تحليل إحصائي لدعم التعرف الآلي

جدول ١٥ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف الثامن – الحرف التاسع

الحرف الثامن	ا	ة	م	ن	منفصل	مجموع
ا					75	75
ة					1	1
ر					8	8
ف					6	6
ق					2	2
م	4				4	8
ن					56	56
ه	2		1		6	9
ي		2		19	4	25

جدول ١٦ : تكرار ظهور الحرف التاسع (أطول مقطع ٩ حروف)

الحرف التاسع	مجموع
ة	2
ا	6
م	1
ن	19

المراجع

Badr Al-Badr, Sabri A. Mahmoud. "Survey and Bibliography of Arabic Optical Text Recognition." J. of Signal Processing, Vol. 41, No.1, pp.49-77 (Jan. 1995).

Asim Nabawi, Sabri Mahmoud. "Arabic Optical Text Recognition: A Classified Bibliography." Engineering Research Bulletin, Minufiyah University, Egypt, Vol. 23, No. 1, Jan. 2000, pp. 79-131.

M. Khorsheed. "Off-line Arabic Character Recognition – A Review." Pattern Analysiss & Applications, 5:31-45, 2002.

Liana M. Lorigo, Venu Govindaraju. "Offline Arabic Handwriting Recognition: A Survey". IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 712-724, May 2006.

M.Z. Khedher and G. Abandah "Arabic Character Recognition using Approximate Stroke Sequence." Third International Conference on Language Resources and Evaluation (LREC2002), 2002.

Yousof S. ElArian, A Lexicon of Connected Components for Arabic Optical Text Recognition, Master thesis, Jordan University of Science and Technology, Amman, Jordan, August, 2006.

Latifa Al-Sulaiti, Designing and developing a corpus of contemporary Arabic, Master thesis, School of Computing, University of Leeds, Leeds, UK, March, 2004.

M.S. Khorsheed. "Recognising handwritten Arabic manuscripts using a single hidden Markov model", Pattern Recognition Letters, 24(2003) pp. 2235-2242.

Issam Bazzi, Richard Schwartz, John Makhoul. "An Omnifont Open-Vocabulary OCR System for English and Arabic", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 6, June 1999.

R. Schwartz, C. LaPre, J. Makhoul, C. Raphael, and Y Zhao. "Language-Independent OCR Using a Continuous Speech Recognition System." Proc. Int. Con$ on Pattern Recognition, Vienna, Austria, pp. 99-103, August 1996.

الحرف السادس	ئ	ا	ب	ة	ت	ث	ح	د	ذ	ر	س	ش	ض	ع	ف	ق	ك	ل	م	ن	ه	و	ي	منفصل	مجموع
أ																								2	2
ئ																								4	4
ا																								2251	2251
ب		1								2							2			1	4	5		29	44
ة																								625	625
ت		3					1							2		3			9		18		14	153	203
ث		1								7												2		27	37
ج		9												2			2	2	2	1			1	1	18
ح		1																			2	4		43	50
خ								1	1					1										1	4
د																								76	76
ذ																								18	18
ر																								420	420
ز																								11	11
س		5	1							1								11			3			27	48
ش										3															3
ص		8		1						1														1	11
ض																					1	1		7	9
ط		3		1																	2			1	7
ظ				3						1														14	18
ع		4		1					6						25					3	10	18	1	113	181
غ															2									4	6
ف	1	2			1					5							1				1	5	12	58	86
ق	1	1		10														1			10	9	7	49	88
ك		1													2				24	1	1			72	101
ل		6	1	11										4	2	1	5	2	2	6	26	18	3	168	253
م		105									1			1	1					2	9	5	22	653	799
ن		69	5	12		4						2				3					20	2	40	481	638
ه		247						4											109	9		1	4	582	956
و																								588	588
ى																								39	39
ي		9	1	1			1	2					1	3		2	2	3	2	378	22	1	10	843	1279

جدول ١٣ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف السادس – الحرف السابع

الحرف السابع	ا	ة	ر	ف	ق	م	ن	ه	ي	منفصل	مجموع
ئ										2	2
ا										475	475
ب										8	8
ة										40	40
ت										1	1
ث			4								4
ح										2	2
د										7	7
ذ										7	7
ر										20	20
س								1			1
ش					2						2
ض										2	2
ع										9	9
ف			2	6					19	6	33
ق										1	1
ك						4				10	14
ل		1						3		17	21
م	48		2							98	148
ن								1	6	394	401
ه	27					4				96	127
و										73	73
ي							56	4		54	114

جدول ١٤ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف السابع – الحرف الثامن

الحرف الخامس	أ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع
آ																																	3	3
أ																																	38	38
ؤ																																	3	3
ئ			3		11	3					2	6													1		13		1		2		28	71
ا																																	10664	10664
ب			11		44	11			6		2		37	2	23	3	1	1	2		12			77	6		11		52		87		777	1199
ة																																	5350	5350
ت	1		33	4		1		2	6	2			6	5	11					1	23	2	1	40	7	74	33	395	37		104		1415	2205
ث			5		1								12												3		2	1	1		3		49	77
ج			3								16		1								1				16		1		4		8		8	58
ح			10	2	27	1		2	6		1		10		1								1	7	8	1	4	14	14		15		170	288
خ				1									2								2		3										23	31
د																																	1401	1401
ذ																																	102	102
ر																																	2392	2392
ز																																	52	52
س			5	8		1			1	1			11											1		16	6	27	8		2		167	273
ش			9										13											1				4	3	3		3	3	36
ص			1		11	2			1		9		17										2	5				1			11		41	101
ض					7	3															20						2		1				72	107
ط	1		18	2	3	1		2					7								16			1	1	1	3	5	17	1	9	3	68	155
ظ			3			12							16											3		2	4	6					95	143
ع			39		29	25	17	2			5	3											3	2	36	7	34	46	43	3	55		335	684
غ																		1					4		1		2	1	4		1		21	36
ف	4		36		127	14	5						67		1			4			7		14	2	5	1	7	24	41	1	8		216	584
ق			19	3	31	3					2	1	6						3	11	1				1	7	11	25	21		37		157	350
ك			3		3	4	9						9										8		1	178	4	30	5		5		449	710
ل			23	6	45	40	11				8				1						1	4	34	40	2	36	45	94	95	3	100		902	1499
م			523		110	18					4		74		8		9	2			8			3	1	1	24	50	214		424		3262	4783
ن			551		85	5	3	1	3	1	1								1		3		1	1	16		4	80	20	17	384		1727	2906
ه			923	1	1						12		3													4	411	62		10	5		2857	4291
و																																	2261	2261
ى																																	1659	1659
ي			33	17	90	59	3		33		14	8	129	4	3		1	1	1	6	87		15	24	9	16	368	119	1		19		2259	3387

جدول ١٢ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف الخامس – الحرف السادس

جدول ١١ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف الرابع – الحرف الخامس

الحرف الرابع	آ	أ	ؤ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع
آ																																			6	6
أ																																			430	430
ؤ																																			55	55
إ																																			1	1
ئ					48		68	45						37	4	2	2									4		24	3	10	19	24	3		7	25	298
ا																																			29471	29471
ب		3		2	415	58	1583	140	7		22		21	24	448	3	32	4	6	16	19		73	3		13	44	157		88	248	171	2	701	2331	6578
ة																																			11377	11377
ت		7		6	106	58	30	26	4	8	109	5	24		82	15	19	13	12	21	16	6	62	9	17	52	202	101	409	138	1243	134	12	165	8750	11861
ث					52	3	13	34							32												6	15		40	31	26		77	1184	1540
ج				2	41	29	4	10							15	9							52					46		15	3	16	1	18	51	968
ح					65	25	44	46	2				23		41	1	9	7	3		3				27	18	7	44	32	28	68	74	10	55	1097	1729
خ					2	7	12	25					4	2	17				4		10		1		14		2	22	5	4	8	4		4	101	248
د																																			16728	16728
ذ																																			426	426
ر																																			15879	15879
ز																																			499	499
س		5			97	28	56	131		13	11	1	5		60		4				4		3		18	21	47	116	74	53	148	103	12	72	1243	2325
ش	1	1			17	10	81	11					8		33						13		1		16	17	5	3	2	1	11	4	7	47	166	466
ص					29	35	138	29			2		28		69						1				12	6	82	16	42	16	17	32	1	39	110	692
ض		3			40	20	32	28			4				9						2		6			14	68	56	64	13	27	41	3	20	288	697
ط				2	284	11	43	35		1	1	1			41								93		13	6	38		27	11	52	6	147	292	1108
ظ					12		10	7							93												103	1	12	4	3	1	2	696	944
ع					695	15	476	273	21	12			23	6	15	2		3			6				14	9	73	61	269	110	215	2	143	1439	3947
غ					4			16							9		3	3		2					48		2	3	33	41	5	109	1	109	208	515
ف		1	1		114		244	171		1	3	3		2	133	3	13		1	3	3	3	69	3	25	41	10	17	39	60	119	7	256	756	2098
ق					190	86	86	176			1		11	12	57	1	4		22	1	22	101	12		3		1	85	16	98	90	201	57	77	942	2352
ك		5			11	33	163	74	15		2		114		73		1					2			49		16	26	1009	62	47	27	2	88	2009	3832
ل	2	3	2	26	390	205	556	350	8	3	21	6	7				49		8		15	1	12	22	120	84	58	33	753	166	487	824	62	133	288	5283	9909
م				13	3789		812	167	2		2	6	85		169		48		1	37	3	2	147			9	32	152	48	134	252	339	8	312	8012	14581
ن		1		3	1954	39	732	72	11	9	6	2	8			3	4	1	5	1	8	13	36	1	20	3	62		88	26	315	98	1337	686	4050	9593
ه					1986	13	14	2		3	10	11	81		31	10				1							7	25	1729	158	19	40	11	26	28364	32520
و																																			7272	7272
ى																																			6550	6550
ي		9		16	323	580	153	337	15	3	104	10	318	19	961	3	85	5	39	26	30	15	114	1	191	68	134	291	410	1649	483	10	45	48	13847	20342

الحرف الثالث	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع
آ																																				22	22
أ																																				2057	2057
ؤ																																				995	995
إ																																				43	43
ئ	3					876		38	138	5	2			16		66	1	1	4										19	1	17	30	23		8	28	1289
ا																																				113405	113405
ب		20		1	11	1338	28	4412	452	80		347	17	594	56	1014	26	324	99	13	99	53	313		78		56	86	531		261	677	299	26	9654	2295	23260
ة																																				8824	8824
ت		110	4		10	856	323	44	73	8	81	443	63	85	6	332	42	140	52	67	13	152	64	185	130	120	210	292	710	742	422	1340	223	10	1009	6772	15133
ث						83	8	2	95	4				134										3		1	7	36	762	118	1183	64	38	5	75	936	3554
ج		21	1		1	543	118	89	45	1	3	3		400	5	710	47	23	9					134		4		2	172	119	61	11	199	15	163	421	3320
ح						477	155	532	173	4	21			239	3	613	5	41	30	41	9	5	1			81	76	165	120	169	88	260	109	267	645	4373
خ						107	28	10	104					5	76	190	33	11	3	12	17	23		17		21			291	17	18	14	22		84	83	1186
د																																				30893	30893
ذ																																				3202	3202
ر																																				44166	44166
ز																																				2960	2960
س		82	2		4	941	195	94	438		642	126	27	47		368		22				120		33		53	271	915	349	738	493	290	397	1115	2068	9913
ش		4			5	295	31	24	43		3	5		31		775	1		1			27		13	3	73	10	26	10	22	26	54	110	88	289	1091	3060
ص						254	138	263	60			39		111		642				24		9		53	8	117	25	257	80	69	25	675	14	164	578	3605
ض		2			5	202	167	97	66			32	2	26		451					2	3		39	9	2	57	270	64	34	369	48	86	216	755	3008
ط		10	3		6	735	236	25	66		42	90	2			389		1	62			6		287	2	57	8	2	199	47	88	91	109	27	245	247	3082
ظ						40		4	38							494								8				90	50	22	40	9	2	53	103	958
ع						639	609	573	3788	266	57			438	49	425	13	33	6	34	101	47	4			80	105	67	1276	670	455	392	783	32	515	3388	14845
غ						94	8	17	131	2				15		220	6	10	5		31	8				127		24	52	14	149	31	56	5	386	213	1604
ف		25	2		4	553		63	323	15	7	20	53	4	9	1708	4	175	3	34	107	32	707	167	1	46	225	29	124	2	82	248	192	63	427	1331	6785
ق		11	2			546	362	136	328	1		16		307	5	315	4	26	7	100	59	107	9	22		126	9	67	261	249	63	1268	287	161	545	1192	6591
ك		11			28	302	138	166	351	69		112		1		212	3	38	14	8				5		77		28	116	1310	119	87	177	69	371	2346	6168
ل	1	52	1			3621	514	664	2183	86	21	133	5	190	4		28	311	23	40		65	2	153	259	383	952	883	77	2524	494	1507	1147	724	1480	8153	26657
م		20	23		21	5249		1672	779	14	4	52	5	6721		2117	23	807	23	22	18	16	1067		10		18	97	746	28	578	359	701	101	789	32716	54796
ن		5	1		39	2944	661	1407	383	49	18	53	10	145	143		170	57	34	42	3	48	135	393		135	60	378		381	167	889	359	200	334	2194	10213	21691
ه						5248	55	2	60	5	19			361		706	41	5	3		57					18	24	5	95	324	26	359	153	237	72280	85099
و																																				15456	15456
ى																																				35409	35409
ي	2	57	16		164	3528	2804	1043	1744	931	44	258	64	7008	44	3998	52	300	111	255	181	387	22	1055	15	577	468	1315	2816	2609	4135	25386	140	4129	56	13599	79313

جدول ١٠ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف الثالث – الحرف الرابع

جدول ٩ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف الثاني – الحرف الثالث

الحرف الثاني	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع
آ																																				1459	1459
أ																																				20129	20129
ؤ																																				1257	1257
إ																																				6277	6277
ئ						19		31	402					17	219	76	1	40			8								168		161	2	10		14	98	1374
ا																																				139833	139833
ب		111	1	4	96	4502	106	291	806	96	36	678	90	13281	45	9453	70	362	139	126	112	187	3	1216	133		362	439	2226	7	1459	431	618	441	15	5805	4486	46774
ة																																				10867	10867
ت		399	6		21	1969	1082	78	141	47	255	723	261	267	41	937	187	359	167	283	47	224	71	518	273	385	782	820	1217	1459	907	1909	759	4100	2830	9949	33473
ث	1					488	63	141		2			3	10		512								19	4		202	8	1647	1094	128	428	226	49	711	2173	7909
ج		19	3			1770	383	221	117	7	21	141	3	817	38	1002	217	45	17	2				915		21		22	558	729	1245	355	326	14	376	1014	10529
ح					131	4003	671	119	391	47	986	8		1860	98	1233	86	876	115	159	142	63	5			128	427	497	981	7451	643	98	972	79	4810	1432	28380
خ						352	289		169	3				423	322	1123	102	38	85	77	66	701				155			682	373	119		151		598	56	5884
د																																				65359	65359
ذ																																				11690	11690
ر																																				53029	53029
ز																																				9087	9087
س	5	816	74		30	2464	704	62	1361		512	197	106	513		860		9				149		919	3	301	210	544	2242	1610	862	323	503	56	1001	3726	20162
ش		57	14			2634	122	2356	217		171	21	19	142		2005	2		6			46		464	47	219	214	115	1	582	27	389	127	62	1282	386	11727
ص						1794	787	86	59			136	49	750		858				24		9		169	42	560	17	8	2994	275	264	50	434	52	740	166	10323
ض		3				734	266	86	54		26	357	9	9		714						48		197	16	9			562	68	8	19	58	222	513	357	4336
ط		76	19		6	1508	224	23	33		67	43	1	1		630		38	18			13		577	2	228	11	14	900	62	220	173	519	178	923	563	7070
ظ					1	79	9	64								392								25		22		4	160	149	84	322	12	2	120	47	1492
ع						7869	3167	580	1461	680	245			3387	231	1295	653	134	529	452	1043	375	124			1189	727	208	2915	2882	2109	1690	633	35	5641	5890	46144
غ						218	45	12	100	7				183	5	281	108	234	65	10	195	52				187			329	40	258	1	66	6	494	36	2932
ف		36	20		38	1026		405	682	15	252	48	58	53	19	1323	87	1190	18	630	419	212	156	585		27	324	20	274	6	119	148	306	27	2653	1281	12457
ق		1				12010	1081	56	771	7		9		1637	57	1805	14	261	39	316	290	426	16	316		75	6	18	2055	381	124	5018	138	2254	1153	30605
ك	1	89				1152	344	644	781	157		91		29	452	4176	3	202	55	4			1	228		415		14	952	1867	1346	76	865	26	641	9015	23626
ل	13	185	19	38	9	7641	887	671	2637	82	67	881	33	283	106	75	54	344	36	123	6	78	40	402	493	1329	1419	930	136	912	1275	29486	34037	11626	184785
م	1	28	735	1	7	10376	190	552	632	1853	243	374	162	1753	10	8988	450	2786	1506	264	101	268	26	5808	555	139	445	532	1698	117	4879	650	764	113	3564	17512	68082
ن	1	30	2			41704	9845	1346	2444	16	238	412	213	2363	175	26	746	646	91	906	331	107	496	873	6	697	392	134	15	201	9244	659	306	8874	113822	198498
ه			15			6930	110	14	64		118			1311	1328	917	139	5	6		14	3				30		49	5	3722	550	401	648	439	833	18224	36817
و																																				56242	56242
ى																																				18616	18616
ي		207	87		949	2163	2885	986	1811	535	83	254	179	1804	56	5485	41	2344	168	229	234	121	20	1614	30	699	1048	681	2973	4387	6161	6834	684	4	599	41500	87855

الحرف الأول	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع
ئ						8	299	373	141	4	6	39		264	89	180	95	10	2347	1	144	113		45	9	165	32	323	488	642	104	412	6	2	378	184	6903
ا																																				213387	213387
ب	58	1009	6	298		13703	115	258	991	1518	168	170	693	275	980	404	5364	80	681	1683	487	91	467	44	5005	355	140	789	4933	1709	2050	6336	56482	9491	175	23447	15540	155995
ة																																				17596	17596
ت		517	176			961	968	14	360	30	474	1174	601	510	150	1623	349	1019	324	681	171	333	49	6038	248	542	1449	1159	1000	1532	919	2091	1033	655	2194	6134	35488
ث						1169	124	515	294		1		2	68		298								25	4	2	132	156	1251	6704	37838	522	559	10	317	2260	52252
ج		9	4			4052	1209	257	738	17	3	143		1101	86	2506	228	81	47					2750		53		60	5186	892	1512	1541	813	24	311	2966	26785
ح						1968	1978	157	4300	50	1109			43452	283	2402	187	1229	137	419	180	58	37			832	548	272	986	5984	306	127	573	92	2207	3256	73129
خ						1927	8282	14	474	27				220	1367	3176	85	87	265	242	81	387		1		248		1	3119	505	53	8	394	39	1808	243	23053
د																																				18951	18951
ذ																																				13529	13529
ر																																				56148	56148
ز																																				8625	8625
س	2	1047	20			2475	1256	59	2519		426	2029	67	247		879		5				188		5270		2990	415	276	26121	7819	1736	1423	15827	1668	1063	5836	81833
ش		193	4			997	372	37	475	1	294	60	15	592		1164	5	4				68		3019	33	255	310	232	6	333	49	2465	154	8	4089	330	15804
ص						2155	828	54	10			1014	63	877		193						65		133	114	523	2	11	23704	620	357	162	626	41	564	481	32597
ض		360	17			329	44	94	153		8	216	21	2		399						95		801	9	6		21	91	75	50	157	420	4616	790	1563	10357
ط		39	2			1013	149	35	16			12				226		18	5					512	10	95	1	17	1002	268	34	179	419	10	426	413	4917
ظ						60	10	3								18								5		44			190	6	200	369	2		15	40	965
ع						5165	17560	1016	1038	938	323				338	2587	518	164	697	224	56	1510	317	1		307	982	406	37005	9684	48249	686	1829	135	3543	1961	138472
غ						280	143	3	174	22					4	270	373	372	51	24	198	67				359			442	100	488	26	42		1867	217	5753
ف	65	6041	6	3381		3762	677	336	2041	68	1385	672	722	923	906	2573	206	1295	315	907	1111	437	79	3834	267	341	13300	1353	5608	1495	1344	1099	1294	181	18251	2334	78655
ق		15				41631	2303	814	3405	2		47		3789	43	2078	54	214	18	287	444	534	24	427		209	10	55	3023	305	187	362	2738	31	1590	4964	69603
ك	3	412		12	6	7297	958	120	1224	1033	4	52	9	57	1045	3793	22	279	61	10	21	8	2	1333	5	554	47	24	3876	2404	2196	348	741	14	1457	3075	32502
ل	1330	9294	40	2586	216	16127	3814	435	7385	1139	3433	7750	2555	3642	2806	6898	3207	4280	2691	3576	482	1261	388	5693	1062	2082	5632	9618	63348	17858	16602	8796	6031	10516	142	9661	68159	310393
م	1	82	230			14233	157	2453	1375	1573	679	7037	425	345	39	5690	196	3486	323	282	602	288	40	5954	238	112	488	1166	1445	718	23598	1475	3176	142	1984	14835	95101
ن		91	28			11267	1235	650	2465	74	370	1174	220	184	156	185	1349	4510	228	1857	119	627	392	1558	58	1713	503	1351	26	1696	274	6614	1092	199	7271	41675	91279
ه			251			3559	2136	52	92		110			1063	3264	4767	61		1566	3	5	103				35	300	3779	623	197	3323	58	4527	10779	44267
و																																				111749	111749
ى																																				2869	2869
ي		1020	473			5695	2014	2385	3276	2763	1734	6268	909	5579	710	5760	2072	2428	969	1323	631	461	120	3740	520	885	5930	1892	1380	2160	5917	1422	5659		95	9801	86024

جدول ٨ : تكرار ظهور الحرف والذي يليه في المقطع: الحرف الأول – الحرف الثاني

الحرف الأخير	متصل في آخر المقطع	منفصل	مجموع
ء		11900	11900
آ	1490	3552	5042
أ	22656	103951	126607
ؤ	2310	792	3102
إ	6321	29677	35998
ئ	185	184	369
ا	296180	213387	509567
ب	9926	15540	25466
ة	37086	17596	54682
ت	27040	6134	33174
ث	4369	2260	6629
ج	1495	2966	4461
ح	3389	3256	6645
خ	264	243	507
د	114464	18951	133415
ذ	15445	13529	28974
ر	115914	56148	172062
ز	12609	8625	21234
س	7231	5836	13067
ش	1646	330	1976
ص	896	481	1377
ض	1481	1563	3044
ط	1171	413	1584
ظ	955	40	995
ع	11174	1961	13135
غ	482	217	699
ف	3654	2334	5988
ق	3496	4964	8460
ك	13901	3075	16976
ل	26149	68159	94308
م	62258	14835	77093
ن	130762	41675	172437
ه	122409	10779	133188
و	81892	111749	193641
ى	62273	2869	65142
ي	72106	9801	81907

جدول ٧ : آخر حرف: حرف متصل في آخر المقطع أو منفصل

جدول ٥ : تكرار الحروف في مقاطعها

الحرف	تكراره في المقطع الذي بطول									مجموع	نسبة
	1	2	3	4	5	6	7	8	9		
م	95101	68082	54796	14581	4783	799	148	8	1	238299	5.40935
ن	91279	198498	21691	9593	2906	638	401	56	19	325081	7.37929
ه	44267	36817	85099	32520	4291	956	127	9		204086	4.63272
و	111749	56242	15456	7272	2261	588	73			193641	4.39562
ى	2869	18616	35409	6550	1659	39				65142	1.47871
ي	86024	87855	79313	20342	3387	1279	114	25		278339	6.31825
أخرى										152325	3.45775

جدول ٦ : تكرار الحرف الأول في المقطع مع طول المقطع

الحرف الأول	تكراره كأول حرف في مقطع بطول									
	1 (منفصل)	2	3	4	5	6	7	8	9	مجموع
ء	11900									11900
آ	3552									3552
أ	103951									103951
ؤ	792									792
إ	29677									29677
ئ	184	2912	3602	160	45					6903
ا	213387									213387
ب	15540	106216	24975	6286	2456	483	29	10		155995
ة	17596									17596
ت	6134	9610	13122	4191	1885	471	60	15		35488
ث	2260	9600	39515	743	121	13				52252
ج	2966	14302	5296	2993	1160	64	4			26785
ح	3256	51019	14474	3682	624	67	7			73129
خ	243	8895	11693	1487	636	77	22			23053
د	18951									18951
ذ	13529									13529
ر	56148									56148
ز	8625									8625
س	5836	25053	34475	13131	2897	328	110	3		81833
ش	330	4793	4529	6015	121	15	1			15804
ص	481	4886	25817	1097	270	40	6			32597
ض	1563	7559	793	349	83	10				10357
ط	413	2025	1374	916	161	28				4917
ظ	40	225	529	146	17	8				965
ع	1961	53599	46091	31891	4676	231	23			138472
غ	217	1365	3346	612	180	32	1			5753
ف	2334	35292	24425	11738	3133	1340	331	58	4	78655
ق	4964	52950	7924	1920	1773	68	2		2	69603
ك	3075	17678	8273	2661	743	51	21			32502
ل	68159	87502	95723	44072	12144	2181	528	64	20	310393
م	14835	47959	14292	15878	1694	376	61	5	1	95101
ن	41675	31471	12424	4324	1076	281	28			91279
ه	10779	23695	8722	948	114	9				44267
و	111749									111749
ى	2869									2869
ي	9801	39753	17471	14698	3017	1188	88	7	1	86024

جدول ٣ : كل المقاطع بطول ٩ ونسبتها

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
لمستضعفين	0.00086	فليقطعهما	0.00009	قسطنطينية	0.00009
ليثنينهما	0.00005	فلتقتلنهم	0.00005	يستبينهما	0.00005
مستقبليها	0.00005	فليلتمسها	0.00005		

جدول ٤ : كل المقاطع بطول ٨ ونسبتها

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
مستخلفكم	0.00005	مستخلفين	0.00005	ستطعمتها	0.00005	فليحملها	0.00005	للمطففين	0.00005
ملتصقتين	0.00005	فسقيتهما	0.00005	لمجنبتين	0.00005	لمستقبلة	0.00005	لمطمئنين	0.00005
فهيجتهما	0.00005	بحبيبتيه	0.00005	فليحلبها	0.00005	تستطيعها	0.00005	فتستقبله	0.00005
فجعلتهما	0.00005	تخفيفهما	0.00005	يستحسنها	0.00005	فغمستهما	0.00005	يستكملها	0.00005
فقبضتهما	0.00005	لقطعتكما	0.00005	ليبتليكم	0.00005	فليمسكها	0.00005	فليمتهما	0.00005
يستعملها	0.00005	فمنعنيها	0.00005	فجمعتهما	0.00005	لمقتسمين	0.00009	فلبستهما	0.00009
يستقبلكم	0.00009	ففظعتهما	0.00009	ليخلعهما	0.00009	فليجعلها	0.00009	فليستجمر	0.00009
فليستغفر	0.00009	فليستنشق	0.00009	فليتحلله	0.00009	فليمنحها	0.00009	يستلمهما	0.00009
فتبيعنيه	0.00009	لمتخلفين	0.00009	بمغنيتين	0.00009	مستضعفين	0.00009	ستقبلهما	0.00009
تسليمتين	0.00009	تصلينهما	0.00014	لخليفتين	0.00014	لجهنميين	0.00014	لينعلهما	0.00014
لمتشبهين	0.00018	فليستئثر	0.00018	تستعملني	0.00018	تستعينها	0.00018	للمحلقين	0.00027
لمتكلفين	0.00027	فليستعفف	0.00027	بسيفيهما	0.00032	فنفختهما	0.00036	ليقطعهما	0.00036
فليطلقها	0.00045	للمسلمين	0.00086						

الحرف	تكراره في المقطع الذي بطول									مجموع	نسبة
	1	2	3	4	5	6	7	8	9		
ء	11900									11900	0.27013
آ	3552	1459	22	6	3					5042	0.11445
أ	103951	20129	2057	430	38	2				126607	2.87396
ؤ	792	1257	995	55	3					3102	0.07041
إ	29677	6277	43	1						35998	0.81715
ئ	6903	1374	1289	298	71	4	2			9941	0.22566
ا	213387	139833	113405	29471	10664	2251	475	75	6	509567	11.56709
ب	155995	46774	23260	6578	1199	44	8			233858	5.30854
ة	17596	10867	8824	11377	5350	625	40	1	2	54682	1.24127
ت	35488	33473	15133	11861	2205	203	1			98364	2.23285
ث	52252	7909	3554	1540	77	37	4			65373	1.48396
ج	26785	10529	3320	968	58	18				41678	0.94608
ح	73129	28380	4373	1729	288	50	2			107951	2.45047
خ	23053	5884	1186	248	31	4				30406	0.69021
د	18951	65359	30893	16728	1401	76	7			133415	3.02850
ذ	13529	11690	3202	426	102	18	7			28974	0.65771
ر	56148	53029	44166	15879	2392	420	20	8		172062	3.90578
ز	8625	9087	2960	499	52	11				21234	0.48201
س	81833	20162	9913	2325	273	48	1			114555	2.60038
ش	15804	11727	3060	466	36	3	2			31098	0.70592
ص	32597	10323	3605	692	101	11				47329	1.07436
ض	10357	4336	3008	697	107	9	2			18516	0.42031
ط	4917	7070	3082	1108	155	7				16339	0.37089
ظ	965	1492	958	944	143	18				4520	0.10260
ع	138472	46144	14845	3947	684	181	9			204282	4.63717
غ	5753	2932	1604	515	36	6				10846	0.24620
ف	78655	12457	6785	2098	584	86	33	6		100704	2.28596
ق	69603	30605	6591	2352	350	88	1	2		109592	2.48772
ك	32502	23626	6168	3832	710	101	14			66953	1.51982
ل	310393	184785	26657	9909	1499	253	21			533517	12.11075

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
عنه	0.1348	سمعت	0.134	ج	0.1338	ى	0.1294	لذ	0.1266	قو	0.1235				
عر	0.1167	لإ	0.1166	فر	0.116	عبا	0.1154	جر	0.113	سا	0.1116				
مة	0.1106	لي	0.1096	هيم	0.1094	حر	0.1083	ية	0.1076	فع	0.1068				
ين	0.1064	ئشة	0.1056	حمن	0.1056	ف	0.1053	ث	0.1019	هل	0.1013				
علي	0.1005	صا	0.0972	هم	0.0964	يث	0.0943	قر	0.0937	يز	0.0935				
حا	0.0888	سفيا	0.0884	ع	0.0884	شعبة	0.0876	سحا	0.0872	خا	0.0869				
بيه	0.0867	عبيد	0.0866	عند	0.0866	عو	0.0825	سما	0.0812	هب	0.0776				
كل	0.0767	نت	0.0763	قلت	0.0753	سى	0.0752	غير	0.0734	تر	0.0732				
شيبة	0.0726	بين	0.072	بعد	0.0714	ض	0.0705	عنهما	0.0687	فلما	0.0674				
شها	0.0665	ته	0.0663	كم	0.0654	فيه	0.0638	مع	0.0634	هما	0.0631				
معا	0.0631	هشا	0.0622	فلا	0.0621	خذ	0.0617	سلمة	0.0615	عيل	0.0615				
لما	0.0608	نز	0.0608	لآ	0.06	بها	0.0593	فو	0.0584	فقلت	0.058				
سعد	0.0566	سمع	0.056	عد	0.0556	بهذ	0.0548	لصلا	0.0541	خل	0.0539				
نها	0.0537	سنا	0.0533	ثا	0.0527	شر	0.0525	قبل	0.0522	ثلا	0.052				
يت	0.0519	لها	0.0517	عنها	0.0497	جد	0.0497	هير	0.0494	قتيبة	0.0493				
نو	0.0493	شي	0.0489	سليما	0.0488	منه	0.048	هد	0.0479	لمثنى	0.0476				
سأ	0.0472	كذ	0.0471	قتا	0.0467	يعني	0.0467	تو	0.0466	يأ	0.046				
عة	0.0458	جعفر	0.0457	طا	0.0457	بأ	0.0455	شا	0.045	نصا	0.0449				
بة	0.0447	بد	0.0442	كنت	0.044	صلا	0.0437	تا	0.0433	ليه	0.0429				
بير	0.0425	فيها	0.0422	نعم	0.0422	لهم	0.0421	عمش	0.0419	لجنة	0.0419				
خير	0.0417	فد	0.0416	لمد	0.0415	لحد	0.0414	عثما	0.0414	فذ	0.0409				
لحا	0.0404	ليس	0.0401	جلا	0.04	لقر	0.0397	سر	0.0396	صد	0.0396				
شيئا	0.039	بت	0.0388	ينة	0.0384	حما	0.0383	ينا	0.0383	نما	0.0382				
حين	0.0382	بني	0.0381	صحا	0.0373	فلم	0.0371	عطا	0.0371	لقا	0.0371				
قة	0.0367	جو	0.0367	حميد	0.0367	للهم	0.0363	لمر	0.0362	ؤ	0.0357				
مه	0.0353	لقد	0.0352	معمر	0.0352	لسا	0.035	منها	0.035	كما	0.0347				
كنا	0.0342	منا	0.0341	يب	0.0341	لليث	0.034	مثل	0.0335	كو	0.0334				
لح	0.0333	يج	0.0332	حمد	0.0332	لمؤ	0.0331	تي	0.033	بيد	0.0325				
حب	0.0324	بنت	0.0322	سف	0.032	يذ	0.032	مي	0.0318	نك	0.0316				
نمير	0.0316	هي	0.0311	نسا	0.0309	لقو	0.0308	سم	0.0306	تى	0.0295				
معه	0.0295	للفظ	0.0293	نة	0.0293	بشا	0.0292	فما	0.0292	تم	0.029				
بما	0.0285	بنا	0.0285	بشر	0.0285	جميعا	0.0284	ليد	0.0284	صو	0.0282				
سلا	0.0281	لمسجد	0.028	جع	0.0279	عشر	0.0275	لعز	0.0273	منصو	0.0272				
لكم	0.0272	كلا	0.027	شد	0.0267	لتي	0.0266	بك	0.0258	حو	0.0258				
فيقو	0.0257	مسلم	0.0257	نحو	0.0257	صم	0.0256	كثير	0.0256	لحسن	0.0254				
ثو	0.0252	ضي	0.0252	لقيا	0.0252	يحد	0.0251	بيع	0.0243	سه	0.0243				
يقا	0.0242	لسما	0.0239	عليها	0.0238	يصلي	0.0235	عز	0.0234	تأ	0.0233				
ثة	0.0232	لبيت	0.0232	لعا	0.0231	منهم	0.0231	نهم	0.0231	تقو	0.0231				
تد	0.023	علم	0.023	لليل	0.0229	يه	0.0229	لسلا	0.0227	فكا	0.0226				

جدول ١ : المقاطع الأكثر تكرارا: أول ٣٥٠ مقطعا

طول المقطع	عدد	تكرار	نسبة
1	57	942097	42.49079
2	537	638359	28.79149
3	3192	418885	18.89270
4	6778	169938	7.664604
5	4896	39026	1.760164
6	2118	7361	0.331998
7	522	1322	0.059625
8	62	162	0.007306
9	8	28	0.001262
المجموع	18170	2217178	100

جدول ٢ : أطوال المقاطع وتكرارها ونسبتها

يشمل الجدول رقم ٥ تكرار كل حرف من حروف العربية في كل مقطع ظهر به حسب طول المقطع. لاحظ أن الهمزة المنفصلة "ء" لم تظهر إلا في المقطع الذي طوله ١ لأنها لا تتصل بما قبلها ولا بما بعدها. بينما حرف الميم على سبيل المثال "م" ظهر منفصلا (في المقطع بطول ١) كما ظهر في كل المقاطع ذات الأطوال المختلفة من ١ إلى ٩. وملاحظة أخرى تبدو لطيفة وهي أن الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتما لن يظهر في مقطع أطول من ذاك المقطع.

يشابه الجدول رقم ٦ سابقه غير أنه يستعرض أول المقطع فقط، ولذا نلاحظ أن الحروف التي لا تتصل من اليسار لا تظهر كحرف أول إلا في منفصلة (طول المقطع ١)، وهذه الحروف هي: ء، ا، أ، ؤ، إ، ا،ة، د، ذ، ر، ز، و، ى. وتتكرر هنا ملاحظة أن الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتما لن يظهر في مقطع أطول من ذاك المقطع.

أما الجدول رقم ٧ فيتعامل مع الحرف الأخير من المقطع. لاحظ أنه في المقطع المنفصل بطول حرف واحد يكون الحرف هو الأول وهو الأخير. ومن فوائد البيانات في هذا الجدول معرفة نسبة ظهور الحرف متصلا في آخر الكلمة، ومثال ذلك: بالنظر إلى الهاء نرى أنها تأتي متصلة بما قبلها عشرة أضعاف ظهورها منفصلة. بينما تأتي التاء المربوطة منفصلة أكثر من ضعفي ظهورها متصلة بما قبلها. وترصد الجداول من ٨ إلى ١٥ تكرار الحرف والحرف الذي يليه في بيانات الدراسة، فيظهر الجدول رقم ٨ تكرار ظهور الحرف الأول والحرف الثاني، ويعرض الجدول رقم ٩ تكرار ظهور الحرف الثاني والثالث، ثم جدول رقم ١٠ يعرض تكرار ظهور الحرف الثالث والرابع وهكذا حتى جدول رقم ١٥ الذي يظهر تكرار ظهور الحرف الثامن والتاسع. ويستعرض الجدول رقم ١٦ تكرار ظهور الحرف التاسع. ونلاحظ من هذه الجداول أن أكثر الحروف الثنائية تكرارا تقع في المقاطع ذات الأطوال ٢ و٣.

النتائج

عرضنا في هذه الدراسة مجموعة من نتائج تحليل كتابي البخاري ومسلم بحيث بينت الدراسة نسبة استخدام حروف اللغة العربية ومقاطعها.ويمكن الاستفادة من هذه النتائج في التعرف الآلي على الكتابة العربية حيث تفيد في زيادة دقة التعرف على الحروف كما تساعد في المعالجة التي تتم بعد التعرف في تصحيح الأخطاء. ويمكن استخدام هذه النتائج لحساب احتمالات ظهور حرف بعد حرف آخر، أو الاستفادة من احتمالية ظهور المقاطع كاملة، كما يمكن الاستفادة منها وهي متوفرة لدى الباحثين كنسخة الكترونية يمكن الاستفادة منها لمن يرغب بذلك. كما يمكن استخدامها في التعرف الآلي على قراءة الكتابة العربية باستخدام نماذج ماركوف المخفاة (Hidden Markov Models).

وينوي الباحثين تطوير الدراسة باستخدام عينات أخري لكتب متنوعة ومواضيع مختلفة ومجلات ومعلومات متوفرة على الشبكة الإنترنت. . كما ينوي الباحثون استخدام نتائج هذه الدراسة في التعرف الآلي على الكتابة العربية.

شكر

يتقدم الباحثان الأول والثاني بالشكر لجامعة الملك فهد للبترول والمعادن على الدعم المتواصل.

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
ا	9.6243	و	5.0401	أ	4.6884	ل	3.0741	لله	2.6366	ر	2.5324
بن	2.4033	حد	1.9598	ن	1.8796	قا	1.8777	عن	1.794	ثنا	1.432
(	1.3445	)	1.3443	إ	1.3385	عليه	1.054	صلى	0.9907	سلم	0.9786
د	0.8547	ة	0.7936	من	0.7781	لا	0.7274	سو	0.7138	ب	0.7009
بي	0.6725	م	0.6691	[	0.6575	]	0.6575	ما	0.6419	با	0.618
ذ	0.6102	عبد	0.5851	في	0.5614	ء	0.5367	نا	0.5082	ه	0.4862
لى	0.4743	فقا	0.4485	ي	0.442	بو	0.4281	لأ	0.4192	لنبي	0.4033
ز	0.389	1	0.3751	خبر	0.3578	2	0.3426	كا	0.3291	على	0.3285
عمر	0.3153	لر	0.3111	لك	0.2902	ثم	0.2846	ني	0.283	محمد	0.279
ت	0.2767	فأ	0.2725	لو	0.272	له	0.2646	س	0.2632	ير	0.2598
يا	0.2569	مر	0.2566	يو	0.2552	يد	0.2516	ثني	0.2469	5	0.244
6	0.2431	بر	0.2419	3	0.2417	4	0.2416	تعا	0.234	عا	0.233
ق	0.2239	هر	0.215	ضى	0.2082	7	0.2014	}	0.2008	9	0.1981
{	0.1961	لت	0.192	8	0.1907	لنا	0.1907	0	0.1891	يحيى	0.1861
نه	0.1833	جا	0.1828	حتى	0.1806	لم	0.1797	سعيد	0.1746	كر	0.1711
قد	0.1709	فا	0.1697	لد	0.1643	ها	0.1605	آ	0.1602	يقو	0.1556
بكر	0.1528	فإ	0.1525	جل	0.152	هو	0.1499	هذ	0.1472	ح	0.1469
نس	0.1452	لز	0.1446	خر	0.1432	مو	0.1432	ك	0.1387	به	0.1352

إن نتائج هذا البحث وخاصة فيما يخص احتمالات تتابع الحروف العربية يمكن استخدامه في التعرف الآلي على الكتابة العربية دون الحاجة إلى تحليل عينات التدريب في كل مرة. كما أنه ونظرا لحجم العينات المستخدمة فإن نتائج هذا البحث من المنتظر أن تكون أكثر دقة.

تحتوي فقرة ٢ من البحث على تعريفات للرموز المستخدمة والفقرة ٣ على تحليل نتائج الدراسة والفقرة ٤ على النتائج والاقتراحات.

تعريفات

نقدم تحت هذا العنوان مجموعة تعريفات للمصطلحات المستخدمة في الدراسة. وهذه التعريفات هي: مقطع، ومنفصل، ومتصل، وأول حرف في المقطع، وآخر حرف في المقطع وطول المقطع.

مقطع: نعرف المقطع على أنه مجموعة الحروف المتصلة ببعضها في كلمة واحدة. فمثلا كلمة "مقطع" مكونة من مقطع واحد طوله أربعة حروف، وكلمة "مجموعة" مكونة من مقطعين هما "مجمو" بطول أربعة حروف و"عة" بطول حرفين، وكلمة "الاستحسان" مكونة من أربعة مقاطع هي: "ا" و"لا" و"ستحسا" و"ن"، وأطوالها ١ و٢ و٥ و١ على الترتيب.

منفصل: نقول أن الحرف منفصل إذا كان غير مرتبط بما بعده، وغير مرتبط بما قبله، أي أنه يكون مقطعا لوحده. فمثلا في كلمة "أن" ذات المقطعين يمثل كل مقطع منها حرفا منفصلا، أي أن "أ" حرف منفصل في هذه الكلمة و"ن" حرف منفصل أيضا فيها، بينما "أ" في كلمة "لأنه" ذات المقطعين يمثل حرفا متصلا، وكذلك حرف "ن" في المقطع الثاني مثالا على الحرف المتصل.

متصل: نقول أن الحرف متصل إذا كان مرتبطا بما بعده، أو مرتبطا بما قبله، أو بكليهما. ومثال ذلك في كلمة "مرتبطا" كل حرف من حروفها مثالا على الحروف المتصلة.

أول حرف: هو الحرف الأول في المقطع. ويمكن أن يكون غير متصل بما بعده فيكون مقطعا لوحده بطول ١. ومثال ذلك كلمة "أول" تتكون من ثلاثة مقاطع، كل حرف منها هو أول حرف في المقطع (وهو آخر حرف في المقطع)، بينما تمثل كلمة "متحابين" مقطعين، أول حرف في المقطع الأول هو "م" وأول حرف في المقطع الثاني هو "ب".

آخر حرف: هو الحرف الأخير في المقطع. وهو غير متصل بما بعده. وإذا لم يكن متصلا بما قبله فيكون مقطعا لوحده بطول ١. ومثال ذلك كلمة "متحابين"، فآخر حرف في المقطع الأول هو "ا" وهو متصل، وآخر حرف في المقطع الثاني هو "ن" وهو أيضا متصل.

طول المقطع: عدد الحروف التي تكون المقطع. ومثال ذلك كلمة "فرحتنا" تتكون من مقطعين: "فر" بطول ٢، و"حتنا" بطول ٤.

التحليل الإحصائي

تم عمل الدراسة على نص كتابي صحيح البخاري وصحيح مسلم كمثال على النصوص العربية المتعلقة بالحديث الشريف. وبلغ تعداد المقاطع في الكتابين ٢٢١٧١٧٨ مقطعا مع التكرار، و١٨١٧٠ مقطعا بعد حذف المكرر. بينما بلغ عدد الحروف والرموز في الكتابين ما مجموعه ٤٤٠٥٣١٨ رمزا. ويشمل الكتابان على ١٠٩٥٢٧٤ كلمة مع التكرار و٥٠٣٦٧ كلمة بعد حذف المكرر.

تبين الجداول من ١ إلى ١٤ مجموعة من النتائج المستخلصة من الدراسة. ويمكن للباحث الراغب في الحصول على الجدول الأساس الكلي والذي يزيد على ٣٠٠ صفحة الاتصال بأحد الباحثين إذ لا يمكن إرفاقه مع هذا البحث. وفي يلي وصفا تحليليا مختصرا لهذه

الجداول

يعرض جدول رقم ١ أول ٣٤٨ مقطعا من المقاطع الأكثر تكرارا من مجموع ١٨١٧٠ مقطعا. ولاحظ أن حرف الألف المنفصل يمثل حوالي ١٠٪ من تعداد المقاطع، يليه حرف الواو المنفصل، فالألف المهموزة، فاللام. ثم المقطع الثلاثي "لله". وهكذا. ويلاحظ المستعرض للجدول أن هناك العديد من المقاطع بطول ٢ و٣ و٤ و٥ وتظهر أكثر شيوعا من مقاطع أخرى بأطوال أقل.

ويعرض جدول رقم ٢ تكرار المقاطع حسب طولها ونسبة وجود هذه المقاطع. ونلاحظ أن أكثر من ٩٠٪ من المقاطع أطوالها ثلاثة حروف فأقل، وأن ما يصل إلى ٩٨٪ من المقاطع أطوالها أربعة حروف فأقل. ونلاحظ أن المقاطع بطول ٩ هي قليلة وتبلغ ٨ مقاطع أغلبها ظهر مرة واحدة، وقد جمعنا هذه المقاطع في جدول رقم ٣. ويبين جدول رقم ٤ كل المقاطع بطول ٨ حروف وعددها ٦٢ مقطعا وهي أيضا قليلة الظهور كما هو واضح في جدول رقم ٢.

تحليل إحصائي لدعم التعرف الآلي على الكتابة العربية

م. حسني المحتسب ود. صبري محمود

قسم علوم الحاسب والمعلومات، جامعة الملك فهد للبترول والمعادن، الظهران، المملكة العربية السعودية

د. رامي قهوجي

قسم التصوير الإلكتروني والاتصال، كلية علم المعلومات، جامعة برادفورد، برادفورد، بريطانيا

الكلمات المفتاحية: مقاطع كلمات اللغة العربية، تكرار الحروف العربية، التعرف الآلي على الكتابة العربية.

ملخص: تقدم هذه الدراسة ملخصا لنتائج دراسة إحصائية لأعداد ظهور حروف ومقاطع الكلمات في اللغة العربية. وتشمل النتائج المعروضة تكرار كل حرف من الحروف العربية في كل مقطع من المقاطع، وتكرار الحرف والحرف الذي يليه في المقاطع المختلفة لكل الحروف. كما تشمل الدراسة على إحصائيات استخدام الحروف والمقاطع ونسبة استخدام كل منها في حالات الاستخدام المختلفة في اللغة العربية. وقد تم تطبيق الدراسة عل كتابي صحيح البخاري ومسلم. وتفيد الدراسة في المساعدة في عملية التعرف الآلي على الكتابة العربية، كما تفيد في عملية تصحيح الأخطاء بعد عملية التعرف.

مقدمة

اتسمت السنوات الخمس الأخيرة بزيادة اهتمام الباحثين بالحوسبة العربية وتطبيقاتها المختلفة. ويلاحظ هذا الاهتمام في العديد من تطبيقات الحوسبة العربية، ومنها التعرف الآلي (أو ما يعرف بالتعرف الضوئي) على الكتابة العربية، سواء كانت كتابة مطبوعة أو بخط اليد . يقدم هذا البحث تحليلا لمقاطع اللغة العربية ولورود الحروف العربية في مواقعها المختلفة من المقاطع في محاولة لدعم التعرف الآلي على الكتابة العربية. وحسب علمنا فإن هنالك محاولات قليلة تمت في هذا الاتجاه ولكنها مختلفة من حيث النتائج وبعضها في النتائج والأهداف. فقد قدم البحث إحصائيات لمقاطع اللغة العربية ونسب المقاطع المتكونة من حرف إلى ثمانية حروف وكذلك نسبة الحروف في بداية الكلمة ووسطها وآخرها للمقاطع المتكونة من حرفين إلى ثمانية حروف. كما عرض البحث معدل عدد الحروف في مقطع الكلمة ٤,٣ وكذلك معدل عدد المقاطع في الكلمة العربية ٢,٢. وقد عرض البحث استخدام هذه المعلومات في التعرف الآلي على الكتابة العربية . وذكر الباحثون أن الكتابة العربية التي جرى تحليلها تم جمعها بشكل عشوائي من كتب قديمة وكتب حديثة وجرائد ومعلومات تم جمعها من شبكة الإنترنت. وتتكون العينات في بحثهم من ٢٦٢٦٤٧ كلمة تتكون من ١١٢٦٤٢٠ حرفا. وفي دراستنا تتضمن الإحصاءات تحليلات إضافية سنستعرضها لاحقا كما أن حجم العينة المستخدمة أكبر. أما البحث الثاني ذا العلاقة فهو يهدف إلى إنتاج معجم لعملية التعرف على الكتابة العربية وذلك في استخدامه للتعرف على مقاطع اللغة العربية وذلك تجنبا لتقطيع الحروف العربية (الذي يؤدي إلى زيادة الأخطاء في التعرف الآلي). اعتمد البحث على تحليل ٢٦٨٠٥ ملفا بصيغة أتش تي ام ال (HTML) لمجلة الدستور الأردنية. وتحتوي تلك العينات على الأخبار المحلية وفيها أخبار زيارات دبلوماسيين أجانب والتي تتضمن كلمات أعجمية منها أسماء الدول والمدن والأسماء الأجنبية. وقد احتوت العينات على ٢٨٥١٩٧٣٤ مقطعا. وقد بينت الدراسة إحصاءات للمقاطع بطول يتراوح من حرف إلى ثلاثة عشر حرفا. وبدراسة النتائج فإن أعداد المقاطع بطول ١٢ و ١٣ قليلة جدا وبنسبة لا تكاد تذكر والراجح أنها تتكون من مقاطع تحتوي على أخطاء. كما بينت رسالة الماجستير محاولة أخرى لتحليل الكتابة العربية بغرض عمل معجم. وقد كانت العينات المستخدمة مستخلصة من شبكة الانترنت بشكل رئيس. وتحتوي العينات على ٨٤٢٦٨٤ كلمة من ٤١٥ عينة مختلفة. وتحتوي العينات على مواضيع مختلفة كالاقتصاد وقصص الأطفال والتعليم والصحة والطب ومقابلات وسياسة ...الخ. وقد بينت الدراسة إحصاءات لمقاطع تتراوح بين حرف واحد وسبعة عشر حرفا. وتحتوي العينات على أخطاء مطبعية ووجود مقاطع غير حقيقية في اللغة نتيجة اندماج أكثر من مقطع كما بين الباحث بأمثلة من هذه المقاطع مما يجعل هذه العينات غير دقيقة المحتوى.

وقد استخدم بعض الباحثين احتمالات تتابع الحروف العربية في التعرف الآلي على الكتابة العربية باستخدام نماذج ماركوف المخفاة (Hidden Markov Models) وذلك بتحليل الكتابة العربية للعينات المستخدمة في التدريب واستخدام تلك النتائج في مرحلة الفحص. وحيث أن العينات المستخدمة في التدريب محدودة نوعا ما فإن احتمالات تتابع الحروف قد لا تكون دقيقة. وقد حسب الباحثون احتمالات وجود حرف يتبعه آخر واحتمالات وجود حرف قبله حرف وبعده حرف. إلا أن هذه الاحتمالات تستخدم داخليا في نظام التعرف الآلي على الكتابة العربية هذا ولم يعرض الباحثين تلك النتائج.