقائمة ثورنديك بدأ اللغويون وكذا التربويون استخدام المدونات اللغوية في وضع المعجمات التعليمية، كما ساهمت بصورة مباشرة في تطوير مناهج علم اللغة التربوي Educational Linguistics[17].

[14] د.محمود إسماعيل صالح ود.إبراهيم الخراشي: الجانب اللغوي للمعجم الحاسوبي للغة العربية، ضمن البحوث المقدمة في اجتماع خبراء المعجم الحاسوبي التفاعلي، ص2.

[15] Martin Thomas, “Electronic Text”, Electronic Text Notes - tx.xml,2002

[16] (السعيد) المعتز بالله "المدونات اللغوية" من كتاب :مدخل إلى علم اللغة الحاسوبي-تحت النشر" ص4، بتصرف.

[17] (السعيد) المعتز بالله "المدونات اللغوية" من كتاب :مدخل إلى علم اللغة الحاسوبي-تحت النشر" ص5، بتصرف.

5. استخدام المدونات اللغوية في المعاجم العربية

رغم أن اللغة العربية ثاني أشهر اللغات المستخدمة حول العالم بعد الإنجليزية، وهي اللغةُ الرسميةُ لِمَا يزيد عن 422.039.637 مليون نسمة[18]. إلا أنها لم تجد طريقها إلى الحوسبة اللغوية والمدونات اللغوية إلا بداية القرن العشرين، وربما يعود هذا التأخر إلى صعوبة معالجة العربية حاسوبيًّا نظرًا لقصور البرامج المتاحة حاليًا عن التعامل مع النظام الصرفي والاشتقاقي والتركيبي للغة العربية، إذ إن اللغة العربية ونظامَها المعجمي يفرض منهجًا معينًا، يختلف عن معالجة باقي اللغات اللصقية كالإنجليزية، على صناع المعاجم بسبب طبيعتها الاشتقاقية التي تتيح لها إنتاج عدد كبير من الكلمات من جذر واحد مثل اشتقاق كاتب ومكتوب ومكتب ومكتبة وكتاب ومكاتبة.. إلخ من الجذر (ك ت ب).

وقد عرفت العربيةُ طريقها إلى توظيف المدونات اللغوية في خدمة مضمار الدرس اللغوي في مطلع القرن العشرين، وكانت تلك المحاولات عبارة عن مشروعات بحثية وأطروحات علمية معدودة، كمدونة نايميخن NIJMEGEN Corpusبين عامي (1990–1996)، والمدونة العربية Corpus Linguae Arabicae"CLARA" عام 1977م، إلى جانب الأطروحات العلمية كمدونة العربية المعاصرة Corpus of Contemporary Arabicللباحثة القطرية لطيفة السليطي عام 2004م، وصولا إلى "مدونة معجم تاريخي للغة العربية" للباحث المعتز بالله السعيد، عام 2010م[19].

ثانيًا: إشكالات الدِّراسة

لم تحظَ اللغة العربية باهتمام صنَّاع المعاجم ومعالجي اللغات الطبيعية بدرجة الاهتمام التي حظيت بها باقي اللغات وعلى رأسها اللغة الإنجليزية. ورغم أنَّ فكرة المعاجم التكرارية ترجع إلى النصف الأول من القرن العشرين، وكذا انتشار المعاجم التكرارية في كثير من لغات الأمم الأخرى، إلا أنها لم تعرف الطريق إلى اللغة العربية بالشكل والصورة المنشودة حتى الآن، وربما يعود ذلك إلى طبيعة اللغة العربية الاشتقاقية ونظامها التصريفي المعقد، إضافةً إلى انفراد العربية ببعض الظواهر الشكلية والتركيبية والبنوية التي تميزها عن غيرها من لغات الفصائل اللغوية الأخرى، بل وتميزها – كذلك– عن شقيقاتها من اللغات السامية[20].

ولعل هذا الإهمال يكمن في صعوبة التعامل مع اللغة العربية ذات الطبيعة الصرفية الاشتقاقية؛ حيث تتمثل صعوبات التعامل مع العربية لبناء معجم فيطبيعة اللغة العربية الاشتقاقية ونظامها التصريفي المعقد؛ حيث تنعكس تلك

[18] http://www.scribd.com/doc/14436546/Languages-Spoken-by-More-Than-10-Million-People.

[19] **لمزيد من المعلومات عن هذه الدراسات، انظر، (السعيد) المعتز بالله "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، أطروحة لنيل درجة الماجستير بكلية دار العلوم، القاهرة، 2008، ص20-28..**

[20] **"المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)" مرجع سابق، ص2 ..**

الطبيعة على المعالجة الآلية لنصوص المدونات المستخدمة في الصناعة المعجمية، إذ تقوم المعالجة على إخضاع الآلة لمتطلبات البحث اللغوي، من خلال أنظمة تفاعلية بإمكانها التعامل مع الإنسان وتنفيذ أوامره[21].

هذا بالإضافة إلى جانب قصور أدوات التحليل عن التعامل مع نظام العربية فهي تتعامل مع المفردات بوصفها مجموعة من الرموز المتلاصقة، دون النظر إلى اللغة التي تنتمي إليها النصوص.

من ناحية أخرى، فإن صناعة المعاجم التكرارية تستلزم الاستعانةَ بمدونة لغوية حاسوبية، "ولا يزال منهج دراسة المدونات اللغوية بكرًا جديدًا غض الإهاب على اللغة العربية، التي لم تعرف الطريق إليها إلا في الثمانينيات من القرن الماضي"[22].

ثالثًا: المنهج

لمَّا كان البحثُ يهدفُ إلى وضع "معجم تكرار للغة العربية المعاصرة" فقد اعتمد المنهجَ الوصفيَّ في رصد مادته اللغوية، إذ يقومُ البحث على رصد مفردات اللغة العربية ومحاولة توصيفها وتنظيمها وفق منهجيةٍ تقوم على ترتيب مفردات المادة المجموعة حسب عدد تكرارها في المدونة.

وقد اتخذ البحثُ من لسانيات المدونة corpus linguistics وسيلةً لبلوغ المعجم المنشود على نحو يساهم في خدمة الدرس اللغوي وتعليم اللغات وصناعة المعاجم على وجه الخصوص، فقد جمع البحثُ بين لسانيات المدونة ومناهج الصناعة المعجمية العربية، إذ يقوم على تطوير منهج ثورنديك التكراري في قائمته للانطلاق إلى معجم يسير وفق مناهج الصناعة المعجمية من حيث ترتيب المواد ومعالجتها داخل المعجم بما يفي بالغرض المنشود من المعجم، وتوظيفه في خدمة أرباب العربية والواردين عليها، تيسيرًا لسير العملية التعليمية وجعلها أكثر نجاعة بين طلاب اللغة العربية حول العالم.

كما تجمعُ الدراسة في جانب معالجة المادة وبناء المعجم بين تقنيات الحاسوب الداعمة للعربية وأساليب التحليل الإحصائي، التماسًا لدقة المخرجات وتحقيقًا للقدر الأكبر من الفائدة، بما يعود بالنفع على ركنيّ العملية التعليمية – المعلم والمتعلم[23].

[21] **Bolshakov , I. and Gelbukh, A(2004). Computational Linguistics "Models, Resources, Applications" p.15.**

[22] **"مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق، ص5.**

[23] **"المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)" مرجع سابق ، ص2.**

تسير منهجية بناء المعجم المنشود وفق مرحلتين رئيسيتين تتقسمان حول طرق معالجة كل منهما، فالأولى تسير وفق الصناعة المعجمية العربية بداية من مرحلة جمع مادة المعجم مرورًا بمرحلة التحرير المعجمي لمادة المدونة وترتيب مداخل المعجم، وانتهاء بمرحلة نشر المخرج النهائي للمعجم المتمثل في "المعجم التكراري للغة العربية المعاصرة" وفق الهدف المنشود من المعجم، وطبيعة الدراسة[24].

أما المرحلة الثانية فهي المعالجة الحاسوبية والإحصائية لمادة المدونة اللغوية، وتتمثل المعالجة الحاسوبية في استخلاص المداخل والوحدات المعجمية مع المعاني والشواهد ومعلومات تردد المفردات داخل المدونة، عن طريق عرض المادة المجموعة على البرامج المتخصصة للتعامل مع العربية مثل المحللات الصرفية، بالإضافة إلى المساعدة في تحرير مادة المعجم وفهرستها عبر المفهرسات الآلية، إلى جانب المساهمة في نشر النتاج النهائي للمعجم المنشود.

كما تتمثل المعالجة الإحصائية في اختبار مدى صلاحية المدونة لتمثيل المجتمع اللغوي، والتأكيد على دقة ضبطها، وإحصاء المفردات وترتيبها في مجموعات متدرجة من الأكثر ورودًا وترددًا إلى الأقل، عن طريق استخدام البرامج والمعادلات الإحصائية[25].

1. المعالجة الآلية للمدونة اللغوية

وتأتي هذه المعالجة متمثلةً في اتباع الخطوات الإجرائية والتنفيذية لوضع المعجم المنشود؛ حيث يمر العمل في المعجم بجملة خطوات تأتي في:

1.جمع المادة وتحديد المصادر التي سيعتمد عليها.

2.التحرير المعجمي لمادة المعجم

1) اختيار الوحدات المعجمية أو وضع قوائم بالكلمات الرئيسية التي ستشكل مداخل المعجم.

2) تأليف المداخل أو معالجتها من نواحيها المختلفة.

3) ترتيب مداخل المعجم بطريقة من طرق الترتيب المعجمي[26].

3.نشر المعجم.

[24] **لمزيد من المعلومات عن هذه الدراسات، انظر، "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق**

[25] **لمزيد من المعلومات عن هذه الدراسات، انظر، "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق**

[26] **(عمر) أحمد مختار "صناعة المعجم الحديث"، ط1، 1418هـ/1998م، عالم الكتب-القاهرة، ص165.**

1.1. جمع مادة المعجم

مادة المدونة اللغوية

لا تُلزم المدونة واضعَها بمادة معينة، إنما هي نصوص عشوائية تخضع لأسس ومعايير يحددها الهدفُ المنشود من المدونة، فمادة المدونة المعدَّة للمعجم التاريخي غير المادة المُعَدَّة للمعجم المعاصر، غيرها للمعجم التكراري، فهي تخضع لمعيارين أساسيين هما المعاصرة لفترة جمع المادة ودراستها حتى تعكس الواقع اللغوي الحقيقي للغة المدروسة وقت وضع الدراسة، والتنوع والشمول، فوجب على صانع المدونة الحرصُ على تنوع مادته لتشمل كافة جوانب اللغة المستخدمة في شتى المجالات من علوم ورياضة وأعمال، كذلك شمول المادة المجموعة لكافة نواحي وألفاظ مفردات اللغة.

يتم تحديد مادة المدونة وفق الهدف المنشود منها، ولمَّا كان الهدفُ من الدراسة إعدادَ معجم تكراري للغة العربية المعاصرة، فقد حاول الباحث تلمس مصادر اللغة التي تعبر عن المستوى اللغوي المدروس، وقد حدد لذلك ثلاثة مصادر:

1.لغة الصحافة المعاصرة: وهي لغة تكتسب خواصها التركيبية من مصادر ثلاثة هي:

أ- الفصحى –كما قعدت لها كتب اللغة– وتعد لغة الصحافة امتدادًا لها وتطورًا لبعض خواصها.

ب– اللغات الأجنبية: بما تسهم به في لغة الصحافة من مفردات وأساليب يتم تعريبها، وما يحدثه ذلك من تغيير في نظام الجملة.

ج- اللغة العامية: بما تفرضه لغة الصحافة منها من مفردات وأساليب[27].

وقد جُمعَت المادة التي تمثل لغة الصحافة من المواقع الإلكترونية للصحف المصرية اليومية.

وقد جاء اختيار مادة الصحافة ضمن مادة المدونة المجموعة للمعجم المنشود لأسباب، منها أنها تعد من أهم أدوات الاتصال وهي بذلك تؤثر تأثيرًا بالغًا في اللغة العربية المعاصرة، كما أنها نمط من أنماط العربية المعاصرة، إذ يطلق عليها الدكتور كمال بشر "العربية المعاصرة"، وكذلك الدكتور السعيد بدوي يسميها "فصحى العصر"[28].

2.المستوي اللغوي الأدبي للعربية المعاصرة، المتمثل في لغة الدساتير المصرية كدستور مصر 1985، ولغة الأدب المتمثلة في بعض الروايات والقصص كقصص يوسف إدريس، والمؤلفات الإسلامية.

3.المادة المأخوذة من موقع الموسوعة الحرة Wikipedia ويكيبيديا، لما يمثله هذا الموقع من شمول لألفاظ اللغة العربية، وتنوع إذ يعمل على تدوينه أكثر من 3000 آلاف مدون من بقاع العالم، ومعاصرة مادتها

[27] انظر، (عبد العزيز) محمد حسن "لغة الصحافة المعاصرة"، دار الفكر العربي-القاهرة، الطبعة الأولى 2002م، ص1.

[28] انظر، "لغة الصحافة المعاصرة"، مرجع سابق، ص10-11، بتصرف.

وتمثيلها للمادة الخام المنشودة لوضع مدونة المعجم، إلى جانب أنها تخضع لرخصة جنو للوثائق الحرة GNU Free Documentation License[29].

ويستمد المعجم مادته من مدونة لغوية محوسبة تعكس الواقع اللغوي والاستخدام الحالي لمفردات اللغة العربية المعاصرة.

2.1 التحرير المعجمي لمادة المدونة

تهدف الدراسةُ إلى وضع معجم تكراري موجَّه لمتعلمي اللغة العربية من أبنائها المحليين أو غير الناطقين بها، ويقوم المعجمُ بحصر أشهر كلمات اللغة العربية المعاصرة، مرتبًا إياها حسب عدد تكرارها ومدى دورانها في اللغة، والغرض من ذلك وضعُ أشهر كلمات اللغة وأكثرها استخدامًا بين يدي المتعلم، ليوجِّه إليها جهدَه ووقته، فبإتقانه لها يكون قد ألمَّ بثلثي اللغة.

وتتم معالجةُ مادة المدونة حتى تكون قابلة للمعالجة الآلية، عن طريق تحويل النصوص إلى صورة إلكترونية يسهل التعامل معها آليًّا، وقد استُخدِمت هذه الطريقةُ في إدخال المادة المجموعة من باب الأدب والمؤلفات كالروايات والمجموعات القصصية والكتب العلمية[30].

1.2.1 المراجعة اللغوية لنصوص المدونة

وجب التخلص من الأخطاء الإملائية الموجودة في نصوص المدونة مادة المعالجة إذ إنها تنعكس على أداء المحللات اللغوية؛ وبما أن مادة المدونة أغلبها جُمعت من مواقع الويب وصفحات الشبكة العنكبوتية وهي بالطبع تحمل قدرًا ليس باليسير إذ إن أكثر القائمين ليسوا على علم باللغة العربية وقواعدها، لهذا فهي في الأغلب تحتوي على أخطاء سواء إملائية أو تركيبية أو صرفية أو حتى المعجمية والأخطاء الشائعة، وهذا من شأنه أن يؤدي إلى خلل في عمل المفهرسات والمحللات الآلية أداة العمل داخل المدونة، حيث إنها قد تؤدي إلى تعدد أشكال المفردة الواحدة ما يؤدي إلى تعدد مداخلها، كذلك كتابة شكل المفردة بصورة خاطئة يجعلها تستقل بذاتها ويتم إدراجها خطأً ضمن مفردات اللغة.

وتتم عملية المراجعة اللغوية إما آليًّا عن طريق أداة التصويب اللغوي والإملائي "Fix Broken Text" الملحقة ببرنامج معالجة النصوص المكتبية "Microsoft Office Word"، ويتمثل عمل هذه الأداة في أنها تطرح عدة احتمالات للمفردة المراد تصحيحها، ويقوم المستخدم باختيار الاحتمال الذي يراه صحيحًا، وإما يدويًّا وهذه لا غنى عنها لضبط المادة

[29] **يمكن الاطلاع على صيغة الرخصة في إصدارها الأخير (نوفمبر 2008) عبر الموقع الإلكتروني**http://www.gnu.org/

[30] **"مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق24، بتصرف.**

المجموعة، فهي العملية الأم في تهيئة مادة المدونة؛ نتيجة أن التعامل مع نصوص المدونة يجب أن يتميز بالحرص ويكتنفه التحري والدقة حتى تكون النتائج صحيحة ودقيقة[31].

وقد قام الباحث بمراجعة النسبة الأكبر من نصوص المدونة المجموعة.

2.2.1 اختيار الوحدات المعجمية وتأليفُ المداخل ومعالجتها

المدخل المعجمي Lexical Entry –أو ما يطلق عليه المادة المعجمية– هو ذلك الحقل الذي تنتمي إليه مجموعة من الكلمات المشتركة في مادة لغوية واحدة، سواء أكانت جذرًا لغويًّا لكلمة عربية أو معربة، أم مادة معجمية لكلمة دخيلة. ويطلق على هذه الكلمات المنسدلة عن المدخل المعجمي مجموعة الكلمات الرأسية Headwords، سواء أتعددت معانيها أم اشتركت في معنى واحد[32].

وقد أخضع البحثُ المادة المجموعة للمعالجة الآلية واستخلاص مداخل المعجم وكلماته الرأسية، عن طريق المحللات الصرفية، حيث إن المحلل الصرفي يقوم بتحليل الكلمة المفردة ويقدم قائمة بكل الاحتمالات الممكنة، فعند إدخال المادة المراد تحليلها إلى المحلل الصرفي يقوم بتفصيل المدخل أو الجذر، إلى جانب المفردات المنسدلة عنها من اشتقاقات وتغييرات ناتجة عن السوابق واللواحق مثل المدخل (ق و ي)، هو المدخل الذي تنسدل عنه المفردات المعجمية {قوة، بالقوة، بقوة، تقوية، تقويتهم، الاستقواء، قوى، القوى، والقوى، قوات، القوات، والقوات، للقوات، وللقوات، قواتها}.

ومن المحللات الصرفية المستخدمة في البحث المفهرس الآلي (Concapp، aConCorde، Concordance، MonoConc)[33].

3.2.1 ترتيب مداخل المعجم بطريقة من طرق الترتيب المعجمي

تم ترتيب مداخل الكلمات داخل المعجم ترتيبًا أبجديًّا دون فصل بين الفعل والاسم والحرف، فقد اكتفى بالنص على نوع الكلمة في معالجة الجانب الصرفي فيذكر نوع الكلمة (اسم تفضيل، اسم فاعل، فعل أمر، فعل ماض، حرف جر.. إلخ)، هذا على الجانب الصرفي.

أما على الجانب الدلالي فكانت على السواء لمداخل الأفعال والأسماء والكلمات الوظيفية، فيأتي المدخل داخل الشاهد مع ذكر معنى المدخل مقدمًا، مع النص على التغيرات السياقية والتصاحبات اللفظية.

3.1 تنسيق المادة داخل المدونة لمعالجتها

[31] "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق62-63، بتصرف.

[32] "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق136.

[33] لمزيد من المعلومات عن المحللات الصرفية انظر، "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق، ص136.

حتى يكون النتاجُ النهائي للعمل المنشود صحيحًا ودقيقًا، وجب تحري هذه الدقة عند إدخال المادة الخام أساس النتاج المنشود، ولطالما احتوت مادة المدونة المجموعة على قدر كبير من الرموز والأشكال والرسوم، مع اختلاف أنواع الخطوط وأحجامها وألوانها، وبالطبع كل هذا له من الآثار السلبية التي تعيق عمل البرامج الحاسوبية المستخدمة في معالجة مادة المعجم، لذا وجب التخلص من كل هذه المعيقات حسبما تقتضي أدوات التحليل الآلي. لذلك لجأ الباحث إلى:

تقسيم المواد اللغوية المجموعة على أكثر من ملف نصي؛ حيث إن المحللات الآلية تعجز عن التعامل مع العديد من الصفحات؛ واتباع:

أ. توحيد نوع الخط ولونه وحجمه (نوع الخط: Courier New، لونه: أسود، حجمه:14).

ب. توحيد الحدود والفواصل بين الصفحات.

ج. تجريد النصوص من علامات الضبط، وذلك للحد من تعدد أشكال المفردة الواحدة.

د. تجريد النصوص من الأرقام والرموز والحروف اللاتينية.

ه. تجريد النصوص من الأشكال والرسوم التوضيحية.

و. تجريد النصوص من علامات الترقيم.

ز. نسخ النصوص المنسقة، وإعادة تجميعها في ملف واحد بامتداد TXT، تمهيدا لمعالجتها آليًّا.[[34]].

1.3.1 أدوات التحليل الآلي لنصوص المدونة

يصعب التعامل مع اللغة العربية بأدوات التحليل الحاسوبي إذ إن طبيعتها الاشتقاقية والبنوية تختلف اختلافًا كليًّا عن اللغات اللصقية كالإنجليزية.

عند معاجلة اللغة العربية آليًّا فإنَّ أول ما يواجه تلك المعالجة التداخل بين المستويات اللغوية، مما يوجب استعمال أنظمة متعددة لمعالجة تلك المستويات، وأن يرتبط كل منها بالآخر، لهذا تم التعامل مع مادة البحث على مستويات التحليل اللغوي من الناحية (الصرفية والبنوية والتركيبية)، فلجأ البحث إلى أدوات التحليل الحاسوبية المتمثلة في:

1.1.3.1 المحلل الصرفي Morphological Analyzer

ويتمثل عملُ المحلل الصرفي الآلي في القيام بنوعين من المعالجة، هما: توليد الكلمة وتحليلها، حيث يقوم المفهرس بتحليل المدخل إلى عناصره الأولية، ليقوم بتحديد الصيغة النهائية للكلمة. فعلى سبيل المثال عند تحليل الفعل (تَقاتَلَ) صرفيًّا، نجد أن المخرجات تتمثل في:

[34]لمزيد من المعلومات عن المحللات الصرفية انظر، "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية" مرجع سابق، 64، بتصرف.

1- الصيغة الصرفية (تفاعل).

2- عناصر الفعل الأصلية (ق ت ل).

3- زوائد تصريفية (ت، ا، َ).

4- زوائد إعرابية (البناء).

كما يقوم المحلل الصرفي بتحليل الكلمة وتفكيكها إلى أجزائها الصرفية الصغرى (المورفيم) مثل السوابق واللواحق التي تلحق أصل الكلمة وغير ذلك؛ حيث يقوم بإعطاء الكلمة بياناتها الصرفية الكاملة، مثل الجذر، والساق، والسوابق, وقسم الكلم، والميزان الصرفي، هذا بالإضافة لوضع السمات الصرفية الخاصة بكل كلمة على حده. فهو أشبه بمعجم صرفي متكامل للكلمة العربية، فهو يستخدم في استرداد جذور المفردات، وتحديد المعلومات الصرفية الخاصة بكل مفردة على حدة، كما يستخدم في توليد المشتقات اللفظية من الجذر اللغوي الوحيد، وهو –بذلك– يتمم عمل المفهرس الآلي[35].

2.1.3.1 المفهرسات الآلية: Electronic Concordance

المفهرس الآلي برنامج إداري يُستخدَم في تنظيم النصوص وفهرستها وترتيب مفرداتها حسبما تقتضي طبيعة الدراسة المعدة؛ ويُعَدّ أحد الأدوات الأساسية المستخدمة في تحليل نصوص المدونات اللغوية، لاسيما المدونات المصنوعة لأغراض معجمية.

1- ويوفر المفهرس الآلي الكثير من الوقت والجهد، إذ يعيد تشكيل النصوص المدرجة لتظهر في صورة منظمة، يسهل التعامل معها آليًا، ومن هذه الفهارس: (Concapp، aConCorde، Concordance، MonoConc)[36].

3.1.3.1 معنون التراكيب العربية Arab Tagger

تنتظم كلمات العربية وفق نظام معين هو المسئول عن هذا التنظيم وربط مكونات الجمل بعضها ببعض، هذا النظام يسمى "النحو العربي"، وهو ما يمثل الجانب التركيبي للغة العربية، وذلك وفق النظام الدلالي للغة القائم على جوانب المعنى والتعبير، إلى الجانب الصرفي المسئول عن المفردات اللغوية.

يقوم عمل معنون التراكيب العربية Arab Taggerعلى أساس المنظمة النحوية؛ حيث يتمثل عمله في مساعدة أدوات التحليل الآلي للنصوص عن طريق إضافة عناوين Tags تعبر عن الصفات النحوية الأساسية لكلمة على حدة[37]، كأن ينص على أن كلمة ما تنتمي إلى قسم الكلام (الاسم أو الفعل أو الحرف)[38].[39].

[35] "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، مرجع سابق، ص110. وانظر المرجع لمزيد من المعلومات عن أنواع المحللات الصرفية ص110-128.

[36] للمزيد عن المفهرسات الآلية وأنواعها، انظر، "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، مرجع سابق، ص88-108.

4.1. نشر المعجم

يتم نشر المنتج النهائي في صورته المنشودة، سواء النسخة المطبوعة ورقيًا، أو النسخة الإلكترونية، بالشكل الذي يُأمَّل أن يُفيد به المعجم أرباب اللغة العربية.

2. المعالجة الإحصائية لمادة المدونة اللغوية

كان ميدان الإحصاء في الدراسات اللغوية هو الميدان الأول لتطبيق اللسانيات الحاسوبية على اللغة العربية؛ حيث صدرت الدراسة الإحصائية للجذور الثلاثية لمفردات اللغة العربية في مايو 1971، وتبعتها دراسة إحصائية للجذور غير الثلاثية في يناير 1972م.

إن المعالجة الإحصائية لمدخلات المدونة تأتي تتمة للمعالجة الآلية؛ حيث إن المعالجة الإحصائية هي الأداة التي عن طريقها يتم التأكد من أن البحث يسير على الطريق الصحيح من حيث دقة مخرجاته وصحتها.

وتتمثل الوظيفة الأولى للإحصاءات التي تخدم البحث، في إيجاز وتلخيص خصائص وأوصاف الوحدات موضوع الملاحظة في خصائص يمكن قياسها وعدَّها. كما أن الإحصاءات الوصفية تخدم البحث في تلخيص المعطيات – البيانات– الكمية[40].

وتتمثل المعالجة الآلية في المدونة –موضع البحث– في تطبيق نظرية العينات الإحصائية، في سبيل تتبع نتائج المعالجة الآلية، ويتم اللجوء لهذه النظرية بحيث تكون تأكيدًا على المادة المجموعة أنها ممثلة تمثيلا صادقًا للمجتمع اللغوي المدروس أو تعتبر صورة مصغرة منه[41]. وقد اتجه البحث إلى استخدام نظرية "العينات الإحصائية العشوائية" نظرًا لأن جمع اللغة أمر صعب للغاية لا يستطيع فرد أيًا كانت إمكاناته حصرها كاملة دون خلل، فحاول البحث استقصاء عينة جزئية من المجتمع الكلي لدراستها، على أن تكون تلك العينة الجزئية ممثلة لكل خصائص اللغة المدروسة والمستوى

[37] **Available from The World Wide Web**: http://www.rdieg . **com/rdi/new/Downloads/Scientific%20Papers/Arabic%20POS%20Tagging%20And%20Applications_NEMLAR%20conference%20Paper_9-2004.pdf.**

[38] **"مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، مرجع سابق، ص127، بتصرف.**

[39] **لمزيد من المعلومات عن المعنونات اللغوية وأنواعها، راجع "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، مرجع سابق، ص126-129.**

[40] **(أحمد) غريب محمد "مدخل إلى الإحصاء"، برنامج دراسة المجتمع، مركز التعليم المفتوح، جامعة بنها، المستوى الأول، 2012، ص12، 13.**

[41] **د.جلال الصياد، د.عبد الحميد محمد ربيع، "مبادئ الطرق الإحصائية"، ط1، 1983م، تهامة/جدة/المملكة العربية السعودية، ص10، ص107، بتصرف.**

اللغوي المستخدم، فضلا عن أن جمع اللغة بصورة كاملة يحتاج إلى إمكانات عالية ورعاية مؤسسة توفر الإمكانات والآليات اللازمة لذلك.

وتكمن أهمية المعالجة الإحصائية في:

1. تساعد البحث على إعطاء أوصاف على جانب كبير من الدقة العلمية.
2. تساعد على تلخيص النتائج في شكل ملائم مفهوم. فالبيانات التي يجمعها الباحث لا تعطي صورة واضحة، إلا إذا تم تلخيصها في معامل أو رقم أو شكل توضيحي، كالرسوم البيانية.
3. تساعد على استخلاص النتائج العامة من النتائج الجزئية. فمثل هذه النتائج لا يمكن استخلاصها إلا تبعًا لقواعد إحصائية، كما يستطيع البحث أن يحدد درجة احتمال صحة التصميم الذي يصل إليه.
4. تمكِّن البحث من التنبؤ بالنتائج التي يحتمل أن يحصل عليها في ظروف خاصة[42].
5. تعين على التخلص من أثر العوامل الأخرى التي لا يستطيع تفاديها في بحوثه، والتي تؤثر دائمًا في نتائج كل بحث كعامل الصدفة واختيار العينات.
6. تساعد المعالجة الإحصائية الباحثَ في تنظيم خطوات بحثه، فهو يحتاج إليها في مرحلة تصميم البحث وتخطيطه، حتى يمكنه في النهاية أن يخرج من بحثه بالنتائج التي يسعى إلى تحقيقها[43].

رابعًا: التَّطبيق

فكرة المعاجم التكرارية حديثة نسبيًّا فقد سبق إليها ثورنديك عام 1921م في اللغة الإنجليزية ومن ثمَّ انتقلت الفكرة إلى لغات عدة حتى أصبحت آليةً معتمدة في تسهيل عملية تعلم اللغة، وللأسف لم تطبق الفكرة على اللغة العربية إلا في محاولة وحيدة متمثلة في دراسة على ألفاظ القرآن الكريم[44].

يعرف المعجم التكراري بأنه نوع من المعاجم اللفظية المتخصصة، تُرتَّب فيه الأبواب والفصول المعجمية بحسب شيوعها، متدرجةً من الأكثر شيوعًا إلى الأقل شيوعًا، ويعتمد المعجم في تحديد نسبة شيوع المواد المعجمية على مدونة لغوية Linguistic Corpusتعكس واقعَ اللغة وتمثل المستوى اللغوي الذي يعنى به المعجم[45]. وعلى هذا تعتمد فكرة المعجم التكراري على حصر تكرار كلمات النصوص المجموعة، ومحاولة تصنيفها في مجموعات مرتبة من الأكثر شيوعًا إلى الأقل شيوعًا.

[42] "مدخل إلى الإحصاء"، مرجع سابق، ص19.

[43] "مدخل إلى الإحصاء"، مرجع سابق، ص20.

[44] (السعيد) المعتز بالله: "المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)"، مرجع سابق، ص1.

[45] (السعيد) المعتز بالله: "المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)"، مرجع سابق، ص1.

وقد قام البحث بتقسيم مفردات المادة المجموعة حسب تكرار كل منها إلى خمس مجموعات، تحتوي المجموعة الأولى على أكثر الكلمات تكرارًا ودورانًا في اللغة، ثم المجموعة الثانية تحتوي على المفردات الأقل، إلى المجموعة الخامسة التي تضم أقل الكلمات استخدامًا في اللغة العربية المعاصرة.

وتأخذ كل مفردة من مفردات المدونة رتبة المجموعة التي تنتمي إليها حسب الجدول التالي:

المجموعة التكرارية	الرمز	تكرار المدخل في المادة المجموعة
الأولى	*	500 > (م)
الثانية	**	100 > (م) > 500
الثالثة	***	50 > (م) > 100
الرابعة	****	10 (م) > 50
الخامسة	*****	1 > (م) > 10

واستمد النموذج المعجمي مادته من الصحافة المعاصرة، فكانت جريدة الأهرام –البوابة الإلكترونية– ليوم 25 يونيو 2012 رافد مادته[46].

واعتمد النموذجُ فكرة المداخل اللغوية لإيراد مادته، فعن طريقها يمكن حصر الألفاظ الواردة التي يجمعها مدخل لغوي واحد. وجاء عدد كلمات المادة المجموعة 4996 كلمة.

وجاء النموذج المعجمي في المدخل اللغوي الواحد مقسمًا إلى حقول:

1) المدخل المعجمي: ويمثل المدخل اللغوي الذي يجمع الكلمات ذات الأصل الواحد.

2) التردد: ويعبر عن عدد تكرار المدخل اللغوي على مدار حصر كلمات المادة المجموعة.

3) نسبة الشيوع: وتحدد المجموعة التي ينتمي إليها المدخل المعجمي إن كانت المجموعة الأكثر شيوعًا أو الأقل أو الأقل، وجاء التعبير عن المجموعات وترتيبها وفقًا للجدول أعلاه.

4) المعنى المعجمي: وفيه حاول البحث استقصاء المعاني المعجمية للمدخل المعجمي عن طريق العودة إلى عدة معاجم: المعجم الوسيط (مجمع اللغة العربية بالقاهرة)[47]، والقاموس المحيط للفيروزابادي[48]، والصحاح في اللغة للجوهري[49]. في محاولة لحصر أفضل تعريف للمدخل المعجمي الوارد في مادة النموذج.

[46]جريدة الأهرام المصرية، البوابة الإلكترونية: ليوم الاثنين 5 من شعبان 1433هـ/ 25 يونيو 2012 / السنة 136/ العدد 45857.

5) الوحدات المعجمية: وتشمل الأشكال اللغوية للمدخل المعجمي الواحد الواردة على مدار المادة المجموعة.

ثم محاولة تفصيل عدد مرات تكرار كل وحدة معجمية على حدة، ونسبة ورودها إلى عدد تكرار المدخل المعجمي (عدد تكرار الوحدة المعجمية/تكرار المدخل المعجمي). مع محاولة تحليل تلك الوحدة المعجمية صرفيًّا وتحديد السوابق واللواحق لها. ومحاولة إيراد المعنى المعجمي للوحدة المعجمية المذكورة، مع التدليل على تلك الوحدة بذكر شاهد من النصوص المجموعة يوضحها ببنيتها المذكورة.

ويضم النموذج المعجمي تمثيلا لمدخل معجمي واحد، في محاولة لتوضيح فكرة المدونة المنشودة موضوع البحث.

النموذج المعجمي

المدخل المعجمي: ع ل و

التردد: 116

[47] مجمع اللغة العربية، ط4، 1425هـ/2004م، مكتبة الشروق الدولية.

المؤلف: مجمع الخالدين للغة العربية بالقاهرة. منهج الكتاب: رتب الكلمات بحسب الحرف الأول من الجذر مع مراعاة الحرف الثاني والثالث، مع ترتيب المواد داخليا بادئا بالفعل ثم الاسم ثم الصفة. مميزات المعجم: وضع بعض الاختصارات مثل (مج) للجمع، استعمال بعض الرسوم التوضيحية. اشتمل المعجم على ثلاثين ألف مادة ومليون كلمة. أدخل في متنه الألفاظ المولدة والمعربة والمحدثة التي تتصف بالشيوع كما حرص على ذكر المصطلحات العلمية الشائعة. والمادة الموجودة في الوسيط أكثر صحة من المنجد، فهو أفضل معجم عربي حديث. راجع د/محمد يوسف حبلص: "معاجم العربية ومصادرها"، دار العلوم–القاهرة، دار الهاني، ط2، 2007م، ص162.

[48] الفيروزابادي (العلامة مجد الدين محمد بن يعقوب الفيروزابادي الشيرازي، ط3، المطبعة الأميرية 1301هـ. الهيئة المصرية العامة للكتاب.

المؤلف: أبو طاهر مجد الدين الفيروزابادي (729–816 هـ). كيفية البحث في المعجم: منهج القافية، الحرف الأخير (باب)، والحرف الأول (فصل)، وما يثلثهما. منهج الكتاب: قسم معجمه أبوابا لكل حرف من الألفباء العادية باب، وقسم كل باب إلى فصول لكل حرف من الأبجدية العادية فصل. مميزات المعجم: حظي القاموس بإعجاب الناس لإيجازه ودقته وشموله، حيث أراد المؤلف أن يتلخص من الضخامة التي تتسم بها المعاجم مما يعوق الباحث عن بلوغ هدفه فلجأ إلى الاختصار والإيجاز؛ حيث جمع ميزتي الشمول والاستقصاء –استقصاء الصيغ ومشتقاتها ومعانيها– وكذلك سهولة المنهج ودقة الطريقة إلى جانب فكرته الجديدة وهي الإيجاز. كما لاحظ أن المعاجم قبله من بداية الخليل قد قسموا المادة تقسيما (كميا)، فقام بتصنيف المادة تصنيفا يخدم اللغوي، قائما على الوظائف النحوية والأنواع الصرفية للكلمات، وجعل هذا معجمه معجما منظما وخادما للغة وكذلك فهو دقيق جدا. راجع المرجع السابق: د/محمد يوسف حبلص.

[49] المؤلف: أبو النصر إسماعيل بن حماد الجوهري (332–393 هـ). كيفية البحث في المعجم: الحرف الأخير (باب)، والحرف الأول (فصل)، وما يثلثهما. منهج الكتاب: قسم معجمه إلى 28 بابا لكل حرف من حروف الألفبائية باب، وقسم كل باب إلى 28 فصلا لكل حرف من حروف الألفبائية فصل عدا (الواو والياء فقد جعلهما بابا واحدا. التزم بفكرة الجذر وجعل لام الكلمة بابا وفاءها فصلا. مميزات المعجم: –أراد أن يجمع في معجمه ما صح عنده من رواية ودراية، ومشافهة عن أصحاب الأصلاء، ومن هنا سمي معجمه الصِّحاح. – كان ينص على ضبط الكلمات تجنبا للتصحيف. –عنى بالإشارة إلى الرديءوالمتروك والمذموم تفريقا لها من الصحاح، وعنى بالإشارة إلى النوادر والمعرب والمولد والمشترك والأضداد واللهجات. –عنى بذكر كثير من مسائل النحو والصرف وفقه اللغةوآراء العلماء ومناقشاتهم فيها. راجع المرجع السابق: د/محمد يوسف حبلص.

درجة الشيوع: **

نسبة الشيوع: 0.02

المعنـى المعجمـي: (عَلاَ) الشيءُ ُ عُلُوًّا: ارتفع. فهو عالٍ، وعليٌّ. ويقال: عَلا النهارُ. ويقال: عَلا فلان في الأرض: تكبر وتجبَّر. وفي التنزيل العزيز :إِنَّ فِرْعَوْنَ عَلاَ فِي الأَرْضِ. و– فلان بالأمر: اضطلع به واستقل. و– بالشيء: جعله عليًّا. و– الشيءَ، وعليه، وفيه–: رَقِيَه وصَعِده. و– الرجلَ قَهَره وغلبه. و– بالسيف: ضَرَبَه. و– فلانٌ حاجتَه: ظَهَرَ عليها[50]. (عَلا) عُلُوًّا، فهو عَلِيٌّ، وعَلا النَّهارُ: ارْتَفَعَ[51]. (عَلا) في المكان يَعْلو عُلوًّا. وعَلِيَ في الشرف بالكسر يَعْلى عَلاءً. ويقال أيضاً: عَلا بالفتح يَعْلى[52].

الوحـدات المعجميـة: {الأعلى، أعلى، اعتلى، العالي، العليا، عالي، على، علي، علينا، عليه، عليها، عليهم، وتعالت، وتعالى، وعلى}.

* الأعلى: ال التعريف+اسم تفضيل (أعلى). (11) (0.2).

المعنـى المعجمـي: (أعْلَى) عن الشيء: نَزَلَ عنه. يقال: أَعلى عن الدابة، إذا نزل عنها. و– الشيءَ: رَفَعَهُ وجعلَهُ عاليا. و– الشيء: صعِده[53].

الشـاهد: وشدد "المجلس الأعلى القوات المسلحة" في بيان بثه على صفحته الرسمية على موقع التواصل الاجتماعي.

* أعلى: ظرف مكان. (2) (0.02).

الشاهد: وأكد الدكتور محمد البلتاجي من أعلى المنصة (...).

* اعتلى: فعل ماض. (1) (0.02).

الشاهد:في حين اعتلى البعض أسطح المباني.

* العالي: ال التعريف+اسم فاعل (عالي). (1) (0.02).

المعنى المعجمي: (العَالِي) يقال: فلانٌ عالي الكعب: شريفٌ. وأَتيته من عالٍ: من فوق[54].

[50] **الوسيط (علا).**

[51] **المحيط باب (الواو) فصل (العين).**

[52] **الصحاح باب (الهمزة) فصل (العين). (علا).**

[53] **الوسيط (علا).**

الشاهد:تقرر زيادةكميات المياه المنصرفة خلف السد العالي لتلبية الاحتياجات القومية من الزراعة والصناعة ومياه الشرب.

* العليا: ال التعريف+صفة مشبهة. (7) (0.1).

المعنى المعجمي: (العُلْيَا): مؤنث الأعلى. وفي الحديث: اليَدُ العُليا خير من اليد السُّفلى. (ج) عُلًى[55].

الشاهد: وأن الكلمة العليا هي للشعب عبر صناديق الاقتراع.

* عالي: اسم فاعل. (1) (0.02).

الشاهد: وهناك تنسيق عالي المستوى بين الشرطة والقوات المسلحة لتأمين مداخل ومخارج المحافظات.

* علي: اسم علم. (1) (0.02).

الشاهد: أما المهندس علي عبدالفتاح القيادي البارز بالجماعة.

* على: حرف جر. (77) (0.7).

المعنى المعجمي: وعَلَى: حَرْف، وعن سِيبَوَيْهِ: اسْمٌ للاسْتِعلاءِ: {وعَلَيْها وعلى الفُلْكِ تُحْمَلُونَ}[56]،[57].

الشاهد: مرسي في كلمة للأمة: أنا رئيس لكلالمصريين .. الثورة مستمرة ودماء الشهداء لن تضيع هدرا .. و سنحافظ علي المعاهدات والمواثيق الدولية.

* وعلى: حرف عطف (الواو)+حرف جر (على). (2) (0.02).

الشاهد: فأنا أحب هؤلاء وأقدر دورهم وأحرص على تقويتهم وعلي الحفاظ عليهم وعلي المؤسسة العريقة التي نحبها ونقدرها جميعا.

* علينا: حرف جر (على)+ضمير المتكلمين (نا). (1) (0.02).

الشاهد: وأن تغلب علينا طباعُنا الأصيلة التي اكتسبناها من حضارتنا عبر آلاف السنين.

* عليه:حرف جر (على)+هاء الغائب. (3) (0.02).

[54] **الوسيط (علا).**

[55] **الوسيط (علا).**

[56] **الآية 22 من سورة المؤمنون**

[57] **المحيط، باب (الواو) فصل (العين).**

الشـاهد: مؤكدا أن مقدرات مصر هي ملك الشعب ويجب عليه حمايتها من العبث من أطراف خارجين على القانون أو المستغلين للأوضاع الداخلية.

* عليها: حرف جر (على)+هاء الغائبة. (3) (0.02).

الشاهد: كما أكد مرسي أن مصر قادرة على الدفاع عننفسها وأن تمنع أي عدوان عليها أو على أراضيها.

* عليهم: حرف جر (على)+هاء الغائبين. (2) (0.02).

الشـاهد: وفي الحوامدية حاول الأهالي إرهاب الملاحظين باللجنة عن طريق قذف الحجارة عليهم كي يقوموا بتسهيل الغش وعدم التشدد مع الطلاب.

* وتعالت: حرف عطف (الواو)+فعل ماض (تعالى)+تاء التأنيث. (2) (0.02).

الشاهد: وتعالتصيحات الفرحة بفوز محمد مرسي كأول رئيس للجمهورية بعد ثورة 25 يناير.

* وتعالى: حرف عطف (الواو)+فعل ماض (تعالى). (1) (0.02).

الشاهد: (...) تحية خالصة من قلبي لهم، وحب لا يعلمه في قلبي إلا الله سبحانه وتعالى (...).

وبناء على التصنيف السابق فبدلا من أن يتحمل متعلم اللغة العربية عبء حفظ ومعرفة كل ألفاظ ومفردات اللغة، يمكنه فقط الاهتمام بأشهر ألفاظ اللغة، ثم يأتي تعلم باقي الألفاظ والمفردات بالتدريج والممارسة.

كذلك فترتيب كلمات اللغة في مجموعات مرتبة من الأشهر إلى الأقل شهرة يعد مرجعًا لمستخدمي اللغة في مختلف المجالات، بدلا من تحمل العبء في البحث عن الكلمة المعبرة عن المقصود وتكون في الوقت نفسه معلومة ومفهومة للمتلقي، وبذلك يمكن توفير الجهد والوقت المبذول لانتقاء الكلمات والمفردات. أما في مجال الترجمة الآلية فيمكن استخدام تلك القوائم وإدخالها إلى البرنامج المترجم في صورة حقول يستدل بها على أكثر المفردات شيوعًا ثم يقوم باستدعاء أشهر معانيها.

خامسًا: نتائج الدِّراسة

كان الباعثُ على إعداد هذه الدراسة عدم وجود دراسات عربية حول المعاجم التكرارية للغة العربية، وإن كان هناك دراسات تناولت التأصيل النظري لفكرة المعجم التكراري، مستهدفة وضع نموذج للمعجم التكراري يساهم في تيسير عملية التعليم والتعليم.

كما هدفت الدراسة إلى الإسهام في مضمار تعليم اللغة وتطويره على المستويين في التعامل مع أبنائها أو الواردين عليها من غير الناطقين بالعربية.

وتأتي نتائج الدراسة في توفيرها المخرجات التالية:

1. إنتاج معجم تكراري للغة العربية المعاصرة، يقوم بحصر أشهر مفردات اللغة العربية وترتيبها في قائمة يسهل الوصول إليها والبحث فيها، مما يوفر على المعلم والمتعلم الإفادة من مخرجات هذا المعجم، بصورة تيسر العملية التعليمية.

2. تفعيل دور المدونات اللغوية الحاسوبية وفق أصوله العلمية النموذجية في تعليم اللغة العربية ونشرها لدى دارسي اللغة العربية من غير الناطقين بها.

3. إمداد معلمي اللغة العربية وباحثيها بما يساعدهم في عملهم، في مجال الصناعة المعجمية.

4. إيضاح الفرْق بين تعلم اللغة العربية بالطريقتين التقليدية القديمة والطريقة الحاسوبية الجديدة.. وأيهما أسهل وأتم.

5. إنتاج مدونة معجم تكراري تحصر ألفاظ العربية المعاصرة، وتكون نواة للمعجم المنشود، مع ترتيبها من الأكثر شيوعًا إلى الأقل مع ذكر نسب وجودها وتكرارها؛ مما يسهم في مجال تعليم اللغة العربية لغير الناطقين بها.

6.الوقوفعلىسبللتطويردورمدونات المعاجم لتفعيل نظم تعلم اللغة العربية.

7.مناقشة التطبيقات العملية لنتائج البحث ومحاولة التوصل إلى بعض التوصيات التي قد تؤدي إلى فاعلية تطبيق مدونات المعاجم في تعلم اللغة العربية.

8.إبراز دور المدونات اللغوية الإلكترونية في خدمة المعجم العربي، ومجال تأليف المعاجم.

سادسًا: الشكر

بِفَيض فَضلِكَ يَحيا العلمُ وَالأَدَبُ وَبِاِسمكَ اليَومَ أَضحَت تَفخرُ الكُتُبُ
فَمِن سَنا فِكْرِكَ التَهذيبُ مُنتَشِرٌ وَمِن ضِيا فَهمِكَ الإرشادُ مُنسَكِبُ

خليل الخوري

إلى مَن نَهَلْتُ مِنْ بَحْرِ عِلْمِهِمَا الفيَّاضِ..

إلى السِّرَاجَيْنِ المُضِيئَيْنِ اللذين نِلْتُ شَرَفَ تَوْجِيهَاتِهِمَا لدِرَاسَتِي..

إلى أسْتَاذَيَّ الجَلِيلَيْنِ...

الأستاذ الدكتور إبراهيم الدسوقي

الدكتور المعتز بالله السعيد

فَائِقُ احترَامِي وتَقْدِيرِي

سابعًا: قائمةُ المراجع

أولا: المراجع العربية

- موسى بريل "قاموس الصحافة العربية"، القدس، الجامعة العبرية 1940.
- (عبده) داود عطية: "المفردات الشائعة في اللغة العربية: دراسة في قوائم المفردات الشائعة في اللغة العربية" 1979م.
- (عبد الكريم) جمعان، تطبيقات استعمال لسانيات المدونات في إصدار معجم تأريخي للغة العربية، بحث منشور في مجلة "دراسات مصطلحية"، مؤسسة البحث والدراسات العلمية، العدد التاسع والعاشر، 1431-1432هـ/2009-2010م.
- (السعيد) المعتز بالله "المدونات اللغوية" من كتاب: مدخل إلى علم اللغة الحاسوبي-تحت النشر".
- (عمر) أحمد مختار "صناعة المعجم الحديث"، ط1، 1418هـ/1998م، عالم الكتب-القاهرة.
- (عبد العزيز) محمد حسن "لغة الصحافة المعاصرة"، دار الفكر العربي-القاهرة، الطبعة الأولى 2002م.
- د.جلال الصياد، د.عبد الحميد محمد ربيع، "مبادئ الطرق الإحصائية"، ط1، 1983م، تهامة/جدة/المملكة العربية السعودية.
- (أحمد) غريب محمد "مدخل إلى الإحصاء"، برنامج دراسة المجتمع، مركز التعليم المفتوح، جامعة بنها، المستوى الأول، 2012.
- جريدة الأهرام المصرية، البوابة الإلكترونية: ليوم الاثنين 5 من شعبان 1433هـ/ 25 يونيو 2012 / السنة 136/ العدد 45857 .
- راجع د/محمد يوسف حبلص: "معاجم العربية ومصادرها"، دار العلوم-القاهرة، دار الهاني ، ط2، 2007م.
- المعجم الوسيط، مجمع اللغة العربية، ط4، 1425هـ/2004م، مكتبة الشروق الدولية.
- الصحاح في اللغة، (الجوهري) إسماعيل بن حماد، تحقيق: أحمد عبد الغفور عطار، دار العلم للملايين، ط4، 1990.
- القاموس المحيط، للفيروز أبادي، ط3، المطبعة الأميرية 1301هـ. الهيئة المصرية العامة للكتاب.

ثانيا: المراجع الأجنبية

- EDWARD L. THORNDIKE "The Teacher's Word Book" NEW YORK, 1921, TEACHER'S COLLEGE, COLUMBIA UNIVERSITY.
- Buckwalter , T. & Parkinson, D. (2011). A Frequency Dictionary of Arabic: Core Vocabulary for Learners.
- Martin Thomas, “Electronic Text”, Electronic Text Notes - tx.xml,2002
- Bolshakov , I. and Gelbukh, A(2004). Computational Linguistics "Models, Resources, Applications"

ثالثًا: الأوراق والأطروحات البحثية

- (السعيد) المعتز بالله: "المعجم التكراري لألفاظ القرآن الكريم (المنهج والنموذج)"، ندوة القرآن الكريم والتقنيات المعاصرة (تقنية المعلومات)، المملكة العربية السعودية، مجمع الملك فهد لطباعة المصحف الشريف، المدينة المنورة، 2009.
- السعيد) المعتز بالله "مدونة معجم عربي معاصر: معالجة لغوية حاسوبية"، أطروحة لنيل درجة الماجستير بكلية دار العلوم، القاهرة، 2008.
- د.محمود إسماعيل صالح ود.إبراهيم الخراشي: الجانب اللغوي للمعجم الحاسوبي للغة العربية، ضمن البحوث المقدمة في الاجتماع الثاني لخبراء المعجم الحاسوبي التفاعلي، الرياض، 2008م.
- (السعيد) المعتز بالله "مدونة معجم تاريخي للغة العربية المعاصرة.. دارسة لغوية حاسوبية"، أطروحة دكتوراه، 2010.

السيرة الذاتية

- باحث ماجستير – قسم علم اللغة والدراسات السامية والشرقية – كلية دار العلوم – جامعة القاهرة.
- حصلت على الدبلوم العام في التربية – كلية التربية – جامعة عين شمس.

- حصلت على دورات "مهارات فن الإعراب"، "مهارات التصحيح اللغوي"، من مركز التدريب اللغوي بكلية دار العلوم، كما حصلت دورات "المهارات اللغوية"، "العروض وتذوق الشعر" من كلية دار العلوم برعاية مؤسسة البابطين.

ICA: The International Corpus of Arabic

Sameh Alansary [*1], Magdy Nagi [**2]

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

***Computer and System Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt*
[2]magdy.nagi@bibalex.org

***Abstract*—This paper sheds light upon the current state of Arabic corpora. Unfortunately, Arabic lackes corpora and corpora tools to affect the quality of Arabic language applications. There are rarely successful Arabic corpora trails in compilation and analysis. So, Bibliotheca Alexandrina (BA) has initiated a big project to build an Arabic corpus in the manner of the International corpus of English (ICE) to be a newly established representative corpus of modern standard Arabic (MSA) that is intended to cover the Arabic language as being used all over the Arab world. ICA was planned to contain 100 million analyzed words with a web query system which allows users to interact with data [ICA website].**

1 INTRODUCTION

As language and linguistics studies cannot rely on intuition or small samples of language, they require empirical analysis of large database of texts as in the corpus-based approach, because of the importance of corpora to language and linguistics studies is aligned to the importance of empirical data. Corpus-based methods can be used to study a wide variety of topics within linguistics. Because Corpora consist of texts, they enable the linguists to contextualize their analyses of language; corpora are very well suited to more functionally based discussions of language and linguistics. Fortunately, modern computers have made it possible to store a large number of texts and to analyze a large number of linguistic features in those texts.

However, Linguists of all persuasions have discovered that corpora can be very useful resources for pursuing various resources of research agendas, due to the importance of corpus in language and linguistic studies such as corpus in lexicography, grammar, semantics, natural Language Processing and other language studies [1].

Due to the increasing need for an Arabic corpus to represent the Arabic language and because the trials to build an Arabic corpus in the last few years were not enough to consider that the Arabic language has a real, representative and reliable corpus, it was necessary to try to build an Arabic corpus that could support the various linguistic research on Arabic. Thus, this work was inspired by the difficulties that encountered Arabic Language researches because of the lack of publicly available Arabic corpora.

Arabic is the largest member of the Semitic language family, most closely related to Aramaic, Hebrew, Ugaritic and Phoenician. Arabic is one of the six official languages of the United Nations[1] and it is the main language of most of the Middle East countries. Arabic ranks fifth in the world's league table of languages, with an estimated 206 million native speakers, 24 million as 2nd language speakers to add up to total of 233 million, whereas World Almanac estimates the total speakers as 255 million. Arabic language is the official language in all of the Arab nations as Egypt, Saudi Arabia and Algeria. Moreover, it is also an official language in non-Arab countries as Israel, Chad and Eritrea. It is also spoken as a 2nd language in other non-Arab countries as Mali and Turkey[2].

The formal Arabic language, known as Classical Arabic is the language in which the Qur'an is written and is considered to be the base of the syntactic and grammatical norms of the Arabic language. However, today it is considered more of a written language than a spoken one [2]. Modern Standard Arabic (MSA) is similar to Classical Arabic, but it is an easier form. It is understood across the Arab world and it is used by television presenters and politicians, it is the form used to teach Arabic as a foreign language. There are different MSA varieties as the rate of similarity between every Arab country version of MSA and Classical Arabic differs. This is one of the issues that this paper will present.

Trials have been conducted to build Arabic corpora, but unfortunately some of them were unsuccessful trials and others were for commercial purposes only. Some of these trials are: KACST Arabic Corpus[3] that is neither analyzed, nor well planned; it also contains a lot of classical Arabic texts [3], CLARA corpus that contains only 15,000 analyzed words [4],

[1] http://www.un.org/en/aboutun/languages.shtml
[2] http://www.un.org/en/aboutun/languages.shtml
[3] http://www.kacstac.org.sa/pages/About.aspx

Al-Nahar Newspaper Text Corpus[4] that is not analyzed and contains texts from one source only, a Corpus of Contemporary Arabic (CCA Corpus)[5] that contains one million words collected from websites as well as online magazines [5], Penn Arabic Treebank that is analyzed but contains one million words only [6], Arabic Gigaword Corpus which is analyzed but is not accessible for free [7], Nemlar project that contains an Arabic written corpus of 500K words; however, its analysis features are limited, in addition it is not accessible for free, and many other incomplete trails[8].

Bibliotheca Alexandrina (BA) has initiated a big project to build the "International Corpus of Arabic (ICA)", a real trial to build a representative Arabic corpus as being used all over the Arab world to support research in Arabic [9]. In order to build a corpus there are a number of factors which need to be taken into consideration. These factors include size, balance and representativeness [10]. The following sections will present more about ICA structure and how the data was collected.

In what follows, Section 2 shows the ICA data design, how it is compiled, discuss the copyrights issue, the lexical density and variety measures that were applied to the ICA. Section 3 refers to the analysis stage of ICA, ICA tag sets and ICA linguistic information. Section 4 gives a brief review on the ICA website for the researchers to query its data. Section 5 describes the conclusion ICA compilation, analysis and the importance of ICA web site.

2 ICA DESIGN & COMPILATION STAGE

ICA is a general and dynamic corpus appropriate for a variety of uses. The collection of samples is limited to written Modern Standard Arabic, selected from a wide range of sources and designed to represent a wide cross-section of Arabic language; it is stimulating the first systematic investigation of the national varieties as being used all over the Arab world. It is important to realize that the creation of ICA is a "cyclical" process, requiring constant reevaluation during the corpus compilation. Consequently, we are willing to change our initial corpus design if there are any circumstances that would arise that requires such changes.

A. ICA Design

As language is infinite but a corpus has to be finite in size, we sample and proportionally include a wide range of text types to ensure the maximum balance and representativeness. A balanced corpus covers a wide range of text genres which are supposed to be representative of the language or the variety under consideration. ICA genre design relied on Dewey decimal classification of documents; however, this has been further classified to suit clear genre distinction rather than classifications for libraries. For example, Dewey decimal classification [11] combines history and geography in one classification, while in ICA they are separated into two sub genres related to humanities genre. It has been designed to reflect a more or less real picture of how Arabic language exists in every field and in every country rather than relying on a theoretical image.

ICA is planned to contain 100 million words. However, currently it is still around 80 million words. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination [12]. Accordingly, ICA includes 11 genres as shown in table (1). Each genre is classified into 24 sub-genres, including; Politics, Law, Economy, Sociology, Islamic, Pros etc. Moreover, there are 4 sub-sub-genres, namely; Novels, Short Stories, Child Stories and plays.

[4] http://www.elda.org/catalogue/en/text/W0027.html

[5] http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

TABLE I

ICA GENRES

Genre	Number of Texts	% of total written
Humanities	1,001	8%
Strategic Sciences	15,359	13%
Social Sciences	7,585	12%
Natural sciences	179	1%
Applied Sciences	2,363	1%
Religion	2,825	14%
Art & Culture	7,074	5%
Biography	116	2%
Literature	3,423	12%
Sports	12,087	7%
Miscellaneous	18,010	25%

ICA determines the size of every genre based on a questionnaire that has been conducted by the Bibliotheca Alexandrina readers along with some statistical studies that have been made on all the available books, magazines, articles, academics, … etc. ICA design focuses on the number of words. However, issues of size are also related to the number of texts taken from the different genres, the number of samples taken from each text, and the number of words in each sample.

Balance in a corpus has not been addressed by having equal amounts of texts from the different sources or genres. The balance is based on the factual distribution of the language real use. For example, literature genre represents 12% while biography genre represents 2% from the corpus data distribution.

In collecting a corpus that represents the Arabic Language, the main focus was to cover the same genres from different sources and from all around the Arab nations. However, we decided to add Arabic data that belongs to the Arabic language even if they had been published outside the Arab world as al-Hayat magazine which is published in London.

1) Text Compilation and Categorization: ICA is regularly updated to be representative. ICA has been compiled according to a methodology has been developed to enabled the corpus compilers to select all and only the MSA data rather than the colloquial Arabic data. The ICA text categorization has also been done according to the topic of the text, depending on distinct semantic features that have been determined for each genre. These features keep the ICA data categorization objective rather than being subjective; depending on the compiler intuition. Accordingly, ICA texts can be considered as a good training data for text categorization system.

ICA data is composed of Modern Standard Arabic (MSA) written texts. There are different resources for compiling the data. It has been decided to compile all available Arabic data written in MSA. ICA composed of four sources, namely; 1. Press source which is divided into three sub-sources, namely; (a) Newspapers such as Al Ahram from Egypt, Addstour from Jordan, Al Hayat from Lebanon … etc (b) Magazines which had been compiled from the official magazines (c) Electronic Press which had been compiled from magazines and newspapers that are written in MSA and have only soft electronic copy through world wide web. (2) Net articles which were compiled from forums and blogs that are also written in MSA. (3) Books which had been compiled from all the available books that are written in MSA and have a soft copy. (4) Academics which had been compiled from the scientific papers, researchers' thesis, PhDs etc.

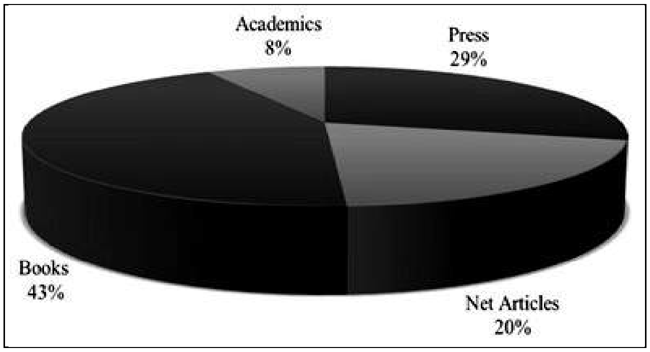


Figure 1: ICA Sources

2) Quantitative Linguistics of ICA: Corpus analysis is both qualitative and quantitative. One of the advantages of corpora is that they can readily provide quantitative data which intuitions cannot be provided reliably. The use of quantification in corpus linguistics typically goes well beyond simple counting, table 2 shows part of the quantitative linguistic analysis for ICA statistics.

TABLE II

QUANTITATIVE LINGUISTIC ANALYSIS FOR ICA STATISTICS

Statistics	**Total Number**
No. of texts	70,022
No. of compiled data	79,569,384
No. of Tokens	76,199,414
No. of Type	1,272,766
No. of ICA sources	4
No. of sub sources	3
No. of genres	11
No. of sub genres	24
No. of sub sub-genres	4
No. of countries	20
No. of covered years	22
No. of writers	1,021

3) ICA Lexical Density: Lexical density is a descriptive parameter which varies according to register and genre to measure the proportion of content words; nouns, verbs, adjectives, and often adverbs, to the total number of words. Written texts have a higher lexical density than spoken texts. Lexical density is determined thus:

Lexical density = (Total Number of lexical words/Total Number of words) * 100

A high lexical density indicates a large amount of information-carrying words and a low lexical density indicates relatively few information-carrying words.

Fig. 2 includes a comparison of the lexical density of Books and Net Articles sources. Random sample was selected from ICA data consisting of 4 million words (2540 texts represent net articles and 154 texts represent books). It has been found that books have high density of information and that net articles have low density of information, because book authors have a much longer time to plan and shape the units of meaning that he or she wishes to use. There is sufficient time to select the most appropriate lexical word, review the text and replace words before one makes the text available. Lexical density, then, can serve as a useful measure of how much information there is in a particular text [13].

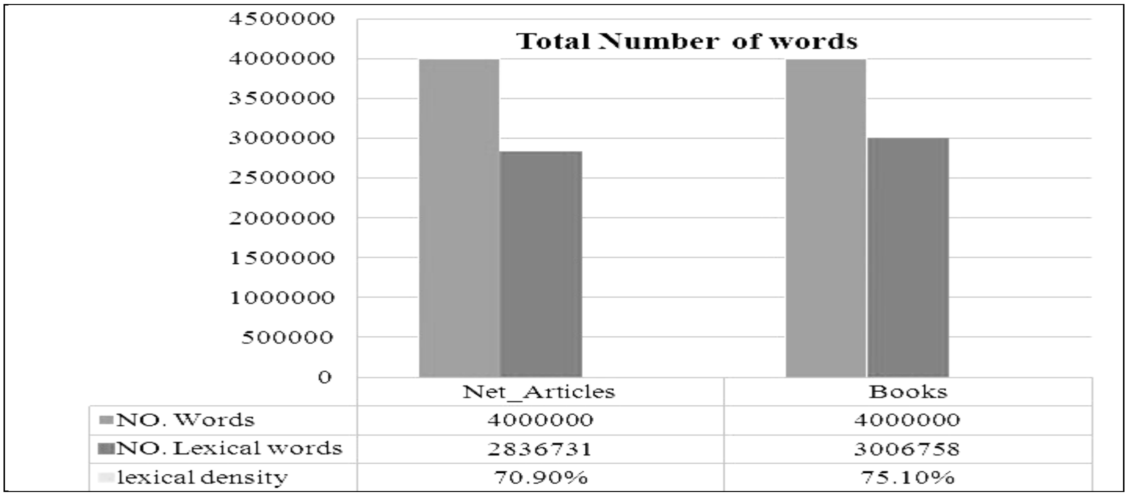
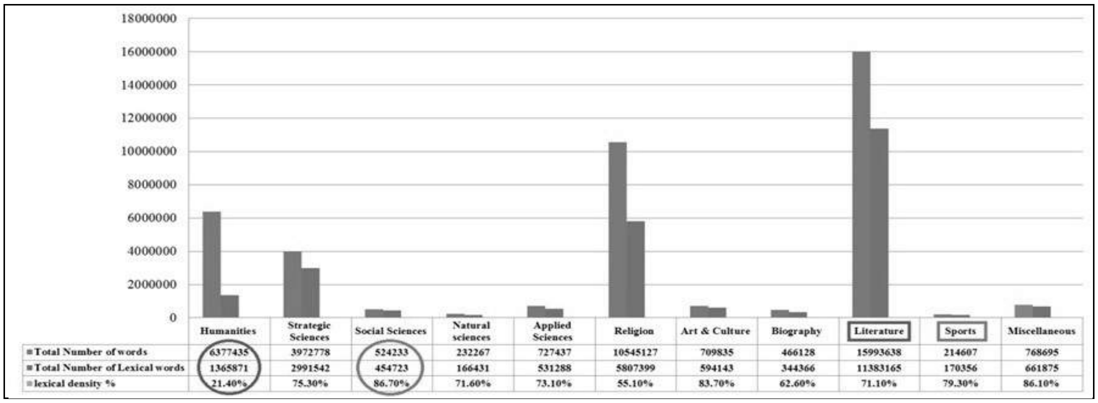


Figure 2: Total number of words and the lexical density of each Source

Fig. 3 shows a distribution of 40,532,180 words among ICA genres, it has been found that the genre that has the largest number of words is the Literature genre and the smallest is Sports. It also includes the lexical density of each genre; it has been found that Social Sciences has high density of information while Humanities has a low density of information.

Humanities is a main genre that includes 4 sub genres (History, Psychology, Philosophy, and Geography) all of which are fields that are purely scientific, resulting in low richness of lexical words.



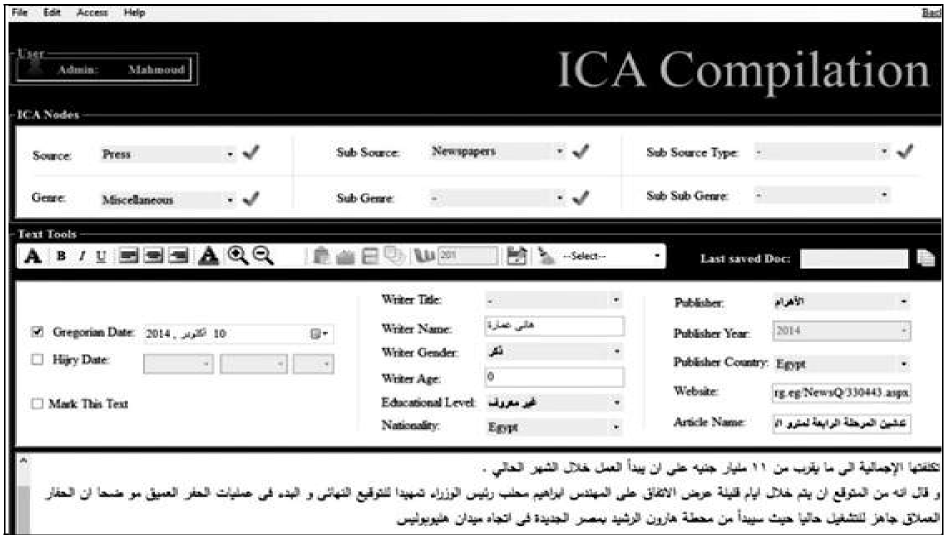
	Humanities	Strategic Sciences	Social Sciences	Natural sciences	Applied Sciences	Religion	Art & Culture	Biography	Literature	Sports	Miscellaneous
■ Total Number of words	6377435	3972778	524233	232267	727437	10545127	709835	466128	15993638	214607	768695
■ Total Number of Lexical words	1365871	2991542	454723	166431	531288	5807399	594143	344366	11383165	170356	661875
■ lexical density %	21.40%	75.30%	86.70%	71.60%	73.10%	55.10%	83.70%	62.60%	71.10%	79.30%	86.10%

Figure 3: Total number of words and the lexical density of each Genre

4) *Copyrights:* One of the serious constraints on developing large corpora and their widespread use is national and international copyright legalizations. According to copyright laws, it is necessary and sensible to protect the authors as well as the publishers' rights of the texts that they had produced. ICA data Copy rights and publishing issues are in progress by Bibliotheca Alexandrina Legal department. For that reason, the ICA data is not available to be downloaded, but the researchers can search the ICA data via the ICA website.

5) *ICA Utilities:* Good corpus needs to meet two major requirements to achieve its goal which are well planning and a huge amount of data[6]. Therefore, we have an application that helps us through the compilation and categorization stage. It is more like a utility containing multiple tools that cover the whole process, from collecting the data, categorizing it into its correct categories to reporting any information or even changing its categorization.

Compiler adds all the available information such as Website, date of publishing, publisher (name and country), writer (name, gender, age, nationality and educational level) which can be very useful to other studies such as sociolinguistics, the documents are stored in UTF format as a txt file to guarantee its compatibility with all platforms, as shown in Fig. 4.



Admin: Mahmoud

Source: Press ✓ Sub Source: Newspapers ✓ Sub Source Type: - ✓

Genre: Miscellaneous ✓ Sub Genre: - ✓ Sub Sub Genre: -

☑ Gregorian Date: 2014 , أكتوبر 10
☐ Hijry Date:
☐ Mark This Text

Writer Title: -
Writer Name: هاني عمارة
Writer Gender: ذكر
Writer Age: 0
Educational Level: غير معروف
Nationality: Egypt

Publisher: الأهرام
Publisher Year: 2014
Publisher Country: Egypt
Website: rg.eg/NewsQ/330443.aspx
Article Name: تدشين المرحلة الرابعة لمترو ال

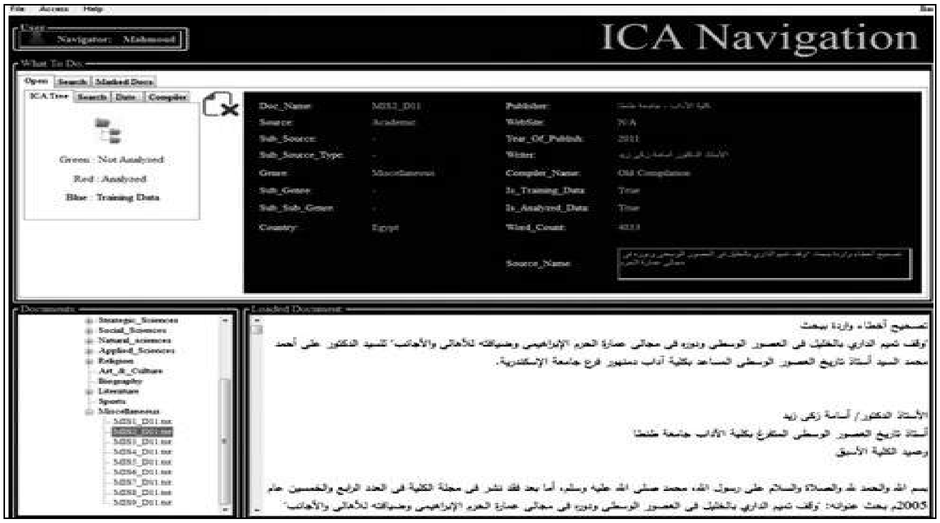
تكلفتها الإجمالية الى ما يقرب من ١١ مليار جنيه على ان يبدأ العمل خلال الشهر الحالي .
و قال انه من المتوقع ان يتم خلال ايام قليلة عرض الاتفاق على المهندس ابراهيم محلب رئيس الوزراء تمهيدا للتوقيع النهائي و البدء في عمليات الحفر العميق مو ضحا ان الحفار
العملاق جاهز للتشغيل حاليا حيث سيبدأ من محطة هارون الرشيد بمصر الجديدة في اتجاه ميدان هليوبوليس

Figure 4: Compilation phase

With the huge amounts of data, navigation tool comes right in place, it enables the user to: A. review the documents in the form of a tree that simulates the ICA hierarchy. Moreover, the compilation date, compilers names or any loaded document can be deleted from the hierarchy and moved to recycling, B. search inside the raw data, the search results will include documents that contain the search input and the context it appears in, C. change the document categorization

[6] http://en.wikipedia.org/wiki/Language_planning

from one source to another or edit the metadata of the document as the complier may have made some errors during the compilation stage, D. choose between three different types of search options: Exact match, Wildcard and Regular Expression, as shown in Fig. 5.



Doc_Name: MSS2_D01
Source: Academic
Sub_Source: -
Sub_Source_Type: -
Genre: Miscellaneous
Sub_Genre: -
Sub_Sub_Genre: -
Country: Egypt
Publisher: كلية الآداب - جامعة طنطا
WebSite: N/A
Year_Of_Publish: 2011
Writer: الأستاذ الدكتور أسامة زكي زيد
Compiler_Name: Old Compilation
Is_Training_Data: True
Is_Analyzed_Data: True
Word_Count: 4033
Source_Name: تصحيح أخطاء واردة ببحث "وقف تميم الداري بالخليل في العصور الوسطى ودوره في مجالي عمارة الحرم

تصحيح أخطاء واردة ببحث
"وقف تميم الداري بالخليل في العصور الوسطى ودوره في مجالي عمارة الحرم الإبراهيمي وضيافته للأهالي والأجانب" للسيد الدكتور علي أحمد
محمد السيد أستاذ تاريخ العصور الوسطى المساعد بكلية آداب دمنهور فرع جامعة الإسكندرية.

الأستاذ الدكتور/ أسامة زكي زيد
أستاذ تاريخ العصور الوسطى المتفرغ بكلية الآداب جامعة طنطا
وعميد الكلية الأسبق

بسم الله والحمد لله والصلاة والسلام على رسول الله، محمد صلى الله عليه وسلم، أما بعد فقد نشر في مجلة الكلية في العدد الرابع والخمسين عام
2005م بحث عنوانه: "وقف تميم الداري بالخليل في العصور الوسطى ودوره في مجالي عمارة الحرم الإبراهيمي وضيافته للأهالي والأجانب"

Figure 5: Navigation

3 CORPUS ANALYSIS STAGE

Corpus linguistics is the analysis of naturally occurring language on the basis of computerized corpora. Usually, the analysis is performed with the help of the computer [14].

B. ICA Tag Sets

Many natural language expressions are ambiguous, and need to draw on other sources of information to be interpreted. To interpret words to be able to discriminate between different usages is needed [15].

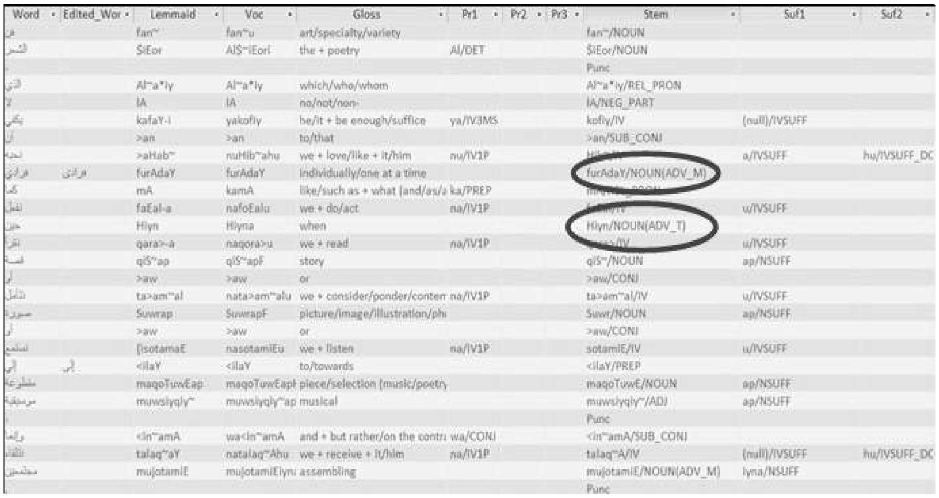
Part-Of-Speech tagging is the process by which a specific tag is assigned to each word of a sentence to indicate the function of that word in the specific context [16]. Traditional Arabic grammar defines a detailed part-of-speech hierarchy which applies to both words and morphological segments. Fundamentally, a word may be classified as nominal, verb or a particle [17]. Arabic is very rich in categorizing words, and contains classes for almost every form of word imaginable. For example, there are classes for nouns of instruments, nouns of place and time, nouns of activity and so on [18].

ICA tag sets consist of 6 categories with 46 types of tags. The tag sets of the VERB category contain 5 tags; Command Verb, Imperfect Verb, Imperfect Passive Verb, Past Verb and Past Passive Verb. NOUN and ADJECTIVE category consists of 10 tags; Adjective, Noun, Adverb of Manner, Adverb of Place, Adverb of Time, Verbal Noun, Proper Noun, Proper Noun(Adverb of Time), Proper Noun(Interjection) and Number. PRONOUN category consists of 3 tags; Demonstrative Pronoun, Pronoun and Relative Pronoun. PARTICLE consists of 10 tags; Focus Particle, Future Particle, Interrogative Particle, Negative Particle, Particle, Verb Particle, Exception Particle, Conjunction, Interjection Particle and Sub Conjunction. META-INFORMATION category consists of 12 tags. Non-linguistic-information category consists of 4 tags. Finally, it is the link category which consists of 2 tags.

Noun category contains:

- (ADV_T): Added right after the basic word class of the nouns. It refers to Adverbs which describe time.
- (ADV_P): Added right after the basic word class of the nouns. It refers to Adverbs which describe place.
- (ADV_M): Added right after the basic word class of the nouns. It refers to Adverbs which describe manner.

Fig. 6 is an example of NOUN (ADV) in ICA.



Word	Edited_Wor	Lemmaid	Voc	Gloss	Pr1	Pr2	Pr3	Stem	Suf1	Suf2
فن		fan~	fan~u	art/specialty/variety				fan~/NOUN		
الشعر		$iEor	Al$~iEori	the + poetry	Al/DET			$iEor/NOUN		
.								Punc		
التي		Al~a*iy	Al~a*iy	which/who/whom				Al~a*iy/REL_PRON		
لا		lA	lA	no/not/non-				lA/NEG_PART		
يكفي		kafaY-i	yakofiy	he/it + be enough/suffice	ya/IV3MS			kofiy/IV	(null)/IVSUFF	
أن		>an	>an	to/that				>an/SUB_CONJ		
نحبه		>aHab~	nuHib~ahu	we + love/like + it/him	nu/IV1P			Hib~/IV	a/IVSUFF	hu/IVSUFF_DO
فرادى	فرادى	furAdaY	furAdaY	individually/one at a time				furAdaY/NOUN(ADV_M)		
كما		mA	kamA	like/such as + what (and/as/a	ka/PREP			mA/REL_PRON		
نفعل		faEal-a	nafoEalu	we + do/act	na/IV1P			foEal/IV	u/IVSUFF	
حين		Hiyn	Hiyna	when				Hiyn/NOUN(ADV_T)		
نقرأ		qara>-a	naqora>u	we + read	na/IV1P			qora>/IV	u/IVSUFF	
قصة		qiS~ap	qiS~apF	story				qiS~/NOUN	ap/NSUFF	
أو		>aw	>aw	or				>aw/CONJ		
نتأمل		ta>am~al	nata>am~alu	we + consider/ponder/conten	na/IV1P			ta>am~al/IV	u/IVSUFF	
صورة		Suwrap	SuwrapF	picture/image/illustration/ph				Suwr/NOUN	ap/NSUFF	
أو		>aw	>aw	or				>aw/CONJ		
نستمع		{isotamaE	nasotamiEu	we + listen	na/IV1P			sotamiE/IV	u/IVSUFF	
إلي	إلى	<ilaY	<ilaY	to/towards				<ilaY/PREP		
مقطوعة		maqoTuwEap	maqoTuwEapl	piece/selection (music/poetry				maqoTuwE/NOUN	ap/NSUFF	
موسيقية		muwsiyqiy~	muwsiyqiy~ap	musical				muwsiyqiy~/ADJ	ap/NSUFF	
.								Punc		
وإنما		<in~amA	wa<in~amA	and + but rather/on the contr	wa/CONJ			<in~amA/SUB_CONJ		
نتلقاه		talaq~aY	natalaq~Ahu	we + receive + it/him	na/IV1P			talaq~A/IV	(null)/IVSUFF	hu/IVSUFF_DO
مجتمعين		mujotamiE	mujotamiEiyna	assembling				mujotamiE/NOUN(ADV_M)	iyna/NSUFF	
.								Punc		

Figure 6: NOUN (ADV) in ICA

C. *Some ICA Linguistic Qualifiers*

The stem-based approach (concatenative approach) has been adopted as the linguistic approach to analyze the ICA.

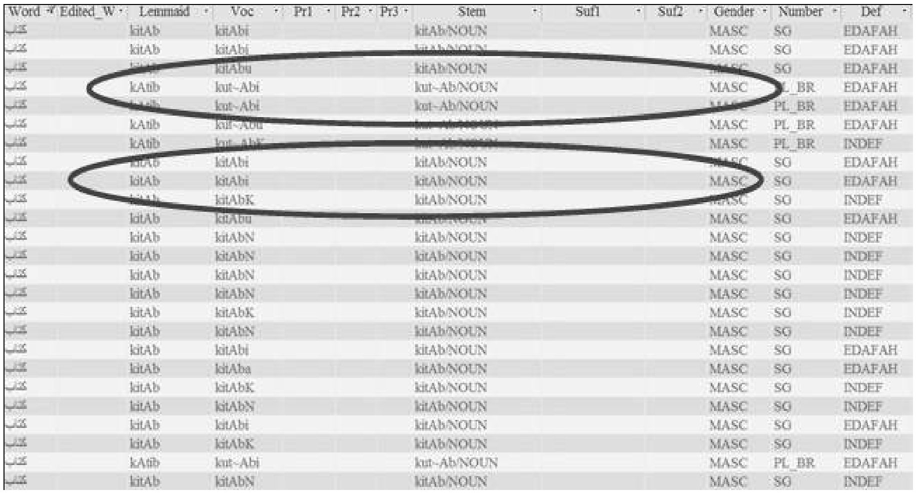
The lexicon includes over 97,000 entries, belonging to 17 part-of-speech (POS) categories. Lexically specified information includes: word, vocalization, lemma, prefix(s) (for nouns, adjectives and verbs), suffix(s) (for nouns, adjectives and verbs), gender (for nouns and adjectives), number (for nouns and adjectives), definiteness (for nouns and adjectives), root and pattern (for verbs and nouns), Arabic stem, and case.

Functional Arabic morphology enables the functional gender and number information thanks to the lexicon that can stipulate some properties as inherent to some lexemes, and thanks to the paradigm-driven generation that associates the inflected forms with the desired functions directly [19].

1) *Number:* Traditional Arab grammarians have established that the Arabic plural system consists of two-mode formation; sound plural and broken plural. Broken plural is undeniably considered, both in morphological and phonological circles, as the most complicated system of nominal plurality because of the great number of patterns due to the overall morphological patterning of the language [20].

ICA Number information includes; broken plural "PL_BR" which indicates the irregularity while conducting plural process; for example, "نواب، مجالس، أنحاء".

Fundamentally, if there is ambiguity such as "كتاب" which can be "kut~Ab" that would be featured as "PL_BR", or "kitAb" that would be featured as "SG", the number feature in ICA is determined according to the vocalization of the words.



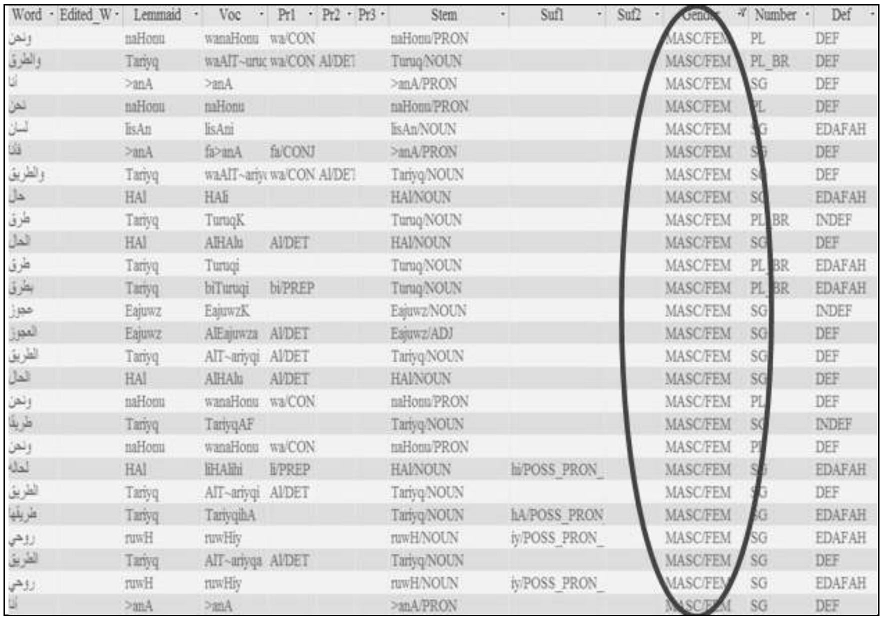
Word	Edited_W	Lemmaid	Voc	Pr1	Pr2	Pr3	Stem	Suf1	Suf2	Gender	Number	Def
كتاب		kitAb	kitAbi				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbi				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbu				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kAtib	kut~Abi				kut~Ab/NOUN			MASC	PL_BR	EDAFAH
كتاب		kAtib	kut~Abi				kut~Ab/NOUN			MASC	PL_BR	EDAFAH
كتاب		kAtib	kut~Abu				kut~Ab/NOUN			MASC	PL_BR	EDAFAH
كتاب		kAtib	kut~AbK				kut~Ab/NOUN			MASC	PL_BR	INDEF
كتاب		kitAb	kitAbi				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbi				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbK				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbu				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbK				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbi				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAba				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbK				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF
كتاب		kitAb	kitAbi				kitAb/NOUN			MASC	SG	EDAFAH
كتاب		kitAb	kitAbK				kitAb/NOUN			MASC	SG	INDEF
كتاب		kAtib	kut~Abi				kut~Ab/NOUN			MASC	PL_BR	EDAFAH
كتاب		kitAb	kitAbN				kitAb/NOUN			MASC	SG	INDEF

Figure 7: ICA Number feature

2) *Gender:* Gender is an unarguable morphosyntactic feature, since it is required for agreement. The realization of the value for gender on the target is the accepted instance of the need for a syntactic rule of agreement [21].

Gender is an inherent feature of nouns, and a contextual feature for any other element that have to agree with the nouns in this feature. Typically, gender is lexically supplied and its value is fixed for the noun. However, in some cases the gender of the nouns can be a semantically selected feature, where one gender value is selected from a set of options. Therefore, the lexical entries of nouns, adjectives and pronouns in a gendered language must specify either that the word has a fixed gender value or that it is capable of taking on different gender values as dictated by the semantics[7].

Some different genders can be chosen to highlight a particular property of the referent. For example, some Arabic words faced in ICA have two genders; masculine, and feminine, as "نحن، حال، طريق". This issue is solved by adding masculine/feminine "MASC/FEM" feature in the column of Gender.



Word	Edited_W	Lemmaid	Voc	Pr1	Pr2	Pr3	Stem	Suf1	Suf2	Gender	Number	Def
ونحن		naHonu	wanaHonu	wa/CON			naHonu/PRON			MASC/FEM	PL	DEF
والطرق		Tariyq	waAlT~uruc	wa/CON	Al/DET		Turuq/NOUN			MASC/FEM	PL_BR	DEF
أنا		>anA	>anA				>anA/PRON			MASC/FEM	SG	DEF
نحن		naHonu	naHonu				naHonu/PRON			MASC/FEM	PL	DEF
لسان		lisAn	lisAni				lisAn/NOUN			MASC/FEM	SG	EDAFAH
فأنا		>anA	fa>anA	fa/CONJ			>anA/PRON			MASC/FEM	SG	DEF
والطريق		Tariyq	waAlT~ariy	wa/CON	Al/DET		Tariyq/NOUN			MASC/FEM	SG	DEF
حال		HAl	HAli				HAl/NOUN			MASC/FEM	SG	EDAFAH
طرق		Tariyq	TuruqK				Turuq/NOUN			MASC/FEM	PL_BR	INDEF
الحال		HAl	AlHAlu	Al/DET			HAl/NOUN			MASC/FEM	SG	DEF
طرق		Tariyq	Turuqi				Turuq/NOUN			MASC/FEM	PL_BR	EDAFAH
بطرق		Tariyq	biTuruqi	bi/PREP			Turuq/NOUN			MASC/FEM	PL_BR	EDAFAH
عجوز		Eajuwz	EajuwzK				Eajuwz/NOUN			MASC/FEM	SG	INDEF
العجوز		Eajuwz	AlEajuwza	Al/DET			Eajuwz/ADJ			MASC/FEM	SG	DEF
الطريق		Tariyq	AlT~ariyqi	Al/DET			Tariyq/NOUN			MASC/FEM	SG	DEF
الحال		HAl	AlHAlu	Al/DET			HAl/NOUN			MASC/FEM	SG	DEF
ونحن		naHonu	wanaHonu	wa/CON			naHonu/PRON			MASC/FEM	PL	DEF
طريقا		Tariyq	TariyqAF				Tariyq/NOUN			MASC/FEM	SG	INDEF
ونحن		naHonu	wanaHonu	wa/CON			naHonu/PRON			MASC/FEM	PL	DEF
لحاله		HAl	liHAlihi	li/PREP			HAl/NOUN	hi/POSS_PRON_		MASC/FEM	SG	EDAFAH
الطريق		Tariyq	AlT~ariyqi	Al/DET			Tariyq/NOUN			MASC/FEM	SG	DEF
طريقها		Tariyq	TariyqihA				Tariyq/NOUN	hA/POSS_PRON		MASC/FEM	SG	EDAFAH
روحي		ruwH	ruwHiy				ruwH/NOUN	iy/POSS_PRON_		MASC/FEM	SG	EDAFAH
الطريق		Tariyq	AlT~ariyqa	Al/DET			Tariyq/NOUN			MASC/FEM	SG	DEF
روحي		ruwH	ruwHiy				ruwH/NOUN	iy/POSS_PRON_		MASC/FEM	SG	EDAFAH
أنا		>anA	>anA				>anA/PRON			MASC/FEM	SG	DEF

Figure 8: ICA words that have MASC/FEM feature

3) *Definiteness:* The definiteness feature in ICA indicates whether the nominal or the adjectival words are definite, indefinite or definite with EDAFAH (DEF_EDAFAH). The feature of Definite with EDAFAH is added to the word according to the context.

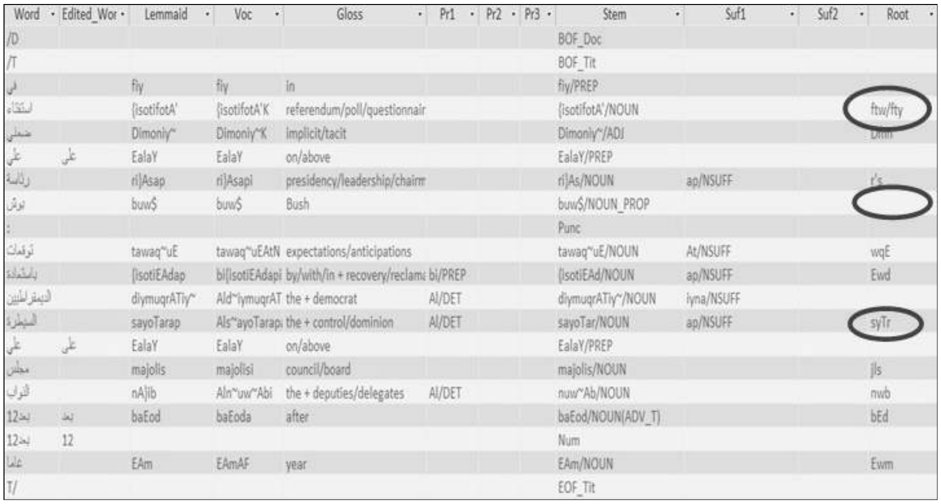
4) *Root:* In Arabic, root is the basic source of all the word forms. The root is not a real word; rather it is a sequence of three consonants that can be found in all the words that are related to it. It is just a sequence of consonants[8]. The root is the primary lexical unit of a word; base word. Moreover, the root of each word is detected according to its lemma.

In ICA, every analyzed noun (including adverbs), adjective and verb has a root related to it even if the root consists of three or four sequenced consonants depending on their lemmas. The ICA data contains 2,435 different distinct roots.

If a specific word has more than one root, it will constitute a problem. This problem is solved by adding the roots separated by "/" in the Root column. Moreover, the root of each word is detected according to its lemma. It is noted that some words have no root as "على، الأمريكي، ديمقراطية"; also some foreign words are used in Arabic orthography such as, "فلوريدا، سوستيه، شارون". Analysts have added the root "NULL" to such words, as shown in Fig. 9.

[7] http://www.features.surrey.ac.uk/features/gender.html

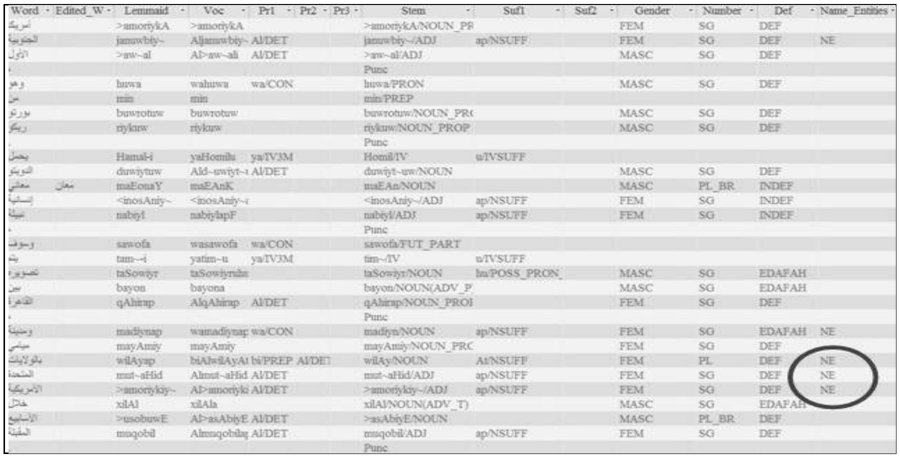
[8] http://arabic.tripod.com/Roots.htm



Word	Edited_Wor	Lemmaid	Voc	Gloss	Pr1	Pr2	Pr3	Stem	Suf1	Suf2	Root
/D								BOF_Doc			
/T								BOF_Tit			
في		fiy	fiy	in				fiy/PREP			
استفتاء		{isotifotA'	{isotifotA'K	referendum/poll/questionnair				{isotifotA'/NOUN			ftw/fty
ضمني		Dimoniy~	Dimoniy~K	implicit/tacit				Dimoniy~/ADJ			Dmn
على	على	EalaY	EalaY	on/above				EalaY/PREP			
رئاسة		ri}Asap	ri}Asapi	presidency/leadership/chairm				ri}As/NOUN	ap/NSUFF		r's
بوش		buw$	buw$	Bush				buw$/NOUN_PROP			
:								Punc			
توقعات		tawaq~uE	tawaq~uEAtN	expectations/anticipations				tawaq~uE/NOUN	At/NSUFF		wqE
باستعادة		{isotiEAdap	bi{isotiEAdapi	by/with/in + recovery/reclama	bi/PREP			{isotiEAd/NOUN	ap/NSUFF		Ewd
الديمقراطيين		diymuqrATiy~	Ald~iymuqrAT	the + democrat	Al/DET			diymuqrATiy~/NOUN	iyna/NSUFF		
السيطرة		sayoTarap	Als~ayoTarap	the + control/dominion	Al/DET			sayoTar/NOUN	ap/NSUFF		syTr
على	على	EalaY	EalaY	on/above				EalaY/PREP			
مجلس		majolis	majolisi	council/board				majolis/NOUN			jls
النواب		nAjib	Aln~uw~Abi	the + deputies/delegates	Al/DET			nuw~Ab/NOUN			nwb
بعد12	بعد	baEod	baEoda	after				baEod/NOUN(ADV_T)			bEd
بعد12	12							Num			
عاما		EAm	EAmAF	year				EAm/NOUN			Ewm
T/								EOF_Tit			

Figure 9: A sample of roots in ICA

5) *Name entity:* This feature is added to words that refer to titles as the title of an institute, ministry, association, country, book, film or conference; it also includes compound names e.g., the compound name 'الولايات المتحدة الأمريكية'.



Word	Edited_W	Lemmaid	Voc	Pr1	Pr2	Pr3	Stem	Suf1	Suf2	Gender	Number	Def	Name_Entities
أمريكا		>amoriykA	>amoriykA				>amoriykA/NOUN_PF			FEM	SG	DEF	
الجنوبية		januwbiy~	Aljanuwbiy~	Al/DET			januwbiy~/ADJ	ap/NSUFF		FEM	SG	DEF	NE
الأول		>aw~al	Al>aw~ali	Al/DET			>aw~al/ADJ			MASC	SG	DEF	
.							Punc						
وهو		huwa	wahuwa	wa/CON			huwa/PRON			MASC	SG	DEF	
من		min	min				min/PREP						
بورتو		buwrotuw	buwrotuw				buwrotuw/NOUN_PR(			MASC	SG	DEF	
ريكو		riykuw	riykuw				riykuw/NOUN_PROP			MASC	SG	DEF	
.							Punc						
يحمل		Hamal-i	yaHomilu	ya/IV3M			Homil/IV	u/IVSUFF					
الدويتو		duwiytuw	Ald~uwiyt~u	Al/DET			duwiyt~uw/NOUN			MASC	SG	DEF	
معاني	معان	maEonaY	maEAnK				maEAn/NOUN			MASC	PL_BR	INDEF	
إنسانية		<inosAniy~	<inosAniy~i				<inosAniy~/ADJ	ap/NSUFF		FEM	SG	INDEF	
نبيلة		nabiyl	nabiylapF				nabiyl/ADJ	ap/NSUFF		FEM	SG	INDEF	
.							Punc						
وسوف		sawofa	wasawofa	wa/CON			sawofa/FUT_PART						
يتم		tam~-i	yatim~u	ya/IV3M			tim~/IV	u/IVSUFF					
تصويره		taSowiyr	taSowiyruhu				taSowiyr/NOUN	hu/POSS_PRON_		MASC	SG	EDAFAH	
بين		bayon	bayona				bayon/NOUN(ADV_P			MASC	SG	EDAFAH	
القاهرة		qAhirap	AlqAhirap	Al/DET			qAhirap/NOUN_PROI			FEM	SG	DEF	
.							Punc						
ومدينة		madiynap	wamadiynap	wa/CON			madiyn/NOUN	ap/NSUFF		FEM	SG	EDAFAH	NE
ميامي		mayAmiy	mayAmiy				mayAmiy/NOUN_PRC			FEM	SG	DEF	
بالولايات		wilAyap	biAlwilAyAt	bi/PREP	Al/DET		wilAy/NOUN	At/NSUFF		FEM	PL	DEF	NE
المتحدة		mut~aHid	Almut~aHid	Al/DET			mut~aHid/ADJ	ap/NSUFF		FEM	SG	DEF	NE
الأمريكية		>amoriykiy~	Al>amoriyki	Al/DET			>amoriykiy~/ADJ	ap/NSUFF		FEM	SG	DEF	NE
خلال		xilAl	xilAla				xilAl/NOUN(ADV_T)			MASC	SG	EDAFAH	
الأسابيع		>usobuwE	Al>asAbiyE	Al/DET			>asAbiyE/NOUN			MASC	PL_BR	DEF	
المقبلة		muqobil	Almuqobilap	Al/DET			muqobil/ADJ	ap/NSUFF		FEM	SG	DEF	
.							Punc						

Figure 10: ICA Name entity (NE)

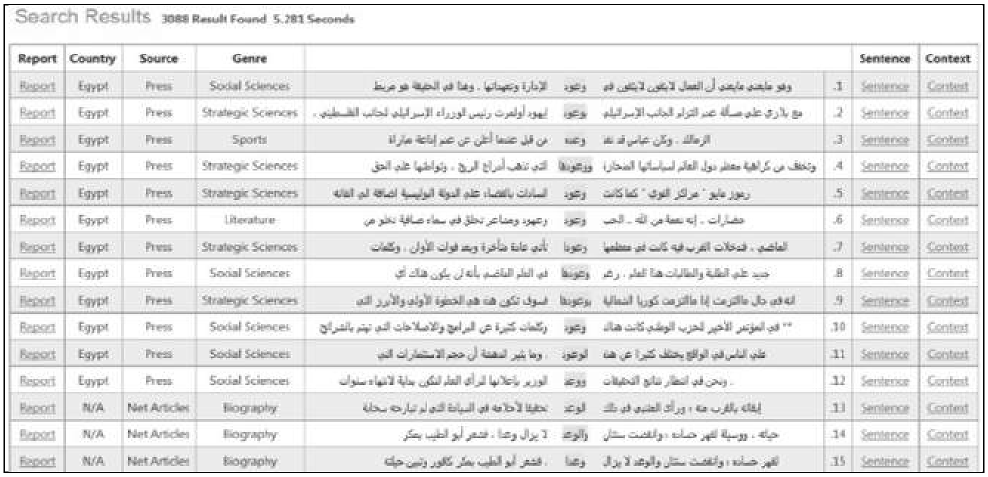
D. *ICA Website*[9]

The ICA data is available to be queried through the ICA website. It is an interface that allows users to interact with the corpus data in a number of ways. The interface provides four options of searching the corpus content; namely, Exact Match Search, Lemma Based Search, Root Based Search and Stem Based Search.

More search options are available; namely, Word Class and Sub Class, Stem Pattern, Number, Definiteness, Gender, Country (Advanced search). Moreover, the scope of search may include the whole corpus, Source(s), Sub-Source(s), Genre(s), Sub-Sub-Genre(s) or Sub-Genre(s).

Fig. 9 presents an example of a query of the analyzed data that states: when the word 'وعد' is searched for using a Lemma-Based search option, the system will highlight all possible lemmas that the word may have, since Arabic is orthographically ambiguous. In this example, the system will highlight several possible lemmas; 'waʕada' 'to promise', 'waʕd' 'Promise' and 'ʕaada' 'return'. If the lemma 'waʕd' 'Promise' is chosen the output search in this case will include all words that have this lemma such as 'وعود' 'Promises', 'alwaʕd'…etc. with all the possible word forms together with concordance lines.

[9] http://www.bibalex.org/ica/en/



Search Results 3088 Result Found 5.281 Seconds

Report	Country	Source	Genre					Sentence	Context
Report	Egypt	Press	Social Sciences	الإدارة وتعهداتها . وهنا في الحقيقة هو مربط	وعود	وهو ماينص مايعني أن العمل لايكون لايكون فيه	.1	Sentence	Context
Report	Egypt	Press	Strategic Sciences	إيهود أولمرت رئيس الوزراء الإسرائيلي لجانب الفلسطيني .	بوعود	مع بلازي على مسألة عدم التزام الجانب الإسرائيلي	.2	Sentence	Context
Report	Egypt	Press	Sports	من قبل عندما أعلن عن عدم إذاعة مباراة	وعده	الزمالك . ولكن عباس قد نقد	.3	Sentence	Context
Report	Egypt	Press	Strategic Sciences	التي تذهب أدراج الريح . وتواطئها على الحق	ووعودها	وتخفف من كراهية معظم دول العالم لسياساتها الضحان)	.4	Sentence	Context
Report	Egypt	Press	Strategic Sciences	السادات بالقضاء على الدولة البوليسية اضافة الى الغائه	وعود	رموز مايو " مراكز القوى " كما كانت	.5	Sentence	Context
Report	Egypt	Press	Literature	وعهود ومشاعر تحلق في سماء صافية تخلو من	وعود	حضارات . إنه نعمة من الله .. الحب	.6	Sentence	Context
Report	Egypt	Press	Strategic Sciences	تأتي عادة متأخرة وبعد فوات الأوان . وكلمات	وعودا	الماضي . فتحالات الغرب فيه كانت في معظمها	.7	Sentence	Context
Report	Egypt	Press	Social Sciences	في العام الماضي بأنه لن يكون هناك أي	وعودنا	جديد على الطلبة والطالبات هذا العام . رغم	.8	Sentence	Context
Report	Egypt	Press	Strategic Sciences	فسوف تكون هذه هي الخطوة الأولى والأرز التي	بوعودها	انه في حال ماالتزمت إذا ماالتزمت كوريا الشمالية	.9	Sentence	Context
Report	Egypt	Press	Social Sciences	وكلمات كثيرة عن البرامج والاصلاحات التي تهتم بالشرائح	وعود	** في المؤتمر الأخير لحزب الوطني كانت هناك	.10	Sentence	Context
Report	Egypt	Press	Social Sciences	. وما يثير الدهشة أن حجم الاستثمارات التي	الوعود	على الناس في الواقع يختلف كثيرا عن هذه	.11	Sentence	Context
Report	Egypt	Press	Social Sciences	الوزير بإعلانها للرأي العام لكن بداية لانتهاء سنوات	ووعد	ونحن في انتظار نتائج التحقيقات .	.12	Sentence	Context
Report	N/A	Net Articles	Biography	تحقيقا لأحلامه في السيادة التي لم تبارحه لحظة	الوعد	إبقائه بالقرب منه ، ورأى المتنبي في ذلك	.13	Sentence	Context
Report	N/A	Net Articles	Biography	لا يزال وعدا . فشعر أبو الطيب بمكر	والوعد	حياته ، ووسيلة لقهر حساده ، وانقضت سنتان	.14	Sentence	Context
Report	N/A	Net Articles	Biography	. فشعر أبو الطيب بمكر كافور وتبين حيلة	وعدا	لقهر حساده ، وانقضت سنتان والوعد لا يزال	.15	Sentence	Context

Figure 11: The lemma 'waʕd' 'Promise' output search

In the search output, there will be some information about the number of search result, country, source, genre, sentence and context for each word. This phase is phase one of ICA website, more enhancements are expected in later phases. The current phase of ICA application does not represent the final release as we are still receiving users' comments and reports till all of them are reviewed and implemented. However, the official phase of ICA application will give the opportunity for the researchers to save their query results to benefit from them in their researches.

4 CONCLUSION

In this view, corpus-based investigations are very helpful; this can be very clear in lexical studies, grammar, semantics, NLP and many other language studies. Hence, this paper presents a quick view of the importance of building Arabic corpora and how most of the existing MSA analyzed Arabic corpora are not sufficient for building Arabic applications which service Arabic NLP. The current status of the ICA was presented along with its design, compilation and the used ICA utilities. This trial can be considered as one of the most successful approaches for analyzing modern standard Arabic (MSA) in comparison with other trials of Arabic analyzed corpora. ICA website plays a role in overcoming the lack of Arabic resources. It is the 1st online freely available, easy access query on 100,000,000 words which reflect the richness and variation of the ICA analyzed corpus whose aim is to help the NLP community in specific and other researchers in general.

REFERENCES

[1] C. Meyer (Ed.) (2002), "English corpus linguistics: An introduction, "Cambridge University Press.

[2] K. Saad, and W. Ashour (2010), "*OSAC: Open Source Arabic Corpora,"In 6th ArchEng Int. Symposiums, EEECS* (Vol. 10).

[3] A. Althubaity, A. Almuhareb, S. Alharbi, A. Al-Rajeh and M. Khorsheed (2008), "KACST Arabic text classification project: Overview and preliminary results.

[4] P. Zemanek (2001), *CLARA (Corpus Linguae Arabica): An Overview*, "*in Proceedings of ACL/*EACL *Workshop on Arabic Language*.

[5] L. Al-Sulaiti and E. Atwell (2004), *Designing and developing a corpus of Contemporary Arabic*, "*in Proceedings of the sixth TALC conference*, Granada, Spain.

[6] M. Maamouri, A. Bies, T. Buckwalter and W. Mekki (2004*), The penn arabic treebank: Building a large-scale annotated arabic corpus*, In NEMLAR conference on Arabic language resources and tools (pp. 102-109).

[7] R. Parker, D. Graff, K. Chen, J. Kong, and K. Maeda (2009), "Arabic Gigaword. Linguistic Data Consortium, "University of Pennsylvania, Philadelphia.

[8] M. Yaseen, M. Attia, B. Maegaard, K. Choukri, N. Paulsson, S. Haamid and A. Ragheb (2006), *Building annotated written and spoken Arabic LR's in NEMLAR project*, *"In Proceedings of LREC*.

[9] A. Olson (1998), Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains, Library trends, 47(2), 233-254.

[10] Biber D. (1993), *Representativeness in corpus design*, Literary and Linguistic Computing 8.243–57.

[11] S. Alansary, M. Nagi and N. Adly (2007), *Building an International Corpus of Arabic (ICA): progress of compilation stage*, In: 7th International Conference on Language Engineering, Cairo, Egypt.

[12] J. Sinclair (2004), *Developing Linguistic Corpora: a Guide to Good Practice Corpus and Text — Basic Principles*, Tuscan Word Centre.

[13] M. Linnarud (1976), "Lexical density and lexical variation-an analysis of the lexical texture of Swedish students' written work, "University of Lund.

[14] G. Bennett (2010), *An introduction to corpus linguistics, Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*, http://www.press.umich.edu/titleDetailDesc.do?id=371534, Michigan ELT, 2010.

[15] T. Yamina (2005), *Tagging by Combining Rules-Based Methods and Memory-Based Learning*, "*proceedings of world academy of science*, engineering and technology, Volume 6 June.

[16] D. Jurafsky and J. Martin (2008), *Speech and Language Processing: An introduction to speech recognition*, computational linguistics and natural language processing. 2nd Edition.

[17] A. Abdelali, J. Cowie and S. Soliman (2005), "Building A Modern Standard Arabic Corpus, *"Workshop on Computational Modeling of Lexical Acqusition*, the spilit meeting, Croatia, 25th to 28th of July.

[18] J. Mace (1999), *Arabic verbs and essential grammar*, London, Hodder and Stoughton.

[19] A. Farghaly (2008), *Arabic Computational Linguistics: Current Implementations*. Center for the study.

[20] D. Zoubir (2010), *The Broken Plural Morphological System in Arabic: A Challenge to Natural Language Processing Models*, Revue Maghrébine des Langues 2010. N°7 – Oran (Algérie).

[21] G. Corbett (2006), Gender, grammatical. In: Brown, Keith (ed.), *The Encyclopedia of Language and Linguistics. 2nd Edition,* Oxford: Elsevier. 749-756.

Biographies

Dr. Sameh Alansary: *Director of Arabic Computational Linguistics Center* Bibliotheca Alexandrina

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

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

Dr. Magdy H. Nagi: Senior Consultant, ICT Sector Bibliotheca Alexandrina.

Dr. Nagi is a Professor in the Computer and Systems Engineering department, Faculty of Engineering, Alexandria University. He obtained his Ph.D. from the University of Karlsruhe, in 1974, where he served as Lecturer for two years and as a Consultant to its Computer Center from 1974-1990. During this period he also served as Consultant to many companies in Germany such as Dr. Otker, Bayer, SYDAT AG, and BEC. He served, since 1995, as Consultant to the Bibliotheca Alexandrina. Among his activities were the design and installation of Bibliotheca Alexandrina's network and information system, namely a trilingual information system that offers full library automation. In 2001, he got appointed as the Head of the Information and Communication Technology (ICT) Sector of the Bibliotheca Alexandrina and occupied that post till 2012. He currently serves as a senior Consultant to the ICT Sector and continues to oversee the various projects and partnerships established between the ICT Sector and many international institutions. Dr. Nagi is a member of the ACM and the IEEE Computer Society as well as several other scientific organizations. His main research interests are in operating systems and database systems. He is author/co-author of more than 100 papers.

المدونة اللغوية العربية (ICA)

سامح الأنصاري[1*] ، مجدي ناجي[2**]

مكتبة الإسكندرية، الشاطبي، الإسكندرية، مصر
*قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر
[1]sameh.alansary@bibalex.org

** قسم هندسة النظم والحاسب، كلية الهندسة، جامعة الإسكندرية، الإسكندرية، مصر
[2]magdy.nagi@bibalex.org

ملخص— نظرا لأهمية اللغة العربية في العالم ومكانتها فقد تم في الآونة الأخيرة عدة محاولات لبناء مدونات عربية وتحليلها، ومع ذلك فإن هناك محاولات ضئيلة ناجحة. وإن هذه المدونات لها أهمية كبيرة في بناء تطبيقات خاصة بمعالجة اللغة الطبيعية. وبناء على ذلك، فإن مكتبة الإسكندرية كمؤسسة عريقة تهتم بشتى العلوم بدأت في بناء واحدة من أكبر المشاريع في الوطن العربي ألا وهو بناء المدونة اللغوية العربية (ICA) وتحليلها على المستوى المورفو-نحوي.

إن بناء المدونة اللغوية العربية جاء محاكيًا للمدونة اللغوية الإنجليزية (ICE) لتمثل اللغة العربية المعاصرة. ولكن تصميم المدونة وتجميعها ليس مطابقا تطابقا تاما للمدونة الإنجليزية ولكن تم تجميعها وتصميمما طبقا لمعايير معينة تناسب بناء المدونات العربية.

فتصنيف المدونة العربية وتصميمها اعتمد على تصنيف ديوي العشري المكتبي وذلك لتحديد الفئات العامة الأساسية التي تمثل اللغة العربية تمثيلا جيدا ومن ثم تم تحديد إحدى عشرة فئة أساسية متفرع من كل منهم أربعة وعشرون فئة فرعية ويتفرع منهم أربعة فئات أخرى فرع فرعية. أما بالنسبة لعملية التجميع فقد تم آليا من أربعة مصادر مختلفة هم الصحافة، المقالات الإلكترونية، الكتب، والدراسات الأكاديمية لكي يصل عدد كلمات المدونة إلى مائة مليون كلمة.

للمدونة اللغوية العربية الأوسمة الخاصة بها التي تستخدم في عملية التحليل، وهذه الأوسمة مكونة من ستة أنواع أساسية متفرع منهم ستة وأربعون نوعا. منها، خمسة لوسم الأفعال، عشرة للأسماء والصفات، ثلاثة للضمائر، عشرة للأدوات، اثنى عشرة للمعلومات الوصفية، أربعة للمعلومات غير اللغوية كالأرقام والعلامات.

أما القاموس الخاص بالتحليل فيتكون من سبعة وتسعين ألف مُدخل وكل منهم مقسم إلى سبعة عشرة معلومة تحلل الكلمة بناء عليهم. وهذه المعلومات هي الكلمة، نطق الكلمة، جذع الكلمة، السوابق، اللواحق، التذكير والتأنيث، العدد، التعريف والتنكير، الجذر، الوزن الصرفي، ساق الكلمة، العلامة الإعرابية.

ولاستفادة الباحثين والطلاب من مزايا المدونة العربية اللغوية فقد تم تصميم موقع إليكتروني خاص بالمدونة للبحث فيها على عدة مستويات منها البحث بمطابقة الكلمة، البحث بأصل الكلمة أو ساقها، البحث بجذع الكلمة، أو البحث بجذر الكلمة. وهناك طرق عديدة أخرى للبحث داخل المدونة.

أثر الفاء الحلقي للفعل الثلاثي المضعف في البنية الصرفية لمضارعه: دراسة لغوية حاسوبية

ا.د/ وفاء كامل فايد

كلية الآداب – جامعة القاهرة

wafkamel@link.net

مقدمة البحث:

خلَصت بحوثي السابقة إلى عدد من القواعد التي تحكم تنافر الأصوات العربية وتآلفها؛ وهو ما يشير إلى أن وراء السلوك اللغوي التلقائي للعربية نظاماً ضمنياً يحدد النماذج المقبولة وغير المقبولة.

ولما كان الفعل الثلاثي المضعف المجرد، مثل الأفعال (جَرَّ، خَفَّ، عَضَّ) لا يتكون إلا من صوتين فحسب وكنا نلحظ أن مضارع الفعل (جَرَّ) هو: يَجُرُّ بضم الجيم، ومضارع الفعل (خَفَّ) هو: يَخِفُّ بكسر الخاء، ومضارع الفعل (عَضَّ) هو: يَعَضُّ بفتح العين،فقد رأيت أن أختبر أثر موقع كل من صوتيهذا النوع من الأفعال على الباب الصرفي لمضارعه. وهذا البحث يحاول تلمس علاقة الصوت الأول من الفعل الثلاثي المضعف المجرد (فاء الفعل)، بالصيغة الصرفية لمضارعه.

أهداف البحث :

يهدف هذا البحث إلى تلمس الإجابة عن التساؤلات الآتية:

1- هل يؤثر مخرج الفاء الحلقي للفعل الثلاثي المضعف في ورود الفعل على باب صرفي بعينه ؟
2- هل يؤثر حيز الفاء الحلقي للفعل الثلاثي المضعف في ورود الفعل على باب صرفي بعينه ؟
3- هل يؤثر اتفاق الصفات أو اختلافها في صوتي الفعل الثلاثي المضعف: الفاء الحلقي مع العين واللام، في ورود الفعل على باب صرفي بعينه ؟
4- هل يمكن تلمس بعض القواعد التي تحكم أثر الفاء الحلقي على الباب الصرفي للفعل الثلاثي المضعف ؟

عينة البحث:

اعتمدت الدراسة القاموس المحيط للفيروزابادي؛ لاستخراج الأفعال الثلاثية الصحيحة التي وردت به؛ لغزارة مادته مع اختصاره، ولحرصه على ضبط حروف كلماته بالشكل، إلى جانب التزامه بتحديد الباب الصرفي لأفعاله بربطها بأوزان الأفعال المعروفة. وقد استقصت الدراسة الأفعال الثلاثية الصحيحة المضعفة به، واتخذتها عينةً للبحث.

خطوات البحث

استقصت الدراسة الأفعال الثلاثية الصحيحة المضعفة التي وردت بالقاموس المحيط، وسجلتها مع تصريفاتها في جدول خاص قام عليه البحث. ومن الأفعال المرصودة في الجدول رصدت الدراسة تصرف المضعف الثلاثي حين يكون أحد الأصوات الحلقية فاءً له، وتتغير أصوات عينه ولامه.

وقد دفعت هذه البيانات إلى مبرمج حاسوبي، وقورنت النتائج اللغوية بالنتائج الحاسوبية.

نتائج البحث:أكَّدت الدراسة ما يلي:

1. أثر مخرج الفاء الحلقي للمضعف فقط، مع مخرج عينه ولامه فقط،في تصرف مضارع الفعل.
2. أثر مخرج (الفاء) مع حيز (العين واللام) في تصرف الفعل على الباب الصرفي.
3. أثر حيز (الفاء) مع مخرج (العين واللام) في تصرف الفعل على الباب الصرفي.

4. أثر حيز (الفاء) وصفته، مع مخرج (العين واللام) وصفته، في تصرف الفعل.
5. أثر مخرج صوتي المضعف مع صفتهما في تصرف مضارع الفعل على باب بعينه.
6. أثر اختلاف صفتي الجهر والهمس في فاء المضعف في تصرف مضارع الفعل.
7. أثر اختلاف صفتي الإطباق والانفتاح في صوتي الثلاثي المضعف على بابه الصرفي.

أهمية البحث وجدواه التطبيقية:

يمكن **الاستفادة من نتائج الدراسةفي** كل من **اللسانيات الحاسوبية واللسانيات التطبيقية**كما يلي:

أولا : في العمل المعجمي الحاسوبي Computational lexicography:

❖ **في بناءقاعدة بيانات معجمية** Lexical database :

فبناء قاعدة بيانات معجمية يتطلب استقصاء للكلمات والأوزان الممكنة والممتنعة، وإحصاء ذلك آليا؛ لتتميم وصف المعجم، والتوصل إلى القواعد الصوتية، والصوتية الصرفية، التي تحكم هذا المعجم.

❖ **في تعرّف الكلام** Speech recognition :

ويكون ذلك ببناء نماذج تشمل قواعد التتابعات الممكنة صوتيا، والتتابعات غير الممكنة، وهو ما يسهل عملية الإدراك الآلي للأصوات؛ بما يرفع نسبة الدقة في التعرف والفهم الآليين.

ثانيا: في اللسانيات التطبيقية Applied linguistics**:**

❖ **في الصناعة المعجمية** Lexicography.

❖ **في تعليم اللغة** Language learning**.**

❖ **في علم المصطلح**Terminology.

السيرة الذاتية:

-أستاذة بكلية الآداب جامعة القاهرة.

- عضو مراسل بمجمع اللغة العربية بدمشق.

- خبيرة بمجمع اللغة العربية بالقاهرة.

- نشرت 53 بحثا باللغتين العربية والانجليزية، وأشرفت على 48 رسالة للماجستير والدكتوراه.

- حصلت على جائزة جامعة القاهرة التشجيعية عام 2004، وجائزة جامعة القاهرة التقديرية عام 2013.

- ألّفت وترجمت 9 كتب منها:(تراكب الأصوات في الفعل الثلاثي الصحيح)- (قصيدة الرثاء بين شعراء الاتجاه المحافظ ومدرسة الديوان: دراسة أسلوبية إحصائية)- (الباب الصرفي وصفات الأصوات)- (بحوث في العربية المعاصرة)- (المجامع العربية وقضايا اللغة)- (معجم التعابير الاصطلاحية في العربية المعاصرة)- (اتجاهات البحث اللساني)- (مدخل إلى اللغة).

Impact of the 1st Consonant Pharyngeal Phoneme in Trilateral Geminate Abstract Verb, on its Morphological Form: a Linguistic Computational Study.

Wafaa Kamel

Faculty of Arts, Cairo University

wafkamel@link.net

Preface:

This article deals with phonotactics, aiming to reveal the relation between phonemes of trilateral geminate abstract verb (TGAV), and its morphological conjugation in present tense (PT).

In a former study, the author concluded some rules dominating phonological behavior of (TAV). These rules demonstrate an internal system underlying phonological formation of Arabic words. This led her to think of (TGAV); to see if this category is submitted to that internal system or not.

Considering pairs of TGAV like [ʢæbbæ] عَبّ [ʢæffæ] عَفّ in past tense, we find the 1st consonant (C1) in both verbs is the same [ʕ] ع and (C2) is labial [b] ب & [f] ف . In (PT), the 1st verb is [jæʢʊbb] يَعُبّ and the 2nd is [jæʢɪff] يَعِفّ . This shows that the point of articulation of (C2):[b] & [f] affected morphological conjugation of both verbs.

We find also pairs like [ðællæ] ذَلّ , [læððæ] لَذّ & [ʢæbbæ] عّبَّ , [bæʢʢæ] بَعَّ , in where the two consonants of each pair interchange their position, leading to a change in morphological form of (PT) of each verb: [jæðɪll] يَذِلّ , [jælæðð] يَلَذّ & [jæʢʊbb] يَعُبّ , [jæbɪʢʢ] يَبِعّ . This indicates that changing the position of consonants leads to a change of its (PT) morphological form.

On the other hand, when (C2) is the same and (C1) changes , we find that the morphological conjugation is affected also, as in pairs [ɣæsˤsˤæ] غَصَّ and [xæsˤsˤæ] خَصَّ : the (PT) of the 1st verb is [jæɣæsˤsˤ] يَغَصّ and of the 2nd is [jæxʊsˤsˤ] يَخُصّ , also in verbs [hæssæ] هَسَّ and [ʕæssæ] عَسَّ , (PT) is [jæhɪss] يَهِسّ and [jæʢʊss] يَعُسّ . The same in verbs [ʔænnæ] أَنَّ [ɣænnæ] غَنَّ, (PT) is [jæʔɪnn] يَئِنّ and [jæɣænn] يَغَنّ , even though

(C1) in the six verbs is pharyngeal Phoneme. This indicates that the 1st consonant of (TGAV) affects the conjugation of the (PT) morphological form.

This article aims to reveal the effect of 1st consonant of (TGAV) on morphological verb form in (PT).

Sample:

All (TGAV) of Fayrouzabady's: al-Qamūs al Muħeet القاموس المحيط للفيروزابادي are considered as a sample.

Questions and aims:

1- Has point of articulation of the pharyngeal 1st consonant of (TGAV) an effect on its conjugation?

2- Has the range of articulation of the pharyngeal 1st consonant of (TGAV) an effect on its conjugation?

3-have manners of articulation of both consonants of (TGAV) an effect on morphological category?

4- Can we identify rules dominating the effect of the pharyngeal 1st consonant on the morphological behavior of (TGAV)?

Results:

1- Finding out rules which dominate the effect of the pharyngeal 1st consonant on (TGAV) morphological conjugation.

2-Revealing the relation between the pharyngeal 1st consonant phoneme and the morphological form .

مجتمع المعرفة وعلوم الانسانيات : نظرة حاسوبية

د/ نبيل على

خبير اللسانيات الحاسوبية

nabialii@gmail.com

الملخص:

تشمل منظومة العلوم الإنسانية عدة فروع علمية أهمها علم الاجتماع وعلم النفس واللسانيات والتاريخ والاقتصاد وإضافة إلى الجماليات.

وكان من الطبيعي أن تتوطد العلاقة بين مجتمع المعرفة وهذه المنظومة العلمية التي تعاظم دورها مع تنامي استخدام شبكات التواصل الاجتماعي وتعدد مظاهر التفاعل بين هذه المنظومة والنقلة النوعية التي نجمت عن بزوغ مجتمع المعرفة وخير شاهد على ذلك تلك النزعة المعرفية التي صبغت بها كل فروع الإنسانيات دون استثناء ومن أمثلة ذلك علم الاجتماع المعرفي وعلم النفس المعرفي وعلم اللسانيات المعرفي وعلم التاريخ المعرفي وعلم الاقتصاد المعرفي إضافة إلى الجماليات المعرفية.

تغري هذه النزعة بالبحث عن مقاربة شبه موحده لفروع الإنسانيات المختلفة ومن ثم ـ وهو الأمر الأكثر طموحا ـ البحث عن معالجات حاسوبية شبه موحده هي الأخرى للأشكاليات العديدة التي تكمن في صلب العلاقة بين مجتمع المعرفة وعلوم الإنسانيات.

تسعى الدراسة الحالية إلى تناول هذه القضية بصورة أولية بهدف إبراز الفرص والتحديات التي تنطوي عليها حاسوبيا وتواصليا أملا في فتح آفاق جديدة لحوسبة الإنسانيات لمقاربات تتمركز حول اللغة بصفتها محور العلوم الإنسانية بلا منازع خاصة بعد الطابع المعرفي المشار إليه سابقا.

لقد حان الوقت أن ينهل أهل حوسبة اللغة من مصادر معرفية أكثر عمقا وأشد ارتباطا بالتفاعل الذهني من جانب والتواصل الاجتماعي من جانب آخر.

إنشاء قاعدة بيانات لحروف القرآن الكريم
(المنهج والنموذج)

محمد عبد الرحمن الخطيب
طالب في مرحلة الدكتوراه
كلية دار العلوم - جامعة القاهرة
Makh2000@hotmail.com

الملخص:

تقدم هذه الدراسة منهجًا علميًّا لإنشاء قاعدة بيانات لحروف القرآن الكريم، يراعي الجوانب القرآنية واللغوية والحاسوبية، وتميِّز هذه القاعدة الحرف المنطوق من غيره، والمكتوب من غيره، والحركة المنطوقة من غيرها، والمكتوبة من غيرها، وحركة الحرف حال الوقف أو الابتداء، كما تراعي ظواهر الرسم العثماني، حيث إنه لا يوجد –فيما اطلعت عليه- قاعدة بيانات لحروف القرآن الكريم، ووجود مثل هذه القاعدة سيقدم خدمة جليلة للدراسات القرآنية واللُّغوية على حدٍّ سواء، إضافة إلى ذلك تقدم الدراسة نموذجًا لقاعدة البيانات لحروف القرآن الكريم، يوضِّح الصورة النهائية لشكلها، كما تقدِّم الدراسة نموذجًا آخر لإحدى الدراسات المستنتجة من قاعدة البيانات، وهي دراسة لشيوع الأصوات في القرآن الكريم، مما يؤكد أهمية إنشاء مثل هذه القاعدة.

الكلمات المفتاحية:

حروف القرآن، حركات القرآن، قواعد البيانات، شيوع الأصوات، اللسانيات الحاسوبية.

أولًا: مقدّمة:

تطوّر الحاسب الآلي والبرامج المرتبطة به تطورًا كبيرًا في القرن الحادي والعشرين، تطورًاله أثرٌ ملموس على جوانب الحياة المختلفة سواءً الاقتصادية أو السياسية أو العلمية التي كانت إحداها العلوم اللغوية.

وتُعدُّ البرامج المرتبطة بقواعد البيانات عنصرًا جوهريًّا في تسيير أمور الحياة اليومية في المجتمع المعاصر، فالعمليات البنكية، والنتائج الدراسية، ووثائق السفر، وغيرها يجب التعامل فيها مع قواعد البيانات، كما أنها تقدّم خدمات جليلة للعلوم المختلفة؛ نظرًا لسهولة تنظيم البيانات فيها، وسهولة التعامل معها، وإمكانية تخزين كميات كبيرة من البيانات، وصِغر المساحة التي تشغلها.

وتهدف هذه الدراسة إلى إنشاء قاعدة بيانات لحروف القرآن الكريم بغرض الاستفادة منها في الدراسات القرآنية واللُّغوية على حدٍّ سواء، فهناك علاقة وطيدة بين القرآن الكريم واللغة العربية، فأصبح نصٍّ لُغوي وصل إلينا هو القرآن الكريم، ولو جُرِّد من إطالة الصوت بالغنة وحروف المد، ومن السكت، ومن التغنّي المأمور به لتحسين الصوت بالقرآن لكان الأداء القرآني يُماثل الأداء الفصيح بأيِّ نصٍّ نثريٍّ آخر [1]، لذلك يوجد كثير من القواعد النَّظرية المتعلقة بعلم التجويد وضعها علماء اللغة عند كتابتهم لقواعد اللغة العربية.

ولم أجد -فيما اطَّلعت عليه- قاعدة بيانات لحروف القرآن الكريم، ووُجِدت قواعد بيانات أخرى لكلمات القرآن الكريم، أو آياته [2] [3]، أو غير ذلك.

وبالنظر إلى البرامج الحاسوبية والمواقع الالكترونية المتعلِّقة بالبحث في القرآن الكريم فإن جميعها -فيما اطَّلعت عليه- تتعامل مع قاعدة بيانات لكلمات القرآن الكريم أو آياته، أو تتعامل مع القرآن الكريم بوصفه نصًّا، منها على سبيل المثال: برنامج المصحف الرقميّ [4]، وبرنامج انتلايز (Intellyze) [5]، وبرنامج آية [6]، وبرنامج قرآن كود (QuranCode) [7]، وبرنامج الباحث في القرآن الكريم [8]،وبرنامج الإحصاء العددي لكلمات وحروف القرآن الكريم [9]، وبرنامج الفرقان [10]، وبرنامج مصحف المدينة النبوية [11]، وموقع علم القرآن الكريم [12]، وموقع الفانوس [13]، وموقع الأوفى [14]، وموقع صفحة القرآن الكريم [15]، وموقع الباحث القرآني [16]، وغيرها.

ويلاحظ على كثير من البرامج الحاسوبية والمواقع الالكترونية المتعلِّقة بالقرآن الكريم الآتي:

- اعتمادها على الرسم الإملائي للقرآن الكريم، مما يعطي نتائج غير دقيقة، ويُخالف رأي جمهور العلماء بوجوب اتباع الرسم العثماني في كتابة المصاحف [17] [18].
- عدم اعتنائها بالحركات.
- عدم وجود قاعدة بيانات لحروف القرآن الكريم.

وتكمن أهمية إنشاء قاعدة بيانات لحروف القرآن الكريم في عدة أمور، منها:

- كونها تتعامل مع أصغر وحدة في القرآن الكريم وهو الحرف، مما يمكّن من التعامل مع الوحدات الأكبر منها، مثل: الكلمة، والآية، والسورة.
- مراعاتها للحركات مع إمكانية التعامل معها بشكل مستقل.

- مراعاتها حال الحرف والحركة من حيث النطق والكتابة.
- اعتمادها على الرسم العثماني، مما يعطي نتائج أكثر دقة عن حروف القرآن الكريم وحركاته من الرسم الإملائي.
- إمكانية استنتاج نتائج وإحصاءات من قاعدة البيانات تخدم دراسات صوتية، ودلالية، ونحوية، وغيرها، منها على سبيل المثال:
  - o دراسة لشيوع الأصوات في القرآن الكريم.
  - o دراسة للائتلاف الصوتي في القرآن الكريم[1].
  - o دراسة لدلالة الأصوات في القرآن الكريم.
  - o دراسة للظواهر النحوية في القرآن الكريم [19].

ثانيًا: قواعد البيانات:

تُعرَّف قاعدة البيانات (Database): بأنَّها مجموعة من عناصرِ البيانات المنطقية المرتبطة مع بعضها بعلاقة معينة، وتتكون قاعدة البيانات من جدول (Table) واحد أو أكثر، ويتكون الجدول من سجل (Record) واحد أو أكثر، ويتكون السجل من حقل (Field) واحد أو أكثر [20].

والهدف الأساسي لقواعد البيانات هو التركيز على طريقة تنظيم البيانات بحيث تكون خالية من التكرار مع إمكانية استرجاعها، وتعديلها، والإضافة عليها، وليس على البرامج أو التطبيقات الخاصة.

فلو أُريد إنشاء قاعدة بيانات لحروف القرآن الكريم فإن من عِناصر البيانات فيها: السورة، والآيِة، والكلمة، والحرف، والحرِكة، والعلاقة بينها أن الحرف والحركة يمثِّلان جزءًا من الكلمة، والكلمة تمثِّل جزءًا من الآية، والآية تمثِّل جزءًامن السورة، والسورة تمثِّل جزءًا من القرآن الكريم، وحتى تكون عناصر البيانات منطقية يجب ترتيب الحروف والحركات في الكلمة، والكلمات في الآية، والآيات في السورة، والسورة في القرآن الكريم.

وحتى يتضح مفهوم قواعد البيانات يعرض جدول (1) قاعدة البيانات لكلمة (مِصْرَ) من قوله تعالى:
(فَلَمَّا دَخَلُواْ عَلَىٰ يُوسُفَ ءَاوَىٰٓإِلَيۡهِأَبَوَيۡهِ وَقَالَ ٱدۡخُلُواْمِصۡرَ إِن شَآءَٱللَّهُ ءَامِنِينَ ٩٩)، وهي الآية التاسعة والتسعون من سورة يوسف التي تحمل الرقم (12) في ترتيب سور القرآن الكريم.

جدول (1): قاعدة البيانات لكلمة (مِصْرَ)

رقم السورة	رقم الآية	رقم الكلمة	رقم الحرف	الحرف	الحركة
12	99	10	1	م	ـِ
12	99	10	2	ص	ـْ

[1] يقوم الباحث بإعداد دراسة بعنوان (الائتلاف الصوتي في القرآن الكريم) لنيل درجة الدكتوراه تحت إشراف الأستاذين الفاضلين: أ.د. إبراهيم الدسوقي عبد العزيز، و أ.د. أحمد شرف الدين.

12	99	10	3	ر	َ

يُلاحظ في جدول (1) الآتي:

- سورة يوسف حملت رقم ترتيبها في القرآن الكريم وهو (12).
- الآية حملت رقم ترتيبها في سورة يوسف وهو (99) .
- كلمة (مِصْرَ) حملت رقم ترتيبها في الآية التي ذكرت فيها وهو (10).
- كل حرف في كلمة (مِصْرَ) حمل رقم ترتيبه في الكلمة.
- لا يمكن أن يتكرر أي صفٍّ (سجل) في قاعدة البيانات بهذه الطريقة.
- يمكن الوصول إلى أيِّ حرف في القرآن الكريم بهذه الطريقة.

ثالثًا: الخطوات المنهجية لإنشاء قاعدة بيانات لحروف القرآن الكريم:

هناك ثلاث خطوات لإنشاء أي قاعدة بيانات، وهي على النحو الآتي:

1- إنشاء جداول قاعدة البيانات.
2- إنشاء العلاقات بين الجداول.
3- إدخال البيانات إلى الجداول.

وسيتم الحديث بالتفصيل عن كيفية القيام بهذه الخطوات لإنشاء قاعدة بيانات لحروف القرآن الكريم.

1، 2- إنشاء جداول قاعدة البيانات لحروف القرآن الكريم، والعلاقات بين جداولها:

سبق أن أُشير إلى الجدول الأساسي في قاعدة البيانات لحروف القرآن الكريم، وهو الجدول الذي يضع كل حرف مع حركته في سجلٍّ مستقِل إلا أنه سيُضاف إليه عمود جديد يُسمَّى (الحركة الإضافية) بغرض تسجيل حركة الحرف عند تغيُّرها حال الوقف أو الابتداء، فمثلًا كلمة (ٱهۡدِنَا) من قوله تعالى: (ٱهۡدِنَاٱلصِّرَٰطَٱلۡمُسۡتَقِيمَ ٦) [الفاتحة: 6] حركة همزة الوصل فيها الكسرة حال الابتداء بها، في حين تسقط همزة الوصل حال وصل الآية بما قبلها، وحركة النون فيها الفتحة حال وصل الكلمة بما بعدها نظرًا لالتقاء الساكنين، في حين أن حركة النون الألف حال الوقف عليها، لهذا الاختلاف الحاصل أُثبِتت حركة الحرف حال الوصل في عمود الحركة، وهو الأصل؛ لأن القرآن الكريم ضُبط على الوصل [18]، ولأن قارئ القرآن بشكل خاص، والقارئ والمتحدِّث بشكل عام، لا يقف على كل حرف، بل يصل كلامه ولا يقف إلا للتَّنفُّس أو رغبة في التوقُّف عن الحديث أو القراءة، وأُثبتت حركة الحرف حال الابتداء أو الوقف في عمود الحركة الإضافية، ويظهر ذلك في جدول (2) الذي يعرض قاعدة البيانات لكلمة (ٱهۡدِنَا).

جدول (2): قاعدة البيانات لكلمة (ٱهۡدِنَا)

رقم السورة	رقم الآية	رقم الكلمة	رقم الحرف	الحرف	الحركة	الحركة الإضافية
1	6	1	1	ٱ		ـِ
1	6	1	2	هـ	ـۡ	
1	6	1	3	د	ـِ	
1	6	1	4	ن	ـَ	ا

ونظرًا لأنه يتوجَّب إضافة وصف لكل حرف وحركة في قاعدة البيانات للتعريف بهما، وتمييز المنطوق من غيره والمكتوب من غيره، ولو أضيف الوصف إلى الجدول نفسه لأحدث تكرارًا كبيرًا في قاعدة البيانات، فمثلًا كلما يُذكر حرف الباء في قاعدة البيانات سيُذكر وصفه، فسيتكرر وصف حرف الباء بعدد مرات تكراره في القرآن الكريم، والمستخدم لن يحتاج إلى وصف حرف الباء إلا مرة واحدة، فهذا التكرار غير مقبول في أنظمة قواعد البيانات؛ لأنه بلا فائدة، لذلك سيُضاف جدول للحروف وجدول للحركات في قاعدة البيانات، تُذكر فيهما تفاصيل الحروف والحركات، ويتم ترميزهما بأرقام؛ حتى يسهل ربطهما بالجداول الأخرى، ثم يتم ربط جدولي الحروف والحركات بالجدول الأساسي، وهو جدول الحروف والحركات في القرآن الكريم مع استبدال الحروف والحركات في هذا الجدول برموزهما المذكورة في جدولهما.

ويتطلَّب إتمام قاعدة البيانات إضافة جدول لكلمات القرآن الكريم وجدول آخر لأسماء سور القرآن الكريم؛ لأن مستخدم قاعدة البيانات سيحتاج إلى تعرُّف الكلمة القرآنية واسم السورة وليس رمزهما، ولو وُضِعت كلمات السورة واسمها في جدول الحروف والحركات في القرآن الكريم لتكررت الكلمة القرآنية بعدد حروفها، وتكرَّر اسم السورة بعدد حروفها، وهذا تكرار غير مقبول في أنظمة قواعد البيانات.

ويوضِّح جدول (3) شكل جدول الحروف والحركات في القرآن الكريم قبل إجراء عملية الاستبدال وإضافة الجداول، ويوضِّح جدول (4) شكل جدول الحروف والحركات في القرآن الكريم بعد إجراء عملية الاستبدال وإضافة الجداول.

جدول (3): جدول الحروف والحركات في القرآن الكريم قبل إجراء عملية الاستبدال وإضافة الجداول

رقم السورة	اسم السورة	رقم الآية	رقم الكلمة	الكلمة القرآنية	رقم الحرف	الحرف	منطوق	مكتوب	الحركة	منطوقة	مكتوبة	الحركة الإضافية	منطوقة	مكتوبة
1	الفاتحة	6	1	ٱهۡدِنَا	1	ٱ	منطوق ابتداءً	مكتوب				ـِ	منطوقة ابتداءً	غير مكتوبة
1	الفاتحة	6	1	ٱهۡدِنَا	2	هـ	منطوق	مكتوب	ـۡ	غير منطوقة	غير مكتوبة			
1	الفاتحة	6	1	ٱهۡدِنَا	3	د	منطوق	مكتوب	ـِ	منطوقة	غير مكتوبة			
1	الفاتحة	6	1	ٱهۡدِنَا	4	ن	منطوق	مكتوب	ـَ	منطوقة وصلًا	غير مكتوبة	ا	منطوقة وقفًا	مكتوبة

جدول (4): جدول الحروف والحركات في القرآن الكريم بعد إجراء عملية الاستبدال وإضافة الجداول

رقم السورة	رقم الآية	رقم الكلمة	رقم الحرف	رمز الحرف	رمز الحركة	رمز الحركة الإضافية
1	6	1	1	351	0	74
1	6	1	2	271	81	0
1	6	1	3	91	71	0
1	6	1	4	261	52	13

وهكذا تشكَّلت جداول قاعدة البيانات لحروف القرآن الكريم، وهي خمسة جداول على النحو الآتي:

1- جدول الحروف.
2- جدول الحركات.
3- جدول الكلمات القرآنية.
4- جدول السور القرآنية.
5- جدول الحروف والحركات في القرآن الكريم.

بعد إنشاء جداول قاعدة البيانات يُلاحظ أن البيانات وُزِّعت على الجداول، ويلزم إنشاء علاقة بين الجداول تمكِّن من الوصول إلى المعلومة من جدول واستكمالها من الآخر عن طريق هذه العلاقة، فمثلًا لو أراد المستخدم تعرُّف عدد مرات تكرار حرف ما في سورة معيَّنة، فهذه المعلومة يمكن استنتاجها من جدول الحروف والحركات في القرآن الكريم، وعند عرض النتيجة لا يمكن عرض اسم السورة إلا بالعودة إلى

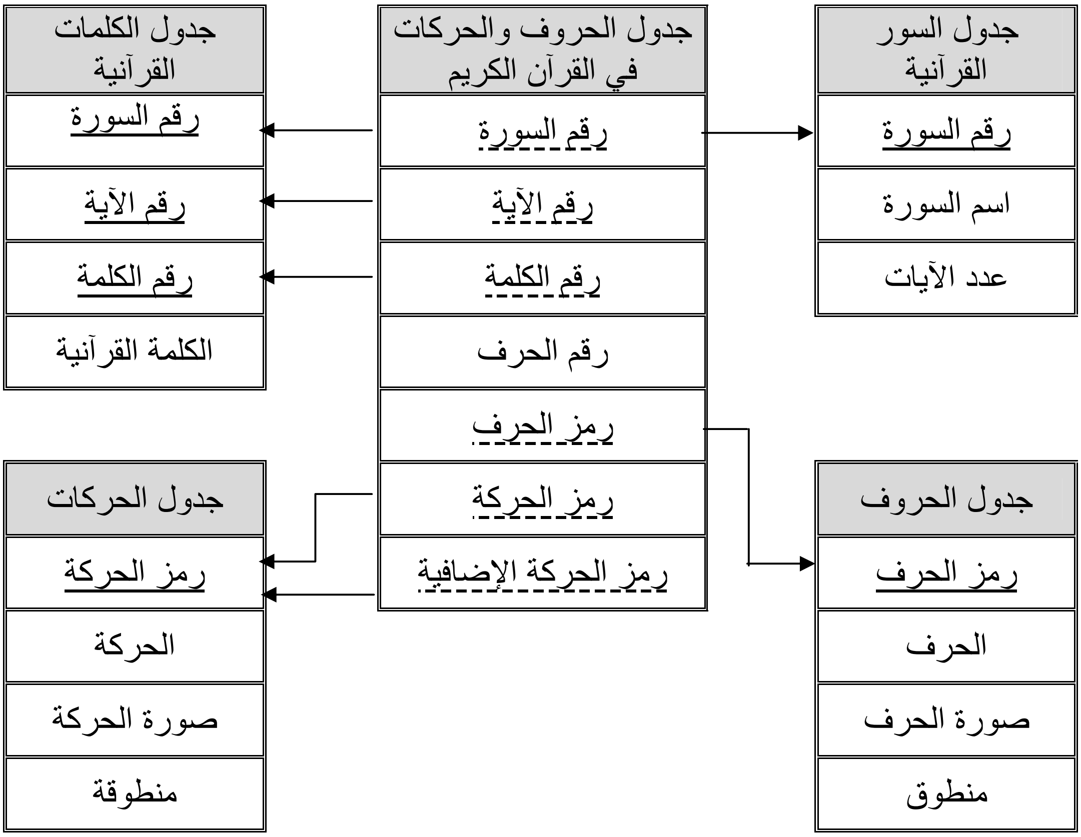
جدول السور القرآنية، ويتم ذلك بإنشاء علاقة بين جدول الحروف والحركات في القرآن الكريم وجدول السور القرآنية.

ولإنشاء علاقة بين جدولين يلزم استخدام مفتاحين –حسب أنظمة قواعد البيانات، المفتاح الأول: المفتاح الأساسي (Primary Key): وهو الحقل الذي يحتوي على بيانات لا تتكرر داخل الجدول، ولا يجوز ترْكه فارغًا، نحو رمز الحرف في جدول الحروف، فكلُّ حرف له رمز مستقلّ، لا يتكرَّر مع حرف آخر، ولا يُترك فارغًا من غير رمز، وأُشير إلى المفاتيح الأساسية في الجداول بوضع خطٍّ مستقيم أسفل اسم الحقل –انظر الشكل (1).

المفتاح الثاني: المفتاح الأجنبي (Foreign Key): وهو الحقل المرتبط بالمفتاح الأساسي في الجدول الآخر بغرض ربط الجدولين، نحو رمز الحرف في جدول الحروف والحركات في القرآن الكريم، وأشير إلى المفاتيح الأجنبية في الجداول بوضع خطٍّ متقطِّع أسفل اسم الحقل [20] –انظر الشكل (1).

ووجود علاقة بين جدولين بحيث يحوي الجدول الأول مفتاحًا أساسيًّا، والجدول الثاني يرتبط مع الجدول الأول بمفتاح أجنبي له نفس نوع بيانات المفتاح الأساسي، يُسمى بالتكامل المرجعي (Referential Integrity) في قواعد البيانات.

ويبيِّن الشكل (1) جداول قاعدة البيانات لحروف القرآن الكريم والعلاقات بينها.



مكتوب
التوضيح

مكتوبة
التوضيح

شكل (1): جداول قاعدة البيانات لحروف القرآن الكريم والعلاقات بينها

3- إدخال البيانات إلى الجداول:

وهي تمثِّل الخطوة الأخيرة لإنشاء قاعدة بيانات لحروف القرآن الكريم، ويتم إدخال البيانات إلى الجداول بشكل يدويٍّ باستثناء جدول الكلمات القرآنية وجدول الحروف والحركات في القرآن الكريم، اللذين سيتم إدخال البيانات إليهما عبر إجراء معالجة آلية لنصِّ القرآن الكريم.

<u>رابعًا: ضوابط تحويل نصِّ القرآن الكريم إلى قاعدة بيانات لحروف القرآن الكريم:</u>

التزم البحث مجموعة من الضوابط عند تحويل نصِّ القرآن الكريم إلى قاعدة بيانات لحروف القرآن الكريم، وهي على النحو الآتي:

- الالتزام برواية حفص عن عاصم من طريق الشاطبية؛ لكونها الرواية الأكثر شهرة في العالم الإسلامي.
- الاعتماد على النسخة الالكترونية من مصحف المدينة النبوية الصادر من مجمع الملك فهد لطباعة المصحف الشريف، والمكتوب بخطِّ الرسم العثماني – حفص، وهو موجود على هذا الرابط (http://fonts.qurancomplex.gov.sa/?page_id=42)؛ لكون هذا المصحف من أضبط المصاحف وأشهرها، إضافة إلى إمكانيِّة التعامل معه حاسوبيًّا.
- مراعاة ظواهر الرسم القرآنيّ، أو ما يُسمَّى بالرَّسم العثماني، نحو: الألف المحذوفة، والواو المحذوفة، والياء المحذوفة، والنون المحذوفة، والألف المزيدة، والواو المزيدة، والياء المزيدة، والألف المبدلة واوًا، والألف المبدلة ياءً، والسين المبدلة صادًا، وغيرها، وذلك بالتمييز بينها وبين أصولها، والتعامل معها في قاعدة البيانات على أنها حروف أو حركات مستقلة.
- بناء قاعدة البيانات على أساس وصل الحرف بما بعده؛ لأن القرآن الكريم ضُبط على الوصل، فيتمُّ إثبات حركة الحرف حال الوصل في خانة الحركة، وإذا كانت حركة الحرف تختلف حال الابتداء به مثل حركة همزة الوصل عند الابتداء بهمزة الوصل، أو حال الوقف عليه مثل حركة الحرف المُنوَّن تنوين نصْبٍ حال الوقف عليه، أو كانت حركة الحرف إحدى الحركات الطويلة وهناك حركة قصيرة من جنسها غير منطوقة ضُبطت في المصحف الشريف، فيتمُّ إثبات كل هذه الحركات التي لها وضع خاص في خانة الحركة الإضافية.

وبناءً على ما سبق يتم تطبيق قاعدة التقاء الساكنين، فيتم إثبات حركة الحرف حال الوصل في خانة الحركة، وإثبات حركة الحرف حال الوقف في خانة الحركة الإضافية، نحو النون في قوله تعالى:(ٱهۡدِنَاٱلصِّرَٰطَ) [الفاتحة: 6]، حركتها الفتحة وصلًا نظرًا لحذف الألف بسبب التقاء الساكنين، وحركتها الألف وقفًا، وبالتَّالي ستكون الفتحة في خانة الحركة، والألف في خانة الحركة الإضافية.

- يتم وضع السُّكون في جداول الحركات وقوائمها من باب التغليب، فإن الحرف إما متحرِّك أو ساكن، ووضع جداول وقوائم خاصة بالحركات وأخرى بالسكون يعقِّد قاعدة البيانات بشكل خاص والبرنامج بشكل عام.

- مراعاة المنطوق والمكتوبمن الحروف والحركات، والحروف في القرآن الكريم بالنسبة للنطق خمسة أقسام، وهي:
  - o **منطوق:** أي أن الحرف منطوق بكامل صفاته ابتداءً ووقفًا وفي دَرْجِ الكلام.
  - o **منطوق ابتداءً:** أي أن الحرف منطوق بكامل صفاته عند الابتداء به، وغير منطوق وقفًا وفي دَرْجِ الكلام.
  - o **منطوق وصلًا:** أي أن الحرف منطوق بكامل صفاته في دَرْجِ الكلام، وغير منطوق ابتداء ووقفًا.
  - o **منطوق وقفًا:** أي أن الحرف منطوق بكامل صفاته عند الوقف عليه شرط أن يكون في آخر الكلمة، حيث إنه لا يصحُّ الوقف على بداية الكلمة أو وسطها، وأن الحرف غير منطوق ابتداءً وفي دَرْجِ الكلام أو منطوق ببعض صفاته[2] أو إحداها في دَرْجِ الكلام[3].
  - o **غير منطوق:** أي أن الحرف غير منطوق ابتداءً ووقفًا وفي دَرْجِ الكلام أو منطوقة إحدى صفاته فقط في دَرْجِ الكلام[4].

والحروف في القرآن الكريم بالنسبة للكتابة قسمان، هما:

1. **مكتوب:** أي أن الحرف كُتِب في أحد المصاحف التي أمر سيدنا عثمان τ بكتابتها، ولا تُعدُّ الحروف التي زيدت بوصفها علاماتللضبط مكتوبة، نحو: الهمزة على ألف محذوفة (◌ٔ◌ٰ)، والواو غير المدية المحذوفة (ۥ)، وغيرها.
2. **غير مكتوب:** أي أن الحرف لم يُكتب في أحد المصاحف التي أمر سيدنا عثمان τ بكتابتها، وتُعدُّ الحروف التي زيدتللضبط غير مكتوبة.

ويتركَّب من أقسام المنطوق والمكتوب للحروف في القرآن الكريم عشرة أقسام، ثلاثة منها لا يندرج حرف من الحروف تحتها، وهي:

1. منطوق ابتداءً وغير مكتوب.
2. منطوق وصلًا ومكتوب.

[2] مِثْل الطاء المُدغمة في التاء إدغامًا ناقصًا.

[3] مِثْل بقاء صفة الغُنَّة للنون الساكنة عند وصلها بأحد حروف الإخفاء.

[4] مِثْل بقاء صفة الغُنَّة للتنوين المتتابع عند وصله بأحد حروف الإخفاء.

3. منطوق وقفًا وغير مكتوب.

وسبعة أقسام تندرج الحروف تحتها، وهي:

1. **منطوق ومكتوب:** وتندرج تحته الهمزة على الألف، والهمزة على الواو، والهمزة على الياء، والهمزة المسهَّلة، والباء، والتَّاء، والتَّاء المربوطة، والثَّاء، والجيم، والحاء، والخاء، والدَّال، والذَّال، والرَّاء، والزَّاي، والسِّين، والشِّين، والصَّاد، والسِّين المبدلة صادًا المنطوقة سينًا، والسين المبدلة صادًا المنطوقة صادًا، والضَّاد، والطَّاء، والظَّاء، والعين، والغين، والفاء، والقاف، والكاف، واللام، والميم، والميم المخفاة، والنَّون السَّاكنة المبدلة ميمًا ساكنة وصلًا، والنَّون، والنُّون المُشمَّة، والهاء، والواو غير المدِّيَّة، والياء غير المدِّيَّة.

وحروف هذا القسم هي الأكثر تكرارًا في القرآن الكريم؛ لأن الأصل في الحروف أن تكون منطوقة ومكتوبة، فمجموع حروف هذا القسم في القرآن الكريم (253262) حرفًا، وهي تمثِّل (81.69%) من حروف القرآن الكريم.

2. **منطوق وغير مكتوب:** وتندرج تحته الهمزة على الألف، والهمزة على السطر، والهمزة على ألف محذوفة، والباء، والتَّاء، والثَّاء، والجيم، والحاء، والخاء، والدَّال، والذَّال، والرَّاء، والزَّاي، والسِّين، والشِّين، والصَّاد، والضَّاد، والطَّاء، والظَّاء، والعين، والفاء، والقاف، والكاف، واللام، والميم، والنَّون، والنُّون المحذوفة، والهاء، والواو غير المدِّيَّة، والياء غير المدِّيَّة، والياء المحذوفة غير المدِّيَّة.

وبشكل عام فإن حروف هذا القسم تمثِّل أحد الأحرف الثلاثة الآتية:

أ. الحرف الأول من الحرف المُشدَّد، ويُلاحظ أن التاء المربوطة والغين لم ترد في هذا القسم؛ لأنها لم ترد مشدَّدة في القرآن الكريم، وكذلك الهمزة لم ترد مشدَّدة في القرآن الكريم.

ب. الحرف المحذوف في القرآن الكريم.

ج. الحرف المنطوق وغير المكتوب من الحروف المقطَّعة في القرآن الكريم.

ومجموع حروف هذا القسم في القرآن الكريم (22620) حرفًا، وهي تمثِّل (7.30%) من حروف القرآن الكريم.

3. **منطوق ابتداءً ومكتوب:** وتندرج تحته همزة الوصل فقط، وقد وردت في (13483) موضعًا من القرآن الكريم، وهي تمثِّل (4.35%) من حروف القرآن الكريم.

4. **منطوق وصلًا وغير مكتوب:** ويندرج تحته التَّنوين المُتَراكِب، والتَّنوين المبدل ميمًا ساكنة.

ومجموع حروف هذا القسم في القرآن الكريم (2250) حرفًا، وهي تمثِّل (0.73%) من حروف القرآن الكريم.

5. **منطوق وقفًا ومكتوب:** وتندرج تحته الباء، والتَّاء، والثَّاء، والدَّال، والذَّال، والطَّاء، والقاف، واللام، والنَّون.

وبشكل عام فإن حروف هذا القسم إما أن تكون مدغمة في غير نفسها، أو أن تكون مخفاة.

ومجموع حروف هذا القسم في القرآن الكريم (7774) حرفًا، وهي تمثِّل (2.51%) من حروف القرآن الكريم.

6. **غير منطوق ومكتوب:** وتندرج تحته الواو المزيدة، والألف من الحروف المقطَّعة، والألف المزيدة، والياء المزيدة.

وبشكل عام فإن حروف هذا القسم تمثِّل الأحرف المزيدة في القرآن الكريم، إضافة إلى الألف غير المنطوقة والمكتوبة من الحروف المقطَّعة في القرآن الكريم.

ومجموع حروف هذا القسم في القرآن الكريم (4000) حرفًا، وهي تمثِّل (1.29%) من حروف القرآن الكريم.

7. **غير منطوق وغير مكتوب:** ويندرج تحته التنوين المتتابع، والسكت.

ومجموع حروف هذا القسم في القرآن الكريم (6648) حرفًا، وهي تمثِّل (2.14%) من حروف القرآن الكريم.

والحركات في القرآن الكريم بالنسبة للنطق خمسة أقسام على غرار ما ذُكر في الحروف، وبالنسبة للكتابة قسمان على غرار ما ذُكر في الحروف أيضًا، ويتركَّب من أقسام المنطوقة والمكتوبة للحركات في القرآن الكريم عشرة أقسام، ثلاثة منها لا تندرج حركة من الحركات تحتها، وهي:

1. منطوقة ابتداءً ومكتوبة.

2. منطوقة وصلًا ومكتوبة.

3. غير منطوقة ومكتوبة.

وسبعة أقسام تندرج الحركات تحتها، وهي:

1. **منطوقة ومكتوبة:** وتندرج تحته الألف، والألف المبدلة واوًا، والألف المبدلة ياءً، والألف الممالة، والواو المدِّيَّة، والياء المدِّيَّة.

وبشكل عام فإن جميع حركات هذا القسم من الحركات الطويلة، ومجموعها في القرآن الكريم (38181) حركة.

2. **منطوقة وغير مكتوبة:** وتندرج تحته الألف، والألف المحذوفة، والواو المدِّيَّة، والواو المدِّيَّة المحذوفة، والياء المدِّيَّة، والياء المدِّيَّة المحذوفة، والفتحة، والفتحة على الحروف المقطَّعة، والضمة، والكسرة، والكسرة على الحروف المقطَّعة.

ومجموع حركات هذا القسم في القرآن الكريم (171059) حركة.

3. **منطوقة ابتداءً وغير مكتوبة:** وتندرج تحته الفتحة، والضمَّة، والكسرة.

وبشكل عام فإن حركات هذا القسم تمثِّل حركة همزة الوصل عند الابتداء بها، ومجموعها في القرآن الكريم (13483) حركة.

4. **منطوقة وصلًا وغير مكتوبة**: وتندرج تحته الفتحة، والفتحة على الحروف المقطَّعة، والضمَّة، والكسرة.

وبشكل عام فإن حركات هذا القسم باستثناء الفتحة على الحروف المقطَّعة تمثِّل الحركات القصيرة الواقعة قبل الحركات الطويلة المنطوقة وقفًا المحذوفة وصلًا التي ستُذكر في القسم الخامس.

ومجموع حركات هذا القسم في القرآن الكريم (6091) حركة.

5. **منطوقة وقفًا ومكتوبة**: وتندرج تحته الألف، وألف العِوَض، والألفات السبعة، والألف المبدلة ياءً، والواو المدِّيَّة، والياء المدِّيَّة.

وبشكل عام فإن جميع حركات هذا القسم من الحركات الطويلة، وتسبِقُها إحدى الحركات القصيرة المذكورة في القسم الرابع.

ومجموع حركات هذا القسم في القرآن الكريم (5905) حركات.

6. **منطوقة وقفًا وغير مكتوبة**: وتندرج تحته ألف العِوَض فقط، وقد وردت في (78) موضعًا من القرآن الكريم.

7. **غير منطوقة وغير مكتوبة**: وتندرج تحته الفتحة، والضمَّة، والكسرة، والسُّكون، والسُّكون على الحروف المقطَّعة، والسُّكون على الحروف المقطَّعة المتحرِّكة وصلًا حال الوقف، والسُّكون (بلا علامة)، والسُّكون (بلا علامة) على الحروف المقطَّعة.

وبشكل عام فإن حركات هذا القسم تمثِّل السكون بصوره المختلفة، أو إحدى الحركات القصيرة الواردة قبل الحركات الطويلة المنطوقة وصلًا ووقفًا التي ذكرت في القسمين الأول والثاني.

ومجموع حركات هذا القسم في القرآن الكريم (127648) حركة.

- مراعاة المكتوب وغير المكتوب من منطوق الحروف المقطَّعة في القرآن الكريم[5]، فمثلاً (صٓ) [ص: 1] سيتم إدخالها في قاعدة البيانات على حرفين، الحرف الأول: الصاد، وهو منطوق ومكتوب، وحركته الألف، وهي منطوقة وغير مكتوبة، والحرف الثاني: الدال، وهو منطوق وغير مكتوب، وهو ساكن، وحتى يُميَّز بين هذا السكون، والسكون الذي تم ضبْطه في القرآن الكريم (ـۡ)، تمت تسميته بالسكون على الحروف المقطَّعة.
- إذا كان للحرف أكثر من صورة تُعدُّ كل صورة للحرف حرفًا مستقِلًّا، فالهمزة على الألف (أ) تُعدُّ

[5] الحروف المقطَّعة في القرآن الكريم: هي (14) حرفًا ابتدأ الله Y بها (29) سورة، الله أعلم بمعناها، مجموعة في قولك: (نَصٌّ حَكِيمٌ قَطْعًا لَهُ سِرٌّ) [22] [23].

حرفًا مستقلًا، وكذلك الهمزة على الواو (ؤ)، والهمزة على الياء (ئ)، ...إلخ.

- لا يتم إثبات البسملة في أوَّل سور القرآن الكريم سوى أوَّل سورة الفاتحة؛ لأن الإجماع انعقد على أنها ليست آية في أوائلهنَّ، قال الداني عن البسملة: "الإجماع لم ينعقد على أنها آية من أول الفاتحة، **وأنه انعقد على أنها ليست آية في سائر السور**، وإن كانت مرسومة في أوائلهنَّ من حيث لم يعدوها مع جملة آيهنَّ" [21]، وعدَّ الكوفيُّون -ومنهم حفص- البسملة آية في أوَّل سورة الفاتحة، وقد التزمت برواية حفص عن عاصم، فأُثبِتت البسملة في أوَّل سورة الفاتحة.
- اعتماد المقطوع في الرَّسْم كلمة مستقلة، نحو قوله تعالى: (أَن لَّا) [الأعراف: 105]، فتُعدُّ (أَن) كلمة مستقلَّة، و(لَّا) كلمة أخرى، بخلاف قوله تعالى: (أَلَّا) [البقرة: 229]، فهي تُعدُّ كلمة واحدة.
- إذا كان لحرفٍ قرآنيٍّ أكثر من وجه سيعتمد الوجه الراجح، نحو وصل هاء(مَالِيَهۜ) بما بعدها في قوله تعالى: (مَالِيَهۜ ٢٨ هَلَكَ) [الحاقة: 28-29]، فيها لحفصٍ وجهان: أحدهما: إظهارها مع السكت، وهو الأرجح، وثانيهما: إدغامها في الهاء التي بعدها في لفظ (هَلَكَ)، واعتُمِدالوجه الأول في قاعدة البيانات؛ لأنه الأرجح [24] [25].
  وإن لم يكن هناك وجه راجح اعتُمِد الوجه الذي ضُبط في المصحف الشريف، نحو وصل آخر سورة الأنفال بأول سورة التوبة، ففيها لحفصٍ ثلاثة أوجه: الأول: القطع، والثاني: السكت، والثالث: الوصل [26]، واعتُمِد الوجه الثالث في قاعدة البيانات؛ لأن ضبط المصحف الشريف على الوصل.

- كل حرف مشدَّد هو عبارة عن حرفين من النوع نفسه، الأول: ساكن، وهو منطوق وغير مكتوب، والثاني: متحرك بحركة الحرف، وهو منطوق ومكتوب، نحو حرف الباء في قوله تعالى: (رَبِّ) [الفاتحة: 2]، فهو يُعد حرفين، الأول: باء ساكنة، وهي منطوقة وغير مكتوبة، والثاني: باء مكسورة، وهي منطوقة ومكتوبة.

  وإذا أتى قبل الحرفِ المشدَّد حرفٌ ساكن ليس من نفس نوع الحرف المشدَّد تُطبَّق القاعدة السابقة مع اعتبار الحرف الساكن الواقع قبل الحرفِ المشدَّد غير منطوق ومكتوب، نحو الراء في قوله تعالى: (ٱلرَّحْمَٰنِ) [الفاتحة: 1]، فهو يُعدُّ حرفين، الأول: راء ساكنة، وهي منطوقة وغير مكتوبة، والثاني: راء مفتوحة، وهي منطوقة ومكتوبة، وتُعدُّ اللام الواقعة قبل الراء المشدَّدة غير منطوقة ومكتوبة.

  ويُستثنى من القاعدة السابقة حالة واحدة لا يُعدُّ فيها الحرف المُشدَّد حرفين بل يُعدُّ حرفًا واحدًا، وهي إذا أتى قبل الحرف المشدد حرف ساكن من نفس نوع الحرف المشدَّد، فيعتبر الحرف الأول: الحرف الساكن، وهو منطوق ومكتوب، والحرف الثاني: الحرف المشدَّد، وحركته حركة الحرف المشدَّد، وهو منطوق ومكتوب أيضًا، نحو اللام المشدَّدة في قوله تعالى: (ٱللَّهِ) [الفاتحة: 1]، فيُعدُّ الحرف الأول: اللام الساكنة الواردة قبل اللام المشدَّدة، وهي منطوقة ومكتوبة، والحرف الثاني: اللام المشدَّدة، وهي منطوقة ومكتوبة أيضًا، وحركتها الفتحة.

- لن يتم تمثيل العلامات الآتية في قاعدة البيانات:
  - o علامة المد (◌ٓ).
  - o علامات الوقف (◌ۖ، ◌ۗ، ◌ۚ،◌ۘ،◌ۛ).

- o علامة موضع السجدة (۩).
- o علامة بداية الأجزاء والأحزاب وأنصافها وأرباعها (۞).

خامسًا: نموذج من جدول الحروف والحركات في القرآن الكريم لثلاث آيات من سورة الفاتحة

يُعدُّ جدول الحروف والحركات في القرآن الكريم الجدول الأساسي في قاعدة البيانات، وسيُكتفى بعرض نموذج منه لثلاث آيات من سورة الفاتحة في جدول (5) بغرض توضيح الفكرة والرغبة في الاختصار.

جدول (5): نموذج من جدول الحروف والحركات في القرآن الكريم لثلاث آيات من سورة الفاتحة

رقم السورة	رقم الآية	رقم الكلمة	رقم الحرف	الحرف(*)	رمز الحرف	الحركة(*)	رمز الحركة	الحركة الإضافية(*)	رمز الحركة الإضافية
1	1	1	1	ب	21	ـِ	71		0
1	1	1	2	س	131	ـَّ	81		0
1	1	1	3	م	251	ـِ	71		0
1	1	2	1	أ	351		0	ـَ	54
1	1	2	2	ل	241	سكون(6)	151		0
1	1	2	3	ل	241	ا	12	ـَ	53
1	1	2	4	ه	271	ـِ	71		0
1	1	3	1	أ	351		0	ـَ	54
1	1	3	2	ل	243	سكون	151		0
1	1	3	3	ر	112	سكون	151		0
1	1	3	4	ر	111	ـَ	51		0

(*): تمَّت إضافة الأعمدة المُظلَّلة رغبة في التوضيح، وهي غير موجودة في أصل جدول الحروف والحركات في القرآن الكريم.

(6) أي سكون –بلا علامة–.

0		81	ـَّ	71	ح	5	3	1	1
53	ـَ	101	◌ٰ	251	م	6	3	1	1
0		71	ـِ	261	ن	7	3	1	1
رمز الحركة الإضافية	الحركة الإضافية	رمز الحركة	الحركة	رمز الحرف	الحرف	رقم الحرف	رقم الكلمة	رقم الآية	رقم السورة
54	ـَ	0		351	أ	1	4	1	1
0		151	سكون	243	ل	2	4	1	1
0		151	سكون	112	ر	3	4	1	1
0		51	ـَ	111	ر	4	4	1	1
73	ـِ	41	ي	71	ح	5	4	1	1
0		71	ـِ	251	م	6	4	1	1
54	ـَ	0		351	أ	1	1	2	1
0		81	ـَّ	241	ل	2	1	2	1
0		51	ـَ	71	ح	3	1	2	1
0		81	ـَّ	251	م	4	1	2	1
0		61	ـُ	91	د	5	1	2	1
0		71	ـِ	241	ل	1	2	2	1
0		151	سكون	242	ل	2	2	2	1
53	ـَ	12	ا	241	ل	3	2	2	1
0		71	ـِ	271	ه	4	2	2	1
0		51	ـَ	111	ر	1	3	2	1
0		151	سكون	22	ب	2	3	2	1

0		71	ـِ	21	ب	3	3	2	1
54	ـَ	0		351	أ	1	4	2	1
0		81	ـُ	241	ل	2	4	2	1
رمز الحركة الإضافية	الحركة الإضافية	رمز الحركة	الحركة	رمز الحرف	الحرف	رقم الحرف	رقم الكلمة	رقم الآية	رقم السورة
53	ـَ	101	◌ٰ	191	ع	3	4	2	1
0		51	ـَ	241	ل	4	4	2	1
73	ـِ	41	ي	251	م	5	4	2	1
0		51	ـَ	261	ن	6	4	2	1
54	ـَ	0		351	أ	1	1	3	1
0		151	سكون	243	ل	2	1	3	1
0		151	سكون	112	ر	3	1	3	1
0		51	ـَ	111	ر	4	1	3	1
0		81	ـُ	71	ح	5	1	3	1
53	ـَ	101	◌ٰ	251	م	6	1	3	1
0		71	ـِ	261	ن	7	1	3	1
54	ـَ	0		351	أ	1	2	3	1
0		151	سكون	243	ل	2	2	3	1
0		151	سكون	112	ر	3	2	3	1
0		51	ـَ	111	ر	4	2	3	1
73	ـِ	41	ي	71	ح	5	2	3	1
0		71	ـِ	251	م	6	2	3	1

سادسًا: شيوع الأصوات في القرآن الكريم

إن إنشاء قاعدة بيانات لحروف القرآن الكريم يفتح بابًا واسعًا للبحث العلمي في القرآن الكريم بشكل خاص واللغة العربية بشكل عام.

ومن ضمن الدراسات المستنتجة من قاعدة البيانات لحروف القرآن الكريم دراسة لشيوع الأصوات في القرآن الكريم سيُتحدَّث عنها هنا بشيء من التفصيل، وهناك العديد من الدراسات التي يمكن استنتاجها من قاعدة البيانات، أو تكون نتائج قاعدة البيانات جزءًا أساسيًّا من دراسة واسعة تتعلَّق بعلم الأصوات، أو الدلالة، أو النحو، أو غير ذلك.

الضوابط التي التُزِمت عند إجراء دراسة شيوع الأصوات في القرآن الكريم:

- الدراسة تتعلَّق بالأصوات فقط، سواء كانت حروفًا أم حركات، مكتوبة أو غير مكتوبة، وبغضِّ النظر عن صورتها التي كتبت بها في القرآن الكريم، وبالتالي لم تتعرَّض الدراسة للحروف المكتوبة وغير المنطوقة مثل الحروف المزيدة.
- تناولت الدراسة الأصوات في سياقها اللغوي، وليس بمعزل عمايجاورها من الأصوات، فأخذت بعين الاعتبار التغيُّرات التي تطرأ على الصوت بتأثير البيئة الصوتية المجاورة، وبناءً عليه لم تُحتسب همزة الوصل في درْج الكلام، وكذلك الأصوات المُدغمة، ونحو ذلك.
- احتساب الحرف المُشدَّد حرفين.

وتجدر الإشارة هنا إلى دراسة قام بها الدكتور محمد علي الخولي لشيوع الأصوات العربية، أجراها على خمسمائة سطر من مئة كتاب حديث النَّشر بواقع خمسة أسطر من الكتاب الواحد، روعي في هذه الكتب أن تكون متنوعة في موضوعاتها، وأن تكون بالعربية الفصحى، والتزم فيها نفس ضوابط هذه الدراسة، وكان مجموع الأصوات التي تم إجراء الدراسة عليها (46029) صوتًا [27]، فأُثبتت نتائج دراسته إلى جانب نتائج دراسة شيوع الأصوات في القرآن الكريم؛ حتى تسهل المقارنة بين نتائج الدراستين.

وهناك دراسات أخرى لشيوع الأصوات، مثل الدراسة التي قام بها الدكتور علي حلمي موسى لمُعْجَمَي الصحاح للجوهري ولسان العرب لابن منظور إلا أنها خاصة بجذور مفردات اللغة العربية، ومثلها إشارة ابن منظور في معجمه لسان العرب إلى ما يتكرر من الحروف ويكثر في الكلام استعماله، وما هو دون ذلك [28].

كما اعتنى بعض علماء علوم القرآن بعدِّ حروف القرآن الكريم إلا أنهم اقتصروا على عدِّ المكتوب دون المنطوق [21] [29]، وبعضهم عدَّ الحرف المشدَّد بحرفين، وعدَّه آخرون حرفًا واحدًا [29]، وكذلك اكتفى بعضهم بذكر إجمالي عدد الحروف، ولم يذكر عدد مرات تكرار كل حرفٍ بشكل مستقلٍّ [21] [30]، ومن فصَّل منهم اقتصر على عدِّ الحروف دون الحركات القصيرة [31] [29] –فيما اطَّلعت عليه، وكل ذلك يخالف الضوابط التي وضِعت للدراسة.

ويُبيِّن جدول (6) الترتيب التنازلي لشيوع الأصوات في القرآن الكريم، موضِّحًا صورة الصوت، وترتيبه، وعدد مرات تكراره، والنسبة المئوية له في القرآن الكريم، إضافة إلى النسبة المئوية للصوت وترتيبه حسب الدراسة التي أجراها الدكتور محمد الخولي لنصوص لغويَّة متنوعة.

جدول (6): الترتيب التنازلي لشيوع الأصوات في القرآن الكريم

الصوت	صورة الصوت	الترتيب	العدد	النسبة المئوية	النسبة المئوية	الترتيب
	في القرآن الكريم				في اللغة[7]	
الفتحة	ـَ	1	96781	19.63	16.74	1
اللام	ل	2	39740	8.06	7.56	3
الكسرة	ـِ	3	38713	7.85	10.63	2
الألف	ا، ◌ٰ، وٰ،ىٰ،◌ٖىٰ	4	29965	6.08	5.89	4
الميم	م،نۢ، ـِۢـُۢ	5	29787	6.04	4.37	8
الضمة	ـُ	6	29344	5.95	5.42	6
النون	ن، ◌ٰن، ◌ۨ، ـًـٌـٍ	7	27876	5.65	4.51	7
الهمزة	أ، إ، ◌ٰٔ◌، ؤ،ۄ، ئ،ي، ء، أ	8	19184	3.89	4.16	9
الواو غير المدية	و	9	15035	3.05	2.79	12
الهاء	هـ	10	14890	3.02	2.36	14
التاء	ت، ة	11	13664	2.77	5.75	5
الراء	ر	12	13479	2.73	3	11
الياء غير المدية	ي، ۦ	13	13199	2.68	3.44	10
الباء	ب	14	12768	2.59	2.27	15
الكاف	ك	15	10707	2.17	1.51	20
الواو المدِّيَّة	و، ۥ	16	10496	2.13	0.87	24
الياء المدِّيَّة	ي، ۦ	17	10032	2.04	2.26	16
العين	ع	18	9414	1.91	2.51	13

(7) حسب الدراسة التي قام بها الدكتور محمد الخولي [27].

الصوت	صورة الصوت	الترتيب	العدد	النسبة المئوية	النسبة المئوية	الترتيب
	في القرآن الكريم				في اللغة	
الفاء	ف	19	8915	1.81	1.92	18
القاف	ق	20	7364	1.49	1.20	21
السين	س،صۜ	21	6689	1.36	1.54	19
الدال	د	22	6577	1.33	2.19	17
الذال	ذ	23	5268	1.07	0.66	27
الحاء	ح	24	4150	0.84	0.89	22
الجيم	ج	25	3435	0.70	0.89	22
الخاء	خ	26	2543	0.52	0.51	29
الصاد	ص،صَۜ	27	2431	0.49	0.74	26
الشين	ش	28	2388	0.48	0.47	31
الزاي	ز	29	1777	0.36	0.29	34
الضاد	ض	30	1766	0.36	0.45	32
الثاء	ث	31	1451	0.29	0.56	28
الطاء	ط	32	1406	0.29	0.8	25
الغين	غ	33	1221	0.25	0.51	30
الظاء	ظ	34	1008	0.20	0.36	33
إجمالي عدد الأصوات في القرآن			493463 صوتًا			

بعد الاطلاع على جدول الترتيب التنازلي لشيوع الأصوات في القرآن الكريم، ومقارنة نتائجه بنتائج دراسة شيوع الأصوات في اللغة يمكن استنتاج الآتي:

- هناك تقارب مقبول بين نتائج دراسة شيوع الأصوات في القرآن الكريم ودراسة شيوع الأصوات في اللغة مع وجود فارق بين نتائج الأصوات الآتية: التاء، والواو المدِّيَّة، والطاء.

ويُحتمل أن تكون النتائج أكثر تقاربًا لو زيدتالأصوات التي أُجريت عليها دراسة شيوع الأصوات في اللغة.

- عند تعارض الدراستين تُقدَّم دراسة شيوع الأصوات في القرآن الكريم لسببين، هما:
  - أن دراسة شيوع الأصوات في القرآن الكريم أُجريت على أبلغ نصٍّ لُغوي وأصحِّه، في حين أن دراسة شيوع الأصوات في اللغة أُجريت على عيِّنة شبه عشوائية.
  - أن عدد الأصوات التي أُجريت عليها دراسة شيوع الأصوات في القرآن الكريم هو (493463) صوتًا، وهو يزيد بأكثر من عشرة أضعاف عن عدد الأصوات التي أجريت عليها دراسة شيوع الأصوات في اللغة، وهو (46029) صوتًا.

- اعتماد ترتيب الأصوات الوارد في جدول (6) لشيوع الأصوات في القرآن الكريم خصوصًا وفي اللغة عمومًا[8]، وبناء عليه يُقترح توزيع الأصوات حسب شيوعها إلى خمس مجموعات على النحو الآتي:
  - **الصوت الأكثر شيوعًا:** الفتحة؛ وذُكرت في مجموعة لوحدها؛ بسبب وجود فارق كبير يزيد عن الضعف بين نسبة ورودها وورود الصوت الذي يليها سواء في القرآن الكريم أو اللغة.
  - **الأصوات الشائعة:** اللام، والكسرة، والألف، والميم، والضمَّة، والنون.
  - **الأصوات متوسطة الشيوع:**الهمزة، والواو غير المدِّيَّة، والهاء، والتاء، والراء، والياء غير المدِّيَّة، والباء، والكاف، والواو المدِّيَّة، والياء المدِّيَّة.
  - **الأصوات متدنِّية الشيوع:** العين، والفاء، والقاف، والسين، والدال، والذال، والحاء، والجيم.
  - **الأصوات الأقل شيوعًا:** الخاء، والصاد، والشين، والزاي، والضاد، والثاء، والطاء، والغين، والظاء.

  وترتيب الأصوات في المجموعات حسب نسبة شيوعها من الأكثر إلى الأقل.

[8] إعداد دراسات أخرى لشيوع الأصوات خاصة في الحديث النَّبوي الشريف وبعض النُّصوص الأدبية سيضيف إلى النتائج التي تمَّ الوصول إليها، ويساعد على تحديد الأصوات الشائعة في اللغة الفصحى بشكل نهائي.

خاتمة:

- إن إنشاء قاعدة بيانات لحروف القرآن الكريم يفتح بابًا واسعًا للبحث العلمي في القرآن الكريم بشكل خاص واللغة العربية بشكل عام، وهناك العديد من الدراسات التي يمكن استنتاجها من قاعدة البيانات، أو تكون نتائج قاعدة البيانات جزءًا أساسيًّا من دراسة واسعة تتعلَّق بعلم الأصوات، أو الدلالة، أو النحو، أو غير ذلك.

  وإنشاء قواعد بيانات أخرى خاصة للحديث النَّبوي الشريف وبعض النُّصوص الأدبية على غرار قاعدة بيانات حروف القرآن الكريم سيضيف إلى النتائج التي تمَّ الوصول إليها، ويمكِّن من الوصول إلى نظريَّات جديدة تتعلَّق بالقرآن الكريم أو اللغة.

- حروف القرآن الكريم بالنسبة للنطق والكتابة سبعة أقسام، وهي: منطوق ومكتوب، ومنطوق وغير مكتوب، ومنطوق ابتداءً ومكتوب، ومنطوق وصلًا وغير مكتوب، ومنطوق وقفًا ومكتوب، وغير منطوق ومكتوب، وغير منطوق وغير مكتوب، ولم يرد في القرآن الكريم حرف منطوق ابتداءً وغير مكتوب، أو منطوق وصلًا ومكتوب، أو منطوق وقفًا وغير مكتوب.

- الحروف المنطوقة والمكتوبة هي الأكثر تكرارًا في القرآن الكريم، فهي تمثِّل (81.69%) من إجمالي حروفه.

- لم تَرِد الحروف الآتية مشدَّدة في القرآن الكريم، وهي: الهمزة، والتاء المربوطة، والغين.

- الحركات في القرآن الكريم بالنسبة للنطق والكتابة سبعة أقسام أيضًا، وهي: منطوقة ومكتوبة، منطوقة وغير مكتوبة، ومنطوقة ابتداءً وغير مكتوبة، ومنطوقة وصلًا وغير مكتوبة، ومنطوقة وقفًا ومكتوبة، منطوقة وقفًا وغير مكتوبة، وغير منطوقة وغير مكتوبة، ولم يرد في القرآن الكريم حركة منطوقة ابتداءً ومكتوبة، أو منطوقة وصلًا ومكتوبة، أو غير منطوقة ومكتوبة.

- توزيع الأصوات حسب شيوعها في القرآن الكريم خصوصًا واللغة عمومًا إلى خمس مجموعات على النحو الآتي:
  - الصوت الأكثر شيوعًا: الفتحة؛ وذُكرت في مجموعة لوحدها؛ بسبب وجود فارق كبير يزيد عن الضعف بين نسبة ورودها وورود الصوت الذي يليها سواء في القرآن الكريم أو اللغة.
  - الأصوات الشائعة: اللام، والكسرة، والألف، والميم، والضمَّة، والنون.
  - الأصوات متوسطة الشيوع: الهمزة، والواو غير المدِّيَّة، والهاء، والتاء، والراء، والياء غير المدِّيَّة، والباء، والكاف، والواو المدِّيَّة، والياء المدِّيَّة.
  - الأصوات متدنِّية الشيوع: العين، والفاء، والقاف، والسين، والدال، والذال، والحاء، والجيم.
  - الأصوات الأقل شيوعًا: الخاء، والصاد، والشين، والزاي، والضاد، والثاء، والطاء، والغين، والظاء.

  وترتيب الأصوات في المجموعات حسب نسبة شيوعها من الأكثر إلى الأقل.

قائمة المصادر والمراجع

[1] الأصوات العربية بين القدماء والمُحدثين: عادل إبراهيم أبو شعر، رسالة ماجستير، جامعة أم القرى- كلية اللغة العربية، 1415هـ.

[2] موقع قرآن داتاباس (Quran Database): http://qurandatabase.org.

[3] موقع إسلام وير: https://www.islamware.com.

[4] موقع مركز تحميل البرامج: http://soft.sptechs.com.

[5] موقع انتلرن (Intellaren): http://www.intellaren.com.

[6] موقع آية: http://a-yah.com.

[7] موقع كود بلكس (CodePlex): http://qurancode.codeplex.com.

[8] موقع شبكة الخدمات الإسلامية: http://www.islamic-services.net.

[9] موقع الأرقام: http://www.alargam.com.

[10] موقع ملتقى أهل الحديث: http://www.ahlalhdeeth.com.

[11] موقع مجمع الملك فهد لطباعة المصحف الشريف – مصحف المدينة النبوية (خدمات حاسوبية للقرآن الكريم وعلومه): http://mushaf-services.qurancomplex.gov.sa.

[12] موقع علم القرآن الكريم: http://www.ketaballah.net.

[13] موقع الفانوس: http://www.alfanous.org.

[14] موقع الأوفى: http://www.alawfa.com.

[15] موقع صفحة القرآن الكريم: http://www.holyquran.net.

[16] موقع الباحث القرآني: http://www.quranicresearcher.com.

[17] المقنع في معرفة مرسوم مصاحف أهل الأمصار: أبو عمرو عثمان بن سعيد الداني، تحقيق: محمد أحمد دهمان، دار الفكر، دمشق 1403هـ-1983م.

[18] سمير الطالبين في رسم وضبط الكتاب المبين: علي محمد الضباع، ط1، المكتبة الأزهرية للتراث، القاهرة 1420هـ-1999م.

[19] قواعد بيانات القرآن الكريم كأساس للمعجم الآلي الموسَّع للغة العربية: أ. د. محمد زكي خضر، ندوة اتحاد المجامع اللغوية العربية، عمَّان، 2002م.

[20] مقدمة في قواعد البيانات: وائل عادل الصلوي، 2012م.

[21] البيان في عدِّ آي القرآن: أبو عمرو الداني، تحقيق: د. غانم قدوري الحمد، ط1، مركز المخطوطات والتراث والوثائق، الكويت 1414هـ-1994م.

[22] المنير في أحكام التجويد: لجنة التلاوة بجمعية المحافظة على القرآن الكريم بالأردن، ط15، جمعية المحافظة على القرآن الكريم، عمَّان 1430هـ-2009م.

[23] التجويد المصور: د. أيمن رشدي سويد، ط2، مكتبة ابن الجزري، دمشق 1432هـ-2011م.

[24] الرعاية لتجويد القراءة وتحقيق لفظ التلاوة: مكي بن أبي طالب القيسي، تحقيق: د. أحمد حسن فرحات، ط4، دار عمار، عمّان 1422هـ-2001م.

[25] إبراز المعاني من حرز الأماني: عبد الرحمن بن إسماعيل بن إبراهيم المعروف بأبي شامة، تحقيق: إبراهيم عطوه عوض، مطبعة مصطفى البابي الحلبي، القاهرة 1402هـ-1982م.

[26] هداية القاري إلى تجويد كلام الباري: عبد الفتاح السيد عجمي المرصفي، ط2، مكتبة طيبة، المدينة المنورة.

[27] الأصوات اللغوية: د. محمد علي الخولي، دار الفلاح، عمَّان.

[28] لسان العرب: ابن منظور، ط3، دار إحياء التراث العربي ومؤسسة التاريخ الإسلامي، بيروت.

[29] القول الوجيز في فواصل الكتاب العزيز: رضوان بن محمد بن سليمان المعروف بالمخلِّلاتي، تحقيق: عبد الرزاق بن علي بن إبراهيم موسى، ط1، مطابع الرشيد، المدينة المنورة 1412هـ-1992م.

[30] منار الهدى في بيان الوقف والابتدا: أحمد بن محمد بن عبد الكريم الأشموني، ط2، مطبعة مصطفى البابي الحلبي، القاهرة 1393هـ-1973م.

[31] فنون الأفنان في عيون علوم القرآن: أبو الفرج عبد الرحمن بن الجوزي، تحقيق: د. حسن ضياء الدين عتر، ط1، دار البشائر الإسلامية، بيروت 1408هـ-1987م.

السيرة الذاتية

محمد عبد الرحمن محمد الخطيب

المؤهلات العلمية: ماجستير آداب لغة عربية، الجمهورية اليمنية، الجامعة الوطنية، كلية الآداب والتربية - بكالوريوس علوم حاسوب، الجمهورية اليمنية، جامعة العلوم والتكنولوجيا، كلية العلوم والهندسة - بكالوريوس آداب لغة عربية، الجمهورية اليمنية، جامعة سبأ، كلية الآداب والتربية.

الإجازات القرآنية : شهادة إجازة في القراءات العشر من طريقي الشاطبية والدرة وشهادة إجازة في قراءتي ابن عامر وعاصم وشهادة إجازة في رواية حفص عن عاصم من طريق الطيبة وشهادة إجازة في رواية حفص عن عاصم من طريق الشاطبية.

الخبرات : موظف في الهيئة العالمية لتحفيظ القرآن الكريم، وخلال عملي تقلدت المناصب الآتية: رئيس القسم التعليمي ورئيس قسم غرب أفريقيا ونائب رئيس لجنة تقنية المعلومات وباحث في الإدارة التعليمية.

الإنجازات : إنشاء قاعدة بيانات لحروف القرآن الكريم - تحليل وتصميم وتنفيذ موقع تعليمي للقراءات القرآنية - تحليل وتصميم وتنفيذ نظام إدارة حلقات تحفيظ القرآن الكريم.

العضويات : عضوية الحفاظ المجازين بالقراءات العشر الصغرى وعضوية الهيئة السعودية للمهندسين.

Creating a Database for the Holy Quran Letters

Mohammed Abdul Rahman AlKhateib
PhD Student
Faculty of Dar El-Ulum - Cairo University
Makh2000@hotmail.com

Abstract: This study introduces a novel scientific approach for generating a database of the Holy Quran letters, taking into consideration all Quranic, linguistic and computing aspects that are related to it. The main advantage of this database is that it recognizes all different cases of the Holy Quran letters and Harakat including spoken, silent, written, and unwritten letter or Harakah. In addition to that it recognizes the special cases that are related to the Othmani style of writing. To the best of my knowledge, there has not been any database of the Holy Quran letters, and I think the presence of a database, such as the one proposed in this study, will be of importance in both Quranic and Linguistic studies. Moreover, this study will show a sample of the database, and how it is presented on its final stage at the application level.

Arabic Ontology Using Different Ontology Learning Techniques

Dalia Fadl[*1], Safia Abas[*2], Mostafa Aref[*3]

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

[1]Dalia_sayed_43@hotmail.com
[2]safiaabas@yahoo.com
[3]Mostafa.m.Aref@gmail.com

***Abstract*— Arabic Language has a set of specialties made it difficult language and may obstruct the development of SW tools for it. Among these specialties, its complex morphological, grammatical, and semantic aspects since it is a highly inflectional and derivational language. Arabic is the official language of hundreds of millions of people in twenty Middle East and northern African countries. It is the religious language of all Muslims of various ethnicities around the world. Surprisingly, little has been done in the field of computerized language and lexical resources. Arabic is a Semitic language which differs from European languages syntactically, morphologically and semantically. The term 'classical Arabic' refers to the standard form of the language used in all writing and heard on television, radio and in public speeches and religious sermons. The goal of this research is to discuss criteria for designing Ontology for Arabic language. It takes an engineering perspective on the development of ontologies in Arabic language. It investigates different types of the ontology development life cycle.**

1 INTRODUCTION

Ontologies are of basic interest in many different fields, largely due to what they promise: a shared and common understanding of some domain that can be the basis for communication ground across the gaps between people and computers. Ontology approaches allow for sharing and reuse of knowledge bodies in computational form. As many traditional activities are changing their manner in the world of today due to the availability of information brought by the World-Wide-Web (WWW), Ontologies are likely to change more when the knowledge is structured in machine readable way, and the abstracts concepts it contains are shared.

In computer Science, ontology is a shared and common understanding of some domain that can be communicated across people and application systems or enabling knowledge sharing. It is a specification of a conceptualization. The rise of linguistic ontologies is a result of two concurrent situations: Information structuring and representation. They facilitate its exploitation by users later. This is the topic of ontologies. In the same time, language is the way to vehicle information and knowledge. So the need for linguistic data is crucial in all research fields. This fields are concerned by the organization of information and its retrieval for the end user [1][2]. In this paper we are going to discuss the ontology development life cycle. State the Arabic Language & Semantic web research and finally list some of the related work.

2 ONTOLOGY DEVELOPMENT LIFE CYCLE

There are six parts in the life cycle in the development of ontology: Creation, Population, Validation, Deployment, Maintenance and Evolution [1]. The 6 parts above can also be subdivided into the following: extracting terms, discovering synonyms, obtaining concepts, extracting concept hierarchies, defining relations among concepts, deducing rules or axioms. These processes are used in order to make the ontology matching become possible. And that the related branches of topics would be available to any users [2]. Manual ontology building is a time consuming activity that requires a lot of efforts for knowledge domain acquisition and knowledge domain modeling. In order to overcome these problems many methods have been developed, including automatically or semi-automatically systems and tools. They use text mining and machine learning techniques to generate ontologies. The research fields which study this issues is usually called "ontology generation", "ontology extraction" or "ontology learning". It studies the methods and techniques used to:

- construct automatically or semi-automatically an ontology, and
- enrich or adapt an existing ontology using different sources.

The ontology learning process is useful for different reasons. First of all, it accelerates the process of knowledge acquisition. Second, it reduces the time for the updating of an existent ontology. Finally, it accelerates the whole process of ontology building [9][10]. There are a few types of ontologies which have different roles. In some cases, discussion goes to a mess because of the ignorance of what type of ontology is under consideration. Some say "ontology is domain-specific like a knowledge base which was a failure". Others say "No, it isn't. Ontology is very generic and hence it is widely applicable and sharable". Both are correct because they are talking about different types of ontology.

Upper Ontologies are harder to design than domain ontologies in a certain respect. They are generally both more granular and more macroscopic. Generally, the concepts they define are more abstract and often epistemological in nature. While someone may be a domain expert in their own field and be able to design a fairly decent ontology about their domain, designing a truly suitable Upper Ontology is a different specialization altogether. Ontology-based linguistic resources are valuable for any natural language processing application, especially Semantic Web applications [4]. Acquiring domain knowledge for building ontologies is highly costly and time consuming. For this reason lots of methods and techniques have been developed for trying to reduce such efforts. The mapping between lexical items (words or multiwords) and concepts can be complex. Due to polysemy, most lexical items can be mapped into more than one concept. Due to synonymy, more than one word can be mapped to a concept.

The absence of free usable lexical and syntactic resources and tools for Arabic makes it a "pi- language" (poorly informative). This constitutes a real difficulty in the process of transferring technology into Arabic. There is a strong need for Arabic language support since the ontology in English cannot be translated to Arabic [7][8]. Ontology has proved their success in multiple domains, such as Medicine, e-Commerce, e-Learning and Biology. To extend this success to the Arabic language, a set of ontology tools and applications needs to be created to fulfill the requirements of the Arabic language. Thus, the research questions we are trying to answer:

- Can existing Ontology tools facilitate building of ontology applications that support the Arabic language?
- What are the challenges to create ontology of the Arabic language?
- What the difference between Arabic language and other language like English and French?

3 ARABIC LANGUAGE AND SEMANTIC WEB RESEARCH

There are various studies conducted on Arabic language in Semantic Web. Zaidi, Laskri and Bechkoum proposed to improve the Arabic information retrieval on the Web in the legal domain by an Arabic search engine supporting the translation of Arabic queries into English or French queries. The aim was to return documents written in Arabic, French or English. Vossen, Pease and Fellbaum worked on Arabic Word Net (AWN) based on the methods developed for EuroWordNet (EWN) and since applied to dozens of languages around the world. The EuroWordNet approach maximizes compatibility across Word Nets and focuses on manual encoding of the most complicated and important concepts. The basic criteria for AWN are connectivity, relevance, and generality, from English to Arabic and from Arabic to English. Hammo surveys on enhancing retrieval effectiveness of search engines for diacritised Arabic documents by building an Arabic– English IR system based on a machine translation approach. AbdulJaleel and Larkey, proposed a statistical transliteration approach for Arabic–English IR.

Grefenstette et al, described the changes required to modify their cross language IR system, which has been designed for European languages to integrate Arabic language. Abdelali et al, described how precision can be improved in query expansion using LSI. Finally, Semmar and Fluhr, presented a new approach to align Arabic–French sentences retrieved from a parallel corpus based on a cross-language IR system. This approach is basically based on building a database of sentences of the target text and considering each sentence of the source text as a query to that database. Guo and Ren highlighted the use of Natural Language Processing (NLP) technology as a significant component in Semantic Web tool. NLP is one branch of the linguistics, which uses the computer technology to realize human language processing effectively. Its ultimate objective is to automatically understand human language with the support of artificial intelligence technology. It is also called as natural language understanding and sometimes is used to transform information to Semantic Web data. Traditional information retrieval also can be turned into knowledge discovery. Al-Khalifa, Hen, Al-Yahya, Bahanshal and Al- Odah proposed a framework for representing a semantic opposition in the Holy Quran using Semantic Web Technologies. Previous research in the field of Computers and the Holy Quran can be classified into six categories, namely: Information Retrieval, Speech Recognition, Optical Character Recognition, Morphology Analysis, Semantic checking and Educational Applications. Very little work has been done toward using semantic web technologies for serving the lexical semantics of the Holy Quran. Hammo, Abu-Salem and Lytinen developed a system QARAB whose main goal is to identify text passages that answer a natural language question. The tasks in QARAB can be summarized as follows: Given a set of questions expressed in Arabic, find answers to the questions under the following assumptions: • The answer exists in a collection of Arabic newspaper text extracted from the Al-Raya newspaper published in Qatar.The answer does not span through documents (i.e. all supporting information for the answer lies in one document)The answer is a short passage.

These are just a few studies conducted directly or indirectly in Semantic web in Arabic language. Based on the information gathered, it can be concluded that work in Arabic language for Semantic Web is still in the infancy. Due to that, it is possible to progress further besides the current available Arabic semantics like those that are used in the Quran [16].

4 ARABIC ONTOLOGY RELATED WORK

Arabic ontology is the foundation of the creation of Semantic Web in Arabic language. Basic categorization of terminologies and meanings in a domain give the semantics. The interrelationship between one word to the other words that matches to its meaning can also result to the stems and branches of semantics. Ontology can be built by using domain experts or learned from information available in a corpus of the domain. The goal of ontology learning is to automatically extract relevant concepts and relations from the given corpus or other kinds of data sets to form Ontology [9].

The process of developing ontology can be subdivided into the following: extracting terms, discovering synonyms, obtaining concepts, extracting concept hierarchies, defining relations among concepts, deducing rules or axioms. These processes are used in order to make the ontology matching become possible and that the related branches of topics would be available to any users. Using the different languages in the study of Ontology can also be a challenge to the many attempts of the Web designs to cater the thousands of users in the World Wide Web. Web information is usually language dependent; and the availability of information related to the language that would be much preferable according to the user would be an increasing need of today. There is a strong need for Arabic language support since the ontology in English cannot be translated to Arabic. Figure 2 shows the ontology for ecommerce domain. Different languages have contained the specific linguistic environment and the cultural context, which has caused the need to develop different ontology for different information language.

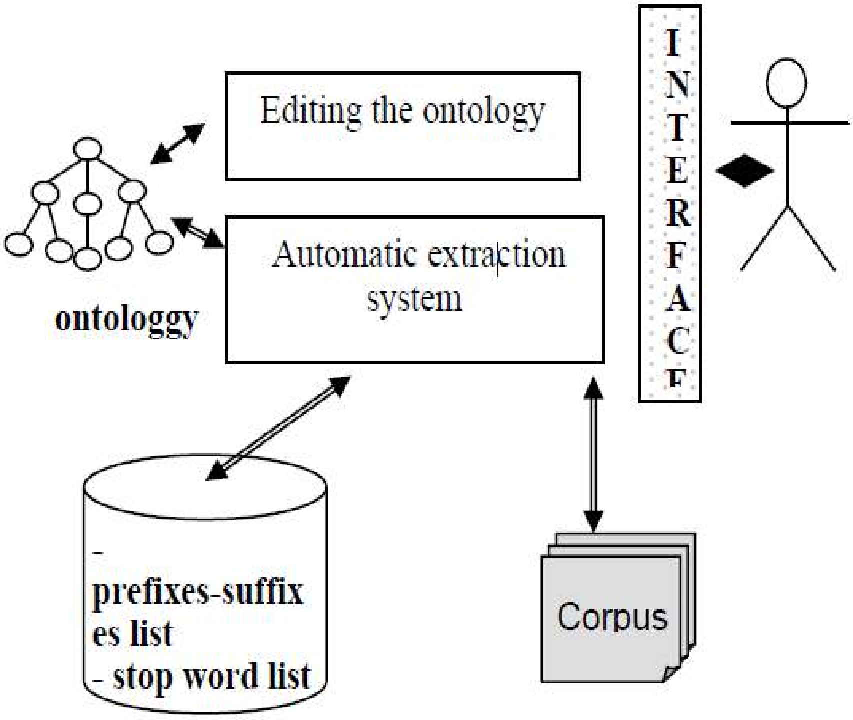


Figure 2: Ontology for ecommerce domain.

A. Manually Developing Ontologies

The first and the most obvious way to build an Ontology from "scratch", i.e. to define classes, relations instants and so on.

1) An Ontological Model For Representing Semantic Lexicons: An Application On Time Nouns In The Holy Quran

Although Arabic is the language of over two hundred million speakers, little has been achieved in regards to computational Arabic resources, especially lexicons. Most of what has been developed was originally tailored for Roman languages, and is not necessarily satisfactory for the Arabic community. In this research, they introduce a computational model for representing Arabic lexicons using ontologies. Ontologies are knowledge representation structures which form the central building block of the Semantic Web. The model is based on the field theory of semantics from the linguistics domain, and the data which drives the design of the model is obtained from the most accurate text that presents superiority and perfection of the Arabic language, the Holy Quran. Creating such lexicons will be invaluable in a number of Arabic applications. This paper presents the design and implementation of the proposed ontological model. Results of its application on "Time nouns" vocabulary of the Holy Quran are presented.

Results show that the model was able to cope with new nouns; however, some semantic dimensions were added to the model to accommodate new features. Lexical relations were also checked to verify that the model captures them sufficiently. Although there exists a dimension of "dynamism" which has two values, *static* and *dynamic*, the model does not capture the element of temporal sequencing in time. For example, *summer* follows *spring*, and *today* comes before *tomorrow*.
the model assumes that if X *isPartOf*Y, then this implies that features associated with word x are also features for word Y, and are added to the componential formula. However, during the evaluation it appears that there exists words for which this statement is not true. For example, *summer isPartOf year;* however, *year hasFeature abstract*, while *summer has Feature concrete*. Therefore, this *Embodiment* feature cannot be inherited. A proposed resolution to this issue is to attach certain properties to *features* which describe the nature of these *features*, whether they are inheritable (shared) or not. Another interesting finding from our evaluation is what we refer to as *the dispersion effect*. This effect occurs when there is minimal or no inclusion relationship within nouns of a specific semantic field, which results in a shallow and wide structure, instead of a deep and narrow one. This means that componential formula will be extremely short; therefore, meaning representation is not sufficient. With regards to "Vague" nouns, this effect is apparent. The componential formulae were very short, thus not giving depth of meaning as is the case with "Day" nouns. When we applied the model on nouns from a different semantic field "Human", we also observed the *dispersion effect*.

the findings of a limited number of features (*semantic richness*) for concrete nouns *vs*abstract nouns, support those reported in the literature. Studies show that words referring to concrete semantic units have richer semantics than abstract ones, and within concrete semantic units, living things have more features than nonliving (artifact) things. Another important finding from our evaluation is that within the "Human" semantic field, it was difficult to identify semantic dimensions. This may be due to two reasons: the fact that the "Human" semantic field is a very large field, and the sample chosen is not focused on a specific domain within the "Human" semantic field. The ontology proposed in this research is unique in representing componential analysis of Arabic vocabulary.

Traditional approaches to Arabic language computational models were based on models of Roman languages. However, our proposed model has originated from an authoritative and rich source of Arabic language, *i.e.*, the Holy Quran. We do not claim that our model is comprehensive. However, we focused on the area where others have not tapped into, that of componential analysis. Additionally, since our model is implemented in OWL, it can easily be extended and linked to other ontologies such as SUMO, LMF, and LexInfo. Furthermore, we believe that such a model for representing Arabic lexicons will enable the creation of a plethora of useful applications for processing Arabic natural language. Such applications include simplifying Arabic language teaching for non-Arabic speakers and building intelligent Arabic dictionaries.
Finally, the results of our work can be summarized as follows:

- Finding appropriate semantic primitives (dimensions) was simpler in concrete concepts and nouns. However, this was not the case with abstract concepts.
- The evaluation also highlights some difficulties associated with this approach to semantics; for example, identifying semantic dimensions, and those which have polarities was difficult.
- Although the lexicon is built based on Time nouns in the Holy Quran, the model is capable of accommodating any Time noun in the Arabic language.

The paper presented an ontological model for a computational lexicon capable of representing Arabic language lexicons in a way which provides a foundation for building useful Arabic language applications using Semantic Web technologies. The model has been implemented on the Arabic language vocabulary associated with "Time" vocabulary in the Holy Quran. Results of the evaluation indicate that the model is capable of representing word semantics in a way that can facilitate semantic analysis of Arabic words and various useful applications. The next natural step is to extend the model into other semantic fields and see how it can accommodate them. Since componential analysis and ontology population are human intensive processes, a major direction in future work is looking into strategies for automated ontology population using technologies such as Latent Semantic Analysis and Formal Concept Analysis. In addition, we plan to develop semantic web applications capable of exploiting the rich structure of the ontological model. We intend to develop an application which automatically performs semantic analysis of words. Another useful application on the horizon is word positioning within the semantic field (classification) based on known features of the word. Classifying a new word in the lexicon is not a simple task. However, using the proposed model, the linguist needs only to select certain features and the application can automatically detect the appropriate classification, and suggest it to the user. Moreover, visualizations of the ontology are useful for linguists in observing semantic field characteristics and language behavior in a certain field, such as word density, word movement, and other attributes of a semantic field.[22]

2) *Al –Khalil: The Arabic Linguistic Ontology Project*

This paper presents a project to building an ontology centered infrastructure for Arabic resources and applications. The core of this infrastructure is a linguistic ontology that is founded on Arabic Traditional Grammar. The methodology they have chosen consists in reusing an existing ontology, namely the Gold linguistic ontology. They discuss the development of the ontology and present our vision for the whole project which aims at using this ontology for creating tools and resources for both linguists and NLP researchers.

Al-khalil is an OWL ontology under development. they have baptized the project Al-Khalil in the sake of the famous grammarian AL-Khalil Ibn Ahmad Alfarahidi because they consider in some sense he was the first to have built an ontology for the Arabic language trough his “kitabalayn” which means the book of the letter ع. the name came from the fact that the dictionary follows a phonetic order starting from the pharyngeal sound ع they have chosen to build our ontology on an existing linguistic ontology namely the Gold ontology. The development of our ontology is two steps:

- Bootstrapping manually the ontology by choosing the linguistic concepts from Arabic linguistics and relating them to the concepts in GOLD.
- Using an automatic extraction algorithm to extract new concepts from linguistic texts to enrich the ontology.

The algorithm is based on the repeated segments calculus method. The general architecture of the In constructing the first prototype of our ontology we have focused on the concepts of Arabic Traditional Grammar that don‟t appear in other linguistic theories such as mital, qiyas, lafda, … and other concepts pertaining to the Neo-khalilean framework which is a modern interpretation of Arabic Traditional Grammar . We make this difference because in the future we aim at:

- Building a community of practice (cope) for the Neo-khalilean school of Arabic traditional grammar. A cope is a subontology that inherits from and extends the overall gold ontology. Subontology classes are distinguished from each other by different name space prefixes, for example gold:noun, hpsg: noun, ATG: noun, ism.
- Extending the content of the ontology. Indeed, as the ontology is intended to be a reference for linguists and NLP researchers in different areas of the field, we aim the ontology to contain exhaustive knowledge about standard Arabic, formal and NLP works on Arabic, dialects and linguistic phenomena relating to Arabic,
- Linking our ontology to projects on Arabic corpus for instance the Algerian Arabic treasury project an building significant applications that use the ontology [23].

Manual acquisition of ontologies is a tedious and cumbersome task. Ontology learning aims to accelerate the time and reduce the effort of building ontology by acquiring concepts and relations semi-automatically or automatically from different information sources such as databases, documents, and/or web pages.

B. Semi-Automatic Developing Ontologies

1) Ontology Learning from Textual Web Documents

Domain ontology plays an important role in annotating web resources with proper semantic information. The underlying assumption behind this work is that the noun phrases appearing in the headings of a document as well as the document’s hierarchical structure can be used to discover the concepts and is-a relations between them in the documents’ domain. In order to verify this assumption a methodology was proposed, and a system was implemented and applied on a set of Arabic agricultural extension documents. The system takes as input a root concept, analyzes all input documents’ heading structure, extracts concepts from headings and builds a taxonomical ontology. The resulting ontology was verified against a modified version of AGROVOC ontology, which is a hand-made ontology developed by Food and Agriculture Organization of the United Nation (FAO). The F-score obtained was 52.29% for lexical evaluation of diseases ontology and 39.64% for lexical evaluation of insects' ontology. Taxonomical F-score was 44.59% for diseases ontology and 31.38% for insects' ontology The objective of our research was to accelerate and improve the Ontology development process by semiautomatically generating a hierarchal ontology. An ontology learning system has been built and tested on web page documents to achieve this objective. Our system generates an ontology from heading titles given a set of web documents using information that exists in the title’s text as well as the HTML structure. The best obtained result (F-score) was 61.14% (precision = 58.18% & recall = 64.65%) in lexical evaluation and 44.9% (precision = 55.43% & recall = 37.73%) in taxonomic evaluation.

Refining the generated ontology did not lead to improvement in lexical and taxonomical evaluation f_score metric. However, it improved the precision significantly at the price of recall. they are now investigating the use of other refinement rules to improve both recall and precision. Another approach that we are investigating is to let a domain expert refine the ontology produced using the merged method then use this ontology as a core ontology that can be extended by analyzing more documents [24].

2) Building a Framework for Arabic Ontology Learning

This paper presents the ArOntoLearn a Framework for Arabic Ontology learning from textual resources. Supporting Arabic language and using domain knowledge or previous knowledge in the learning process are the main features of the framework, besides it represents the learned ontology in Probabilistic Ontology Model (POM), which can be translated into any knowledge representation formalism, and implements data-driven change discovery, therefore it updates the POM according to the corpus changes only, and allows user to trace the evolution of the ontology with respect to the changes in the underlying corpus. the framework analyses Arabic textual resources, and matches them to Arabic Lexico-syntactic patterns in order to learn new Concepts and Relations. They developed a framework for incremental ontology learning, using Arabic natural language processing, machine learning and text mining techniques, in order to extract ontology from Arabic textual resources. The novel aspects about the framework are: (i) the flexibility with respect to use other Arabic linguistic analyzers, and add new Lexico-syntactic patterns to reach more accuracy, (ii) the independence of a concrete ontology representation language, (iii) benefits from the previous knowledge by using an assistant ontology, (iv) using the probability for capturing uncertainty and enhancing user interaction, (v) the integration of data-driven change discovery strategies increasing the efficiency of the system, as well as the traceability of the learned ontology with respect to changes in the corpus, making the whole process more transparent[6].

3) Arabic WordNet Current State and Future Extensions

AWN is a free lexical resource for modern standard Arabic. It is based on the design and contents of Princeton WordNet (PWN)and can be mapped onto PWN as well as a number of other wordnets, enabling translation on the lexical level to and from dozens of other languages. Moreover, the mapping of WordNet to the Suggested Upper Merged Ontology (SUMO) provides opportunities to use the semantic side in some Arabic NLP applications. Constructing AWN presents challenges not encountered by established wordnets. These include the script on the one hand and the morphological properties of Semitic languages, centered around roots, on the other hand. The foundations for meeting these challenges have been laid. An innovation with significant consequences for wordnet development is the proposal to substituteEnglish WN as the ILI with SUMO.

Following EuroWordNet, AWN is developed in two phases by first building a core wordnet around the most important concepts, the so-called Base Concepts, and secondly extending the core wordnet downward to more specific concepts using additional criteria. The core wordnet should thus become highly compatible with wordnets in other languages that aredeveloped according to the same approach. For the core wordnet, The Common Base Concepts(CBCs) of the 12 languages in EWN and BalkaNet are being encoded as synsets in AWN; other Arabic language-specific concepts are added and translated manually to the closest synset. The same procedure is performed for all English synsets that currently have an equivalence relation in the SUMO ontology. Synset encoding proceeds bi-directionally: given an Englissynset, all corresponding Arabic variants (if any) will be selected; given an Arabic word, all its senses are determined and for each of them the corresponding English synset is encoded. The Arabic synsets will be extended with hypernym relations to form a closed semantic hierarchy. SUMO will be used to maximize the semantic consistency of the hyponymy links. This will represent the core wordnet, which is a semantic basic for the further extension. Thework is mostly done manually. When a new Arabic verb is added, extensions are madefrom verbal entries, including verbal derivates, nominalizations, verbal nouns, and so on. We also consider the most productive forms of deriving broken plurals. This is done by applying lexical and morphological rules iteratively.

The database is further extended downward from the CBCs. First, a layer of hyponyms is chosen based on maximal connectivity, relevance, and generality. Two major pre-processing steps are required, preparation and extension. Preparation entails compiling lexical and morphological rules and processing available bilingual resources from which we construct a homogeneous bilingual dictionary containing information on the Arabic/English word pair. This information includes the Arabic root, the POS, the relative frequencies and the sources supporting the pairing. The Arabic words in these bilingual resources must also be normalized andlemmatized while maintaining vowels and diacritics. We next apply 17 heuristic procedures, previously used for EWN, to the bilingual dictionary in order to derive candidate Arabic words/English synsets mappings. Each mapping includes the Arabic word and root, the Englishsynset, the POS, the relative frequencies, a mapping score, the absolute depth in AWN, the number of gaps between the synset and the top of the AWN hierarchy, and attested tokens of the pair. The Arabic word/English synset pairs constitute the input to a manual validation process. We proceed by chunks of related units (sets of related WN synsets, e.g. hyponymy chains and sets of related Arabic words, i.e., words having the same root) instead of individual units (i.e., synsets, senses, words).

Finally, AWN will be completed by filling in the gaps in its structure, covering specific domains, adding terminology and named entities, etc. Each synset construction step is followed by a validation phase, where formal consistency is checked and the coverage is evaluated in terms of frequency of occurrence and domain distribution. The total coverage of AWN will be around10,000 synsets. Although the construction of AWN has been manual, some efforts have been made to automate part of the process using available bilingual lexical resources. Using lexical resources for the semiautomatic building of wordnets for languages other than English is not new. In some cases a substantial part of the work has been performed automatically, using PWN as source ontology and bilingual resources for proposing correlates. An early effort along these lines was carried out during the development of Spanish WordNet within the framework of EuroWordNet project. Later, the Catalan WordNet and Basque WordNet were developed following the same approach.ForAWN, they have investigated two different possible approaches. On the hand, they produce lists of suggested Arabic translations for the different words contained in the English synsets corresponding to the set of Base Concepts. In this case the input to the lexicographical task is the English synset, its set of synonyms and their Arabic translations. On the other hand, they derive new Arabic word forms from already existing, manually built, Arabic verbal synsets using inflectional and derivational rules and produce a list of suggested English synset associations for each form. In this case the input is the Arabic verb, the set of possible derivates and the set of English synsets which would be linked to corresponding Arabic synset. In both cases, the list of suggestions is manually validated by lexicographers.

This methodology takes advantage of one of a central characteristic of Arabic, namely that many words having a common root (i.e. a sequence of typically three consonants) have related meanings and can be derived from a base verbal form by means of a reduced set of lexical rules. Since AWN entries must be manually reviewed, our aim is once again not to automatically attach new synsets but rather to suggest new attachments and to evaluate whether these suggestions can help the lexicographer. As with previous approach, we are more interested in getting a broad coverage than high accuracy, although an appropriate balance between these two measures is nonetheless desirable.[22]

C. Automatic Developing Ontologies

1) Automatic construction of ontology from Arabic texts

The work proposes an approach of automatic construction that is using statistical techniques to extract elements of ontology from Arabic texts. Among these techniques they use two; the first is the “repeated segment” to identify the relevant terms that denote the concepts associated with the domain and the second is the “co-occurrence” to link these new concepts extracted to the ontology by hierarchical or nonhierarchical relations. The processing is done on a corpus of Arabic texts formed and prepared in advance. They use statistical methods, since these methods do not require these types of annotated corpora and NLP1 analyzers (such as the lexical analyzer and parser). These methods are based on two criteria: the relevance of a term from a domain that is defined by the number of occurrences of the word in the corpus and the co-occurrence of two terms at a frequency more high. they started the initialization of the ontology manually, by the general (generic) concepts retrieved from the ontology of GOLD (General Ontology for Linguistic Description).they have formed a domain corpus by the recovery of text from articles of journals and books of the domain and also the collection of documents over the Web. This corpus was preprocessed to remove some ambiguity, reduce the number of transactions and adapt the corpus according to their aim.

After preparing the corpus, they move to the extraction step of ontology elements. The processing is done in two passages. In the first; they extract all the terms (one or more words) used to denote concepts in the domain, using the method of “repeated segments” based on the following prepositions: *A significant term is used several times in a specialized text.*

- Terms can be complex, that are composed of several words used individually (ex. . (جملةاسمية
- Complex terms are constructed using a finite number of sequences of words. In the second passage; we will seek the pairs of terms that co-occur more frequently in the corpus. The result of this processing provides them with a list of pairs of terms that will be used to update the ontology.

Many perspectives are offered based on our work, among them; we proposed an ontology that represents the fundamentals notions of Arabic linguistics, this ontology can be useful for developing NLP tools that analyze Arabic texts. A second perspective would be to use our techniques and statistical methods for information extraction on Arabic texts for other works (e.g. terminology extraction, creation of electronic dictionaries and thesaurus ...) [19].

5 CONCLUSION

The increasing interest in Ontologies for many natural language applications in the recent years has led to the creation of ontologies. These Ontologies are for different purposes and with different features systems. Also the recent work in Artificial Intelligence is exploring the use of formal ontologies. Its use is as a way of specifying content-specific agreements for the sharing and reuse of knowledge among software entities. There are various studies conducted on Arabic language in Semantic Web. The propose of this studies is to improve the Arabic information retrieval on the web. The ontology development life cycle had many questions around it in the last few years. We have discussed some of these criteria’s and methods and give some examples. This paper is part of an ongoing research to develop a frame work for building an Arabic ontology.

REFERENCES

[1] P.Saariluoma , K. Nevala, From Concepts to Design Ontologies, Cognitive Science, University of Jyväskylä, Finland, 2009.

[2] C.e Roche, ONTOLOGY: ASURVEY, University of SavoieEquipeCondillac - Campus Scientifique,73 376 Le Bourget du Lac cedex – France,2002.

[3] Fortuna B., M. Grobelnik, D. Mladenic. "System for Semi-automatic Ontology Construction". Demo at ESWC 2006, June 11-14, ,Budva, Montenegro 2006.

[4]A. Lieto, Manually vs semiautomatic domain specific ontology building,Annoaccademico, spain, 2007-2008.

[5] H. Aliane, Z.Alimazighi, MazariA.Cherif, "Al –Khalil: The Arabic Linguistic Ontology Project, Semantic web and Arabic Language" Team, Research Center on Scientific and technical Information, Algiers. 2010.

[6] N. Ghneim, W. Safi, M. Al Said Ali," Building a Framework for Arabic Ontology Learning", Damascus University, Damascus, Syria,2008.

[7]T. R. Gruber, Toward Principles for the Design of Ontologies, Stanford Knowledge Systems Laboratory,1996.

[8] A. Gangemi,Ontology Design Patterns for Semantic Web Content, Laboratory for Applied Ontology, ISTC-CNR, Rome, Italy,2006.

[9] Zhan Cui, Dean Jones and Paul O'Brien Intelligent Business Systems Research Group Intelligent Systems Lab BTexact Technology Issues in Ontology-based Information Integration , 2002.

[10] L. Al-Safadi, M. Al-Badrani, M. Al-Junidey, "Developing Ontology for Arabic Blogs Retrieval", International Journal of Computer Applications (0975 – 8887)Volume 19– No.4, April 2011.

[11] Annika O¨ hgren,Ontology Development and Evolution: Selected Approaches for Small-Scale Application Contexts, Information Engineering Group Department of Computer and Electrical Engineering School of Engineering, J¨onk¨oping University .J¨onk¨oping, SWEDEN ,ISSN 1404-0018, 2005

[12] W. N. Borst. Construction of Engineering Ontologies for Knowledge Sharing and Reuse.PhD thesis, University of Twente, Enschede, 1997.

[13] C. Brewster, F. Ciravegna, and Y. Wilks. User-Centred Ontology Learning for Knowledge Management. In Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers,2002.

[14] N. Noy and C. Hafner, The State of the Art in Ontology Design, AI Magazine Volume 18 Number 3,1997.

[15] Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C.,
Introducing the Arabic WordNet Project, in *Proceedings of the Third International WordNet Conference*, Sojka, Choi, Fellbaum and Vossen eds. 2006.

[17] Maryam Hazman, Samhaa R. El-Beltagy, Ahmed Rafea A Survey of Ontology Learning Approaches, International Journal of Computer Applications (0975 – 8887)Volume 22– No.9, May 2011

[18] Viviana Mascardi1, Valentina Cordì1, Paolo Rosso2, A Comparison of Upper Ontologies (Technical Report DISI-TR-06-21).

[19] Ahmed CherifMazari, HassinaAliane, and ZaiaAlimazighi,Automatic construction of ontology from Arabic texts, Proceedings ICWIT 2012

[20] Scott Farrar and Terry Langendoen , A Linguistic Ontology for the Semantic Web, 2003

[21]M. Attia, M. Rashwan, A. Ragheb, M. Al-Badrashiny, H. Al-Basoumy, A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields
SpringerLink Home Aug 28 2008.

[22] Maha Al-Yahya*, Hend Al-Khalifa , Alia Bahanshal, Iman Al-Odah and Nawal Al-Helwah AN ONTOLOGICAL MODEL FOR REPRESENTING SEMANTIC LEXICONS: AN APPLICATION ON TIME NOUNS IN THE HOLY QURAN, *The Arabian Journal for Science and Engineering, Volume 35, Number 2C, December 2010*

[23]HassinaAliane, ZaiaAlimazighi, Ahmed CherifMazari, Al - Khalil : The Arabic Linguistic Ontology Project. In proceeding of: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta

[24] Maryam Hazman , Samhaa R. El-Beltagy , Ahmed Rafea Ontology Learning from Textual Web Documents, INFOS2008, March 27-29, 2008 Cairo-Egypt.

[25] Hend S. Al-Khalifa, Areej S. Al-Wabil, The Arabic language and the semantic web: Challenges and opportunities, The 1st International Sysmposium on Computers and Arabic Language & Exhibition 2007

[26] Thomas R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing

[27] Automatic Ontology Construction, *Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University March, 1993.*

[28] MithunBalakrishna, Dan Moldovan, Marta Tatu, Marian Olteanu , Semi-Automatic Domain Ontology Creation from Text Resources, Lymba Corporation Richardson TX 75080 USA

[29] A. Maedche and S. Staab, Semi-Automatic Engineering of Ontologies from Text, Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany, 2000.

[30] SilvanaHartmann ,GyörgySzarvas , IrynaGurevych Mining Multiword Terms from Wikipedia, ISBN 978-1-4666-0189-5, , IGI Global. 2012

[31] MehrnoushShamsfard, Towards Semi-Automatic Construction of a Lexical Ontology for Persian, NLP Research Laboratory, Faculty of Electrical & Computer Engineering, ShahidBeheshti University, Tehran, Iran,2008

[32] Karoui, L., Aufaure, M., and Bennacer, N.. Ontology Discovery from Web Pages: Application to Tourism. In ECML/PKDD: Knowledge Discovery and Ontologies KDO-2004.

[33] Maedche, A. and Staab, S..Ontology Learning for the Semantic Web.In IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2). 2001

[34] Fellbaum, C.. Ed.: “WordNet: An Electronic Lexical Database”, MIT Press. 1999

[35] Soergel, D. Lauser, D., Liang, A., Fisseha, F., Keizer, J., and Katz, S..Reengineering Thesauri for New Applications: the AGROVOC Example. In Journal of Digital Information, 4, 4 (Mar. 2004).

[36] Shamsfard, M. and Barforoush, A. A.. The state of the art in ontology learning: A framework for comparison. The Knowledge Engineering Review, Vol. 18 No.4 pp. 293-316. 2003

[37] Gomez-Perez, A., Manzano-Macho, D..OntoWeb Deliverable 1.5: A Survey of Ontology Learning Methods and Techniques. Universidad Politecnica de Mad+rid. 2003

[38] Sabou, M., Wroe, C., Goble, C., and Mishne, G..Learning Domain Ontologies for Web Service Descriptions: an Experiment in Bioinformatics. In Proceedings of the 14th International World Wide Web Conference (WWW2005), Chiba, Japan. 2005

[39] Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R.W., and Musen, M.A. 2001. Creating Semantic Web Contents with Protege-.In IEEE Intelligent Systems, Vol. 16, No. 2, pp. 60-71. 2000

[40] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., and Wenke, D. 2002. OntoEdit: Collaborative ontology development for the semantic web. In International Semantic Web Conference (ISWC 2002), Sardinia, 2002

[41] Sanchez, D., and Moreno, A. 2004. Creating ontologies from Web documents.In Recent Advances in Artificial Intelligence Research and Development. IOS Press, Vol. 113, pp.11-18.2004

Biography

Dalia Fadl is PhD student. She works as assistant lecturer since 2007. She has M.Sc. degree in Computer science on September 2012.

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

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

ملخص:

أنطولوجيا عربية باستخدام تقنيات مختلفة لتعلم الأنطولوجيا

داليا فاضل[1] – صفية عباس[2] – مصطفى عارف[3]

قسم علوم الحاسب – كلية الحاسبات والمعلومات – جامعة عين شمس

[1]Dalia_sayed_43@hotmail.com
[2]safia_abbas@yahoo.com
[3]mostafa.m.aref@gmail.com

تحتوي اللغة العربية على مجموعة من الخواص جعلتها من اللغات الصعبة وعرقل هذا تطوير أدوات وبرمجيات للغة العربية. ومن بين تلك الخواص الصرفية المعقد، النحوية والجوانب الدلالية. اللغة العربية هي اللغة الرسمية لمئات الملايين من الناس في منطقة الشرق الأوسط ودول شمال أفريقيا. وهي اللغة الدينية لجميع المسلمين من أعراق مختلفة في جميع أنحاء العالم. على الرغم من ذلك تم القيام بقليل في مجال اللغة المحوسبة والموارد المعجمية. اللغة العربية هي لغة سامية والتي تختلف من اللغات الأوروبية نحويا ولغويا وشكليا. يشير مصطلح 'العربية الفصحى' إلى النموذج القياسي من اللغة المستخدمة في الكتابة وتسمع في التلفزيون والإذاعة وفي الخطابات العامة والخطب الدينية. الهدف من هذا البحث هو مناقشة معايير لتصميم علم الموجودات للغة العربية. واتخاذ وجهة نظر الهندسة على تطوير ذلك العلم. يعرض البحث أنواع مختلفة من دورة حياة تطوير علم الموجودات.

Evaluation of Semantic Networks

P. Elkafrawy[*1], M. Rafea[**2], M. Nasef[*3], R. Elnemr [**4]

[*]*Mathematics and CS Department, Faculty of Science and Menofia University*
Shebin ELKom, Egypt

[1]basant.elkafrawi@science.menofia.edu.eg

[3]mnasef81@yahoo.com

[**]*Central Lab. for Agricultural Expert System*
Agricultural Research Center, Egypt

[2]mahmoudrafea@gmail.com

[4]rashaelnemr82@yahoo.com

***Abstract*— a semantic network is a network, which represents semantic relations among concepts. This is often used as a form of knowledge representation. It is a directed graph consisting of nodes, which represent concepts, and edges which represent relations between nodes. A semantic network is used when one has knowledge that is best understood as a set of concepts that are related to one another. More generally, most semantic networks are cognitively based. In this paper, we evaluate semantic networks and present its different types and variations.**

1 INTRODUCTION

A semantic network (SN) is widely used knowledge representation technique. Semantic network is a graphical knowledge representation scheme consisting of nodes, and links between nodes. Computer implementations of semantic networks were first developed for artificial intelligence and machine translation, but earlier versions have long been used in philosophy, psychology, and linguistics [1]. The nodes of the net represent objects or concepts and the links represent relations between nodes. The links are unidirectional and labelled; therefore, a semantic network corresponds to a directed graph. From the graphical point of view, the nodes are usually represented by circles or boxes and the links are drawn as arrows or simple connectors between the circles. The structure of the network defines its meaning, depending on which nodes are connected to which other nodes. KR techniques are divided in to two main categories one is declarative and other is procedural. The declarative representation techniques are used to represent objects, facts, relations. Whereas the procedural representation are used to represent the action performed by the objects. Semantic net is a declarative KR technique that can be used either to represent knowledge or to support automated systems for reasoning about knowledge [2].

Semantic network have different application such as practical knowledge representation for the Web [3], [4]. Semantic modelling and knowledge representation in Multimedia Database is another application of SN [5]. SN is used to model trouble shooting's knowledge. In Pattern-recognition semantic net can be used to help the computer to identify how objects to be analyzed are related to one another. It is used heavily in Natural language processing. Bootstrapping knowledge representation uses semantic nets to make the web more intelligent. Finally, SN is an excellent reasoning mechanism [6].

The paper is organized after this introduction. Section 2 reviews the basics of SN. In section 3 we explain how SN inferencing is performed and inheritance criteria. Types of SN are listed in section 4. Then an evaluation of SN is presented in section 5 to study the pros and cons of SN. In section 6 partitioned semantic nets are defined. Finally, the conclusion of the paper is given in section 7 with future work.

2 SEMANTIC NETWORK REVIEW

In order to have a concrete example of what a semantic network is, let us look at figure 1 in [7], as can be seen, the node which labeled "person" is linked to the node which labeled “living being”. The link is labeled “is-a”. Indeed, technically speaking, the diagram represents the fact that there is a binary relation between a living being, such as a person, and the concept of person itself. Another node with the label “cat”, as well as a “is-a” link from this node to the “living being” node, again representing that a cat is a type of living being.

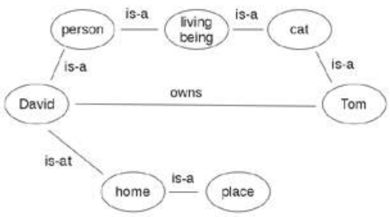


Figure 1: Example of Semantic Network

Also, there is a person called “David” and a cat called “Tom”, and David owns Tom, the structure of the network becomes apparent as shown in figure 1. Clearly, a new link labeled "owns" would need to be added as well, in order to represent that David owns Tom. Indeed the nodes labeled “living being”, “person” and “cat” represent the generic or class concept of a living being, a person and a cat, respectively; in practice, they represent just abstract concepts. Instead, the nodes “David” and “Tom” represent an individual instance of the nodes “person” and “cat”, respectively; in fact David is a person and Tom is a cat. In conclusion it is crucial to notice that there are two types of context, classes and individuals, although they are represented in the same way. Now let us add more information as shown in figure 2. The information now being represented is that David is a person and home is the place he is at.

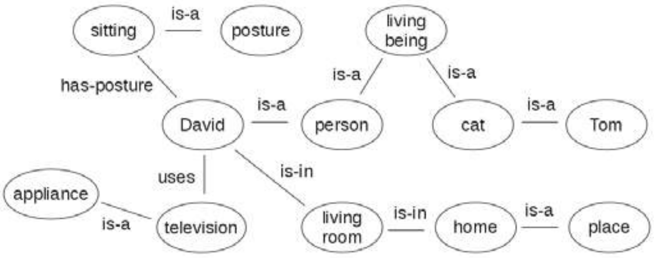


Figure 2: Example of Semantic Network

As the number of nodes increases, the meaning of the respective links needs to be considered. It should be apparent that not all links are alike. Indeed, some links express only relationships between nodes, and are therefore assertions of the nature of the relationship between two different nodes. For example, the link “is-at” in figure 2, which describes the relationship that the person David is at the place home. The “is-a” links in figure 2, instead, are structural links, in that they provide “type” information about the node. It is clear since this information is about the node itself and not about the relationship it has to be a different type of node. For instance, the node “home” is an individual instance of the class node labeled “place”. The network in figure 2 now provides a representation for information about the nodes belonging to it. For instance, a person called David is the owner of a cat called Tom, and at the moment he is sitting in the living room, using a television.

Another important characteristic of the node-link representation is the implicit “inverse” of all relationships represented by a link. Indeed, if there is a link going from one node to another, this also implies the reverse, and it means that there is a link from the second node to the first. For example, there are two nodes labeled “David” and “television” with the link labeled “uses”. The direction of the relationship is that “David uses a television”. In practice “David” is a subject and “television” is the object, and “uses” is the verb or action or link between them. This “David uses television” relation implies the inverse relationship that “television is-used- by David”, as shown in figure 3.

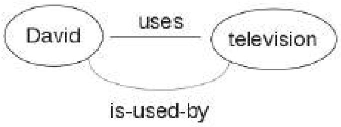


Figure 3: Symmetric relationships in Semantic Networks

Non-binary relationships can be represented by “turning the relationship into an object”. This process in knowledge representation system is known as “reification”. As shown in figure 4, we can represent the generic give event as a relation involving three things: a giver, a recipient and an object, give (john, mary, book)[8], [9].

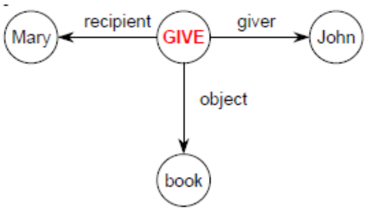


Figure 4: Non-binary relations

Semantic networks are very good at representing events, and simple declarative sentences, by basing them round an “event node” [10]. For example: “John gave lecture w6 to his students” as shown in figure 5. In fact, several of the earliest semantic networks were English-understanding programs.

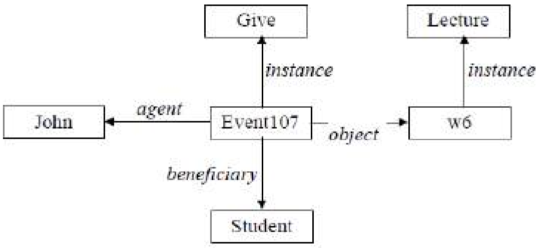


Figure 5: Example of representing events

3 Knowledge Inferencing in Semantic Networks

With any kind of knowledge representation scheme, it is possible to infer knowledge that is not directly represented by the scheme. To give an example of what can be found out from the semantic network in figure 2 that is not directly represented, let us consider figure 6. By tracing the path from the node “living room” to the node “David” via the link labeled “is-in” and then from the node “David” to the node “television” via the link labeled “uses”, it is possible to infer that the television is in the living room by inferring a link labeled “is-in” between the node “television” and the node “living room”, as shown in figure 6. This means that this information does not need to be explicitly represented in the original network, for it can be easily inferred later.

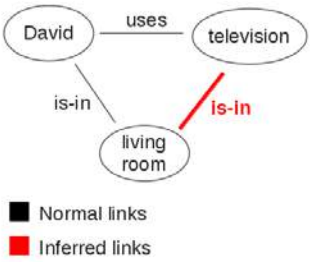


Figure 6: Example of knowledge inferring in Semantic Networks

From a mathematical point of view, composing links occurs by placing them end-to-tail. This composition creates a new link. It is not possible to compose every pair of links, only those whose destinations and sources correspond. The destination of the first must be the source of the second. By composing links, new relationships between nodes can be found and described. Such a process is also called chasing links and the terminology introduced comes from a branch of mathematics called Category Theory. Looking at figure 7 in [3] and formalizing the whole lot from a logical point of view, we can say that if x is an individual and y is class, the link “is-a” between them can be interpreted as the following formula: y(x).

Figure 7: Simple example of instancing in Semantic Networks

E.g.: cat (Tom).

Instead, if x and y are classes, the link between them can be interpreted as the following formula:

$$\forall Z\ x(Z)\ \Rightarrow\ y(Z)$$

Category theory is an area of study in mathematics that examines in an abstract way the properties of particular mathematical concepts, by formalizing them as collections of objects and arrows, where these collections satisfy some basic conditions.

E.g.: $\forall Z\ cat(Z) \Rightarrow living\_being(Z)$.

Finally, if a class or an individual has some properties, these can be translated to binary predicates:

$\forall Z\ y(Z) \Rightarrow property(Z, value)$ *class*

$property(x, value)$ *individual*

In conclusion, coming back to our original example, figure 8 shows the results of more link chasing. As you can see, additional relationships are derived, e.g., a person has a posture, may own a cat and may use appliances.

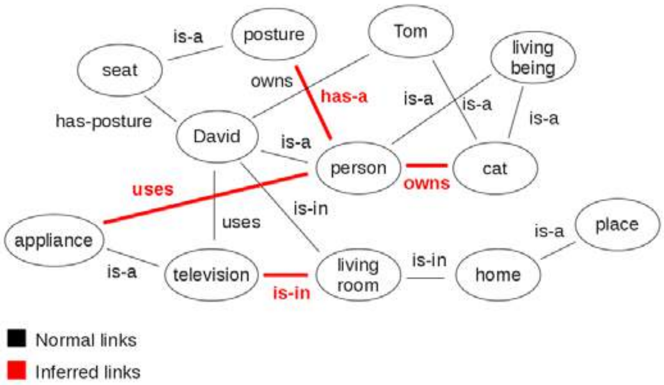


Figure 8: A more complicated example of inference in Semantic Networks

A. Inheritance in Semantic Net

Semantic network are generally used to represent the inheritable knowledge. Inheritance is most useful form of inference. Inheritance is the belongings in which element of some class inherit the attribute and values from some other class. To support inheritance object must be organized into classes and classes must be arranged in a generalization hierarchy. Because there is an association between two or more nodes the Semantic nets are also known as associative nets. These associations are proved to be useful for inferring some knowledge from the existing one. If user wants to get any knowledge from the knowledge base they need not to put any query. The activated association or relation provides the result directly or indirectly only need to follow the links in the semantic net. IS-A, and A-KIND-OF are generally used to represent the value of a link in semantic net as shown in figure 9. The searching algorithms of the semantic net are Intersection Search, Inheritance, Breadth First, Depth First, and Heuristic Search [6], [9].

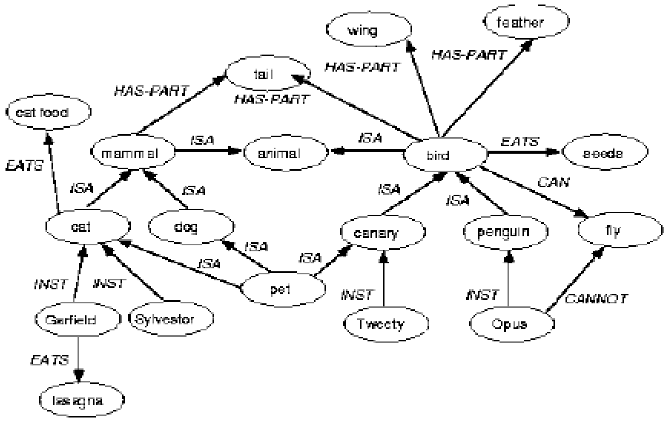


Figure 9: Represents of IS-A, HAS, INSTANCE

Two important features of semantic networks are the ideas of default (or typical) values and inheritance. We can assign expected/default values of parameters and inherit them from higher up the hierarchy. This is more efficient than listing all the details at each level.

B. Multiple Inheritance

A node can have any number of super-classes that contain it, enabling a node to inherit properties from multiple parent nodes and their ancestors in the network. Sometimes it may cause conflicting inheritance. With simple trees, inheritance is straight-forward. However, when multiple inheritance is allowed, problems can occur. For example, consider this famous example: Question: "Is Nixon a pacifist?" [8].

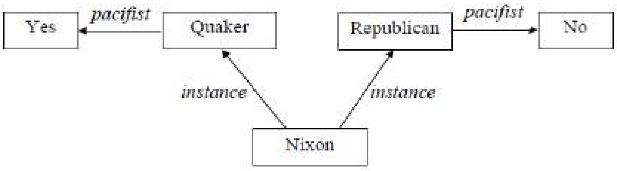


Figure 10: Multiple Inheritance

Conflicts like this are common is the real world. It is important that the inheritance algorithm reports the conflict, rather than just traversing the tree and reporting the first answer it finds. In practice, we aim to build semantic networks in which all such conflicts are either over-ridden or resolved appropriately [8], [10].

4 SEMANTIC NET TYPES

According to [1], [11] the known semantic networks can be divided into six kinds depending on the used techniques:

A. Definition network

Definition network emphasizes the subtype or is-a relation between a concept type and a newly defined subtype. The resulting network, also called a generalization hierarchy, supports the rule of inheritance to copy properties defined for a supertype to all of its subtypes. Since definitions are true by definition, the information in these networks is often assumed to be necessarily true. Such systems can be useful for many applications, but they can also create problems of conflicting defaults as shown in figure 11.

The Nixon diamond on the left shows a conflict caused by inheritance from two different supertypes: by default, Quakers are pacifists, and Republicans are not pacifists. Does Nixon inherit pacifism along the Quaker path, or is it blocked by the negation on the Republican path? On the right is another diamond in which the subtype Royal Elephant cancels the property of being gray, which is the default color for ordinary elephants. If Clyde is first mentioned as an elephant, his default color would be gray, but later information that he is a Royal Elephant should caused the previous information to be retracted.

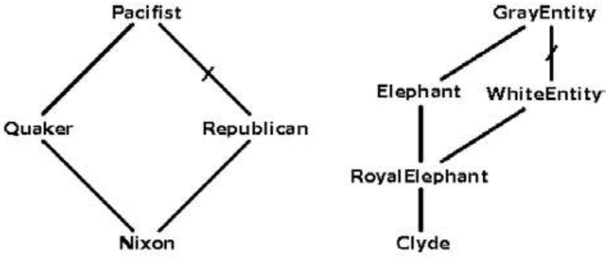


Figure 11: Conflicting defaults in a definitional network

To resolve such conflicts, many developers have rejected local defaults in favor of more systematic methods of belief revision that can guarantee global consistency.

B. Assertion network

Assertion networks are designed to assert propositions. Unlike definition networks, the information in networks of this kind is assumed to be contingently true, unless it is explicitly marked with a modal operator. Some assertion networks have been proposed as models of the conceptual structures underlying natural language semantics. The most successful approach was the method of adding explicit nodes to show propositions. Logical operators would connect the propositional nodes, and relations would either be attached to the propositional nodes or be nested inside them.

C. Implication networks

Implication networks are a special case of a propositional semantic network in which the primary relation is implication. Other relations may be nested inside the propositional nodes, but they are ignored by the inference procedures. They may be used to represent patterns of beliefs, causality, or inferences. Depending on the interpretation, such networks may be called belief networks, causal networks, Bayesian networks, or truth-maintenance systems.

D. Executable networks

Executable networks include some mechanisms which can perform inferences, pass messages, or search for patterns and associations. Executable semantic networks contain mechanisms that can cause some change to the network itself. The executable mechanisms distinguish them from networks, which are static data structures and can only change through the action of programs external to the net itself.

Three kinds of mechanisms are commonly used with executable semantic networks:
1) *Message passing networks* can pass data from one node to another. For some networks, the data may consist of a single bit, called a marker, token, or trigger; for others, it may be a numeric weight or an arbitrarily large message.
2) *Attached procedures* are programs contained in or associated with a node that perform some kind of action or computation on data at that node or some nearby node.
3) *Graph transformations* combine graphs, modify them, or break them into smaller graphs. In typical theorem provers, such transformations are carried out by a program external to the graphs. When they are triggered by the graphs themselves, they behave like chemical reactions that combine molecules or break them apart.
These three mechanisms can be combined in various ways. Messages passed from node to node may be processed by procedures attached to those nodes, and graph transformations may also be triggered by messages that appear at some of the nodes.

The simplest networks with attached procedures are dataflow graphs, which contain passive nodes that hold data and active nodes that take data from input nodes and send results to output nodes. Petri nets are considered as the most widely-used formalism that combines marker passing with procedures.

E. Learning networks

Learning networks build or extend their representations by acquiring knowledge from examples. The new knowledge may change the old network by adding and deleting nodes and arcs or by modifying numerical values, called weights, associated with the nodes and arcs. A learning system, natural or artificial, responds to new information by modifying its internal representations in a way that enables the system to respond more effectively to its environment. A learning system, natural or artificial, responds to new information by modifying its internal representations in a way that enables the system to respond more effectively to its environment. Systems that use network representations can modify the networks in three ways:

1) *Rote memory*: The simplest form of learning is to convert the new information to a network and add it without any further changes to the current network.
2) *Changing weights*: Some networks have numbers, called weights, associated with the nodes and arcs. In an implicational network, for example, those weights might represent probabilities, and each occurrence of the same type of network would increase the estimated probability of its recurrence.
3) *Restructuring*: The most complex form of learning makes fundamental changes to the structure of the network itself. Since the number and kinds of structural changes are unlimited, the study and classification of restructuring methods is the most difficult, but potentially the most rewarding if good methods can be found.

Systems that learn by rote or by changing weights can be used by themselves, but systems that learn by restructuring the network typically use one or both of the other methods as aids to restructuring. Neural nets are a widely-used technique for learning by changing the weights assigned to the nodes or arcs of a network.

F. Hybrid networks

Hybrid networks combine two or more of the previous techniques, either in a single network or in separate but closely interacting networks. Conceptual graphs, for example, include a definitional component for defining types and an assertion component that uses the types in graphs that assert propositions. The most widely used hybrid of multiple network notations is the Unified Modeling Language (UML). Although UML is not usually called a semantic network, its notations can be classified according to the categories of semantic networks discussed in this research.

5 Semantic Network Evaluation

As we saw so far, Semantic networks are characterized by a high representational and expressive power, which is why they constitute a powerful and adaptable method of representing knowledge. In particular, semantic networks present some advantages that can be summarized as follows. Many different types of entities can be represented in Semantic Networks. SN is easy to visualize. Semantic Networks provide a graphical view of the problem space. So they allow an easy way to explore the problem space and therefore they are relatively easy to understand. They can be used as a common communication tool between different fields of knowledge, e.g., between computer science and anthropology. Efficient in space requirements; objects represented only once, relationships handled by pointers. Semantic Networks provide a way to create clusters of related elements. They resonate with the ways in which people process information.

They are a more natural representation than logic (using meaning axioms). They are characterized by a higher cognitive adequacy than logic-based formalisms. Semantic Networks allow the use of efficient inference algorithms (graph algorithms). They have a higher expressiveness than logic (e.g., they allow properties overriding). They allow us to structure the knowledge to reflect the structure of that part of the world, which is being represented. There are very powerful representational possibilities as a result of “is a” and “is a part of” inheritance hierarchies. The semantics, i.e. real world meanings, are clearly identifiable. They can accommodate a hierarchy of default values They can be used to represent events and natural language sentences. Knowledge engineers can easily define the relationship. Scalable and modular structure i.e. easy to build and maintain. Formal definitions of semantic networks have been developed. Simplicity, naturalness, and clarity

Semantic Network also has some limitations, which frequently lead to some epistemological problems. First, a distinction between classes and individuals does not exist. The system is limited by the user’s understanding of the meanings of the links in a semantic network. As pointed out previously, links between nodes are not all alike in function or form.
Indeed, we need to differentiate between links that constitute some relationship and links that are structural in nature. As shown in figure1, the link “is-a” behaves in two different ways: between the nodes “Tom” and “cat” it specifies an instance of a cat; instead, between the nodes “cat” and “living being” it specifies a category, a hierarchy [7].

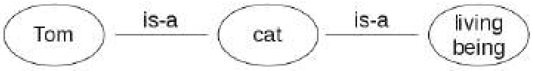


Figure 12: Example of a link used with different meanings

On the subtleties of the "is-a" link revealed even more distinctions in the uses of this link. A possible work-around to this problem could be to specify in a more detailed way the name of the links, distinguishing between relational and structural ones, as shown in figure 13. In this case we re-wrote the link between the nodes “Tom” and “cat” as an “instance-of” link; and the link between the nodes “cat” and “living being” as a “subtype-of” link.

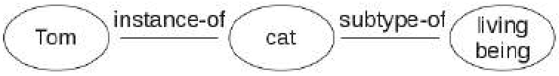


Figure 13: Removing ambiguity from a link

Second, SN lacks a link name standard [6]. Third, there are no standards about node and arc values [4]. Forth, there is no internal structure of nodes [8]. Fifth, binary relation is easy to represent, however, sometimes it is difficult. For example: the sentence “John causes trouble to the party” [6].

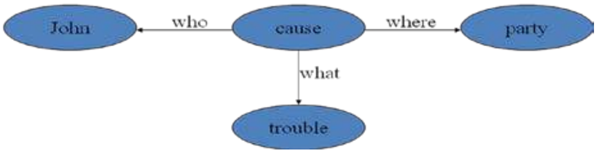


Figure 14: Binary relation in semantic network

Sixth, quantified statements are very hard to represent by Semantic net [6] for example: "Every dog has bitten a postman" and "Every dog has bitten every postman". The solution to this problem could by using Partitioned semantic networks to represent quantified statements. Moreover, negation "John does not go fishing" and disjunction "John eats pizza or fish and chips" produce discrepancies. Seventh, the context of a word is not clear, as if a node is labeled "Table," for example [6], does it represent? A specific Table, The class of all Table, or The concept of a Table.
A distinction between attributes associated to a class and attributes inherited by the individuals of the class does not exist [7]. A formal semantic does not exist, so there is not an agreed-upon notion of what a given representational structure means. Indeed, semantic networks do tend to rely upon the procedures that manipulate them. A solution to this problem could be using conceptual graphs, formalism for knowledge representation, or a knowledge representation system such as KL-ONE, which allows overcoming semantic indistinctness in semantic network representation [7].
Inheritance particularly from multiple sources and when exceptions in inheritance are wanted can cause problems [8]. Difficult implementation of some operations, and difficult control of the inheritance, to solve this problem, Nisenbowm [12] proposes an algebraic method.

No easy way to represent heuristic information [8]. Search may lead to combinatorial explosion especially for queries with negative results [2]. There is much formalism under the name semantic networks with different expressive capabilities but always with a formal reasoning model. So, a formal semantic model for reasoning is necessary. A more structured formalism is necessary. Poor representation of arbitrary relations exists; Insufficient expressiveness; and unclear semantics.

6 PARTITIONED SEMANTIC NET

The semantic net can be divided into two more networks. The semantic net is to be partitioned to separate the various nodes and arcs in to units and each unit is known as spaces. Using partitioned semantic net user can define the existence of the entity. One space is assigned to every node and arc and all nodes and arcs lying in the same space are distinguishable from those of other spaces. Nodes and arcs of different spaces may be linked, but the linkage must pass through the boundaries, which separate one space from another. The central idea of partitioning is to allow groups, nodes and arcs to be bundled together into units called spaces. Every node and every arc of a network belongs to one or more spaces. Universal and existential quantifier can be represent by the Partitioning semantic net Partitioning semantic nets can be used to delimit the scopes of quantified variables [2], [6]. While working with quantified statements, it will be help full to represent the pieces of information consist some event. Suppose that we wish to make a specific statement about a dog, Danny, who has bitten a postman, Peter: "Danny the dog bit Peter the postman". Hendrix’s Partitioned network would express this statement as an ordinary semantic network:

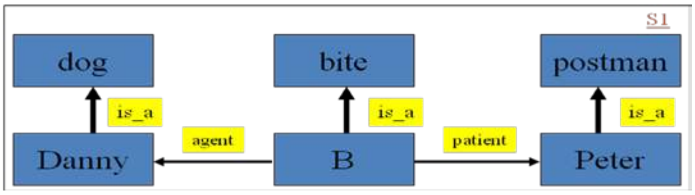


Figure 15: Partitioned Semantic Net

Suppose that we now want to look at the statement: "Every dog has bitten a postman". Hendrix partitioned semantic network now comprises two partitions SA and S1. Node G is an instance of the special class of general statements about the world comprising link statement, form, and one universal quantifier ∀

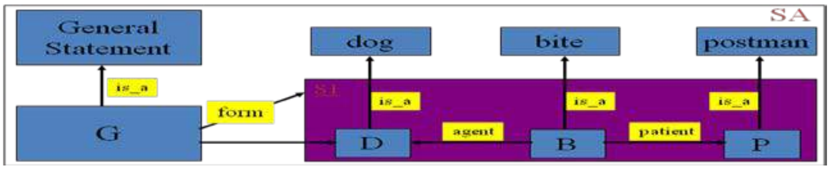


Figure 16: Represents Partitioned Semantic Net for Quantifiers

Suppose that we now want to look at the statement: "Every dog has bitten every postman".

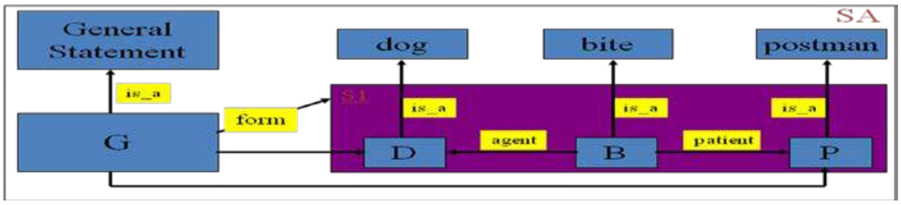


Figure 17: Represents Partitioned Semantic Net for Quantifiers

Suppose that we now want to look at the statement: "Every dog in town has bitten the postman".

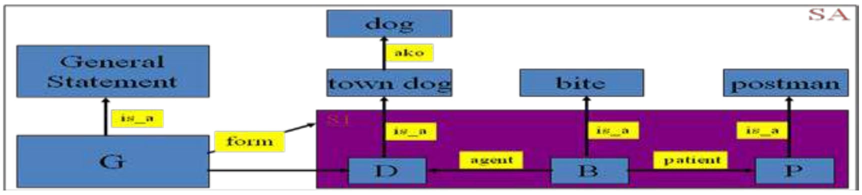


Figure 18: Represents Partitioned Semantic Net for Quantifiers

7 CONCLUSION

There are various knowledge representation techniques in AI. Semantic net is commonly used KR technique that represents the connection between objects or class of objects. It is a directed graph in which nodes / vertices represent the objects/ class of objects and edges and links (unidirectional) represent the semantic relations between the objects. Semantic net are used to represent the inheritable knowledge. Inheritance is most useful form of inference. Semantic nets have some advantage such as simplicity, naturalness. However, they have some disadvantages as poor representation of arbitrary relations, difficult implementation of some operations, and difficult control of inheritance. All KR techniques have their own semantics, structure as well as different control mechanism and power. Combination of two or more representation technique may be used for making the system more efficient and improving the knowledge representation. So, in the future we are trying to build the intelligent system that can learn itself by the query and have a power full mechanism for representation and inference. The aim is to take the advantage of the semantic net and another knowledge representation technique under one umbrella.

References

[1] J.F.Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks-Cole, Pacific Grove, CA, 2000.

[2] P.Tanwar, T.V. Prasad and Kamlesh Datta, "Hybrid technique for effective knowledge representation and a comparative study", *International Journal on Computer Science and Engineering (IJCSE)*, vol.3, no.4, pp.43-57, 2012.

[3] J.H.M.Tah, H.F.Abanda, "Sustainable building technology knowledge representation: using semantic web techniques", *Advanced Engineering Informatics*, 25, pp. 547-558, Journal home page: www.elsevier.com/locate/aei, 2011.

[4] Hendler, Jim, Van Harmelen, Frank, "Foundations of Artificial Intelligence", *Elsevier*, vol.3, pp: 821-839, ISBN: 1574-6526, first edition, http://www.sciencedirect.com, 2008.

[5] W.Khatib, "Semantic modelling and knowledge representation in Multimedia", available at ieeeexlore.ieee.org, 1999.

[6] P. Tanwar, Dr. T.V Prasad, Dr.Mahendra, S. Aswal, "Comparative study of three declarative knowledge representation techniques", *International Journal on Computer Science and Engineering (IJCSE)*, vol. 02, no. 07, pp. 2274-2281, 2010.

[7] D.Sorrentino, "Integrating semantic networks and object oriented model to represent and manage context", *Master thesis of science in computer science*, 2012.

[8] K.R.Chowdhary, "Artificial intelligence (semantic network)", Web Site: http://www.krchowdhary.com, 2011.

[9] M.Huntbach, "AI: Notes on semantic net and frames", *School of Electronic Engineering and Computer Science, Queen Marry University of London*,Web Site: http://www.eecs.qmul.ac, 1996.

[10] J.A.Bullinaria, "IAI: Semantic Networks and Frames", *School of Computer Science, University of Birmingham*, Web Site: http:// www.cs.bham.ac.uk, 2005.

[11] M.Marinov, I.Zheliazkova, "An Interactive Tool based on priority semantic networks" *Knowledge-Based Systems* 18, pp.71-77, 2005.

[12] L.Nisenbowm, "An algebraic method of inference in semantic nets", *Computer and Artificial Intelligence* 9 (3), pp. 257-271, 1990.

BIOGRAPHY

Passent ElKafrawy, Associate Professor, Faculty Science, Menofia University

Dr. Passent M ElKafrawy is an Associate Professor since 2013, she got her PhD from the University of Connecticut in United states on 2006 in Computer Science and Engineering; in the field of computational geometry as a branch of Artificial Intelligence. Then she taught in Eastern State University of Connecticut for one year. In 2007 she worked as a Teacher in Faculty of Science, Menoufia University, Mathematics and Computer Science department since that time till now.

Mohammed M. Nasef, Lecturer, Faculty of Science, Menofia University.

Dr. Mohammed M. Nasef is a Lecturer since 2011, he got his PhD from Faculty of Science, Menofia University, Egypt on 2011 in computer science. Supervising five research studies between PhD and MSc. Member of the faculty projects for education development as DSAP, CIQAP. He is the manager of It-Unit, Faculty of science, Menofia University since 2013 until now.

Rasha Elnemr, works at Central Laboratory for Agricultural Expert System, Agricultural Research Centre, was born in Cairo, Egypt, in 1982. She received her Bachelor degree in Computer

Science from Helwan University in 2003 and M. Sc. in Computer Science from the Helwan University in 2009. She is currently pursuing her PH.D degree in the Department of Mathematics and Computer Science, Faculty of Science, Menofia University.

تقييم الشبكة الدلالية

, رشا خليل السيد النمر*, محمد مصطفى ناصف, محمود عبدالواحد رافع*بسنت محمد الكفراوى

*قسم الرياضيات, كلية العلوم– جامعة المنوفية

**المعمل المركزى للنظم الزراعية الخبيرة

مركز البحوث الزراعية

. وكثيرا ما يستخدم هذا كشكل من أشكال تمثيل المعرفة إن الشبكة الدلالية هي شبكة تقوم بتمثيل العلاقات الدلالية بين المفاهيم وتستخدم الشبكة .والشبكة الدلالية هي مخطط موجه يتكون من العقد والتي تمثل المفاهيم والحواف والتي تمثل العلاقات بين هذه العقد وفى العموم معظم .الدلالية عند وجود معرفة ويكون افضل طريقة لفهمها هو تمثيلها كمجموعة من المفاهيم المرتبطة بعضها بعض الشبكات الدلالية تعتمد على الادراك المعرفى. وفى هذة الورقة البحثية نحن نقيم الشبكة الدلالية ونعرض انواعها المختلفة.

Comparative Study of Case Based Reasoning Software and Natural Language

Passent ElKafrawy[*1], Rania Mohamed[**2]

[*]*Mathematics and CS Department, Faculty of ScienceandMenofiaUniversity*
ShebinElkomMenofia, Egypt

[1]`basant.elkafrawi@science.menofia.edu.eg`

[**]*Faculty Computer Science, Modern University for Technology & Information*
Cairo, Egypt

[2]`rania.a.mohamed@gmail.com`

Abstract—**Case-Based Reasoning (CBR) is a problem-solving paradigm that solves a new problem by remembering a previous similar situation and by reusing the information and knowledge of that situation. A Case-Based Reasoning (CBR) tool is software that can be used to develop several applications that require case-based reasoning methodology. However, large volumes of information can make it a complex task to gain useful insight from historic datasets. This paper gives background information about CBR software and introduces the most used CBR tools (CBR Shell, FreeCBR, jCOLIBRI, myCBR and eXiTCBR). Then it introduces a comparative analysis study based on some determined factors that affect the CBR software including noisy data or missing values in the cases. Finally, an evaluation of the retrieving phase of each software is introduced. The process of Natural Language Generation for a Conversational Agent translates some semantic language to its surface form expressed in natural language Case Based Reasoning technique show which is easily extensible and adaptable to multiple domains and languages, that generates coherent phrases and produces a natural outcome in the context of a Conversational Agent that maintains a dialogue with the user.**

Keywords: - Artificial Intelligence, Case-Based Reasoning, Natural Language.

1 INTRODUCTION

The AI Engine is the core of the Personality Forge. It uses both Natural Language Processing (NLP) and Case-Based Reasoning (CBR) which are two philosophies of artificial intelligence which had previously not been mixed (to my knowledge). In Natural Language Processing, sentences are parsed and broken down to reveal the structure of the sentence and information about individual words and their relation to other words in the sentence. In Case-Based Reasoning, sentences are searched for keyphrases which trigger pre-programmed responses. The AI Engine does both- it first breaks down the sentences using NLP into their most basic elements, finds relationships between those elements, finds the meaning of individual words, and then passes all this information forward to the keyphrase, or CBR section. Responses are matched against both specific and broad categories of statements, and then the response is constructed using both the bot's own original words and a wealth of information available from the other chatter's message and memories of the other chatter. The Personality Forge's own scripting language, AIScript, takes this flexibility even further by providing the ability to create if-statements and responses based on memories, emotion, sex, time, and date.

Case-based Reasoning is an emerging field in Artificial intelligence. It is mostly used in problem solving in the artificial intelligence applications. Case-based reasoning is an approach which utilizes the experience gained from solving past problems [1]. This approach maintains all information of past-solved problems, where this experience is stored as a case. The collection of all these past cases is stored in the form of case-base. There are various factors which define the efficiency of this approach [2]. The major factor is the number of past experiences stored in a case base. The new problem should be identified in term of the experience of the problems faced before. The new upcoming problem is considered as a new case. The strategy of finding a similar case for the new problem under investigation stored in the case base is another major factor of defining the efficiency of the case-based reasoning approach. The evaluation of the selected case and indexing of the suggested case for future use are other factors that define the performance of case-based reasoning system.

Case-based reasoning has many advantages over other reasoning approaches such as rule based reasoning [5]. This reasoning approach bears a resemblance to human reasoning. It provides the facility of taking the decision such as human beings take decision in real time. Case-based reasoning is a machine learning mechanism as the solutions of precedent problems faced are stored in the case base. This approach learns from both success & failure of solutions of the previous problems. These past experiences are being reused for solving the coming problems. The process of knowledge acquisition is easily handled in this approach. But in the case of other reasoning approaches, the knowledge acquisition process is not so simple & alsocostly. The other major advantages of this approach over the other reasoning approaches are the worth of the solution. In the revise phase of case-based reasoning approach, the proposed solution is revised according constraints of the problem. Then the proposed solution is repaired according to constraints. It is also modified for fulfilling the constraints of the problem. This phase of case-based reasoning boosts the excellence of the solutions & extends the effectiveness of this approach. The errors of the previous solutions do not propagate in the future of problem's solutions [5]. It can also be applied in those domains where the information about problems is incomplete & insufficient for finding the adequate rules or algorithms to solve them.

After this introduction, a theoretical background is illustrated in section 2. Section 3 introduces a brief description of each software used in this paper. The section 5 introduces a comparative study after testing and comparing the CBR applications. The conclusion andfuture workareintroduced in section 7..

2 THEORETICALBACKGROUND

Case-Based Reasoning (CBR) is a problem-solving paradigm that solves a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation [1]. More specifically, CBR uses a database of problems to resolve new problems. The database can be built through the knowledge Engineering (KE) process or it can be collected from previous cases.

In a problem-solving system, each case would describe a problem and a solution to that problem. The reasoning engine solves new problems by adapting relevant cases from the library [3]. Moreover, CBR can learn from previous experiences. When a problem is solved, the case-based reasoning can add the problem description and the solution to the case library. The new case that in general represented as a pair <problem, solution> becomes immediately available and can be considered as a new piece of knowledge.

According to Doyle et al. [4], Case-Based Reasoning is different from other Artificial Intelligence approaches in the following ways:Traditional AI approaches rely on general knowledge of a problem domain and tend to solve problems on a first-principle while CBR systems solve new problems by utilizing specific knowledge of past experiences.

CBR supports incremental, sustained learning. After CBR solves a problem, it will make the problem available for future problems. The CBR Cycle can be represented by a schematic cycle, as shown in Figure 1. First phase is the retrieve phase, which identifies features via noticing the feature values of a case, initially match a list of possible candidates and select the best match from the cases[7].

Second phase is the reuse phase, where the difference between the new and the old case is determined by copying the old case and adapting by transforming or reusing the old solution. The third phase is the revise phase, if the solution from the last phase is incorrect, then this solution must be evaluated in a real environment setting and the errors/flaws of the solution must be found if the solution was evaluated badly.Finally, the Retain phase which incorporates the lesson learned from the problem-solving experience into the existing knowledge by extracting or indexing. By extracting we mean if the problem was solved using an old case, the system can build a new case or generalize an old case. By indexing we mean via deciding what types of indexes can be used in the future by integrating and modifying the indexing of existing cases afterthe experience.

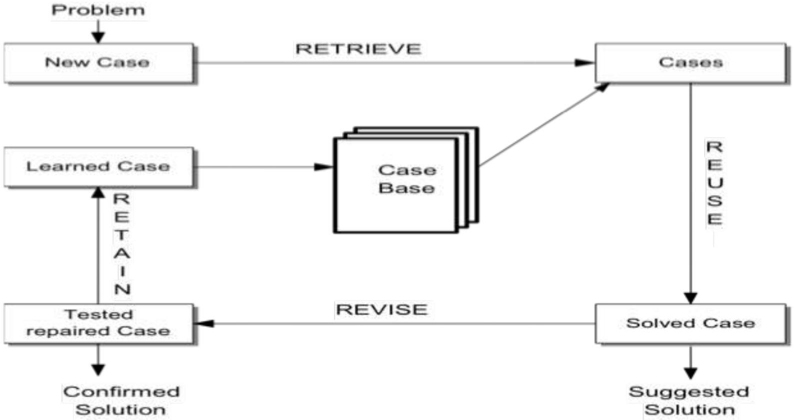


Figure 1: Case-based reasoning

There are three main types of CBR that differ significantly from one another concerning case representation and reasoning. The first one is called Structural in which a common structured vocabulary is developed, i.e. ontology. The second is textual in such way cases are represented as free text, i.e. strings. The third is the Conversational CBR in which a case is represented through a list of questions that vary from one case to another; knowledge is contained in customer / agent conversations [5].

During the past twenty years, many CBR applications have been developed, ranging from prototypical applications built in research labs to large-scale fielded applications developed by commercial companies[6].

The common application areas of CBR include help-desk and customer service, recommender systems in electronic commerce, knowledge and experience management, medical applications and applications in image processing, applications in law, technical diagnosis, design, planning and applications in the computer games and music domain [6].

CBR shells are kinds of application generators with graphical user interface. Nonprogrammer users can use them but the extension or integration of new components in these tools are not possible.

There is a clear difference between a CBR application and a CBR shell. A CBR application is a direct implementation of CBR methodology to a specific domain problem in order to solve this problem. On the other hand, a CBR shell is an application that enables developers to develop a CBR application.

3 CBR SOFTWARE

This section introduces a brief description of each software used in this paper.

TABLE I
CBR SOFTWARE USED

CBR Shell	**Author**
CBR Shell	AIAI, Stuart Aitken
FreeCBR	Lars Johanson
jCOLIBRI	University Complutense Madrid, GAIA group
myCBR	German Research Center for Artificial Intelligence
eXiTCBR	University of Girona

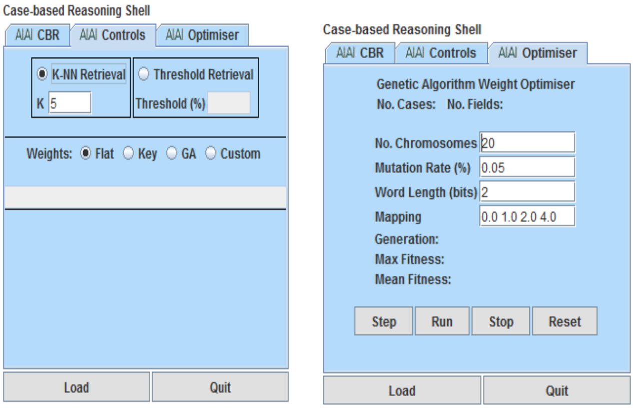
A. CBR Shell

The AIAI CBR [8] Shell is a generic tool for case-based reasoning. The tool performs classification based on case comparison. The parameters of the algorithm can be varied: the number of nearest neighbors considered can be specified, the weights can be set manually, or the weights can be optimized by genetic algorithm. The accuracy of the algorithm is measured by a leave-one-out evaluation.

The data must be in comma-delimited form, where a newline delimits a case. The first line of the case base must contain the name of the key file, the second states the goal field. The key file defines the type of matching that is done on each field in each case. The matching types include:

- Num-numerical comparison by evaluating the ratio of 2 numbers
- Stringexact-string comparison (equality test)
- Trigram-comparison of strings/sentences/paragraphs by trigram matching

Figure 2 shows a sample screen shots of the application while testing a sample case base.



● K-NN Retrieval ○ Threshold Retrieval
K 5
Threshold (%)
Weights: ● Flat ○ Key ○ GA ○ Custom

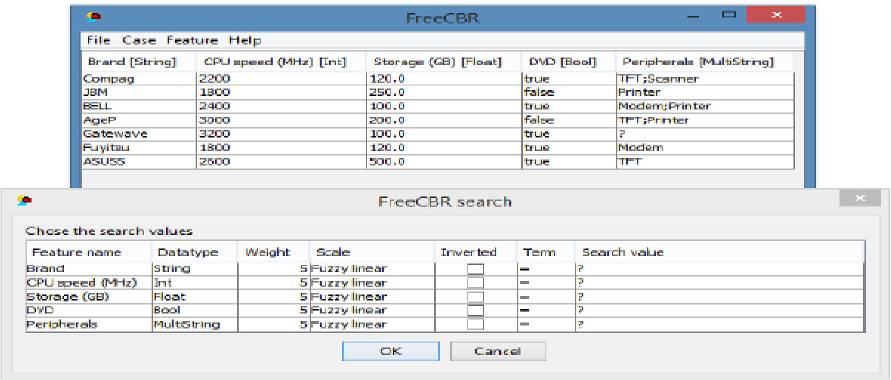
Genetic Algorithm Weight Optimiser
No. Cases: No. Fields:
No. Chromosomes 20
Mutation Rate (%) 0.05
Word Length (bits) 2
Mapping 0.0 1.0 2.0 4.0
Generation:
Max Fitness:
Mean Fitness:

Figure 2: CBR shell GUI interface

B. FreeCBR

FreeCBR[9] is a free open source Java implementation of a Case Based Reasoning tool. Cases are stored as text cases; each case is a set of features. Each case consists of a predefined set of features. It finds the closest match among cases in a case set. The closestmatch is calculated using weighted Euclidian distance, Normal Distance algorithm, and Logarithmic Distance algorithm. It only supports selection and retrieval phases.

Author details must not show any professional title (e.g. Managing Director), any academic title (e.g. Dr.) or any membership of any professional organization (e.g. Senior Member IEEE).



Brand [String]	CPU speed (MHz) [Int]	Storage (GB) [Float]	DVD [Bool]	Peripherals [MultiString]
Compag	2200	120.0	true	TFT;Scanner
JBM	1800	250.0	false	Printer
BELL	2400	100.0	true	Modem;Printer
AgeP	3000	200.0	false	TFT;Printer
Gatewave	3200	100.0	true	?
Fuyitsu	1800	120.0	true	Modem
ASUSS	2600	500.0	true	TFT

Chose the search values

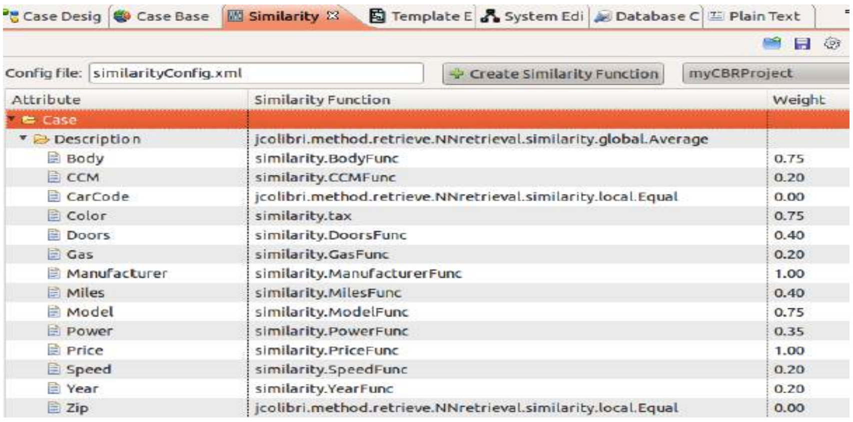
Feature name	Datatype	Weight	Scale	Inverted	Term	Search value
Brand	String	5	Fuzzy linear	□	–	?
CPU speed (MHz)	Int	5	Fuzzy linear	□	–	?
Storage (GB)	Float	5	Fuzzy linear	□	–	?
DVD	Bool	5	Fuzzy linear	□	–	?
Peripherals	MultiString	5	Fuzzy linear	□	–	?

Figure 3: FreeCBR

C. jCOLIBRI

jCOLIBRI [10] is a framework for developing various CBR applications. It is Java-based and uses JavaBeans technology for case representation and automatic generation of user interfaces.

jCOLIBRI supports full CBR cycle. At Retrieve stage, the nearest N cases are retrieved and there are 5 retrieval strategies, 7 selection methods and over 30 kinds of similarity functions (SF) in the spheres of text formatting and ontology. At the Refusal stage, several methods for adaptation are available (direct proportion) and also in ontology. At Revise stage, methods for revision of cases are realized, as well methods for new indexes (IDs) generation and methods for decision making (preference elicitation). At Retain stage, there are methods for query retaining as a new case. The maintenance algorithms such as RENN, BBNR, etc. are also supported. jCOLIBRI allows retrieval form clustered and indexed case bases and submits program interfaces (connectors) to access text and XML files, as well standard and DL (descriptive logic) data bases. These interfaces can be used for diagnostic systems databases access. All CBR cases can be represented graphically. There are lots of CBR applications, developed on jCOLIBRI base: additional shells (abstract levels) for distributed CBR systems, statistical CBR systems, multiagent supervisor systems [10, 11,12], systems for text files classification, and a lot of CBR recommender systems intended to trip, car type or restaurant choice, and other of discrete-event type. Figure 4 shows a sample screen shot of jCOLIBRI



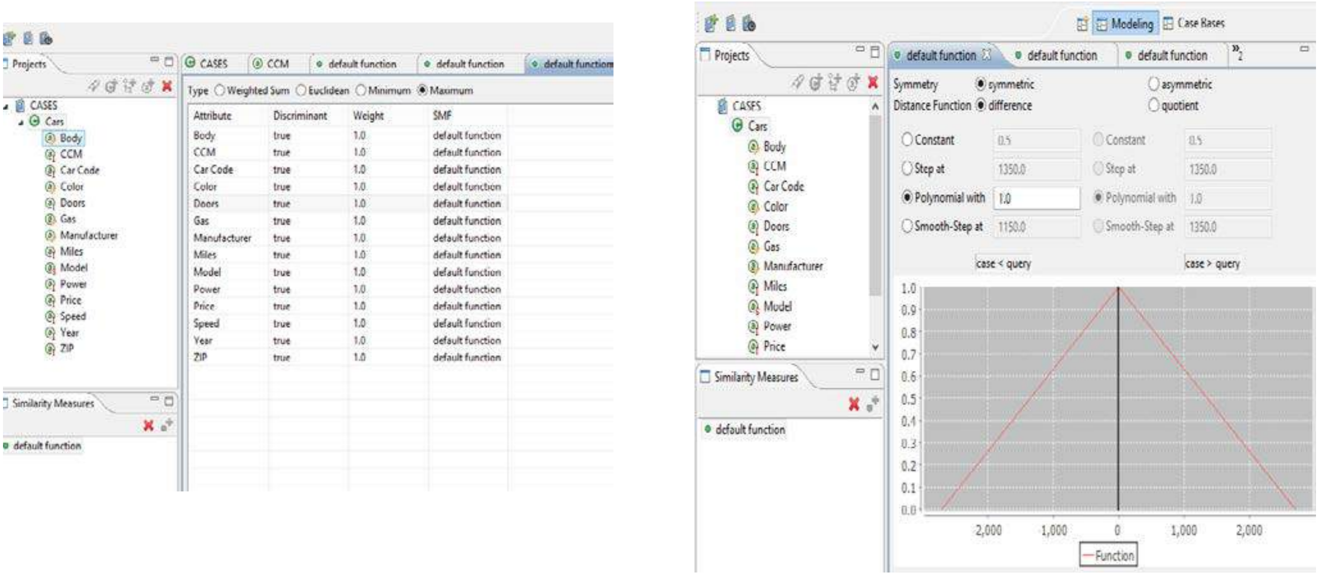
Config file: similarityConfig.xml

Attribute	Similarity Function	Weight
Case		
Description	jcolibri.method.retrieve.NNretrieval.similarity.global.Average	
Body	similarity.BodyFunc	0.75
CCM	similarity.CCMFunc	0.20
CarCode	jcolibri.method.retrieve.NNretrieval.similarity.local.Equal	0.00
Color	similarity.tax	0.75
Doors	similarity.DoorsFunc	0.40
Gas	similarity.GasFunc	0.20
Manufacturer	similarity.ManufacturerFunc	1.00
Miles	similarity.MilesFunc	0.40
Model	similarity.ModelFunc	0.75
Power	similarity.PowerFunc	0.35
Price	similarity.PriceFunc	1.00
Speed	similarity.SpeedFunc	0.20
Year	similarity.YearFunc	0.20
Zip	jcolibri.method.retrieve.NNretrieval.similarity.local.Equal	0.00

Figure 4:jCOLIBRI GUI

D. myCBR

myCBR [13] is an open-source similarity-based retrieval tool and software development kit (SDK). The framework my CBR supports description of cases with various attributes: numeric, character and string, logical, class type, etc. The templates of the cases are generated as classes or subclasses with a number of attributes, called slots (Figure 5b).



Type ○ Weighted Sum ○ Euclidean ○ Minimum ● Maximum

Attribute	Discriminant	Weight	SMF
Body	true	1.0	default function
CCM	true	1.0	default function
Car Code	true	1.0	default function
Color	true	1.0	default function
Doors	true	1.0	default function
Gas	true	1.0	default function
Manufacturer	true	1.0	default function
Miles	true	1.0	default function
Model	true	1.0	default function
Power	true	1.0	default function
Price	true	1.0	default function
Speed	true	1.0	default function
Year	true	1.0	default function
ZIP	true	1.0	default function

(a)

Symmetry ● symmetric ○ asymmetric
Distance Function ● difference ○ quotient

	case < query		case > query
○ Constant	0.5	○ Constant	0.5
○ Step at	1350.0	○ Step at	1350.0
● Polynomial with	1.0	● Polynomial with	1.0
○ Smooth-Step at	1150.0	○ Smooth-Step at	1350.0

(b)

Figure 5: a:myCBR case base, b: Editing the local similarity function viaGUI

The CBR Cases are objects of the class described by its attributes. Each attribute can participate in the class with its value and weight that determine the significance of the attribute in relation to others. Attributes with weight of zero (0) are not considered when searching the case-base DB.

Usually, case decisions have attributes with zero weight. In myCBR are given the opportunity to edit the similarity functions (SF) on class level (global SF) and on an attribute level (local SF). At the class level the SF are: weighted sum, Euclidean difference, maximum or minimum. On attribute level, the SF can be modified through the GUI, as shown in the middle of Figure 5a and they can be symmetrical, asymmetrical, step-type or smooth step-type, linear or polynomial.

In myCBR the case and their attributes can be created manually or automatically. The automatic generation of attributes (slots) is done during the import procedure of the Comma Separated Value (CSV) file. Then to each column name from the CSV file is assigned an attribute with the same name. To each row of the file the new case (instance of the class) is created in the case-base DB.

E. eXiTCBR

eXiTCBR [14] is a case-based reasoning tool developed at the eXiT research group of the University of Girona. It goes beyond pure CBR prototyping and aims to support experimentation.

The eXiTCBR framework was designed to bring together CBR methods currently used in medical applications and data mining and visualization techniques that can be plugged into it. eXiTCBR also facilitates the incorporation of new techniques, if required.

eXiTCBR architecture follows a modular approach based on the different phases in a CBR system. Each CBR step is implemented as generic class. When a new method is required but not provided in the system, it can be assembled as a particular instance of a generic class. When other techniques need to be integrated or hybridized, the corresponding executable codes should also be included.

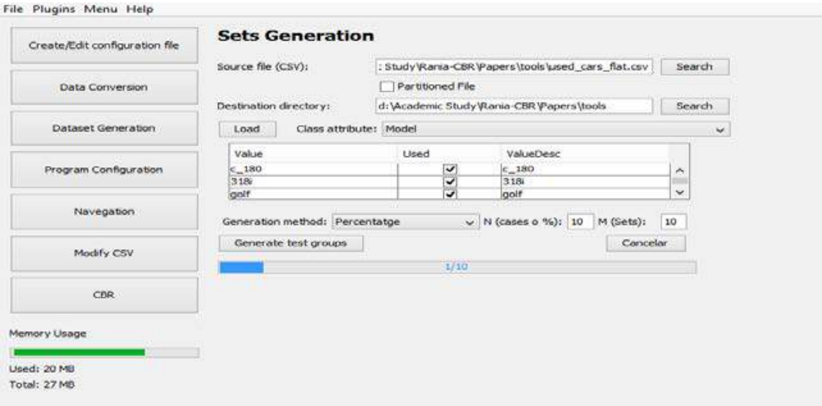
Input file Requirements:[15]
- First row: description of the attributes
- Second row: short name of the attributes, in a single word
- Third row: attribute type:
    0: discrete (as Nause, LumbarPain, Urine, Micturition, Burning)
    1: numeric (as Temp)
    2: text
    -1: do not take into account (as identifier, or the classes).

-Fourth row: attribute weight
    - First Column: case identifier. There should be a different identifier for all of the cases
    - Class attribute: it should be numeric (0-positive, 1-negative, or the other way around). In the example, Inflammatory is a diagnosis (class attribute), as well as nephritis.
    - No empty lines at the end of the file
    - Decimal numbers expressed with dots (39.0). Do not use dots for thousands.

-Current eXITCBR version provides support for the first two steps.

In selection the retrieved cases, the selection methods select the best cases from with the solution of a new case is provided at the reuse phase. The selection methods available are the following:

- Select1K. The best / most similar case will only be used.
- Select NK. The N best similar cases will be used.
- Select Threshold. The cases close to the new case will be used. “Close” is modeled by a threshold that needs to be specified at the right box. Use this option as a first attempt to use the tool, indicating in the right box 0.7 this option illustrates in a nice way the plots that result after running the experiment.
- Null. No method is applied. All the cases are used.



Source file (CSV): : Study\Rania-CBR\Papers\tools\used_cars_flat.csv

□ Partitioned File

Destination directory: d:\Academic Study\Rania-CBR\Papers\tools

Class attribute: Model

Value	Used	ValueDesc
c_180	☑	c_180
318i	☑	318i
golf	☑	golf

Generation method: Percentatge N (cases o %): 10 M (Sets): 10

Memory Usage

Used: 20 MB

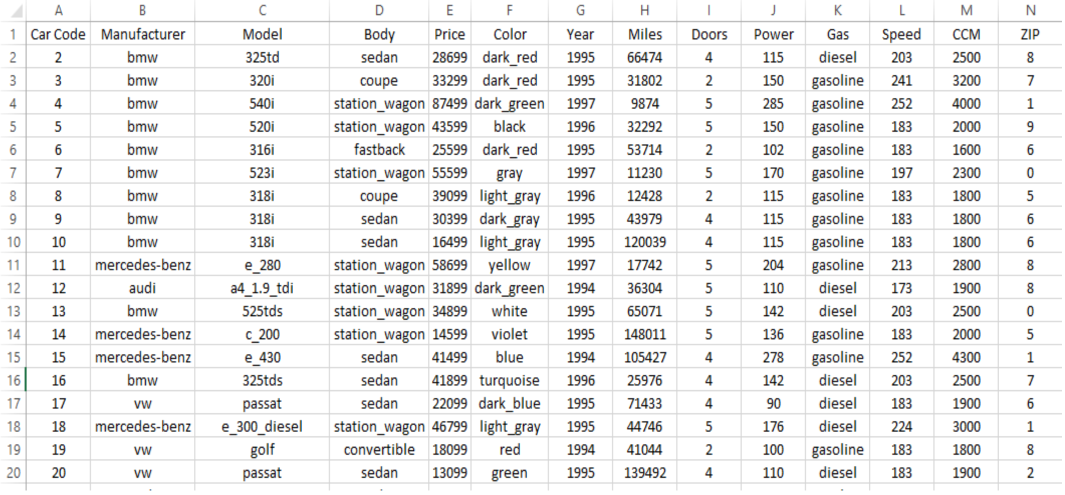
Total: 27 MB

Figure 6: eXiTCBR CBR

4 Comparative Study

This section introduces a comparative study after testing and comparing the CBR applications mentioned previously in table 1 using the same case base.

The case base used for testing the previously mentioned software obtained from the UC Irvine Machine Learning Repository, which contains details for 1000 cases for used cars [16].



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Car Code	Manufacturer	Model	Body	Price	Color	Year	Miles	Doors	Power	Gas	Speed	CCM	ZIP
2	2	bmw	325td	sedan	28699	dark_red	1995	66474	4	115	diesel	203	2500	8
3	3	bmw	320i	coupe	33299	dark_red	1995	31802	2	150	gasoline	241	3200	7
4	4	bmw	540i	station_wagon	87499	dark_green	1997	9874	5	285	gasoline	252	4000	1
5	5	bmw	520i	station_wagon	43599	black	1996	32292	5	150	gasoline	183	2000	9
6	6	bmw	316i	fastback	25599	dark_red	1995	53714	2	102	gasoline	183	1600	6
7	7	bmw	523i	station_wagon	55599	gray	1997	11230	5	170	gasoline	197	2300	0
8	8	bmw	318i	coupe	39099	light_gray	1996	12428	2	115	gasoline	183	1800	5
9	9	bmw	318i	sedan	30399	dark_gray	1995	43979	4	115	gasoline	183	1800	6
10	10	bmw	318i	sedan	16499	light_gray	1995	120039	4	115	gasoline	183	1800	6
11	11	mercedes-benz	e_280	station_wagon	58699	yellow	1997	17742	5	204	gasoline	213	2800	8
12	12	audi	a4_1.9_tdi	station_wagon	31899	dark_green	1994	36304	5	110	diesel	173	1900	8
13	13	bmw	525tds	station_wagon	34899	white	1995	65071	5	142	diesel	203	2500	0
14	14	mercedes-benz	c_200	station_wagon	14599	violet	1995	148011	5	136	gasoline	183	2000	5
15	15	mercedes-benz	e_430	sedan	41499	blue	1994	105427	4	278	gasoline	252	4300	1
16	16	bmw	325tds	sedan	41899	turquoise	1996	25976	4	142	diesel	203	2500	7
17	17	vw	passat	sedan	22099	dark_blue	1995	71433	4	90	diesel	183	1900	6
18	18	mercedes-benz	e_300_diesel	station_wagon	46799	light_gray	1995	44746	5	176	diesel	224	3000	1
19	19	vw	golf	convertible	18099	red	1994	41044	2	100	gasoline	183	1800	8
20	20	vw	passat	sedan	13099	green	1995	139492	4	110	diesel	183	1900	2

Figure 7: Case-base snapshot

There are a number of major concerns when studying case-based reasoning approach. These major concerns are listed below:

- What is the structure of the cases?
- What are the selection strategies for finding similar cases?
- How is the case being retrieved?
- How is the selected case being revised?
- How is the suggested case being stored in case base?
- How is the suggested case being indexed for faster access?

- How to deal with noisy data or missing values?

According to the previous mentioned points, a comparative study between the CBR software mentioned previously in table 1 in section 3 is done. Next paragraphs describe the effect of each factor to each CBR software respectively.

After applying the same query to all CBR software, the researcher has discovered the following: CBR Shell, very simple interface, the retrieval can use KNN or Threshold and weights can be specified manually, it uses genetic algorithm for optimization, no case revised and cases are stored in custom text files, no case indexing.

FreeCBR has a very simple GUI interface, find the "closest" match of the stored cases, the closest match is calculated using weighted Euclid distance. FreeCBR finds the closest match among cases in a case set. Each case consists of a predefined set of features. The features are defined by a name and a data type where the data type may be String, MultiString, Float, Int and Bool.
jColibri has a very simple and powerful GUI, it represents cases in a very simple way. jColibri allows retrieving cases using a SQL query and then it organizes cases after they load into memory and the case can be graphically presented.

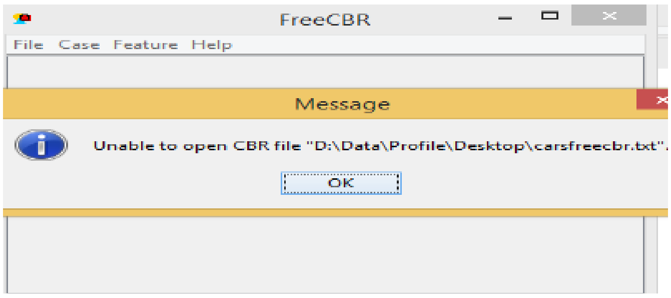
There are a number of case retrieval algorithms applicable in case based reasoning. These algorithms are based on the similarity metric that allows resemblance between cases stored in case base. The nearest neighbor retrieval algorithm & induction retrieval algorithms are two chief algorithms used in this process. Nearest-neighbor retrievalis a straightforward approach that computes the similarity between relevant cases found through indexing. The case is elected on worth of weighted computation of its feature. When the value of weighted calculation of its features is greater than other cases, then meticulous case is elected from the case base.
jCOLIBRI can be used as a basis for complex CBR applications development with full CBR R4 cycle, using various data bases. jCOLIBRI supports working with external Database and external sources.

myCBR has a simple GUI. The cases are very simple. They support only Retrieve and Retain phases. During the Retrieve phase, all precedents are extracted. They are sorted by degree of similarity based on the chosen global SF. The Query to the case-base DB could be done on the basis of all or part of the attributes, describing the case.
myCBR does not work with external DB. It stores the cases in text file or in XML file. That's way it cannot support the case indexation and categorization. The case cannot be graphically presented in the GUI, but it is possible to present the distribution of values of a selected attribute for all cases in the database.

No interfaces to external systems and DB are available in myCBR. It is valid regarding the interfaces to real-time or diagnostic systems. On Retain phase, myCBR allows saving the Query as a new case, also to use an old case as a basis for new Query. MyCBR is entirely based on GUI, providing a ready-windows templates and forms for defining classes, attributes, SFs, queries to the case-base DB, visualization of found results and more.

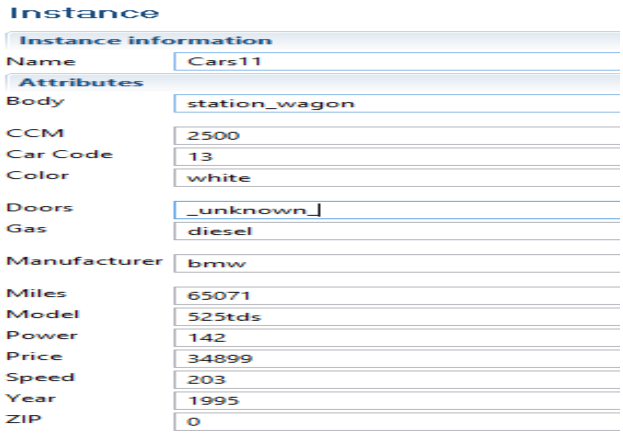
myCBR platform can be used for non-complex CBR applications development with partial CBR R4 cycle and with small number of cases in text file. For CBR application development, no time for programming is needed but it is needed only for case configuration. MyCBR is not suitable to be applied with large number of attributes with text solution, especially when they must be visually presented in one window. Table 2 summarizes the comparisons between the selected CBR software.

Concerning the missing data and unknown values in the case base software, CBR Shell, FreeCBR, and eXiTCBR can’t load the case base if modified or edited with removing values as shown in figure 8. myCBR handles this type of noise by adding two extra values (_unknown_ and _undefined_) to an attribute as shown in figure 9.



Unable to open CBR file "D:\Data\Profile\Desktop\carsfreecbr.txt".

Figure 8: FreeCBR and missing values



Instance

Instance information

Name: Cars11

Attributes

Body: station_wagon
CCM: 2500
Car Code: 13
Color: white
Doors: _unknown_
Gas: diesel
Manufacturer: bmw
Miles: 65071
Model: 525tds
Power: 142
Price: 34899
Speed: 203
Year: 1995
ZIP: 0

Figure 9::myCBR and missing values

As an addition to the comparative study, it must be taken in consideration the evaluation of each software, using some statistical measures like precision, recall, and F-Measure.Precision is the probability that a retrieved case is relevant, Recall is the probability that a relevant case is retrieved in a search, and F-Measure is the harmonic mean of recall and precision together which appeared in table 2.

Table 2 also shows a statistical view after applying the same query to the five CBR softwares. The results show that the highest accuracy reached and the number of cases retrieved and matched through the jCOLIBRI followed by myCBR, FreeCBR,CBR Shell and eXiTCBR respectively.

TABLE 2
CBR SHELL COMPARISION

CBR Shell	Case Structure	Selec-tionstrate-gies	Case retrieval	Case revised	Case stora-ge	Case indexed	Graphical User Interface(GUI)	Dealing with uncertain data	Correct Match Cases	Precision	Recall	F-Measure	Accuracy
CBR Shell	Text-ual	distance method	Two methods KNN Threshold	Manual	Text	No	Very simple GUI	Can’t handle	70	0.666	0.7	0.682	0.703
Free CBR	Text-ual	weighted Euclid distance	Simple matching	Manual	Text	No	Simple and easy but limited	Can’t handle	81	0.764	0.81	0.786	0.813
jCOLIBRI	Xml /text	Similari-tyfunc-tions	method k-NN, Threshold, Ontology, Textual, OpenNLP and GATE Recommender	Automatic	CSV XML	Yes	Simple and power-ful Use wizard to simplify	Handle as null	99	0.933	0.99	0.961	0.994
myCBR	Object	similarity functions	Query model	Manual	CSV XML	No	user can customize the GUI and handle most of things	Handle as _unknown_ or _undefined_	93	0.902	0.93	0.916	0.934
eXiT CBR	Custom CSV	distance method or similarity measure	Simple Querying	Manual	Text	No	Very simple , no options	Can’t handle	61	0.603	0.61	0.606	0.613

5 CONCLUSIONS

This paper introduces a comparison among most common used CBR software. It also mentions the advantages and disadvantages of each software. Moreover, this paper applies the same case base to the five CBR software to compare and evaluate the results using the predetermined factors and calculating Precision, Recall ,F-Measure and Accuracy for each one. As a conclusion CBR, Free CBR and eXit CBR are very simple software including simple GUI and only include the selection and retrieval of similar cases using traditional techniques. On the other hand, both myCBR and jCOLIBRI are more complex and can be used for complex CBR. myCBR interfaces over matches jCOLIBRI's and provides more options as weights and Similarity functions, type modification of attributes and cases. This is of great importance for query adjustment and refining the case base.There should be a new technique to improve the CBR process, which needs to be developed and tested. The new idea is to use a semantic approach to store and retrieve cases with added meta-data to the cases itself to help in measuring similarity instead of the traditional techniques.

REFERENCES

[1]A. Aamodt and E. Plaza, “Case-based reasoning: foundational issues, methodological variations, and system approach” *AI Communications* 7(1), 39–59, 1994.

[2] Janet L. Kolodner, “An Introduction to Case-Based Reasoning” *Artificial Intelligence Review* 6, 3-34, 1992.

[3] Riesbeck, C. and Schank, R. *Inside Case-Based Reasoning*, Lawrence Erbaum Associates, Inc., 1989.

[4] Doyle M., Hayes, C., Cunningham, P. and Smith, B. *CBR Net: Smart Technology over a Network*. Department of Computer Science, Trinity College Dublin., 1998.

[5] Zhi-We Ni, Shan-Lin “Integrated Case-based Reasoning” *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, XI; 2-5 November 2003.

[6] Leake, David, "CBR in Context: The Present and Future”, http:/ / www. cs.indiana.edu/~leake/papers/p-96-01_dir.html/paper. html, In Leake, D., editor, *Case-Based Reasoning: Experiences, Lessons, and Future Directions.* AAAI Press/MIT Press, 1-30, 1996.

[7] Simon C.K. Shiu, “Case-Based Reasoning: Concepts, Features and Soft Computing” *Applied Intelligence* 21, 233–238, 2004.

[8] http://www.aiai.ed.ac.uk/project/cbr/CBRDistrib

[9] http://freecbr.sourceforge.net/

[10] Bello-Tomás,J.J., González-Calero, P. A., Díaz-Agudo, B. “JColibri: An Object-Oriented Framework for Building CBR Systems”. In *Advances in Case-Based Reasoning, Lecture Notes in Computer Science*.Springer Berlin/ Heidelberg, Vol. 3155/2004, p. 32-46,2004.

[11]K.Boshnakov,C.Boishina,M. Hadjiiski, “Multiagent fault-tolerant supervising control of wastewater treatment plants for wastewater”, *International Conference, Automatics and in formatics'11*,Bulgaria, Sofia,2011.

[12] M.Hadjiski, V.Boishina, “EnhancingFunctionality of Complex Plant Hybrid Control System Using Case-Based Reasoning”,*IntelligentSystems*(IS), 5th IEEE International Conference, 7-9 July 2010, London,25-30,2010.

[13] http://www.mycbr-project.net/index.html

[14] http://exitcbr.udg.edu/

[15]http://exit.udg.edu/download.aspx?id=13eXiT*CBR user tutorial - eXiT*CBR 3.0 user tutorial

[16] http://archive.ics.uci.edu/ml/

BIOGRAPHY

Passent elKafrawy, Associate Professor, Faculty Science, Menofia University

Dr. Passent M ElKafrawy is an Associate Professor since 2013, she got her PhD from the University of Connecticut in United states on 2006 in Computer Science and Engineering. In the field of computational geometry as a branch of Artificial Intelligence. Then she taught in Eastern State University of Connecticut for one year. In 2007 she worked as a Teacher in Faculty of Science, Menoufia University, Mathematics and computer science department since that time till now.

She has over 20 publications and member of ACEE, ECOLE and TIMA research organizations. One of the organizing members of the following conferences: SPIT and ESOLEC. Supervising over 10 research studies between PhD and MSc. Member of the faculty projects for education development as CIQAP, DSAP, and Question Bank.

Rania Ahmed ,Assistant Lecture in Modern University for Technology and Information,

was born in Giza, Egypt, in 1985.She received the Bachelor in Computer Science degree from the Helwan University in 2006 and the Master in Computer Science degree from the Helwan University in 2010.She is currently pursuing the PH.D degree with the Department of Computer Science.

دراسةمقارنة لحالة البرمجيات القائمة على المنطق واللغة الطبيعية

بسنت محمد الكفراوى*,رانيا احمد محمد**

*أستاذ مساعد , كلية العلوم جامعة المنوفية

**مدرس مساعد بالجامعة الحديثة للتكنولوجيا والمعلومات*

الملخص

المنطق القائم على الحالة (CBR) هو نموذج حلا لمشكلة التي يحل مشكلة جديدة بتذكروضع مماثلا لسابق وإعادة استخدام المعلومات والمعرفة من هذا الوضع. المنطق القائم على الحالة (CBR) هي البرمجيات التي يمكن استخدامها لتطويرالعديد من التطبيقات التي تتطلب منهجية التفكيرالقائم على القضية. ويستند CBR على الحدس أن المعلومات المكتسبة من التجارب السابقة (الحالات أوالحالات) يمكن أن يكون أداة هامة لتوفير حلول لها، وتعزيزالعمليات ذات الصلة، والمشكلة في متناول اليد. وبالتالي، فإنه يساعد على تحسين النتائج وتوفير الموارد القيمة. ومع ذلك، يمكن أن كميات كبيرة من المعلومات يجعل من مهمة معقدة لاكتساب المعرفة المفيدة من قواعد البيانات التاريخية. CBR هوأداة فعالة للحصول على معلومات مفيدة في مثل هذه الظروف وهي واحدة من تقنيات الذكاء الاصطناعي تطبق معظم بنجاح في السنوات الأخيرة. تعطي هذه الورقة معلومات أساسية عن برنامج التأهيل المجتمعي ويقدم الأكثر استخداما لأدوات التأهيل المجتمعي (CBR شل، FreeCBR، jCOLIBRI، myCBR وeXiTCBR). ثم انه يقدم دراسة تحليلية مقارنة على أساس بعض العوامل المحددة التي تؤثر على برنامج التأهيل المجتمعي بما في ذلك البيانات صاخبة أوالقيم المفقودة في الحالات. فإنه يقارن أيضا بين من لهم أخيرا، وقدم تقييما للمرحلة استرجاع كل البرامج.عملية توليد اللغة الطبيعية لوكيل المحادثة يترجم بعض العبارات الدلالية لشكل سطحه المعبر عنها في اللغة الطبيعية القضية استنادا أسلوب التفكيرالمعرض الذي ينزلق بسهولة وقابلة للتكيف مع المجالات وبلغات متعددة، أن يولد العبارات متماسكة وتنتج نتيجة طبيعية في سياق وكيل المحادثة التي تحافظ على الحوارمع المستخدم.

A Proposed Standardization for Arabic Sign Language Benchmark Database

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

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

[1]ahmed.new80@hotmail.com

[2]fahmytolba@gmail.com

***Abstract*- This The lack of a visualized representation for standard Arabic Sign Language (ArSL) makes it difficult to do something as commonplace as looking up an unknown word in a dictionary. The majority of printed dictionaries organize ArSL signs (represented in drawings or pictures) based on their nearest Arabic translation; so unless one already knows the meaning of a sign, dictionary look-up is not a simple proposition. In this paper we introduce the ASL database, a large and expanding public dataset containing video sequences of thousands of distinct ArSL signs. This dataset is being created as part of a project to develop an Arabic sign language translator. At the same time, the dataset can be useful for benchmarking a variety of computer vision and machine learning methods designed for learning and/or indexing a large number of visual classes especially approaches for analyzing gestures and human communication.**

Key words: **Arabic Sign Language (ARSL), Arabic Sign Language Database, Database Benchmark.**

1 INTRODUCTION

Arabic Sign language is different in each Arab region or/and country with many dialects. This difference gives the difficulty of communicating and dealing between deaf people in different Arabian countries. A need appeared to unify Arabic sign language in all Arabian countries. This derived the Council of Arab Ministers of Social Affairs (CAMSA) to take a decision of developing a unified Arab sign language dictionary and publish it to all countries, in an attempt to help Arab deaf people to have a common language in addition to their local language [1]. This dictionary is mostly used in education and in common communication such as sign language interpreters in television. Arabic sign language like other known sign languages depends on three basic factors that are used to represent the manual features: hand shape, hand location and orientation. In addition to the non-manual features that are related to head, face, eyes, eyebrows, shoulders and facial expression like puffed checks and mouth pattern movements. ASL is limited to represent nouns, adjectives and verbs. Prepositions and adverbs are represented in the context of articulation by specifying locations, orientations and movement. Intensifiers represented by iteration [1]. Signs forming and sequencing in the articulation, are done depending on the Arabic sign language grammar and rules.

Arabic sign languages (ARSLs) are still in their developmental stages. Only in recent years has there been an awareness of the existence of communities consisting of individuals with disabilities; the Deaf are not an exception. Arab Deaf communities are almost closed ones. Interaction between a Deaf community and a hearing one is minimal and is basically concentrated around families with deaf members, relatives of the deaf, and sometimes play friends and professionals [2]. As in other communities, communication with a deaf person is polarized within such circles. This situation has led to the emergence of many local means of sign communication. Until recently, such signs have not been gathered or codified. Signs are starting to spread, forming acknowledged sign languages. By and large, the view held vis-a-vis disability, including hearing, in the Arab society is still one of accommodation rather than assimilation [2].

Sign languages all over the world are not a new invention. They existed on par with the spoken languages. Their invention cannot be attributed to any person. Rather, they developed naturally just as other verbal languages. Similarly, ARSLs have been developing naturally. In their ‘‘natural context,’’ ARSLs developed as in-dependent systems of communication. They are not interpretations of standard Arabic or spoken vernaculars [2].

2 ARABIC SIGN LANGUAGE

ARSLs share many similarities and manifest certain features of difference. After all, this is true for all languages; indeed, trace features of universality can be traced among the sign languages of the world. Basically, ARSLs developed independently, although some have benefited from the pioneer experience of the others. The possible sources of ARSLs could be traced to the following:

- Borrowings, especially European and American.
- Creations, which are initialization of conceptual signs usually by gestural repertoire of spoken varieties.
- Miming actions, shapes, and things in nature.
- Expanding means, such as compounding and blending.
- ‘‘Dumb’’ regional signs, which are basically signs inherited over centuries, used by ‘‘mute’’ people, and of a local nature.

Finger spelling is fairly new and is mostly a combination of creation and miming source. It is used to spell out proper nouns and words that do not have sign correspondence. Finger spelling, however, is not used to read out or communicate the standard form of Arabic. Therefore, there is no ‘‘manual Arabic’’ yet; perhaps such form of signed standard Arabic might develop if the deaf are to be educated through sign language and if need arises to have a signed Arabic that corresponds to the standard. Further, there has been no attempt so far to write down ARSLs (sign writing). ASL, for example, has established writing systems, but these have not been widely used to record ASL literature; however, there is a large body of ASL literature available in movies, videotapes, and compact disks [3].

Arabic, on the other hand, has a considerable body of signed literature mainly in movies, TV series, and news bulletins; this body has been neither recorded nor utilized for the development of Arabic sign vernaculars.

Arabic sign languages are not particularly different from other known sign languages, such as BSL. In fact, the Arabic varieties in use have undergone some lexical influence from other sign languages [4]. ARSLs are basically manual languages made from cheremes that involve the three recognized elements: configuration of hands (hand shape), placement/space (position of hand in relation to body), and movement (directions and contacts within space). In addition to these manual shapes, ARSLs make use of other non-manual features, like those of the face, mouth, and tongue.

Arabic sign languages also exhibit similar forms to other established sign languages, such as links between form and meaning that may be iconic, pictorial, conventional, or arbitrary [5]. Arabic sign languages’ word correspondence (i.e., signs) is limited to two basic classes, nouns/adjectives and verbs, and lacks, unlike standard Arabic, many of the particles (e.g., prepositions and some adverbs or intensifiers). However, the relationships and concepts represented by prepositions and intensifiers, for example, can be expressed by other means. This could be done by the position and direction of one sign in relation to another in the case of prepositions and by repetition of sign regarding intensifier [6]. Other vocabulary items can be explained under the following categories: synosigns, antosigns, homosigns, and compounds.

- Synosigns: usually two different signs with one meaning are not common in ARSLs. However, they do exist and mostly evolve as a result of shifting from one sign to another, and when the first sign is not totally abandoned, the two signs continue to coexist for some time until one, usually the second, dominates. Examples from Jordanian Sign Language are girl and rich.
- Antosigns: The type of antosigns present in ARSLs is mostly complementary pairs, which is different only in one element: movement. This makes an to signs in sign language different from antonyms in spoken languages, in which the sounds and meaning are different
- Homosigns: Arabic sign languages use some homo-signs. There is no difficulty in understanding the referential meaning of such signs, which is usually clear from the context
- Compounds: A very important method to expand vocabulary is through compounding. This is also true for sign languages, including Arabic. Whenever two signs can give the meaning of another concept when combined, they are employed to do so, especially in developing sign languages such as those of Arabic. Indeed, it is much easier to understand a concept in relation to another rather than to invent one; consider these examples: dentist, internist, vet, and dream.

Arabic sign languages are similar to other sign languages of the world in that they are basically spatial–gestural languages. This makes it difficult to compare sign languages with their spoken counter-parts; Arabic in this regard is not an exception. As a matter of fact, many concepts used to describe spoken languages are inadequate for the description of sign languages. Nevertheless, inevitably, one system should be mapped practically into the other.

Generally, ARSLs do not follow the same order of their spoken or written counterparts. Usually, a reversed order is used. This is because sign languages are highly schematized and indeed more pragmatic than the spoken ones. In Arabic, emphasis is given to content signs, those representing nouns and verbs. The nominal ‘‘sentence’’ is usually made up from a subject and a predicate, such as ‘‘she/he deaf ’’ [6]. And, unlike spoken and written varieties, there is no singular, dual, or plural agreement in ARSLs.

Signed sentences, on the other hand, do not make use of tense/aspect as in spoken and written varieties. Tense is simply and practically used. Past, present, and future times are indicated at beginnings of conversation chunks and only shifted when there is need to indicate a different tense (e.g., worked). Negatives and interrogatives have more than one way of expression. While in some cases non manual gestures are important (e.g., raised eyebrows, head and shoulders leaning forward, signed question mark), in other cases signs are used, for instance, ‘‘red not’’.

As for other grammatical features like emphasis and adverb position, emphasis is done by repetition, longer signing time, and facial expressions and dramatization; adverbs are explained manually, by one hand's position in relation to the other. Other features, such as passivization, declension, and indeclension, are nonexistent. Conditional expressions, sentence boundaries, and turn taking are usually achieved by non-manual features of facial expressions and context.

Sign languages show greater link between form and meaning than spoken languages [5]. Arabic word order is so flexible that it allows for one meaning to be expressed in different formal structures, such as V-S-O (verb-subject-object), S-V-O, O-V-S, V-O-S. This makes the structure of ARSLs familiar, especially to hearing learners, and easily comprehensible to the uneducated (most deaf people in the Arab countries are) because of their grammatical simplicity, which does not exist in standard Arabic. All this in my opinion makes sign language in general and Arabic in particular more ‘‘pragmatic’’ than the spoken varieties of language, which adds to the advantages of sign language more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

3 RELATED WORK

For evaluation and benchmarking of automatic sign language recognition, large corpora are needed. Recent research has focused mainly on isolated sign language recognition methods using video sequences that have been recorded under lab conditions using special hardware like data gloves. Such databases have often consisted generally of only one speaker and thus have been speaker-dependent, and have had only small vocabularies. Most databases used in sign language processing so far do not provide or include what is important for the evaluation of sign language processing algorithms.

The National Center for Sign Language and Gesture Resources at Boston University has published an expanding database of American Sign Language (ASL). Dreuw and colleagues from the RWTH Aachen University created several subsets for the evaluation of isolated and continuous sign language recognition: RWTH-BOSTON-50 [7, 8], RWTHBOSTON- 104 [9], and the new RWTH-BOSTON-400.

The new RWTH-BOSTON-400 is the largest publicly available benchmark corpus for video-based continuous sign language recognition. It contains 843 sentences, several speakers, and separate splits for training, development, and testing of automatic sign language recognition systems.

The RWTH-BOSTON-400 database is created from a subset of the larger data set available through Boston University. The BU ASL corpus has been used previously in evaluation of computer vision and pattern recognition methods, including detection of head gestures [10], recognition of facial expressions [11], hand tracking and recognition of hand shapes and movements [12, 13, 14].

The National Center for Sign Language and Gesture Resources (NCSLGR) at Boston University has been engaged in the collection of ASL data (including sets of individual utterances, narratives, and dialogues) from Deaf native signers.

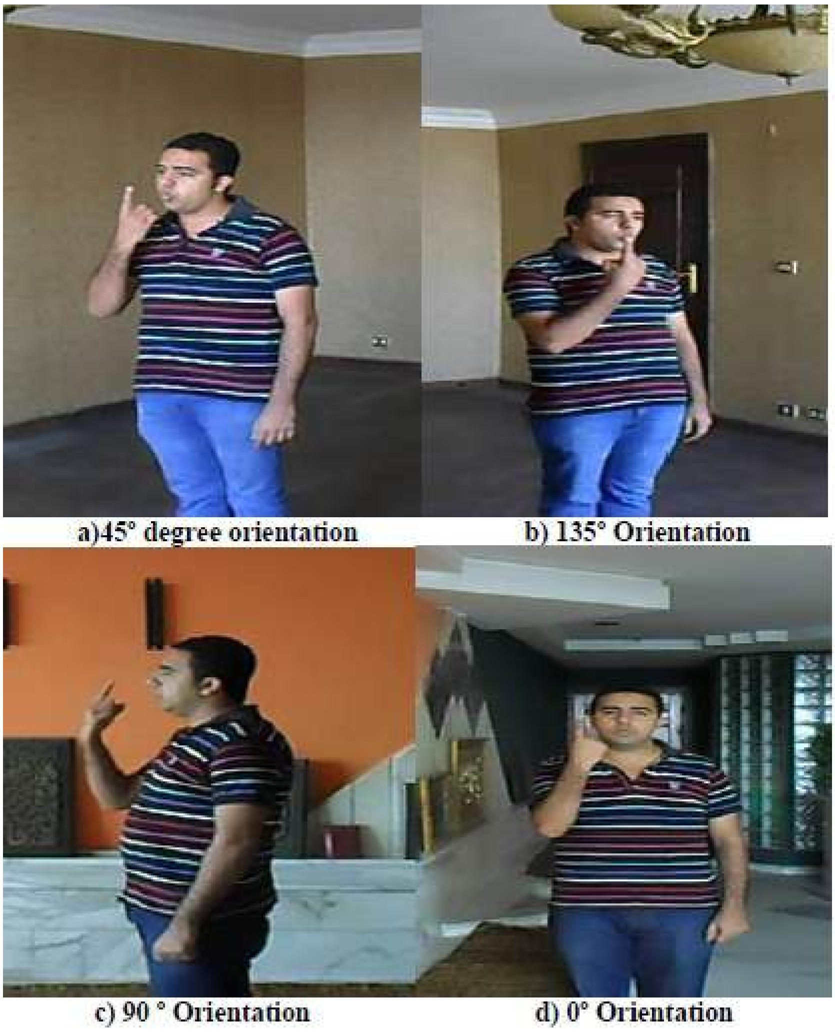
The NCSLGR makes available high-quality video files showing the signing from multiple angles, including a close-up of the face, in a variety of video formats, along with linguistic annotations that have been carried out in conjunction with the American Sign Language Linguistic Research Project (ASLLRP) at Boston University, using Sign Stream [15, 16].

4 ARSL DATABASE BENCHMARK

This benchmark database is being created as part of a project to develop an Arabic sign language translator. This project aims to provide a credible tool of communication between the deaf sector and the community. The first milestone in the project is to build a national digital database for standard Arabic sign language.

The Standard Arabic sign language dictionary can be categorized in 27 categories of words. The number of signs is: 1216 signs including the alphabets and numerical. Four sign language experts have captured the complete dictionary and two different experts have reviewed the signs validity. The four sign experts are deliberately picked as: two are a left-handed person and the other are a right-handed. For each captured person:

- Videos are captured from 4 different angles (Fig. 1).
- Recording signs' videos using different viewing angles (0, 270 then 315, 225).We recorded videos of 5, 10, 30 and 50 frames per second. The recording has been done for 2 different persons; each sign is recorded 3 times.


a)45° degree orientation

b) 135° Orientation

c) 90 ° Orientation

d) 0° Orientation

Figure 1: The word (one) captured by different orientation angles.

This concludes the database size as: 4*(1216*4) = 19,456 signs videos. The facial expressions represent a major non-ignorable key feature in identifying the meaning. (Ex: divorce and marriage have the same hand signs but differ in the facial expressions). A study has been conducted on the captured signs database conclude that nearly 72% of the signs mainly depend on facial expressions and body language to deliver the right meaning to the recipient .

The database benchmark has several orientations:

- The performance benchmark.
- The data validity benchmark.
- The data variation and generalization benchmark.

The formal Arabic sign language dictionary approved by the league of Arab states is used to build Arabic sign language database .In order to validate the created database; a survey was conducted with a sample of 80 deaf students who were participating in the various user studies. The sample of deaf people is chosen from 4 different certified NGOs with different levels of education.

The main objective of interface development is to ease the access of the required video(s) or sign of a specific word. The Database Access Interface is implemented in ASP.Net and runs in any modern browser. As discussed before, the procedure of building ASL at Boston University is used to build our database. The Interface provides high-quality video files showing the signing from multiple signers, multiple angles, including a close-up of the face, different lighting conditions,

in a variety of video formats as shown in figure 2. Summary of the number of recorded videos available in the interface with different acquisition techniques is outlined in table 1.

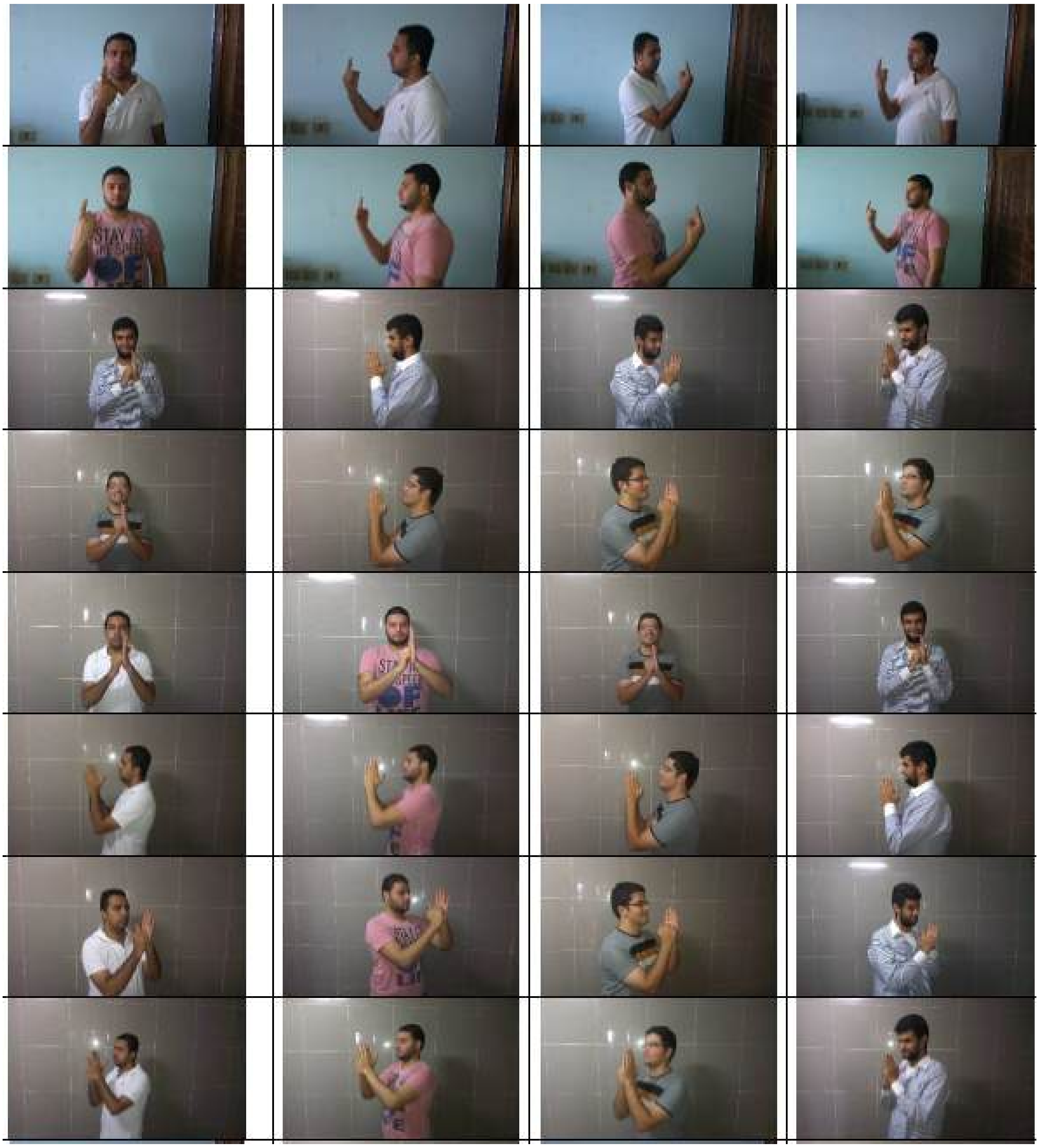


Figure 2: Different samples of the database with multiple signers, positions and lighting conditions

TABLE 1
NUMBER OF RECORDED VIDEOS FOR EACH SIGNERS AND DIFFERENT RECORDING CONDITIONS

	Signer 1	**Signer 2**	**Signer 3**	**Signer 4**
Number of Videos	1216	1216	1216	1216
Top-Side orientation	1216	1216	1216	1216
45-135 degrees	1216	1216	1216	1216
Close up face	980	682	733	822
Body Only	1216	720	1216	1216
Normal Lighting	1216	1216	1216	1216
Low Lighting	912	941	891	951

In addition, the interface provides with linguistic annotations that have been carried out in XML format. The interface allows users to query the data (or some user-specified subset of the data) in search of specific signs (or types of signs, e.g. finger-spelled signs), non-manual behaviors, or combinations thereof, while facilitating transfer of video files and annotations from the web site to the user's computer without the need of third-party software.

The annotations are available as XML files. Video files are available in a variety of formats that offer different trade-offs between file size and video quality. The original, uncompressed video sequences have resolution of 600x800 pixels, and were recorded at 60 frames per second. Grayscale and color cameras were used for recording the sequences. Each sequence was captured simultaneously by multiple (two to four) synchronized cameras: one or two cameras showing a front view of the upper body of the signer, one camera zooming in on the face from the front, and in many cases a camera showing the signer's upper body from the side. Calibration sequences are available for most of the recording sessions. The calibration sequences show a chessboard-like calibration pattern at a variety of 3D orientations, as seen from multiple cameras.

5 CONCLUSIONS

The In this paper, we described the recording of a new sign language corpus which meets the requirements for an Arabic sign language translator. The database is based on a vocabulary of 1216 basic signs in Arabic sign language and comprises 531 sentences each articulated by 4 different signers. The whole database will be made available for interested researchers in order to establish the first benchmark. The currently extracted features produce good recognition performance for a single trained signer. The experimental results reveal that they are robust enough for signer-independent sign language recognition.

BIOGRAPHY

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

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

REFERENCES

[1] S. Samreen and M. Benali,"قواعد لغة الإشارة العربية القطرية الموحدة ", 2009.

[2] M. A. Abdel-Fattah, "Arabic Sign Language: A Perspective". Journal of Deaf Studies and Deaf Education, vol. 10 no. 2, 2005.

[3] Wilcox, S. P., & Kreeft, J. "American Sign Language as a foreign language". ERIC Digest. Retrieved April 20, 2004, from http://www.ericfacility.net/databases/ERIC_Digest/ed49464.htm.

[4] Miller, C. "Disc: Arabic Sign Language", Re: 7.1101. Retrieved March 14, 2004, from http://www.linguislist.org/issues/7/7–1110.html.

[5] Brennan, M. "British Sign Language: The Language of the Deaf Community". In T. Booth and W. Swann (Eds.), Including Pupils with Disabilities: Curricula for All.Milton Keynes, UK: Open University Press, 1987.

[6] Suwed, A. A. " لغة الإشارة العربية لغة الصم – [Arabic Sign Language, Deaf language as in Libya]". Tripoli, Libya: AI-Mansha'ah AI Aamah Lin-Nasher wal I'lam, 1984.

[7] M. Zahedi, D. Keysers, and H. Ney. "Pronunciation clustering and modeling of variability for appearance-based sign language recognition". In International Ges-ture Workshop 2005, volume 3881, Vannes, France, May 2005.

[8] M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney. "Continuous sign language recognition approaches from speech recognition and available data resources". In Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios, pages 21–24, Genoa, Italy, May 2006.

[9] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. “Speech recognition techniques for a sign language recognition system”. In Interspeech 2007, pages 2513–2516, Antwerp, Belgium, August. ISCA best student paper award of Interspeech 2007.

[10] U. M. Erdem and S. Sclaroff. “Automatic detection of relevant head gestures in American Sign Language communication”. In International Conf. on Pattern Recogni-tion (ICPR), volume 1, pages 460–463, 2002.

[11] C. Vogler and S. Goldenstein. “Facial movement analysis in ASL”. Springer Journal on Universal Access in the Information Society, 2007.

[12] C. Vogler and D. Metaxa. “Handshapes and movements: Multiple-channel ASL recognition”. Springer Lecture Notes in Artificial Intelligence, (2915):247–258, 2004.

[13] Q. Yuan, S. Sclaroff, and V. Athitsos. “Automatic 2d hand tracking in video sequences”. In IEEE Workshop on Applications of Computer Vision, 2005.

[14] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. “Tracking using dynamic programming for appearance-based sign language recognition”. In IEEE Intl. Conf. on Automatic Face and Gesture Recognition, pages 293–298, Southampton, April 2006.

[15] C. Neidle, S Sclaroff, and V. Athitsos. “Signstream: A tool for linguistic and computer vision research on visual-gestural language data”. Behavior Research Methods, Instruments, and Computers, 3(33):311–320, 2001.

[16] C. Neidle. “Signstream: A database tool for research on visual-gestural language”. Journal of Sign Language and Linguistics, 1/2(4):203–214, 2002.

توحيد قياسي مقترح لبناء قاعدة بيانات موحدة للغة الاشارة العربية

أحمد سمير - قسم الحسابات العلمية - كلية الحاسبات و المعلومات - جامعة عين شمس
محمد فهمي طلبة - قسم الحسابات العلمية - كلية الحاسبات و المعلومات- جامعة عين شمس

خلاصة:

تعتبر الصعوبة في التمثيل الشكلي للغة الاشارة من أكبر العوائق و التحديات للبحث عن اشارة في القاموس مما يسبب منع انتشار مترجمات لغة الاشارة و وصول قاموس لغة الاشارة لمستخدميه. يهدف هذا البحث لاقتراح و تنفيذ أول قاموس للغة الاشارة العربية معتمدا علي فيديوهات متنوعة الزوايا, الاضاءة و منفذس الاشارات. نهدف من هذا وضع قاعدة بيانات علي أساس توحيد قياسي للبحث. يمكن الاستفادة من قاعدة البيانات في مجالات البحث الاكاديمي و الاستخدام العادي.

Continuous Space Language Modeling

Mohamed Talaat[*1], Sherif Abdou[**2], Mahmoud Shoman[*3]

[*]*Information Technology, Faculty of Computers and Information, Cairo University*
Cairo, Egypt

[1]mtalaat@fci-cu.edu

[3]m.essmael@fci-cu.edu

[**]*Information Technology, Faculty of Computers and Information, Cairo University*
Cairo, Egypt

[2]s.abdou@fci-cu.edu

Abstract—**This paper presents a long distance continuous language model (LM) based on a latent semantic analysis (LSA). In the LSA framework, the word-document co-occurrence matrix is commonly used to tell how many times a word occurs in a certain document. Also, the word-word co-occurrence matrix is used in many previous studies. In this research, we introduce a different representation for the text corpus, this by proposing long-distance word co-occurrence matrices. These matrices to represent the long range co-occurrences between different words on different distances in the corpus. By applying LSA to these matrices, words in the vocabulary are moved to the continuous vector space. In the LSA space, we represent each word with a continuous vector that keeps the word order and position in the sentences; this may help to attack the long-range dependencies problem occurs in many LMs. The word continuous vector is constructed by concatenating the reduced continuous vectors for that word from each projection matrix. We use tied-mixture HMM modeling (TM-HMM) to robustly estimate the LM parameters and word probabilities. Experiments on Giga Words corpus show improvements in the perplexity results compared to the conventional n-gram.**

1 Introduction

A language model (LM) plays an important role in automatic speech recognition systems (ASR). It determines how likely a word sequence would occur in terms of probability. In other words, the LM task is to estimates the probability$P(W)$for a given word sequence $W = w_1, w_2, \ldots, w_N$. Together with the acoustic model, LM has been used to reduce the acoustic search space and resolve acoustic ambiguity. Without it, ASR systems would not understand the language and it would be hard to find the correct word sequence.

N-gram model[1], [2],[3] is the most frequently used LM technique in all types of natural language processing, speech recognition and machine translation applications. There are several factors contributing to this reality. First, n-gram models are easy to build; all it requires is a plain text. Second, the computational overhead to build an n-gram model is virtually negligible given the amount of typically used data in many applications. Last, n-gram models are fast to use during decoding as it does not require any computation other than a table look-up. It defines the probability of an ordered sequence of n words by using an independence assumption that each word depends only on the last n-1 words. In case of trigram (n=3), the probability for the word sequence $W = w_1, w_2, \ldots, w_N$ is:

$$P_{trigram}(W) = \prod_{i=1}^{N} P(w_i | w_{i-2}, w_{i-1}) \tag{1}$$

In spite of this success, the n-gram suffers from some major problems. One of the key problems in n-gram modeling is the inherent data sparseness of real training data. If the training corpus is not large enough, many actually possible word successions may not be well observed, leading to many extremely small probabilities. This is a serious problem and frequently occurs in many LMs. Assigning all strings a nonzero probability helps prevent errors in speech recognition.

A technique called smoothing is partially solved the problem by ensuring that some probabilities are greater than zero for words which do not occur or occur with very low frequency in the training corpus. The basic idea of smoothing techniques is to subtract probability mass from the relative frequent seen events and distribute it to the unseen events. Smoothing methods can be categorized according to how the probability mass is subtracted (discounting) and how it is redistributed (back-off)[4], [5], [6], [7],[8],[9], [10].

Another way to avoid data sparseness problem is by mapping words into classes which is called a class-based LM, resulting a LM with less parameters. Class-based LM gives for infrequent words more confidence by relying on other more frequent words in the same class. The simplest class-based LM is known as class-based n-gram LM[11]. A common way to improve a class-based n-gram LM is by combining it with a word-based n-gram LM using interpolation method[12], [13]. Another approach is using a class-based n-gram LM to predict the unseen events, while the seen events are predicted by a word-based n-gram LM. This method is known as word-to-class back-off[14].

In addition to the data sparseness problem, the n-gram also suffers from the adaptability problem[15]. N-gram language model adaptation (to new domain, speaker, and genre) is very difficult using a relatively small amount of data, simply because of the huge number of parameters, for which large amount of adaptation data is required. The typical practice for this problem is to collect data in the target domain and build a domain specific language model. The domain specific language model is then interpolated with a generic language model trained on a larger domain independent data to achieve robustness[16].

Based on the Markov assumption, the word-based n-gram LMs are very powerful in modeling short-range dependencies but weak in modeling long-range dependencies, this because it uses only short-range dependencies and does not care about the long-range information. Many attempts were made to capture long-range dependencies. The cache-based LM[17] used a longer word history (window) to increase the probability of re-occurring words. Then there was a trigger-based LM[18], a generalization of the cache-based model. In this model, related words can increase the probability of the word that we are trying to predict. However, the training process (finding related word pairs) is computationally expensive. There are also n-gram variants known as skip n-gram LM[19], [5] that tries to skip over some intermediate words in the context, or a variable-length n-gram LM[20] that uses extra context if it is considered to be more predictive.

Based on the n-gram problems that are briefly introduced above, several studies are introduced to build LMs in the continuous parameter space. This may help to overcome the data sparseness, adaptability, and long range dependencies problems of the conventional n-gram LM. In the continuous space, the words are treated as vectors of real numbers rather than of discrete entities. As a result, long-term semantic relationships between the words could be quantified and can be integrated into the model.

Bellegarda et al[21], [22], [23]. introduced latent semantic analysis (LSA) to language modeling. The concept of LSA was first introduced by Deerwester et al[24]. for information retrieval. It maps words into a semantic space where two semantically related words are placed close to each other. Recently, LSA has been successfully used in language modeling to map discrete word into continuous vector space (LSA space). Bellegarda combines the global constraint given by LSA with the local constraint of n-gram language model. The same approach is used in[25], [26], [27], [28], [29], [30] but using neural network (NN) as an estimator. Gaussian mixture model (GMM) could also be trained on this LSA space[15]. Also, the tied-mixture LM (TMLM) is proposed in the LSA space[16]. Context dependent class (CDC) LM using word co-occurrence matrix is proposed in[31]. Instead of a word-document matrix, a word-phrase co-occurrence matrix is used in [32]as a representation of a corpus.

To apply the LSA, the text corpus must be represented by a mathematical entity called matrix. LSA is usually used together with the word-document matrix[33] to represent the corpus. Its cell contains the frequency of how many times a word occurs in a certain document in the corpus. Also, the word-word co-occurrence matrix is used in some previous studies; its cell a_{ij} contains the frequency of word sequence $w_j w_i$ in the corpus.

We introduce a LM in the continuous parameter space based on LSA by proposing a different representation for the text corpus and taking into consideration the long range dependencies between words. We represent the text corpus by creating long-distance word co-occurrence matrices. These matrices represent the co-occurrences between different words on different distances in the corpus. And then applying the LSA to each one of these matrices separately. A tied-mixture HMM model is trained on the LSA results to estimate the LM parameters and word probabilities in the continuous vector space.

2 CONTINUOUS SPACE LANGUAGE MODELING

The concept of language modeling in continuous space is introduced in several studies before as discussed above. The underlying idea of this approach is to attack the data sparseness, adaptability, and long range dependencies problems of the conventional n-gram models by performing the language model probability estimation in a continuous space. In the continuous space, words are not treated as discrete entities but rather vectors of real numbers.

Thus, what we need to build a continuous space LM is: a mapping from the discrete word space to a representation in the continuous parameter space in the form of vectors of real numbers, and then training a statistical parametric model (classifier) which decides the next word given the mapped history in the resulting space.

As a result, long-term semantic relationships between the words could be quantified and can be integrated into the model, where in the continuous space we hope that there is some form of distance of similarity between histories such that histories not observed in the data for some word are smoothed by similar observed histories. This help to attack the data sparseness issue discussed above for n-gram LMs.

By moving to the continuous space, we can cast the language modeling problem as an acoustic modeling problem in speech recognition. In the acoustic modeling, large models can be efficiently adapted using a few utterances by exploiting the inherit structure in the model by techniques like maximum likelihood linear regression (MLLR)[34]. So, we can re-call the acoustic modeling adaptation tools to adapt language models in the continuous space. This addresses the adaptability issue discussed above for n-gram LMs[15], [16].

In this study we propose a method to address the long range dependencies issue discussed above for the n-gram LMs, in addition to the data sparseness and adaptability problems as well.

3 LATENT SEMANTIC ANALYSIS

LSA is a study that mainly aims to reveal the hidden meaning behind the text. It is also able to represent the text in a low-dimension of continuous space.

The concept of LSA was first introduced by Deerwester et al.[24]for information retrieval. Since then there has been an explosion of research and application involving LSA. It was brought to the field of LM for ASR by Bellegarda et al.[21] LSA extracts semantic relations from a corpus, and maps them to a low dimension vector space. The discrete indexed words are projected into LSA space by applying singular value decomposition (SVD) to a matrix that representing a corpus.

The first step is to represent the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column. In the original LSA, the representation matrix is a term-document co-occurrence matrix.

Next, LSA applies singular value decomposition (SVD) to the matrix. In SVD, a rectangular matrix is decomposed into the product of three other matrices. One component matrix describes the original row entities as vectors of derived orthogonal factor values, another describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed. For a matrix C with $M \times N$ dimension, SVD decomposes the matrix Cas follows:

$$C \approx USV^T \tag{2}$$

Where U is a left singular matrix with row vectors and dimension$M \times R$, the matrix U is corresponding with the rows of matrix C. $S$is a diagonal matrix of singular values with dimension $R \times R$. $V$is a right singular matrix with row vectors and dimension $N \times R$, the matrix V is corresponding with the columns of matrix C. $R$is the order of the decomposition and $R \ll \min\,(M, N)$.These LSA matrices are then used to project the words into the reduced R-dimension LSA continuous vector space. In case of a term-document matrix used as a representation matrix, matrix U contains information about words while matrix Vcontains information about the documents. So, the matrix U is used to project words in the LSA space.

Assume we have a vocabulary of size V, each word $\{\mathrm{i}\ 1 \leq \mathrm{i} \leq \mathrm{V}\}$ can be represented by an indicator discrete vector $\mathrm{w_i}$ having one at the $\mathrm{i^{th}}$ position and zero in all other $\mathrm{V} - 1$ positions. This vector can be mapped to a lower dimension Rvector $\mathrm{u_i}$, using a projection matrix A of dimension $\mathrm{V} \times \mathrm{R}$ according to the following equation:

$$u_i = A^T w_i \tag{3}$$

In other words, a continuous vector for word $\mathrm{w_i}$ is represented by the $\mathrm{i^{th}}$ row vector of matrix A. So each word $\mathrm{w_i}$ has a continuous representation vector $\mathrm{u_i}$.

Based on the word mapping in Equation 3, each history h consists of a set of $\mathrm{N} - 1$ words for an n-gram can be represented as a concatenation of the appropriate mapped words. The history vectors are of dimension$\mathrm{R}(\mathrm{N} - 1)$. According to this mapping for word histories h, we can train a statistical parametric model on these histories continuous vectors and build a model to estimate the word probabilities $\mathrm{p}(\mathrm{w}|\mathrm{h})$in the continuous LSA space.

4 Proposed Long-Distance Matrices

LSA starts from representing the corpus through a mathematical entity called a representation matrix. In the original LSA, the used representation matrix is a term-document co-occurrence matrix, where its cell $\mathrm{C}(\mathrm{w_i}, \mathrm{d_j})$ contains co-occurrence frequency of word $\mathrm{w_i}$ in document$\mathrm{d_j}$. In[15], [16], [31], the word-word co-occurrence matrix is used to represent the corpus, where each cell $\mathrm{C}(\mathrm{w_i}, \mathrm{w_j})$ denotes the counts for which word $\mathrm{w_i}$ follow word $\mathrm{w_j}$ in the corpus.

In this study, we propose a representation for the corpus using many word-word co-occurrence matrices, where each matrix will represent the co-occurrence relation between each word and the previous words on different distances in the corpus as we will show below.

The distance-one word co-occurrence matrix is a matrix representation where each row represents a current word$\mathrm{w_i}$, and each column represents the 1^{st} preceding word $\mathrm{w_{i-1}}$ as illustrated by Fig. 1. Each cell $\mathrm{C}(\mathrm{w_i}, \mathrm{w_j})$ is a co-occurrence frequency of word sequence $\mathrm{w_j w_i}$. This is a square matrix with dimension $\mathrm{V} \times \mathrm{V}$, where V is the vocabulary size. It represents the co-occurrence relations between each word and the first preceding words to that word appeared in the corpus.

$$C = \begin{array}{c} \overbrace{\hphantom{c_{11}\ c_{12}\ \cdots\ c_{1j}}}^{1^{st}\ prceding\ word} \hphantom{\cdots\ c_{1v}} \\ \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1v} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2v} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{iv} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ c_{v1} & c_{v2} & \cdots & c_{vj} & \cdots & c_{vv} \end{bmatrix} \end{array} Current\ word$$

Figure 1: Distance-One Word Co-occurrence Matrix

The distance-two word co-occurrence matrix is a matrix representation where each row represents a current word $\mathrm{w_i}$, and each column represents the 2^{nd} preceding word $\mathrm{w_{i-2}}$ as illustrated by Fig. 2. Each cell $\mathrm{C}(\mathrm{w_i}, \mathrm{w_j})$ is a co-occurrence frequency when the word $\mathrm{w_j}$ occurs as the 2^{nd} preceding word of word $\mathrm{w_i}$. This is a square matrix with dimension $\mathrm{V} \times \mathrm{V}$, where V is the vocabulary size. It represents the co-occurrence relations between each word and the 2^{nd} preceding words to that word appeared in the corpus. And the same for distance-three matrix and so on.

$$C = \begin{array}{c} \hphantom{c_{11}\ c_{12}}\overbrace{\hphantom{\cdots\ c_{1j}\ \cdots\ c_{1v}}}^{2^{nd}\ prceding\ word} \\ \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1v} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2v} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{iv} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ c_{v1} & c_{v2} & \cdots & c_{vj} & \cdots & c_{vv} \end{bmatrix} \end{array} Current\ word$$

Figure 2: Distance-Two Word Co-occurrence Matrix

By using these long-distance co-occurrence representation matrices, we hope to collect more information about each word and its relation with the previous words in the corpus on different distances; this may help to attack the long-dependencies problem of the conventional n-gram model. These matrices are large, but very sparse ones. Because of their large size and sparsity, we can apply the second step of LSA, by making SVD to each one of them separately to produce a reduced-rank approximation to each matrix of them.

Before proceeding in the SVD step, the entries of the co-occurrence matrices are smoothed according to Equation 4, this because the co-occurrence matrices typically contain a small number of high frequency events and a large number of less frequent events, and the SVD derives a compact approximation of the co-occurrence matrix that is optimal in the least square sense, it best models these high frequency events, which may not be the most informative[16].

$$\hat{C}(w_i, w_j) = \log\left(C(w_i, w_j) + 1\right) \tag{4}$$

Based on the SVD results from Equation 2, we construct a projection matrix A of dimension $\mathrm{V} \times \mathrm{R}$ corresponding to each word co-occurrence matrix by using the left singular matrix U, and the diagonal matrix S of the SVD results as follows:

$$A_{V\times R} = U_{V\times R} S_{R\times R} \tag{5}$$

Now, we have a projection matrix A for each constructed long-distance word co-occurrence matrix. For example, if we create distance-one, distance-two, and distance-three word co-occurrences matrices then after the SVD for each one of them, we will have three projection related matrices. We can map the discrete word vector $\mathrm{w_i}$ $\{\mathrm{i}\ 1 \leq \mathrm{i} \leq \mathrm{V}\}$ into the continuous space using Equation 3 to get a word vector $\mathrm{u_i}$ in the continuous space for each word from each projection matrix A.

As a result from the previous step, we have more than one mapped continuous vector $\mathrm{u_i}$ for each word $\mathrm{w_i}$, then we concatenate these continuous vectors of each word to construct the final word vector that uniquely represent the word in the LSA continuous vector space. The final word vector will contain information about the relation of the word with the previous words appeared on different distances in the corpus. In the next section, we introduce the tied-mixture model used to estimate the word probabilities using these word vectors in the continuous space.

5 EXPERIMENTS

Two primary metric used to evaluate the language model is perplexity (PP) on test data as defined by the following equation:

$$PP = 2^{-\frac{1}{N}\log_2 P_L(W)} \tag{6}$$

Minimizing perplexity means maximizing the log-likelihood function. Although perplexity is not always agree with word-error rate[37], but it is the first approximation towards better language model. It tells how many word choices during the recognition process. Small number of word choices will make speech recognition system easier to choose the correct word.

As a baseline, a statistical bigram language model using Modified Kneser-Ney smoothing has been built using SRILM toolkit[38], which is referred to as Word-2gr. -The language model data has about 85K sentences comprising about 2M words. The vocabulary size is 91K words. First, we limit the construction of TMMs for words that occur 100 times or more, so we ended up in 2800 words including the beginning sentence and end sentence symbols. We mapped all the remaining words into one class, a sort of filter or unknown word, so the vocabulary size become 2801 words from the original 91K vocabulary.

The normal tied-mixture language model (TMLM) is trained by constructing the word co-occurrence matrix of dimensions 2801×2801. Each element $\mathrm{C}(\mathrm{w_i}, \mathrm{w_j})$ in the co-occurrence matrix is smoothed using Equation4. Singular value decomposition (SVD) is performed on the resulting smoothed co-occurrence matrix. The SVDLIBC toolkit[40] with Lanczos method is used to compute the 50 (R=50) highest singular values and their corresponding singular vectors for the smoothed co-occurrence matrix. The resulting singular vectors are used to construct the projection to a 50-dimensional space. Each word in the vocabulary is represented by a vector of size 50. In another words, each bigram history is represented by a vector of size 50 representing that word. Then a TM-HMM is built and trained.

The proposed long-distance tied-mixture model (LD-TMM) with max distance (D=5) is trained as follows: we construct distance-one, distance-two, distance-three, distance-four, and distance-five word co-occurrence matrices, these are sparse matrices each of dimensions 2801×2801. Each element $C(w_i, w_j)$ in each co-occurrence matrix is smoothed using Equation 4. Singular value decomposition (SVD) is performed on each of the resulting smoothed co-occurrence matrices. The SVDLIBC toolkit with Lanczos method is used to compute the 10 (R=10) highest singular values and their corresponding singular vectors for each smoothed co-occurrence matrix. The resulting singular vectors of each matrix are used to construct the projection to a 10-dimensional space. Each word in the vocabulary is represented by a vector of size 50, this by concatenating the five word vectors of size 10 resulting from the SVD step for the five co-occurrence matrices. In another words, each bigram history is represented by a vector of size 50 representing that word. Thus, a document can be represented by a sequence of 50-dimensional vectors corresponding to the history of each of its constituent words. Then a TM-HMM is built and trained.

For the TMLM and LD-TMLM, we use the HTK toolkit[39] for building and training the TMM-HMM model, and the total number of the shared Gaussian densities (Gaussians pool) used is set to 200. Also, when calculating the TMM score, the TMM likelihood probability generated by the model is divided by 40 to balance its dynamic range with that of the n-gram model.

Table 1 shows the log probabilities and perplexity results for: the baseline word bigram (Word-2gr) with Modified Kneser-Ney smoothing, the normal TMLM, the proposed long-distance trigram TMM (LD-TMM) with max distance

(D=5), and an interpolation between the baseline bigram and the proposed LD-TMM. The interpolated LM uses uniform weights. These results for the same 53 reference sentences (3677 words) used in the rescoring results before.

TABLE 2
LANGUAGE MODELS PERPLEXITY RESULTS

Language Model (LM)	**Log prob.**	**Perplexity(PP)**
Word-2gr	-6264.47	47.80
TMLM	-1474.40	2.48
LD-TMLM (D=5)	-1026.57	1.88
Word-2gr + LD-TMM (D=5)	-3645.52	9.50

The first two rows in the table show the perplexity of the baseline bigram (Word-2gr) model and the perplexity of the tied-mixture continuous language model (TMLM), where the perplexity of the baseline Word-2gr model is 47.80 and the perplexity of the TMLM is 2.48. The continuous TMLM shows improvement in the perplexity results over the baseline Word-2gr model. The third row in the table shows that the proposed LD-TMLM (D=5) improves the perplexity to 1.88. An interpolation between the baseline bigram (Word-2gr) and the proposed LD-TMM (D=5) using uniform weights improves the perplexity results to 9.50 compared to the perplexity results of the baseline bigram (Word-2gr) model.

6 CONCLUSION

In this paper, we first point out the problems and drawbacks of the widely used n-gram language model. We have proposed a different representation for the text corpus, this by constructing more than one word-co-occurrence matrix that cover long distance dependencies between different words in the corpus. Also, we introduced a continuous space language model based on LSA using these matrices. We used the tied-mixture HMM modeling to robustly estimate model parameters. The proposed corpus representation may help to address the n-gram drawbacks in the continuous vector space. Our initial experimental results validated the proposed approach with encouraging results compared to the traditional n-gram LM.

REFERENCES

[1] A. A. Markov, “An example of statistical investigation in the textof ‘Eugene Onyegin’ illustrating coupling of ‘tests’ in chains,” inProceedings of the Academy of Sciences, vol. 7 of VI, St. Petersburg,1913, pp. 153–162.

[2] F. Damerau, Markov models and linguistic theory: an experimental study of a model for English, ser. Janualinguarum: Series minor. Mouton, 1971.

[3] F. Jelinek, Statistical Methods for Speech Recognition, ser. Language, Speech, & Communication: A Bradford Book. MIT Press, 1997.

[4] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in Acoustics, Speech, and Signal Processing, 1995.ICASSP-95., 1995 International Conference on, vol. 1, 1995, pp. 181–184 vol.1.

[5] H. Ney, U. Essen, and R. Kneser, “On structuring probabilistic dependencies in stochastic language modelling,” Computer Speech and Language, vol. 8, pp. 1–38, 1994.

[6] I. J. Good, “The population frequencies of species and the estimation of population parameters,” Biometrika, vol. 40(3 and 4), pp. 237–264, 1953.

[7] F. Jelinek and R. L. Mercer, “Interpolated estimation of markov source parameters from sparse data,” in In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May 1980, pp. 381–397.

[8] S. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 35, no. 3, pp. 400–401, 1987.

[9] G. Lidstone, “Note on the general case of the Bayes–Laplace formula for inductive or a posteriori probabilities.” Transactions of the Faculty of Actuaries, vol. 8, pp. 182–192, 1920.

[10] T. C. Bell, J. G. Cleary, and I. H. Witten, Text Compression. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1990.

[11] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, “Class-based n-gram models of natural language,” Comput. Linguist., vol. 18, no. 4, pp. 467–479, Dec. 1992.

[12] S. Broman and M. Kurimo, “Methods for combining language models in speech recognition,” Interspeech, pp. 1317–1320, September 2005.

[13] Y. Wada, N. Kobayashi, and T. Kobayashi, “Robust language modeling for a small corpus of target tasks using class-combined word statistics and selective use of a general corpus,” Systems and Computers in Japan, vol. 34, no. 12, pp. 92–102, 2003.

[14] T. Niesler and P. Woodland, “Combination of word-based and categorybased language models,” in Spoken Language, 1996.ICSLP 96.Proceedings., Fourth International Conference on, vol. 1, 1996, pp. 220–223 vol.1.

[15] M. Afify, O. Siohan, and R. Sarikaya, “Gaussian mixture language models for speech recognition,” in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.IEEE International Conference on, vol. 4, 2007, pp. IV–29–IV–32.

[16] R. Sarikaya, M. Afify, and B. Kingsbury, “Tied-mixture language modeling in continuous space,” in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, ser. NAACL’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 459–467.

[17] R. Kuhn and R. De Mori, “A cache-based natural language model for speech recognition,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 12, no. 6, pp. 570–583, 1990.

[18] R. Rosenfeld, “A maximum entropy approach to adaptive statistical language modeling,” Computer Speech and Language, vol. 10, no. 3, pp. 187 – 228, 1996.

[19] S. Nakagawa, I. Murase, M. Zhou, . L, and ., “Comparison of language models by stochastic context-free grammar, bigram and quasi-simplifiedtrigram,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2008, 0300-1067.

[20] T. Niesler and P. Woodland, “A variable-length category-based ngram language model,” in Acoustics, Speech, and Signal Processing, 1996.ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 1, 1996, pp. 164–167 vol. 1.

[21] J. Bellegarda, J. Butzberger, Y.-L.Chow, N. Coccaro, and D. Naik, “A novel word clustering algorithm based on latent semantic analysis,” in Acoustics, Speech, and Signal Processing, 1996.ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 1, 1996, pp. 172–175 vol. 1.

[22] J. Bellegarda, “A multispan language modeling framework for large vocabulary speech recognition,” Speech and Audio Processing, IEEE Transactions on, vol. 6, no. 5, pp. 456–467, 1998.

[23] ——, “Latent semantic mapping [information retrieval],” Signal Processing Magazine, IEEE, vol. 22, no. 5, pp. 70–80, 2005.

[24] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, vol. 41, no. 6, pp. 391–407, 1990.

[25] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” JOURNAL OF MACHINE LEARNING RESEARCH, vol. 3, pp. 1137–1155, 2003.

[26] F. Blat, M. Castro, S. Tortajada, and J. Snchez, “A hybrid approach to statistical language modeling with multilayer perceptrons and unigrams,” in Text, Speech and Dialogue, ser. Lecture Notes in Computer Science, V. Matouek, P. Mautner, and T. Pavelka, Eds., vol. 3658. Springer Berlin Heidelberg, 2005, pp. 195–202.

[27] A. Emami, P. Xu, and F. Jelinek, “Using a connectionist model in a syntactical based language model,” in Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03). 2003 IEEE International Conference on, vol. 1, 2003, pp.I–372–I–375 vol.1.

[28] H. Schwenk and J. Gauvain, “Connectionist language modeling for large vocabulary continuous speech recognition,” in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, vol. 1, 2002, pp.I–765–I–768.

[29] H. Schwenk and J.-L.Gauvain, “Neural network language models for conversational speech recognition,” in ICSLP, 2004.

[30] ——, “Building continuous space language models for transcribing european languages.” in INTERSPEECH. ISCA, 2005, pp. 737–740.

[31] W. Naptali, M. Tsuchiya, and S. Nakagawa, “Language model based on word order sensitive matrix representation in latent semantic analysis for speech recognition,” in Computer Science and Information Engineering, 2009 WRI World Congress on, vol. 7, 2009, pp. 252–256.

[32] Fumitada, “A linear space representation of language probability through SVD of n-gram matrix,” Electronics and Communications in Japan (Part III: Fundamental Electronic Science), vol. 86, no. 8, pp. 61–70, 2003.

[33] T. Rishel, A. L. Perkins, S. Yenduri, F. Zand, and S. S. Iyengar, “Augmentation of a term/document matrix with part-of-speech tags to improve accuracy of latent semantic analysis,” in Proceedings of the 5th WSEAS International Conference on Applied Computer Science, ser. ACOS’06. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2006, pp. 573–578.

[34] C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” Computer Speech and Language, vol. 9, no. 2, pp. 171 – 185, 1995.

[35] J. Bellegarda and D. Nahamoo, “Tied mixture continuous parameter modeling for speech recognition,” Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 38, no. 12, pp. 2033–2045, 1990.

[36] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains,” The Annals of Mathematical Statistics, vol. 41, no. 1, pp. 164–171, 1970.

[37] D. Klakow and J. Peters, “Testing the correlation of word error rate and perplexity,” Speech Commun., vol. 38, no. 1, pp. 19–28, Sep. 2002.

[38] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in Proceedings of ICSLP, vol. 2, Denver, USA, 2002, pp. 901–904.

[39] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, “The HTK book, version 3.4,” in Cambridge University Engineering Department, Cambridge, UK, 2006.

[40] The SVDLIBCToolkit Web Sitehttp://tedlab.mit.edu/_dr/SVDLIBC/

BIOGRAPHY

Dr. Sherif Abdou received his B.Sc. and M.Sc. degrees in computer science and automatic control from University of Alexandria, Egypt in 1993 and 1997, respectively. He received a Ph.D degree in Electrical and Computer Engineering from University of Miami, USA in 2003. In 2003 Dr. Abdou joined BBN Technologies as a senior staff scientist in the Arabic language team. Currently DrAbdou is Associate Professor at the Information Technology department at the Faculty of Computers and Information. Abdou is a member of the review committee in several conferences andjournals in the HLT fieldsand is the Principal Investigator and Co-Principal Investigator of several research projects in the areas of Language learning, Virtual tutors, Web monitoring and Intelligent Contact Centers.

Mahmoud A. Ismael, Associate Prof. Faculty of Computers and Information, Cairo University, Egypt. PhD in Computers and Systems Engineering, Ain Shams University 1998. MSC in Electronics and Communication Engineering, Faculty of Engineering, Cairo University, 1994. BSC in Electronics and Communication Engineering, Faculty of Engineering, Cairo University, 1990. He is currently in the Information Technology department of the faculty of Computers and Information, Cairo University. His research interests include: Speech recognition, pattern classification, signal processing and artificial intelligence.

Mohamed Talaat, Teaching Assistant at Faculty of Computers and Information, Cairo University, Egypt. B.Sc in Information Technology, Faculty of Computers and Information, Cairo University, in 2008. His research interests include: Speech recognition, pattern classification, and language modeling.He joined Fawry Integrated Systems as a software advisory architect.

بناء نماذج لغوية فى الفضاء المتصل

محمد طلعت، شريف عبده، محمود شومان
كلية الحاسبات و المعلومات- جامعة القاهرة

ملخص

فى هذا البحث نقدم طريقة لبناء نماذج لغوية بعيدة المدى فى الفضاء المتصل. وهذا العمل يعد تطوير للنماذج اللغوية فى الفضاء المتصل التى أثبتت كفائتها بالمقارنة بنماذج ال ngram الشهيرة التى تواجه صعوبة حساب العدد الضخم من المتغيرات. فى هذا البحث نقدم تطوير لهذه النماذج اللغوية مع توفير القدرة على أدخال العلاقات بين الكلمات بعيدة المدى. ولقد اثبتت النتائج قدرة هذه النماذج المطورة على تحسين مقياس ال perplexity للنصوص المختبرة.

أقسام الكلم العربي

نحو تقييم المحللات الصرفية العربية في ضوء منهج تمام حسان

عمرو حمدي الجندي

مجمع اللغة العربية

فريق حاسوبويه

Amr25@hotmail.com

ملخص

تسعى هذه الورقة إلى محاولة تقييم تحليل المحللات الصرفية العربية لأقسام الكلم في ضوء منهج العالم اللغوي تمام حسان؛ وذلك من خلال تقديم بعض الحقائق التي تراها أساسية في مجال تقنيات اللغة، ثم توضيح الدعائم التي قام عليها النظام الصرفي العربي، مبينة أهمية تقسيم الكلم وحاجة تقنيات اللغة إلى معيارية هذا التقسيم، متناولة منهج النحاة العرب في تقسيم الكلم وعلى الأخص منهج تمام حسان، ومنهج عدد من المحللات الصرفية العربية في تقسيم الكلم موضحة أهم ما يميز منهجها وما يعيبه، وأخيرا تقترح الورقة عددا من التوصيات في هذا السياق.

الكلمات المفتاحية

المحللات الصرفية – أقسام الكلم – النظام الصرفي العربي.

1. مقدمة

هناك عدة حقائق أود التأكيد عليها في البداية؛ أولها: أن لكل لغة خصائصها وقواعدها المتميزة، وثانيها: أن معالجة اللغة يعني – ضمن ما يعني – نقل هذه الخصائص وتلك القواعد إلى الحاسوب بهدف إنتاج التقنية، وثالثها: أن وسائل الذكاء الاصطناعي لا تستطيع وحدها إنتاج تقنيات عالية الدقة (تحاكي الإنسان) دون تعلم جميع هذه الخصائص وتلك القواعد والتقيد بها.

إذًا ومن خلال ما سبق يمكن القول إن فهم طبيعة اللغة العربية وما انبنى عليه نظامها الصرفي، مسألة استراتيجية عند بناء ما يعرف بالمحلل الصرفي العربي. وإذا أخذنا في الاعتبار أن المحلل الصرفي هو أحد البنى التحتية لمنظومة تقنيات اللغة؛ فلا مفر أمام طرق الذكاء الاصطناعي من تعلم هذا النظام برُمّته.

إن ما يعرف بأقسام الكلم هي إحدى أركان هذا النظام الصرفي العربي، ومن ثم لزم التعرف عليه من خلال علمائه القدامى والمحدثين، ومن خلال المحاولات الجادة التي حاولت وصف هذا النظام واستكشاف قواعده التي تحكمه، كما لزم رصد مدى تقيد المحللات الصرفية العربية بهذا النظام، خاصة وأن الدرس الصرفي هو أحد الدروس التي تدخل ضمن عدد كبير من تطبيقات التعلم الإلكتروني، فضلا عن حاجة الحاسوب نفسه إلى تعلم هذا النظام كما هو.

2. النظام الصرفي العربي

يرى الدكتور تمام حسان أن النظام الصرفي العربي مبني على ثلاث دعائم هامة:[1]

(أ) المعاني الصرفية:

- من حيث تقسيم الكلم: كالاسمية والفعلية.
- من حيث تقسيم الصيغ: كالإفراد والتعريف.

(ب) المباني الصرفية: وهي الصيغ المجردة واللواصق والزوائد ومباني الأدوات والحذف والاستتار، وهي التي تدل على المعاني الصرفية، وتتحقق هذه المباني عن طريق العلامات التي تندرج تحتها، فـ(زيد) تدل على صيغة الاسم، و(ضرب) تدل على صيغة الفعل، و(الـ) تدل على ال المعرفة، وهكذا. وتنقسم هذه المباني إلى نوعين:

- مباني التقسيم: وهي أقسام الكلم وما يندرج تحتها من صيغ صرفية للأسماء والأفعال والصفات، وصور للضمائر والظروف والأدوات والخوالف.
- مباني التصريف: وهي المقابلات الموجودة بين مباني التقسيم؛ فتسند الأفعال إسنادات مختلفة بحسب التكلم والخطاب والغيبة وبحسب الإفراد والتثنية والجمع وبحسب التذكير والتأنيث، وتتصرف الأسماء تصريفات مختلفة باختلاف الإفراد والتثنية والجمع، والتذكير والتأنيث، والتعريف والتنكير، وهكذا بقية مباني التقسيم.

(ج) العلاقات العضوية الإيجابية والقيم الخلافية: وهي وجوه الارتباط والاختلاف بين هذه المباني.

3. أهمية تقسيم الكلم العربي

لا يمكن اعتبار تقسيم الكلم العربي نوعا من فضول العلم الذي ليس بالضرورة الاهتمام به فضلا عن التقيد به، وإنما هو ضرورة ملحة في فهم العربية سواء على المستوى اللغوي أو التقني.

3.1 على المستوى اللغوي

لقد عدّ الدكتور تمام حسان ستة أسباب تجعل تقسيم الكلم أمرا هاما وضروريا، (As-Saqy, F., 1977) أصوغها على النحو التالي:

1) توقف جزء من المعنى النحوي على البنية الصرفية؛ كتوقف الفاعل والمفعول على كون البنية اسما، وتوقف الحال والتمييز على الاشتقاق والجمود، وتوقف المفعول المطلق على بنية مصدر من مادة الفعل، بينما المفعول لأجله على بنية مصدر من غير مادة الفعل، ...إلخ.
2) عدم إمكانية استخراج الكثير من المعاني الصرفية الهامة إلا بتقسيم الكلم؛ كالمسمى، والموصوف بالحدث، واجتماع الحدث والزمن، ...إلخ.

[1] اعتمدنا في هذه الدراسة على منهج الدكتور تمام حسان

3) أمن اللبس في فهم أقسام الكلم التي تنقل إلى استعمال أقسام أخرى؛ كنقل الفعل والوصف إلى العلمية، والاسم إلى الظرفية، ...إلخ.
4) اتضاح ظاهرة "تعدد المعنى الوظيفي للمبنى الواحد"، كصيغة فعيل، وهنا، وإذا، وما، ...إلخ.
5) تحديد المعرب والمبني من الكلم، ومن ثم الانتفاع بقرينة الإعراب في الكشف عن المعنى.
6) التعرف على المباني التي تحدد القرائن اللفظية المركبة في بعض أنماط الجمل، كقرائن الربط والتضام والرتبة.

3.2 على المستوى التقني

تقنيات اللغة هي إحدى فروع الذكاء الاصطناعي الذي يُعنى بمحاكاة الإنسان في أنشطته ومساعدته في حياته، وبالتالي فهدفها هو محاكاة الإنسان في تعامله مع اللغة، وهذا يستلزم النظر إلى كيفية استخدام الإنسان للغة وتواصله بها وتعلمه إياها.

وبعبارة مختصرة فإنه مما لا شك فيه أن معظم الناس لا يفكرون كثيرا عند تعاملهم مع اللغة في أقسام الكلم – وإن كانوا يطبقون قواعدها – بل يهمهم في نهاية المطاف الفهم والتواصل، أما متعلم اللغة، إنسانا كان أم آلة، فهو يهتم بأقسام الكلم، ومن ثم فإن الداعي إلى التقيد بأقسام الكلم أهم من الداعي إلى إهمال هذا الأمر.

4. أقسام الكلم الرئيسة عند النحاة العرب

4.1 مرحلة ما قبل تمام حسان

منذ الخليل بن أحمد الفراهيدي وحتى ما قبل ظهور كتاب "اللغة العربية معناها ومبناها" عكف النحاة العرب على تقسيم الكلمة العربية إلى أقسام ثلاثة (اسم، فعل، حرف)؛ ولكن تخلل هذه الفترة الطويلة بعض العلماء الذين أضافوا قسما رابعا على اختلاف بينهم:

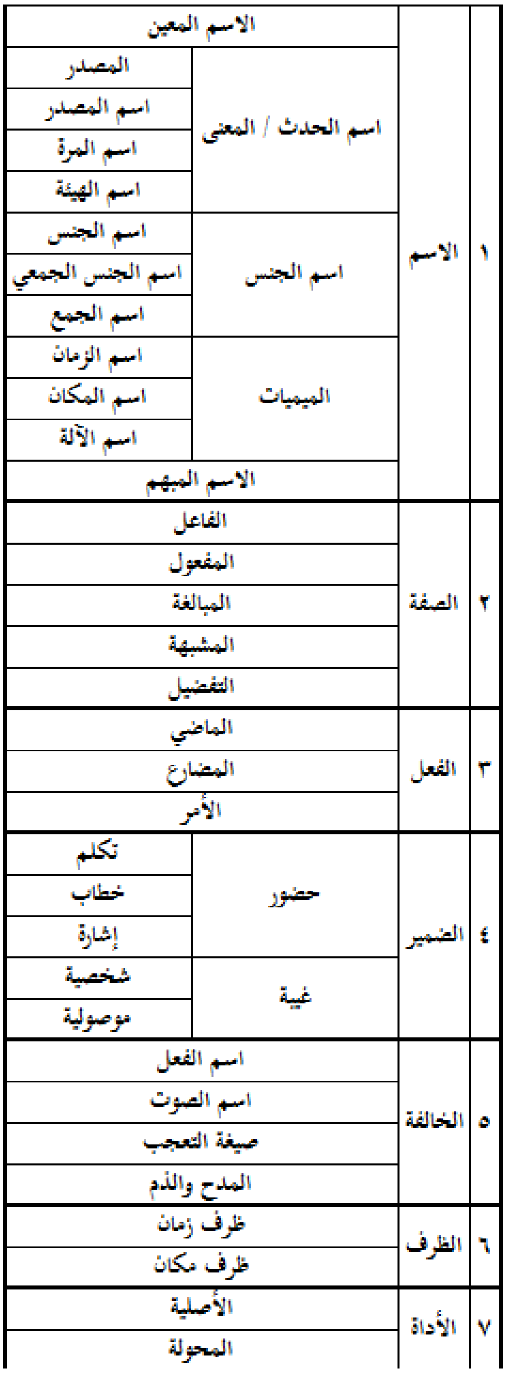
- بعض القدماء: (اسم، فعل، حرف، اسم الفعل).
- د.إبراهيم أنيس: (اسم، فعل، ضمير، أداة).
- د.مهدي المخزومي: (اسم، فعل، كنايات، أداة).

4.2 مرحلة تمام حسان

اعتمد د. تمام حسان على المنهج الوصفي في التقسيم الدقيق للكلم العربي، مستخدما أساسي المعنى والمبنى معا؛ أما أساس المعنى فيتضمن كلا من: (التسمية، الحدث، الزمن، التعليق، المعنى الجملي)، وأما أساس المبنى فيتضمن كلا من: (الصورة الإعرابية، الرتبة، الصيغة، الجدول، الإلصاق، التضام، الرسم الإملائي).

لقد اكتفى د. تمام حسان باختلاف واحد في المعنى والمبنى معا، كي يجعل الكلم العربي سبعة أقسام هي: (اسم، صفة، فعل، ضمير، أداة، ظرف، خالفة)، ثم جعل لكل قسم أقساما فرعية، وتحت كل قسم فرعي أقساما أخرى.

جدول 1: أقسام الكلم العربي عند تمام حسان



١	الاسم	الاسم المعين	
		اسم الحدث / المعنى	المصدر
			اسم المصدر
			اسم المرة
			اسم الهيئة
		اسم الجنس	اسم الجنس
			اسم الجنس الجمعي
			اسم الجمع
		الميميات	اسم الزمان
			اسم المكان
			اسم الآلة
		الاسم المبهم	
٢	الصفة	الفاعل	
		المفعول	
		المبالغة	
		المشبهة	
		التفضيل	
٣	الفعل	الماضي	
		المضارع	
		الأمر	
٤	الضمير	حضور	تكلم
			خطاب
			إشارة
		غيبة	شخصية
			موصولية
٥	الخالفة	اسم الفعل	
		اسم الصوت	
		صيغة التعجب	
		المدح والذم	
٦	الظرف	ظرف زمان	
		ظرف مكان	
٧	الأداة	الأصلية	
		المحولة	

5. أقسام الكلم الرئيسة في المحللات الصرفية العربية

5.1 المحللات الصرفية العربية

ثمة أداتان إحداهما تقوم بالتحليل الصرفي وتسمى "المحلل الصرفي"، والثانية تقوم بتعيين أقسام الكلم وتسمى "معيّن أقسام الكلم"، ويعتبر البحث أن كلتا الأداتين تنتميان إلى شيء واحد هو "التحليل الصرفي"، على اعتبار أن التحليل الصرفي للكلمة يتضمن أيضًا تحديد القسم الكلامي الذي تنتمي إليه.

وعلى كلٍّ فإن المحللات الصرفية هي إحدى البنى التحتية في عملية معالجة اللغات الطبيعية بشكل عام، واللغة العربية بشكل خاص، ويرجع ذلك إلى أن معالجة النص تعتمد بشكل رئيس على عملية التحليل الصرفي، بالإضافة إلى كون العربية – كما هو معلوم – لغة الاشتقاق والتصريف.

وظيفة المحللات الصرفية هي القيام بتحديد العناصر الآتية للكلمة: (الجذر، الوزن، الجذع، السوابق، اللواحق، مباني التقسيم، مباني التصريف).

ومن هنا كانت المحللات الصرفية العربية مهمة في كل تطبيق أو تقنية أو معالجة تتصل بالنص العربي من قريب أو بعيد، مثل: تطبيقات الوسائط المتعددة، تدقيق النص العربي من حيث الضبط والإملاء، صناعة المعاجم الإلكترونية، البحث النصي الذكي، دعم كل من أنظمة تحويل النص العربي إلى كلام منطوق وأنظمة التعرف البصري على الحرف العربي، استخراج الطابع الصرفي للكتّاب العرب، دعم المستويات العليا للتحليل اللغوي العربي. (Attia, M., 2000)

وعلى كل فسوف نقدم نبذات مختصرة عن المحللات الصرفية (صخر، الميزان، باك ولتر، الخليل) التي تم استخدامها في إجراء التجارب على مدونة الاختبار، والتي بلغت حوالي 7500 كلمة من العربية المعاصرة في مجالات: الفن والاقتصاد والأدب والسياسة والعلوم والاجتماع والرياضة، كما نطلع في الفقرات اللاحقة على أقسام الكلم العربي فيها.

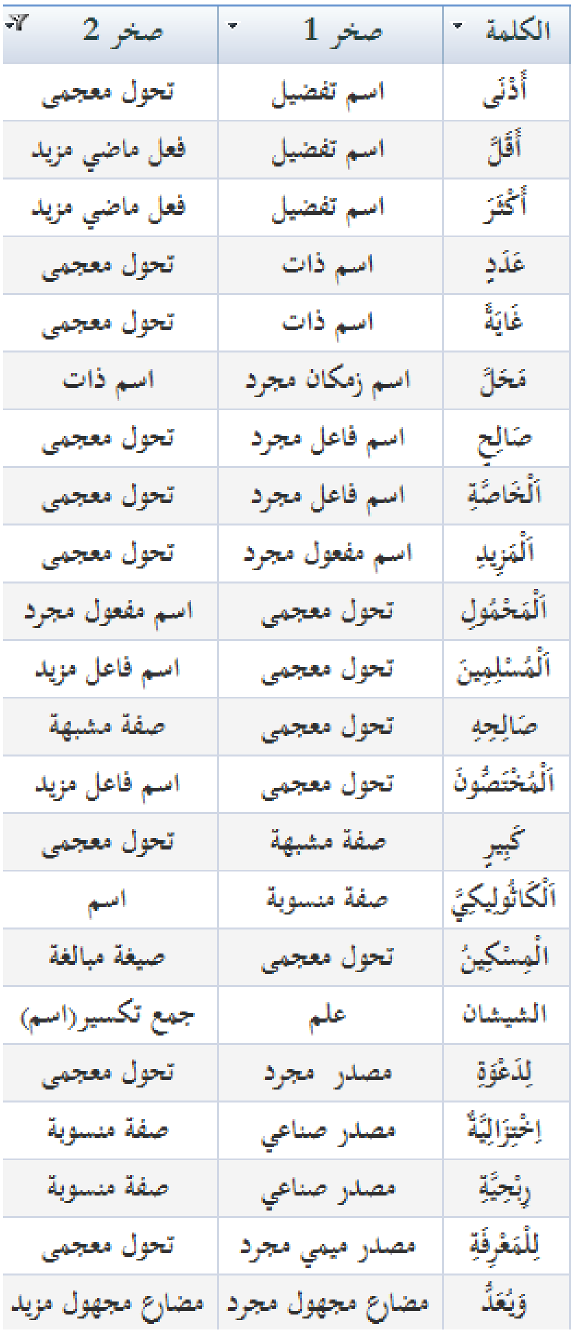
5.1.1 صخر

يقوم المحلل الصرفي التابع لشركة صخر بإنتاج العديد من البيانات اللغوية، على نحو مما يأتي:

1. بيانات صرفية: وتتضمن قسم الكلم، الميزان الصرفي، الساق، الجذر ونوعه، السوابق، اللواحق، العدد ونوعه، الجنس، التعريف والتنكير.
2. بيانات نحوية: وتتضمن قابلية الحالة الإعرابية ونوعها، قابلية التنوين ونوعه، الضمير المتصل ونوعه، نوع الفاعل.
3. بيانات معجمية: وتتضمن المعنى بالعربية، المعنى بالإنجليزية، السمات الدلالية، كون الكلمة تشير إلى عاقل أو غير عاقل.

يقوم صخر أحيانا بإعطاء أكثر من قسم كلامي لبعض الكلمات التي تحتمل ذلك، مثل: صيغة أَفْعَلَ، الكلمات المحولة دلاليا، المشترك اللفظي، صيغة المبني للمجهول للفعل الثلاثي المضعف المجرد أو المزيد، ...إلخ.

جدول 2: تعدد الأقسام الكلامية لبعض الكلمات في محلل صخر الصرفي



الكلمة	صخر 1	صخر 2
أَدْنَى	اسم تفضيل	تحول معجمى
أَقَلّ	اسم تفضيل	فعل ماضي مزيد
أَكْثَرَ	اسم تفضيل	فعل ماضي مزيد
عَدَدٍ	اسم ذات	تحول معجمى
غَايَةٌ	اسم ذات	تحول معجمى
مَحَلّ	اسم زمكان مجرد	اسم ذات
صَالِحٍ	اسم فاعل مجرد	تحول معجمى
اَلْخَاصَّةِ	اسم فاعل مجرد	تحول معجمى
اَلْمَزِيدِ	اسم مفعول مجرد	تحول معجمى
اَلْمَحْمُولِ	تحول معجمى	اسم مفعول مجرد
اَلْمُسْلِمِينَ	تحول معجمى	اسم فاعل مزيد
صَالِحِهِ	تحول معجمى	صفة مشبهة
اَلْمُخْتَصُّونَ	تحول معجمى	اسم فاعل مزيد
كَبِيرٍ	صفة مشبهة	تحول معجمى
اَلْكَاثُولِيكِيِّ	صفة منسوبة	اسم
اَلْمِسْكِينُ	تحول معجمى	صيغة مبالغة
الشيشان	علم	جمع تكسير(اسم)
لِدَعْوَةِ	مصدر مجرد	تحول معجمى
اِخْتِزَالِيَّةٌ	مصدر صناعي	صفة منسوبة
رِبْحِيَّةٍ	مصدر صناعي	صفة منسوبة
لِلْمَعْرِفَةِ	مصدر ميمي مجرد	تحول معجمى
وَيُعَدُّ	مضارع مجهول مجرد	مضارع مجهول مزيد

بلغ متوسط الحلول الصرفية التي قدّمها محلل صخر لكل كلمة، ما قيمته 1.25 حل صرفي / كلمة، وبلغت نسبة الكلمات التي لم يستطع إيجاد حل لها 1.98%

5.1.2 الميزان

محلل الميزان هو أحد أدوات معالجة اللغة في الشركة الهندسية لتطوير النظم الرقمية RDI، التي تقول إن نسبة تغطيته تتجاوز 99.5% معتمدة في ذلك على 7,500 مورفيم.

ويفرق الميزان بين نوعين من التحليل الصرفي، على النحو التالي:

1. التحليل الصرفي: ويعرض فيه 5 معلومات عن الكلمة (النوع، السابق، الجذر، الوزن، اللاحق)، ويقسم نوع الكلمة إلى أنواع 5 هي: (مصرفة منتظمة، جامدة، مستثناة صرفيا، معربة، غير عربية).
2. أقسام الكلم: وتقدم فيه بقية المعلومات الصرفية المتعلقة بكل من مباني التقسيم ومباني التصريف.

لدى الميزان إمكانية عرض جميع الحلول الممكنة للكلمة أو اختيار حل واحد فقط من الحلول المتاحة، ولذلك بلغ متوسط الحلول الصرفية التي قدّمها الميزان لكل كلمة، ما قيمته 1.00 حل صرفي / كلمة، وبلغت نسبة الكلمات غير المحللة 4.29%.

5.1.3 باكولتر

محلل صرفي إنجليزي للغة العربية، يقوم بعرض جميع الصور المختلفة للكلمة من ناحية الضبط، عن طريق الكتابة الصوتية، فمثلا كلمة "بَابُ" يعرض جميع احتمالاتها على النحو التالي:

جدول 3: التحليل الصرفي لكلمة (باب) لدى باكولتر

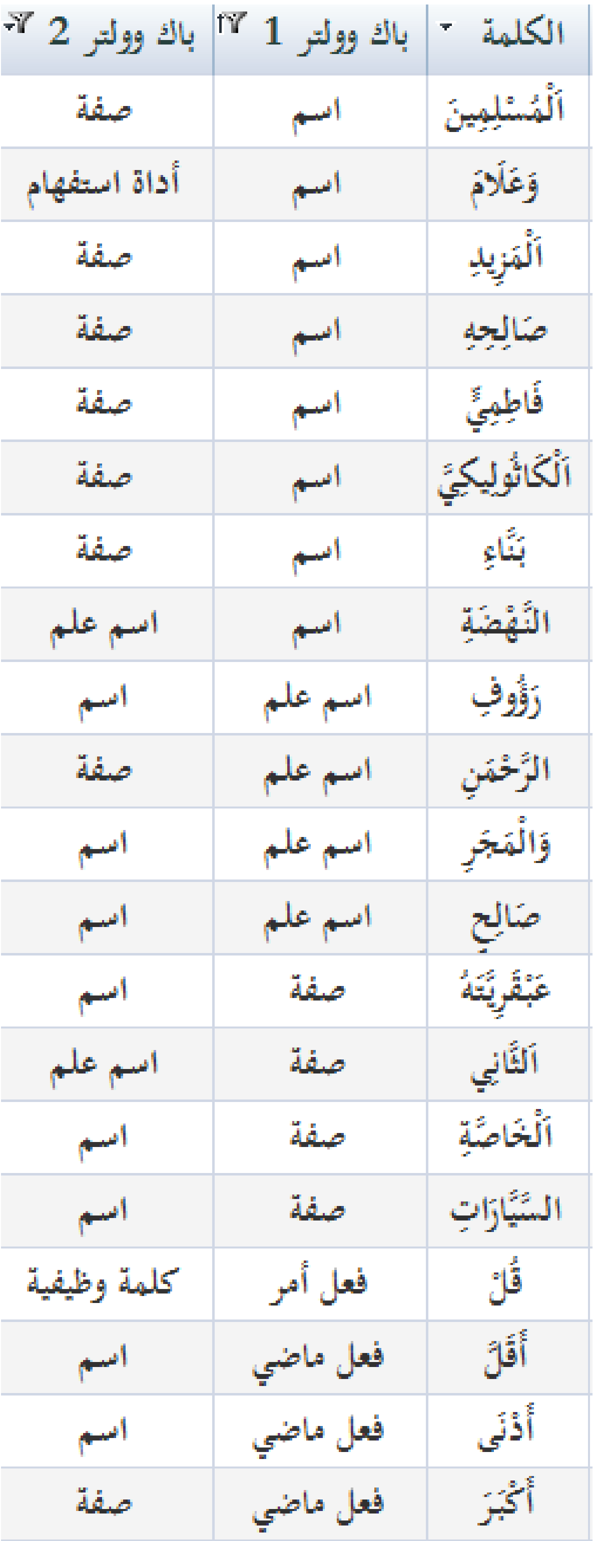
الكلمة	bAb	bAb	bi|b	bi>ab
الجذع	bAb_1	bAb_2	|b_1	<ab_1
	bAb	bAb		
السابق			bi	bi
النوع	NOUN	NOUN	PREP	PREP
المعنى 1	Door	category	by	by
المعنى 2	Gate	rubric	with	with
			|b	<ab
المعنى 1			NOUN_PROP	NOUN
المعنى 2			August	father

كما يقدم المعلومات الآتية:

1. جذع الكلمة وقسمها ومعانيها.
2. السوابق واللواحق ونوعها ومعانيها.

بلغ متوسط الحلول الصرفية التي قدّمها باكولتر لكل كلمة، ما قيمته 2.53 حل صرفي / كلمة، أما الكلمات التي لم يستطع باكولتر إيجاد حلول لها فلم تتجاوز 1.19%

جدول 4: تعدد الأقسام الكلامية لبعض الكلمات في باكولتر



الكلمة	باك وولتر 1	باك وولتر 2
اَلْمُسْلِمِينَ	اسم	صفة
وَعَلَامَ	اسم	أداة استفهام
اَلْمَزِيدِ	اسم	صفة
صَالِحِهِ	اسم	صفة
فَاطِمِيٍّ	اسم	صفة
اَلْكَاثُولِيكِيِّ	اسم	صفة
بِنَاءِ	اسم	صفة
النَّهْضَةِ	اسم	اسم علم
رَؤُوفِ	اسم علم	اسم
الرَّحْمَنِ	اسم علم	صفة
وَالْمَجْرِ	اسم علم	اسم
صَالِح	اسم علم	اسم
عَبْقَرِيَّتَهُ	صفة	اسم
اَلثَّانِي	صفة	اسم علم
اَلْخَاصَّةِ	صفة	اسم
السَّيَّارَاتِ	صفة	اسم
قُلْ	فعل أمر	كلمة وظيفية
أَقَلَّ	فعل ماضي	اسم
أَدْنَى	فعل ماضي	اسم
أَكْبَرَ	فعل ماضي	صفة

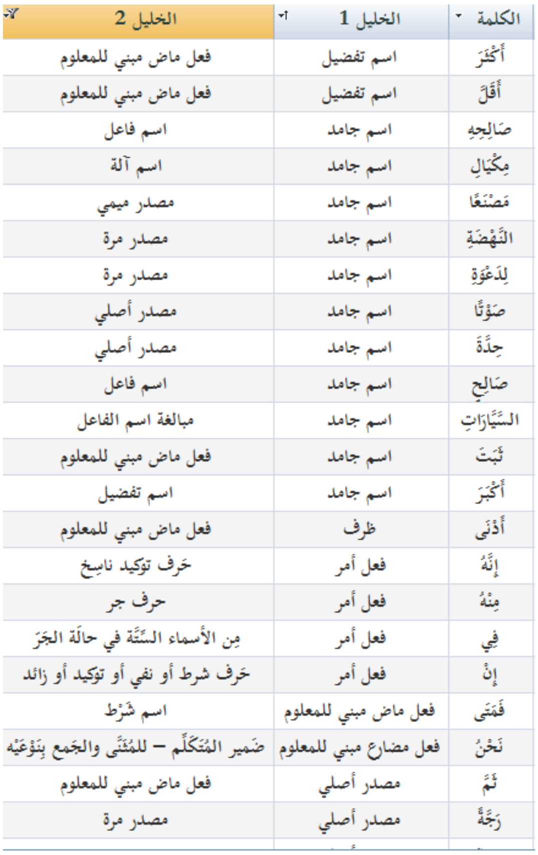
5.1.4 الخليل

يقوم هذا المحلل الصرفي بالتعرف على جميع الحلول الممكنة للكلمة، وتحديد لائحة الخصائص الصرفية لهذه الحلول، وتشمل هذه اللائحة بالنسبة لكل حل:

1. تشكيل الكلمة
2. السوابق واللواحق
3. نوع الكلمة
4. الوزن
5. الجذع
6. الحالة الإعرابية

بلغ متوسط الحلول الصرفية التي قدّمها محلل الخليل الصرفي لكل كلمة، ما قيمته 1.79 حل صرفي / كلمة، وبلغت نسبة الكلمات التي ليست لها أية حلول صرفية 31.47%

جدول 5: تعدد الأقسام الكلامية لبعض الكلمات في محلل الخليل الصرفي



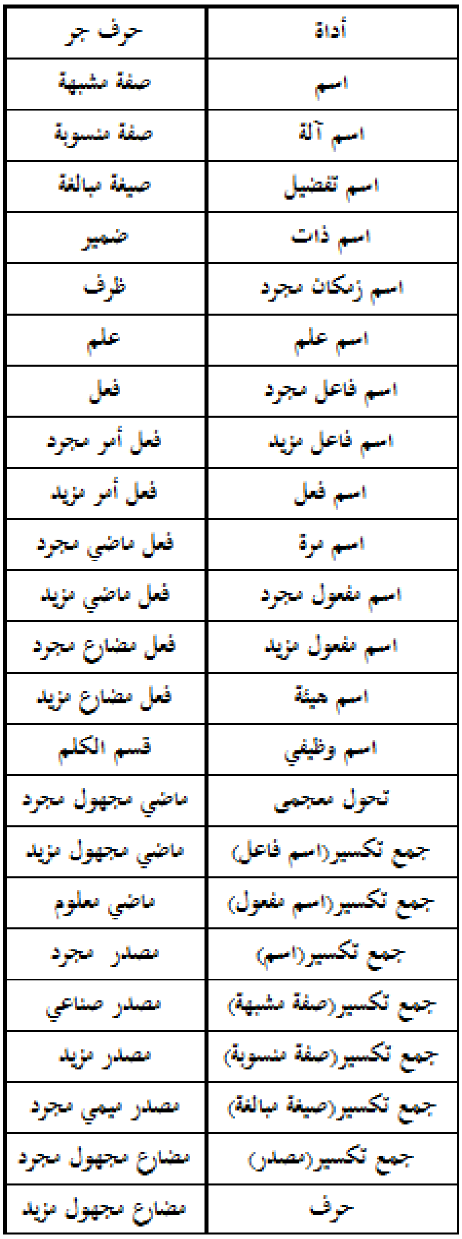
الكلمة	الخليل 1	الخليل 2
أَكْثَرَ	اسم تفضيل	فعل ماض مبني للمعلوم
أَقَلَّ	اسم تفضيل	فعل ماض مبني للمعلوم
صَالِحِهِ	اسم جامد	اسم فاعل
مِكْيَالِ	اسم جامد	اسم آلة
مَصْنَعًا	اسم جامد	مصدر ميمي
النَّهْضَةِ	اسم جامد	مصدر مرة
لِدَعْوَةِ	اسم جامد	مصدر مرة
صَوْتًا	اسم جامد	مصدر أصلي
جِدَّةَ	اسم جامد	مصدر أصلي
صَالِحِ	اسم جامد	اسم فاعل
السَّيَّارَاتِ	اسم جامد	مبالغة اسم الفاعل
ثَبَتَ	اسم جامد	فعل ماض مبني للمعلوم
أَكْبَرَ	اسم جامد	اسم تفضيل
أَدْنَى	ظرف	فعل ماض مبني للمعلوم
إِنَّهُ	فعل أمر	حَرف توكيد ناسِخ
مِنْهُ	فعل أمر	حرف جر
فِي	فعل أمر	مِن الأسماء السِّتَّة في حالَة الجَرّ
إِنْ	فعل أمر	حَرف شرط أو نفي أو توكيد أو زائد
فَمَتَى	فعل ماض مبني للمعلوم	اسم شَرْط
نَحْنُ	فعل مضارع مبني للمعلوم	ضَمير المُتَكَلِّم – للمُثَنَّى والجَمع بِنَوْعَيْه
ثَمَّ	مصدر أصلي	فعل ماض مبني للمعلوم
رَجَّةً	مصدر أصلي	مصدر مرة

5.2 أقسام الكلم في المحللات الصرفية العربية

5.2.1 صخر

بلغ عدد أقسام الكلم في محلل صخر 48 قسما، ويلاحظ أن هذه الأقسام تتضمن أقساما رئيسية مثل (اسم، فعل، ضمير، ...إلخ)، كما تتضمن أقساما فرعية مثل (اسم آلة، صفة مشبهة، فعل أمر، جمع تكسير، ...إلخ).

جدول 6: تعدد الأقسام الكلامية لبعض الكلمات في محلل صخر

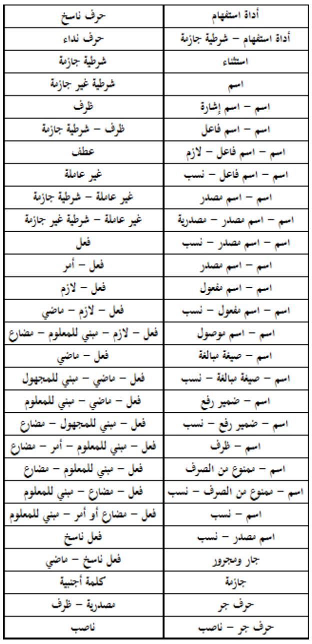


أداة	حرف جر
اسم	صفة مشبهة
اسم آلة	صفة منسوبة
اسم تفضيل	صيغة مبالغة
اسم ذات	ضمير
اسم زمكان مجرد	ظرف
اسم علم	علم
اسم فاعل مجرد	فعل
اسم فاعل مزيد	فعل أمر مجرد
اسم فعل	فعل أمر مزيد
اسم مرة	فعل ماضي مجرد
اسم مفعول مجرد	فعل ماضي مزيد
اسم مفعول مزيد	فعل مضارع مجرد
اسم هيئة	فعل مضارع مزيد
اسم وظيفي	قسم الكلم
تحول معجمي	ماضي مجهول مجرد
جمع تكسير(اسم فاعل)	ماضي مجهول مزيد
جمع تكسير(اسم مفعول)	ماضي معلوم
جمع تكسير(اسم)	مصدر مجرد
جمع تكسير(صفة مشبهة)	مصدر صناعي
جمع تكسير(صفة منسوبة)	مصدر مزيد
جمع تكسير(صيغة مبالغة)	مصدر ميمي مجرد
جمع تكسير(مصدر)	مضارع مجهول مجرد
حرف	مضارع مجهول مزيد

5.2.2 الميزان

يعتمد الميزان في تحديد أقسام الكلم على الوزن في تحديد خصائص جذع الكلمة، وعلى السوابق واللواحق في تحديد خصائص لواصق الكلمة، وقد بلغت أقسام الكلم من خلال عينة الاختبار 56 نوعا.

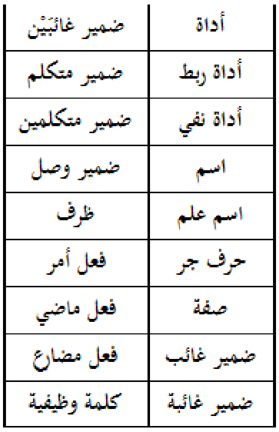
جدول 7: الأقسام الكلامية في الميزان



أداة استفهام	حرف ناسخ
أداة استفهام – شرطية جازمة	حرف نداء
استثناء	شرطية جازمة
اسم	شرطية غير جازمة
اسم – اسم إشارة	ظرف
اسم – اسم فاعل	ظرف – شرطية جازمة
اسم – اسم فاعل – لازم	عطف
اسم – اسم فاعل – نسب	غير عاملة
اسم – اسم مصدر	غير عاملة – شرطية جازمة
اسم – اسم مصدر – مصدرية	غير عاملة – شرطية غير جازمة
اسم – اسم مصدر – نسب	فعل
اسم – اسم مصدر	فعل – أمر
اسم – اسم مفعول	فعل – لازم
اسم – اسم مفعول – نسب	فعل – لازم – ماضي
اسم – اسم موصول	فعل – لازم – مبني للمعلوم – مضارع
اسم – صيغة مبالغة	فعل – ماضي
اسم – صيغة مبالغة – نسب	فعل – ماضي – مبني للمجهول
اسم – ضمير رفع	فعل – ماضي – مبني للمعلوم
اسم – ضمير رفع – نسب	فعل – مبني للمجهول – مضارع
اسم – ظرف	فعل – مبني للمعلوم – أمر – مضارع
اسم – ممنوع من الصرف	فعل – مبني للمعلوم – مضارع
اسم – ممنوع من الصرف – نسب	فعل – مضارع – مبني للمعلوم
اسم – نسب	فعل – مضارع أو أمر – مبني للمعلوم
اسم مصدر – نسب	فعل ناسخ
جار ومجرور	فعل ناسخ – ماضي
جازمة	كلمة أجنبية
حرف جر	مصدرية – ظرف
حرف جر – ناصب	ناصب

5.2.3 باك ولتر

لم تتجاوز أقسام الكلم على عينة اختبار محلل باكولتر 18 قسما.

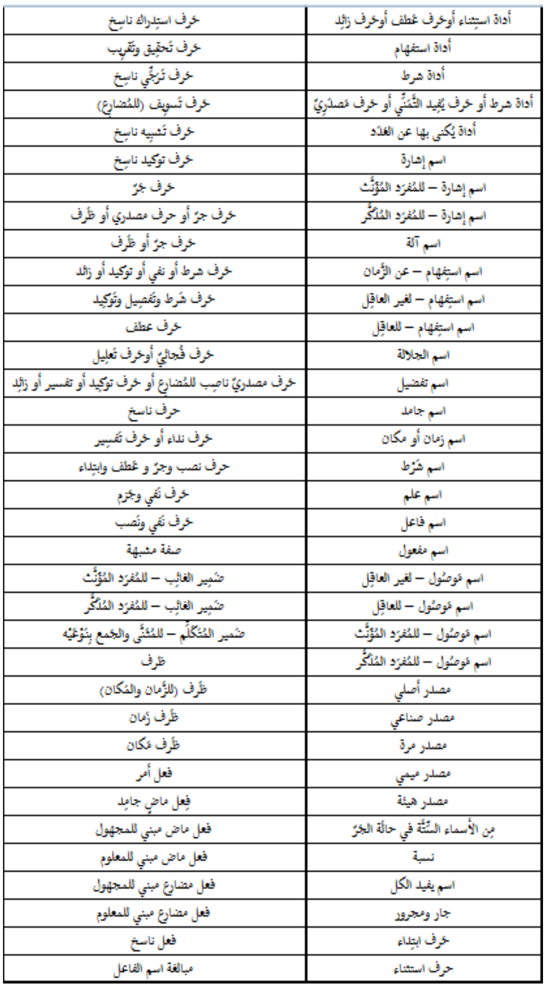


أداة	ضمير غائبَيْن
أداة ربط	ضمير متكلم
أداة نفي	ضمير متكلمين
اسم	ضمير وصل
اسم علم	ظرف
حرف جر	فعل أمر
صفة	فعل ماضي
ضمير غائب	فعل مضارع
ضمير غائبة	كلمة وظيفية

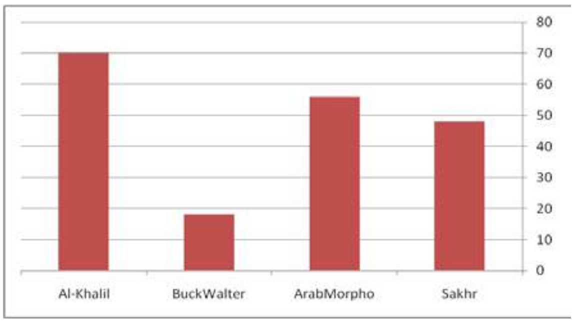
5.2.4 الخليل

محلل الخليل الصرفي بلغت أقسام الكلم في عينة اختباره 70 قسما

جدول 9: الأقسام الكلامية في محلل الخليل الصرفي



أداة استِثناء أوحرف عَطف أوحرف زائِد	حَرف استِدراك ناسِخ
أداة استفهام	حَرف تَحقِيق وتَقرِيب
أداة شرط	حَرف تَرَجِّي ناسِخ
أداة شرط أو حَرف يُفِيد التَّمَنِّي أو حَرف مَصدَرِيّ	حَرف تَسوِيف (للمُضارِع)
أداة يُكنى بها عن العَدَد	حَرف تَشبِيه ناسِخ
اسم إشارة	حَرف توكِيد ناسِخ
اسم إشارة – للمُفرَد المُؤَنَّث	حَرف جَرّ
اسم إشارة – للمُفرَد المُذَكَّر	حَرف جَرّ أو حرف مصدري أو ظَرف
اسم آلة	حَرف جَرّ أو ظَرف
اسم استِفهام – عن الزَّمان	حَرف شرط أو نفي أو توكيد أو زائِد
اسم استِفهام – لِغير العاقِل	حَرف شَرط وتَفصِيل وتَوكِيد
اسم استِفهام – للعاقِل	حَرف عطف
اسم الجلالة	حَرف فُجائِيّ أوحرف تَعلِيل
اسم تفضيل	حَرف مصدريّ ناصِب للمُضارِع أو حَرف توكِيد أو تفسير أو زائِد
اسم جامد	حرف ناسخ
اسم زمان أو مكان	حَرف نداء أو حَرف تَفسِير
اسم شَرْط	حرف نصب وجرّ و عَطف وابتِداء
اسم علم	حَرف نَفي وجزم
اسم فاعل	حَرف نَفي ونَصب
اسم مفعول	صفة مشبهة
اسم مَوصُول – لِغير العاقِل	ضَمِير الغائِب – للمُفرَد المُؤَنَّث
اسم مَوصُول – للعاقِل	ضَمِير الغائِب – للمُفرَد المُذَكَّر
اسم مَوصُول – للمُفرَد المُؤَنَّث	ضَمير المُتَكَلِّم – للمُثَنَّى والجمع بِنَوْعَيْه
اسم مَوصُول – للمُفرَد المُذَكَّر	ظَرف
مصدر أصلي	ظَرف (للزَّمان والمَكان)
مصدر صناعي	ظَرف زَمان
مصدر مرة	ظَرف مَكان
مصدر ميمي	فعل أمر
مصدر هيئة	فِعل ماضٍ جامِد
مِن الأسماء السِّتَّة في حالَة الجَرّ	فعل ماض مبني للمجهول
نسبة	فعل ماض مبني للمعلوم
اسم يفيد الكل	فعل مضارع مبني للمجهول
جار ومجرور	فعل مضارع مبني للمعلوم
حَرف ابتِداء	فعل ناسخ
حرف استثناء	مبالغة اسم الفاعل



شكل 1: عدد الأقسام الكلامية في المحللات الصرفية

5.3 تقريب منهج المحللات الصرفية في تقسيم الكلم من منهج د. تمام حسان

5.3.1 صخر

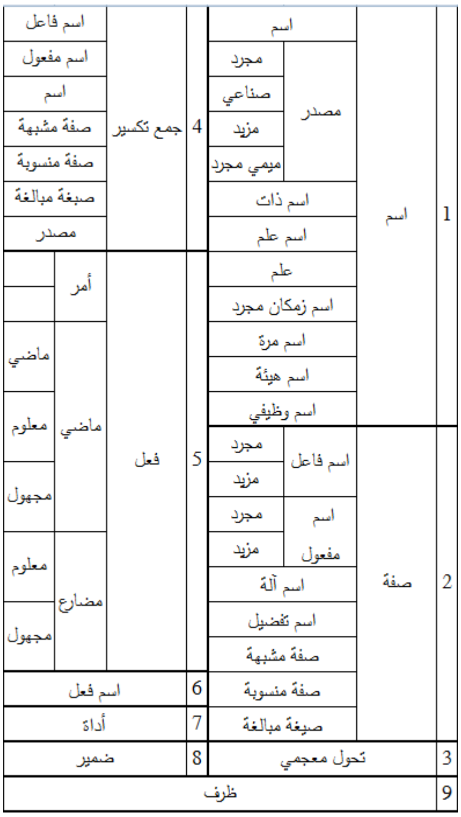
يمكن تقسيم الكلم داخل محلل صخر إلى 9 أقسام رئيسة: الاسم، الصفة، الفعل، الضمير، الأداة، اسم الفعل، الظرف، جمع التكسير، التحول المعجمي.

الأقسام السبعة الأولى تتفق تماما مع أقسام د. تمام حسان؛ بينما القسمين الأخيرين هما قسمان جديدان ليس بالنسبة لمنهج تمام حسان وإنما لمنهج الصرف العربي كله.

فجمع التكسير هو الجمع الذي تغيرت صورة مفرده، هذا المفرد قد يكون اسما وقد يكون صفة، ولذا يتم وضعه تحت قسمي الاسم أو الصفة. والتحول المعجمي ليس سوى دلالة جديدة للبنية الصرفية، هذه الدلالة قد تكون علما أو وظيفة أو أي شيء آخر غير المعنى الصرفي الذي يعبر عنه هذا المبنى.

إلا أن التطبيقات التي تقوم صخر بتنفيذها كالترجمة الآلية وغيرها، قد تكون هي السبب في إدخال هذه الأقسام الجديدة على الكلم العربي.

جدول 10: أقسام الكلم العربي لدى صخر بعد إعادة هيكلتها



<table>
<tr><td rowspan="12">1</td><td rowspan="12">اسم</td><td colspan="2">اسم</td><td rowspan="7">4</td><td rowspan="7">جمع تكسير</td><td colspan="2">اسم فاعل</td></tr>
<tr><td rowspan="4">مصدر</td><td>مجرد</td><td colspan="2">اسم مفعول</td></tr>
<tr><td>صناعي</td><td colspan="2">اسم</td></tr>
<tr><td>مزيد</td><td colspan="2">صفة مشبهة</td></tr>
<tr><td>ميمي مجرد</td><td colspan="2">صفة منسوبة</td></tr>
<tr><td colspan="2">اسم ذات</td><td colspan="2">صبغة مبالغة</td></tr>
<tr><td colspan="2">اسم علم</td><td colspan="2">مصدر</td></tr>
<tr><td colspan="2">علم</td><td rowspan="12">5</td><td rowspan="12">فعل</td><td rowspan="2">أمر</td><td></td></tr>
<tr><td colspan="2">اسم زمكان مجرد</td><td></td></tr>
<tr><td colspan="2">اسم مرة</td><td rowspan="6">ماضي</td><td rowspan="2">ماضي</td></tr>
<tr><td colspan="2">اسم هيئة</td></tr>
<tr><td colspan="2">اسم وظيفي</td><td rowspan="2">معلوم</td></tr>
<tr><td rowspan="9">2</td><td rowspan="9">صفة</td><td rowspan="2">اسم فاعل</td><td>مجرد</td></tr>
<tr><td>مزيد</td><td rowspan="2">مجهول</td></tr>
<tr><td rowspan="2">اسم مفعول</td><td>مجرد</td></tr>
<tr><td>مزيد</td><td rowspan="4">مضارع</td><td rowspan="2">معلوم</td></tr>
<tr><td colspan="2">اسم آلة</td></tr>
<tr><td colspan="2">اسم تفضيل</td><td rowspan="2">مجهول</td></tr>
<tr><td colspan="2">صفة مشبهة</td></tr>
<tr><td colspan="2">صفة منسوبة</td><td>6</td><td colspan="3">اسم فعل</td></tr>
<tr><td colspan="2">صيغة مبالغة</td><td>7</td><td colspan="3">أداة</td></tr>
<tr><td>3</td><td colspan="3">تحول معجمي</td><td>8</td><td colspan="3">ضمير</td></tr>
<tr><td>9</td><td colspan="7">ظرف</td></tr>
</table>

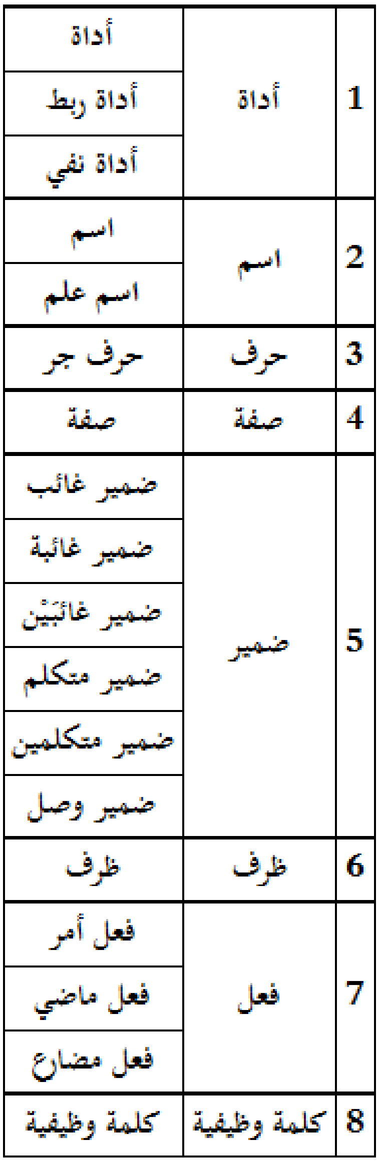
5.3.2 الميزان

يمكن تقسيم الكلم في الميزان إلى 6 أقسام، هي: (اسم، فعل، صفة، ضمير ، أداة، ظرف)

5.3.3 باك ولتر

أما أقسام الكلم في محلل باك وولتر فيمكن القول إنها ثمانية أقسام هي: اسم – صفة – حرف – ضمير – أداة – ظرف – كلمة وظيفية.

جدول 11: أقسام الكلم العربي لدى باكولتر بعد إعادة هيكلتها



1	أداة	أداة
		أداة ربط
		أداة نفي
2	اسم	اسم
		اسم علم
3	حرف	حرف جر
4	صفة	صفة
5	ضمير	ضمير غائب
		ضمير غائبة
		ضمير غائبَيْن
		ضمير متكلم
		ضمير متكلمين
		ضمير وصل
6	ظرف	ظرف
7	فعل	فعل أمر
		فعل ماضي
		فعل مضارع
8	كلمة وظيفية	كلمة وظيفية

5.3.4 الخليل

يمكن جعل أقسام الكلم في الخليل سبعة هي: اسم – صفة – فعل – ضمير – حرف – أداة – ظرف.

جدول 12: أقسام الكلم العربي لدى محلل الخليل بعد إعادة هيكلته

1	اسم	اسم إشارة	اسم إشارة
			للمُفرَد المُؤنَّث
			لمُفرَد المُذكَّر
		اسم آلة	
		اسم استفهام	عن الزَّمان
			لغير العاقِل
			للعاقِل
		اسم الجلالة	
		اسم جامد	
		اسم زمان أو مكان	
		اسم شَرْط	
		اسم علم	
		اسم موصول	لغير العاقِل
			للعاقِل
			للمُفرَد المُؤنَّث
			للمُفرَد المُذكَّر
		مصدر	أصلي
			صناعي
			مرة
			ميمي
			هيئة
		مِن الأسماء السِّتَّة في حالة الجَرّ	
		نسبة	
		اسم يفيد الكل	
		جار ومجرور	

2	أداة	استِثناء أوحَرف عَطف أوحَرف زائِد	
		استفهام	
		شرط	
		شرط أو حَرف يُفيد التَّمنِّي أو حَرف مَصدَرِيّ	
		يُكنى بها عن العَدَد	
3	ضمير	ضَمِير الغائِب – للمُفرَد المُؤنَّث	
		ضَمِير الغائِب – للمُفرَد المُذكَّر	
		ضَمير المُتكَلِّم – للمُثنى والجَمع بِنَوعَيه	
4	ظرف	ظَرف	
		ظَرف (للزَّمان والمَكان)	
		ظَرف زمان	
		ظَرف مَكان	
5	فعل	فعل أمر	
		ماض	جامِد
			مبني للمجهول
			مبني للمعلوم
		مضارع	مبني للمجهول
			مبني للمعلوم
		فعل ناسِخ	
6	صفة	مبالغة اسم الفاعل	
		صفة مشبهة	
		اسم تفضيل	
		اسم فاعل	
		اسم مفعول	

7	حرف	ابتِداء
		استثناء
		استِدراك ناسِخ
		تَحقِيق وتَقرِيب
		تَرَجِّي ناسِخ
		تَسوِيف (للمُضارِع)
		تَشبِيه ناسِخ
		تَوكِيد ناسِخ
		جَرّ
		جَرّ أو حرف مصدري أو ظَرف
		جَرّ أو ظَرف
		شرط أو نفي أو توكيد أو زائد
		شَرط وتَفصِيل وتَوكِيد
		عطف
		فُجائِيّ أوحَرف تَعلِيل
		مَصدرِيّ ناصِب للمُضارِع أو حَرف تَوكِيد أو تفسير أو زائِد
		ناسخ
		نداء أو حَرف تَفسِير
		نصب وجَرّ و عَطف وابتِداء
		نَفي وجَزم
		نَفي ونَصب

6. الاستنتاجات والتوصيات

- لا يمكن قياس كفاءة المحلل الصرفي العربي بناء على ما يقدمه من معلومات صحيحة فقط، ولكن ينبغي أن تقاس كفاءته بمدى مطابقة هذه المعلومات لما اتفق عليه علماء الصرف، وبعبارة أخرى ينبغي أن تتوافق المحللات الصرفية على نظام معياري للتحليل الصرفي، يأخذ في اعتباره منطلقات علم الصرف العربي ومنتهياته.
- يظل تحليل أقسام الكلم الرئيسة والفرعية داخل المحللات الصرفية بحاجة إلى عينة اختبارية أكبر حجما؛ كي تستوعب جميع تلك الأقسام، بالإضافة إلى عناصر التحليل الصرفي الأخرى.
- يكفي محلل صرفي عربي معياري واحد يستخدم في جميع التطبيقات التي تحتاج إلى التحليل الصرفي.
- إدخال أقسام صرفية جديدة على الكلم العربي ينبغي أن يعتمد على المنهج الوصفي الذي يقرر هذا بناء على أساسي المعنى والمبنى، خاصة إذا كان هذا القسم الجديد قد تم تناوله ضمن الدلالة.
- أثبتت محاولة تقريب مناهج المحللات الصرفية إلى منهج د. تمام حسان أنه من الممكن وببعض الخوارزميات البسيطة والعنونة الصرفية اليدوية، أنه من الممكن الوصول إلى محلل صرفي معياري للغة العربية، يراعي متطلبات التطبيقات التكنولوجية ولا يهمل أركان النظام الصرفي العربي العبقري.

الأعمال المستقبلية

- الاستفادة من المحللات الصرفية للغات الاشتقاقية الأجنبية كالعبرية والروسية والصربية والفرنسية.
- محاولة تحديد مواصفات المحلل الصرفي المعياري للغة العربية المفيد لتطبيقات معالجة اللغة العربية آليا.

شكر وتقدير

أود التوجه بخالص الشكر إلى الدكتور ياسر حفني الذي ساعدني بأفكاره وتوجيهاته في كتابة هذه الورقة، كما أشكر كلا من شركتي صخر – آر دي آي على إمدادي بالتحليلات الصرفية التي احتجت إليها.

المراجع

[1] As-Saqy, F., 1977. *Arabic parts of speech in terms of form and function*, Khanji Library, Cairo.

[2] Attia, M., Rashwan, M., 2004. *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*, Proceeding of the Arabic Language Technology and Resources Int'l Conference: NEMLAR, Cairo.

[3] Attia, M., 2000. *A Large-Scale Computational Processor Of The Arabic Morphology, And Applications*, A Thesis Submitted to the Faculty of Engineering, Cairo University in Partial Fulfillment of the Requirements for the Degree of Master Of Scince in Computer Engineering, Giza.

[4] Hassaan, T., 1994. *Arabic language its sense and structure*, Dar Ath-Thaqafa, Casablanca.

[5] Sawalha, M., Atwell, E., 2009. *Adapting Language Grammar Rules for Building a Morphological Analyzer for Arabic Text*, Proceedings of ALECSO Arab League Educational Cultural and Scientific Organization workshop on Arabic morphological analysis, Tunisia.

عمرو حمدي الجندي

- معيد بمجمع اللغة العربية بالقاهرة.
- مؤسس فريق حاسوبويه لتقنيات اللغة العربية.
- محاضر لغويات حاسوبية بجامعة عين شمس وأكاديمية حاسوبويه.
- باحث لغويات حاسوبية سابقا بالشركة الهندسية لتطوير الأنظمة الرقمية.
- عضو لجنة التحكيم في أكثر من مجلة علمية ومؤتمر دولي في مجال تقنيات اللغة العربية.
- باحث ماجستير في اللغويات الحاسوبية بقسم علم اللغة، كلية دار العلوم، جامعة القاهرة.
- استضيف لدى العديد من البرامج التليفزيونية والإذاعية متحدثا عن حوسبة اللغة العربية.
- شارك في العديد من المؤتمرات واجتماعات الخبراء بأبحاث متخصصة في تقنيات اللغة العربية.
- شارك في أكثر من مشروع بحثي ممول في مجالي الأنطولوجيا والتنقيب عن المعلومات داخل النصوص.
- درّب عشرات اللغويين على استخدام أدوات معالجة اللغات الطبيعية NLTK في معالجة النصوص.
- أسهم في بناء عدة مدونات لغوية ضخمة في مجالات الدلالة والشعر والتعرف البصري على الحروف OCR

Abstract

ARABIC PARTS OF SPEECH

Towards The Arabic Morphological Analyzers Evaluation in Spot of Tammam Hassan's Approach

Amr Hamdy El-Gendy

Academy of the Arabic Language
Hasoubawayh Team
Amr25@hotmail.com

Abstract: This paper proposes to evaluate the Arabic Morphological Analyzers analysis to the part of speech in spot of Tammam Hassan's Approach. Paper offers some of the facts which think are essential in the language technologies, then describes the Arabic morphological system props, indicating the importance of the speech division and the need of the language technologies to the normative division, explaining the Arabic grammarians and number of Arabic morphological analyzers approach in the division of speech, pointing out the most important characteristic, finally, the paper proposes a number of recommendations in this context.

Building an Arabic Part of Speech Tagger

Hadeer Abdelrazik[1], Sameh Alansary[2]

Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University
El Shatby, Alexandria, Egypt.

[1]hadeer.abdelrazik@bibalex.org
[2]sameh.alansary@biballex.org

***Abstract*— This paper focuses mainly on building rule-based Arabic part of speech tagger (APOS) which is a major component in higher-level analysis of text corpora. Before building the tagger some steps have to be taken. The first step is to compile representative corpus that represents the modern Standard Arabic (MSA). Suggesting new tag sets is one of the most important steps. Collecting Arabic rules from different Arabic grammatical books and handling these rules to be suitable for MSA. The purpose of building a POS tagger is to assign part of speech tags automatically to words reflecting their syntactic category. The final stage is designing the APOS to disambiguate Arabic words. Its output can also be used in many NLP applications, such as: Speech synthesis, Spelling correction, searching large text databases, information extraction, Question Answering, Speech Recognition, Text-to-speech conversion, Machine Translation, Grammar Correction and many more. It is also one of the main tools needed to develop any language corpus. The evaluation stage of the APOS tagger contributes 84%.**

1 INTRODUCTION

Natural Language Processing (NLP) is a research discipline related to artificial intelligence, linguistics, philosophy, and psychology. The aim of this discipline is building systems that are capable of understanding and interpreting the computational mechanisms of natural languages.

Part-of-speech tagging (POS tagging or POST) disambiguation, is the process of automatic tags assignment of words in a text according to the word features and its relationship with the adjacent and related words in a phrase, sentence, or paragraph; the word in context[1]. Fig. 1 shows an example of POST for words of the sentence "the girl kissed the boy on the cheek":

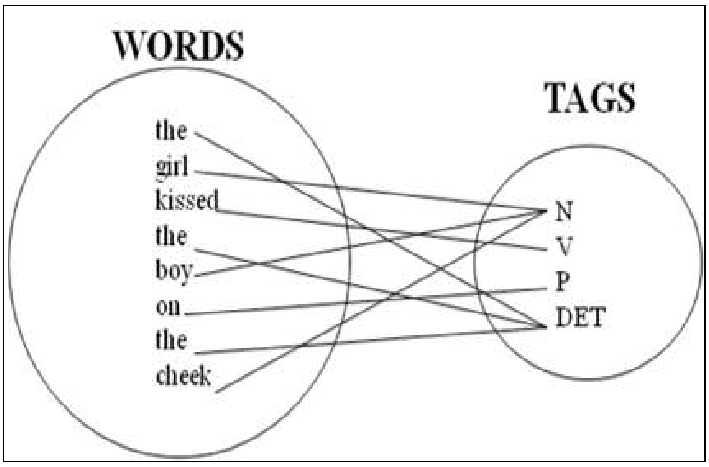


Figure 1: POST for the sentence "the girl kissed the boy on the cheek"

There are three general approaches to deal with tagging: first is the rule-based approach where there are fundamentals that any system depends upon such as dictionary or lexicon to obtain all the possible tags for every word and a set of hand-written rules to identify the correct tag from the multiple choices of tags for each word. Then, the linguistic rules are analyzed in order to disambiguate words. Second, the statistical approach which belongs to the mainstream approach in natural language processing and computational linguistics including POST, syntactic parsing, semantic interpretation, lexical acquisition, machine translation, information retrieval, and also in language learning such as automatic grammar induction and syntactic or semantic word clustering [1]. Third, the transformation-based learning (TBL) approach which allows having linguistic knowledge in a readable form, because it is a rule-based algorithm for automatic POST by transforming from one linguistic level to another using transformation rules in order to find the suitable tag for each word. It consists of algorithms, and rules and processing[2]. In probabilistic approaches only few tags are possible for any

[1] http://en.wikipedia.org/wiki/Part-of-speech_tagging
[2] http://en.wikipedia.org/wiki/Brill_tagger

given word; the list of tags can be found in the lexicon. Contextual rules that define the valid sequences of tags which permit choosing the correct tag from the local context [2].

There are three approaches to deal with the tagging problem: Rule-based approach: consists of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags. Statistical approach: consists of building a trainable model and to use previously-tagged corpus to estimate its parameters. Hybrid approach: Consists in combining rule-based approach with a statistical one. Most of the recent study uses this approach as it gives better results [3].

In what follows, section 2 presents some Arabic taggers and analyzers such as Arabic Part-of-Speech Tagger (APT), ASVM Tagger Tag set, Statistical Arabic Part-of-Speech Tagger (APOS), RDI Arab tagger (ArabMorpho), MAPSSeman Lite PoS Tagger, and Alkhalil Morhphological analyzer (Alkhalil Morpho Sys). Section 3 describes the levels of building APOS tagger which included suggesting new tag sets, data preparation, and designing APOS tagger. Conclusions and results are presented in section 4.

2 SOME ARABIC TAGGERS AND ANALYZERS

A Part-Of-Speech Tagger (POS Tagger) is software that reads text in some language and assigns parts of speech to each word as well as any other token[3].

A. Arabic Part-of-Speech Tagger (APT)

APT was built by Shereen Khoja; it was developed using a combination of both statistical and rule-based techniques since hybrid taggers seem to produce the highest accuracy rates. The APT Tag sets were 131 derived from the Arabic grammatical tradition rather than from an Indo-European based Tag sets, the reason for this is that Arabic is a very different language from Indo-European languages [4]. These parts-of-speech are further sub-categorized into more detailed parts-of-speech which collectively cover the whole of the Arabic language [5].
However, there are some disadvantages that have appeared; the transliteration of the tagger has no vowels to differentiate between the pronunciations of words such as “ksrna ,كسرنا we broke”. The tags are treated as one compound without separating its components such as “VPPl2MJ”. There is no differentiation between “taa marbuta” and “taa maftuha "ة"” such as “mdrst “مدرسة” which is pronounced as “مدرست””. The initial version of the lexicon contains all the corpus words without removing any clitics. Finally, the tags in APT tag sets are insufficient does not provide full description of the language [6].

B. ASVM Tagger Tag set

ASVM Tagger is based on a supervised machine learning technique which is Support Vector Machine (SVM)[4], by Diab M. et al. ASVM-POS system uses a compact tag set of simple RTS tags. RTS is the collapsed tags available in the Arabic Treebank distribution, this collapsed tag set is a manually reduced form of the 135 morpho-syntactic tags created by AraMorph, and a rule based morphological analyzer by Buckwalter. They consist of 24 tags, by adding definiteness, number and gender information to enrich the number of tags; they comprise the ERTS tag set. ERTS tag set is comprised of 75 tags, but only 57 tags are instantiated for the current system. It also reflects some of Modern Standard Arabic morphological features [3]. The ASVM-POS tags each word with only a single tag specifying the word type as NN for noun, VBD for Verb in past tense. It also separates each clitic (prefix or suffix) as a single token with separate tag [7].

C. Statistical Arabic Part-of-Speech Tagger (APOS)

Arabic part-of-speech tagger (APOS) is the development that can be used for analyzing and annotating traditional Arabic texts, especially the Quran text and relatively old books (from the third century Hijri) [8]. Moreover, this tagger is responsible for assigning to each word the most appropriate morphological tag by using 13 tags: 3 subcategories of verbs, 6 subcategories of nouns, and 4 subcategories of particles with investigating the principle aspects of Arabic morphology and grammar. It counts a total words of 21882 with a 3565 unique words ranged in more than 1600 sentences. Among these counts, there are 10258 nouns, 2587 verbs, and 9037 particles.

[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://www.ldc.upenn.edu/myl/morph/buckwalter.html

D. RDI Arab tagger (ArabMorpho)

RDI Arab tagger is based on long n-grams probability estimation in addition to powerful and efficient tree search algorithms and search-based disambiguation technique. However, it does not have specific tags, for example, proper nouns may be assigned to the same general tag “NN” and ”NNS” for plural noun. r. RDI tag set consists of 62 tags. Each word is tagged with multiple detailed tag vectors; these vectors represent some morphological features of this word such as word type, gender, number, syntactic case and definiteness. Also, a word containing prefix or suffixes or both, is tagged as a single word [7].

The Arabic POS tagging process are implemented in the following steps: 1) The Arabic strings sequence to be POS tagged are morphologically analyzed and disambiguated in combinatorial manner using ArabMorpho©. These results are presented in a disambiguated quadruples sequence where each string is substituted by either one quadruple or a mark of transliterated string. 2) For the prefix, pattern, and suffix morphemes of each quadruple in the sequence, the Arabic POS labels; APOS (p) APOS (t:f) APOS(s) are retrieved from the Arabic lexical knowledge base. 3) The Arabic POS tags vector of each word in the sequence is then composed using the formula:

$$\mathbf{APOS\ (w)=Concat(APOS(p),\ APOS(t{:}f),\ APOS(s))\ (2)}$$

Where the “Concat” function simply concatenates the POS sub vectors of the constituting morphemes after eliminating any mutual redundancy among their tags.

E. MAPSSeman Lite PoS Tagger[5]

Kalmasoft's Deep PoS Tagger (MAPSSeman® Lite PoS Tagger) returns semi context-free solutions for each token using comprehensive set of syntactic rules. It is designed to prepare Arabic corpus, since tagged corpus is more useful than an untagged corpus, because it contains more information than in raw texts. Once a corpus is tagged, it can be used to extract information. Then, this can be used for creating dictionaries and grammars for a language using real language data. Tagged corpora are also useful for detailed quantitative analysis of texts. Unfortunately, this tagger is not available.

F. Buckwalter’s Morphological Analyzer (BAMA)

Tim Buckwalter Morphological analyzer 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 [9].

In Buckwalter analyzer, Arabic words are segmented into prefix, stem and suffix strings[6]. Buckwalter morphological analyzer has two versions:

The first is Buckwalter Arabic Morphological Analyzer Version 1.0. It was produced by Linguistic Data Consortium (LDC) (2002); representing a stage of development somewhere between the Penn Arabic Treebank part 1 (the AFP corpus) and the Penn Arabic Treebank part 2 (The Ummah Corpus). The Buckwalter Arabic Morphological Analyzer is being used for POS-tagging a considerable amount of text data, and various orthographic anomalies have been observed in the source text.

The data consists primarily of three Arabic-English lexicon files: prefixes (299 entries), suffixes (618 entries), and stems (82,158 entries representing 38,600 lemmas). The lexicons are supplemented by three morphological compatibility tables used for controlling prefix-stem combinations (1,648 entries), stem-suffix combinations (1,285 entries), and prefix-suffix combinations (598 entries). The actual code for morphology analysis and POS tagging is contained in a Perl script[7].

The second is Buckwalter Arabic Morphological Analyzer Version 2.0. This is the version that was used for morphological annotation and POS tagging of the Penn Arabic tree bank part 3 (The Annahar Corpus). This release is available only to LDC members.

The data consists primarily of three Arabic-English lexicon files: prefixes (548 entries), suffixes (906 entries), and stems (78,839 entries representing 40,219 lemmas). The lexicons are supplemented by three morphological compatibility tables used for controlling prefix-stem combinations (2,435 entries), stem-suffix combinations (1,612 entries), and prefix-suffix combinations (1,138 entries). The actual code for morphology analysis and pos tagging is contained in a Perl script (aramorph.pl). Sample input (infile.txt) and corresponding output file (outfile.xml) are provided[8].

In this Tag set there is no distinction between categories and features for POS. The particle classification has no attributes. He does not distinguish between attached pronouns or other clitics and inflection of the word (suffixes) [10].

[5] http://www.kalmasoft.com/index.htm

[6] http://www.ldc.upenn.edu/Catalog/docs/LDC2004L02/readme.txt

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

[8] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004L02

G. Alkhalil Morhphological analyzer (Alkhalil Morpho Sys)

AlKhalil Morpho Sys could be considered as the best Arabic morphological system. Actually, AlKhalil won the first position, among 13 Arabic morphological systems around the world, at a competition held by (المنظمة العربية للتربية والعلوم) The Arab League Educational, Cultural and Scientific Organization (ALECSO) [12]. The system was developed in collaboration with the Arab League Educational, Cultural and Scientific Organization (ALECSO) and King Abdul Aziz City for Science and Technology (KACST).

For each given word, Alkhalil Morpho Sys enables the identification of all of the possible solutions associated with their morphosyntactic features.

For nouns these features are as follows:

1) For non derivable words, the system gives:

- Vocalization
- The proclitic and the enclitic associated whenever they exist,

2) For derivable words, the system generally proposes several solutions. For each of these solutions, the system displays:

- Vocalization
- The proclitic and the enclitic associated whenever they exist,
- The nature of the noun:
  - Different verbal noun types (مصدر أصلي ، مصدر ميمي).
  - Active participle (اسم فاعل).
  - Passive participle (اسم مفعول).
  - Time and place nouns.
  - Instrumental noun (اسم آلة).

Al khalil also provides the concordance between proclitics and enclitics along with the output syntactic features:

- To check the concordance of the stem ultimate character's diacritical short vowel (العلامة الإعرابية) with the proclitic syntactic function, e.g., the prepositions "ب" and "ك"appear only with nouns in genitive case (الأسماء المجرورة).
- To check the concordance of the word nature with the enclitic, e.g., No concordance between the enclitic pronoun "هم" and passive verbs.
  - Concordance of the hamza allography (ء,أ , إ, ئـ, or ؤ) in the system's proposed solutions with that of the input word, e.g., the hamza "ؤ"cannot be followed by the short vowel kasra "ِ"
  - Concordance of the vocalizations of the system's proposed solutions with those that may exist in the input word. (Manual reference).
  - Of benefit to users, Open Source software is licensed so you can download and use the software free-of-charge. The source code for this software is made available free-of-charge, you (or a programmer you hire) can make changes to this software to better meet your needs, and you can release your changed code back to the community passing the benefit on to other users[9].

3 PLANNING TO BUILD APOS TAGGER

Arabic language processing (ALP) is a field of research in which many linguistic specialists are interested. Such interest has increased with the written Arabic documents' proliferation which is partly due to the Web popularization and the increase of the means of communication in Arabic. Researchers achievements in the recent years, in this field have led to a variety of important applications such as automatic indexing, information retrieval, machine translation, automatic summarization, automatic word generation, syntactic analysis, morphological analysis, automatic vocalization, spell checking and text analysis systems [11]. Furthermore, tagging is one of the important applications that has gone through a lot of development. However, tagging is a challenging task. The degree of difficulty depends on the language under consideration. In this research, Arabic is the language that is considered which is a highly inflected language. Arabic language is morphologically quite regular; it includes very few irregular forms. Moreover, Arabic is highly inflectional which leads to changes in the structure of the words in many cases; ultimately causing high degree of complexity of tagging. The other problem is the lack of Arabic language resources such as corpora and tools.

Modern Standard Arabic (MSA) processing is highly affected by the missing diacritical which makes it more complex to both syntactic and semantic analysis. That is due to the fact that diacritics reduce the number of possible classes for each word.

However, most of the current written texts are without diacritics. Thus, the most appropriate solution is to remove the diacritics for all the diacritized words of the corpus in order to uniform the data. Also, experimental results on undiacritized Arabic were proven to be useful [2]. The Arabic word is composed of stem and affixation that indicate

[9] http://alkhalil.sourceforge.net/

tense, gender and number. Clitics can also be attached to the beginning of the stem, end or both. Clitics are segments that represent an independent syntactic role; mainly conjunctions, preposition and pronouns. Prepositions and conjunctions are attached to beginning of the word while pronouns at the end [6]. Bar-Haim referred to each unit of the word that represents an independent tag as segment [12].

A. Suggested New POS Tag Set

Alkhalil analyzer is the best analyzer, although it has some problems with its database [13]. It has some limitations; it does not provide POS tags in a format that is reusable. Neither does it fully differentiate between clitics and affixes nor it detects proclitics or enclitics, but they are referred to either as prefix or suffix [14]. Accordingly the built tagger has adopted Alkhalil Marphological analyzer tag sets with some editing to disambiguate the data that have been used to test the built tagger.

This tag set is quite different from AlKhalil Sys tag sets in some points:

- Pronoun category is separated from the noun category, because the pronouns have different features from that of the nouns. Most Arabic nouns are derived from trilateral lexical roots, and all the nouns that are derived from the same root are clustered under that root entry in any Arabic or Arabic-English dictionary. However, some nouns have more than one root. Also, Arabic nouns are usually derived from lexical roots through application of particular morphological patterns [15].
- Adjective category is separated from the noun category, because an adjective modifies a noun by describing, identifying, or quantifying words[10]. An adjective is preceded by the noun which it modifies[11]. It's an attribute, a characteristic, or a description originally refers to a continuing adjective. Adjective has specific patterns such as "?afaal أَفّْعَل", "faala? فَعْلاء", "faalan فَعْلان" , "faala فَعْلى", "fael فَعِل" , "faela فَعِلَة", and "fail فَعِيل" , "faila فَعِيلة"[12]. As shown in the following example:

Examples of Arabic sentences that contains adjectives:

هذا كتابٌ مفيدٌ.
قرأت الكتابَيْن المفيدَيْن.

In this tag set, Alkhalil's tag set is further modified to be smaller and more inclusive, because, generally smaller tag sets perform better on unknown words [16].

TABLE 1

FINAL MAIN POS TAG SET

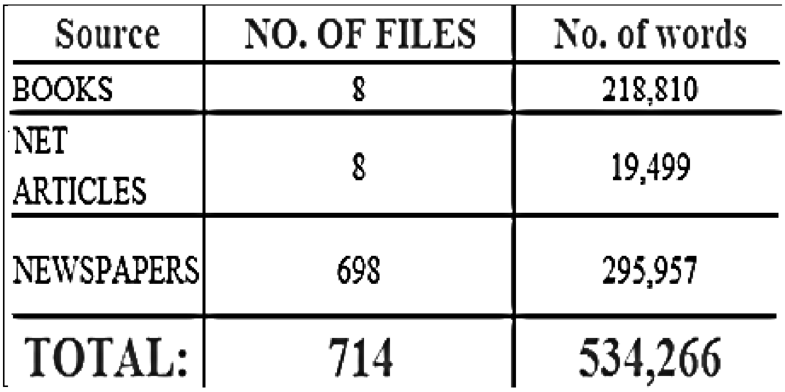
Main POS Tag set	**أقسام الكلام الرئيسية**	**POS Tag set Abbreviation**
Noun	اسم	NOUN
Adjective	صفة	ADJ
Verb	فعل	VERB
Pronoun	ضمير	PRON
Particle	أداة	PART

The previous tag sets are decided according to EGALS recommendations.

[10] http://www.writingcentre.uottawa.ca/hypergrammar/adjectve.html
[11] http://www.reefnet.gov.sy/education/kafaf/Bohoth/Naet.htm
[12] http://www.reefnet.gov.sy/education/kafaf/Bohoth/SefaMushabaha.htm

B. Data Preparation

1) *Collecting data:* The main task of compiling a corpus is collecting and editing texts13. This corpus was collected in 2011 from different sources that represent Modern Standard Arabic (MSA) to be used in morphological analysis process later. It was collected from three sources: Books, Net articles and Newspapers; every source consists of four main classes of Dewey's classification: Arts & recreation which contain Arts (CLA) and Sports (SPT), Social sciences which contains Economic (ECN) and Politics (PLT), Religion (RLG), Social sciences (SCT) and Computer science, information & general works and Miscellaneous (MSL) which contains all other sciences.
The whole collected data consists of approximately 534,266 words with 714 files distributed over the three sources; 218,810 words in 8 files of Books, 19,499 words in 8 files of Net Articles, and 295,957 words in 698 files of Newspapers. As shown in the Fig. 1:

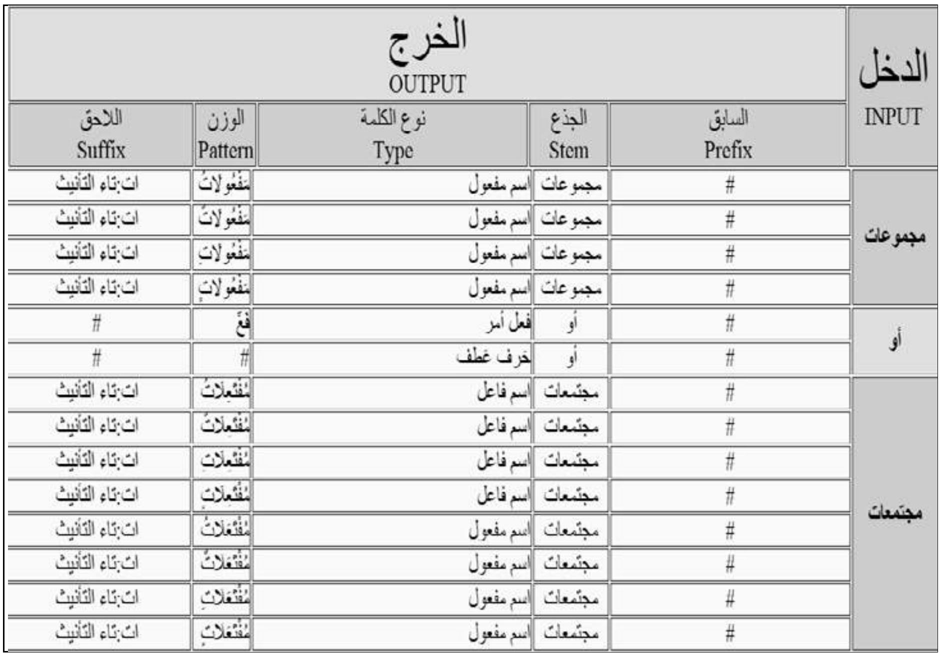


Source	NO. OF FILES	No. of words
BOOKS	8	218,810
NET ARTICLES	8	19,499
NEWSPAPERS	698	295,957
TOTAL:	714	534,266

Figure 2: The corpus distribution with numbers

Presenting the repressiveness of a language variety through a set of linguistic samples is a controversial issue. The discussion stems from the complexity found in defining representativeness itself and achieving that a fraction contains the components which confer the representative nature of the whole [17].

2) *Building lexicon:* The collected data was analyzed by AlKhalil morphological analyzer. Then, the researcher extracted the distinct solutions and enhance them to be compatible with the new tag sets for building the lexicon. Fig. 3 is an example of AlKhalil output:

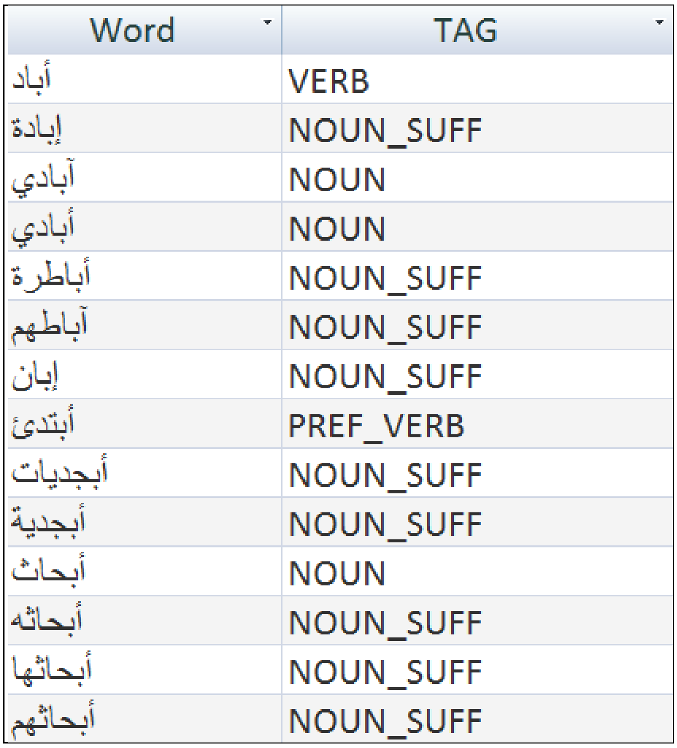


الخرج OUTPUT					الدخل INPUT
اللاحق Suffix	الوزن Pattern	نوع الكلمة Type	الجذع Stem	السابق Prefix	
ات:تاء التأنيث	مفعُولاتُ	اسم مفعول	مجموعات	#	مجموعات
ات:تاء التأنيث	مفعُولاتٌ	اسم مفعول	مجموعات	#	
ات:تاء التأنيث	مفعُولاتِ	اسم مفعول	مجموعات	#	
ات:تاء التأنيث	مفعُولاتٍ	اسم مفعول	مجموعات	#	
#	فعْ	فعل أمر	أو	#	أو
#	#	حرف عطف	أو	#	
ات:تاء التأنيث	مفتعِلاتُ	اسم فاعل	مجتمعات	#	مجتمعات
ات:تاء التأنيث	مفتعِلاتٌ	اسم فاعل	مجتمعات	#	
ات:تاء التأنيث	مفتعِلاتِ	اسم فاعل	مجتمعات	#	
ات:تاء التأنيث	مفتعِلاتٍ	اسم فاعل	مجتمعات	#	
ات:تاء التأنيث	مفتعَلاتُ	اسم مفعول	مجتمعات	#	
ات:تاء التأنيث	مفتعَلاتٌ	اسم مفعول	مجتمعات	#	
ات:تاء التأنيث	مفتعَلاتِ	اسم مفعول	مجتمعات	#	
ات:تاء التأنيث	مفتعَلاتٍ	اسم مفعول	مجتمعات	#	

Figure 3: An example of AlKhalil output

[13] https://www.uni-due.de/CP/compile_corpus.htm

Fig. 4 is the example of the created lexicon after enhancement:



Word	TAG
أباد	VERB
إبادة	NOUN_SUFF
آبادي	NOUN
أبادي	NOUN
أباطرة	NOUN_SUFF
آباطهم	NOUN_SUFF
إبان	NOUN_SUFF
أبتدئ	PREF_VERB
أبجديات	NOUN_SUFF
أبجدية	NOUN_SUFF
أبحاث	NOUN
أبحاثه	NOUN_SUFF
أبحاثها	NOUN_SUFF
أبحاثهم	NOUN_SUFF

Figure 4: Example of the created lexicon after handling

3) *Collecting rules:* The rules stage is one of the most important stages in the tagging and disambiguation processes. These rules were collected from different Arabic grammatical books and were developed according to the use of MSA. Rules are useful in increasing the accuracy by reducing the number of solutions for the selected words that have more than one solution or if the selected word has no solutions at all within the lexicon.

Examples of rules to assign tags to the words:

- If the current word is “خلال”, the following word is tagged as “NOUN”.
- If the current word is “لن”, the following word is tagged as “VERB”.
- If current word ends with “ة” or “ات”, assign the tag “NOUN_SUFF” to this word.

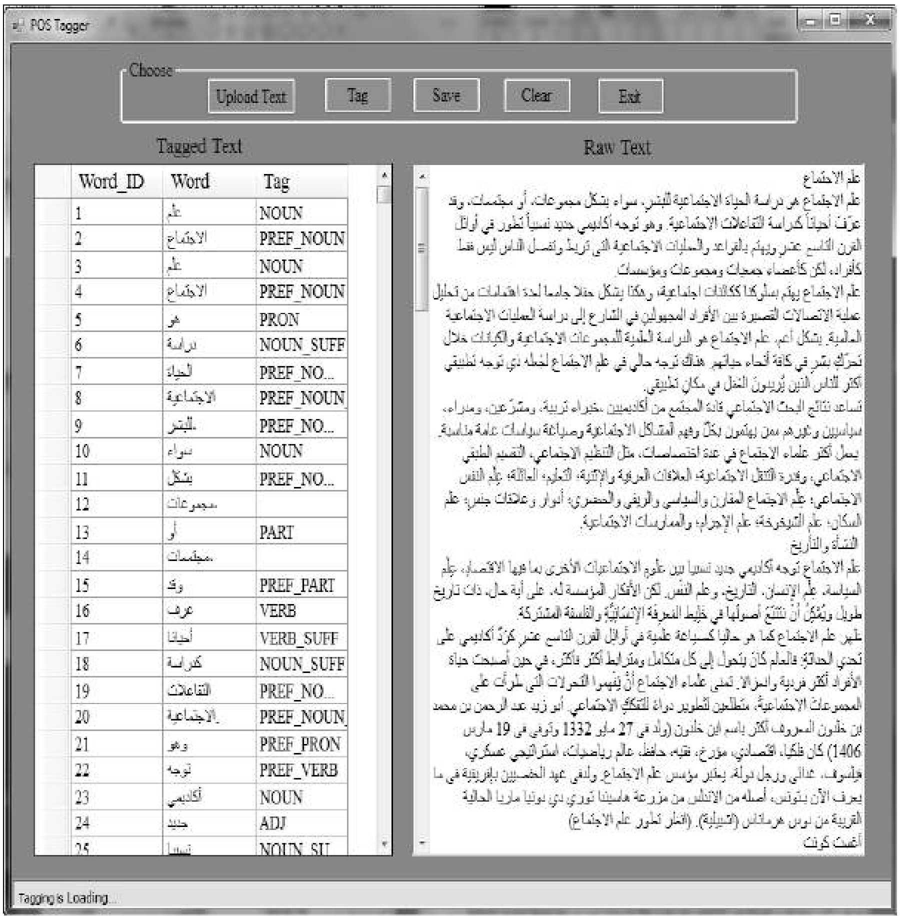
C. *Designing of the POS Tagger Application*

NLP, simply put, is to make computers understand and generate human language. Therefore, it requires both linguistic knowledge and programming skills[14]. So, the implementations of all system stages were done in C# programming language and the final results of the tagged text were automatically extracted in an access database file.

In spite of the fact that statistical models are less accurate than rule-based models, most existing POS analyzers have been based on a probabilistic model, because these systems are very robust and can be automatically trained. The limitations of the rule-based taggers are that they are non-automatic, costly and time-consuming [18].

The researcher presents a rule-based POS tagger which automatically infers rules from a training corpus. Thereby it avoids most of the limitations of traditional rule-based taggers in this system. The interface of built APOS is as shown in Fig. 3:

[14] http://nlpdotnet.com/



Word_ID	Word	Tag
1	علم	NOUN
2	الاجتماع	PREF_NOUN
3	علم	NOUN
4	الاجتماع	PREF_NOUN
5	هو	PRON
6	دراسة	NOUN_SUFF
7	الحياة	PREF_NO...
8	الاجتماعية	PREF_NOUN
9	للبشر.	PREF_NO...
10	سواء	NOUN
11	بشكل	PREF_NO...
12	مجموعات،	
13	أو	PART
14	مجتمعات،	
15	وقد	PREF_PART
16	عرف	VERB
17	أحيانا	VERB_SUFF
18	كدراسة	NOUN_SUFF
19	التفاعلات	PREF_NO...
20	الاجتماعية.	PREF_NOUN
21	وهو	PREF_PRON
22	توجه	PREF_VERB
23	أكاديمي	NOUN
24	جديد	ADJ
25	نسبيا	NOUN_SU

علم الاجتماع
علم الاجتماع هو دراسة الحياة الاجتماعية للبشر، سواء بشكل مجموعات، أو مجتمعات، وقد
عرّف أحياناً كدراسة التفاعلات الاجتماعية. وهو توجه أكاديمي جديد نسبياً تطور في أوائل
القرن التاسع عشر ويهتم بالقواعد والعمليات الاجتماعية التي تربط وتفصل الناس ليس فقط
كأفراد، لكن كأعضاء جمعيات ومجموعات ومؤسسات
علم الاجتماع يهتم بسلوكنا ككائنات اجتماعية، وهكذا يشكل حقلا جامعا لعدة اهتمامات من تحليل
عملية الاتصالات القصيرة بين الأفراد المجهولين في الشارع إلى دراسة العمليات الاجتماعية
العالمية. بشكل أعم، علم الاجتماع هو الدراسة العلمية للمجموعات الاجتماعية والكيانات خلال
تحرّك بشر في كافة أنحاء حياتهم. هناك توجه حالي في علم الاجتماع لجعله ذي توجه تطبيقي
أكثر للناس الذين يُريدون العمل في مكان تطبيقي.
تساعد نتائج البحث الاجتماعي قادة المجتمع من أكاديميين، خبراء تربية، ومشرّعين، ومدراء،
سياسيين وغيرهم ممن يهتمون بحلّ وفهم المشاكل الاجتماعية وصياغة سياسات عامة مناسبة.
يعمل أكثر علماء الاجتماع في عدة اختصاصات، مثل التنظيم الاجتماعي، التقسيم الطبقي
الاجتماعي، وقدرة التنقل الاجتماعية، العلاقات العرقية والإثنية، التعليم، العائلة، عِلْم النفس
الاجتماعي، عِلْم الاجتماع المقارن والسياسي والريفي والحضري، أدوار وعلاقات جنس، علم
السكان، علم الشيخوخة، علم الإجرام، والممارسات الاجتماعية.
النشأة والتاريخ
علم الاجتماع توجه أكاديمي جديد نسبيا بين علوم الاجتماعيات الأخرى بما فيها الاقتصاد، عِلْم
السياسة، عِلْم الإنسان، التاريخ، وعلم النفس. لكن الأفكار المؤسسة له، على أية حال، ذات تاريخ
طويل ويُمْكِنُ أَنْ تتتبّع أصولها في خليط المعرفة الإنسانيّة والفلسفة المشتركة.
ظهر علم الاجتماع كما هو حاليا كـصياغة علمية في أوائل القرن التاسع عشر كردٍّ أكاديمي على
تحدي الحداثة: فالعالم كان يتحول إلى كل متكامل ومترابط أكثر فأكثر، في حين أصبحت حياة
الأفراد أكثر فردية وانعزالا. تمنى علماء الاجتماع أن يفهموا التحولات التي طرأت على
المجموعات الاجتماعية، متطلعين لتطوير دواء للتفكك الاجتماعي. أبو زيد عبد الرحمن بن محمد
بن خلدون المعروف أكثر باسم ابن خلدون (ولد في 27 مايو 1332 وتوفي في 19 مارس
1406) كان فلكيا، اقتصاديا، مؤرخ، فقيه، حافظ، عالم رياضيات، استراتيجي عسكري،
فيلسوف، غذائي ورجل دولة. يعتبر مؤسس علم الاجتماع ولد في عهد الحفصيين بإفريقية في ما
يعرف الآن بتونس، أصله من الأندلس من مزرعة هاسيندا توري دي دونيا ماريا الحالية
القريبة من دوس هرماناس (إشبيلية). (انظر تطور علم الاجتماع)
أغست كونت

Tagging is Loading...

Figure 5: APOS tagger interface

The POS tagger is tested on data that consists of 150,000 words collected from different sources that were tagged manually according to Arabic context rules; because Arabic grammatical structures by themselves are rather useless. Like road signs, grammatical structures take on meaning only if they are situated in a context and in connected discourse [19]. Fig. 3 shows the APOS workflow:

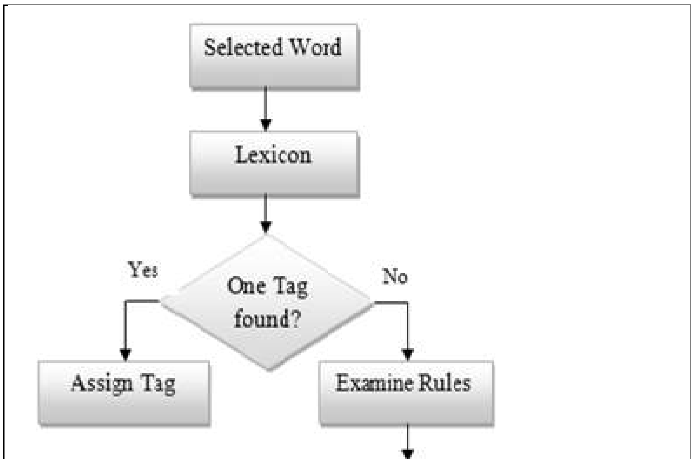


Figure 6: The APOS workflow

4 CONCLUSION

The purpose of building POS tagger is to assign part of speech tags automatically to words reflecting their syntactic category. POS tagger is a system that uses various sources of information to assign possibly unique POS to words. Automatic text tagging is an important step in discovering the linguistic structure of large text corpora. This project will make using Arabic language easier for native Arabs or non-Arabic speakers. It will serve search engines, database engines, Information Extraction applications and any other applications of Artificial Intelligence (AI) that makes use of Arabic Language Processing. POS tagger is one of the basic tools and components necessary for any robust NLP infrastructure of a given language.

Before building the POS tagger, some steps were taken: a) In the process of tagging a corpus, it is important, to distinguish a "tag-set"; a group of symbols represents various parts of speech, from tagger software program that inserts the particular tags making up a tag-set. This distinction is important because tag-sets differ in the number and types of tags that they contain, and some taggers can insert more than one type of tag-set.

The tag-set comprises of 5 main tags. This tag-set follows EAGLS guidelines, b) some data was compiled to be used as the testing corpus. It was compiled according to the criteria of corpus linguistics. Afterwards, some parts of the data have been disambiguated manually to be a training data to test the POS tagger, c) AlKhalil Arabic morphological analyzer was used to analyze the collected corpus and to obtain a lexicon containing different solutions for the corpus tokens. The tags were handled to agree with the new tag-set, d) setting Arabic rules is one of the most challenging steps of this project. These rules help in words disambiguation. Finally, the tagger has passed through three stages of tests; the rules were improved to enhance the tagging and disambiguation results. After comparing the results with the manually analyzed data, the percentage of the correct solutions was 84%.

REFERENCES

[1] Ch. Samuellsson (2003), *inMitkov R. (ed.)* “The Oxford handbook of Computational Linguistics. “ Oxford University Press, P358.

[2] M. Hadni, A. Ouatik, A. Lachkar and M. Meknassi (2013), “IMPROVING RULE-BASED METHOD FOR ARABIC POS TAGGING USING HMM TECHNIQUE. “ Sidi Mohamed Ben Abdellah University, Morocco.

[3] Y. Elhadj (2009), “Statistical Part-of-Speech Tagger for Traditional Arabic Texts.” *Journal of computer science 5 (11):794-800, ISSN 1549-3636*, science publications.

[4] Sh. Khoja (2001), “APT: Arabic part-of-speech tagger. *“in a Proc. of the Student Workshop* at the. 2nd Meeting of the NAACL.

[5] J. Haywood and H. Nahmad (1962), *A new Arabic grammar*, Lund Humphries Publishers Ltd.

[6] T. Yamina (2005), “Tagging by Combining Rules-Based Methods and Memory-Based Learning. *“ In Proc. of world academy of science, engineering and technology*, Volume 6 June.

[7] A. Diab, K. Hacioglu and D. Jurafsky (2004), “Automatic tagging of Arabic text: from raw text to base phrase chunks. *“ In 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference* (HLT-NAACL04, 74 - 2 self.

[8] M. Rashwan, E. Khalil, A. Rafea (2009); “Comparison between two Arabic tagsets, *“International Conference on Natural Language Processing and Knowledge Engineering*, NLP-KE.

[9] T. Buckwalter (2002), “Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, “University of Pennsylvania, LDC Catalog No.: LDC2002L49.

[10] A. Aliwy (2013), “ Arabic Morphosyntactic Raw Text Part of Speech Tagging System, “ Faculty of Mathematics, Informatics and Mechanics, University of Warsaw.

[11] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. BEBAH and M. SHOUL (2010),” Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts. “*In International Arab Conference on Information Technology*.

[12] S. AlGahtani, W. Black, & J. McNaught (2009), “Arabic part-of-speech tagging using transformation-based learning. “ *In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo (pp. 66-70).

[13] M. Altabba, A. Al-Zaraee, and M. Shukairy (2010),” An Arabic morphological analyzer and part-of-speech tagger. “Doctoral dissertation, Master’s thesis, Arab International University, Damascus, Syria.

[14] H. Rabiee (2011), “Adapting Standard Open-Source Resources to Tagging a Morphologically Rich Language: A Case Study with Arabic. “ *InRANLP Student Research Workshop (pp. 127-132), September 2011*.

[15] K. Ryding (2005), “A reference grammar of modern standard Arabic. “ Cambridge University Press.

[16] A. Feldman (2008), “TagsetDesign,InflectedLanguages, and N-gram Tagging, *“The Linguistics Journal* Vol. 3 (No. 1).

[17] C. Rizzo (2010), *Getting on with corpus compilation: from theory to practice*, ESP World, Issue 1 (27), Volume 9, 2010, http://www.esp-world.info.

[18] B. Megyesi (1998), “Brill’s POS tagger with Extended Lexical Templates for Hungarian. “ Department of Linguistics, Stockholm University, Sweden.

[19] M. Jiyad, (2006), “A Hundred and One Rules!, “ A Short Reference for Arabic Syntactic, Morphological & Phonological Rules for Novice & Intermediate Levels of Proficiency.

Hadeer Abdelrazik: *Senior Corpus Developer, Arabic Computational Linguistics Center Bibliotheca Alexandrina*

Bachelor of Arts from the faculty of Arts, Phonetics Department, Alexandria University in 2005. She obtained pre-masters certificate, Phonetics Department, Alexandria University in 2006, and she has been studying to complete the master degree in computational linguistics since 2009 up to now. She also participated with a team in building a tool for morphological generation and analysis of Arabic roots with excellent degree (field study). Her main areas of interest are concerned with corpus work, morphological, semantic and syntactic analysis in specific and natural Language Processing (NLP) in general. She has been working at Bibliotheca Alexandrina as a corpus developer since 2006 up to now. She obtained a certificate for participation in “The first annual forum for graduate students at the Faculty of Arts, University of Alexandria” in 2012.

Dr. Sameh Alansary*: Director of Arabic Computational Linguistics Center* Bibliotheca Alexandrina

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

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

بناء نظام حاسوبي لوسم أقسام الكلمة العربية لتطبيقات المعالجة الآلية للغة الطبيعية

هدير عبد الرازق[1]، سامح الأنصاري[2]

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر

[1]hadeer.abdelrazik@bibalex.org
[2]sameh.alansary@biballex.org

ملخص— يتركز هذا البحث على بناء واسم لغوي صرفي لوسم كلمات اللغة العربية والذي يعد من أهم العناصر الأساسية في المراحل المتقدمة لتحليل الذخائر والمدونات العربية. ويجب قبل بناء الواسم المرور بعدة خطوات من أهمها: تجميع ذخيرة لغوية تمثل اللغة العربية المعاصرة، وتشكيل أوسمة جديدة لوسم كلمات الذخيرة لتناسب كل أشكال الكلمات الموجودة، والبحث عن وتجميع قواعد عربية من مختلف الكتب النحوية العربية، وأخيرًا تصميم الواسم الصرفي كأداة لوسم الكلمات.

إن الهدف الأساسي من بناء الواسم الصرفي هي إلحاق وسم كل كلمة تبعًا لموقعها في الجملة آليا. أما عن النتائج التي تم الحصول عليها من عملية الوسم فيمكن استخدامها في عدة تطبيقات أخرى لخدمة معالجة اللغة الطبيعية. وبعد تقييم هذا الواسم الصرفي واختباره على عينة من الكلمات فإن النتائج التي تم الحصول عليها 84%.

بِنَاءُ بَنْكٍ شَجَرِيٍّ نَحْوِيٍّ لِلُّغَةِ الْعَرَبِيَّةِ الْفُصْحَى الْمُعَاصِرَةِ
(لُغَةُ الصّحافةِ الْإِلِكْتُرُونِيَّةِ الْمِصْرِيَّةِ نَمُوذَجًا)(*)

أحمد روبي محمد

*باحث ماجستير في قسم علم اللغة والدراسات السامية والشرقية

*كلية دار العلوم، جامعة الفيوم، مصر.

Ahmedruby757@yhoo.com
aruby@istnetworks.com

ملخص:

يعدُّ البنك الشجري النحوي موردًا هامًّا لبناء التطبيقات الإحصائية لمعالجة اللغات الطبيعية NLPكما يعدُّ أيضا أداة للبحث في الظواهر اللغوية التي تصفُ الواقع اعتمادًا على مجموعة من النصوص التي تمثل ذلك الواقع اللغوي. فالبنك الشجري النحوي العربي Syntactic Arabic TreeBankمجموعة من التحليلات النحوية لجمل عربية مستقاة من موقع إسلام أون لاين.

ويختلف البنك الشجري SATBعن البنوك الشجرية الأخرى من حيث المعلومات اللغوية وطريقة تمثيلها؛ وذلك لطبيعة الهدف المنشود من بناء محلّلٍ نحويٍّ يقوم على الوظائف النحوية الدلالية؛ لتحقيق المهمة الأساسية للتحليل النحوي الآلي، وهي توفير المعطيات اللازمة للتحليل اللغوي الأعمق،للفهم الأتوماتي للنصوص اللغوية،واستخدمتتمثيلَ بِنية العبارة Phrase structure representationفي التحليل مع مراعاة تحديد العلاقات النحوية لكل وحدة Token في الجملة؛وذلك لسببين رئيسيين: الأوّل سهولة استخلاص السمات المميزة Features Extractionللوحداتالمكونة للجملة، والآخر الحصول على سمات هجينة بين الاعتمادية وهيكلة السين البارية X-barمن حيث تحديد العلاقات النحوية التبعية بين المركبات ورؤوسها ومكمّلاتها.

وتقدّم هذه الورقة توصيفا لمراحل البناء المتسلسلة ثم مقارنتها بغيرها من البنوك الشجرية الأخرى، ثم مدى الاستفادة من استخدام البنك الشجري في بناء المحلل النحوي.

الكلمات المفتاحية:

البنوك الشجرية – التحليل النحوي – العنونة النحوية –المحلل النحوي–معالجة اللغات الطبيعية .

1- مقدمة:

لقد أصبحت هندسة اللغة العربية من أهم مجالات تقنية المعلومات والاتصالات، ولقد دعا ذلك إلى تعاظم الحاجة إلى المعالجة الآلية للغة العربية بمستوياتها المختلفة الصوتية والصرفية والنحوية والدلالية، بسبب الفيض المعلوماتي من جانب وإلحاقها بالثورة التكنولوجية

(*)بحث تكميلي لنيل درجة الماجستير في علم اللغة بإشراف: الأستاذ الدكتور فريد عوض حيدر أستاذ علم اللغة بكلية دار العلوم-الفيوم، ونائب رئيس الجامعة لشؤون البيئة وتنمية المجتمع مشرفا رئيسيا، والأستاذ الدكتور محسن رشوان أستاذ الاتصالات والإلكترونيات بكلية الهندسة جامعة القاهرة مشرفا مشاركا، والدكتور خالد أبو غالية مدرس علم اللغة بكلية دار العلوم-الفيوم مشرفا مشاركا.

من جانب آخر، وتساعد الأنظمة الخبيرة في مجال الذكاء الاصطناعي على تمثيل اللغة بكافة مستوياتها وفق قواعدَ للمعرفة وقواعد تجريبية على نحو يستطيع الحاسوب التعامُلَ معها.

تمثل معالجة النحو آليا –حاليًا على الأقلّ- صلب اللسانيات الحاسوبية، وتشهد ساحتها أقصى درجات الامتزاج بين اللسانيات والحاسوبيات، بجانب ذلك فالمعالجة النحوية الآلية هي قنطرة الوصل التي تعبر خلالها مسارات الاقتراض المتبادل بين علوم اللغة وعلوم الحاسب.[5] ويلزم لمعالجة النحو آليا وضعقواعده في صياغة رسمية مكتملة وفقا للنموذج النحوي المتبع[5]، وانطلاقا من هذا سعت المؤسسات الأوروبية والأمريكية المعنية بحوسبة اللغات إلى بناء مدونات معنونة نحوياannotated Parsedcorporaللاستفادة من أساليب التعلم الإحصائي Machine Learning في بناء نماذج إحصائية لهذه المدونات.

وتعدُّ مجموعة التحليلات النحوية للجمل المعدّة يدويًّا أو ما تسمّى بالبنوك الشجريةTreebanksمصدرا مهمًّا لِبناء المحلِّلات الإحصائية وتقييمها بشكل عامّ، واستخدمت هذه التحليلات أو البنوك الشجرية الغنية بالتذييل التفصيلي في العديد من التطبيقات منها، تجزئة النصوص، والتشكيل الآلي، والعنونة بأقسام الكلام، وفك اللبس الصرفي، وتحديد أبنية المركبات، والعَنونة الدلالية، وتُستخدم أيضا هذه التحليلات لبناء المعايير الذهبية Gold Standards؛لتقييم دقة الأنظمة المحوسبة وقياسها،وكذلك لإيجاد أوجه التشابه والاختلاف في نتائج التحليل مبينة الحالات التي تتفق عليها والتي تختلف فيها الأنظمة المحوسبة.

ويعود تاريخ أوَّل محاولة لبناء مُدوَّنة مُعَنونة نحويا-للغة الإنجليزية- إلى ثلاثة عقود مضت، حيث كانت في الأغلب قائمة على الطرق اليدوية، وكان هدفها هو محاولة إيجاد منهج شامل للترميز يصلح للتطبيقات المتنوعة[6]،ثم تبنت المؤسسات المعنية بحوسبة اللغات فكرة إنشاء مشروع البنوك الشجرية وكان على رأسها مركز اتحاد البيانات اللغوية(Linguistic Data Consortium) بجامعة بنسلفانيا الذي تولى مهمة إنشاء مشروع البنك الشجري للغة الإنجليزية (PATB) عام 2001م، ثم بدأ يتوسع ليشمل اللغة الصينية والعربية ولغات أخرى.

وهناك مجهودان كبيران في المدوّنات اللغوية العربية المعُنونة نحويًّا وهما: بنك بنسلفانيا الشجري(PATB)، وبنك براغ الشجري الاعتمادي (PADB)،وهذان المجهودان قدما تمثيلات لغوية معقدة ومعلومات لغوية غنية التفاصيل؛ لتسمح هذه المعلومات بالبحث في التطبيقات العامة لمعالجة اللغة الطبيعية، إلّا أن كثيرا من هذه المعلومات غير مستخدم حاليا في التطبيقات العربية[15]، ويتطلب هذا النوع من المدوّنات الكثير من الوقت والجهد البشري.وقدم مؤخرا مركز أنظمة التعلم الحاسوبي(Center For Computational Learning Systems) بجامعة كولومبيا بنكًا شجريًّا نحويًّا ثالثًا (CATiB) بهدف تسريع عملية الترميزannotationوتجنب المعلومات التي لا فائدة منها، وكان غرضه الترجمة الآلية[15].

وفي هذا الإطار قدم الباحث بناءً نحويًّا آخر (SATB)، بهدف بناء محلل نحوي إحصائي يقوم على تحديد المعاني الوظيفية النحوية لوحداتالجملة Tokensمن خلال تحديد العلاقات التركيبية النحوية؛ لما لأهميتها في التفهم الآلي للنصوص، وتحديد استقلالية الجملة من خلال العلاقات النحوية عن طريق الربط والارتباط بين مكونات الجملة،وبهذا تختلف طبيعة بناء البنك الشجري (SATB) عن بناء البنوك الشجرية الأخرى من حيث المعلومات اللغوية وطريقة تمثيلها، أما المعلومات اللغوية فتتمثل في تجزئة النصوص، والعنونة بأقسام الكلام، والمحتوى النحوي من حيث المركبات والعلاقات النحوية، أما تمثيلها فكان باستخدام النظرية النحوية الوصفية Descriptive theoryمستخدمة القوالب التخطيطية ل السين البارية X-[11Bar]لعرض أبنية الجمل من خلال تمثيل شجرة بنية العبارة Phrase Structure Tree Representation.

وفي هذه الورقة، سأقدم توصيفًا لمراحل بناء البنك الشجري النحوي، مع تزويدها بالمعلومات الإحصائية،والأدوات الحاسوبية المساعدة، ثم عرض مقارنة بينه وبين البنوك الشجرية الأخرى،واستخدامه، والأعمال المستقبلية والخاتمة.

2- اختيار المدونة اللغوية:

فالمدونة اللغوية مجموعة من النصوص يمكن التعامل معها آليا والتحكم في بياناتها ومدخلاتها بالإضافة أو الحذف أو التعديل من خلال قواعد البيانات التي تتعامل مع هذه النصوص[1]، وتهدف هذه المدونة إلى وصف الواقع اللغوي اعتمادا على مجموعة من النصوص التي تمثل ذلك الواقع، أو تأكيد فرضيات قائمة حول لغة معينة[1].

واختيار المدونة اللغوية المعنية بالتحليل يعتمد على الهدف المنشود منها،وهذا الهدف يتعلق بالزمان والمكان والمستوى اللغوي المطلوب، وفي معظم الحالات تتكون هذه النصوص من الصحف أو الجرايد أو المجلّات المعاصرة، وعلى هذا فاخترت صحيفة إسلام أون لاين– للمدة الزمنية المحددة من 2002 إلى 2010 - مصدرا للمدونة اللغوية المعنيةبالتحليل[*]لعدة أسباب منها:

-إسلام أون لاين صحيفة إلكترونية، والصحافة الإلكترونية أصبحت من أهم الوسائل الإعلامية المعاصرة، وتنتشر بين كافة الشرائح المجتمعية بصورة متسارعة.

-التزامها بالكتابة العربية الفصحى غير الهجينة بالعامية.

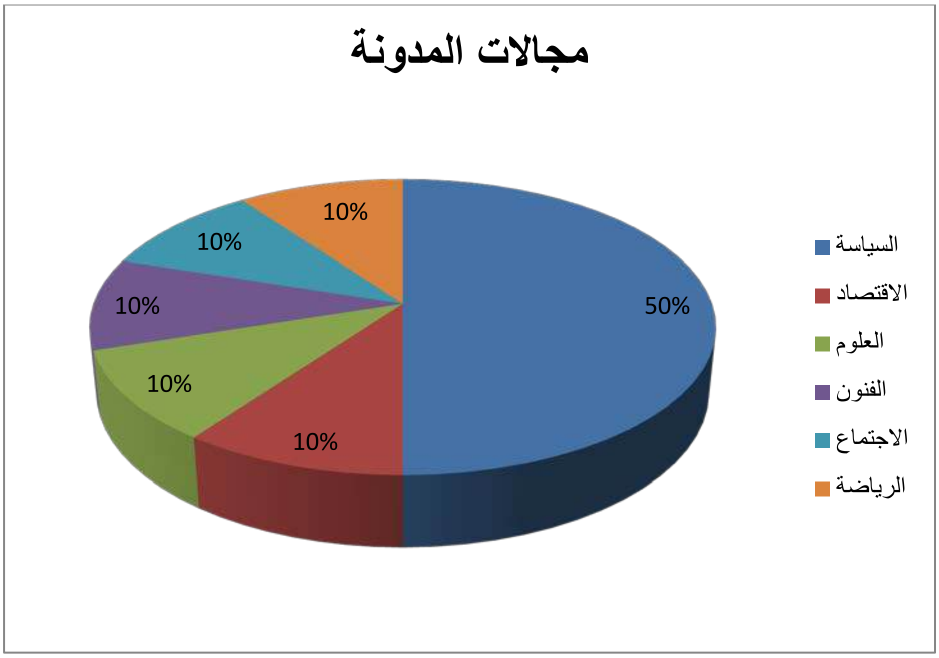
-معظم المحررين العاملين بها من اللغويين، وهذا يعكس نقاء المفردات والأساليب اللغوية وخلوها من الأخطاء الإملائية والنحوية.

-كثرة عدد زوّار موقع الصحيفة عن المواقع الأخرى.

ويخضع اختيار النصوص – وفقا لبناء المدونة اللغوية – على نظرية العينات الإحصائية Statistical Sampling Theory، ومن خلالها يقوم صنّاع المدونة باختيار عينة من النصوص التي تتفقوأهدافهم البحثية سواء أكانت عينة عشوائية Probabilistic Samples(Random)، أم عينة غير عشوائية-Non-Probabilistic Samples (Non . [1Random)]

وتم تصنيف هذه النصوص حسب المجالات التي تنتمي إليها، وبلغ عدد كلماتها 100,000كلمة وبلغ عدد الواحدات123 ,000وحدة، وعدد المقالات 98 مقالا، وتم تحريرها في ملفات نصية وتسمية كل ملف باسم المجال الذي ينتمي إليه، ويوضح الشكل رقم(1) تصنيف المدونة المستخدمة وحجمها.

[*]هذه النصوص مأخوذة من مدونة الشركة الهندسية لتطوير النظم الرقمية RDI والتي يبلغ عدد كلماتها مليون كلمة.



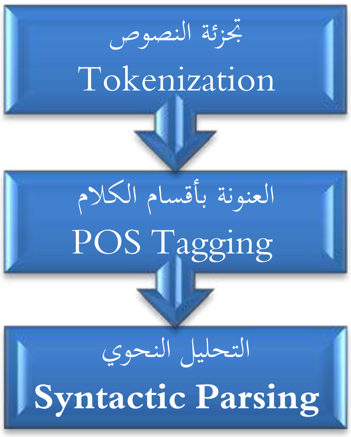
الشكل رقم (1)
تصنيف المدونة المستخدمة إلى مجالات.

3- تجزئة النصوص**Tokenization**:

تعد عملية تجزئة النصوص- آليا- مرحلة أساسية قبل المعالجة الآلية للنصوص اللغوية، وفي ظل التطور السريع لمعالجة اللغات الطبيعية أصبحت عملية تجزئة النصوص خطوة حاسمة في تنقيب النصوص واستخلاص المعلومات. وتقول الحكمة الشائعة في معالجة اللغة الطبيعية إن تجزئة النص العربي إلى كلمات من خلال التجريد وتقليص الاحتمالات الهجائيةOrthographic Normalizationمفيد للعديد من التطبيقات مثل نمذجة اللغة واسترجاع المعلومات والترجمة الآلية الإحصائية. [15]

وتعمل معظم تطبيقات معالجة اللغات الطبيعية مثل معنونات أقسام الكلام Part-Of-Speech taggersوالمحلّلات النحوية Syntactic Parsersوالتجذيع Stemmingعلى النصّ المقسّم إلى أجزاء/عناصر، وتشمل هذه الأجزاء الكلمات والأرقام وعلامات الترقيم والرموز، وغيرها من الواحدات المكوّنة للنصّ.

وتتوقف دقة هذه التطبيقات وفقا لدقة عملية تجزئة النصوص؛لأنها مرحلة أولية أو أساسية في التحليل اللغوي تقوم عليها بقية المراحل التحليلية، فهي مطلب أساسي للتحليل النحوي الذي ترسي دعائمه تلك الوحدات النحوية، ويوضح الشكل التالي رقم (2) مراحل البناء الشجري.



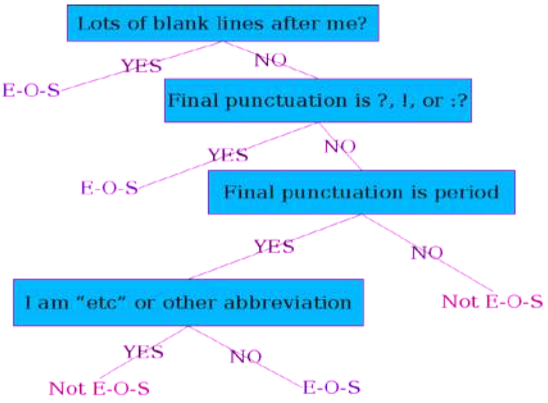
الشكل رقم (2)
مراحل بناء البنك الشجري

وتجزئة النصوصTokenizationهي عملية تقسيم تُجرى على النص لتقسيمه إلى كلمات أو وحدات أو جمل،حتى تتمكن آليات معالجة اللغات الطبيعية من معالجة النصوص حسب ما تهدف إليه طبيعة التطبيقات المنشودة[8]، وتتم عملية تجزئة النصوص –بالنسبة للغة العربية- على ثلاثة مستويات:

1) التجزئة على مستوى الجملة

وينجز التقسيم -في تحديد أبعاد الجملة في المدونة اللغوية موضوع الدراسة- حسب الإسناد والتركيب التام المفيد وما بين الجمل من علاقات الربط بواسطة أدوات الاستئناف والعطف[3]، فكانت علامات الترقيم خير سبيل لتحديد هذه الأبعاد الجملية -تحديدا شكليًّا- في النصّ.

وتتم عملية تجزئة الجمل أو تحديدها -آليا- في النص من خلال وجود فواصل الأسطر وعلامات الترقيم، وتستطيع آلية تجزئة النصوص Tokenizerأن تحدد حدود الجملة من خلال علامة الترقيم النقطة(.) التي توضع في نهاية الجملة، وكذلك بعض الجمل التكميلية مثل الجملة الاستفهامية التي تنتهي بعلامة استفهام(؟)، والجملة التعجبية التي تنتهي بعلامة تعجب(!)، إلّا أن هذه الآلية التي تعتمد على هذه المدخلات تواجه عدة مشكلات في تحديد حدود الجمل بسبب تعدد وظائف بعض علامات الترقيم مثل النقطةالتي توضع في نهاية الجملة، وتوضع أيضا بين الاختصارات مثل أ.د، ص.د.ب وغيرها، وكذلك الفاصلة (,) التي تعد ملمحا مميزا للفصل بين الوحدات أوالمكونات، توضع أيضا في حال الأرقام العشرية مثل 11,4، ومع ذلك يتغلب اللسانيون الحاسوبيون على هذه المشكلات عن طريق بناء مصنفات ثنائية binary classifiersمثل أشجار القرار Decision Trees، وتعتمد هذه التقنية على تحديد خواص العلامات وإعادة ترتيبها في صورة شجرية متدرجة، حيث يشكل التنسيق بين هذه الخواص تصورا يؤدي إلى تحديد الفئات من حيث التجزئة أو عدم التجزئة(ع)، ويبيّن الشكل التالي رقم (3) الخواص التي يعتمد عليها مصنِّف أشجار القرار في تحديد الكلمة النهائية للجملة.


Lots of blank lines after me?
YES
NO
E-O-S
Final punctuation is ?, !, or :?
YES
NO
E-O-S
Final punctuation is period
YES
NO
Not E-O-S
I am "etc" or other abbreviation
YES
NO
Not E-O-S
E-O-S

الشكل رقم (3)

تحديد الكلمة النهائية في الجملة باستخدام أشجار القرار.

واستخدمت في البحث آلية تجزئة النصوصTokenizerالمدرجة في المحلل النحويStanford parserالتابع لفريق معالجة اللغات الطبيعية بجامعة ستانفورد، إلّا أنه لم يستخدم هذه التقنيات في إزالة لبس حدود الجملة، ونتج عن ذلك أخطاء في تحديد حدود الجمل، قمت بتصحيحها يدويا، ومن هذه الأخطاء:

-يفصل في الحالات الآتية مثل: د.كلثم، 2,4 سم.

-يفصل في حالة تعدد النقاط الدالة على الكلام المحذوف ...، ويعتبر كل نقطة نهاية جملة.

-إذا لم ينته السطر بعلامة(.) أو (؟) يضمه إلى السطر التالي له.

2) التجزئة على مستوى الوحدات/العناصر الرئيسية

فالعنصر اللغوي Token هو أصغر وحدة نحوية، يمكن أن تكون كلمة أو جزءا من الكلمة، أو تعبيراصطلاحي، أو مركبا، أو علامة ترقيم،ومادامت العناصر اللغوية الرئيسية هي الجزء الملموس من التحليل فيمكن أن نطلق عليها أيضا"وحدات التحليل النحوي"[2]تلك الوحدات الرئيسية التي تعتبر عنصرا أساسيا في النص اللغوي، فالوحدة الرئيسة هي البناء اللغوي المتكامل سواء أكانت كلمة أو علامة أو رقما.[8]

فالتجزئة على مستوى الوحدات أو العناصر الرئيسية تشمل ثلاثة متسويات:

أ- الكلمة:

فالكلمة هي أصغر وحدة مستقلة في النص، ولعل أشهر من عرف الكلمة من علماء اللغة المحدثين هو العالم الأمريكي"بلومفيلد" Bloomfield، الذي قال " الكلمة أصغر وحدة حرة" ومعنى هذا أن الكلمة عنده هي أصغر وحدة لغوية يمكن النطق بها معزولة[2]، كما يمكن استعمالها لتركيب الجملة أو كلام، ويجب أن تتكون من مورفيم حر Free Morpheme على الأقل، وتعتبر عنصرا تحليليا بسيطا على المستوى النحوي في بعض الأنحاء رغم تركيبها مع مورفيمات أخرى.

ب- المركب غير الكلامي:

هو انضمام كلمة إلى كلمة فأكثر، وتكون بحكم المفرد نحويا ودلاليا مثل: عبد_الله، أبو_عيد، إسلام_ أون_ لاين، الصهيو_أمريكي، الجيو_ إستراتيجية.

ت- الرمز أو العلامة:

يشمل جميع الرموز المستخدمة في النص العربي، مثل علامات الترقيم والأرقام، وغيرها من الرموز.

وهناك فرق بين تجزئة النصوص للغات ذو النظام الألفبائي واللغات ذو النظام الفكري مثل الصينية، فعادة في اللغات ذو النظام الألفبائي يتم الفصل من خلال حدود الكلمات أي الفراغات أو المساحات البيضاء، أما في اللغات ذو النظام الفكري فهي لا تحتوي على معلومات حول حدود الكلمة؛ لذلك التجزئة أصعب بكثير من النظام الألفبائي.

وتعتمد آلية تقطيع النصوص Tokenizer على الفراغات البيضاء وعلامات الترقيم والأرقام كعلامات مميزة لتجزئة هذه الوحدات الرئيسية[9]، أما الوحدات التي تحتوي على كلمتين أو أكثر مثل عبدالله، قمت بوضع شرطة بدلا من الفراغات البيضاء لتصبح : عبد_الله.

3) التجزئة على مستوى الوحدات/العناصر الفرعية

يمكن أن نعرّف العنصر اللغوي أيضا بأنه" بناء لغوي يحدده مستوى التحليل" ، فنجد أن العنصر اللغوي الرئيسي قد يكون مكونا من مورفيم / عنصر فرعي واحد أو أكثر من مورفيم [2]، فعلى سبيل المثال، حيث يمكن للكلمة المفردة (العنصر الرئيسي) أن تشمل على ما يصل إلى أربع وحدات فرعية سواء سوابق أو لواحق[8].

وتتوقف حدود عملية تجزئة العناصر الرئيسية إلى عناصر فرعية إلى طبيعة الغرض من البحث، أي ما هي العناصر الفرعية المراد تجزأتها من العناصر الرئيسية؟

وللإجابة على هذا السؤال، نبيّن -أولا- أنواع المورفيمات اللصقية Concatenative Morphemes في اللغة العربية،فهناك ثلاثة أنواع من المورفيمات المتسلسلة وهي : الجذع(Stem) واللواصق(affixes) والزوائد(Clitics).

أ- فالجذع: هو أساس الكلمة بعد حذف الزوائد واللواصق منها، ووجوده ضروري لكل كلمة، ومن أمثلته: الجذع (كتب) الذي تكون عنه التركيب في(وسيكتبونها) والجذع (مكتب) في صيغة الجمع(المكتبات).

ب- اللواصق: هي مورفيمات تتعلق بجذع الكلمة، وهناك نوعان من اللواصق:

1) السوابق(Prefixes): وهي مورفيم يسبق الجذع في أوله ومن أمثلته: نون في الفعل المضارع في "نفعل-نعمل-نشكر".

2) اللواحق(Suffixes): وهي مورفيم يلحق الجذع في آخره ومن أمثلته: الواو والنون في جمع المذكر السالم في" المسلمون-العاملون".

ت- الزوائد:هي مورفيمات نحوية تكون مقيدة بكلمات أخرى، و تتعلق بجذع الكلمة بعداللواصق،وهناك نوعان من الزوائد:

1) الزوائد في بداية الكلمة (Proclitics)فهي تشبه اللواصق، ولكنها تختلف اختلافا واضحا عن اللواصق التي تمثل جزءا من الكلمة صوتيا وبنيويا، ومن أمثلتها:حروف العطف، وحروف الجر، والنداء.

2) الزوائد في نهاية الكلمة(Enclitics)وهي التي تعقب الكلمة، مثل الضمائر المتصلة.

وقد يكون القرار مربكا أحيانا في جعل المورفيم لاصقة صرفية أو زائدة نحوية، ومع ذلك نقول –عموما- أن اللواصق تحمل ملامح صرفية نحوية مثل(الزمن-الشخص-الجنس-العدد) بينما الزوائد تخدم الوظائف النحوية مثل(النفي-التعريف- العطف أو الجر).

وطبيعة ما نصبو إليه يجعلنا نقف أمام الوحدات النحوية (الزائدة بنوعيها) باعتبارها عنصرا مستقلا للتحليل النحوي، فهي عنصر تحليلي ذات علاقات نحوية نظمية بغيرها، وسنبيّن منهجنا في التجزئة من خلال الأمثلة التالية:

- للمدرسة ----> ل+المدرسة
- وسيكتبونهم ------>و+سيكتبون+هم
- مستشفاهم ------> مستشفى+هم
- سمائه -------> سماء+ه
- مكتبتنا ------> مكتبة+نا

واستخدمت آلية MADA+TOKANلتجزئة النصوصباتباع منهج ATBفي التجزئة كخطوة أولية للتحليل، ثم قمت بتعديل الخارج يدويا ليتناسب مع المنهج المقترح.

4- العنونة بأقسام الكلام POS Tagging:

تعد عنونة الأقسام الكلامية عملية أساسية من عملية التحليل اللساني حيث تهدف إلى استخلاص الانواع الكلامية وهي تلك السمات التي تحمل الخواص النحوية البدائية المميزة لكل كلمة منفردة بمعزل عن سياقها الإعرابي في النص محل الدراسة[7].

ومن البديهي أن الأقسام الكلامية لكلمات نص ما هي من أهم المدخلات الابتدائية لأي عملية تحليل نحوي لهذا النص، فلا نستطيع مثلا على الإطلاق أن نعرف أن كلمة(تحليل) في الجملة السابقة مضاف إليه دون أن نعرف أولا أنها اسم[7]؛ ولذلك تقوم المحللات النحوية بشكل عام على التحليل الصرفي للكلمات.

فالعنونة بأقسام الكلام هي مهمة تعيين السمات الصرفية والنحوية لكل وحدة في النص[10] عن طريق إلحاق كل مفردة بالنص برمز أو عدة رموز تشير إلى سماتها الصرفية والنحوية.ويمكن أن تكون سمات أقسام الكلام للغة العربية كبيرة جدابسبب غنى اللغة العربية صرفيا، ومع ذلك يفضل كثيرمن الباحثين العاملين في معالجة اللغة العربية العمل على مجموعات أصغر حجما؛ حتى يمكن التنبؤ بها بدقة عن طريق أساليب التعلم الإحصائي.

وتتم عملية ترميز الخصائص الصرفية للكلمة المحللة باستخدام مجموعة العناوين للخصائص الصرفية للكلمة المحللة POS Tag set ، ويمكن أن تصل السمات الصرفية للغة العربية نظريا إلى 330 ألف سمة، وقد اعتمد باكولتر على آلاف السمات (للنص غير المقطع)، بينما اعتمد على مجموعة من السمات(للنص المقطع) تصل إلى حوالي 500 سمة، وهذا ما استخدمه بنك بنسلفانيا الشجري.

واعتمدت على مجموعة العناوين الصرفية POS Tag set -وبلغ عدد سمات هذه المجموعة 62 سمة-التي اقترحها الباحث محمد عطية في رسالته للدكتوراه لتطوير آلية التشكيل، لما أراه مناسبا لعملية التحليل النحوي، كما تتميز بأنها ليست مستعارة من اللغات الأخرى، بل صممت خصيصا للغة العربية.وتم استخدام آلية Arab Tagger©(*)لعنونة الأقسام الكلاميةللمدونة اللغوية المعنية بالتحليل ثم مراجعة الخارج يدويا ، ومن أمثلة التحليل لمعنون الأقسام الكلامية كما في المثال التالي:

{(وَ); (NullSuffix) (Conj) (NullPrefix) }

(*)الاسم التجاري للمعنون العربي للأقسام الكلامية في الشركة الهندسية لتطوير النظم الرقمية**RDI**.

{(جَاءَ); (NullPrefix) (Active) (Verb) (Past) (NullSuffix) }

{(اَلْعُدْوَانُ); (Definit) (Noun) (NullSuffix) }

{(اَلْإِسْرَائِيلِيُّ); (Definit) (Noun) (NoSARF) (RelAdj) }

{(اَلْأَخِيرُ); (Definit) (Noun) (ExaggAdj) (NullSuffix) }

{(عَلَى); (NullPrefix) (Prepos) (NullSuffix) }

{(لُبْنَانَ); (NullPrefix) (Noun) (NoSARF) (Femin) (Single) (NullSuffix)) }

{(عَلَى); (NullPrefix) (Prepos) (NullSuffix) }

{(قَاعِدَةِ); (NullPrefix) (Noun) (SubjNoun) (Femin) (Single) }

{(اِسْتِمْرَارِ); (NullPrefix) (Noun) (NounInfinit) (NullSuffix) }

{(حِرَاكِ); (NullPrefix) (Noun) (NullSuffix) }

{(اَلْجَبْهَةِ); (Definit) (Noun) (Femin) (Single) }

{(اَلْمَفْتُوحَةِ); (Definit) (Noun) (ObjNoun) (Femin) (Single) }

{(فِي); (NullPrefix) (Prepos) (NullSuffix) }

{(اَلْجَنُوبِ); (Definit) (Noun) (ExaggAdj) (NullSuffix) }.

الشكل رقم (4)

تحليل POS باستخدام معنون أقسام الكلام ©Arab Tagger

5- التحليل النحوي Syntactic Parsing:

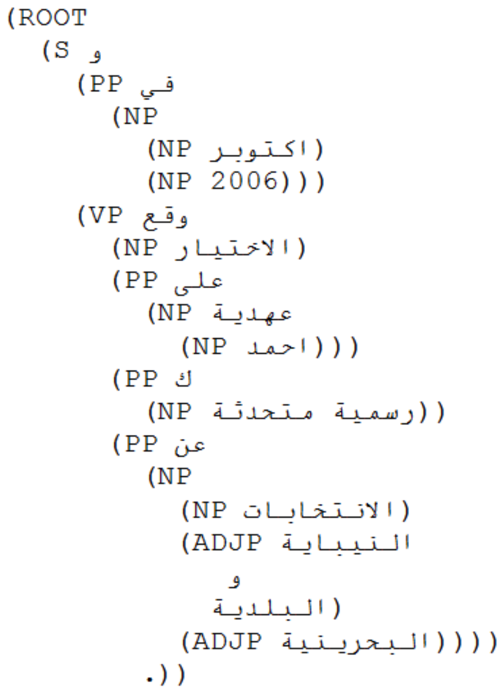
وتظل المهمة الرئيسية للتحليل النحوي الآلي، هي توفير المعطيات اللازمة للتحليل اللغوي الأعمق، ألا وهو الفهم الأتوماتي للنصوص اللغوية[5]، وتقوم هذه المرحلة على عدة خطوات أساسية في التحليل وهي التقويس النحوي أو التمثيل النحوي، والنظرية النحوية، والمحتوى النحوي.

1) التَّقوِيس النحوي Syntactic bracketing

هو نموذج رياضي ينظم الكلمات بطريقة متماسكة، بحيث يظهر العلاقات بين الكلمات في الجملة[13]، وهو ما يشبه صيغة باكوس نور BNF.واعتمده تشومسكي في إعادة كتابة القواعد النحوية الشكلية من خلال النحو المتحرر من السياق[12].

والنظام الرياضي المستخدم في الترميز النحوي للمدونةهو النحو المتحرر من السياق اعتمادًا على مخطط السين الباريةX-barمن حيث الرأس والمكمل والوصف، مع عدم الاعتماد على الفئات الفارغة، وهذا ما يعتمده محلل ستانفورد النحوي، وبايكل مع الاعتماد على الفئات الفارغة.

واعتمدت على آلية Stanford Parserكخطوة شبه آلية في تحليل جمل المدونة، والمدخل للمحلل نصوص مجزئة(Tokenized)، ثم قمت بتعديل نتائج التحليلعنطريق استخدام التعبيرات النمطيةRegular Expression، ونجد في الشكل التالي رقم (5) مثالا لجملة محللة باستخدام Stanford Parser.

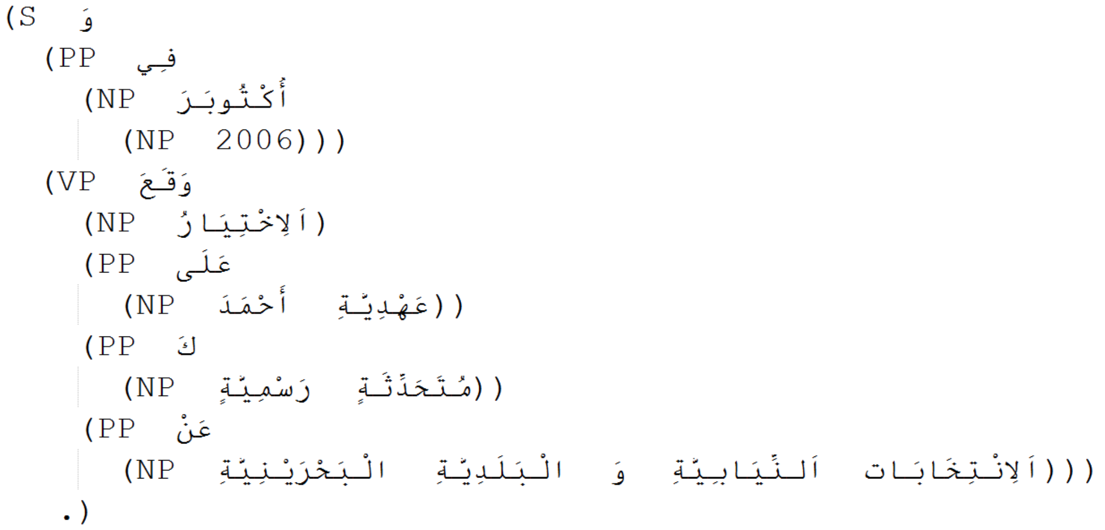


```
(ROOT
  (S و
    (PP في
      (NP
        (NP اكتوبر)
        (NP 2006)))
    (VP وقع
      (NP الاختيار)
      (PP على
        (NP عهدية
          (NP احمد)))
      (PP ك
        (NP رسمية متحدثة))
      (PP عن
        (NP
          (NP الانتخابات)
          (ADJP النيابية
            و
            البلدية)
          (ADJP البحرينية))))
    .))
```

الشكل رقم (5)

جملة محلل باستخدام Stanford Parser

وتم تعديل الخارج من محلل ستانفورد؛ ليتوافقَ مع النظرية النحوية من حيث المحمولوأبنية العواملPredicate-argument structureباستخدام التعبيرات النمطيةRegex، ويبين الشكل التالي رقم(6) الجملة معدَّلة من حيث المحمول وأبنية العوامل.



```
(S   وَ
  (PP   فِـي
    (NP   أُكْـتُـوبَـرَ
      (NP   2006)))
  (VP   وَقَـعَ
    (NP   اَلإِخْـتِـيَـارُ)
    (PP   عَـلَـى
      (NP   عَـهْـدِيّـةِ   أَحْـمَـدَ))
    (PP   كَ
      (NP   مُـتَـحَـدِّثَـةٍ   رَسْـمِـيّـةٍ))
    (PP   عَـنْ
      (NP   اَلإِنْـتِـخَـابَـات   اَلـنِّـيَـابِـيّـةِ   وَ   الْـبَـلَـدِيّـةِ   الْـبَـحْـرَيْـنِـيّـةِ)))
  .)
```

الشكل رقم (6) نتائج Stanford parser مجردة من POS بعد التعديل اليدوي.

2) الرموز الوظيفية **Functional Tags**:

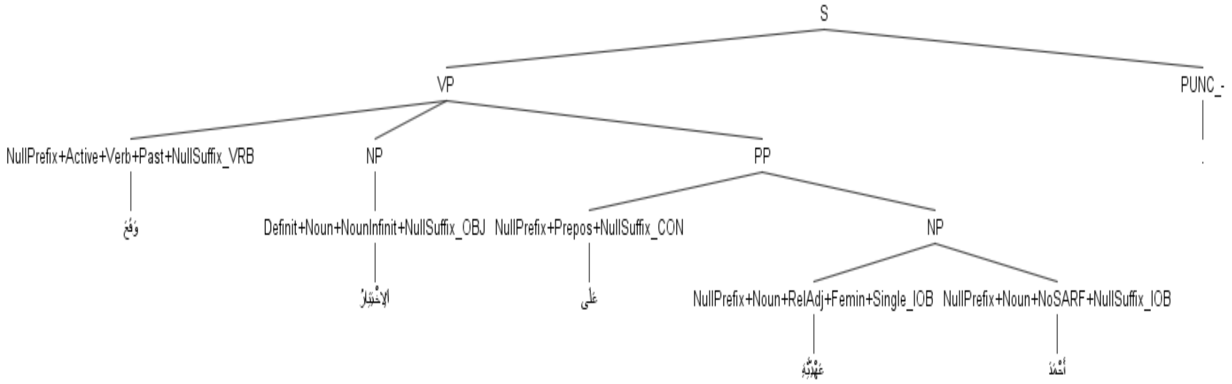
ولا يستغنى التحليل اللغوي عن البيانات التصنيفية والوظيفية؛ لأن المكونات مع نفس خصائصها التصنيفية يمكن أن تقع في علائق وظيفية أخرى، والعلائق الوظيفية نفسها يمكن أن تنطبق على مكونات أخرى مع اختلاف خصائصها التصنيفية[4]، وانطلاقا من الهدف المنشود جُعلت هذه الرموز الوظيفية لوحدات الجملة وليس مركباتها، وهذه الرموز مصنفة كما في الجدول رقم(1):

الجدول رقم (1)

الرموز الوظيفية المقترحة

المصطلح الإنجليزي	الرمز	المصطلح العربي	الفئة
Predicate	SBJ	المسند-	الوظائف النحوية Grammatical functions
Subject	PRD	المسند إليه-	
Topic	TPC	الموضوع-	
Object	OBJ	المفعولية-	
Verb	VRB	فعل-	
Tamyiyz	TMZ	تمييز-	
Modifier	MOD	وصف-	
IDafa	IDF	إضافة-	
Indirect object	IOB	مفعول غير مباشر-	
Locative Temporal	TMB	الزمان والمكان-	الأدوار الدلالية semantic roles
Purposive	PRB	السببية-	
movement	MOV	الحركة-	
direction	DIR	الاتجاه-	
IDah	IDH	الإيضاح-	
Depending	DPN	المتلازم-	التلازم التركيبي Phrase depended
Verb_depend	VRD	فعل_متلازم-	
Conjunction	CON	ربط-	
For punc and sym	-	-	

وباستخدام المراحل البنائية الثلاث، يظهر التحليل كما بالشكل رقم(6)، مستخدما أداة عرض التحليل الشجريstanford-tregex.


S
VP
PUNC_-
NullPrefix+Active+Verb+Past+NullSuffix_VRB
NP
PP
.
وَقَعَ
Definit+Noun+NounInfinit+NullSuffix_OBJ
NullPrefix+Prepos+NullSuffix_CON
NP
الاِخْتِيارُ
عَلَى
NullPrefix+Noun+RelAdj+Femin+Single_IOB
NullPrefix+Noun+NoSARF+NullSuffix_IOB
عَهْدِيَّةٍ
أَحْمَدَ

الشكل رقم (6)

التحليل الشجري باستخدام أداة العرض الشجري.

6- أوجه التشابه والاختلاف بينSABT والبنوك الشجرية:

ويحتوي الجدول التالي رقم(2) على أوجه التشابه والاختلاف بين البنوك الشجرية العربية.

الجدول رقم (2)

أوجه التشابه والاختلاف بين البنوك الشجرية العربية

البنك النحوي	المدونة اللغوية	التمثيل النحوي	سمات أقسام الكلام	عددالوظائف النحوية المعنونة للكلمات	عدد الوظائف النحوية المعنونة للمركبات	الفئات الفارغة
ATB	**صحيفة النهار**	**شجرة بنية العبارة**	**500 سمة**	**-**	**20**	✓
PADT	**صحيفة وكالة فرانس برس+وكالة أنباء الحياة**	**شجرة بنية التبعية**	**أكثر من 500 سمة**	**20**	**-**	**-**
CATiB	**صحيفة وكالة فرانس برس+وكالة أنباء الحياة+الحياة**	**شجرة بنية التبعية**	**6 سمات**	**7**	**-**	**-**
SATB	**صحيفة إسلام أون لاين**	**شجرة بنية العبارة**	**62 سمة**	**18**	**-**	**-**

7- ترميز المدونة باستخدام لغة التوصيف XML:

تم ترميز البنك النحوي بلغة الترميز القابلة للامتداد Extensible Markup Languageباستخدام‏DTDالخاص بالبنك الشجري الألمانيTIGER،حيث يحولنظام

التقويسbracketingإلى تنسيقNegra export format،وهذا التنسيق يجعل البيانات على شكل مصفوفة سهلة القراءة؛مما يساعدنا على سهولة استخلاص السمات Features Extractionمن البناء الشجري.

8- استخدام البنك الشجريSATB:

وتحقيقا للغرض من البناء، قمت ببناء محلل نحوي قائم على الوظائف النحويةالدلاليةFunctional Tagsباستخدام أساليب التعلم الإحصائي الموجهة supervised learningمستخدما مصنّف الحقول العشوائية المشروطة Conditional random fields (CRFs)وتم استخراج سمات البناء لكل جملة عن طريق استخدام ملفاتXMLوكانت السماتFeatures هي الأقسام الكلامية والفئة النحوية للكلمة، ووضع الكلمة في البناء الشجري من حيث التسلسل العائلي(الأب-الأخوات)، نوع الكلمة في هيكلة السين البارية من حيث الرأس والمكمل والوصف، والمصنفات Classes هي الرموز الوظيفية.

وبعد تدريب المدونة النحوية(البنك الشجري) على مصنف الحقول العشوائية المشروطة++CRFوكانت نتيجة الإحصائيات كالآتي، كما هو مبيّن في الجدول رقم (3).

الجدول رقم (3)

إحصائيات عن البنك الشجري والتدريب الآلي

Words	Tokens	Sentence	Training	Testing	Class error rate
100,000	123K	5,045	%90	%10	%11,2

9- الخاتمة:

عرضنا في هذا البحث، موردا لغويا جديدا بهدف بناء محلل نحوي قائم على الوظائف النحوية الدلالية،لتحقيق المهمة الأساسية للتحليل النحوي الآلي وهي الفهم الأتوماتي للنصوص اللغوية ثم ذكرت مراحل البناء من اختيار المدونة اللغوية، وتجزئة النصوص، والتحليل النحوي، ثم بيّنت استخدام البنك الشجري في بناء محللٍ نحويٍّ، وآمل أن تزيد عدد كلمات البنك الشجري ليصل إلى مليون كلمة.

10- المراجع:

[1]السعيد، المعتزبالله (2010) مدونة معجم تاريخي للغة العربية، معالجة لغوية حاسوبية، أطروحة دكتوراه، كلية دار علوم-القاهرة.

[2] شمس الدين، جلال (د ت) الأنماط الشكلية لكلام العرب، مؤسسة الثقافة الجامعية-الإسكندرية.

[3]عاشور، المنصف (1991) بنية الجملة العربية بين التحليل والنظرية، منشورات كلية الآداب بمنوبة-تونس.

[4] عبادة، محمد(2007) الجملة العربية مكوناتها-أنواعها-تحليلها، مكتبة الآداب-القاهرة.

[5] علي، نبيل(1988) اللغة العربية والحاسوب، تعريب-القاهرة

[6]Abeillé, A.(2003) Building and Using Parsed Corpora, Springer.
[7]Attia, Mohamed(2004) Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications, Faculty of Engineering, Cairo University

[8]Attia, Mohammed (2007) Arabic Tokenization System in Proceedings of 5th Workshop on important unresolved Matters.

[9]Habash, Nizar and Farag, Reem, and Roth, Ryan.(2009) Syntactic Annotation in Columbia Arabic Treebank, In proceedings of the 2nd International Conference on Arabic Language Resources Tools.

[10] Habash , Nizar(2010), Introduction to Arabic Natural Language Processing, A Publication in the Morgan & Claypool Publishers series.

[11]Jurafdky, Daniel & Martin, James H (2006). Speech and language processing: An introduction to natural language processing, computational linguistics, and speechrecongnition, Formal Grammars of English.

[12]Slonneger,Kennth and others.(1995). Formal Syntax and Semantics of Programming Languages.

[13]Pustejovsky, James and Stubbs, Amber (2012) Natural Language Annotation for Machine Learning, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol.

[14]Rambow, Owen (2010) The Simple Truth about Dependency and Phrase Structure Representations, In proceedings of the 2010 Annual Conference Chapter of the ACL.

أحمد روبي محمد عبد الرحمن.

باحث لغوي في معالجة اللغة العربية آليا بإحدى شركات تقنية اللغات الطبيعية بالقاهرة، درست اللغة العربية التراثية والمعاصرة في مرحلة الليسانس بكلية دار العلوم جامعة الفيوم، وفي مرحلة الدراسات العليا ركزت على اللغويات بشكل عام واللغويات الحاسوبية بشكل خاص، وعملت على مشروعات في معالجة الكلام آليا(تحويل المكتوب إلى منطوق، وتحويل المنطوق إلى مكتوب)، والتشكيل الآلي، والمدقق الإملائي للغة العربية، ومهتم بالنظريات اللغوية الحاسوبية وبعنونة المدونات اللغوية.

Building Syntactic TreeBank for Modern Standard Arabic "Egyptian Newswire Language as a Model "

Ahmed Ruby Mohammed
Linguistics M.Sc. Researcher
Faculty of Dar-Oulum
Ahmedruby757@yhoo.com
aruby@istnetworks.com

Abstract-The syntactic Tree bank is an important resource for building applications for statistical natural language processing NLP syntactic Arab Treebank SATB is a set of grammatical analysis of Arabic sentences, has been relying on Islam on line corpus as a source of texts to be analyzed, and SATB differs from other tree banks in terms of linguistic information and the method of representation; and that the nature of the objective of building Parser based on semantic grammatical functions; to achieve the primary task of the automated analysis of grammar, which provide the necessary linguistic analysis to Natural Language Understanding, I used the Phrase Structure Tree representation in the analysis taking into account to determine the grammatical relations per Token in the sentence, and so for two main reasons: first easily extract the distinctive features of tokens constituent of the sentence, and the other is to get a hybrid attributes between dependency and structure X-barin terms of identifying grammatical dependency relations between the heads and complements phrases. This paper is a description of the stages of the constructions sequent and then compare it to other tree banks and how to use it.

Automatic Extraction of Subcategorization Frames of Arabic Verbs

Marwa Saber [1], Sameh Alansary [2]

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

Abstract—*The prime purpose of this study is to automatically extract the syntactic arguments of the modern standard Arabic verbs by designing a parser for analyzing the syntactic structures of Arabic verb phrases based on Chomsky's X-bar theory. This study also aims to test to what extent is X-bar theory an appropriate syntactic theory that is able to reveal the related syntactic arguments of specific predicates. In order to fulfill the purposes of the study, the researcher adopts an analytical descriptive approach where a corpus containing 600 sentences that was taken from Arabic Parkinson corpus for 60 verbs are analyzed using the IAN tool. IAN is the Interactive Analyzer and it is chosen for modelling X-bar(to model) theory owing to its close association with the X-bar approach whichindicates that all languages share the same underlying syntactic structure. Acquiring verb subcategorization is a fundamental issuein several NLP tasks, for instance, in parsing where theavailability of knowledge related to subcategorizationframes (SCFs) and the complement/adjunctdistinctionmeaningfully increases the accuracy of results.*

Keywords: Syntactic arguments, Subcategorization frame (SCF), X-bar theory, F-measure

1 Introduction

Automatic acquisition of lexical knowledge is the milestone in building the computational lexicons and grammars which enrich the work in Natural language processing applications (NLP) [1]. Knowledge about the verb is highly important because it is the primary source of the relational information in the sentence. The lexical information of the verb should specify the information about the subcategorization frame that represents the number,types, and the syntactic realization of the arguments which are the participants of the event described by the verb [2].
An argument is an expression that helps complete the meaning of a predicate. Most predicates take one, two, or three arguments. The discussion of predicates and arguments is associated the most with (content) verbs and noun phrases (NPs), although other syntactic categories can also be construed as predicates or as arguments. Arguments must be distinguished from adjuncts. While a predicate needs its arguments to complete its meaning, the adjuncts that appear with a predicate are optional; they are not necessary to complete the meaning of the predicate. Although, most of the syntactic and semantic theories acknowledge arguments and adjuncts, their definitions vary, and the distinction exists in all languages. The area of grammar that explores the nature of predicates, their arguments, and adjuncts is called valency theory. Predicates have a valence; they determine the number and type of arguments that can or must appear in their environment. Identifying the syntactic arguments of the predicate leads to the acquisition of the subcategorization frame of this predicate which plays a vital role in many natural language processing (NLP) applications.For example,it is important in improvingthe parsing results, solving the problems of parsing (PP-attachment,distinction between arguments and adjuncts) and the construction of lexicons. The parser that is enhanced withsubcategorization information is able to recognize the correct predicate – arguments relations [3]. Subcategorization frames specify the number and syntacticcategory of the predicate arguments;it also describes the predicateargumentstructure that is associated with it. SCFs is a well-studied linguistic phenomenon from a theoretical perspective. SCFs are of immediate utility in natural language processing (NLP) for electronic dictionaries and statistical parsers. Particularlywhen the language which is processed is a free wordorder language, where complements can freely appear on the leftor right side of the verbal head, Information about the subcategorization will play a crucial role in building applications serving this language [4].

Therefore, the representation of subcategorization plays an important role in tree bank annotation. Tree banks usually annotate subcategorization, both for free word order languages, likeTIGER Corpus for German (http://www.ims.unistuttgart.de/projekte/TIGER/), Alpino Dependency Tree bank for Dutch (http://www.let.rug.nl/˜vannoord/trees/),and Italian Syntactic Semantic Tree bank [5], and for fixed word order, like the English and Chinese Penn Tree banks ([6], [7]) that associate the resource with a repository, i.e. Prop Bank, where SCFs are collected.

The task of collection SCFs is a very time-consuming, because of the relative unportability of SCFs across corpora that feature different kinds of text andliterary genres. Therefore, various scholars proposedthe development of automatic systems for the extraction of subcategorization knowledge from linguistic corpora. [1] claimed that the task of syntactically analyzing substantial corpora of naturally occurring text has become a focus ofrecent work. Analyzed corpora would be of great benefit in the gatheringof statistical data regarding language use. Predicate subcategorization is a key component of any lexical entry, because most, if not all, recent syntactic theories extract syntactic structure from the lexicon [8]. A wide-coverage parser utilizing such a lexicalist grammar must have access to an accurate and comprehensive dictionary encoding the number and category of a predicate's arguments and also information about control with predicative arguments, semantic selection preferences on arguments, and so forth, to allow the extraction of the correct predicate-argument structure.

It has been very crucial in the recent years for the NLP researchers in order to develop any NLP application is to recognize the predicat’s syntactic arguments because it has been found that several substantial machine-readable subcategorization dictionaries exist for English, either built largely automatically from machine-readable versions of conventional learners' dictionaries, or manually by (computational) linguists (e.g. the Alvey NL Tools (ANLT) dictionary [9]; the COMLEX Syntax dictionary [10]). Unfortunately, neither approach can yield an accurate or comprehensive computational lexicon, because both rest ultimately on the manual efforts of lexicographers, therefore, prone to errors of omission and commission which are hard or impossible to detect automatically [8]. Furthermore, manual encoding of the lexical entries with the predicat’s argument is labour intensive.

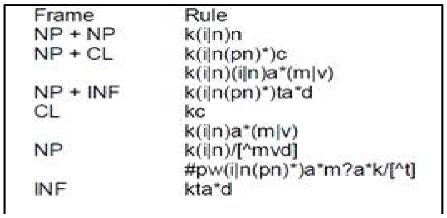
The subcategorization of a lexical item is one of the most important pieces of information associated with it. It is vital for both theoretical linguistics and in practical applications. It is indispensable in computational lexicons in order to be useful for natural language processing. Parsing can be greatly enhanced by providing the parser with the subcategorization frames of the verbs[11].

Several methods have been suggested to automatically extract the subcategorization frames from text corpora (e.g. [8] [12]-[14]; [15],[16],[17], [18]).

The architecture of the Brent system [12] consists of three modules: 1) verb detection: finds some occurrences of verbs using the case filter ([19]), a proposed rule of grammar, 2) SF detection: finds some occurrences of five subcategorization frames using a simple, finite state grammar for a fragment of English, 3) SF decision: determines whether a verb is genuinely associated with a given SF, or its apparent occurrences in that SF are due to error. This is done using statistical models of the frequency distributions. Brent uses the untagged brown corpus as input. The syntactic frames are: NP only, Tensed clause, Infinitive, NP & Clause, NP & Infinitive and NP&NP (these phrases types yield three syntactic frames with a single argument and three with two arguments. The cues used for identifying these frames are: lexical categories used in the definitions of the cues.

[15] suggested a method for the automatic extraction of the subcategorization frame by collecting as much co-statistics about the occurrences as possible from the text corpus, and then use statistical filtering (e.g., significance test ora mutual information measure) to get rid of false cues. He used the Kupiec’s stochastic part-of-speech tagger to tag 4 million words of the New York Times newswire. Then,he suggested a program to detect the SF consisting of two parts: 1) a finite state parser ran through the text to parse auxiliary sequences noting whether a verb is active or passive, and then it parses complements that follow the verb until something recognized as a terminator of subcategorized arguments is reached ([15] usedaperiod and subordinating conjunctions as frame terminators). Whatever has been found is entered in the histogram. 2) a process of statistical filtering is performed on the raw histograms to decide the best guess for what frames each observed verb actually had. The parser does not learn from participles since an NP after them may be subject rather than the object. The program acquired a dictionary of 4900 frames for 3104 verbs (an average of 1.6 per verb) [2].

[16]also make use of a PoS tagged corpus and a finite-state NP parser to calculate the relative frequency of the same six syntacticframes Brent used. The procedure of [16] to automatically determine subcategorization frame frequencies is to make a list of verbs out of the tagged corpus and then tokenize each sentence that contains the target verb; all the noun phrases except the pronouns are tokenized as “n” by a noun phrase parser. Then, they apply a set of frame extraction rules to the tokenized sentences. These rules are written as regular expressions as in Fig.1.A regular grammar is used to estimate the appropriate syntactic frame for each verb token in the corpus.



Frame	Rule
NP + NP	k(i|n)n
NP + CL	k(i|n(pn)*)c
	k(i|n)(i|n)a*(m|v)
NP + INF	k(i|n(pn)*)ta*d
CL	kc
	k(i|n)a*(m|v)
NP	k(i|n)/[^mvd]
	#pw(i|n(pn)*)a*m?a*k/[^t]
INF	kta*d

Figure 1: A frame extraction rules to the tokenized sentence

In an experiment involving the identification of these frames, the system showed an accuracy rate of 83%. The most frequent source of errors in frame identification by this system was errors in NP boundary detection. The second most frequent source was misidentification of infinitival purpose clauses.

[8] proposed a system for distinguishing 160 verbal frame classes and their relative frequency in English. B & C's system consists of six components, which are applied in sequence to sentences containing a specific predicate in order to retrieve a set of frame classes for that predicate:

1. A tagger, a first-order HMM part-of-speech and punctuation tag disambiguator, is used to assign and rank tags for each word and punctuation token in a sequence of sentences;
2. A lemmatizer is used to replace word-tag pairs with lemma-tag pairs;
3. A probabilistic LR tagger, trained on a tree bank, returns ranked analyses;
4. A pattern set extractor which extracts frame patterns, including the syntactic categories and head lemmas of constituents from sentence subanalyses which begin/end at the end of specified predicates.
5. A pattern classifier which assigns patterns in patternsets to frame classes or rejects patterns as unclassifiable on the basis of the feature values of syntactic categories and the head lemmas in each pattern;
6. A pattern set evaluator which evaluates sets of patternsets gathered for a (single) predicate constructing putative frame entries and filtering the latter on the basis of their reliability and likelihood.

The system of [8]achieved a token recall of 80.9%, which is comparable to previous approaches. B & C have attributed most of the errors to the filtering phase, which they describe as the 'weak link' in the system.

[17] presented an unsupervised learning method for subcategorization acquisition that considers what the shortcomings of the previous methods; that is, they are knowledge-based and thus require either existing tools (e.g., a wide-coverage parser in the case of [8]) or an important amount of time and linguistic expertise to write the necessary patterns, regular expressions, finite-state NP parsers etc (e.g., [12], [15], [16]). In contrast, [17] method only requires PoS-tagged text as input. This method is based on the assumption that subcategorized constituents differ from non subcategorized ones in terms of frequency [10]. This means that the subcategorization property of a verb should somehow show up when enough sentences containing this verb are collected. The idea of unsupervised learning then is to model the global behavior of each verb, and group verbs that behave syntactically similar. These groupings should then ideally correspond to groups of verbs with similar subcategorization properties. The information about the group membership of a verb could therefore be used by a parser when making local decisions, e.g., about the complement- or adjuncthood of a constituent. The global subcategorization behavior of a verb is extracted using hierarchical clustering ([20],[21])

There is an Arabic attempt to automatically extract the Arabic subcategorization frames (or predicate-argument structures) from the Penn Arabic Treebank (ATB) for a large number of Arabic lemmas, including verbs, nouns and adjectives [18]. The results have been compared against a manually constructed collection of subcategorization frames designed for an Arabic LFG parser [22].

In English,the construction and extraction of subcategorization frames received a lot of attention [18],one example is the specialized lexicon COMLEX [10] which is an extensive computational lexicon containing syntactic information for approximately 38,000 English headwords, with detailed information on subcategorization, containing 138 distinct verb frames for 5,662 active verbs lemmas. For Arabic, the attention has been directed, for the most part, to the construction and automatic extraction of semantic roles [22]. According to the researcher knowledge, there areonly one resource that exists for Arabic subcategorization frames which is the lexicon that is manually developed for the Arabic LFG Parser [22]. It is published as an open-source resource under the GPLv3 license[1]. It contains 64 frame types, 2,709 lemmas types, and 2,901 lemma-frame types, averaging 1.07 frames per lemma. The resource incorporates control information

[1] http://arasubcats-lfg.sourceforge.net

and details of specific prepositions with obliques and the automatically lexicon of Arabic subcategorization frames [18] (the automatic extraction of syntactic information, or subcategorization frames, from the Arabic Treebank (ATB) [23].) which uses the first manual one in the evaluation of it.

In this paper, the researcher will present an implemented system that takes a raw, untagged text corpus as its only input to generate a partial list of verbs that occur in the text and the subcategorization frames (SFs) in which they occur. Verbs are detected by the tagger, through searching in the dictionary and tokenizeing the sentence. Section 2 explains the notion of the syntactic arguments and the differences between the main arguments, the difference between arguments and adjuncts. Section 3 includes the definition of the subcategorization frame and sheds light on all different frameworks works in the area of the automatic extraction of the subcategorization frame. Section 4 details the description of the used Arabic corpus.Section 5 represents the analysis tool (IAN) and the proposed system to automatically extract the SCFs for modern standard Arabic (MSA) verbs. Section 6 illustrates an experiment on an Arabic sentenceto explain the methodology of the automatic extraction of subcategorization frames. Section 7 evaluates the output. Finally, section 8 concludes the paper.

2 SYNTACTIC ARGUMENTS

An argument is an expression that helps complete the meaning of a predicate. Most predicates take one, two, or three arguments. The discussion of predicates and arguments is associated the most with verbs and noun phrases (NPs), although other syntactic categories can also be consideredas predicates oras arguments. Recognizing the syntactic arguments of the predicate leads to the acquisition of the subcategorization frame of this predicate. In syntax, the terms argument and complement overlap in meaning and use to a large extent. In dependency grammar,arguments are sometimes calledactants.

Languages usually have one privileged syntactic argument per sentence which should always be a subject.There are certain tests that could be done to detect the subject, one involving agreement. The form of the verb depends on a certain entity which is the subject.The object is another syntactic argument which typically is the entity in which the action is done to. For example,in sentence (1), the subject of that sentence is "الولد", while the entity that the action was performed onis the word “الدرس”, which makes it the object of the sentence. The syntactic arguments such as subject and object are not the same as semantic roles of agent and patient.There is also the third argument, which goes by various names. Some call it the indirect object, dative, recipient. Inexample (2),“علي” is the third argument. This is used whenever there is a verb that takes three arguments.

1) كتب الولد الدرس.
2) أعطى محمد علي هدية.

A. The difference between arguments and adjunct

In order to represent accurate subcategorization information, a distinction should be made between complements and adjuncts. Complements are taken to be syntactically specified and required by the head (The predicate needs its arguments to complete its meaning), whereas adjuncts can only modify a head (the adjuncts are not necessary to complete the meaning of the predicate; their appearance with a predicate are optional), according to almost all different frameworks (e.g. the Minimalist Program [24], Lexical-Functional Grammar [25], Head-Driven Phrase Structure Grammar [26], Categorial Grammar [27], and Tree-Adjoining Grammar [28]. Constituents have to be either selected as complements or adjuncts. The distinction between arguments and adjuncts certainly exists in all languages. The distinction between arguments and adjuncts is essential in the basic analysis of the syntax and semantics of clauses.Sentencein (3) contains the first noun phrase asthe subject, while the object argument is the prepositional phrase (*على الجائزة*). Verbal predicates that require an object argument are described as transitive. Moreover, there are verbal predicates that require two object arguments are described as ditransitive. In (4) additional information has been added which is considered as an adjunct.

3) حصلت المرأة العجوز على الجائزة
4) حصلت المرأة العجوز على الجائزة الكبيرة

The added phrase “الكبيرة” is adjunct because it provides additional information that is not necessary to complete the meaning of the predicate “حصل”. One key difference between arguments and adjuncts is that the appearance of a given argument is often obligatory, whereas adjuncts appear optionally. The PP in (3)is an argument because when it isomitted, the remaining part makes an incomplete sentence.Subject and object arguments are known as core arguments; core arguments can be suppressed, added, or exchanged in different ways.

B. Obligatory vs. optional arguments

Many arguments behave like adjuncts with respect to another diagnostic which is the omission diagnostic (sign). Adjuncts can always be omitted from the phrase, clause, or sentence in which they appear without rendering the resulting expression unacceptable. Whereas, the obligatory arguments cannot be omitted.“البيت” ‘the house’ is an optional argument,see sentences in (5) and (6).

5) الأم نظفت البيت.
6) الأم نظفت.

C. Representing arguments and adjuncts

The distinction between arguments and adjuncts is often indicated in the tree structures used to represent syntactic structure. In phrase structure grammar, an adjunct is "adjoined" to a projection of its head predicate in a manner that distinguishes it from the arguments of that predicate. The distinction is very clear in the theories that employ the X-bar schemaas in Fig.2.

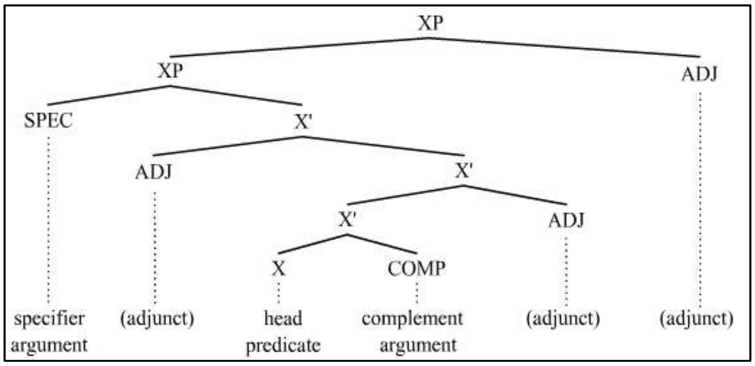


Figure 2: The schema of X-bar theory

The complement argument appears as a sister of the head X, and the specifier argument appears as a daughter of XP. The optional adjuncts appear in different positions adjoined to a bar-projection of X or to XP.

D. Semantic Intuitions Concerning the Argument-Adjunct Distinction

The task of making the distinction between arguments and adjuncts of a verb can be described as a way of capturing a basic intuition. For example, if a world event or activity must be described, such an event will necessitate participants or other relevant information that is salient to the setting (completion of meaning).In any sentence, some information will be more crucial to the described event and other information will be less important. Thus, the linguistic intuition is that in an event described by the verb, there will be key participants without it the event would not be complete and other peripheral information that provides descriptors of the general condition or circumstance of the state or event, which are not as central to the meaning of the verb [29].

When we invoke our intuitions of which are the “necessary” participants in a given state or event described by the verb, we are referencing the semantics side of the issue [29]. When we need to know the arguments of a given event or state we ask about the participants of that event, our intuition is responsible for that.For example:

7) هو أعطاها كتاب عن السيارات من باريس الليلة الماضية لنجاحها في دراستها.

This sentence includes five elements or concepts: “هو”, “ها”, “كتاب عن السيارات من باريس”,“الليلة الماضية” and “ لنجاحها في دراستها”. In such an event as “أعطى”‘giving’, there are certain participants that would be considered necessary to make the meaning complete. We would first require the mention of the entity who gives, the entity who receives, and the object that is transferred between the two entities. That is, intuition would tell us that for in a sentence like (7) there are three participants, namely “هو” ‘he’, “ها” ‘her’ and “كتاب” ‘a book’, each expression plays a central role in the “أعطى”‘giving’ event and are required by the verb.

The other two expressed elements in the sentence, namely the adverbial “الليلة الماضية” ‘last night’ and the prepositional phrase “لنجاحها في دراستها” ‘for succeeding in his studies’, provide a general setting for the event of “أعطى” ‘giving’.

In the case of the first three elements, the relationships they have with the verb are often referred to as thematic relations. This concept of thematic relations has been discussed by numerous studies in the linguistics literature (cf. [30], [31], [32]). Thematic relations describe the roles these participants play in the event or state created by the verb. Then, in example (7), “هو”‘he’ is the AGENT or GIVER as he takes on the role of the giving entity, “كتاب”‘book’would be the THEME or TRANSFERRED ITEM, and “ها”‘her’ is the RECIPIENT of the “كتاب” ’book’. Furthermore, these participating roles are considered required or obligatory in such a way that if they were removed from the sentence as in (8) and (9), they will result in incomplete sentences:

8) هو أعطى.

9) هو أعطاها.

For such utterances as in (8) and (9) to be meaningful, the missing participant(s) would have to be cited elsewhere and recoverable in the context. Thus, these participants are considered to play a direct role in the relational information conveyed by the verb, and therefore, necessary components of the semantics of the verb. Those participants that have thematic relationships with the verb are considered to be semantic arguments of the verb. In contrast to the arguments, the last two elements in example (7) would be considered semantic adjuncts. Unlike arguments, adjuncts do not rely on the relational information conveyedby the verb. Rather they comment on the general action or state of the predicating unit – the verb and its arguments. The adverbial “الليلة الماضية”‘last night’and the prepositional phrase “لنجاحها في دراستها” ‘for succeeding in his studies’ are present because they comment on the event described by the verb and its arguments: the adverbial sets the time in which the giving takes place and the prepositional phrase describes causal events leading up to the event. Thus, in general, the elements in the sentence that are in a thematic relationship with the verb, and play a central role in the event or state presented by the verb are considered to be arguments. These arguments are licensed and required by the verb to realize its full meaning but the elements in the sentence that do not hold a specific relationship to the verb and provide contextual information.

E. Problem of Semantic Intuition

The semantic intuition in determining the argumenthood may be tricky. Example in (10),the "أكل"‘eating’event has two participants: the one who eats and the entity that is eaten. The prepositional phrase provides a general location or setting in which eating takes place. From such example one can extrapolate that prepositional phrases could always be considered adjuncts as in example (10). However, this is not always the case. Consider the example in (11):

10) أكلنا الطعام في المطبخ.

11) وضعت الكتاب على الطاولة.

The locative prepositional phrase in (10) is distinguished from the same prepositional phrase inexample (11), which would generally be recognized as the argument of the verb ‘put’ as it is thelocation in which the book is placed. The event would not be complete without the prepositional phrase “على الطاولة” so it would have to be classified as an argument.Finally, in certain cases, the distinction seems to depend on the lexical items present in theSentence [33].

F. Challenges in NLP

Dealing with the distinction between arguments and adjuncts constitutes a clear semantic challenge in the NLP community. There is no single set of rules by which we could say that a certain phrase is an argument or an adjunct, as such a decision would depend on the verb in the sentence. The distinction depends on the semantic and syntactic context and world knowledge. There are numerous automatic tasks in NLP that would benefit from a clear distinction between arguments and adjuncts, including tasks like automatic parsing, machine translation, text summarization and text simplification. In order for such tasks to successfully benefit from the argument/adjunct distinction, the syntactic, lexical, and semantic resources on which these NLP tasks rely have to do an accurate and consistent job in identifying as well as describing the distinction [29]. Finally, one of the well-known challenges in creating and maintaining NLP resources, especially the creation of labeled corpora, is that annotations are costly. Establishing clear guidelines for any NLP resource is crucial, as they are the key factor in facilitating quick but consistent decisions in annotating text.

Due to the influences of transformational syntax (e.g. Principles & Parameters (P&P), Minimalist Program (MP)) and its view that semantics can be mapped onto a hierarchical syntactic structure in a systematic and deterministic manner, much of the discussion of argument and adjunct distinction cannot be made without making close reference to the syntactic concept of complements, and core and oblique arguments. Complements are phrases that are obligatorily selected or subcategorized by the verb. Since objects, which are core arguments of the verb, are obligatory in transitive/ditransitive sentences, they are also considered to be complements of the verb. In a similar manner, since oblique arguments (e.g. adverbial phrases) are not required like the core arguments are, they are considered to be syntactic adjuncts (i.e. non-complements) of the verb [29].

3 SUBCATEGORIZATION FRAME

A. What is subcategorization

The subcategorization frame (SCF) of a verb specifies the number and categories of syntactic arguments a verb takes. It includes all the complements of a given word. For instance, the verb "أكل"'eat'can be used as both transitive or intransitive, respectively, as in "أكل الولد"'the boy eat' or "أكل الولد الطعام"'the boy eat the food' both are valid frames associated with the verb "أكل"'eat'.

B. Valency vs. subcategorization

The theory that explores the nature of predicates, their arguments, and adjuncts is called the valency theory. Predicates have a valence; they determine the number and type of arguments that can or must appear in their environment. The valence of predicates is also investigated in terms of subcategorization frames. Valency includes all of the verb arguments, including the subject. The linguistic usage of the term valence is derived from the definition of valency in chemistry. This scientific metaphor is developed by Lucien Tesnière, who rendered verb valency into a major component of his dependency grammar theory of syntax and grammar. The notion of valency first appeared as a comprehensive concept in Tesnière's book[34].

C. The status of subjects

The subcategorization notion is similar to the notion of valency. Although subcategorization has originated with phrase structure grammars in the Chomskyan tradition, while valencyhas originated with Lucien Tesnière of the dependency grammar tradition. The primary difference between the two concepts concerns the status of the subject. As it was originally conceived, subcategorization did not include the subject, meaning that a verb is subcategorized for its complement\complements only; whereas, valency includes the subject. Tesnière used the word actants to mean what are now widely called arguments (and sometimes complements). An important aspect of Tesnière's understanding of valency was that the subject is an actant (=argument, complement) of the verb in the same manner that the object is. The concept of subcategorization which is related to valency, but associated more with phrase structure grammars than with the dependency grammar that Tesnière has developed, did not originally view the subject as part of the subcategorization frame.

D. Verb Subcategorization in Linguistic Theory

The treatment of subcategorization varies across linguistic theories [11]. In the following section, we will offer an overview of the different approaches and compare their relevance for subcategorization acquisition.

*1) Government-Binding and related approaches:*The government and binding theory was developed by Chomsky and others in 1980's. One of the central parts of GB is the X-bar theory [11]. GB seeks to capture the similarities between different categories of lexical phrases by assigning the same structure to them. Rather than having different phrase structure rules for VPs, NPs, etc., just the two basic rules in (1) cover all the lexical categories.

(1) Phrase Structure Rules:
(For any lexical category X, X0=Head)
XP ➔Specifier X′
X′ ➔X0 Complements

In the trees generated by these rules, the top node (corresponding to left side of the rule) is known as the mother, with the two daughters introduced by the right side of the phrase structure rule. The daughter nodes at the same level are known as sisters. In (2) one of the daughters, X′, is also a mother with daughters of her own, just as in normal family relationships.

(2) Basic X-bar Structure
XP➔maximal projection
specifier X′➔intermediate projection
X0➔head complement(s)

This schemaclaims that all phrases are projected from lexical categories in the same way. For adjunction: X*n*➔Y*m*X*n*.A head (=X0) subcategorizes for all and only its sisters. The subcategorized complements are always phrases.Heads and their maximal projections share features, allowing heads tosubcategorize for the heads of their sisters (i.e. *rely*).In general, specifiers are optional. Evidently, specifiers may be words or phrases. The following trees illustrate how X-bar theory works. We apply the X-bar rules to specific categories. First, we have to find the head, which determines the type of phrase, then look for specifiers, complements, adjuncts, and conjunctions. An important feature of GB is the fact that subjects are not subcategorized for by the verbal head. The domain of subcategorization is limited to the maximal projection containing the head. In GB subjects are typically outside of VP, i.e. they are not sisters to the verbal head. This leads to GB predicting a number of subject/object asymmetries in syntax [35].

2) Lexical-Functional Grammar: The LFG model of syntax consists of two parts, the c-structure and the f-structure. c-structure encodes such inter linguistically variable properties as word order and phrase structure. F-structure expresses the relations between the functional constituents of a phrase. Those constituents are grammatical functions such as SUBJ (subject), OBJ (object), or XCOMP (open complement). Thus, LFG accords theoretical, primitive status to the notion of grammatical function, which GB treats as reducible to phrase structures. LFG accords to the notion of grammatical function. Although c-structures, together with the lexicon, determine the f-structures there is no direct mapping from c-structures to f-structures, and each obey their own specific constraints. F-structures are built based on information from two sources. One is functional annotations associated with c structures. For example, see Fig.3.

$$\begin{array}{lcc} \mathrm{S} \longrightarrow & \mathrm{NP} & \mathrm{VP} \\ & (\uparrow \mathrm{SUBJ}){=}\downarrow & \uparrow{=}\downarrow \end{array}$$

Figure 3: Functional annotations associated with c structures

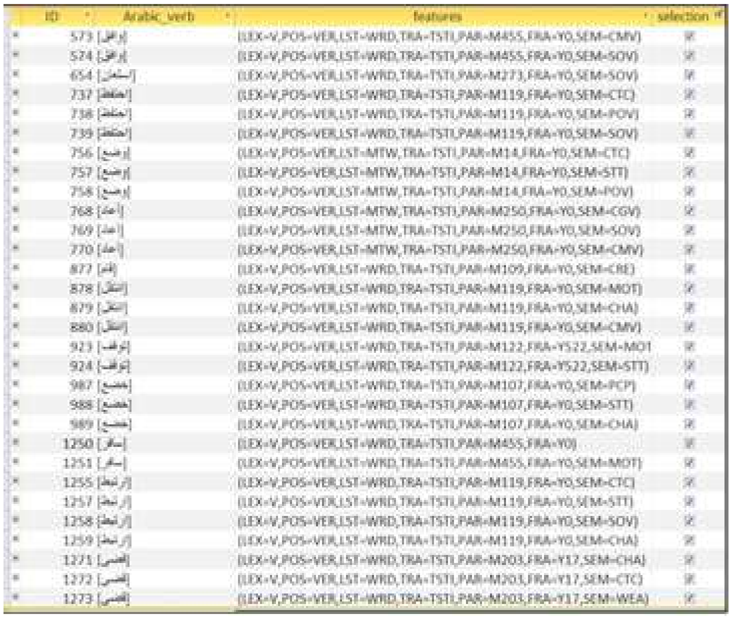
The arrows in the annotation refer to the function of the annotated constituent. The up-arrow means that the function refers to the mother of the node, while the down-arrow indicates the node itself. So the first NP annotated as (↑SUBJ) means that this NP is the SUBJ of its mother, i.e. the S, or more precisely, that the f-structure carried by the NP goes to the S's SUBJ attribute. Similarly, the VP's annotation (↑=↓) indicates that the VP's f-structure is also S's f-structure – which can be paraphrased as VP being the functional head [35]. The other source of information is the lexicon. Lexical forms subcategorize for forms rather than categories. This allows for non-standard categories to realize functions in a sentence (e.g. non-NP subjects). Functions are also linked to arguments of the Predicate-Argument Structure. In contrast to GB, in LFG; subject forms part of the verb's subcategorization frame.

*3) Head-Driven Phrase-Structure Grammar:*This theory of grammar combines insights from a variety of sources, most notably GPSG, CG and GB. It stresses the importance of precise formal specification. In HPSG subcategorization, information is specified in lexical entries as exposed in [26], the subject is treated in a way similar to other arguments. Verbs have a SUBCAT feature whose value is a list of synsem objects corresponding to values of the SYNSEM features of arguments subcategorized for by the head. The order of these objects corresponds to the relative obliqueness of the arguments, with the subject coming first, followed by the direct object, then the indirect object, then PPs and other arguments.

4 THE CORPUS

A. Compiling some of Arabic verbs

Subcategorization is used to refer to the subdivision of major syntactic categories, particularly verbs, according to what other constituents they co-occur with. Thus, the category of verbs can be split into subcategories such as transitive, intransitive, ditransitive or other kinds of verbs based on the number and type of syntactic arguments they requires. A single verb may belong to more than one subcategorization frame (SF); which can be described as the order and category of the constituents that are co-occurring with the verb. The Arabic verbs were selected by the researcher according to the types of transitivity: transitive, ditransitive since they are the most commonly used verbs in the Arabic. The transitivity (TRA) consists of two major classes and each oneconsists of subclasses; firstly, the transitive category (TST) (requires object) which consists of the following subclasses: direct monotransitive (TSTD): one direct object, indirect monotransitive (TSTI): one indirect object, ditransitive (TST2): one direct object and one indirect object and tritransitive (TST3): three objects. Secondly, the intransitive (NTST): (does not require object). This class consists of the following subclases: unergative (NERG): the subject is the agent and unaccusative (NACC): the subject is not the agent. The latter (intransitive verbs)is not included in the researcher study.



ID	Arabic_verb	features	selection
573	[رافق]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M455,FRA=Y0,SEM=CMV}	☑
574	[رافق]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M455,FRA=Y0,SEM=SOV}	☑
654	[استعمل]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M273,FRA=Y0,SEM=SOV}	☑
737	[احتفظ]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=CTC}	☑
738	[احتفظ]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=POV}	☑
739	[احتفظ]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=SOV}	☑
756	[وضع]	{LEX=V,POS=VER,LST=MTW,TRA=TSTI,PAR=M14,FRA=Y0,SEM=CTC}	☑
757	[وضع]	{LEX=V,POS=VER,LST=MTW,TRA=TSTI,PAR=M14,FRA=Y0,SEM=STT}	☑
758	[وضع]	{LEX=V,POS=VER,LST=MTW,TRA=TSTI,PAR=M14,FRA=Y0,SEM=POV}	☑
768	[أعاد]	{LEX=V,POS=VER,LST=MTW,TRA=TSTI,PAR=M250,FRA=Y0,SEM=CGV}	☑
769	[أعاد]	{LEX=V,POS=VER,LST=MTW,TRA=TSTI,PAR=M250,FRA=Y0,SEM=SOV}	☑
770	[أعاد]	{LEX=V,POS=VER,LST=MTW,TRA=TSTI,PAR=M250,FRA=Y0,SEM=CMV}	☑
877	[قام]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M109,FRA=Y0,SEM=CRE}	☑
878	[انتقل]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=MOT}	☑
879	[انتقل]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=CHA}	☑
880	[انتقل]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=CMV}	☑
923	[توقف]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M122,FRA=Y522,SEM=MOT	☑
924	[توقف]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M122,FRA=Y522,SEM=STT}	☑
987	[خضع]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M107,FRA=Y0,SEM=PCP}	☑
988	[خضع]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M107,FRA=Y0,SEM=STT}	☑
989	[خضع]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M107,FRA=Y0,SEM=CHA}	☑
1250	[سافر]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M455,FRA=Y0}	☑
1251	[سافر]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M455,FRA=Y0,SEM=MOT}	☑
1255	[ارتبط]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=CTC}	☑
1257	[ارتبط]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=STT}	☑
1258	[ارتبط]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=SOV}	☑
1259	[ارتبط]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M119,FRA=Y0,SEM=CHA}	☑
1271	[قضى]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M203,FRA=Y17,SEM=CHA}	☑
1272	[قضى]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M203,FRA=Y17,SEM=CTC}	☑
1273	[قضى]	{LEX=V,POS=VER,LST=WRD,TRA=TSTI,PAR=M203,FRA=Y17,SEM=WEA}	☑

Figure 4: A list of the selected Arabic verbs from the analyzed data

B. Corpus description and Classification

The researcher has selected60 verbs belonging to different types of transitivity sub-categories. 100 sentences were extracted for each verb, the selected sentences were chosen according to the length; the length of sentences range from five to 12 words.The selected sentences were filtered to be 10 sentences for each verb. The filtration was done to achieve some criteria; the criteria were assigned on a linguistic and systematic basis.

The corpus is intended to be representative of the contemporary standard use of the written Arabic language. The corpus is segmented into sentences and tagged for POS. All the required linguistic attributes are assigned to each word. The maximum length of the sentences is 12 words. This corpus is collected according to the following:

- The occurrences of the syntactic arguments of the predicate (verb).
- Does the syntactic argument occur after the predicate immediately or preceded by a constituent as in (12)?

  12) تمسكت الدولة العبرية امس[2] بموقفها الذي قال بار ايلان انه ينطوي

- Does the same predicate have one specific type of subcategorization frame in all its context?

Moreover, the frequency of each structure is documented. Since the researcher's data is compiled from different genres (as the corpus was extracted from arabi Corpus[3]), the coverage rate is high and the opportunity of the occurrence of different structure is extremely high, as a result the data is considered more robust. Example from the collected corpus can be seen in table 1.

TABLE I

EXAMPLES OF THE CONTEXTS SELECTED FOR THE ARABIC VERB "حصل"

ID	Sentences	sentence_ref
142	حصل أحمد حميد الطاير وزير المواصلات الاماراتي بالوكالة، في مؤتمر وزراء	126
143	يحصل الأردن على جزء من النفط العراقي بسعر خاص يقل عن	126
144	حصل قطاع السكة الحديد عام 1996 على استثمارات قدرت ب	126
145	حصلت الجزائر امس على قرض من صندوق أبو ظبي للتنمية مقداره	126
146	حصل خطأ من قبل الادارة السابقة في عدد محدود من العلب	126
147	حصلت شركة »الرضوان للهندسة والمقاولات« على شهادة نظام الجودة العالمية الأيزو	126
148	حصلت »نومورا« التي تنشط في أسواق المال والاستثمار على مقعد في	126

[2] Note that it was wrote wrongly in the corpus and will be corrected automatically in the processing phase

[3] http://arabicorpus.byu.edu/

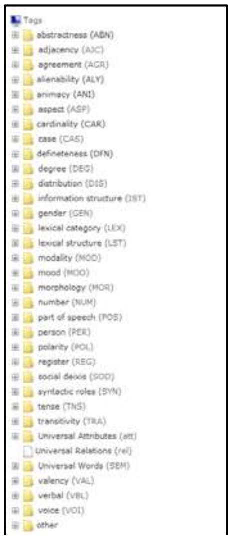
ID	Sentences	sentence_ref
149	حصلت الشركة الخليجية الدولية للاستثمار على ترخيص من السلطات الكويتية بالعمل	126
150	يحصل زبون المصرف المشترك في الخدمة المصرفية الالكترونية على الاشتراك مجانا	126
151	حصل الجمهوريون ايضا على ما طلبوه، أي على نصوص تشمل خفض	126

In table 1, there are examples of the contexts selected for the Arabic verb “حصل” which was extractedfrom the corpus. There are ten sentences that were selected very carefully, taking into account the diversity in the structures, for example sentence in (13):is not selected bythe research to be analyzed because it is an incomplete sentence, but sentence (14) can be analyzed manually and automatically because the researcher can extract the syntactic argument of the predicate “حصل”.Moreover, in sentence (14) the complement of the verb occurs after the subject, but in sentence (15) the complement of the verb occurs after the adverb “امس”and not after the subject. In sentence (16), the subject is modified by a phrase that separate the main predicate “حصل” from its syntactic argument.

13) حصل أحمد حميد الطاير وزير المواصلات الاماراتي بالوكالة، في مؤتمر وزراء
14) يحصل الأردن على جزء من النفط العراقي بسعر خاص يقل عن
15) حصلت الجزائر امس على قرض من صندوق أبو ظبي للتنمية مقداره
16) حصلت نومورااالتي تنشط في أسواق المال والاستثمار على مقعد في

C. Tagging the data

The researcher used a list of features extracted from the UNDL Foundation tagset4. This tagset is a set of features in the universal networking language (UNL[5]) dictionary that depend on the structure of the natural language. However, in order to boost the standardization of the lexical resources used in the UNL framework, the UNDL Foundation6 recommends adopting the following tags for some specific and pervasive grammatical phenomena. The hierarchy of the tagset is shown in Fig.5.



Tags
abstractness (ABN)
adjacency (AJC)
agreement (AGR)
alienability (ALY)
animacy (ANI)
aspect (ASP)
cardinality (CAR)
case (CAS)
definateness (DFN)
degree (DEG)
distribution (DIS)
information structure (IST)
gender (GEN)
lexical category (LEX)
lexical structure (LST)
modality (MOD)
mood (MOO)
morphology (MOR)
number (NUM)
part of speech (POS)
person (PER)
polarity (POL)
register (REG)
social deixis (SOD)
syntactic roles (SYN)
tense (TNS)
transitivity (TRA)
Universal Attributes (att)
Universal Relations (rel)
Universal Words (SEM)
valency (VAL)
verbal (VBL)
voice (VOI)
other

Figure 5: List of tags in alphabetical order

Several of those linguistic constants have been already proposed in the Data Category Registry (ISO 12620)[7], and represent widely accepted linguistic concepts. The purpose of this tag set is providing the technical means for describing any linguistic behavior which should be done in a highly standardized manner, so that others could easily understand and exploit the data for their own benefit. The main intention is to create a harmonized system in order to make language resources as easily understandable and exchangeable as possible.

[4] http://www.unlweb.net/wiki/Tagset
5 http://www.unlweb.net/unlweb/
[6] www.undlfoundation.org
[7] http://media.dwds.de/clarin/userguide/text/concepts_ISOcat.xhtml

D. The analysis of the data

The theory which is adopted in the linguistic analysis process is the X-bar[8] theory; it postulates that all human languages share certain structural similarities, including the same underlying syntactic. Constituency grammars are a method of sentence analysis that divides a sentence into major parts, which are in turn further divided into smaller parts in a process that continues until irreducible constituents are reached, i.e., until each constituent consists of only a word or a meaningful part of a word. The end result is presented in a visual diagrammatic form that reveals the hierarchical immediate constituent structure of the sentence at hand. For example sentence (17) is represented in Fig.6.

17) وافق شارون على الخطة المصرية.

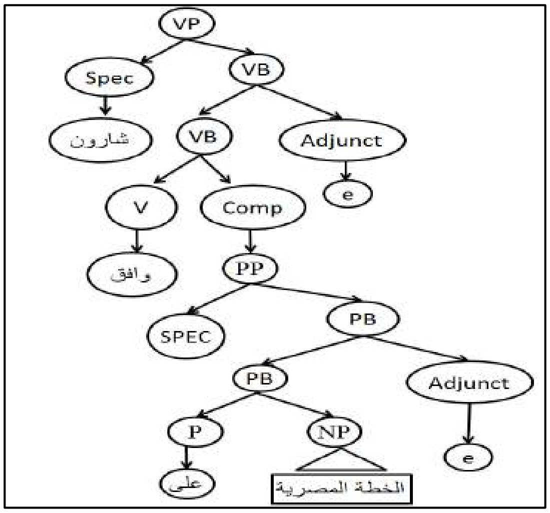


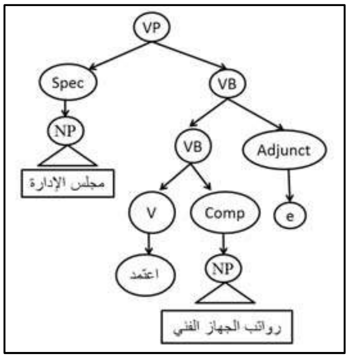
Figure 6: The deep representation of sentence (17) by X-bar theory

وافق: V, +[—PP]

The researcher introducesthe linguistic analysis of two sentences for the verb “اعتمد” based on X- bar theory,the verb “اعتمد” has two different subcategorization frames, as shown in in Fig .7 and Fig.8.

18) تعتمد فكرة الحفار علي شفط الاتربة الناتجة عن الحفر بالبنطة والحلزون

19) اعتمد مجلس الإدارة رواتب الجهاز الفني



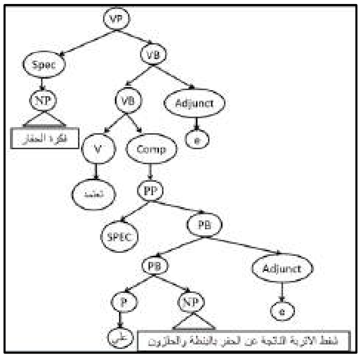


Figure 7: The representation of sentence in (19) **Figure 8: The representation of sentence in (18)**

اعتمد: V, +[—NP]اعتمد: V, +[—PP[علىNP]]

[8] http://www.unlweb.net/wiki/X-bar_theory

5 THE ANALYSIS TOOLUSED INTHE AUTOMATIC EXTRACTION OF THE SUBCATEGORIZATION FRAME

This section discusses the linguistic design and the implementation of the tool used in the automatic extraction of the subcategorization frame. The UNL system has developed a tool that is capable of performing a tokenization, disambiguation and deep/shallow syntactic analysis of the Arabic structures. The tool is called Interactive Analyzer (IAN). IAN is a natural language analysis systemthatcan represent natural language sentences into syntactic trees. In its current release, it is a web application developed in Java and available at the UNLdev[9]. The lexicon of the tool is designed to includethe lexical items of a given natural language along with a set of assigned linguistic features. The grammar of the tool consists of different steps to syntactically analyze the structures: tokenization (the identification of the tokens (lexical items) of each sentence of the input sentence) and parsing. These steps are responsible for both of the lexical and the deep syntactic analysisof the Arabic sentences. They can be implemented through three different types of rules; Normalization rules (N-rules), Disambiguation (D-rules) and transformation rules (T-rules).

A. UNL Lexicon

The dictionary assignsa list of features for each lexical item. The features cover different linguistic levels: morphological information, morpho-syntactic information and syntactic information. UNL framework uses a standard and universal list of features (Tagset) to describe all types of linguistic information concerning every Arabic word. Part of speech feature; used to classify words into main classes and each class may include subclasses. The classes are nouns, verbs, adjective, adverb, affix, classifier, conjunction, determiner, interjection, numeral, particle and pronoun. Morpho-syntactic information is concerned with the grammatical categories and linguistic units that have both morphological and syntactic properties,gender, number, person and many other features are involved in the grammatical agreement in the lexicon.Transitivity is a feature of verbs which indicates the number of objects a verb requires or takes in a given instance. The transitivity of a verb can be basically classified into intransitive (NTST) and transitive (TSTD). Moreover, the lexicon is enhanced by information about person, tense, case and voice.

B. The grammar

Grammars are sets of rules used to transform the natural language into a parsed tree. There are two different types of rules: transformation rules that are used to make changes to the nodes or relations and disambiguation and tokenization rules that are used to control the changes over nodes or relations.

1) Normalization Rules: N-rules constitute the pre-processing module that is applied to the input text prior to the processing phase; before the dictionary lookup phase. They are concerned with normalizing the input text; replace the abbreviations by their extended forms, assign the spaces if not found (some texts are written wrongly with no space boundary between the words). Finally,correct the wrong written words toassistthe identificationof the words from the dictionary.

2) Disambiguation rules: The function of D-rules is to tokenize and prevent wrong lexical choice from the dictionary.The tokenization algorithm is strictly dictionary-based. The D-rules also control the segmentation of the tokens.

3) *Transformation rules:* This type of rules is used for normalization and syntactic analysis. The syntactic moduleis responsible for transforming the list of nodes into the tree structure using binary relations based on X-bar theory. The syntactic processing is done automatically.The general design for this module is that “It starts by composing small trees for the small phrases in the sentence and combining these small trees together to form a bigger tree”. While building the syntactic trees for the 600 sentences a lot of linguistic issues have been faced.

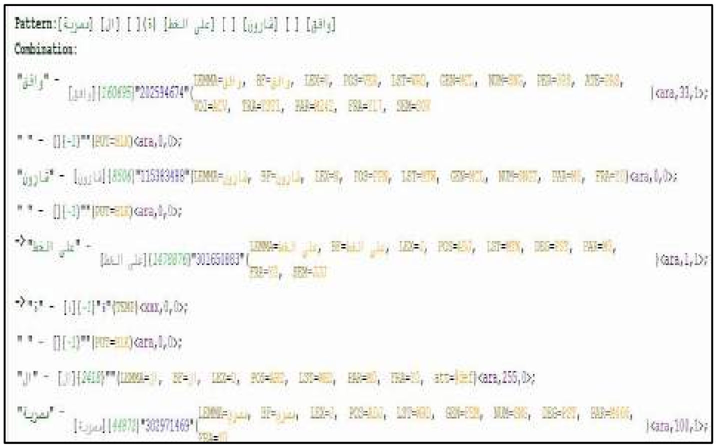
6 A WALK THROUGH AN EXAMPLE

In order to parse the sentence in (17) “وافق شارون على الخطة المصرية”, First, the sentence should be tokenized according to the dictionary.Thus, the sentence will be tokenized to the following pattern:

[وافق][][شارون][][على الخط](ة)[][ال][مصرية]

[9] http://dev.undlfoundation.org/index.jsp

”وافق”, space, ”شارون”,space, “على الخط”,”تاء مربوطة”,space,definite article,”مصرية”and each node is assigned with the appropriate tag, so “وافق”is VER (verb),space is assigned tothe tag BLK (blank), “شارون” is assigned tothe tag PPN (proper noun), “ على الخط“is wrongly tokenized hence assigned to the tag ADJ (adjective), the definite article " ال" is assigned to the tag ART (article) and finally the word “مصرية” is assigned tothe tag ADJ (Adjective), see Fig.9.,whichrepresents the automatic output of this stage.



```
Pattern:[مصرية] [ال] [ ](ة) [على الخط] [ ] [شارون] [ ] [وافق]
Combination:
"وافق" - [وافق]{160695}"202594674"(LEMMA=وافق, BF=وافق, LEX=V, POS=VER, LST=WRD, GEN=MCL, NUM=SNG, PER=3PS, ATE=PAS,
VOI=ACV, TRA=TSTI, PAR=M242, FRA=Y17, SEM=SOV)<ara,33,1>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
"شارون" - [شارون]{8506}"115383488"(LEMMA=شارون, BF=شارون, LEX=N, POS=PPN, LST=MTW, GEN=MCL, NUM=SNGT, PAR=M0, FRA=Y0)<ara,0,0>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
->"على الخط" - [على الخط]{1478876}"301650863"(LEMMA=على الخط, BF=على الخط, LEX=J, POS=ADJ, LST=MTW, DEG=PST, PAR=M0,
FRA=Y0, SEM=[illegible])<ara,1,1>;
->"ة" - [ة]{-1}"ة"(TEMP)<xxx,0,0>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
"ال" - [ال]{2418}""(LEMMA=ال, BF=ال, LEX=D, POS=ART, LST=WRD, PAR=M0, FRA=Y0, att=@def)<ara,255,0>;
"مصرية" - [مصرية]{44872}"302971469"(LEMMA=مصري, BF=مصري, LEX=J, POS=ADJ, LST=WRD, GEN=FEM, NUM=SNG, DEG=PST, PAR=M466,
FRA=Y0)<ara,100,1>;
```

Figure 9: The tokenization and tagging stage

Theadjective[على الخط]should be retokenized as [على], [ال], [خط]and this is the role of the disambiguation rules. A rule should be added to block the sequence of the adjective and”تاء المربوطة ”if it is preceded by a masculine noun (MCL) “شارون”.Such rule has been added as in (1).The rule states that the adjective is blocked; the blocking is stated in the rule bythe symbol (=0).

(1) ({J|V}, %01) (^BLK , ^STAIL, ^ACC) = 0;

So the output would be as shown in Fig.10

```
Blocked at segment:[مصرية][ال][ ](ة)[على الخط][ ][شارون][ ][وافق]
Position index:11
Re-tokenizing: "على الخطة"
Position index:11
Pattern: [مصرية][ال][ ][خطة][ال][ ][على][ ][شارون][ ][وافق]
Current State
Pattern:[مصرية][ال][ ][خطة][ال][ ][على][ ][شارون][ ][وافق]
"202594674"{160695}[وافق] - "وافق"(LEMMA=وافق, BF=وافق, LEX=V, POS=VER, LST=WRD, GEN=MCL, NUM=SNG,
PER=3PS, ATE=PAS, VOI=ACV, TRA=TSTI, PAR=M242, FRA=Y17, SEM=SOV)<ara,33,1>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
"115383488"{8506}[شارون] - "شارون"(LEMMA=شارون, BF=شارون, LEX=N, POS=PPN, LST=MTW, GEN=MCL,
NUM=SNGT, PAR=M0, FRA=Y0)<ara,0,0>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
"2991}[على] - "على}""(LEMMA=على, BF=على, LEX=P, POS=PRE, LST=WRD, PAR=M0, FRA=Y0, att=@on,
rel=PLC)<ara,255,0>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
"2418}[ال] - "ال}""(LEMMA=ال, BF=ال, LEX=D, POS=ART, LST=WRD, PAR=M0, FRA=Y0, att=@def)<ara,255,0>;
"105898568"{39202}[خطة] - "خطة"(LEMMA=خطة, BF=خطة, LEX=N, POS=NOU, LST=WRD, GEN=FEM, NUM=SNG,
PAR=M1, FRA=Y0, ABN=ABT, ALY=ALI, ANI=NANM, CAR=CTB, SEM=CGN, SFR=K0)<ara,169,15>;
" " - [ ]{-1}""(PUT=BLK)<ara,0,0>;
"2418}[ال] - "ال}""(LEMMA=ال, BF=ال, LEX=D, POS=ART, LST=WRD, PAR=M0, FRA=Y0, att=@def)<ara,255,0>;
"302971469"{44872}[مصرية] - "مصرية"(LEMMA=مصري, BF=مصري, LEX=J, POS=ADJ, LST=WRD, GEN=FEM,
NUM=SNG, DEG=PST, PAR=M466, FRA=Y0)<ara,100,1>;
```

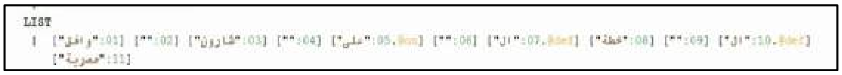
Figure 10:Re-tokenization of the wrongly tokenized adjective

The output of the tokenization stage is as shown in Fig.11.

```
Pattern:[مصرية][ال][ ][خطة][ال][ ][على][ ][شارون][ ][وافق]
```

Figure 11: The output after the tokenization stage

The stage that comes after the tokenization is the transformation stage. In the transformation stage, the researcher has built a set of T-rules to transform the natural language in Fig.13 into a parsed tree.

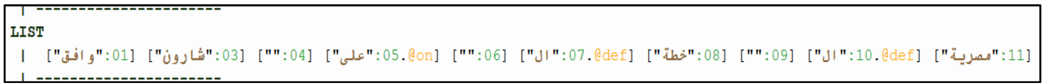


```
LIST
 | ["وافق":01] ["":02] ["شارون":03] ["":04] ["على":05.@on] ["":06] ["ال":07.@def] ["خطة":08] ["":09] ["ال":10.@def]
   ["مصرية":11]
```

Figure 13:The natural language input before starting the parsing

The blank space should be deleted from the input in Fig.13 to prepare the sentence for the parsing step.So, a rule for omittingthe blank space is used as in (2) which states that if there is a blank space beside a word, it should be deleted as in Fig.14. The blank space between the verb “وافق” and “شارون” is suppressed,and the rules will be applied recursively; the blank space beside each node is deleted, so the final output after deleting the blank space from the input will be as in Fig.15.

(2) (Word , %y , ^blk) (BLK , %02) := (%y , +blk);



```
| ----------------------
LIST
 |  ["وافق":01] ["شارون":03] ["":04] ["على":05.@on] ["":06] ["ال":07.@def] ["خطة":08] ["":09] ["ال":10.@def] ["مصرية":11]
 | ----------------------
```

Figure 14: Deleting the space after the word “وافق”

[01:"وافق] [03:"شارون"] ["05:"على"] ["07:"ال] [08:"خطة.]["10:"ال"] ["11:"مصرية]

Figure 15: Deleting the space in the NL input

In this phase, small constituents or trees are constructed for the small phrases (usually noun phrases) in the sentence and then combined to form a bigger tree gradually until the whole sentence is analyzed. First, the noun “خطة” ‘plan’ will be projected to the intermediate constituent (NB) as it is the head of noun phrases as in rule (3). Then, the adjective “مصرية” ‘Egyptian’ will be projected to the intermediate constituent (JB) as in rule (4) then this intermediate constituent will be linked to the definite article “ال”to form the maximal projectionadjective phrase (JP) as in rule (5).Once the adjective phrase is projected to this maximal projection, it will leave the list structure and constitute a part of the syntactic tree.

(3) (N , Word , ^NB , ^PROJ , %x) := (%x , +NB , +PROJ) ;
(4) (J , ^PTP , ^proj , %x) := (%x , +JB , +proj);
(5) (ART,%y)(JB , %x , ^pro) := (JP(“ال” , %y ;%x , +pro) , +JP , +GEN = %x , +DEF = %x , %01) ;

The constructed (JP) will be linked to intermediate constituent (NB)”خطة” ‘plan’ to build a bigger (NB) by the rule in (6) as shown in Fig.16. As there are no other modifiers for the constructed (NB) “الخطة مصرية” ‘Egyptian plan’ will be projected to thecorresponding maximal projection (NP)by the rule in (7) as shown in Fig.17.

(6) (NB , %n) (JP , GEN = %n , def , %adjc) := (NB(%n ; %adjc) , +rel = mod , +NB , +GEN = %n , %01) ;
(7) (ART, %z) NB(NB(%x ; %y),+NB, %01):= (NP(%x , -NB ; %z), NP, %01) (%y) ;

```
String View:
| <SHEAD>
| #L("03:"شارون",01:"وافق)
| #L("05:"على",03:"شارون.@on)
| #L("05:"على.@on,"07:"ال.@def)
| #L("07:"ال.@def,:02)
|   NB:02("01:,08:"خطة)
|    JP:01("13:"ال",11:"مصرية)
| <STAIL>
```

```
| String View:
| <SHEAD>
| #L("03:"شارون",01:"وافق)
|#L("05:"على",03:"شارون.@on)
| #L("05:"على.@on,:03)
|   NP:03(:02,"07:"ال.@def)
          NB:02("01:,08:"خطة)
          JP:01("13:"",11:"مصرية)
```

Figure 16: Building the intermediate constituent “خطة مصرية” (NB)Figure 17: Building the noun phrase “الخطة المصرية” (NP)

The (NP) “خطة مصرية” ‘Egyptian plan’ is preceded by the preposition “على”’on’in the Arabic input sentence, so it will be linkedwith this preposition to form the intermediate projection (PB) by the rule in (8). There are no other modifiers for the constructed (PB) “على الخطة مصرية” ‘on the Egyptian plan’ that will be projected to thecorresponding maximal projection (PP)by rule in (9) as shown in Fig.18.

(8) (P , %p) (NP %adjc) := (PB(%p ; %adjc) , +PB , %01);
(9) (PB , %x , ^pro) := (PP(%x , +pro ; +e , %y) , +PP , %01) ;

```
|  <SHEAD>
|  #L("03:"شارون",01:"وافق)
|  #L("05:,03:"شارون)
|    PP:05(:04,"":18)
|      PB:04("05:"على.@on,:03)
|        NP:03(:02,"07:"ال.@def)
|          NB:02("01:,08:"خطة)
|            JP:01("13:"",11:"مصرية)
|  <STAIL>
| ----------------------
LIST
|  ["05: [03:"شارون"] [01:"وافق{}
```

Figure 18: Building the prepositional phrase "على الخطة المصرية" (PP)and the remaining nodes in the list

The remaining nodes (the processing units) that are not linked to the tree structure are the proper name "شارون"'Sharon' and the verb "وافق" 'agree'. "شارون"'Sharon' will be projected to themaximal projection (NP).The (PP) is marked as in rule(10). The (PP) will be linked to the verb "وافق" 'agree' to form the intermediate constituent (VB) by rule (11), then this intermediate constituent (VB) will be linked to themaximal projection (NP)"شارون"'Sharon' to form the maximal projection (VP) by rule (12).

(10) (V , TSTI , Y1, %v) (NP , ^subj , %x) (PP , %y) := (+subj , %x) (%v ,V , TSTI) (%y , +Arg0) ;
(11) (V, TSTI , %v) (PP , %comp) := (VB(%v ; %comp , +comp) , +verb = %v , +VB , %01) ;
(12) (%x , NP , subj) (VB , %v) := (VP(%v ; %x) ,%01) ;

```
| UW View:
|  <SHEAD>
|    VP:08(:07,:06)
|      VB:07(وافق:01,:05)
|        PP:05(:04,"":18)
|          PB:04(05:على.@on,:03)
|            NP:03(:02,07:ال.@def)
|              NB:02(خطة:08,:01)
|                JP:01(مصرية:11,"ال":13)
|      NP:06(شارون:03,"":20)
|  <STAIL>
```

Figure 19: The automaticsyntactic representation of sentence in (17)

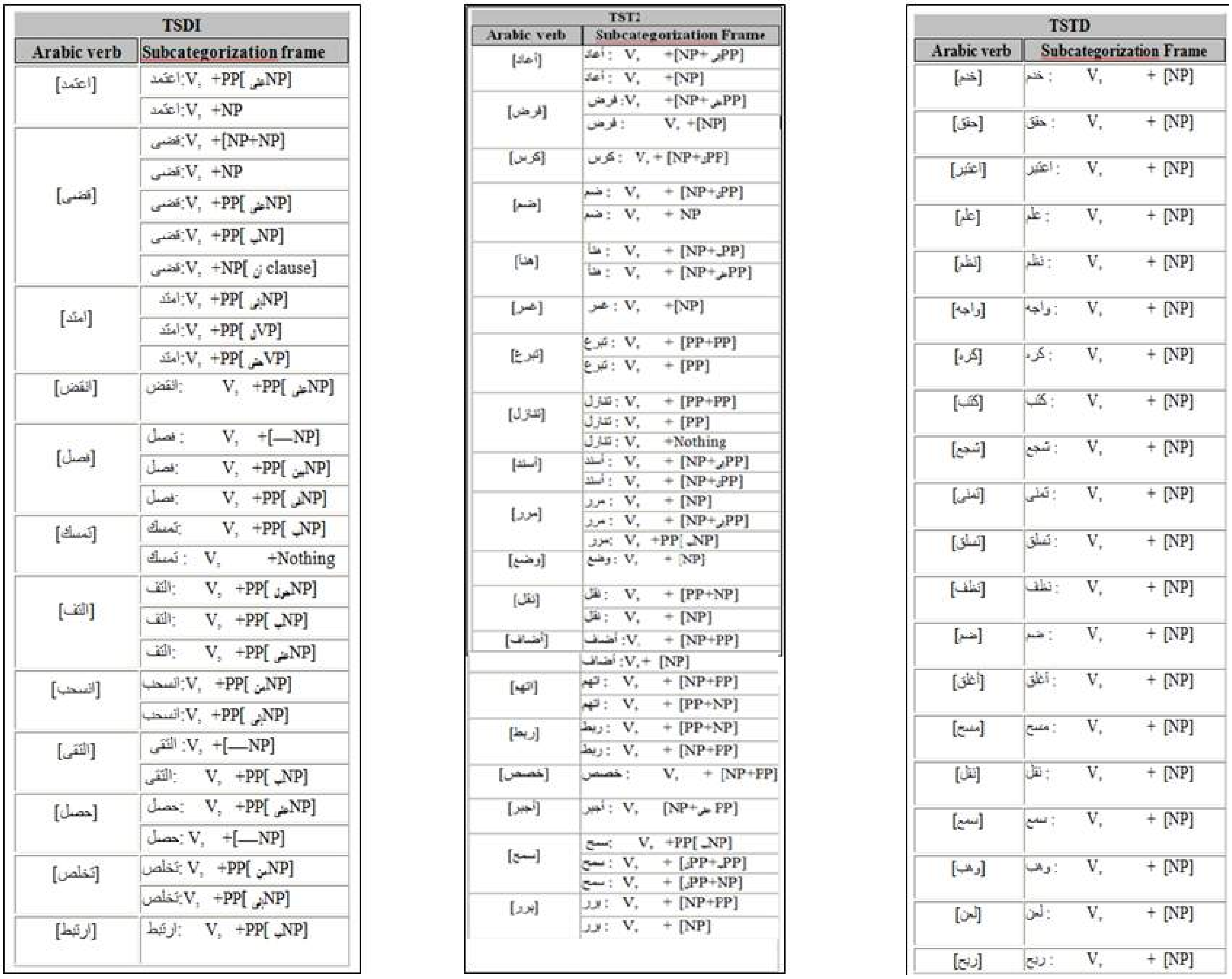
```
----------------------
| Scope Reference: :05
| Current NL string:""
| Original NL string:[]
| Attributes: PP, SCOPE,
,Arg0,comp
| Parent scope::07
| ----------------------
```

Figure 20: The PP which is the complement of the verb "وافق" 'agree'

The (PP) "على الخطة المصرية" 'on the Egyptian plan' is the verb's complement(Arg0)orthe argument of the verb "وافق" 'agree' which is indirect transitive verb (TSTI)as shown in Fig.20. The first NP "شارون"'Sharon' is the subject of the verb(V) "وافق" 'agree'. The subcategorization frame of the verb "وافق" 'agree' was extracted automatically as:
[V, + [—PP]].

7 EVALUATIONS AND RESULTS

Currently, 17 SFs are detected and withlargerdata it is expected to detect more SFs. Ultimately, the researchers expect to provide a large SF dictionary to the NLP community and to train dictionaries for specific corpora. The output has been evaluated,the method that is adopted for the evaluation of the results is the F-measure. It considers both the precision and the recall of the grammar to compute the percentage, according to the formula: F-measure = 2 x ((precision x recall) / (precision + recall)).The accuracy level was high for the TSTI verbs, it was 98%, for the TST2 verbs, the accuracy was also %98; however, for the TSTDverbs, the accuracy was 88%. The research faces many problems in the automatic extraction of the subcategorization frame of the verbs with the TSTD category because the tool fails in detecting the boundaries of the phrases because of the apposition phenomena for instance.Examples of the extracted subcategorization frames are shown in Fig.21.



TSDI	
Arabic verb	Subcategorization frame
[اعتمد]	اعتمد:V, +PP[على NP]
	اعتمد:V, +NP
[قضى]	قضى:V, +[NP+NP]
	قضى:V, +NP
	قضى:V, +PP[على NP]
	قضى:V, +PP[ب NP]
	قضى:V, +NP[ان clause]
[امتد]	امتد:V, +PP[الى NP]
	امتد:V, +PP[ل VP]
	امتد:V, +PP[حتى VP]
[انقض]	انقض: V, +PP[على NP]
[فصل]	فصل : V, +[—NP]
	فصل: V, +PP[بين NP]
	فصل: V, +PP[في NP]
[تمسك]	تمسك: V, +PP[ب NP]
	تمسك : V, +Nothing
[التف]	التف: V, +PP[حول NP]
	التف: V, +PP[ب NP]
	التف: V, +PP[على NP]
[انسحب]	انسحب:V, +PP[من NP]
	انسحب:V, +PP[الى NP]
[التقى]	التقى :V, +[—NP]
	التقى: V, +PP[ب NP]
[حصل]	حصل: V, +PP[على NP]
	حصل: V, +[—NP]
[تخلص]	تخلص: V, +PP[من NP]
	تخلص:V, +PP[الى NP]
[ارتبط]	ارتبط: V, +PP[ب NP]

TST2	
Arabic verb	Subcategorization Frame
[أعاد]	أعاد : V, +[NP+ على PP]
	أعاد : V, +[NP]
[فرض]	فرض :V, +[NP+ على PP]
	فرض : V, +[NP]
[كرس]	كرس : V, + [NP+ل PP]
[ضم]	ضم : V, + [NP+ل PP]
	ضم : V, + NP
[هنأ]	هنأ : V, + [NP+ب PP]
	هنأ : V, + [NP+على PP]
[عمر]	عمر : V, +[NP]
[تبرع]	تبرع : V, + [PP+PP]
	تبرع : V, + [PP]
[تنازل]	تنازل : V, + [PP+PP]
	تنازل : V, + [PP]
	تنازل : V, +Nothing
[أسند]	أسند : V, + [NP+الى PP]
	أسند : V, + [NP+ل PP]
[مرر]	مرر : V, + [NP]
	مرر : V, + [NP+الى PP]
	مرر: V, +PP[ب NP]
[وضع]	وضع : V, + [NP]
[نقل]	نقل : V, + [PP+NP]
	نقل : V, + [NP]
[أضاف]	أضاف :V, + [NP+PP]
	أضاف :V,+ [NP]
[اتهم]	اتهم : V, + [NP+PP]
	اتهم : V, + [PP+NP]
[ربط]	ربط : V, + [PP+NP]
	ربط : V, + [NP+PP]
[خصص]	خصص : V, + [NP+PP]
[أجبر]	أجبر : V, [NP+على PP]
[سمح]	سمح: V, +PP[ب NP]
	سمح : V, + [ل PP+ب PP]
	سمح : V, + [ل PP+NP]
[برر]	برر : V, + [NP+PP]
	برر : V, + [NP]

TSTD	
Arabic verb	Subcategorization Frame
[ختم]	ختم : V, + [NP]
[حقق]	حقق : V, + [NP]
[اعتبر]	اعتبر : V, + [NP]
[علم]	علم : V, + [NP]
[نظم]	نظم : V, + [NP]
[واجه]	واجه : V, + [NP]
[كره]	كره : V, + [NP]
[كتب]	كتب : V, + [NP]
[شجع]	شجع : V, + [NP]
[تمنى]	تمنى : V, + [NP]
[تسلق]	تسلق : V, + [NP]
[تلطف]	تلطف : V, + [NP]
[ضم]	ضم : V, + [NP]
[أغلق]	أغلق : V, + [NP]
[مسح]	مسح : V, + [NP]
[نقل]	نقل : V, + [NP]
[سمع]	سمع : V, + [NP]
[وهب]	وهب : V, + [NP]
[لحن]	لحن : V, + [NP]
[ربح]	ربح : V, + [NP]

Figure 21: Examples of the extracted SCFs lists.

8 CONCLUSION

This paper examined a new technique to automatically extract the subcategorization frames of Arabic verbs. This paper showed that the technique works through tokenizing the input, then parse it and identify the cueswhich will be helpful in identifying the main arguments of the verbs in each context. The identification and encoding of syntactic subcategorization frames is an essential requirement in the construction of computational lexicons.The accuracy level was high for the TSTI verbs, it was 98%, for the TST2 verbs, the accuracy was also %98; however, for the TSTDverbs, the accuracy was 88%.

REFERENCES

[1] T. Brisco and J .Carroll, "Generalized probabilistic LR parsing of natural language (corpora) with unification-based methods". computational linguistics 19:25-59. 1993.

[2] K. Elghamry, "A Generalized Cue-Based Approach to the Automatic Acquisition of Subcategorization Frames", PhD thesis, Indiana University, Bloomington, Indiana. 2004.

[3] C. Manning, "Automatic acquisition of a large subcategorization dictionary from corpora". *In Proceeding of the 31st annual meeting of the association for computational linguistic*s, Columbus,Ohio, pp.235-242.1993.

[4] M. Collins, "Head-driven statistical models for natural language parsing". Computational Linguistics , 29(4). 2003.

[5] S. Montemagni, F. Barsotti, M. Battista, N. Calzolari,O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli,M. Massetani,R. Raffaelli, R. Basili, M.T. Pazienza,D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte, "Building the Italian Syntactic-Semantic Treebank". In Anne Abeill´e, editor, *Building and using Parsed Corpora*. Kluwer, Dordrecht. 2003.
[6] M. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank". Computational Linguistics, 19. 1993.
[7] N. Xue, "Annotating the predicate-argument structure of Chinese nominalizations", *In Proceedings of LREC'06*. 2006.
[8] T. Brisco and J .Carroll, "Automatic extraction of subcategorization from corpora". *In proceeding of the 5th ACL Conference on Applied natural Language Processing*, Washington, DC.356—363.1997
[9] B. Boguraev and T. Brisco, "Large lexicons for natural language processing utilising the grammar coding system of the Longman Dictionary of Contemporary English". Computational Linguistics, 13(4):219–240. 1987
[10] R. Grishman, C. Macleod, and A. Meyers, "Comlex syntax: building a computational lexicon". *In Proceedings International Conference on Computational Linguistics*, COLING-94, pages 268–272. 1994.
[11] G.Chrupala , "Acquiring Verb Subcategorization from Spanish Corpora", PhD program Cognitive Science and Language, Universitat de Barcelona, Department of General Linguistics. 2003.
[12] M. Brent, "Automatic extraction of subcategorization frames from untagged texts". *In proceeding of the 29th Annual Meeting of the ACL.* 209-214. 1991.
[13] M. Brent,, "From grammar to lexicon: Unsupervised learning of lexical syntax". Computational Linguistics.19: 243-262. 1993.
[14] M. Brent, "Surface cues and robust inference as a basis for the early acquisition of subcategorization frames". Lingua 92: 433-470 . 1994.
[15] C. Manning, "Automatic acquisition of a large subcategorization dictionary from corpora*". In Proceeding of the 31st annual meeting of the association for computational linguistics*, Columbus,Ohio, pp.235-242.1993.
[16] A.Ushioda, D. Evans, T.Gibson and A. Waibel, "Estimation of verb subcategorization frame frequencies based on syntactic and multi-dimensional statistical analysis". *In H. Bunt and M. Tomita (eds.), Recent Advances in Parsing Technology*. Dordrecht: Kluwer. 241-254. 1996.
[17] S. Buchholz, "Distinguishing Complements from Adjuncts Using Memory-Based Learning". *In B. Keller ed. 'Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*, 1998.
[18] M. Attia, PavelPecina, LamiaTounsi, Antonio Toral and Josef van Genabith, "Lexical Profiling for Arabic". *In Proceedings of eLex* ,Dublin City University, Dublin, Ireland, pp. 23-33. (2011)
[19] A.Rouvret and J. Vergnaud, "Specifying Reference to the Subject". Linguistic Inquiry, 11(1), 1980.
[20] H. Schütze, "Distributional part-of-speech tagging". *In Proceedings of the 7th Conference of the European Chapter of The Association for Computational Linguistics*, Dublin, Ireland. 1994.
[21] J. Zavrel and J. Veenstra, "The language environment and syntactic word class acquisition". *In F.Wijnen, and C. Koster (eds.), Proceedings of Groningen Assembly on Language Acquisition* (GALA95), Groningen. 1995.
[22] M. Attia, "Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation", Ph.D. Thesis. The University of Manchester, Manchester, UK. 2008.
[23] M. Maamouri and A. Bies, "Developing an Arabic Treebank: Methods, guidelines, procedures, and tools". *In Workshop on Computational Approaches to Arabic Script based Languages*, COLING. (2004).
[24] N. Chomsky, The Minimalist Program. MIT Press.1995.
[25] J. Bresnan, "Lexical-functional syntax". Oxford: Blackwell. 2001.
[26] C. Pollard, and I. Sag, "Head-Driven Phrase Structure Grammar". Chicago: Chicago University Press. 1994.
[27] G. Morrill, "Type-logical grammar". Dordrecht: Kluwer. 1994.
[28] A. K Joshi, and Y. Schabes, *Tree-adjoining grammars*. *In Handbook of formal languages*. G. Rozenberg and A.Salomaa (eds.). Berlin: Springer-Verlag, pp. 69-123.1997.
[29] J. Hwang, "Making Verb Argument Adjunct Distinction in English", Synthesis Exam paper, University of Colorado at Boulder. 2011.
[30] J. Fillmore, "The case for case". In Emmon Bach and R. Harms, editors, Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York, 1968.
[31] R. Jackendoff, "Semantic Interpretation in Generative Grammar". The MIT Press, Cambridge, Massachusetts,1972.
[32] D. Dowty. "On the semantic content of the notion 'thematic role'. Properties, Types and Meaning", 2:69–130, 1989.
[33] J. Fillmore, "Under the circumstances". *In Proceedings of the 20th Annual Meeting of the Berkeley Linguistics Society*, Berkeley, California, 1994.
[34] D. Tesnière, L, *Éléments de syntaxestructurale*. Paris: Klincksieck. 1959.
[35] P. Sells, "Lectures on Contemporary Syntactic Theories", Center forthe Study of Language and Information, Stanford. 1985.

BIOGRAPHIES

Marwa Saber Selim Arafat

Head of Grammar Development unit, Arabic Computational Linguistics Center, Bibliotheca Alexandrina, Alexandria, Egypt.She graduated from Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Egypt. Her MA thesis is in the "The Automatic Extraction of the syntactic arguments of the Arabic verbs in modern standard Arabic". Her main areas of interest are Arabic morphology, syntactic parsing of MSA, semantic analysis, summarization, machine translation and working on Interlingua-based Machine Translation Systems.

She obtained the universal networking language (UNL) certificates; CLEA250, CLEA750, CUP250, and CUP500. She attended the X UNL School organized by the UNDL foundation at Bibliotheca Alexandrina (7-11 October 2012).

She is Developer at UNDL foundation, Geneva- Switzerland. She is a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Universal Networking Language foundation, United Nations, Geneva, Switzerland. She attended many conferences such as Language Engineering Conferences, Ain-Shams University, Cairo-Egypt, (2006,2007,2008,2009,2010, 2011,2012 and 2013), http://www.esole.org (Presence), Arabic Language Technology International Conference (ALTIC), Bibliotheca Alexandria, Alexandria-Egypt 2011, http://www.altec-center.org/conference (Presence), Human Language Technology for Development Conference (HLTD 2011), Bibliotheca Alexandrina, Alexandria, Egypt, May 2 - 5 2011 and The fourth international Arabic linguistic symposium (ALS), Cairo, Egypt.

Dr. SamehAlansary*: Director of Arabic Computational Linguistics Center* Bibliotheca Alexandrina

Dr. SamehAlansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

استخلاص الإطارالنحوي للأفعال العربية آلياً

مروة صابر[1]،سامح الأنصاري[2]

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الإسكندرية، مصر

[1]marwa.saber@bibalex.org
[2]sameh.alansary@bibalex.org

ملخص ـــ الغرض الرئيسي من هذه الدراسة هو الاستخراج الآلي للمتعلقات النحوية للأفعال العربيةمن خلال تصميم محلل نحوي لتحليل التراكيب النحوية من العبارات الفعلية على أساس نظرية *X-bar* .وتهدف هذه الدراسة أيضا إلى اختبار مدى ملاءمةالنظرية النحوية*X-bar* للكشف عن المتعلقات النحويةللمسند.لتحقيق أغراض الدراسة، يعتمد الباحث على المنهج الوصفي التحليلي حيث يتم تحليلمدونة تحتوي على 600جملةمختارة من المدونة العربية باركنسون باستخدام أداة التحليل IAN الذي تم اختياره في هذا الصدد نظرا لارتباطه الوثيق بنظرية *X-bar*والتي*تشير إلى أن* جميع اللغات تشترك في نفس التركيب النحوي العميق.إن استخراج الإطار النحوي(Subcategorization Frame) يعتبرخطوة أساسيةمن أجلالمعالجة الآلية للغات الطبيعية؛ على سبيل المثال، إتاحة معلومات عن الإطار النحوي والتفريق بين ما يتم اعتباره متعلقا أو ملحقا نحويا بالنسبة للمحلل النحوي يساعد في زيادة دقة النتائج.

Syntax - Semantics Interface for Arabic Natural Language Processing

Israa Elhosiny [1], Sameh Alansary [2]

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

Abstract—**The prime purpose of this study is to build prototype rules for analysing Arabic to enable computers to understand the Arabic language. An important step in the language understanding process is constructing a representation of the meaning of a sentence, so the study aims to mapping the syntactic relations with its correspondence semantic relation. Universal Networking Language (UNL) as a language for computer enables computers to process information and knowledge of a language. In order to fulfil the purposes of the study, the researcher adopts an analytical descriptive approach where a grammar of three components using UNL; Morphological, syntactic and semantic components was developed after building a dictionary and a corpus containing 210 verb phrases are analysed using the IAN tool. The linguistic and coverage limitations have been discussed. Finally, the F-measure, as a statistical analysis method is used to measure the accuracy of the grammar as well as its level of adequacy and grammatical competence.**

1 INTRODUCTION

Natural language understanding (NLU) falls into the interdisciplinary field of computational linguistics. Allen (1987) describes several sides to this field. On the one hand, the technological aspect which is concerned with building systems that are able to understand and produce natural language texts in order to make computers smarter and more intuitively usable. The technological aspects are defined by sub-areas such as the development of proper grammatical representations, parsing techniques and knowledge representation mechanisms. On the other hand theoretical linguistics is more interested in producing a structural description of natural language [1].

Understanding a natural sentence should come up with possible interpretations and choose one or more that are most preferred in the current context by integrating information from various knowledge sources. The temporal aspects that determine when each type of information is used is still unresolved issues in modeling human sentence interpretation. Two alternative views have been proposed in psycholinguistic comprehension models. The first view is characterized by serial or syntax-first models (first analysis) which hold that syntax is processed autonomously prior to semantic information [2]. The second view, represented by interactive or constraint-satisfaction models (immediate interaction) which claims that all types of information interact at each stage of language comprehension [3]. Psycholinguistic studies of human sentence processing address the temporal issues of knowledge application attempting to support one side of the modularity debate. The modularity debate is a debate over whether certain decisions such as in syntactic analysis are shielded from the effects of semantic or contextual information or whether they are subject to immediate effects of such types of information.

An important step in the language understanding process is constructing a representation of the meaning of a sentence, given the syntactic structure. Mapping from syntactic structures into a meaning representation is referred to as semantic interpretation or semantic mapping. For this type of representation we need a set of interpretation rules that specifies how to create a meaning representation from the syntax representation [4].

There are many syntactic theories, many semantic theories, and the interface questions look different for all of them. Jackendoff (2002) suggests a view on which semantic structures and syntactic structures are independently generated, and the interface conditions may be quite complex [5]. Most of the theories which do use a compositional formal semantics are non-transformational. In spite of the many advantages of transformational grammars, and their central role in the development of modern linguistic theory, they were never computationally very tractable, nor formally elegant, nor easy to work with for models dealing with how we process language word by word. Non-transformational grammars are compatible with compositional semantics include GPSG [6], HSG, and their descendants, several modern versions of Categorical Grammar [7], Joshi's Tree-Adjoining Grammar [8], and Bresnan and Kaplan's Lexical Functional Grammar [9]. Such theories are particularly popular within the computational linguistics community, where great progress is being made, including much progress in computational formal semantics.

Present-day research is not much concerned with general issues concerning semantic relations. It focuses mostly on specific domains. Systems are designed to analyze texts in a certain field. They use lists of semantic relations specifically tailored to capture salient connections between concepts in the domain. Sometimes they also use lexical resources developed to describe the concepts in the field in question. One of the best known examples is the FrameNet project at the University of Berkeley, California [10],[11], [12].It proposes case frames used to analyze texts pertaining to law. The case frames developed label each participant in a specific type of legal event. The participants are extracted at the intra-clause level as arguments of the verb. For example, the Criminal process frame includes reference to a Suspect which has been arrested by an Authority, and against which are pressed Charges. Another project that has focused on a specific domain is BioText, also at the University of Berkeley [13]. The project aims to identify the relations between the entities in bioscience texts. The authors make use of an ontology of concepts built from medical texts - MeSH (Medical Subject Headings). They mostly focus on the relations between components of nominal compounds. Rosario and Hearst (2001) propose 38 relations, more specific than the generic Agent , Object , etc., specific enough to be useful for their task, for example activity/physical process (virus reproduction),change(disease development), cause (1-2) (food infection), cause (2-1) (flu virus),defect (hormone deficiency), procedure (blood culture), etc.

SNOWY is a knowledge acquisition project developed by Fernando Gomez at the University of Central Florida [14], that processes general texts. The semantic interpretation part of the project is based on a list of thematic roles, an ontology of predicates connected to WordNet's verb classes, in addition to connections between these predicates and WordNet's ontology of nouns. The thematic roles (agent, theme, instrument, etc.) apply to links between verbs and their arguments, and also to nominalized verbs along withtheir modifiers.

Rapid Knowledge Formation (RKF) is a recently concluded project at the University of Texas at Austin whose goal was to develop a system for building complex knowledgebase through the combination of components (events, entities and modifiers) [15]. To describe the relations between these components, the project makes use of a dictionary of relations that describe the interaction between two events (e.g. causal relations), an event and the entities involved (e.g. agent, instrument), an entity and an event (e.g. capability) two entities (e.g. part) or an event or entity and their properties (e.g. duration, size) . These relations cover three syntactic levels and stem from Propbank. The system is used by experts in a certain domain to encode the knowledge in their specific field.

AnCora, is a multilingual corpus annotated at different linguistic levels consisting of 500,000 words in Catalan (AnCora-Ca) and in Spanish (AnCora-Es). Currently, AnCora is the largest multilayer annotated corpus of these languages that is freely available. The two corpora consist mainly of newspaper texts annotated at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses). All of the resulting layers are independent of each other, thus making it easier to manage the data. The annotation can be performed manually, semi-automatically, or fully automatically, depending on the encoded linguistic information. The development of these basic resources constituted a primary objective, since there was a lack of such resources for these languages [16].

Abstract Meaning representation (AMR) makes extensive use of PropBank framesets [17]. For example, it represents a phrase like "bond investor" using the frame "invest-01", even though no verbs appear in the phrase. It is agnostic about how we might want to derive meanings from strings, or vice-versa. In translating sentences to AMR, there is no dictation of a particular sequence of rule applications or provide alignments that reflect such rule sequences. This makes sembanking very fast, and it allows researchers to explore their own ideas about how strings are related to meanings. AMR is heavily biased towards English; it is not an Interlingua [18].

This paper is divided into four sections; section 2exhibits the formal framework of the study; formal rules and tool, section 3 discusses the linguistic framework behind the syntax semantics interface; syntactic and semantic approach as well as relations, section 4 presents the linguistic resources used to make the study; corpus, dictionary and grammar, section 5 discusses the limitations of the study and evaluation. Finally, section 6 concludes the paper.

1 FORMAL FRAMEWORK

Human languages have three main components which enable people to communicate with each other, using their brains as the language processor for those components. The adopted framework called Universal Networking Language (UNL) follows the same logic; consisting of three components that simulate the same components as human languages. The first component is words which are used to express concepts, in the UNL framework they are called "Universal words", also referred to as UWs that are inter-linked with each other to form the UNL expressions of sentences. The second component is the links, which is called "relations" in the framework; they specify the role of each word in a sentence. The third component which is the subjective meanings that are intended by the author are expressed through "attributes" [19]. Every language has its own grammar which describes and governs the linguistic behaviour of the structures of that language. UNL as a language for computers should have its grammar that describes languages in a way that computers can understand; it is the UNL formal grammar.

In order to form a semantic UNL graph, nodes; words are inter-related by relations. Inside each relation, nodes are isolated by a semicolon (;). In the UNL framework, there can be three different types of relations; they are explained in table (1):

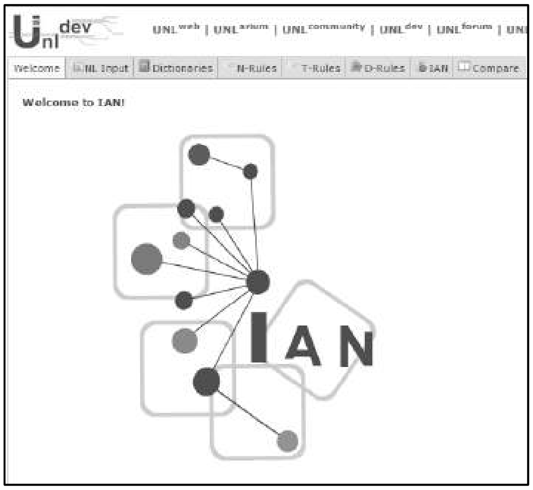
TABLE I

THE DIFFERENT TYPES OF RELATIONS

Relation type	Definition	Form	format
Linear relations (L)	Express the surface structure of natural language sentences.	binary	L(X;Y), or (X)(Y)
Syntactic relations	Express the deep (tree) structure of the natural language sentences, they are not predefined, due to the free use of syntactic theory.	n-ary	rel(X;Y)
Semantic relations	Express the structure of UNL graphs, they constitute a predefined and closed set that are stated in the UNL specs [14].	binary	rel(X;Y)

The system needs a tool that encloses the three components in order to simulate the understanding process that takes place in the human brain. The analysis engine that enables the computers from NLU. IAN[1] [20] is a natural language analysis system that represents natural language sentences morphologically, syntactically and semantically in the UNL format, the application's interface is shown in Fig. 1.

[1] http://dev.undlfoundation.org/analysis/index.jsp



Welcome to IAN!

Figure 1: The interface of "IAN"

2 LINGUISTIC FRAMEWORK

IAN is a flexible environment for linguistic description which can provide a linguistic description for Natural language texts using any linguistic theory [20]. On the syntactic level, any sentence can be analyzed using either the constituency based approach as shown in Fig. 2a, or the dependency based approach as shown in Fig. 2b.

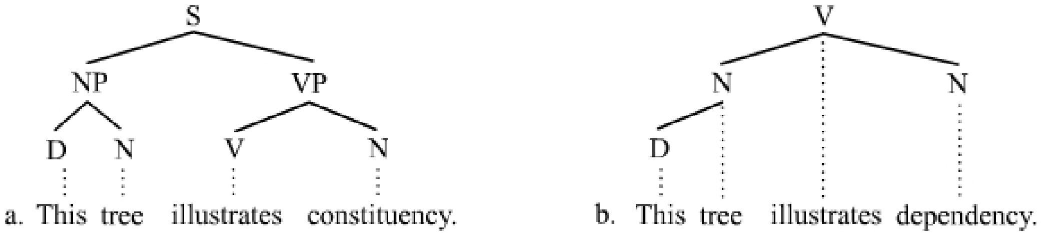

S
NP
VP
D
N
V
N
a. This tree illustrates constituency.
V
N
N
D
b. This tree illustrates dependency.

Figure 2: constituency and dependency representations

IAN's grammar environment allows both approaches for writing grammars. For the paper in hand, the used approach is the dependency based approach [22].

The potential benefits of using dependency-based representations in syntactic parsing, as opposed to the more traditional representations based on constituency. According to Coving ton (2001), dependency parsing offers six advantages [23] :

- Dependency links are close to the semantic relationships needed for the next stage of interpretation; it is not necessary to "read off" head-modifier or head-complement relations from a tree that does not show them directly.
- Dependency tree contains one node per word. Because the parser's job is only to connect existing nodes, not to postulate new ones, the task of parsing is in some sense more straightforward.
- Dependency parsing lends itself to word-at-a-time operation, i.e., parsing by accepting and attaching words one at a time rather than by waiting for complete phrases.
- Dependency relations are close to semantic relations, which facilitate semantic interpretation.
- Dependency representations are more constrained (less complex), which facilitates parsing.
- Dependency representations are more suitable for languages with free or flexible word order.

3 LINGUISTIC RESOURCES

The following sub-sections explain the linguistic resources used to conduct the study. The data collected for the study should have some specific parameters and size which is described in the corpus subsection. The dictionary built for that corpus follows some specifications in terms of the dictionary format and the features needed in building the grammar is explained in the dictionary sub-section. The grammar subsection describes the grammar modules, explained with the example in (2).

A. Corpus

The data is composed of verb phrases that include the selected Arabic verbs. The selected verbs are translated from AnCora 2.0 corpora (see section 1). The semantic annotation of verbal predicates in this corpora implies the systematic mapping between syntax and semantics, basically expressed in the argument structure in Spanish and Catalan.

The corpus is collected from the Egyptian newspaper; Al-Ahram 1999 as it is considered as being representative of modern standard Arabic. The pages of Al-Ahram are collected on the Arabicorpus website; arabiCorpus[2] allows the researcher to search in large, untagged Arabic corpora. 'Untagged' means that the words in the corpora have not been assigned to a particular part of speech. ArabiCorpus is divided into five main categories or genres: Newspapers, Modern Literature, Nonfiction, Egyptian Colloquial, and Pre-modern. User can search any text individually by using the Advanced Search mode. You can even search all of the texts at the same time. It allows search in combined, individual or all texts. The total number of words of the whole ArabiCorpus is: 173,600,000.

In order to collect an appropriate size of data for linguistic analysis, the size of the corpus to be analyzed has to be precisely estimated: it should not be too small, because it would raise the risk of not containing enough data. On the other hand, the corpus should not be too big either, since the time needed for analysis has to be also taken into account when planning corpus building. Sentences are 8 words long to contain all verbs arguments with their modifiers. The average number of sentences is 5 sentences to study each verb differs according to the nature of the verb itself; if it is an intransitive verb, the number of its arguments is less than the transitive verb. The corpus is divided to two sub-corpora; training corpus and test corpus. The training corpus is 150 sentences which cover the proposed verbs classification (see section B). The test corpus contains 60 sentences to test the grammar that is built using the trained corpus.

B. Dictionary

The analysis dictionary is linking nodes of the natural language text to the entries of the NL dictionary in the UNL dictionary format in Fig. 3. It is a word-based dictionary.

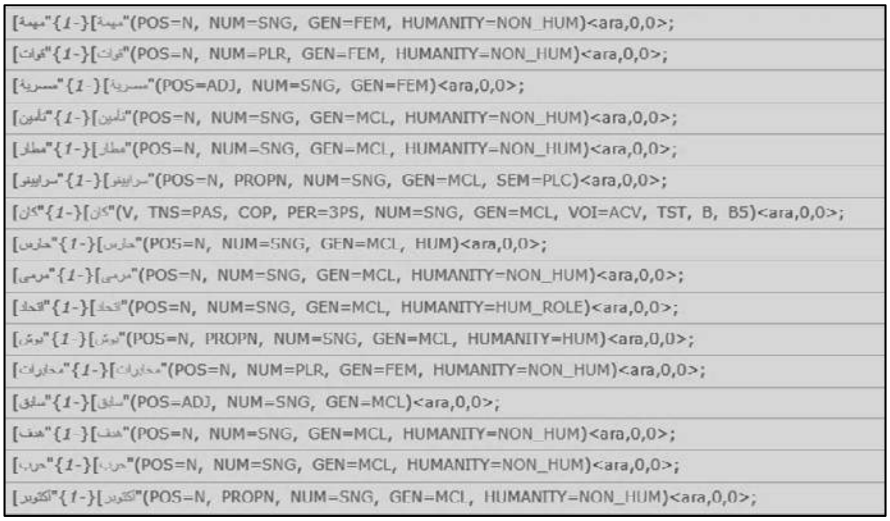
[NLW] {ID} “UW” (ATTR , ...) < FLG , FRE , PRI >; COMMENTS

Figure 3: The Format of the UNL dictionary

where:[HW] is the lexical item of the natural language,[ID] is the unique identifier (primary-key) of the entry, [UW]: is the Universal Word of UNL, For this study, it contains an Arabic word to be appeared in the final output as an understandable word,[ATTR]: is the list of features of the NLW. It can be a list of simple features: NOU, MCL, SNG or a list of attribute-value pairs: POS=NOU, GEN=MCL, NUM=SNG Attributes are separated by “,”, [FLG] is the three-character language code according to ISO 639-3; 'ara', "Arabic" for instance. [FRE] is the frequency of HW in natural texts. It is used for natural language analysis (NL-UNL). It can range from 0 (least frequent) to 255 (most frequent), [PRI]: is the priority of the HW. It is used for natural language generation (UNL-NL). It can range from 0 to 255, and [COMMENT] is any comment necessary to clarify the mapping between NL and UNL entries.

[2] http://arabicorpus.byu.edu/

The researcher has choose to build this dictionary as word-based because it is more appropriate for the analysis task. The dictionary includes all of the corpus words with their part of speech tags and the other needed attributes as shown in Fig. 4.



```
[مهمة"{1-}[مهمة"(POS=N, NUM=SNG, GEN=FEM, HUMANITY=NON_HUM)<ara,0,0>;
[قوات"{1-}[قوات"(POS=N, NUM=PLR, GEN=FEM, HUMANITY=NON_HUM)<ara,0,0>;
[مسرية"{1-}[مسرية"(POS=ADJ, NUM=SNG, GEN=FEM)<ara,0,0>;
[تلبين"{1-}[تلبين"(POS=N, NUM=SNG, GEN=MCL, HUMANITY=NON_HUM)<ara,0,0>;
[مطار"{1-}[مطار"(POS=N, NUM=SNG, GEN=MCL, HUMANITY=NON_HUM)<ara,0,0>;
[سرايفو"{1-}[سرايفو"(POS=N, PROPN, NUM=SNG, GEN=MCL, SEM=PLC)<ara,0,0>;
[كان"{1-}[كان"(V, TNS=PAS, COP, PER=3PS, NUM=SNG, GEN=MCL, VOI=ACV, TST, B, B5)<ara,0,0>;
[حارس"{1-}[حارس"(POS=N, NUM=SNG, GEN=MCL, HUM)<ara,0,0>;
[مرمى"{1-}[مرمى"(POS=N, NUM=SNG, GEN=MCL, HUMANITY=NON_HUM)<ara,0,0>;
[اتحاد"{1-}[اتحاد"(POS=N, NUM=SNG, GEN=MCL, HUMANITY=HUM_ROLE)<ara,0,0>;
[بوش"{1-}[بوش"(POS=N, PROPN, NUM=SNG, GEN=MCL, HUMANITY=HUM)<ara,0,0>;
[محاولات"{1-}[محاولات"(POS=N, NUM=PLR, GEN=FEM, HUMANITY=NON_HUM)<ara,0,0>;
[سابق"{1-}[سابق"(POS=ADJ, NUM=SNG, GEN=MCL)<ara,0,0>;
[هدف"{1-}[هدف"(POS=N, NUM=SNG, GEN=MCL, HUMANITY=NON_HUM)<ara,0,0>;
[حرب"{1-}[حرب"(POS=N, NUM=SNG, GEN=MCL, HUMANITY=NON_HUM)<ara,0,0>;
[اكتوبر"{1-}[اكتوبر"(POS=N, PROPN, NUM=SNG, GEN=MCL, HUMANITY=NON_HUM)<ara,0,0>;
```

Figure 4: The Arabic Dictionary

Attributes concerning, part of speech 'POS', gender 'GEN', number 'NUM' and humanity 'HUMANITY' are assigned to all nouns in the dictionary. While verbs follow a syntactic – semantic classification described below, in table (2).The selected verbs follow the specifications of the semantic annotation of verbal predicates in AnCora 2.0 corpora (see section 1). There are 24 Lexical Semantic Structures (LSSs) compiled and described as appeared in the corpus, grouped around the 4 general event classes; states, activities (or processes), accomplishments and achievements. According to the UNL formalism and relations, verbs divide the event classes to three classes; states, activities (or processes), and achievements. In the accomplishment verbs the subject is mapped to causer semantic relations which does not exist in the UNL semantic relations. According to the UNL framework, especially with relations, the event classes are divided to only three classes; states, activities (or processes), and achievements, as the subject of the accomplishment verbs is mapped to the causer semantic relations which is not exist in the UNL semantic relations; can be expressed by the agent relation and subsequently it is combined to the activities semantic class. Also, LSSs became 15 instead of 24. An example of the omitted LSSs, the B12 "unaccusative-passive-ditransitive" as passive verbs are not included in the corpus.

A, B, and C represent the semantic classes of the verbs, in other words which semantic verb class requires its subject to be mapped to an agent, experiencer, or object. Ax (A1, A2,…,A7) represent the description of the syntactic structure of the semantic class. For example, A4 requires a subject, direct object (N), and indirect second object (PP).

TABLE II

THE ARABIC SYNTAX-SEMANTICS VERBS CLASSIFICATION

Semantic class	Syntactic structure	Mapping schema	example
Activities (A)	subject (A1)	Agent	صام المسلمون
	subject-direct object (A2)	Agent - object	فتح الطالب الباب
	subject-indirect object (A3)	Agent - object	وافق المجلس على المشاركة
	subject- direct object – indirect 2nd object (A4)	Agent – object - goal	الاتحاد يحث المواطنين على الموافقة
	subject- indirect object –direct 2nd object (A5)	Agent-goal -object	طلب من الطلاب المشاركة
	Subject- direct object –indirect 2nd object (A6)	Agent – object - coobject	تفصلهم الحدود عن قراهم
	Subject- PP (A7)	Agent - place	سافر كلينتون إلى نيوجيرسي

States(B)	subject (B1)	Experiencer	بكى الأطفال
	subject-direct object (B2)	Experiencer - object	أحب المصري الأرض
	subject-indirect object (B3)	Experiencer- object	يحتوي القصر على استراحتين
	subject-direct object [amount] (B4)	Experiencer- extension	استغرقت الرحلة ثلاث ساعات
	Subject– predicate (B5)	Attribute (aoj)	كان بوش رئيس المخابرات
Achievements (C)	subject (C1)	Object	هبطت القاعدة
	subject-indirect object (C2)	Object-goal	يتسببالجفاففيالخسائر
	subject-indirect object (C3)	Object-coobject	انفصلتفنلندةعنروسيا

C. Grammar

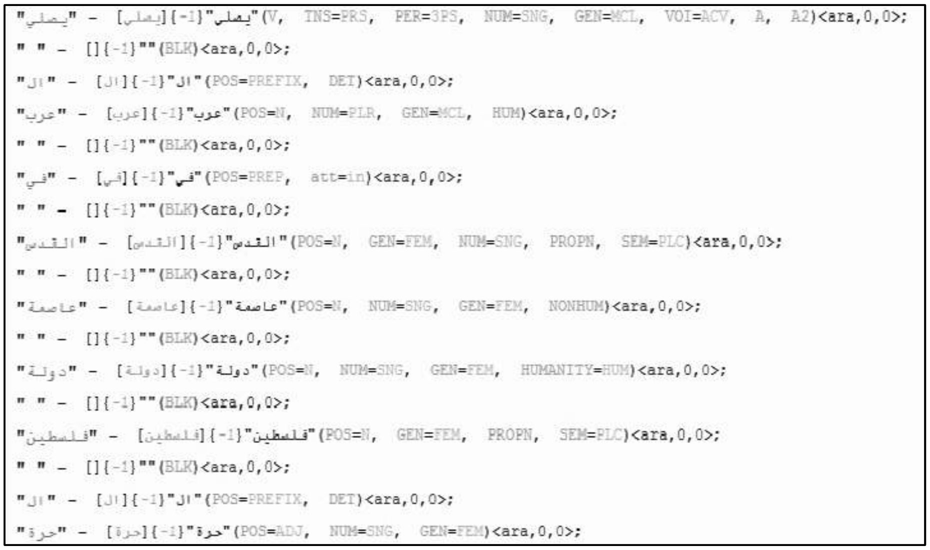
The grammar has several modules such as; morphological, syntactic, and syntax-semantic mapping. The following sub-sections will describe each of the aforementioned modules.

1) *Morphological Module:* As Arabic displays a wide range of inflection and derivation, it gives rise to a large space of morphological variation. The UNL formalism is designed to segment any Natural Language input according to the morphemes stored in the dictionary. This means that UNL deals with any input as a sequence of morphemes (linearly). It cannot deal with the derivational aspect of Arabic in a two-level approach (root + morphological pattern) [24]. Consequently, it is not possible to derive a word from a root although the nature of Arabic morphology is non-linear. Therefore, both of the inflectional and derivational aspects of Arabic should be dealt with concatenatively to be able to adapt Arabic to UNL.This module is perform two tasks. First, for extracting the deep morphological form out of the surface form, for example the two attached prefixes “لل”which are the surface form for “لال”.Rule (1-a) states that, if there are two prepositions “ل” that appear together as prefixes, change the second one to “ال”.Another two examples of extracting the deep form, first, the inflected verb ending “وا” when the connected pronoun attached to it and became “ و”; the “ا” is removed, as in the verb “فصلوها”. Rule (1- b) is dealing with an undefined word in the dictionary as it is in the surface form, so this undefined word which ends with “و” and followed by the connected pronoun “ها” should be changed to “وا” and retrieved from the dictionary using the operator “?”. Second, when a connected pronoun is attached to a real feminine noun that ends with "ة", the "ة" becomes "ت" in the surface form. Rule in (1-c) extracts the deep form and retrieves it as a noun from the dictionary. For example, the noun "مدرسة" is extracted out of "مدرستها" which include both the noun and the connected pronoun "ها".

(1)
a. (“ل” ,%x) (“ل” ,%y)(N,%w):= (%x)(“ال”, DET)(%w);
b. (TEMP,”/…+و/”,%x)(CPRON,%y):= ("وا"<"و",%x,?V)(%y);
c. (TEMP,”/..+ت/”)(%y,CPRON,%y):=(%x,"ة"<"ت",?N)(%y);

The second task of this module is to remove the linguistic obstacles between words to make them ready to be linked in the following module which is the syntactic module. These obstacles are blank spaces, punctuations, the accusative suffix “ا” , and definite article. Example in (2) explains how this module works:

(2) يصلي العرب في القدس عاصمة دولة فلسطين الحرة



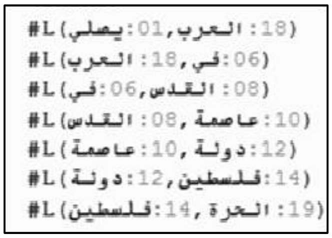
```
"يصلي" - [يصلي]{-1}"يصلي"(V, TNS=PRS, PER=3PS, NUM=SNG, GEN=MCL, VOI=ACV, A, A2)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"ال" - [ال]{-1}"ال"(POS=PREFIX, DET)<ara,0,0>;
"عرب" - [عرب]{-1}"عرب"(POS=N, NUM=PLR, GEN=MCL, HUM)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"في" - [في]{-1}"في"(POS=PREP, att=in)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"القدس" - [القدس]{-1}"القدس"(POS=N, GEN=FEM, NUM=SNG, PROPN, SEM=PLC)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"عاصمة" - [عاصمة]{-1}"عاصمة"(POS=N, NUM=SNG, GEN=FEM, NONHUM)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"دولة" - [دولة]{-1}"دولة"(POS=N, NUM=SNG, GEN=FEM, HUMANITY=HUM)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"فلسطين" - [فلسطين]{-1}"فلسطين"(POS=N, GEN=FEM, PROPN, SEM=PLC)<ara,0,0>;
" " - []{-1}""(BLK)<ara,0,0>;
"ال" - [ال]{-1}"ال"(POS=PREFIX, DET)<ara,0,0>;
"حرة" - [حرة]{-1}"حرة"(POS=ADJ, NUM=SNG, GEN=FEM)<ara,0,0>;
```

Figure 5: The sentence words mapped to their dictionary entries

In the beginning, the words of the input sentence will be mapped with their dictionary entries as in Fig. 5. to prepare the nodes to be linked with a syntactic relation, the obstacles should be removed to facilitate the assignment of the relations. For the sentence in hand, blank spaces and definite article represent those obstacles. First, blank spaces will be removed using the rule in (3a). Second, definite articles "ال" will be combined to the nouns as in "عرب" or adjectives as in "حرة" and assigns the attribute 'DEF' to the nodes using the rule in (3b). After applying the rules of the morphological module to the nodes as shown in Fig.5, nodes will be ready to be linked by the syntactic relations.

(3) a- (%x)(BLK):=(%x);
b- (DET,"ال",%x)({N|J},^DEF,%y):=(%y,-DET,DEF);

The output of the morphological module in Fig.6 represents the words of the sentence in the form of a hypothetical list relations '#L'. ':01', ':18', ..etc and unique IDs are assigned automatically to each word in the sentence.



```
#L(يصلي:01,العرب:18)
#L(العرب:18,في:06)
#L(في:06,القدس:08)
#L(القدس:08,عاصمة:10)
#L(عاصمة:10,دولة:12)
#L(دولة:12,فلسطين:14)
#L(فلسطين:14,الحرة:19)
```

Figure 6: The output of the morphological module

2) *Syntactic Module:* this module depends on the morphological module; it uses the output of the morphological analysis as its input. It is responsible for linking the words of the sentence with syntactic dependency relations. The set of syntactic relations (syntactic tags) that are used in the grammar are those used in the Quranic Arabic Dependency Treebank [25]as shown in table (3). The reason for choosing this set of relations is that it is well-equipped to provide the technical means for describing any syntactic behaviour properly.

TABLE III

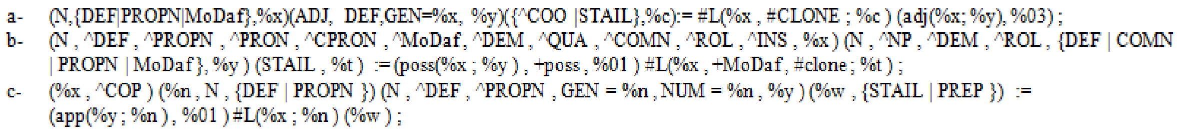
THE DIFFERENT TYPES OF RELATIONS

Relation	Arabic Name	Dependency Relation	Example
adj	صفة	Adjective	فلسطين الحرة
poss	مضاف إليه	Possessive construction	دولة فلسطين
app	بدل	Apposition	القدس عاصمة
spec	تمييز	Specification	ثلاثون جنيها
subj	فاعل	Subject of a verb	أكل الولد
obj	مفعول به	Object of a verb	**أكل** الولد **تفاحة**
subjx	اسم كان	Subject of a special verb or particle	**كانالولد** نشيطا
predx	خبر كان	Predicate of a special verb or particle	**كان**الولد**نشيطا**
gen	جار ومجرور	Preposition phrase	في الحديقة
link	متعلق	PP attachment	**الولدفي** الحديقة
conj	معطوف	Coordinating conjunction	**الولد والبنت**
circ	حال	Circumstantial accusative	يأكل دائما
emph	توكيد	Emphasis	قد أعلن
sub	صلة	Subordinate clause	**المشكلة التي** يعاني منها الشعب

The grammar deals with four main word classes as heads; nouns, prepositions, adjectives and verbs as ordered sub-modules of the syntactic module; nouns sub-modules are located first in the grammar rules, then prepositions, adjectives and finally verbs sub-module, . So, nouns in the sentence in (2) will be linked together first. Then, head nouns will be linked to the verb as an optional or obligatory argument. The grammar is designed to link heads with their modifiers; dependents, then heads with each other.

The grammar is designed in a way that considers that all definite nouns are heads, regardless of how it is defined. It includes proper names, nouns with definite article, and the head of the possessive constructions. For the sentence in hand, the relation 'adj' between the proper noun "فلسطين" as the head and the dependent adjective "الحرة" will be represented first using the rule in (4a) as "فلسطين" is a head and the adjective is at the end of the sentence and it is not modified by any other nodes after. As for the noun "دولة", it is indefinite noun that is followed by a proper noun "فلسطين", so it will be considered as a head in the possessive construction. Thus, they will be linked by the 'poss' relation using the rule in (4b). The noun "دولة" became a definite noun through Idafa. Moreover, the same rule in (4b) will be applied to the nouns "عاصمة" and "دولة". The noun "عاصمة" is also definite and it is preceded by another definite noun "القدس"; a proper name. Both definite nouns have the same gender and number and definiteness, so the grammar will consider this construction as an apposition and link them with 'app' relation as in the rule in (4c).

(4)



```
a-  (N,{DEF|PROPN|MoDaf},%x)(ADJ,  DEF,GEN=%x,  %y)({^COO |STAIL},%c):= #L(%x , #CLONE ; %c ) (adj(%x; %y), %03) ;
b-  (N , ^DEF , ^PROPN , ^PRON , ^CPRON , ^MoDaf, ^DEM , ^QUA , ^COMN , ^ROL , ^INS , %x ) (N , ^NP , ^DEM , ^ROL , {DEF | COMN
    | PROPN | MoDaf}, %y ) (STAIL , %t )  := (poss(%x ; %y ) , +poss , %01 ) #L(%x , +MoDaf, #clone ; %t ) ;
c-  (%x , ^COP ) (%n , N , {DEF | PROPN }) (N , ^DEF , ^PROPN , GEN = %n , NUM = %n , %y ) (%w , {STAIL | PREP })  :=
    (app(%y ; %n ) , %01 ) #L(%x ; %n ) (%w ) ;
```

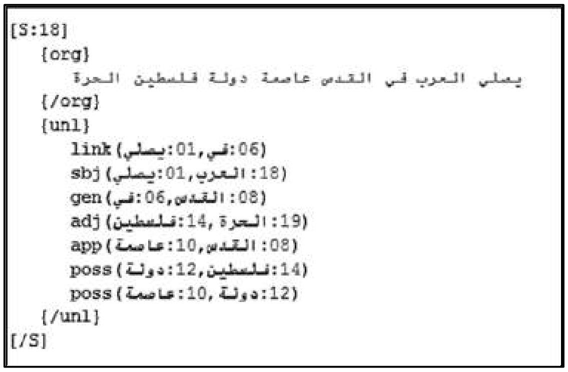
Now, there are four unmodified and unlinked nodes in the sentence; “يصلي”, “العرب”, “في”, and “القدس”. The second sub-module will be applied to the sentence; prepositions as heads. The preposition “في”will be linked to the noun “القدس” with the ‘gen’ relation as stated in rule (5). As the dependent “القدس” is a name of a place, the grammar will be able to predict that the preposition “في” in this sentence is locative. Subsequently, the LOC feature will be assigned to it.

(5) (PREP , {"ب" | "في" }, %x , ^att=%v) (N , {TIM | PLC }, %n) (%w , {PREP | STAIL | N , MoDaf | N , DEF | N , PROPN }) := (gen(%x,+LOC ; %n) , %02) #L(%w , #CLONE ; e) ;

After applying the first two models, three nodes are still unlinked which are the verb “يصلي”, the noun “العرب”, and the preposition “في”. Thus, the rules of the third sub-modules will to be applied; verbs as heads module.As the noun “العرب” agrees with the verb “يصلي” in gender, they will be linked by a ‘sbj’ relation using the rule in (6a). Finally, the remaining unlinked head nodes “يصلي” and “في” will be linked by ‘link’ relation, as the grammar considers that the locative preposition should be linked to the verb using the rule in (6b). The final dependency syntactic graph is shown in Fig. 7.

(6) a- (V , ^COP , ^sbj_assigned, %x) (N , GEN = %x , %y) ({PREP | PART | STAIL | ADV | N , TIM | N , ^NUM = %x | N , MoDaf | N , DEF | PROPN | ADJ , ^DEF }, %e) := #L(%x , +sbj_assigned, #CLONE ; %e) (sbj(%x ; %y) , %03) ;

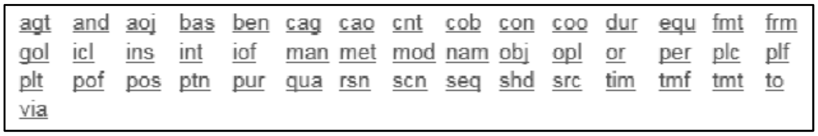
b- (v,%v)(PREP , {"ب" | "في" }, %x , ^att=%v, LOC) (%w , {PREP | STAIL | N , MoDaf | N , DEF |N , PROPN }) := (link(%v ; %x) , %01) #L(%w , #CLONE ; e) ;



```
[S:18]
   {org}
      يصلي العرب في القدس عاصمة دولة فلسطين الحرة
   {/org}
   {unl}
      link(يصلي:01,في:06)
      sbj(يصلي:01,العرب:18)
      gen(في:06,القدس:08)
      adj(فلسطين:14,الحرة:19)
      app(عاصمة:10,القدس:08)
      poss(دولة:12,فلسطين:14)
      poss(عاصمة:10,دولة:12)
   {/unl}
[/S]
```

Figure 7: The syntactic dependency graph for “يصلي العرب في القدس عاصمة دولة فلسطين الحرة”

3) *Syntax-semantic mapping Module:* the task of this module is to transform the syntactic graph to a semantic graph; map each syntactic relation to its corresponding UNL semantic relation. There are 46 UNL semantic relations in the UNL framework which are mentioned below in Fig. 8. The study does not include the nominal semantic relations, as the focus is on the verb arguments rather than the noun modifiers. Therefore, the final result will not include all of the relations that are mentioned below.



agt	and	aoj	bas	ben	cag	cao	cnt	cob	con	coo	dur	equ	fmt	frm
gol	icl	ins	int	iof	man	met	mod	nam	obj	opl	or	per	plc	plf
plt	pof	pos	ptn	pur	qua	rsn	scn	seq	shd	src	tim	tmf	tmt	to
via														

Figure 8: the UNL semantic relation labels

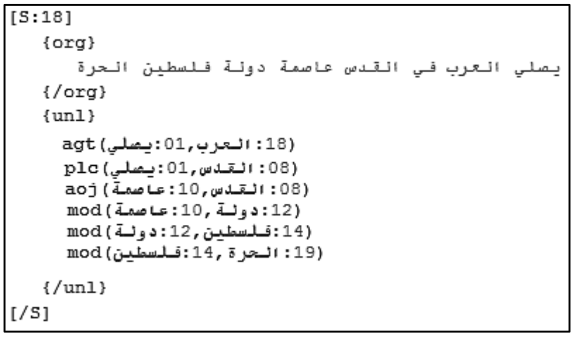
There are three cases in mapping: one relation to one relation mapping, two relations to one relation mapping and one relation to one word with adding a UNL attribute. For the sentence in hand, only the first two cases are presented.

- *One Syntactic Relation to One Semantic Relation Mapping:* since that the adjective and the second element in the possessive construction; 'مضاف إليه' are considered to be a modification of an entity, all ‘poss’ relations between nouns will be mapped with ‘mod’ semantic relation; the relation between “عاصمة” and “دولة” as well as the relation between “دولة” and “فلسطين” as stated in rule (7a).Furthermore, the ‘adj’ relation between “فلسطين” and “الحرة”will be mapped to the ‘mod’ relation as in rule (7b). The ‘app’ relation between "القدس" and "عاصمة" will be mapped with "aoj" semantic relation, since 'aoj' is used to express the predicative relation between the predicate and the subject using the rule in (7d). For the 'sbj' relation between the verb "يصلي" and the noun "العرب", it depends on the syntax-semantic verb classification in the dictionary. This verb is from the class A which implies that its subject 'sbj' is mapped to the agent semantic relation 'agt' as in rule (7c).

(7) a- adj (N , %x ; ADJ , %y) := mod(%x ; %y) ;
b- poss (N , %x ; {N | PRON }, %y) := mod(%x ; %y) ;
c- sbj (V , A , %x ; %y) := agt(%x ; %y) ;
d- app (%x ; %y) := aoj(%x ; %y) ;

- *Two Syntactic Relations to One Semantic Relation Mapping:* the prepositional head in the relation between "في" and "القدس" is a dependent in the relation between "يصلي" and "في", so both of these relations will be mapped to the 'plc' semantic relation between "يصلي" and "القدس". The preposition is expressed through the place semantic relation by using the rule in (8). Fig. 9 shows the final semantic representation that is obtained by IAN tool for the sentence in (2), while Fig. 10 is the graphical view for Fig. 9.

(8) link (V , %x ; PREP , ^att = %x , %y) gen (PREP , LOC, %r ; N , ^TIM , %t) := plc(%x ; %t) ;



```
[S:18]
   {org}
      يصلي العرب في القدس عاصمة دولة فلسطين الحرة
   {/org}
   {unl}
     agt(يصلي:01,العرب:18)
     plc(يصلي:01,القدس:08)
     aoj(عاصمة:10,القدس:08)
     mod(عاصمة:10,دولة:12)
     mod(دولة:12,فلسطين:14)
     mod(فلسطين:14,الحرة:19)

   {/unl}
[/S]
```

Figure 9: The UNL semantic representation (IAN output)

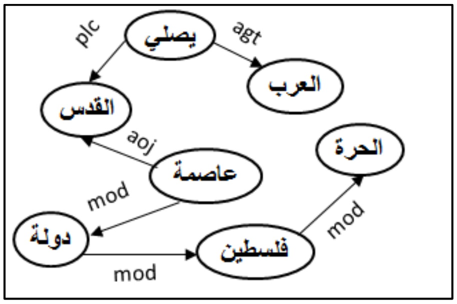


Figure 10: the UNL semantic graph

- *One Syntactic Relation to One word with attribute Mapping:* if the head is an adverb or a quantifier and the dependent is a noun that is linked by the syntactic relation 'poss', the noun will take the UNL attribute of the head. For example, in a structure like "بعض الكتب" and "قبل الشروق", the UNL attributes '@paucal' and '@before' will be assigned to the following words; "الشروق.@before" and "الكتب.@paucal".

Finally, the researcher can conclude that good syntactic representation leads to good semantic representation. Table (4) represent the final syntax-to semantics results for mapping as found in the corpus contains 210 sentences.

TABLE IV

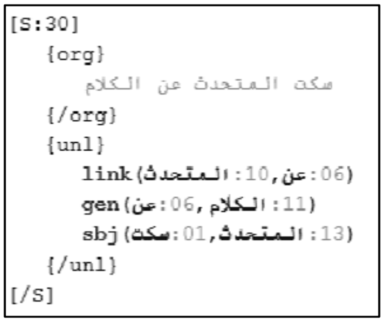
THE MAPPING BETWEEN SYNTACTIC AND SEMANTIC RELATIONS

Syntactic Relation	Semantic Relation	example
sbj	agt	أكلها الإنسان الفقير
	exp	بكى الحاضرون
	obj	سقط المبنى
obj	obj	كتب الولد الدرس
	ext	تستغرق الرحلة ثلاث ساعات

obj2	gol	أعطى أحمد محمد هدية
Link (V) - gen	src	يأكل من شجرة معينة
	pur	يصلي من أجل العثور على المفقودين
	plc	سافر إلى مصر
	coa	يصلي البريء مع المذنب
	tim	نام أطفالها في المساء
	obj	وافق على الاقتراح
	coobj	فصل الليل عن النهار
	gol	طلبوا من العراق التعاون
Link (N) -gen	obj	العثور على المفقودين
	plc	تشتمل خطة التصنيع في المصانع على إنتاج مركبات صغيرة
	mod	يتسبب في خسائر للبنك
poss	mod	هبط الجنود من فوهةالبركان
	@across,@same,@multal, @ paucal, @before	يصلي في أي مكان طاهر - أقنع الصهيونيونبعضمندوبي الدول - لعب حازم في نفسمركزه - سافروا إلى العراق قبلأزمة الخليج .
spec	qua	سيلعب المنتخب ثلاث مباريات
app	@proximal	أظن أن هذاالبورتريه لتشارلز كينيدي
	aoj	يصلي العرب في القدسعاصمة فلسطين
SUBJX-PREDX	aoj	كانت مهمة القوات المصرية تأمين مطار سراييفو
sub	mod	أصبحت هذه المشكلة الإنسانية التي يعاني منها شعب العراق معقدة للغاية
	obj	طلبوا إليه أن يشارك
adj	mod	لعب الأهلي كرةتجارية
Conj-conj	and	أصبح من المشاهد المألوفة والمتكررة
emph	@confirmation	كان حزب العمال قدأعلن هدنة

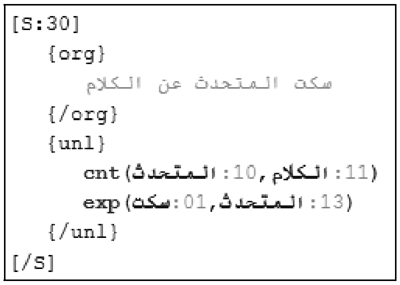
4 LIMITATIONS AND EVALUATION

The limitations of this study can be divided into two categories which are coverage and linguistic limitations. Concerning for the problem of coverage, the study is focusing on verbs and the syntax-semantic mapping of their arguments, this why not all of the UNL nominal semantic relations have appeared in the selected corpus such as 'iof'; "an instance of" and 'pof'; "part of" relations. As for the linguistic limitations, there are some syntactic relations that were not mapped with their corresponding semantic relations. The verb "سكت" is an intransitive verb; however, an unexpected occurrence have been found in the corpus "سكت المتحدث عن الكلام" which is syntactically expressed as shown in Fig. 12. The preposition "عن" should be linked to the verb "سكت" and not with the noun "المتحدث"; however, this verb is encoded as an unergative verb and doesn't have a sub-categorization frame introduced by "عن", so the preposition is linked to the nearest head node after the verb. Subsequently, the link relation between "المتحدث" and "عن" -gen relation between "عن" and "الكلام" going to be mapped to the ‘cnt’; content relation which is not representing the suitable meaning, as in Fig. 11. The corpus includes similar cases of wrong mapping for the verb arguments which cause a mistake ratio about 5 %.



```
[S:30]
   {org}
      سكت المتحدث عن الكلام
   {/org}
   {unl}
      link(المتحدث:10,عن:06)
      gen(عن:06,الكلام:11)
      sbj(سكت:01,المتحدث:13)
   {/unl}
[/S]
```

Figure 11: The syntactic graph for “سكت المتحدث عن الكلام”



```
[S:30]
   {org}
      سكت المتحدث عن الكلام
   {/org}
   {unl}
      cnt(المتحدث:10,الكلام:11)
      exp(سكت:01,المتحدث:13)
   {/unl}
[/S]
```

Figure 12: The semantic mapping for “سكت المتحدث عن الكلام”

The F-measure is primarily considered for the purpose of measuring the mapping grammar accuracy and precision. It integrates two folds: precision and recall, according to which calculations are performed. Precision is the number of correct results divided by the number of all returned results; whereas, recall is the number of correct results divided by the number of results that should have been returned. The upcoming formula explicates the way F-measure is computed:

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

$$= 2 * ((0.98 \times 0.91) / 0.98 \times 0.91)) = 94.8\ \%$$

The grammar displays a high level of success and performance; accuracy of results amounts to 94.8% of the total number of structures analyzed. That is, merely 5.2 % of the corpora fails to be correctly mapped to the semantic representation.

5 CONCLUSION

Building a complete computational system of language understanding is a difficult problem. The task of sentence understanding requires a variety of different types of knowledge; morphological, syntactic and semantic knowledge. In this paper, a text understanding system has been presented which have a flexible architecture that allows any and all available knowledge to be exploited to produce the best interpretations from the available sentence and knowledge available. We have presented the algorithm it is based upon, by examples and discussed the cognitive and computational motivations for the system. The pyshco-linguistic temporal view of Frazier and Fodor which is characterized by serial or syntax-first models (first analysis) has been applied automatically by building the syntax grammar module using the UNL formalism to work alone first, then the mapping module to be applied second, to enable computer from natural language understanding. The built grammar has to be tested in a larger corpus to be more reliable and robust to be more useful as a Natural Language understanding system.

REFERENCES

[1] J. Allen, “Natural Language Understanding”. CA: CA: Benjamin/Cummings, 1987.

[2] L. Frazier and J.D. Fodor, “The sausage machine: A new two-stage parsing model”. Cognition 6, 291-325, 1978.

[3] W. Marslen-Wilson, and L.Tyler, “Against modularity.” In J.L.Garfield (Ed.), *Modularity in Knowledge Representation and Natural Language Understanding*. Cambridge, Mass: MIT Press,1987.

[4] A. Hauptmann, “From Syntax to Meaning in Natural Language Processing*”,inProceeding of the 9th national conference on Artificial intellegance*,1991.

[5] R. Jackendoff, *Foundations of Language: Brain, meaning, grammar, evolution.* Oxford and New York: Oxford University Press, 2002.

[6] G. Gazdar, H. Klein; K. Pullum; A. Sag, “Generalized Phrase Structure Grammar”. Oxford: Blackwell, and Cambridge, MA: Harvard University Press, 1985.

[7] P. Jacobson, “Towards a variable-free semantics”, journal of Linguistics and Philosophy, vol 22, Issue 2 , pp 117-185, 1999.

[8] D. Jurafsky, J.H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, Prentice-Hall, Englewood Cliffs, 2009.

[9] J. Bresnan, The mental representation of grammatical relations, The MIT Press, Cambridge, 1982.

[10] J. Fillmore and B. Atkins, "FrameNet and Lexicographic Relevance"*. In proceedings of the first international conference on language Resources and Evaluation*, Granada, Spain, 1998.

[11] F.Baker,J.Fillmore,andB.Lowe, "The Berkeley framenetproject" , *In COLING-ACL'98:ProceedingsoftheConference, Montreal*.AssociationforComputationalLinguistics, 1998.

[12] D. Gildea and D. Jurafsky. "Automatic Labeling of Semantic Roles" , *In Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, 2000.

[13] B. Rosario, and M. Hearst, " Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy" , *in the Proceedings of EMNLP '01*, Pittsburgh, PA, 2001.

[14] F. Gomez, "Linking WordNet Verb Classes to Semantic Interpretation" .*In Proceedings of the COLING-ACL Workshop on the Usage of WordNet on NLP Systems.*Universite de Montreal, Quebec, Canada, 1998.

[15] P.Clarke, and B.Porter, "Building ConceptRepresentations from Reusable Components", *In Proceedings of the Fourteenth National Conference on Artificial Intelligence*,Menlo Park, 1997.

[16] M. Bertran, , O. Borrega, , M. Marti, , and M.Taule, "AnCoraPipe: A new tool forcorpora annotation", *(Working paper 1: TEXT-MESS 2.0)*. Universitat de Barcelona, 2010.

[17] P. Kingsbury and M,Palmer, "From TreeBank toPropBank" ,*in proceeding of Third International Conference on Language Resources and Evaluation, LREC-02* , Las Palmas, Canary Islands, Spain, 2002.

[18]L.Banarescu, C.Bonial, S. Cai, M. Georgescu, K.Griffitt, U.Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representationfor sembanking" . *In Linguistic Annotation Workshop (LAW VII-ID), ACL*, 2013.

[19] S. Alansary, M. Nagi, N. Adly, "UNL+3: The Gateway to a Fully Operational UNL System", *in Proceeding of 10th Conference on Language Engineering*, Cairo, Egypt, 2010.

[20] S. Alansary, M. Nagi, N. Adly, "IAN: A tool for Natural Language Analysis", *in Proceeding of 12th Conference on Language Engineering*, Cairo, Egypt, 2012.

[21] UNL dictionary specs: http://www.unlweb.net/wiki/Dictionary_Specs (accessed 27 October 2014)

[22] I. Mel′čuk, (1987),"Dependency syntax : theory and practice", Albany: State University Press of New York, 2012.

[23] M. Covington, " A fundamental algorithm for dependencyparsing*", in Proceedings of the 39th Annual ACM Southeast Conference*,ed.John A. Miller and Jeffrey W. Smith, pp. 95-102, 2001.

[24]S. Alansary, "A morphological Analyzer and Generator for Arabic: Covering the Derivational part ", *in Proceeding of NEMLAR International Conference on Arabic language Resources and Tools*, Cairo, Egypt, 2004.

[25] K. Dukes and T. Buckwalter (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In Proceedings of the 7th International Conference on Informatics and Systems (INFOS). Cairo, Egypt.

Israa Elhosiny

Principal Grammar Developer in the Arabic Computational Linguistics Center Bibliotheca Alexandrina.

MA. Student in the Department of Phonetics and Linguistics. Faculty of Arts, Alexandria University. She obtained her BA, Department from Phonetics and Linguistics 2004. She is also Principal Grammar Developer of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. She is working in the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now. She has an experience in morphological analysis and generation and text tokenization. She participated in building grammars using UNL for library information system (LIS) and Knowledge Extraction sYStem (Keys). She is a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Dr. Sameh Alansary*: Director of Arabic Computational Linguistics Center* Bibliotheca Alexandrina

Dr. Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

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

التفاعل بين النحو والدلالة من أجل بناء نظام لفهم اللغة العربية آليا

إسراء الحسيني[1] ,سامح الأنصاري[2]

قسم الصوتيات واللسانيات، كلية الآداب جامعة الإسكندرية

[1]israa.elhosiny@bibalex.org

[2]sameh.alansary@bibalex.org

ملخص – إن الهدف الأساسي لهذا البحث هو بناء نموذجا لقواعد تقوم بتحليل اللغة العربية لتمكين الحاسوب من فهمها آليا. وحيث أن تمثيل المعنى يعتبر ركنا مهما وخطوة أساسية في عملية فهم اللغة، فإن هذه الدراسة تهدف إلى ربط العلاقات النحوية للجملة العربية بنظيرتها من العلاقات الدلالية. وقد قام الباحث بإستخدام لغة الشبكات العالمية (UNL) – وهي نظام مصمم خصيصا لتمكين الحاسوب من معالجة اللغات الطبيعية- كإطار صوري يتم إجراء الدراسة من خلاله، حيث ساعدت في بناء مكونات قواعد التحليل اللغوي ومن ثم بناء نظام الفهم (الشبكات الدلالية). هذه المكونات هي المكون الصرفي والمكون النحوي والمكون الدلالي. وقد تم بناء هذه المكونات بعد اختيار العينة اللغوية التي تحتوي على مائتي وعشرة جملة فعلية وبناء المعجم اللازم لها. أما بالنسبة لقواعد التحليل فقد تم كتابتها باستخدام أداة التحليل اللغوي (IAN) والتي يتم من خلالها تحويل اللغة العربية إلى شبكات دلالية. عند اختبار قواعد التحليل كانت نتائج ربط العلاقات النحوية بنظيرتها الدلالية دقيقة بنسبة % 94.8.

Building a Spoken Arabic Corpus for Egyptian Children Data Collection and Transcription

Heba Salama, Sameh Alansary

Phonetics and linguistics Department, Faculty of Arts Alexandria University
Heba.Salama.slp@gmail.com
Phonetics and linguistics Department, Faculty of Arts Alexandria University
s.alansary@link.net

***Abstract*—This paper aims to build a spoken Arabic corpus for Egyptian children. This corpus is special in many ways.It is the first corpusofa spoken Arabic forEgyptianchildren. It is acollection of longitudinal child language data.It is a speech- based corpus transcribed from recordings of spontaneous conversations. This spontaneous speech transcribed later using the CHAT format as described in the CHILDES(Child Language Exchange System)database. It provides data in consistent fully documented transcription system. The corpus text files transcribed from 10 children (5 boys -5 girls) aged in range from 1.6 (one year and half) to 4 years with about 5 hour recordings.We obtained audio recording of thirty minutes spontaneous speech which produced by children in natural settings. The children divided into five Age groups according to their age. Each group was increase by five months.The recording of children was by a transcriber and parents. The transcripts of spoken interactions providea vast amount of useful data for linguistic, psychological, and sociological studies of child language.Audio data presented in WAV file format. Broad phonetic transcription wasmanually by using CHILDESUnicode and chat program codes of transcription. Transcription based on orthographic conventions of English using IPA symbols. Approximately 15 GB of audio file transcribed.The size of the corpus is nearly 25,645 utterances based on audio files by 10 children.**

Key words:child corpus, CHILDES database.

1 INTRODUCTION

The primary motivation for corpus building has been to provide the data needed to address certain theoretical issues. In particular, corpora have been useful for examining the language as well as characteristics of the input language learner typicallyhears. Corpora of child are invaluable in supporting specific claims within theories of language acquisition. It is possible to use general language corpora as afirst-degree approximation to the input that children receive. [1]Make effective use of a corpus of the wall street journal in this capacity. The shared database can led to advances in methodologies and theory.The easy availability of language corpora and their processing tools have opened up many new areas of language research, which were unknown to us even a few decades ago. Language corpora and the results obtained from them have put intuitive language study under strong challenge. In most cases, intuitive observation provedwrong or inadequate when compared with the findings from corpora. Thus, corpora have proved their utility in empirical language analysis, theory making, as well as in theory modification that were missing in intuitive language study. Corpora have great applications in language research. The importance of corpora to language and linguistics studies aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those, which are subjective, or based upon the individual's own internalized cognitive perception of language. As language and linguistics studies cannot rely on intuition or small samples of language, they require empirical analysis of large database of texts as in the corpus-based approach. Corpus-based methods can used to study a wide variety of topics within linguistics.

Language corpora have long provided a rich source of information about child language. It was appeared in form of diary studies [2] and continue development till today in form of database [3],[4] and tools for computerized corpora (e.g.,[5]), which has led to a recent surge in child language corpora. All the research trends now use computerized data exchange system. Build child language corpora raised as the revolution of computer technology and tools. Thirty-two languages around the world build its child corpora and contribute it to CHILDES(CHIld Language Data Exchange System)database. CHILDES assist in the development of the field of language acquisition research in two major ways. First, the process of systematizing the database for the field will produce a variety of methodological contributions. Secondly, analysis based on processing of the database should be able to classify a wide variety of empirical issues that can lead to the advance of theory construction.The availability of The CHILDES project has great effect on the field of language acquisition research.In addition to, the increase affordability of audio recording equipment, computers and memory that made revolutionary changes in the way of research conduction in the child language field and enhanced the usability of these corpora for addressing research questions at multiple level of linguistic structure (e.g. phonology, morphology, and the lexicon). The CHILDES project is an initiative, which collects transcription datasets from different studies of child language. Any researcher who has made transcriptions of child language can contribute their data to the CHILDES database. It becomes freely available to the entire research community via the project’s website [6]. Moreover, the contributed datasets was available in a standardized encoding format CHAT. There are significant contributions to the literature on child language based on some dataset that is now part of CHILDES. The replication of the findings in these studies is very high. We can download any CHILDES data and check any claims. CHILDES database is valuable for two possible reasons. First, child language data is difficult to collect and the reuse is not easy. Second, the extensive ethical and legal permissions that obtained when a researcher collects samples of child language.

In Arab countries, there are only two Arab countries such as Qatar and Emirates built their child language corpus database and make it available through websites. Unfortunately, Egyptian Arabic corpora for children not built. One reason makes us build Egyptian corpus for children is that there is no previous Egyptian spoken corpus for Egyptian children. This enthuse us to work on this research. We need to develop child corpora in Egyptian Arabic for two needs.The first need is to develop Corpora of child language, whichare essential for investigate the development of child language. The Second need is to document child languagein a standard transcription format CHAT that helps to share data. Transcription was originally a process carried out manually, i.e. with pencil and paper, using an analogue sound recording stored on, e.g., a Compact Cassette. Nowadays, most transcription is on computers. Recordings are usually digital audio or video files, and transcriptions are electronic documents. Specialized computer software exists to assist the transcriber in efficiently creating a digital transcription from a digital recording. The most widely transcription tools used in linguistic research is CLAN (Computerized Language Analysis).CLAN is produce by the CHILDES, whichprovides tools for studying conversational interactions, as well as serving as a repository for language corpora from around the world. CLAN is a software program that used to transcribe sound files using a standard set of rules called “CHAT format.”

In this paper, we construct a spoken Arabic corpus for Egyptian children that based on spontaneous conversation between young children and/or parents, researcher. This work will contribute withthree main contributions.First,it is the first Egyptian corpus for children. Second, all the files are in separated text file in a standard transcription format CHAT. Third,it is constitute a corpus that compile to CHILDES database

2 DATA COLLECTION

A speech sample was a spontaneous speech in unstructured interview. Data elicited through conversation, naming object, pictures around the child in his environment.We usedthe things that children normally use rather than something new, and describe what they were doing while playing. We encourage natural interaction to include all styles such as set with child in his class, playing with the child, interacting while mother, and or teacher teaching various things. Interview increasingly structured when the child was able to produce morphemes e.g. when the child was produced singular noun the investigator and/or mother asked him about plural competence… and so on. Data collected in nursery by investigator (six children), and four at home by mother and/or mother and investigator.

A. Bases of Corpus Data Collection

The corpus data collectedin terms of gender control.We took five boys and fivegirls.The corpus data collected based on spontaneous speech between investigator and /or mother and children. Therecorded corpus based on longitudinal data.

B. Materials

The materials used by investigator and/ or mother to facilitate spontaneous speech production were toys, objects, picture books and stories or just talking without any specific topics. In young children, the activities were children’s daily life at home such astaking shower, wearing clothes. The child must recognize whatever the materials used. We use anything that children normally use rather than using something new.

C. Participants

Ten children (Five Boys and five) were participated in this research. The children selected randomly with no history of delayed language orhealth problems. All children were normal and their first language is Arabic. The children ranged in age from 1.6 to 4 years (mean age 2.77). They divided into five Age groups according to their age. Each group was increase by five months. Group one: from one year and half to two years, Group two: from two years to two years and half, Group three: from two years and half to three years, Group four: from three years to three years and half, Group five: from three years and half to four years.These groups shown in table1 below:

TABLE I: AGE GROUP RANGE

No	Age range	Number of children
1	1.6 -2	One boy one girl
2	2 -2.5	One boy one girl
3	2.5-3	One boy one girl
4	3-3.5	One boy one girl
5	3.5 - 4	One boy one girl
Means	10	

D. Settings

We took audio recording of spontaneous speech produced by children in natural settings, whether in child home or kindergarten.

3 THE RECORDING OF THE CORPUS

Recording took place in a quiet room. The presumed time of recording for each child is to be 30 minutes with total approximately five hours. The length ofinteraction was varying. In young children below two years, recording was on intervals because young children were easily frustrated and much moveable and needs many things during recording such as go toilet or eat. In only one youngboy who is age below two years his mother recorded his speech at home and investigator continue recording at nursery. Another young girl who is age below two her mother recorded 15 minutes and continues 15 in the next day. Recording held by using high quality recorderSony/WM-GX322 and tapes for seven children and three children directly recorded digitally through phones. Each child was informed that heor/she will recorded his or/her speech. All children were happy during recordingtheir speech, play with cassette and/or phone, and wait to hear their voicesafter recording. After recording, the children's audio data saved on computer. Transcripts of the recorded speech made later.

4 PROBLEMS IN DATA COLLECTION

Since collecting data is very large task, because the process of transcribing naturalistic samplesis extremely time-consuming activities. Each child took about 32 hours for transcription with total 320 hours of transcription, with average32hours. We were manually transcribed approximately 25,645 words for all 10 children. This inevitably restricts the amount of spontaneous data that collected and result in the researcher relying on relatively ten samples of child's data.We sum the total number of transcribed words for each child, as shown in table **2**.

TABLE II
SUMMARY OF STATISTICS OF THE CORPUS DATA

CHILD	Age	Line number	Number of investigator items	Number of child items	Total
1	1.7.2	1049	1070+ 311 Mot + Inv	384	1765
2	1.9.20	961	1122	385	1507
3	2.2.18	1345	1448+ 760	706	2914
4	2.4.19	1304	1882	1105	2987
5	2.10	833	1351	627	1978
6	3.0	1272	1949	1277	3226
7	3.5.9	737	657	2402	3059
8	3.5.20	2491	1380	1321	2701
9	3.7.12	1116	1686	1158	2844
10	3.8.1	960	1252	1412	2664
Total		12068	14868	10777	25,645

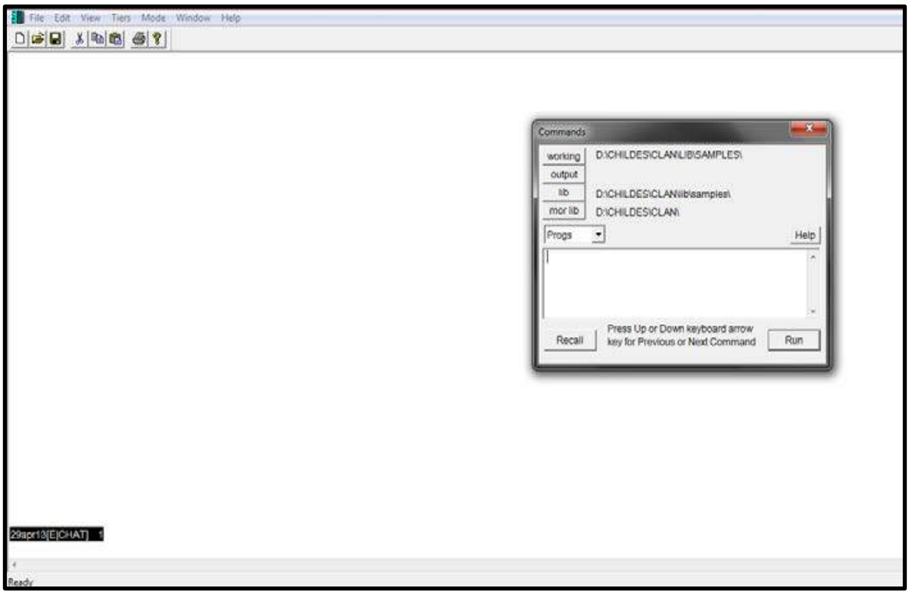
5 METHOD

In this section, we will explain the procedures used and summarize of the steps used in data collection. First we had a permission from the child's mother and/or headmaster of kindergarten fortunately all the mothers and all the headmasters of kindergartensagreed to let their children participate. We spent half an hour in the class interacting with all children gave them candies and sweets to feel familiar. Then we took the child and set in quite room. We audio recorded one young childwho is age1.7 spontaneous speech with his mother. We record one, and/or two child per time. There is no pressure on the child to continuerecording every child show pleasure and acceptance to record. They were happy with cassette recorder especially and the wait to hear their voice after recording. We took notesafter recordingregarding comments on child pronunciation or articulation. In one subject, the investigator set with mother to assure and revise the pronunciation of some words produced by her child. In another child, the researcher asks the mother to write the script of the conversation and the situation of her child recording who is an age1.9 year to check the pronunciation.

6 DATA PRE-PROCESSING

A. Preparing CLAN Software

First, we download the installation files for CLAN from [7]. When we download CLAN, we get CHAT along with it.The CHILDS directory has subfolders: the main ones are CLAN, LIB, and MOR. We also put media file inside the CHILDS directory.After we opened CLAN program we set the working and lib directories before running CLAN. The CLAN window and command window looks like as shown in Fig. 1.



working D:\CHILDES\CLAN\LIB\SAMPLES\
output
lib D:\CHILDES\CLAN\lib\samples\
mor lib D:\CHILDES\CLAN\

Press Up or Down keyboard arrow key for Previous or Next Command

29apr13[E|CHAT] 1

Figure 1: CLAN window.

B. Sound File Editing

We have data from tapes and memory cards. We transferseven recorded children's audio data from tapes in computer (laptop) by using a cable from cassette to computer (laptop) and saved the sound files. We install three reordered children's audio data from mobile memory by using cable. Audio data file should be in .wav format. We used cool edit program to save the file in wav and reduce the big size of file.For example, the size of one child sound file before cool edit was1.34GB and after we applied, become 345MB.In addition, the file rates become 48 Hz, which required for open the file in CLAN program. We use audacity program to remove noise from some audiotaped recording. The 10 digitized sound wave files saved in the subdirectory folder of CHILDES. Moreover, we installed quick time program to enable listen to audio file.We put all the sound files inside the subdirectory folder of CHILDES. Therefore, the program easily finds files when we transcribe it.

C. Text File in CHILDES Format

The sound files transcribed into text files in CHAT format. Our transcriptions conventions follow the CHAT-conventions of the CHILDES database. CHAT is an acronym for Codes for the Human Analysis of Transcripts. The CHAT format specifies a set of rules for transcription and specifies which additional meta-information provided with a transcript file. CHAT provides both basic and advanced formats for transcription and coding. The CHAT system provides a standardized format for producing computerized transcripts of face-to-face conversational interactions. These interactions may involve children and parents, doctors and patients, or teachers.

There are three major components of a CHAT transcript: the file headers, the main tier, and the dependent tiers Headers.

A. Headers

CHAT uses five types of headers: hidden, obligatory, participant-specific, constant, and changeable. We will mention below:

1) Hidden Headers:

There are three hidden headers appear before this header. These are the @Font header, the @UTF-8 header where all files in the database use this header to mark the fact that they are encoded in UTF-8., and the @Color Words which appear in that order.

2) Obligatory Headers:

CHAT has seven initial headers. The first six headers @Begin, @Languages, @Participants, @ID, and @Media @End are obligatory for each file. The last one @End appears at the end of the file as the last line.

```
@Begin
@Language:      ara
@Participants:  CHI     Merna   Target_Child, INVInvestigator
@ID:    ara|sample|CHI|3.8.1|female|||Target_Child||
```

@ID: ara|sample|INV|||||Investigator||
@Media: Merna audio
@End

3) Participant Specific Headers:

The third set of headers provides information specific to each participant. Most of the participant specific information is in the @ID tier.
@Birthplace of #:

4) Constant Headers:

Constant headers mark the name of the file and the background information of the children.
@Location: kindergarten
@Time Duration: 12:30-13:30
@Transcriber: Investigator

5) Changeable Headers:

Changeable headers contain information that can change within the file. Changeable headers occur at the beginning of the file along with theconstant headers.
@Activities: Asking Question, Naming objects, telling stories
@Date:
@Situation

B. Tiers

The content of a file presented in CHILDS as tiers. There is a main tier and several dependent tiers for each line (utterance).

1) Main Tiers:

The main tier is the most important tier because it is where the utterances listed. It is marked with an asterisk (*). After the asterisk, there is a three letters speaker ID, a colon and a tab. The main tiers used in our Arabic corpus include the following:
*INV: utterance of investigator
*CHI: utterance of child
*MOT: utterance of mother
*OTH: utterance of another child during recording

C. Dependent Tiers

Dependent tiers are optional additions to the transcript. These tiers begin with the % symbol and can contain codes and commentary regarding what was said in the Main Tier directly above it. There are some basic rules for coding dependent tiers. We will mention it in the following points.
1- Dependent tiers wrote on separate lines, because the extensive use of complex codes in the main line make it unreadable.
2- All dependent tiers begin with the percent symbol (%) followed by a three-letter code in lowercase letters.
3-The dependent tier code is the percent (%)symbol, followed by a three-letter code ID and a colon such as”mor” for morphology.
4- The text of the dependent tier begins after the tab.

This is an example of Sample file coded in CHAT format:
@Begin
@Language: ara
@Participants: CHI abdrahmanfawzy Target_Child, INV Investigator
@ID: ara|sample|CHI|3.0|male|||Target_Child||
@ID: ara|sample|INV|||||Investigator||
@Location: kindergarten
@Transcriber: Investigator
@Birth of CHI: 8-june-2000
@Age of CHI: 3.0
@Date of Recording: 8-June-2003
@Time duration: one hour and ten minutes
@Activities: asking Question, naming Objects, telling stories, conversation
@Source of transcription: audio tapes
@Media: abdrahmanfawzy audio

*INV: ʔe tæ:ni ?
*CHI: bos̳s̳i
*INV: ʔe: do:l ?
*CHI: 0 [+ trn]
%act: the child looks attoys
*INV: ʔe: do:l ?
*CHI: huh?
@End

7 SAVING .CHA FILE

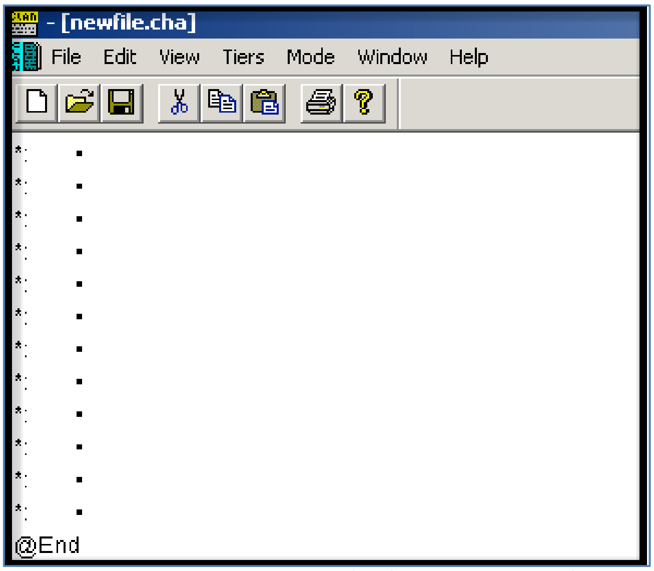
We saved transcriptfile with a complete set of header information with the exact same name as the audio, except with.cha where .cha stand for CHAT fileformat. We saved the transcript file in the same folder. For example, Farah. Cha, Farah. Wav. Then we started the full transcript.

8 LINKING TEXT TO AUDIO

The CLAN editing links transcripts with audio files.Linking text to audio is very important aspect of CHAT/CLAN. It increases transcription accuracy and allows an infinite number of reviews or analyses of our transcript. In addition, it allows us to export a given utterance to COOL EDIT, or PRAAT to do acoustical analysis of what we have recorded. As the acoustic analysis reveals features such as speech rate, articulation of individual words, prosodic contour.

A. The Process of Linking

We will summarize the steps of linking text to audio:
1- First, we press the F5 key to begin linking causes a “bullet” to appear and the cursor move to the next line. We can insert 10 to 20 bullets to start after that, we wrote a transcription related to the recording.
2- At the end of the recording, we type @End.
3- Each time we place the cursor by a “bullet” and press F4, we hear the segment of the recording linked to that line.
4- We press escape-8 for continuous playback. The CLAN highlights each utterance as it played.
5- We press F5 to see the transcript of the segments of the recording linked to the lines transcribed.
When we finished linking the file, the transcript will look like as shown in Fig.2.



```
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
*:    ▪
@End
```

Figure 2: The transcript shape after finished linking the file.

9 GOALS OF TRANSCRIPTION

A transcription system is address two different audiences. One audience is the human audience of transcribers, analysts, and readers. The other audience is the digital computer and its programs. To deal successfully with these two audiences, a system for computerized transcription needs to achieve three goals.These three goals are clarity, readability, and ease data entry.Our transcription conform the three major goals, which the CHAT format designed to achieve [8]. The three goals of transcription are as the following:

A. Clarity

Every symbol used in the coding system should be clear anddefine real world referent. The relation between the referent and the symbolis consistent and reliable. Symbols that sign particular words always spelled in a consistent manner.

Symbols that sign particular conversational patterns should refer to actual patterns consistently observable in the data. Another way of clarity is through the systematicity. Systematicity is a simple extension of clarity across transcripts or corpora. Codes, words, and symbols used in a consistent manner across transcripts. Each code should have a unique meaning independent of the appearance of other codes.

B. Readability

The goal of readability is more important than the goal of clarity of marking. The human language needs to be easy to process, so transcripts need to be easy to read. In the CHILDES system, there are varieties of CHAT options that allow a user to maximize the readability of a transcript.

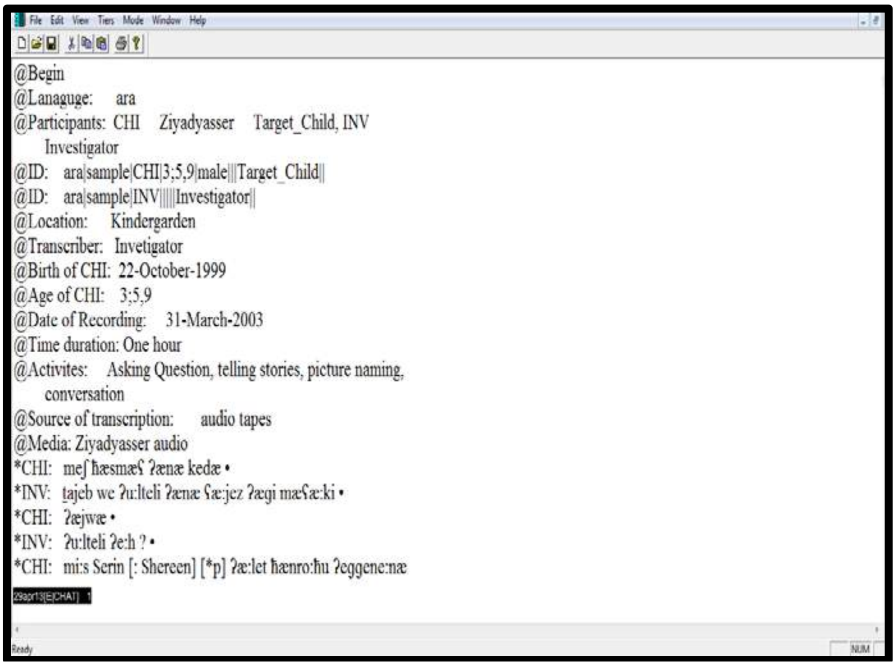
D. Ease Data Entry

When the distinctions increase within a transcription system, data entry becomes increasingly difficult and sensible to error. There are two ways todeal with this problem. The first method tries to simplify the coding schemeand its categories. The problem with this approach is that the loss of clarity. Thesecond method tries to help the transcriber by providing computational aids.Therefore,The CLAN programs follow this path provides systems for the automaticchecking of transcription accuracy.

10 TRANSCRIPTION PROCESS OF AUDIO FILE

A. From sound File to text File

After we preparedall text file, all sound filestranscribed into text files. There are rules for making a transcription. The programs expect transcripts to be in a particular way, or they will not run. Thus, we download a Unicode and use the IPA symbol found in 'keyman' program from CHILDS site keyman [9] program to use in making transcription.Our data transcribed and coded in the CHILDS format [10]. The keyman windows work on screen so it facilitates transcription process. Our recordings transcribed based on orthographic conventions of English using IPA symbols to be easy to access through database in internet.We used broad phonetic transcriptionto look at overall gross structure of the conversation or the relative distribution of turns-at-talk amongst the participants[11].Most of the data included in the CHILDES database document the development of a particular child without focusing on a specific issue. However, the transcripts have been primarily prepared to investigate the development of lexical and morphosyntax phenomena.The transcriber was the investigator and the checker. Extra speaker system used to amplify the sound for more clear sound during transcription.The transcript sample looks like, asshown in Fig.3.



```
@Begin
@Lanaguge:    ara
@Participants: CHI    Ziyadyasser    Target_Child, INV
      Investigator
@ID:   ara|sample|CHI|3;5,9|male|||Target_Child||
@ID:   ara|sample|INV|||||Investigator||
@Location:    Kindergarden
@Transcriber:  Invetigator
@Birth of CHI: 22-October-1999
@Age of CHI:   3;5,9
@Date of Recording:    31-March-2003
@Time duration: One hour
@Activites:    Asking Question, telling stories, picture naming,
      conversation
@Source of transcription:      audio tapes
@Media: Ziyadyasser audio
*CHI:   meʃ ħæsmæʕ ʔænæ kedæ •
*INV:   t̲ajeb we ʔu:lteli ʔænæ ʕæ:jez ʔægi mæʕæ:ki •
*CHI:   ʔæjwæ •
*INV:   ʔu:lteli ʔe:h ? •
*CHI:   mi:s Serin [: Shereen] [*p] ʔæ:let ħænro:ħu ʔeggene:næ
```

Figure 3: Transcriptsample.

As the major goals of CHAT is to maximize systematicity and minimize inconsistency. Mapping the speech of language learners into standard adult forms is not an easy task.Therefore, CHAT provides various tools that mark some of these divergences of child forms from adult standards. CHAT is a powerful program as it tracks a wide variety of structures, compute automatic indices, and analyze morphosyntacx. The basic units of CHAT transcription are the morpheme, the word, and the utterance. There are rules to achieve the goal of consistency for word-level transcription. The programs expect transcripts to look a certain way. We will explain in the following sections the rules of consistency and codes of divergences.

B. Transcription Code for Main Line and Basic Word

In this section, we will discuss the principles for transcribing words and morphemes in the main line. The main Lines begin with a * and the three letters CHI indicate the speaker. After the three letters code is a colon and tab. CLAN automatically put in an asterisk and a colon followed by a tab and then the bullet. The reminder of the main tier line is composed primarily of a serious of words. Words defined as a series of ASCII characters separated by space. All characters that are not punctuation markers areparts of words. The default punctuation set includes the space and these characters are **{, - . -; -?- !- [ ]-**<>}. Not all these characters or the space not used within words. While non-letter characters such as the plus sign (+) or the sign (@) can be used within words to express special meanings. The basic words on Main lines are composed of words and other markers. Words are pronounceable forms, surrounded by spaces. Most words are entered are found in the dictionary. The first word of a sentence not capitalized, unless it is a proper noun. We will present some examples for CHAT conventions used in transcript files, as shown in table 3. To see more example of conventions visit [12].

TABLE III
EXAMPLE OF CHAT CONVENTIONS

CHAT conventions for main line	**Code**	**Example**
Special form marker	@s: second-language	*CHI: merci@s:f
Unintelligible speech	xxx	*MOT: ʔe: dæ ? *CHI: xxx . *MOT: ʔe:h ?
Untranscribed Material	www	*MOT: www. %exp: talks to neighbor on the telephone
Action without speech	0	*MOT: beteʕmeli ʔe fi Fathalla ? *CHI: 0 %act: the child is looking at the book
Phonological Fragments	&	*CHI: &f &f &f fo:ʔʔ.
Non completion of word	text(text)text	*MOT: we dæ ʔe: dæʔ dæ ʔesmu ʔe: dæʔ *CHI: soʕ(bæ:n)
Letters	@l	*INV: betæxo:d ʔe:h ? *CHI: gi:m@l we zæ:l@l zeʔb
Assimilations	[:]	ʔekko:rsi [: ʔelko:rsi]
CHAT conventions forspecial utterance terminators	**Code**	**Example**
Trailing Off	+...	*CHI: ʔæxo ʔelʔerd dæ. bijet̲laʕ foʔʔ +…. *INV: t̲ab bos̲s̲ ʔelfi:l ʕæ:mel ʔe:h
Interruption	+/	*INV: bijeʕmel ʔe: +/ *CHI: we dæ bo:bi kæmæ:n
Explanation	[= text]	*CHI: ʔænæ ʔækteb dæ [= pen]
Retracing	[//]	*CHI: <ʔænæ bælæwwen> [//]bælæwwen beʔælæm [*belʔælæm] [*f:w]
Errors	[*]	*CHI: ʕænd ʔettemr [: ʔennemr] [*p] we ʔettemr [: ʔennemr] [*p] ʕæmmæ:l ʔælʕæb mæʕæ:h. [*p].means phonological error.

11 Conclusions

We introduceda construction ofspoken Arabiccorpus for Egyptian children. We introducedhow to transcribe data with the most widely transcription tools used in linguistic research CLAN program.CLAN is a transcription editor with a large functions produce by the CHILDES. CLAN used to transcribe sound files using a standard set of rules called "CHAT format."

comments that I needed through all stages of this research , for his constant encouragement to do this research the best way I can. I am very lucky to work under his supervision and I am proud to be his student, he was so generous to teach me exactly how to be a linguistics researcher

REFERENCES

[1] Pullum, G. K. and Scholz, B. C. 2002. "Empirical assessment of stimulus poverty arguments".The Linguistic Review 19 (1–2): 9–50.

[2] Darwin, C. 1877. “A Biographical Sketch of an Infant.” Mind 2:285–94.
[3] Deville, G. 1891. “Notes sur le developpement du langage II.” Revue de linguistique et de philologycomparée 24:10–42, 128–43, 242–57, 300–20.
[4]B. MacWhinney, *The CHILDES Project. Tools for Analyzing Talk*. Third Edition.Mahwah, NJ: Lawrence Erlbaum Ass, 2000.
[5]B. MacWhinney, *The CHILDES Project. Tools for Analyzing Talk*. Third Edition.Mahwah, NJ: Lawrence Erlbaum Ass, 2000.
[6]CHILDES database website: http://www.childes.psy.cmu.edu/CLAN/, (accessed 1 March 2013)
[7]MacWhinney (2012). The CHILDES Project: Tools for Analyzing Talk. ElectronicEdition. The CLAN Programs Carnegie Mellon University, available from: http://childs.psy.cmu.edu/manuals/clan/, (accessed 14 March2013).
[8] Machinery (2012). The CHILDS project. Tool for analyzing talk Electronic Edition. The CHAT Transcription format. Carnegie Mellon University available from:http://childs.psy.cmu.edu/manuals/clan/, (accessed 14 March 2013).
[9]Keyman unicodes:http://www.tavultesoft.com/keymanweb/, (accessed 1 March 2013)
[10]CHILDESdatabase website:http://www.childes.psy.cmu.edu/CLAN/, (accessed 1 March 2013)
[11] (Williamson, 2009).
[12] CHAT manual: http://www.childes.psy.cmu.edu/chat/,(accessed 1 March2013)

BIOGRAPHY

Heba Salama is postgraduatestudent in the faculty of Arts phonetics and linguistics department Alexandria University.She is interested in child language research. Her main interest is to collect corpus data to study child language development.She is searching for standard criteria to collect and transcribe data.She likescorpus linguistic because it ismore methodology that is powerful,scientific and open objective verification of results. Electronic corpora have advantages, which is unavailable to their paper based equivalents. The availability of data exchange allows the researcher to answer question by looking for the transcript of spontaneous speech of many data, rather than single study. Sharing data make a revolution in study child language.She was making longitudinal language study with the use of paper. She found that the most obvious advantage of using computer for language study is the speed of processing and the ease of data manipulation. e.g., searching, sorting, and formatting. Advances in computer technology enable to share child language data more readily.The database is very important in helping the researcher to manage the problem they faced and wishes to test a detailed theoretical prediction on naturalistic samples.

Dr. SamehAlansary*: Director of Arabic Computational Linguistics Center*Bibliotheca Alexandrina

Dr. SamehAlansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

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

بناء مدونة لغوية عربية للأطفال المصريين

جمع البيانات

هبه سلامة، سامح الأنصارى

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر

يهدف هذا البحث بناء مدونة لغوية عربية منطوقة للأطفال المصريين وتتميز هذه المدونة بالعديد من الخصائص منها (1) انها اول مدونة لغوية خاصة باللغة المصرية للأطفال. (2) عبارة عن دراسة طولية للغة الطفل.(3) تعتمد علي الكلام التلقائي. تم نسخ الكلام الموجودة في قاعدة البياناتCHAT format التي تقدم نظام كتابة صوتية موحد لسهولة وضعها علي الأنترنت ومشاركة البيانات بين الباحثين. تحتوي المدونة علي 10 ملفات تم تسجيلها من 10 اطفال (5 اولاد– 5بنات) من عمر 1,6 الي 4 سنوات. اجمالي عدد ساعات التسجيل 5 ساعات. تم تسجيل بواسطة الباحث ووالدة الطفل من خلال الانشطة التي يقوم بها الطفل.حجم المدونة 15 جيجيا بايت حوالي 25,645 كلمة. تفيد هذه البيانات في دراسات التحليل اللغوي والسيكولوجي والأجتماعي للأطفال.