Language Variation as a Context for Information Retrieval

Ahmed Abdelali[1], Jim Cowie [1], Hamdy S. Soliman [2]

[1] Computing Research Laboratory New Mexico State University
Box 30001/MSC 3CRL
Las Cruces, NM 88001
{ahmed,jcowie}@crl.nmsu.edu
[2] Computer Science Department, New Mexico Institute of Mining and Technology
Socorro, NM 87801-0389, USA
{hss}@nmt.edu

Abstract. Speakers of widespread languages may encounter problems in information retrieval and document understanding when they access documents in the same language from another country. The work described here focuses on the development of resources to support improved document retrieval and understanding by users of Modern Standard Arabic (MSA). The lexicon of an Egyptian Arabic speaker and the lexicon of an Algerian Arabic speaker overlap, but there are many lexical tokens which are not shared, or which mean different things to the two speakers. These differences give us a context for information retrieval which can improve retrieval performance and also enhance document understanding after retrieval.

The availability of a suitable corpus is a key for much objective research. In this paper we present the results of experiments in building a corpus for Modern Standard Arabic (MSA) using data available on the World Wide Web. We selected samples of online published newspapers from different Arabic countries. We demonstrate the completeness and the representativeness of this corpus using standard metrics and show its suitability for Language engineering experiments. The results of the experiments show that is possible to link an Arabic document to a specific region based on information induced from its vocabulary.

1 Introduction

Arabic is currently the sixth most widely spoken language in the world. The estimated number of Arabic speakers is 250 million of which roughly 195 million are first language speakers and 55 million are second language speakers [23, 24]. Arabic is an official language in more than 22 countries. Since it is also the language of religious instructions in Islam, many more speakers have at least a passive knowledge of the language [25].

It is important to realize that what we typically refer to as "Arabic" is not single linguistic variety; rather, it is a collection of different dialects and sociolects. Classical Arabic is an older, literary form of the language, exemplified by the type of Arabic used in the Quran. MSA is a version of Classical Arabic with a modernized vocabulary. MSA is a formal standard common to all Arabic-speaking countries. It is the language used in the media (newspapers, radio, TV), in official speeches, in courtrooms, and generally speaking, in any kind of formal communication. However, it is

not used for everyday, informal communication which is typically carried out in one of the local dialects.
With the huge amount of data published daily in Arabic over the net and other media [19], it becomes necessary to develop a tool that would help query Arabic text. Studying the language variations or identifying the different geographical sub-areas within the area where the language is being used will help us know more about the language. The study will provide guidance for the right way to search or query in Arabic, avoiding some of the ambiguity that can be generated from the different senses and the different usages the word might have. The first step is to identify these areas and study how much the differences are compared to the national common language used all over the Arab world.

2 Users and Context

Any language speaker using a search engine has his/her own set of beliefs about the language and content of the documents he/she is searching. However, for the most commonly used languages there are large divergences between usage in different countries and regions. This despite the existence of standardization bodies and accepted standards for teaching.
We use the term Search Language Context (SLC) in this paper to identify the pairing(s) of the language of the searcher and the language of the materials being searched. SLC applied to both cross language and monolingual search. For Arabic, for example a speaker of Egyptian Arabic searching sources from Oman and Algeria has a SLC which consist of ($E_{Speaker}O_{Target}$ | $E_{Speaker}A_{Target}$). Making the assumptions that there is a 1 to 1 mapping from speaker vocabulary to target vocabulary can cause reduction in the recall, due to missing terms (Speaker Words != Target Words) and reduce imprecision due to terms having different usage, (Speaker words = Different sense target words).
Equivalent English examples would be American English (AE) and British English (BE) for example:

- Sidewalk does not exist in BE
- Pavement (road surface) does exist in BE but has sense of AE sidewalk.

Identifying these differences is important to improve searching and also to avoid confusion in interpretation of documents. If we know the SLC for a user/collection and have appropriate resources we can enhance searches and analyse the language of retrieved documents to detect possible sources of confusion.
This paper discusses methods for detecting the SLC, it also describes methods for developing resources which can be used to support a user both in searching and browsing.

3 Language Identification and Recognition

Anyone with a limited knowledge of foreign languages quickly identifies lexical differences between the languages which can be either orthographical or morphologi-

cal. In English it is typical to see the, and, and of. In Spanish de, los, el, or in French -ueux or –isme. But –ation is found frequently in both French and English. Many scientists have been interested in these phenomena and have explored the issue from different perspectives. One of the major approaches is based on statistical language modeling. Markov Models are used to estimate the probability of a sentence. By training a different language model for each language, the conditional probability for a new sentence to be generated from the language model can be computed. These probabilities can be compared to make a prediction. Almost all of language modeling research is based on the fact that it is difficult to estimate probabilities for large order Markov models; this is called the data sparseness problem. To overcome this difficulty, typically trigram models are used and probabilities are smoothed to overcome issues with unseen terms.

This processing can be done on the word or character level. While Grefenstette [11] compared common words and common trigrams he did not fully describe his methodology; however from context it would appear to have been a zeroth-order Markov model based on either words or trigrams. Dunning [8] made a more exhaustive comparison using models of order zero-through six on characters (i.e., from single letters to sequences of 7 letters); he also found that trigrams work well. Cavnar and Trenkle [5] tested an n-gram text categorization system on a collection of Usenet newsgroup articles written in different languages and a set of articles from different computer oriented newsgroups. The language classification system achieved 99.8% correct classification. McNamee [21] described a system to identify language using data obtained from the World Wide Wed which achieved an accuracy approaching 100% on a test comprised of ten European languages.

4 Statistical Language Modeling and N-gram Models

The goal of Statistical Language Modeling is to build a model that can estimate the distribution of natural language as accurate as possible. A statistical language model (SLM) is a probability distribution P(s) over strings S that attempts to reflect how frequently a string S occurs as a sentence [6].

The original (and is still the most important) application of SLMs is speech recognition, but SLMs also play a vital role in various other natural language applications as diverse as machine translation, part-of-speech tagging, intelligent input method and Text To Speech system, among the classic models are:

- N-gram models
- N-class models
- Probabilistic context-free grammars (PCFG)

In n-gram language models, each word depends probabilistically on the n-1 preceding words:

$$P(w1...wn) = \Pi P(wi \mid wi\text{-}n+1...wi\text{-}1)$$

The n-gram model assigns to each string the probability of that string in relation to all other strings of the same length (i.e., probabilities for strings of the same length sum to 1), which means that it will overestimate sentence probabilities in comparison to, e.g., a probabilistic context-free grammar. The n-gram model also allows the whole

test corpus to be "parsed" as a single string (including "end-of-sentence" words), which will allow the model to take into account dependencies that span sentence boundaries. Alternatively, sentences can be "parsed" one by one and the probabilities multiplied afterwards.

To evaluate a language model (LM), it is common to use the information theory quantity of entropy to get an estimate of how good a LM might be. Entropy and perplexity which are defined in terms of probability, the corpus probability can be computed as the product of the sentence probabilities:

$$P(\text{Corpus}|\text{Model}) = P(S1,...,Sn|\text{Model}) = \prod P(Si|\text{Model})$$

It is also possible to view the entire corpus as a single string and compute the corpus probability as a simple string probability:

$$P(\text{Corpus}|\text{Model}) = P(w1,...,wn|\text{Model})$$

The entropy in communication is a guide to determine the efficient codes for sending messages. This could be related to language by returning to the idea that the entropy is a measure of our uncertainty. The more we know about something the lower the entropy will be because we are less surprised by the outcome of our trail. In the speech recognition community, people tend to refer to perplexity rather than entropy [20]. The relation between the perplexity and entropy is:

$$\text{perplexity}(x1n,m) = 2H(x1n,m)$$
$$= m(x1n)1/n$$

5 Modern Standard Arabic Regional Variations

Very little is known about Modern Standard Arabic, the assumption is that Modern Standard Arabic is the uniform language used over the Arab speaking countries. Local dialects and colloquial languages have been extensively studied as early as 19th century.

Kirchhoff [14] reports “Arabic dialects are classified into two major groups: Western Arabic, which includes the dialects spoken in Morocco, Algeria, Tunisia, and Libya, and Eastern Arabic, which can be further subdivided into Egyptian, Levantine, and Gulf Arabic. These various dialects differ considerably from each other and from Modern Standard Arabic. Differences affect all levels of language, i.e. pronunciation, phonology, vocabulary, morphology, and syntax. Table 1 lists examples of the differences between Egyptian Colloquial Arabic (ECA) and Modern Standard Arabic…. However, widely differing dialects, such as Moroccan Arabic and the Iraqi dialect, may hinder communication to the extent that speakers adopt Modern Standard Arabic as a lingua franca”.

Table 1. Some differences between Modern Standard Arabic and Egyptian Colloquial Arabic

MSA	ECA	Change	Gloss
/thalatha/	/talata/	/th/→/s/,/t/	Three
/dhahab/	/dahab/	/dh/→ /z/,/d/	Gold
/sayf/	/seif/	/ay/→ /ei/	Summer
yatakallam(u)	yitkallim	inflections	he speaks
Tawila	Tarabeeza	vocabulary	Table

The assumption about MSA could not be verified in the absence of an adequate corpus that would support the assumption or reject it. Minimal work has been done in the area, for either assessing the language itself or comparing its features to other languages. Goweder A and De Roeck A. [10] produced an Arabic corpus using 42591 articles from Al-Hayat newspaper archive of the year 1998. The experiment was mainly to reproduce and confirm results made on small-scale corpus about the sparseness of Arabic compared to English. In 2001 LDC released the Arabic Newswire, a corpus composed of articles from AFP Arabic Newswire. The corpus was tagged using SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from 13 May 1994 to 20 December 2000 with approximately 76 Million tokens and 666,094 unique words. Abdelali et al. [2] discussed issues related to AFP corpus used in TREC conferences; although the size of the corpus is significant. There were many reservations about its representativeness for MSA. Issues about language style, compositions and inconsistency in translation/transliteration were highlighted by examples from the corpus in Table 2. Xu, Fraser, and Weischedel [26] used additional articles from Al-Hayat and An-Nahar newspapers to get terms for automatic query expansion in addition to terms from the AFP corpus. In 2003 LDC also released Arabic Gigaword a bigger and richer corpora compiled from different sources that includes Agence France Presse, Al Hayat News Agency, Al Nahar News Agency and Xinhua News Agency. There is no significant information about this new collection.

Table 2. Translation/Transliteration examples fro AFP

Words	English	Occur.	Words	English	Occur.
لوس انجليس	Los Angeles	21	جوهانس	Johannes	4
لوس انجلوس	Los Angeles	23	يوهانس	Johannes	74
لوس انجيلس	Los Angeles	2	يوهانيس	Johannes	8
لوس انجيليس	Los Angeles	34	جوهانز	Johannes	1
كارولاينا	Carolina	26	جوهانسبورغ	Johannesburg	173
كارولينا	Carolina	14	جوهانسبرغ	Johannesburg	15
ويسكونسين	Wisconsin	8	جوهانسوبورغ	Johannesburg	1
ويسكنسن	Wisconsin	2	جوهانسورغ	Johannesburg	1
ويسكونسن	Wisconsin	16	فايمار	Weimar	3
نيوهامبشير	New Hampshire	15	فيمار	Weimar	10
نيوهامبشر	New Hampshire	9			

Each Arabic speaking country or region also has its own variety of colloquial spoken Arabic. These colloquial varieties of Arabic appear in written form in some poetry, cartoons and comics, plays and personal letters. The colloquial variations are reflected at some level in the Standard Arabic language itself. Many things such as other native languages, accessibility of the area, and the events happening in the region over the time affect the language significantly.

Some of the effects are reflected in lexicon variants, proper names and word usage in different areas. Some of the Arabic literatures have recognized the existence of these variants. But no work has been done to study the phenomena or its implications.

Filali [9] discussed different translations of the Latin word "space". Some linguist in the Middle East translated it as "المكان". In morocco it was translated to "الفضاء" and in Algeria "الحيز".

Al Samrae [3] in his visit to Tunisia noticed that they use different naming and terminology than the way he used to know it back in Iraq. In his book he recorded his experience and titled a chapter of the book as "The Tunisian Arabic".

Maamouri [18] reports that the emergence today, of alarabiyya as a culturally defined set of linguistic resources including the sum of old and new linguistic varieties in use in each given Arab country, shows that there are going to be as many varieties of official 'Arabics' in the region and therefore, as many 'fushas' or fusha standards as there are Arab countries. The vagueness of the language officialization and the unclear definition of the legal status of the term used may not prove to be detrimental after all. It may lead to the individual choice of each Arab country to adapt its language officialization and its status planning policy and measures to the specific requirements of its own diglossia situation.

Arabic readers could notice some words that are used in one region rather than others, or are used in different meanings, Tables 3 and 4 show examples of term usage to refer to the same subject. Tables 5 and 6 refer to names used in different regions for the same object or entity. Tables 7 show examples of words that carry different meanings in different regions.

Table 3. Example of "dormitory"

English Word	El-khabar Algeria	Addustur Jordan	Hayat London
Dormitory	مراقد	عنبر	إقامة

Table 4. Examples of "arrest" and "tend to fall"

English Word	El-khabar Algeria	Al-anbaa Morocco
Arrest	توقيف	حجز
Tend to fall	معرضة للسقوط	آيلة للسقوط

Table 5. Example of naming differences "Ministry of Education"

Egypt, Saudi Arabia	Qatar, Kuwait, Bahrain, Jordan	Mauritania
وزارة المعارف	وزارة التربية والتعليم	وزارة التهذيب

Table 6. Example of naming differences "Ministry of religious/Islamic affairs"

Egypt, Algeria, UAE	Kuwait, Qatar
وزارة الشئون الدينية	وزارة الأوقاف والشئون الإسلامية

Table 7. Example of usage differences for word "ملاحم" and "دوار"

Sense	Word	Sentence	Country
fierce battles; epics	الملاحم	ولكن افتقار الأدب العربي لهذا اللون من الشعر لا يعني عدم احتواءه على المعاني والمفردات كالبطولة والشجاعة والفخر والحماسة التي تتصف بها الملاحم المعروفة	Algeria
butcheries	الملاحم	كما ينفذ قسم المراقبة الصحية حملات تفتيش واسعة النطاق على محلات الاستهلاك الأدمى المتمثلة في بقالات بيع المواد الغذائية والملاحم والمخابز والمطاعم ومحلات الحلاقة	Oman
rotating, turning	دوار	دوار برج الصحوة من المعالم والمواقع الهامة والحيوية حيث يربط معظم ولايات السلطنة بمحافظة مسقط	Oman
Vertigo	الدوار	كذلك الدوار , الناتج عن أمراض عصبية , كالإصابات المركزية في الدماغ و المخيخ	Syria
Bedouin camp, village	الدوار	وعلى بعد خطوات من مسكن أبويه اقترب منه شخصان، اعتقد في البداية أن الأمر يتعلق باثنين من أبناء الدوار،	Morocco

For our purpose, we mined text from newspapers and news services from different Arab speaking countries. We encountered some difficulties in getting common newspapers in parts of the region: either they were not available in electronic format or if they were available, we could not obtain the text in an appropriate format to analyze. Quite a few websites publish their content in PDF files, from which Arabic text cannot be easily reconstituted. In these cases, we had to replace the most common newspapers or news sources in an area with other less common, which were at least available in reasonable quantity. The Table 8 shows the countries from which we collected newspapers.

We did not consider the number of readers, or the popularity of the selected papers selected. Our choices were mainly governed by the considerations of availability already mentioned. This indeed must affect the analysis and conclusions, but we considered that for this preliminary study we could establish some initial results from this small survey, with an eye on improving this analysis with a larger and more representative corpus.

Table 8. Information about news sources and countries of origin

Newspaper	URL	Country	# of files	Size (Kb)
Ahram	www.ahram.org.eg	Eygpt	1567	10348
Alraialaam	www.alraialaam.com	Kuwait	390	15784
Alwatan	www.alwatan.com	Oman	10932	141636
Aps	www.aps.dz	Algeria	7408	68508
Assafir	www.assafir.com	Lebanon	13914	77290
Jazirah	www.al-jazirah.com	Saudi Arabia	3723	28296
Morocco	www.morocco-today.info	Morocco	17196	165266
Petra	www.petra.gov.jo	Jordan	3567	20960

Raya	www.raya.com	Qatar	270	7740
Teshreen	www.teshreen.com	Syria	33703	403228
Uruklink	www.uruklink.net	Iraq		

6 Corpus Collection and Assessment

We used a locally developed spider program to get the data from each site. The spider was initialized with one of the main links in the top hierarchy of the site along with the level of depth to which it should collect document from. The spider will traverse the links and save the pages linked to the main page in a top-down fashion until it reaches the depth specified. The spider runs every morning, (basically evening in the Arab world), which avoids peak traffic time, when people will be reading the newspaper, and also avoid creating problems that could be caused to the server by successive hits from the spider robot. We kept the spider running for a period of more than 3 months in the year of 2002 and collected 107 days of daily issues. Details about the size/number of files per newspaper are shown in the Table 9.

Before indexing the data, we reviewed all the data to check for specific formats that were added for general formatting of the text, such as the link character kasheeda (known also as taweel), which may be added for cosmetic purpose and has no effect on the text, for example, "مـدة" "الأحـداث" "صـاحب السـمو" which are same as " صاحب مدة" "الأحداث" "السمو" respectively.

We also considered removing all the diacritics because Modern Standard Arabic is generally written without diacritics, though in very rare cases people may use them in this type of media primarily for clarification purposes. Contrary to previous experiments [4, 10,12, 15,16], we kept the text close to its original format other than the previous mentioned changes and we did not apply any further processing of what is called Normalization.

We believe that some of these normalizations will hide a lot of features and create more ambiguity knowing that replacing initial إ or أ with bare alif ا means ان could be أنْ , أنَّ , إنْ , إنَّ or آن. The same normalization could hide local variants of the same word as the case for the word "انترنت". Usually in the Middle East they use "إنترنت" in contrast to North Africa where they use "أنترنت" bearing in mind that there are reasons behind this; in the Middle east they use a transliteration of the word "Internet" from English versus in North Africa where the transliteration of the French word for Internet is used [1].

A corpus by itself can do nothing at all; being nothing other than a store of used language [20]. Corpus access software can re-arrange that store so that observations of various kinds can be made[13, 22]. Using available tools we first experimented by applying some statistical and probability tests, such as Zipf's law and the Mandelbrot formula. These tests are useful for describing the frequency distribution of the words in the corpus. Also they are well-known tests for gauging data sparseness and providing evidence of any imbalance of the dataset.

Table 9. Information about the collection

Source	Total Words	Distinct Words	Ratio

Ahram	455,366	16,569	3.639
Alraialaam	1,160,203	97,580	8.411
Alwatan	4,714,199	122,467	2.598
Aps	2,512,426	52,481	2.089
Assafir	3,448,639	121,911	3.535
Jazirah	1,405,083	84,638	6.024
Morocco	3,306,137	19,092	0.577
Petra	989,140	45,896	4.640
Raya	612,409	55,868	9.123
Teshreen	1,467,368	49,067	3.344
Uruklink	2,378,499	32,899	1.383

As a result, from Table 9, which presents a summary of the collection, for number of this datasets, there is no reason to believe that the datasets are imbalanced; Except the Moroccan dataset and the Iraqi one which, we believe to be replaced either by collecting more data or looking for an alternative source from the same area, the rest of the datasets we believe are a real complete representative corpus for the area and that a serious study on these corpuses would bring and reveals very important information about this corpus and the Arabic language in general.

Table 10. Perplexity results of language models and test documents

Test LM	Ah-ram	Al-railaam	Alwa-tan	Aps	Assafir	Jazirah	Mo-rocco	Petra	Raya	Te-shreen	Uruk-link
Ahram	3.47	11.23	12.79	17.01	12.43	11.95	17.24	12.70	12.30	14.17	14.46
Alraialaam	13.05	4.19	12.34	16.85	11.78	11.74	15.89	12.18	11.79	13.50	13.87
Alwatan	12.93	10.61	2.88	16.40	11.95	11.68	16.27	11.91	11.69	13.45	13.73
Aps	13.01	11.67	13.08	1.36	12.38	12.60	16.05	12.66	12.12	14.03	14.31
Assafir	11.48	11.37	10.77	16.86	3.32	11.25	16.37	10.90	11.94	14.20	11.91
Jazirah	13.27	11.66	13.08	17.45	13.05	1.56	17.19	12.71	12.35	14.27	14.58
Morocco	13.13	11.76	13.04	16.89	13.03	12.21	0.40	12.71	12.40	14.20	14.12
Petra	13.28	11.53	13.10	17.27	12.76	12.37	15.45	2.23	12.55	14.08	14.47
Raya	13.43	11.11	12.84	16.74	12.59	11.88	18.28	12.71	1.49	14.06	14.48
Teshreen	13.19	12.08	13.22	18.00	13.17	13.00	16.44	12.69	12.63	1.64	14.66
Uruklink	12.89	11.56	12.84	16.40	12.76	11.84	13.54	12.08	12.42	13.76	0.97

7 Analysis

To assess the collections for variations based on regional distribution we divided each collection to ten parts. We used one tenth for testing and the remaining 9/10 for building the model language. After building the 10 combinations for each collection, we computed the entropy and perplexity value for each document in test data. Table 10 shows the average perplexity for the test samples for each model.

From Table 10 we can see clearly how language structures represented by n-grams differ from place to another. Also we can quantify the language change between neighboring countries and others further away. For example, if we consider the Saudi Jazirah collection, the closest collection to it with the smallest perplexity is Alraialaam, Raya, Petra, Assafir and Alwatan, for the exception of Assafir, all the other collection are from countries that border Saudi Arabia. The other example, we can see clearly how the Algerian APS and the Morrocan Collection varies from the rest of the data, which reflects exactly the distance between these countries and the remaining Arab speaking countries.

To carry this test further we selected a set from AFP documents and computed the perplexity for the set and the Models built from the other collections.

The first observation about the results is that the distribution of the perplexity values are more uniform compared to the former results. This could be explained by one of two things. First could mean that the AFP collection uses news from the different Arabic world (i.e. covers the styles and the structures used in the Arab media) or could mean that the contributors are from different backgrounds and that influences the quality of the collection. To confirm either hypothesis we contacted the Chief Editor for the Arabic Desk of AFP-Middle East HQ in Nicosia – Cyprus –which was the source of the LDC collection-, we asked the Chief Editor about the background of the staff employed by AFP. Surprisingly, the staff was diverse enough to contain almost sample from each country. The Chief Editor reported "They are all Arabs and mainly Lebanese. But we have also Egyptian, Syrian, Jordanian, Tunisian and Algerian journalists. Some of them have a university diploma in journalism or translation and other in different majors".

The test proves and validates our assumption about the language variation of MSA. See Table 11.

Table 11. Perplexity results for AFP test documents

Test LM	AFP	Test LM	AFP
Ahram	11.77	Morocco	12.03
Alraialaam	10.81	Petra	11.71
Alwatan	10.79	Raya	11.82
Aps	11.99	Teshreen	12.2
Assafir	11.17	Uruklink	12.73
Jazirah	12.05		

8 Conclusion and future work

In this paper we discussed methods for detecting SLC, also we described methods for developing resources which can be used to support a user both in searching and browsing. We showed that is possible to classify an Arabic document based on the information computed from its text.

Our aim is to explore further these variations and assess them for building a regional lexicon. Deploying such lexicon in IR systems will improve precisions by reducing the ambiguity of words generated from different usages in different regions.
We also plan to use the same data to produce glossaries dynamically which will alert a user that the SLC they are operating in diverges from the SLC they would be operating in using documents in their own variant of a language.
These methods should also work for English and any other language where a suitable corpus of language variation is available

References

[1] Abdelali, A. (2004) Localization in Modern Standard Arabic. Journal of the American Society for Information Science and technology (JASIST), Volume 55, Number 1, 2004. pp. 23-28.

[2] Abdelali, A. Cowie, J. Soliman S. H. (2004) Arabic Information Retrieval Perspectives. Proceedings of JEP-TALN 2004 Arabic Language Processing, Fez 19-22. April 2004.

[3] Al Samarae I. (1981). The historical linguistic evolution, 2nd edition. Dar Al Andalus. Beirut, Lebanon. (Book in Arabic).

[4] Al-Kharashi, I. A. and Evans, M. W. (1994) Comparing words, stems, and roots as index terms in an Arabic information retrieval system. Journal of the American Society for Information Science (JASIS) 45(8), pp 548-560.

[5] Cavnar, W. B., and Trenkle, M. J., (1994) N-Gram-Based Text Categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, US. pp. 161-175. 1994.

[6] Clarkson P.R. and Rosenfeld. R. (1997) Statistical Language Modeling Using the CMU-Cambridge Toolkit. Proceedings ESCA Eurospeech 1997

[7] Cowie, J.; Ludovik; Y., and Zacharski. R. (1999) Language recognition for mono- and multi-lingual documents. Proceedings of the Vextal Conference. Venice pp. 209-214.

[8] Dunning, T., (1994) Statistical Identification of Language. Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University, March 1994.

[9] Filali, H. (2001) Studies on the poem of Mafdi Zakaria El-Kasida magazine, Issue 9, Algeria. (Document in Arabic).

[10] Goweder, A. and De Roeck, A. (2001) Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001.

[11] Grefenstette, G. (1995) Comparing Two Language Identification Schemes.3rd International Conference on Statistical Analysis of Textual Data. Rome, 1995.

[12] Hmeidi, I., Kanaan, G. and M. Evens (1997) Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. Journal of the American Society for Information Science, 48/10, pp. 867-881.

[13] Hunston, S. Corpora in applied linguistics Cambridge University Press May 2002.
[14] Kirchhoff, K. (2002) Novel Speech Recognition Models for Arabic. Johns-Hopkins University Summer Research Workshop 2002, Final Report.
[15] Larkey, L. S. and Connell, M. (2002) Arabic Information Retrieval at UMass in TREC-10 In Voorhees, E.M. & Harman, D.K. (Eds.) The Tenth Text Retrieval Conference, TREC 2001 NIST Special Publication 500-250, pp. 562-570.
[16] Larkey, L. S., Ballesteros, L., and Connell, M. (2002) Improving Stemming for Arabic Information Retrieval, Proceedings of SIGIR 2002, pp. 275-282
[17] Ludovik, Y., and Zacharski. R. (1999) Multilingual document language recognition. Proceedings of the Machine Translation Summit VII. Singapore. pp. 317-323.
[18] Maamouri, M., (1998) Arabic Diglossia and its Impact on the Quality of Education in the Arab Region HUMAN DEVELOPMENT: MOVING FORWARD WORKSHOP. Mediterranean Development Forum. Marrakech, Morocco. September 3 - 6, 1998
[19] Madar Research - In Focust Article (2004) http://www.madarresearch.com/news/newsdetail.aspx?nwsId=6 Retrieved Sept 22, 2004
[20] Manning, C., Schütze, H. (1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA. May 1999. ISBN 0-262-133600-1
[21] McNamee, P., (2004) Language Identification: A Solved Problem Suitable for Undergraduate Instruction. Proceedings of the 20th Annual Consortium for Computing Sciences in Colleges East (CCSCE-04), pp. 94-101, October 2004.
[22] Meyer, C. F. (2002) English corpus linguistics: an introduction Cambridge University Press July 2002.
[23] Moreh, S. (1988) Studies in Modern Arabic Prose and Poetry, Leiden, E.J. Brill, 1988.
[24] Stetkevych, J., (1970) The Modern Arabic Literary Language Lexical and Stylistic Developments University of Chicago 1970.
[25] Worldwide Internet Population. (2002) www.commerce.net/other/research/stats/wwstats.html Retrieved Sept 14, 2002.
[26] Xu, J. Fraser, A. Weischedel M. R. (2001) TREC 2001 Cross-lingual Retrieval at BBN NIST Text RE-trieval Conference TREC10 Proceedings, Gaithersburg, MD, pp. 68-77.