Influence of Stemming on Clustering of Arabic Texts: Comparative Study in Document Retrieval

Abdessalem Kelaiaia
Management sciences department
University of May 8, 1945, Guelma, Algeria

Hayet Farida Merouani
LRI Laboratory
Computer sciences department
University of Badji Mokhtar, Annaba, Algeria

ABSTRACT
Initially, this paper, sets out to study the influence of stemming on the quality of the Arabic text clustering, and then describes the testing the application of an approach based on this clustering to improve Document Retrieval (DR). A classical local document system generally, employs statistical methods for calculating the similarity between the introduced query and each document in the target collection to finally provide an ordered list of documents (hit list). In the present approach, the collection is submitted to the clustering process, and then the list of documents returned is constructed from formed clusters based on the nearest representative among the representatives of clusters compared to the user's query. The choice of the Arabic language is motivated by its very particular morpho-syntactic characteristics.

General Terms
Arabic text processing, Information search and retrieval, Clustering, Text indexing.

Keywords
Text, Arabic language, stemming, preprocessing, clustering, local document retrieval.

1. INTRODUCTION
This paper begins with the study of the influence of stemming on Arabic text clustering and then improvements that can be made by this classification within a local document retrieval system. The choice of the Arabic language is essentially due to its very specific morpho-syntactic characteristics [1][2]. The nature of the Arabic language, the writing system, writing orientation, omission of vowels and morphological structure has slowed research into this language, especially in automatic classification. In the literature, most of research is focused on the morphological aspect of this language [2] via developing preprocessing tools such as stemming and their influence on information retrieval or on supervised classification (categorization). But only a small number of research projects focus on document clustering**.** Two major works are identified, on a morphological analysis based on the language and using the n-gram, the authors [3] used a statistical approach (based on the technique of entropy maximization) for the clustering of an Arab-based articles covering several areas such as politics, economics, etc. [4] developed an algorithm (integrated into the standard software clusters TEMIS Insight Discoverer) that, from descriptors in Arabic, contains similar documents in classes according to their semantic similarity and proximity topic.

In order to evaluate the results, a comparison is made with a classical local document retrieval system, which is based in general on the calculation of similarity indices between the user's query and each of the documents in the target collection. Depending on the values calculated for these indices, an ordered list of documents (hit list) is provided to the user. Unfortunately, documents relevant to this query are generally poorly positioned or not present on this list, which does not allow the user to explore them.

2. DOCUMENT CLUSTERING
Clustering is a process of grouping objects represented in the same form in uniform groups (clusters) as a measure of similarity. In document clustering objects become documents (texts) represented, for example, as a bag of words. The need for such classification is explained by the large number of texts that are often contained in a documentary database, and subsequently, the difficulty of document retrieval in this database and the organization and the rapid and efficient exploration of the structure of the database.

Two large categories of clustering algorithms are generally used [5][6]. Hierarchical clustering algorithms, which generate a complete tree from grouping all these elements to be classified in one cluster at the root, to the repartition of each element in its own cluster that represents leaves of the tree. The second category concerns the partitioning clustering algorithms that partition directly into a number of clusters specified in advance.

2.1 Clustering methods
In the present research, two different methods of clustering are used. The first is the clustering by agglomeration, which belongs to the family of hierarchical algorithms; the second is the k-medoïds algorithm and its variant PAM "Partitioning Around Medoïds", which belongs to the family of partitioning algorithms. The choice of these two methods is due to their popularity in the automatic classification community [6].

2.2 Document clustering evaluation
Generally, there are two measures that could be used [5][7] in statistical approaches to assess the quality of clustering. The first is to calculate the overall similarity to measure the density of each cluster. The second is to use measures such as the entropy or the F-measure to compare the structure obtained with others developed in advance.

2.3 Document distance
The similarity between two documents, represented by their respective vectors in the vector space, is calculated using a correlation between their two vectors. This correlation may be one of the two distances widely used: euclidean distance and cosine distance.

The cosine distance is a technique that results from the observation of two vectors. If they have approximately the same attributes, then they point in the same direction in space representation. So to calculate the similarity between two

documents represented by their two respective vectors d_i and d_j using the cosine, it suffices to calculate the cosine of the angle between them, which varies from 0 if the two vectors are the same, to 1 if they are orthogonal to each other [8]:

$$sim(d_i, d_j) = cos(\alpha) = \frac{\sum_k d_{ik}.d_{jk}}{\sqrt{\sum_k d_{ik}^2 . \sum_k d_{jk}^2}} \qquad (1)$$

3. ARABIC LANGUAGE

The Arabic language has an alphabet containing 28 consonants that change their layout according to their position. Unlike English, Arabic is an agglutinative language; articles, prepositions and pronouns stick to adjectives, nouns, verbs and particles which they relate, which creates ambiguities during morphological analysis.

An Arabic word can represent a phrase in English; it may be composed of a stem (base), proclitics such as prepositions or conjunctions, prefixes and suffixes, which express grammatical features and indicate the functions of cases, verb mode and modalities (number, gender, etc.) and enclitics, which are personal pronouns [9]. For example the word أتأكلونها, which mean: "do you eat it?" is decomposed as follows:

Table 1. Arabic word decomposition

Enclitic	Suffix	Stem	Prefix	Proclitic
هَا	ونَ	أكُلُ	تَ	أ

The collage of flexional elements (proclitics, prefixes, suffixes, enclitics) creates patterns [9]. The flexion of a root may generate up to 150 different patterns حمل→ محمل، حامل ، محمول….. This property makes the application of preprocessing techniques such as stemming very useful especially in Information Retrieval (IR) Systems.

4. TEXT PREPROCESSING

Preprocessing aims to standardize the representation of texts to be classified. Various tools such as stemming and lemmatization are used to reduce the ambiguity inherent in natural language. Then a numerical representation (indexing and representation) is given to these texts.

4.1 Stemming

Stemming aims to obtain the lexical root or stem for words in natural language, by removing affixes attached to them, i.e. it's regrouping under a single identification words whose root is common. For example, the words حملة ,محمل ,يحمل are flexions of stem حمل. For this, stemmers are developed; they are generally designed for a specific language on which a certain expertise should be developed. Reference [2] considers that the use of a dictionary for stems and morphological analysis are other forms of stemming. Several stemming algorithms have been studied for different languages. For Arabic, there are several stemmers, the most famous are Al-Stem [10] and StemmerLight10 [2].

In the present research Al-Stem was used due to its performance [11] and re-implemented to work with entire text.

4.2 Indexing and document representation

Whatever the processing to be undergone by a collection of texts, some advance preparation of these texts is needed. The goal is to transform the texts, which are a succession of strings, into a numerical representation easily interpretable and manipulated by this processing. For this purpose, the vector space model wherein each document is represented by a set of indexing terms (attributes) and every word is a dimension of the vector space is chosen. To measure the importance of these terms in the texts, a weighting is given to each. For this, the TF-IDF (Term Frequency, Inverse Document Frequency) weighting "Eq. (2)" [8] is used.

$$d_{ji_{tfidf}} = tf_i^{dj} \bullet log\left(\frac{|D|}{df_i}\right) \qquad (2)$$

tf_i^{dj} represents the weight factor of the term t_i in the document d_j.

$log\left(|D|/df_i\right)$ represents the weight of the term t_i in whole collection.

5. RESEARCH METHODOLOGY

The present methodology is divided into three phases; the first one prepares the parameters around which the present evaluation is built. It attempts to determine the relevance of documents to every query. During this phase the preprocessing required for each text (text cleaning, transliteration, tokenization, removing stop words, stemming and representation as a TF-IDF vector) is also done.

The second phase is devoted to the clustering process applied to the evaluation corpus with the goal of generating different partitions (5, 7, 9, 11, 13 and 15 clusters) with the two selected methods (hierarchical agglomerative method and PAM method). To assess the influence of stemming, the process of clustering is launched, initially, on the raw corpus then on the same corpus that had gone though stemming.

The third phase is dedicated to the studied approach. Here the answer of both following questions must be given :
How deal with natural language queries?
How construct the list of documents to be returned?

5.1 Query treatment

Each query must undergo the same preprocessing as the corpus. For example, the user query

ما هي العلاقات التي قد توجد بين السياحة وعالم الطيران و السفر الجوي

becomes after the preprocessing operations :

Transliterated form:
AlElAqAt Alty twjd AlsyAHp EAlm AlTyrAn Alsfr Aljwy
العلاقات التي توجد السياحة عالم الطيران السفر الجوي

Stemmed form:
ElAq twjd syAH EAlm Tyr sfr jw
علاق توجد سياح عالم طير سفر جو

5.2 Formulation of Retrieved Document

The list of documents returned in response to the query is built according to following algorithm:

a) submit the query to the same preprocessing as the text in the documents;

b) calculate the distance between all cluster representatives and the query: the representative closest to the query wins;
c) the list is constituted of documents in order of their appearance in the cluster of the winning representative; and
d) the list is completed by adding the documents remaining in the cluster in order of their distances from the representative of the winning cluster.

Algorithm 1. Constitution of the list of documents returned

Example
For the user query described above and the partition into 13 clusters (k = 13), obtained by the hierarchical agglomerative method applied on the raw corpus. The representatives of the clusters are: To32, Spo02, Aut09, Rec04, Rel16, Chd03, Chd26, Hm19, SC45, Sc34, To35, TO01 and To37. The nearest representative (determined by distance) from the user query is To37, so the list of returned documents begins with this document and will include the texts that form the cluster with To37 as representative.

Here**,** the distance measure used to measure the distance between two texts and between a text and query is the cosine measure defined in Eq. (1).

6. EVALUATION CRITERIA

6.1 Partitions produced by document clustering

To measure the quality of the structure produced by the clustering approach, the measure called overall similarity, which is equal to the similarity between all the documents of the same cluster taken two-by-two [7] is used. The weighted average similarity is calculated as follows:

$$Q_S = \frac{1}{|S|^2} \bullet \sum_{\substack{d_i \in S \\ d_j \in S}} sim(d_i, d_j) \quad (3)$$

d_i and d_j are two documents of cluster *S*.

6.2 Document retrieval

For assessment of the quality of research carried by the studied approach, the two popular measures used in this field, namely, precision and recall defined, respectively, as in "Eq. (4)" and "Eq. (5)" are employed.

$$Precsion = \frac{RDR}{DR} \quad (4)$$

$$Recall = \frac{RDR}{TRD} \quad (5)$$

RDR represents the number of relevant documents returned by retrieval system.
DR represents the number of documents returned.
TRD represents the total number of relevant documents contained in whole corpus.

7. CORPUS

All experiments conducted in the present work were carried out on CCA corpus (Corpus of Contemporary Arabic) compiled by Latifa El Sulaiti [12] that includes 432 texts.

8. RESULTS AND DISCUSSIONS

For both clustering methods used in the present experiment, the major difficulty lies in choosing the number of clusters that must be made a priori. Then, the studied approach is tested on several partitions.

8.1 Influence of stemming on the quality of the generated clusters

In calculating the overall similarity of the different partitions produced, the first observation is that the stemming has improved this similarity in the clusters obtained (improvement is about 3% with both methods). This is due mostly to the effect of stemming, which helped remove the flexions of words that have the same root, so the documents relating to the same topic will have a greater chance of being in the same cluster.

Table 2. Average overall similarity of clusters

Nb of clusters	5	7	9	11	13	15
Hier. Aggl. with stemming	0,081	0,088	0,093	0,123	0,145	0,152
Hier. Aggl. without stemming	0,046	0,068	0,074	0,087	0,091	0,114
PAM with stemming	0,083	0,085	0,099	0,112	0,142	0,140
PAM without stemming	0,054	0,059	0,084	0,098	0,101	0,108

Also, when the number of clusters is increased overall, the similarity increases and clusters become denser, this tells that the corpus is very divers. Another observation is that neither of the two clustering methods used has produced a clear improvement.

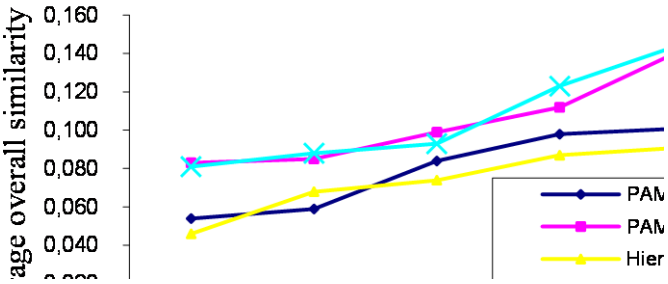

Fig 1: Average overall similarity of clusters

8.2 Influence of number of clusters on the quality of document retrieval with the studied approach

To evaluate the influence of number of clusters on document retrieval with the studied approach, the test with several partitions of different sizes is conducted and the results below are achieved.

For numbers of clusters between 5 and 15 (5, 7, 9, 11, 13, 15), the lowest average precision is achieved with 7 clusters (0.18) and the best with 13 clusters (0.50) (Table 3, Figure 2), when the hierarchical agglomerative method is used on a stemmed corpus. When the corpus is clustered with the PAM method, the worst average precision is obtained with 9 clusters (0.21) and the best with 13 clusters (0.41).

Table 3. Average precision of the top 10 returned documents with the studied approach on some clusters

Clustering methods	Number of clusters					
	5	7	9	11	13	15
Hier. Aggl. with stemming	0,22	0,18	0,21	0,44	0,50	0,48
PAM with stemming	0,25	0,23	0,21	0,40	0,41	0,36
Hier. Aggl. without stemming	0,13	0,16	0,21	0,34	0,37	0,38
PAM without stemming	0,14	0,14	0,27	0,34	0,33	0,35

In Figure 2, the precision is generally improved when the number of clusters is increased. However, the best precision is achieved with the partition within 13 clusters, which means that the best result is not always obtained by a high number of clusters.

With raw texts, the worst average precision is obtained with 5 clusters (0.13) and the best with 15 clusters (0.38) (Table 3, Figure 2) when using the hierarchical agglomerative method. With the PAM method, the lowest average precision is obtained with 5 and 7 clusters (0.14) and the best with 15 clusters (0.35). This confirms the result for stemmed texts, i.e., the precision improves as the number of clusters increases.

Finally, in both of the cases (stem texts and raw texts), the influence of the clustering process on the quality of a document retrieval system is obvious; this is mainly due to the quality of clusters obtained with both clustering techniques used.

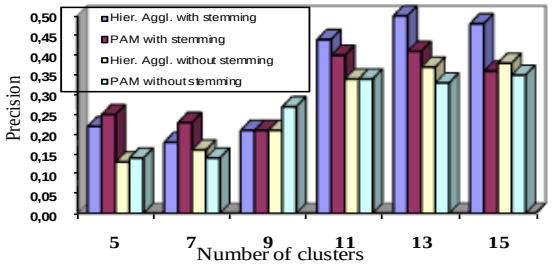

Fig 2: Influence of the number of clusters on the quality of document retrieval with the studied approach

8.3 Influence of stemming on the quality of document retrieval with the studied approach

Whatever the method used for clustering, the comparison between the results obtained on stemmed texts and raw texts, demonstrate the positive impact of stemming on the quality of the document retrieval. That is, there is an improvement in the average precision equal to 7.3% in the hierarchical agglomerative method and 4.8% using the PAM method (Table 3, Figure 2). This is directly linked to the quality of the generated clusters.

8.4 Improvement of the studied approach on document retrieval

From the results obtained with different partitions, a comparison between an average of precision and recall obtained with the 3 best scores (average of 11, 13 and 15 clusters) of the top 10 documents returned by the studied approach (using the two methods of clustering) and those obtained by a classical document retrieval (CDR) system is conducted. An improvement in precision estimated at 13% for the first method (Table 4, Figure 3) and 5% for the second is noticed. On the other hand, there was an improvement in recall estimated at 5% for the first method (Table 5, Figure 5) and 1% for the second.

These results were obtained on the corpus after stemming; the results for the raw corpus represent an improvement in precision estimated at 5% for the hierarchical agglomerative method (Table 4, Figure 4) and 3% for the PAM method, compared to the classical document retrieval system. On the other hand, there was an improvement in recall estimated at 3% for the hierarchical agglomerative method (Table 5, Figure 6). Unfortunately, there is no improvement with the PAM method.

8.4.1 Average precision

Table 4. Average precision of top returned documents

Top returned documents	5	10	20	30	40	50	100	200
Hier. Aggl. with stemming	0,47	0,47	0,35	0,35	0,30	0,26	0,15	0,08
PAM with stemming	0,40	0,39	0,35	0,33	0,29	0,26	0,15	0,08
CDR system	0,43	0,34	0,26	0,21	0,20	0,19	0,12	0,06
Hier. Aggl. without stemming	0,42	0,36	0,34	0,32	0,27	0,25	0,15	0,08
PAM without stemming	0,35	0,34	0,33	0,32	0,29	0,26	0,17	0,09
CDR system	0,39	0,31	0,24	0,20	0,17	0,15	0,10	0,05

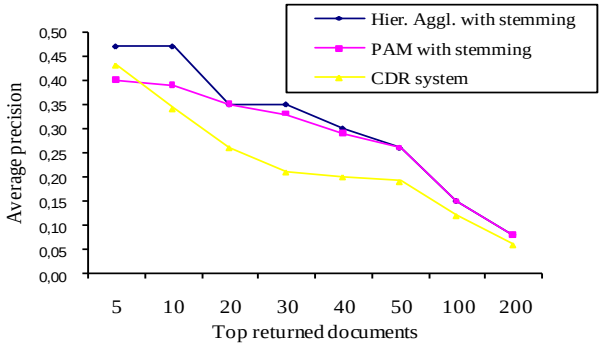

Fig 3: Average precision of top returned documents (with stemming)

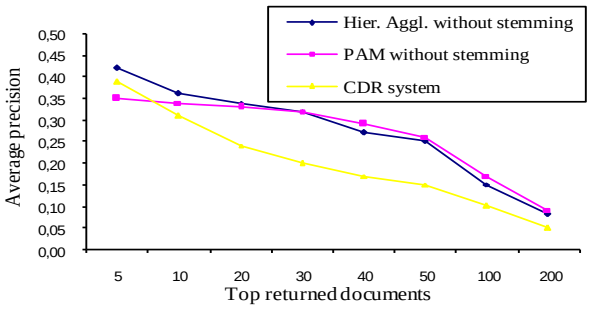

Fig 4: Average precision of top returned documents (without stemming)

8.4.2 Average recall

Table 5. Average recall of top returned documents

Top returned documents	5	10	20	30	40	50	100	200
Hier. Aggl. with stemming	0,15	0,30	0,41	0,61	0,69	0,76	0,88	0,98
PAM with stemming	0,13	0,26	0,42	0,60	0,68	0,75	0,87	1,00
CDR system	0,17	0,25	0,37	0,43	0,51	0,60	0,75	0,82
Hier. Aggl. without stemming	0,16	0,26	0,45	0,60	0,67	0,75	0,91	0,96
PAM without stemming	0,12	0,23	0,42	0,58	0,70	0,79	0,91	0,92
CDR system	0,15	0,23	0,33	0,40	0,45	0,49	0,58	0,63

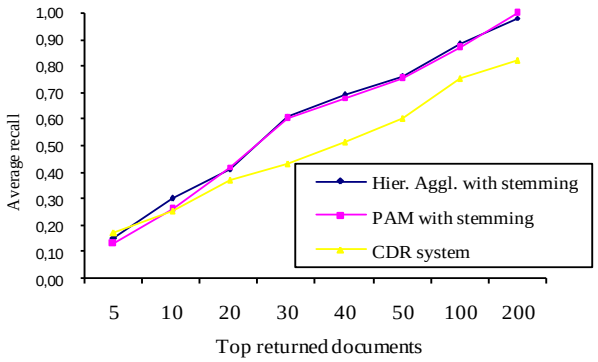

Fig 5: Average recall of top returned documents (with stemming)

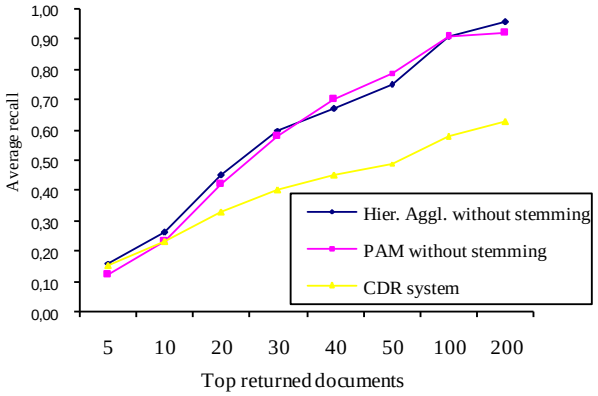

Fig 6: Average recall of top returned documents (without stemming)

8.5 The effect of clustering and stemming on retrieval performance

The improvements in retrieval performance described in the previous section are due mainly to the process of clustering, but partly also to the stemming process. Indeed, with a clustering of the raw corpus, documents relating to the same topic have a high probability of being in the same cluster. With stemming, the probability is increased, so if a representative of a cluster is relevant to a query, its neighborhood will be as well, since this neighborhood was built by involving all the attributes representing documents. This increases the chances of having more relevant documents at the top of the returned list; this is unlike a classical system in which only the terms in the queries are used in the search. Thus, the use of stemming and clustering in combination limits the extent of the search needed to find a predetermined number of documents the search for relevant documents.

It must to stress the importance of the number of clusters, which as seen, is a very important factor in the studied approach. Indeed, it was noted that given an adequately large number of clusters, the results are satisfactory. This is explained by the "specialization of clusters", i.e., with the increase in this number the topics fall into distinct clusters, making rapprochement with a query easier and more fruitful.

9. CONCLUSION AND FUTURE PLANS

During this study authors found that the Arabic language text in experimental corpus reacted well to the stemming process in generating clusters that are of better quality after stemming than in the raw state. Such improvement in performance is due to the fact that stemming attenuates the flexional characteristics of the Arabic language despite the ambiguities that may result in some cases. These cases are more than compensated for by the high rate of correct stems extracted.

It has been demonstrated that the improvement that clustering has made in the two methods of clustering used here can make in document retrieval on raw texts and even on texts having undergone stemming. This is supported by the improvement of the precision and recall. The experiments were conducted with several partitions in clusters producing different results, reflecting the importance of the choice of the number of clusters.

The present work was conducted on a relatively small corpus. It is needed to test the present approach on a larger corpus, which requires more techniques for selecting the number of clusters and the number of documents to be returned, since the clusters may be bigger than in the present experiments so far.

The authors have not yet been able to explore the effect of reducing the size of the representation vectors on the quality of clusters in the retrieval of Arabic documents; this task remains first on future work.

10. REFERENCES

[1] Aljlayl, M., and Frieder, O. 2002. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. In the International Conference on Information and Knowledge Management (CIKM), Virginia, USA.

[2] Larkey, L.S., Ballesteros, L., and Connell, M.E. 2007. Light Stemming for Arabic Information Retrieval. Arabic Computational Morphology, book chapter, Springer.

[3] Sawaf, H., Zaplo, J. and Ney, H. 2001. Statistical Classification Methods for Arabic News Articles. In proceedings of the ACL/EACL Workshop on ARABIC Language Processing: Status and Prospects, Toulouse, France.

[4] Huot, Ch., and Coupet, P. 2005. Le Text Mining sur la langue Arabe : application au traitement des sources ouvertes. TEMIS SA, Paris, France.

[5] Jain, A.K., Murty, M.N., and Flynn, P.J. 1999. Data Clustering: A Review. ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323.

[6] Jardino, M. 2004. Recherche de structures latentes dans des partitions de textes de 2 à K classes. 7es Journées

internationales d'Analyse statistique des Données Textuelles, France, pp. 661-671.

[7] Steinbach, M., Karypis, G., and Kumar, V. 2000. A Comparison of Document Clustering Techniques. In KDD Workshop, Text Mining, Minnesota, USA.

[8] Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management, Vol. 24 (5), pp. 513-523.

[9] Diab, M., Hacioglu, K., and Jurafsky, D. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL'04), USA, pp. 149-152.

[10] Darwish, K., and Oard, D. W. 2002. Evidence combination for Arabic-English retrieval. In TREC, Gaithersburg: NIST, USA, pp. 703-710.

[11] Darwish, K., Hassan, H., and Emam, O. 2005. Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval. In proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, USA, pp. 25–30.

[12] El Sulaiti, L. 2003. L'arabe contemporain. Radio Qatar, Qatar.