The Writer Authentication by Using Syllables Frequency

Amer A. Abdulrahman*

Date of acceptance 8/1/2008

Abstract

An approach is depended in the recent years to distinguish any author or writer from other by analyzing his writings or essays. This is done by analyzing the syllables of writings of an author. The syllable is composed of two letters; therefore the words of the writing are fragmented to syllables and extract the most frequency syllables to become trait of that author.

The research work depend on analyzed the frequency syllables in two cases, the first, when there is a space between the words, the second, when these spaces are ignored. The results is obtained from a program which scan the syllables in the text file, the performance is best in the first case since the sequence of the selected syllables is higher than the same syllables in the second case.

Introduction.

The main objective of this research is to establish recognition strategy to differentiate the writings of the writers and confirm that these writings is belong to one author and does not belong to other. Any writer has characteristics in his writings, when he writes his writings, poetries, essays or prose's, he has some special lineament that may be found in lower degree in other writers.

In this research, take several writings for some famous writers [4] , [3] and analyzed their writings.

The kind of the of these writings is essays or prose's because the poetries have rhyme and this rhyme may effect on the selected syllables [1]

Distinguishing Strategy

This research is depending on extract the frequency of the syllables in the text file. Table (1) and Table (2) described the syllables were used in this research. The writing is converted to a text file, and the syllables of it is checked sequentially.

Example:- Figure (1) is one of the writing belong to WELLIAM SHAKESPEARE, the first word is THE has the syllables TH , HE and E , the second word is has the syllables Wo , or , rl and ld and in the same manner, all the text file is converted to syllables, obtain the number of frequency of each syllables depending on the syllables were exist in Table (1) and Table(2).

THE World is too much with us; late and soon,
Getting and spending, we lay waste our powers:
Little we see in Nature that is ours;
We have given our hearts away, a sordid boon!
This Sea that bares her bosom to the moon,
The winds that will be howling at all hours
And are up-gather'd now like sleeping flowers,
For this, for everything, we are out of tune;
It moves us not.Great God! I'd rather be
A pagan suckled in a creed outworn,
So might I, standing on this pleasant lea,
Have glimpses that would make me less forlorn;
Have sight of Proteus rising from the sea;
Or hear old Triton blow his wreathèd horn.

Figure (1) writing for SHAKESPEARE [2]

Fragmenting Algorithm

Input: I as integer, text1 as a sequential file
Out put: syllable as string

Step1-set I to zero.
Step2- convert a writing to a text1.
Step3- do until end of file
Step4- I = I + 1
Step5- syllable(i) = text1(i) +text1(i+1)
Step6- loop

Table (1) the syllables of two letter from aa to mz

aa	ba	ca	da	ea	fa	ga	ha	ia	ja	ka	la	ma
ab	bb	cb	db	eb	fb	gb	hb	ib	jb	kb	lb	mb
ac	bc	cc	dc	ec	fc	gc	hc	ic	ic	kc	lc	mc
ad	bd	cd	dd	ed	fd	gd	hd	id	jd	kd	ld	md
ae	be	ce	de	ee	fe	ge	he	ie	je	ke	le	me
af	bf	cf	df	ef	ff	gf	hf	if	jf	kf	lf	mf
ag	bg	cg	dg	eg	fg	gg	hg	ig	jg	kg	lg	mg
ah	bh	ch	dh	eh	fh	gh	hh	ih	jh	kh	lh	mh
ai	bi	ci	di	ei	fi	gi	hi	ii	ji	ki	li	mi
aj	bj	cj	dj	ej	fj	gj	hj	ij	jj	kj	lj	mj
ak	bk	ck	dk	ek	fk	gk	hk	ik	jk	kk	lk	mk
al	bl	cl	dl	el	fl	gl	hl	il	jl	kl	ll	ml
am	bm	cm	dm	em	fm	gm	hm	im	jm	km	lm	mm
an	bn	cn	dn	en	fn	gn	hn	in	jn	kn	ln	mn
ao	bo	co	do	eo	fo	go	ho	io	jo	ko	lo	mo
ap	bp	cp	dp	ep	fp	gp	hp	ip	jp	kp	lp	mp
aq	bq	cq	dq	eq	fq	gq	hq	iq	jq	kq	lq	mq
ar	br	cr	dr	er	fr	gr	hr	ir	jr	kr	lr	mr
as	bs	cs	ds	es	fs	gs	hs	is	js	ks	ls	ms
at	bt	ct	dt	et	ft	gt	ht	it	jt	kt	lt	mt
au	bu	cu	du	eu	fu	gu	hu	iu	ju	ku	lu	mu
av	bv	cv	dv	ev	fv	gv	hv	iv	jv	kv	lv	mv
aw	bw	cw	dw	ew	fw	gw	hw	iw	jw	kw	lw	mw
ax	bx	cx	dx	ex	fx	gx	hx	ix	jx	kx	lx	mx
ay	by	cy	dy	ey	fy	gy	hy	iy	jy	ky	ly	my
az	bz	cz	dz	ez	fz	gz	hz	iz	jz	kz	lz	mz

Table (2) the syllables of two letter from na to zz

na	**oa**	**pa**	**qa**	**ra**	**sa**	**ta**	**ua**	**va**	**wa**	**xa**	**ya**	**za**
nb	**ob**	**pb**	**qb**	**rb**	**sb**	**tb**	**ub**	**vb**	**wb**	**xb**	**yb**	**zb**
nc	**oc**	**pc**	**qc**	**rc**	**sc**	**tc**	**uc**	**vc**	**wc**	**xc**	**yc**	**zc**
nd	**od**	**pd**	**qd**	**rd**	**sd**	**td**	**ud**	**vd**	**wd**	**xd**	**yd**	**zd**
ne	**oe**	**pe**	**qe**	**re**	**se**	**te**	**ue**	**ve**	**we**	**xe**	**ye**	**xe**
nf	**of**	**pf**	**qf**	**rf**	**sf**	**tf**	**uf**	**vf**	**wf**	**xf**	**yf**	**zf**
ng	**og**	**pg**	**qg**	**rg**	**sg**	**tg**	**ug**	**vg**	**wg**	**xg**	**yg**	**zg**
nh	**oh**	**ph**	**qh**	**rh**	**sh**	**th**	**uh**	**vh**	**wh**	**xh**	**yh**	**zh**
ni	**oi**	**pi**	**qi**	**ri**	**si**	**ti**	**ui**	**vi**	**wi**	**xi**	**yi**	**zi**
nj	**oj**	**pj**	**qj**	**rj**	**sj**	**tj**	**uj**	**vj**	**wj**	**xj**	**yj**	**zj**
nk	**ok**	**pk**	**qk**	**rk**	**sk**	**tk**	**uk**	**vk**	**wk**	**xk**	**yk**	**zk**
nl	**ol**	**pl**	**ql**	**rl**	**sl**	**tl**	**ul**	**vl**	**wl**	**xl**	**yl**	**zl**
nm	**om**	**pm**	**qm**	**rm**	**sm**	**tm**	**um**	**vm**	**wm**	**xm**	**ym**	**zm**
nn	**on**	**pn**	**qn**	**rn**	**sn**	**tn**	**un**	**vn**	**wn**	**xn**	**yn**	**zn**
no	**oo**	**po**	**qo**	**ro**	**so**	**to**	**uo**	**vo**	**wo**	**xo**	**yo**	**zo**
np	**op**	**pp**	**qp**	**rp**	**sp**	**tp**	**up**	**vp**	**wp**	**xp**	**yp**	**zp**
nq	**oq**	**pq**	**qq**	**rq**	**sq**	**tq**	**uq**	**vq**	**wq**	**xq**	**yq**	**zq**
nr	**or**	**pr**	**qr**	**rr**	**sr**	**tr**	**ur**	**vr**	**wr**	**xr**	**yr**	**zr**
ns	**os**	**ps**	**qs**	**rs**	**ss**	**ts**	**us**	**vs**	**ws**	**xs**	**ys**	**zs**
nt	**ot**	**pt**	**qt**	**rt**	**st**	**tt**	**ut**	**vt**	**wt**	**xt**	**yt**	**zt**
nu	**ou**	**pu**	**qu**	**ru**	**su**	**tu**	**uu**	**vu**	**wu**	**xu**	**yu**	**zu**
nv	**ov**	**pv**	**qv**	**rv**	**sv**	**tv**	**uv**	**vv**	**wv**	**xv**	**yv**	**zv**
nw	**ow**	**pw**	**qw**	**rw**	**sw**	**tw**	**uw**	**vw**	**ww**	**xw**	**yw**	**zw**
nx	**x**	**px**	**qx**	**rx**	**sx**	**tx**	**ux**	**vx**	**wx**	**xx**	**yx**	**zx**
ny	**oy**	**py**	**qy**	**ry**	**sy**	**ty**	**uy**	**vy**	**wy**	**xy**	**yy**	**zy**
nz	**oz**	**pz**	**qz**	**rz**	**sz**	**tz**	**uz**	**vz**	**wz**	**xz**	**yz**	**zz**

Analysis the text

The text is fragmented to syllables. According to Table (1), the existence of the first syllable aa is checked, and the frequency of this syllable is summed. Convert the first letter of this syllable to A to became Aa , repeat the checking to added to the same sum. Convert the second letter of the syllable to A to became AA and repeat the checking to added to the same sum.

Analysis Algorithm

Input: I, pcount as integer, syllable as string, text1 as a sequential file
Output: frequency number to each syllable in text1

Step1- Set I to zero.
Step2- Do until end of file.
Step3- I=I+1
Step4- Take the ith syllable (see table (1)).

Step5- Perform Step8.
Step6- Convert the left letter of the ith syllable to capital form and perform Step8.
Step7- Convert the left and right letter of ith syllable to a capital form.
Step8- Check this syllable with all the fragmented syllables of text1,
If is it found then pcounti = pcounti + 1.
Step9- Loop.
Step10- End.

By taking 3-7 writings to any writer, and see the biggest frequency syllables, the syllable th is the biggest in all writings of all writes, therefore, this syllable is ignored because it cannot used for recognitions process. Check the writings of any writer to deduce the two characteristically frequency syllables. These two syllables represent as lineament of that writer and used them for recognizing his writings. These pair of syllables are used for recognize that writer only and must not use the same pair to recognize another writer, for example, the syllables that selected for Shakespeare are In and An , the syllable In must be in the sequence range (1-7), the syllable An must be in the sequence range (1-5).

A master file is created to each writer that contains all his writings to arrange the frequency syllables of these writing in decreasing order. This file is used to propping the recognizing process by ensuring the two characteristically syllables.

Determination of Writing Authentication Algorithm

Input: text, master-file as a sequential file
Out put: syllable as string

Step1:- Enter the characteristically syllables of this writer.
Step2:- If the pair of syllables were not exist in the determining range then go to step5.
Step3:-If the syllables were not exist in master-file then go to step5.
Step4:-This writing is for this writer, go to step6
Step5:- This writing is not for this writer
Step6:- End.

Result and Discussion

The first step is to determine how the syllables must be fragmented? There two ways, the first is take in account the spaces between the words, the second, is to ignored the spaces between the words.

When determining the permitting range, the sequence is used instead of frequency, this will avoid the problem of the size of the writing, i.e when the size of writing is large, the frequency is large too, and when the size of writing is small, the frequency is small too, but in the same case, the sequence is fixed, for example, as illustrated in Table(3), the syllables that selected for Shakespeare are In and An , the permitting range for finding In is 7<= seq <=1 and the permitting range for finding An is 5<= seq <=1.

As see in table (3) and table(4), the selected syllables must exist and have nearly sequences in all writings of the writer.

Analysis the Text Which Has Space between the Words

Table (3) illustrates the biggest frequency syllables in decreasing sequentially arrangement for three writings to three writers.

Table (3) three writers with their writings Se = sequence, Ph= syllables, Re= frequency

Se	Shakespeare						Tennyson						Worworth					
	Writing1		**Writing2**		**Writing3**		**Writing1**		**Writing2**		**Writing3**		**Writing1**		**Writing2**		**Writing3**	
	Ph	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**
1	Th	82	Th	49	In	23	Th	87	Th	82	Er	17	Th	72	Th	46	He	466
2	He	48	In	44	Th	22	Nd	56	He	64	Ve	13	He	65	He	43	Th	445
3	Ou	38	He	39	Nd	20	He	54	Nd	34	Th	13	In	35	Er	26	In	276
4	An	33	Re	38	An	20	In	51	An	32	He	12	An	30	Ea	22	Re	212
5	Re	32	An	36	On	17	An	51	De	29	Or	9	Er	30	Re	20	Er	210
6	Or	32	Nd	33	Ve	15	Re	40	Re	21	Nd	9	Ea	27	Ro	20	An	172
7	In	31	Es	30	Er	14	Er	37	Ll	20	An	9	Re	25	In	17	Ng	172
8	Nd	29	Ve	27	Es	13	To	35	Er	19	Fo	8	At	23	As	16	Nd	162
9	Ha	28	Ha	26	Or	12	Ve	33	He	18	Ev	8	Nd	22	Ed	16	Ar	153
10	At	28	Er	26	Re	11	On	32	La	18	Re	7	Ed	21	Or	16	Ea	145
11	To	27	Is	23	Ne	11	Le	29	In	17	Ll	7	Ou	21	An	15	St	144
12	Ar	27	Or	22	Me	11	At	28	No	17	Ee	7	Ar	20	Ar	15	Le	143
13	Er	26	on	21	Ll	11	En	27	Ro	17	Iv	6	As	20	At	15	Ro	136
14	Ea	24	Al	20	Me	11	Es	27	To	17	By	5	Ha	20	Ha	15	Ed	131
15	Ng	24	Hi	20	Ne	11	Ll	25	Of	16	Al	4	Ou	20	It	13	Es	130

Table (4) illustrates the master file for the same writers that descried in Table(3)

Table (4) the master file for three writersSe = sequence, Ph= syllables, Re= frequency

Se	Shakespeare		Tennyson		Worworth	
	Ph	**Re**	**Ph**	**Re**	**Ph**	**Re**
1	**Th**	**153**	**Th**	**182**	**He**	**574**
2	**In**	**98**	**He**	**130**	**Th**	**563**
3	**He**	**97**	**Nd**	**99**	**In**	**328**
4	**An**	**89**	**An**	**92**	**Er**	**266**
5	**Nd**	**82**	**Er**	**73**	**Re**	**257**
6	**Re**	**81**	**In**	**70**	**An**	**217**
7	**Or**	**66**	**Re**	**68**	**Nd**	**196**
8	**Er**	**66**	**To**	**53**	**Ng**	**196**
9	**Ha**	**65**	**Ll**	**52**	**Ea**	**194**
10	**Ou**	**64**	**Ve**	**52**	**Ar**	**188**
11	**Es**	**62**	**Le**	**50**	**Ro**	**178**
12	**Ve**	**59**	**On**	**50**	**St**	**169**
13	**On**	**57**	**Or**	**48**	**Ed**	**168**
14	**Is**	**55**	**De**	**45**	**Le**	**166**
15	**Ng**	**52**	**Ha**	**44**	**Es**	**161**

Analysis the Text After Ignored the Space between the Words

Table (5) illustrates the biggest frequency syllables in decreasing sequentially arrangement for three writings to three writers.

Table (5) three writers with their writings Se = sequence, Ph= syllables, Re= frequency

Se	Shakespeare						Tennyson						Worworth					
	Writing1		**Writing2**		**Writing3**		**Writing1**		**Writing2**		**Writing3**		**Writing1**		**Writing2**		**Writing3**	
	Ph	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**	**Ph**	**Re**
1	Th	84	Th	50	In	23	Th	91	Th	83	Er	17	Th	75	Th	49	He	467
2	He	48	In	46	Th	22	Nd	58	He	64	Re	15	He	65	He	43	Th	453
3	Ou	39	He	39	Nd	21	He	54	Nd	35	Th	14	In	32	Er	28	In	276
4	An	34	Re	39	An	20	An	53	An	32	Ve	13	Er	30	Ea	24	Er	234
5	Ea	34	An	36	Et	17	In	51	De	29	He	12	An	29	Ro	21	Re	212
6	Re	33	Es	36	On	17	Er	42	To	25	An	9	Ea	25	Re	20	Es	210
7	Or	32	Nd	34	Ve	15	Re	40	Re	21	Nd	9	Es	25	As	17	St	202
8	At	31	Ha	29	Er	14	To	37	Er	20	Or	9	Re	24	In	17	Ng	174
9	In	31	Er	27	Es	13	Es	36	Le	20	Ev	8	At	24	St	17	An	172
10	To	31	Ve	27	Ha	12	Ea	33	No	19	Fo	8	Nd	24	At	16	Ea	167
11	Es	30	Et	24	Ll	12	Ve	33	Ha	18	Ea	7	Ro	23	Ed	16	Nd	166
12	Ha	30	Ea	23	Ne	12	On	32	Le	18	Ll	7	St	22	Or	16	Ar	154
13	St	30	Is	23	Or	12	Et	31	In	17	Iv	6	Ed	21	An	15	Ed	151
14	Nd	29	Li	22	Re	12	Le	30	Of	17	By	5	As	21	Ar	15	Le	145
15	Ar	27	Or	22	Al	11	En	29	Or	17	Ea	5	Ha	21	Es	15	Ro	141

Table (6) illustrates the master file for the same writers that descried in Table(5)

Table (6) the master file for three writers Se = sequence, ph= syllables, re= frequency

Se	Shakespeare		Tennyson		Worworth	
	Ph	Re	**Ph**	Re	**Ph**	**Re**
1	**Th**	**156**	**Th**	**188**	**Th**	**577**
2	**In**	**100**	**He**	**130**	**He**	**575**
3	**He**	**97**	**Nd**	**102**	**In**	**328**
4	**An**	**90**	**An**	**94**	**Er**	**249**
5	**Nd**	**84**	**Er**	**79**	**Re**	**257**
6	**Re**	**84**	**Re**	**76**	**Es**	**250**
7	**Es**	**79**	**In**	**70**	**St**	**242**
8	**Ha**	**71**	**To**	**63**	**Ea**	**220**
9	**Er**	**68**	**Ea**	**54**	**An**	**217**
10	**Or**	**66**	**Es**	**53**	**Nd**	**203**
11	**Ea**	**65**	**Ll**	**53**	**Ng**	**199**
12	**Ou**	**65**	**Or**	**52**	**Ar**	**189**
13	**Et**	**63**	**Ve**	**52**	**Ed**	**186**
14	**Is**	**59**	**Le**	**51**	**Ro**	**168**
15	**St**	**59**	**On**	**50**	**Le**	**157**

Reference

1. Jump .J . D , 1979 , The Tragical History of the Life and Death.
2. Marzban K.B, (1990). The Plays of Shakespeare: Selected with Introduction to Each Pieces , Efficient Offset Printers, New Delhi, India.
3. Mayer M.,(1998). Poetry an Introduction, University of Connecticut ,USA, 2nd Edition.
4. Shakespeare W, Mittulme .M.A PH.D, (1980), Juluis Caeser, Long mans Green and Cold.

استخدام تقنية تكرار مقاطع الكلمة في أثبات هوية الكاتب

عامر عبد المجيد عبد الرحمن*

*مدرس مساعد – قسم الحاسبات كلية التربية للبنات - جامعة بغداد.

الخلاصة:

ان طرق تحليل الكتابات والمقالات وغيرها تم اعتمادها في السنين الاخيرة وذلك لتمييز أي مؤلف حيث يتم من خلال تحليل مقاطع الكتابات والمقالات , حيث يتكون المقطع من حرفين وعليه فأن كلمات النص تقطع الى مقاطع لاستخراج المقطع الاكثر تكرارا ليكون بمثابة الميزة للمؤلف .

تم في هذا البحث استخدام طريقة تحليل تكرار المقاطع ولمرحلتين , الاولى عندما يكون هنالك فراغ بين الكلمات والثانية عندما يهمل الفراغ ولا يؤخذ بنظر الاعتبار.

بينت النتائج المستحصلة من هذا البحث من خلال عمل برنامج لمسح المقاطع أن أداء الطريقة لاولى هو أفضل حيث أعطى نسبة تكرار أكبر من الثانية