

Localization in Modern Standard Arabic

Ahmed Abdelali

Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-8001. E-mail: ahmed@crl.nmsu.edu

Modern Standard Arabic (MSA) is the official language used in all Arabic countries. In this paper we describe an investigation of the uniformity of MSA across different countries. Many studies have been carried out locally or regionally on Arabic and its dialects. Here we look on a more global scale by studying language variations between countries. The source material used in this investigation was derived from national newspapers available on the Web, which provided samples of common media usage in each country. This corpus has been used to investigate the lexical characteristics of Modern Standard Arabic as found in 10 different Arabic speaking countries. We describe our collection methods, the types of lexical analysis performed, and the results of our investigations. With respect to newspaper articles, MSA seems to be very uniform across all the countries included in the study, but we have detected various types of differences, with implications for computational processing of MSA.

Introduction

The common history of Arabic countries has played an essential role in the unity of the written language from the Gulf to the Atlantic for several centuries. In the last century differences have appeared dividing the region into parts consisting of small sets of countries. To analyze these differences we examined daily newspapers as reference material. Grounding our analysis in actual data is essential for its objectivity. Our intention is to examine the actual language used to communicate to a wide range of the population, rather than specialized papers or magazines, directed to specialist readers. This work is important in terms of application of information science to Arabic. It has implications for Information Retrieval, Information Extraction and Machine Translation. Since these technologies are primarily form-based, small changes in form may seriously affect the results of these applications.

Received February 25, 2003; revised June 2, 2003; accepted June 2, 2003

© 2003 Wiley Periodicals, Inc.

The article first describes some facts about Arabic and then proceeds to examine different aspects of language differences: spelling, word usage, transliteration, etc.

Evolution of the Arabic Language

Arabic as a language can be divided into three separate categories: classical written Arabic; written Modern Standard Arabic; and spoken Arabic.

Classical written Arabic is principally defined as the Arabic used in the Holy Quran and in the earliest literature from the Arabian peninsula, but it forms the core of much literature to the present day.

Modern Standard Arabic is a modernization of the structures of classical Arabic and includes words for modern phenomena as well as additions from the many dialects used all over the Arabic world.

Spoken Arabic is a mixed form, with many variations, and often with a dominating influence from local languages (prior to the introduction of Arabic) or from colonial languages. Differences between the variants of spoken Arabic can be large enough to make them incomprehensible to one another. Hence, it would be correct to refer to them as separate languages named according to the area where they are spoken, like Moroccan, Cairo Arabic, North Syrian Arabic, etc. (Kjeilen, 2002).

Modern Standard Arabic (MSA) is the formal Arabic that is written and spoken throughout all 21 Arab countries. Also known as *Fus'ha*, it is used in various forms by approximately 290 million (2002 est.) Muslim and Christian Arabs. In the Arab world, MSA is the language of the news media, intellectual life, and literature. MSA is the form of Arabic universally taught in schools of the Arab world. In addition, books, newspapers, journals, reports, and most other printed material are printed in Modern Standard Arabic. Educated speakers also use it in formal discussions or giving oral presentations. News on radio and TV, speeches by presidents and ministers, and discussions by intellectuals are conducted in Modern Standard Arabic (Godlas, 2002).

The nineteenth and the twentieth centuries have made clear impacts on Modern Standard Arabic. All the areas

TABLE 1. Colonial influences.

Country	Control	Type	Dates
Algeria	France	Colonization	1830–1962
Bahrain	Britain	Protectorate	1835–1971
Egypt	Britain	Colonization	1882–1952
Iraq	Britain	Colonization	1917–1932
Jordan	Britain	Protectorate	1922–1946
Kuwait	Britain	Protectorate	1899–1961
Lebanon	Britain	Protectorate	1920–1943
Libya	Italy	Colonization	1929–1949
Morocco	France	Protectorate	1912–1956
Oman	Britain	Protectorate	1820–1976
Qatar	Britain	Protectorate	1916–1971
Saudi A.			
Sudan	Britain	Colonization	1873–1954
Syria	Britain	Colonization	1922–1946
Tunisia	France	Colonization	1881–1956
U.E.A	Britain	Protectorate	1819–1971
Yemen	Britain	Colonization	1839–1965

TABLE 2. List of newspapers.

	Newspaper	Country
1	Al-jazirah	Saudia Arabia
2	Alraialaam	Kuwait
3	Alwatan	Oman
4	Aps	Algeria
5	Assafir	Lebanon
6	Iraq2000	Iraq
7	Morocco-today	Morocco
8	Petra	Jordan
9	Raya	Qatar
10	Thawra	Syria

where the Arabic language is spoken have been affected by the colonial movement. The area was largely divided between the powerful countries of that era, and for around a half-century or more every single country was under some kind of colonial control. The period has affected the language locally in terms of borrowing words and in the way the language grew later on. Table 1 shows the colonial influence in the countries examined in this paper.

TABLE 3. Statistics on word distribution by source.

Newspapers	Size (Kb)	Number of files	Number of words (tokens)	Number of word types (T)	Words only found in this source (OS)	Percent of OS in T
Al-jazirah	35,046	1,713	220,099	35,815	8,155	22.8
Alraialaam	2,634	432	58,287	17,792	2,267	12.7
Alwatan	29,710	2,206	534,308	43,564	9,757	22.4
Aps	1,624	44	29,190	8,122	1,076	13.2
Assafir	22,998	1,762	916,015	100,313	46,266	46.1
Iraq2000	10,444	914	49,619	9,679	1,655	17.1
Morocco-today	9,206	808	90,613	17,707	1,234	7.0
Petra	4,338	632	61,231	14,508	1,879	13.0
Raya	19,148	2,118	265,875	42,197	2,577	6.1
Thawra	33,688	2,471	2,779	1,565	860	54.9
All	168,836	13,100	2,228,016	156,153	75,726	48.4

In addition, the nature of each society helped to shape the language. A rich terminology characterizes every country or region. For example countries with economies based on agriculture could have many more words and terms used to express all the states and the behaviors related to agricultural activity. Fewer terms relating to technical or scientific fields are found in the local media or in addressing local issues. By the middle of the last century many of these countries were engaged in a serious struggle between their identity and deculturalization. For this reason many regional movements were established to restore and modernize Arabic, which was in different stages of derivation or assimilation (Stetkevych, 1970; Wehr, 1976).

Analysis

We have gathered samples of newspapers from different countries for the purpose of comparing the language used in different parts of the Arabic world. Using this data we have tried to produce an analysis of the facts of localization.

While acquiring our resources we encountered some difficulties in getting common newspapers in parts of the region for two reasons: either they were not available in electronic format or if they were available we could not obtain the text in an appropriate format for analysis. So we had to replace these newspapers with other less common or widely read ones, which were available with a considerable amount of usable text. Table 2 shows the countries from which we collected newspapers. The countries were spread throughout the region. We didn't take the number of readers into consideration, or how popular the papers are. Our choice was mainly governed by the problems we have already mentioned above. This could affect the analysis and conclusions, but we felt that for this preliminary study, we could establish some results about the actual status of varieties of Modern Standard Arabic and thus provide a motive to continue to improve and expand this analysis.

The techniques we used to acquire this data were as follows:

1. Data spidering. We used a locally developed spider program that runs on a daily basis to get the data from

TABLE 4. Growth of number of unique words found in one resource.

Newspapers	Words			Words			Words			Words		
	3,000 Word	Number of unique word types	found only in this source	5,000 Word	Number of unique word types	found only in this source	10,000 Word	Number of unique word types	found only in this source	15,000 Word	Number of unique word types	found only in this source
Al-jazirah	3,225	1,511	768	5,045	2,380	1,266	10,025	4,222	1,823	15,067	5,759	2,275
Alraialaam	3,013	1,337	669	5,312	2,641	1,313	10,329	4,863	2,117	16,466	7,527	3,243
Alwatan	3,119	1,555	723	5,017	2,183	934	10,142	4,021	1,564	15,000	5,152	1,803
Aps	3,325	1,791	1,016	5,304	2,676	1,469	11,797	4,692	2,168	14,781	5,557	2,265
Assafir	3,121	1,670	884	5,192	2,554	1,254	10,375	4,588	2,010	15,094	6,205	2,392
Iraq2000	2,962	1,691	1,106	5,602	2,616	1,481	10,498	4,974	2,719	14,787	6,792	3,505
Morocco-today	3,237	2,008	1,222	5,117	2,930	1,705	10,265	5,109	2,727	14,104	6,569	3,315
Petra	3,147	1,836	946	5,097	2,685	1,256	10,038	4,565	1,916	15,027	6,203	2,431
Raya	3,424	841	351	5,378	1,380	514	10,730	2,293	680	14,979	3,317	885
Thawra ^a	2,779	1,565	860	—	—	—	—	—	—	—	—	—

^a The unique data from “Thawra” was small. Although the size of the collection and the number of files was large, the website was not updated frequently and so the collection represents multiple copies of the same data.

each site. Initially, 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 go. The spider traverses the links and saves the pages linked to the main page in a top-down fashion until it reaches the depth specified.

2. Data indexing. The data collected has to be indexed into one database. The purpose of indexing is to recognize every token in the resources, to enable frequency counts and other types of data.
3. Analysis of the results. The previous indexing step produces different sets of data, word lists, and frequency lists that are used in the analysis. (See Tables 3 and 4.)

One of the first results to notice in this initial analysis is the ratio of the number of the unique words found only in one resource to the common words shared by at least two resources (Fig. 1). The number of unique words provided a great incentive for more investigation into the nature of the differences.

Although the differences appear huge compared to the size of the collection, using linguistic metrics such as Zipf’s law (which suggests that up to 50% of any corpus should consist of words that occur only once) could explain some of these behaviors. By looking at the results in Tables 5, 6, 7, and 8 it can be seen that the frequency of the unique words is mostly less than three occurrences. If we set the threshold to a higher number, the total number decreases dramatically.

We proceeded to analyze the nature of these differences using a simple technique of comparing retrieved documents.

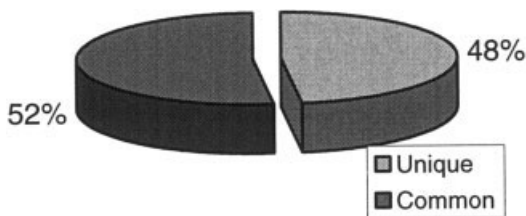


FIG. 1. Ratio of common words to unique.

We queried the system interactively and analyzed the retrieved document in terms of:

- The relevance of the document.
- The structure of the sentences retrieved.
- The new words appearing along with the query words.

We chose a variety of types and subjects for the queries to ensure broad coverage of the data. Queries were made in

TABLE 5. Frequency of unique words in the collection of about 3,000 words.

Newspaper	3,000 Word	Occurrence				
		3 or more	5	7	10	15
Al-jazirah	768	105	43	24	13	10
Alraialaam	669	72	7	2	1	0
Alwatan	723	69	30	14	6	2
Aps	1016	103	35	14	6	1
Assafir	884	48	16	7	4	3
Iraq	1106	54	6	2	1	0
Morocco	1222	86	44	11	6	2
Petra	946	55	19	10	4	3
Raya	351	71	34	11	8	5
Thawra	860	58	34	30	1	1

TABLE 6. Frequency of unique words in the collection of about 10,000 words.

Newspaper	1,0000 Word	Occurrence				
		3 or more	5	7	10	15
Al-jazirah	1,823	150	47	30	16	9
Alraialaam	2,117	124	37	18	6	3
Alwatan	1,564	147	56	32	14	9
Aps	2,168	256	79	41	13	6
Assafir	2,010	125	32	10	3	3
Iraq	2,719	125	20	9	7	3
Morocco	2,727	205	95	52	21	8
Petra	1,916	111	35	19	5	2
Raya	680	98	38	23	16	11
Thawra	0	0	0	0	0	0

TABLE 7. Frequency of unique words in the collection of about 5,000 words.

Newspaper	5,000 Word	Occurrence 3 or more				
			5	7	10	15
Al-jazirah	1,266	120	43	23	10	9
Alraialaam	1,313	89	23	6	2	0
Alwatan	934	108	39	20	14	7
Aps	1,469	136	47	23	10	2
Assafir	1,254	77	22	10	4	3
Iraq	1,481	92	18	7	4	3
Morocco	1,705	109	64	30	13	6
Petra	1,256	69	22	13	4	3
Raya	514	78	41	25	13	9
Thawra	0	0	0	0	0	0

fields of politics, religion, economics, sports, etc. The results explaining some of the reasons for the high rate of the unique words in the different newspapers can be grouped under the following categories:

1. Differences of spelling: As Arabic is a highly inflected language, the spelling of words was one of the consistent differences noticeable in the different newspapers. This consistency is probably due to historical facts about the development of the Arabic grammar itself and also to the different narrations of the Quran, one of the references for the Arabic language (Larkey, 2002). Table 9 contains examples of different spellings of the same word in different newspapers. More examples are reported in Appendix A.
2. Differences in transliteration: The influence of foreign languages on Modern Standard Arabic could be seen in the way some words are spelled in Arabic based on the spelling in the language they were derived from, mainly English and French. The pronunciation of the English letter "i" differs from the French "i" and similarly for the a, e, ch, etc. Table 10 contains examples of these differences. More examples are provided in Appendix A.
3. Differences in usage: The usage of words marks another distinction between regions in the Arabic world. Even if all the words are native Arabic words, their usage remains rare or nonexistent in some parts of the Arabic

TABLE 8. Frequency of unique words in the collection of about 15,000 words.

Newspaper	15,000 Word	Occurrence 3 or more				
			5	7	10	15
Al-jazirah	2,275	187	53	31	12	7
Alraialaam	3,243	156	44	25	8	3
Alwatan	1,803	190	75	43	22	8
Aps	2,265	262	90	44	18	10
Assafir	2,392	159	47	18	3	3
Iraq	3,505	195	34	10	5	1
Morocco	3,315	235	93	57	24	11
Petra	2,431	136	48	21	9	3
Raya	885	117	43	21	15	10
Thawra	0	0	0	0	0	0

TABLE 9. Spelling differences between APS (Algeria) and Petra(Jordan).

Word	Translation	APS	Petra
إدارة	Administration	4	0
أدارة	Administration	5	33
أيام	Days	14	0
أيام	Days	0	64
الأربعاء	Tuesday	19	0
الأربعاء	Tuesday	1	6
أحمد	Ahmed	13	3
أحمد	Ahmed	5	33

world while they are very often used in others (Filali, 2001). Tables 11 and 12 contain examples of words that are used in some regions, while in other regions other words are used to convey the same meaning. (Appendix A contains more examples).

4. Names: Due to the tribal social structure being strong in some areas, some tribes remain in place for centuries and the standard names used by the tribe never appeared outside the area where they were living. Examples are shown in Tables 13 and 14.
5. Imported words and derivations: A final class of differences in Modern Standard Arabic occurs in foreign loan words, which vary according to the source language and the adoption process. The Arabic language uses different ways for acquiring new terminology and concepts; either deriving or adopting or transliterating words from other languages as a means to develop the language and keep an up-to-date terminology. This leads to differences of words for the same concepts or objects in different areas as shown in Table 15. One other element that contributed significantly to the same issue are misspellings and ty-

TABLE 10. Spelling differences originated from French and English.

Word	Translation	Source	Occurrence
الإنترنت	Internet	English	125
الانترنت	Internet	French	49
الأوبك	OPEC	English	19
الأوبيب	OPEP	French	18
الإيدز	AIDS	English	20
السيدا	SIDA	French	3

TABLE 11. Example of "Dormitory."

English Word	El-khabar Algeria	Addustur Jordan
Dormitory	مرآقد	عنبر

TABLE 12. Example of "Tend," "Arrest."

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

TALE 13. First names common in Saudi and not used in Jordan.

Name	Petra	Al-jazirah	
فهد	Fahad	2	331
فوزية	Fawziah	0	109
موضي	Moudha	0	1
نورة	Nourah	0	189
سطام	Sattam	0	28

TABLE 14. Family names common in Saudi and not used in Jordan.

Name	Petra	Al-jazirah
آل Al- (of the tribe, . . .)	1	270
الغامدي Alghamedi	0	203
العتيبي Alotaibi	0	207
الحربي Alharbi	0	173
الشمري Ashamari	0	126
القحطاني Alkahtani	0	125
الزهراني Alzahrani	0	125

pographical errors. They do contribute to the overall ratio of differences, but these problems are not specific to Arabic and do not reflect characteristics of the language itself.

Conclusion

This analysis done on the selected newspapers reports some preliminary results and concludes from this set of resources that this type of media are remarkably similar, but there are still some important differences, which once cataloged, can be useful to NLP applications and publishers of various Arabic language resources. More intensive investigation is needed to discover the full scope of localization in MSA. The next step is to investigate different type of resources such as technical texts and educational publications. For reasons of availability, we were unable to consider some countries such as Egypt, the Arabic speaking country with the highest population, even though the impact of Egyptian Arabic is very clear in recent century on politics, religion, and literature. It must be considered in further research.

Acknowledgments

I would like to thank my advisor Dr. Hamdy Soliman (New Mexico Tech), Dr. Jim Cowie, and Dr. Stephen Helm-

TABLE 15. Words adopted or derived.

English word	Imported or derived word	Other word
Bridge	كبري	جسر
Cadre	كوادر	إطارات
Cable	كابيل	حبل
Globalization	عولمة	العالمي النظام
Humanism	أنسنة	النزعة الإنسانية
Privatization	الخصخصة	الخصوصية
Agenda	مذكرة	

reich (Computing Research Laboratory) for their fruitful discussions and comments.

References

- Filali, H. (2001). Studies on the poem of Mafdi Zakaria. El-Kasida magazine, 9, Algeria. Retrieved October 2, 2002, from http://www.aljahidhiya.asso.dz/ALKASSIDA/kassia-9/alkassida_9_2.htm
- القصيدة. 2001. 9 فيلالي ح, مستويات الموت في شعر مفدي زكريا Retrieved October 2, 2002, from http://www.aljahidhiya.asso.dz/ALKASSIDA/kassia-9/alkassida_9_2.htm
- Godlas, A. Modern Standard Arabic (MSA). (2002). Retrieved December 1, 2002, from: <http://www.arches.uga.edu/~godlas/MSA.html>
- Kjeilen T., Abubakr S., & Negahban D. J. (2002). Encyclopedia of the Orient. Retrieved December 1, 2002, from <http://i-cias.com/e.o>
- Larkey, L.S., Ballesteros, L., & Connell, M. (2002). Improving stemming for Arabic information retrieval. Proceedings of SIGIR 2002 (pp. 275–282).
- Stetkevych, J. (1970). The modern Arabic literary language; lexical and stylistic developments. Chicago: University of Chicago Press.
- Wehr, H. (1976). A dictionary of Modern Written Arabic (Arabic-English). Edited by J. Milton Cowan. New York: Ithaca Press.

Appendix

1. Spelling Differences

الايام Days	
APS	0
Al-jazirah	50
Raya	10
Petra	64
Iraq2000	17
Althawra	11
Morocco-Today	1
Alwatan	0
Total	155
الاربعاء Tuesday	
APS	0
Al-jazirah	138
Raya	33
Petra	6
Iraq2000	2
Althawra	0
Morocco-Today	0
Alwatan	56
Total	235
ادارة Administration	
APS	0
Al-jazirah	61
Raya	86
Petra	47
Iraq2000	1
Althawra	0
Morocco-Today	2
Alwatan	78
Total	273

أحمد / Ahmed	
APS	13
Al-jazirah	210
Raya	72
Petra	5
Iraq2000	2
Althawra	0
Morocco-Today	1
Alwatan	110
Total	413

أيام / Days	
APS	14
Al-jazirah	36
Raya	86
Petra	0
Iraq2000	5
Althawra	4
Morocco-Today	1
Alwatan	45
Total	191

الأربعاء / Tuesday	
APS	19
Al-jazirah	15
Raya	58
Petra	1
Iraq2000	0
Althawra	0
Morocco-Today	0
Alwatan	13
Total	106

إدارة / Administration	
APS	4
Al-jazirah	64
Raya	64
Petra	1
Iraq2000	1
Althawra	1
Morocco-Today	0
Alwatan	18
Total	155

أحمد / Ahmed	
APS	3
Al-jazirah	269
Raya	75
Petra	33
Iraq2000	13
Althawra	0
Morocco-Today	0
Alwatan	149
Total	542

Word	Petra	Al-jazirah
الإسلام Islam	0	14
الإسلام Islam	2	9
الإسرائيلي Israeli	2	7
الإسرائيلي Israeli	32	17
الأضرار damages	0	6
الأضرار damages	1	0
الأمريكية American	2	40
الإنترنت American	32	15
الإنترنت Internet	24	25
الإنترنت Internet	2	123
شيئ something	1	63

	APS	ALWATAN
الأستاذ Teacher	8	0
الأستاذ Teacher	0	3
الإفريقي African	12	0
الإفريقي African	0	5
الأمية Illiteracy	29	11
الأمية Illiteracy	9	0
بالإضافة In addition to	9	1
بالإضافة In addition to	0	127

2. Word Usage Differences

Word	APS	ALWATAN
التدريب Training	10	0
التدريب Training	0	62
التدابير Arrangement	10	1
الإجراءات Arrangement	0	38
أشغال Sessions	14	0
أعمال Sessions	13	30

Word	Petra	Al-jazirah
معالي highness	1	61
السيد Mister	276	22