



جامعة السودان للعلوم والتكنولوجيا

كلية الدراسات العليا

أداة مكملة لضبط الكلمات العربية بالشكل

## A Complementary Tool for Diacritizing Arabic Words

إعداد:

الروضة عبد اللطيف عبد الحليم حامد

إشراف:

أ.د. عز الدين محمد عثمان

بحث جزئي مقدم لنيل درجة الماجستير في علوم الحاسوب

أغسطس 2011

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

اقْرَأْ وَرَبُّكَ الْأَكْرَمُ ﴿٣﴾ الَّذِي عَلَّمَ بِالْقَلَمِ ﴿٤﴾

سورة العلق

# إهداء

إلى الحب الأول .. أمي

إلى الحاضر الغائب .. أبي

إلى من شاركوني لحظات السعادة والشقاء طول السنين الماضية .. إخواني

إلى كل من علمني حرفاً .. أساتذتي وأستاذاتي

إلى كل من دعمني معنوياً وشجعني على المضي قدماً

إلى زملائي في كلية النيل والذين كانوا معي قلباً وقالبا ..

إلى من لأجد سواهن عندما تضيق بي الدنيا .. صديقاتي

أهدي هذا العمل

# شكر و عرفان

من لايشكر الناس لايشكر الله. فله الحمد والمنّة من قبل ومن بعد.

أتقدم بجزيل الشكر إلى كل من:

1. أ.د. عز الدين محمد عثمان: المشرف على البحث وصاحب الفكرة

والذي استمتعت جداً بالعمل والنقاش معه.

2. د. هويدا علي عبد القادر أحمد والتي كان لها دور مقدر في إخراج

البحث بصورته النهائية.

3. أ. أشواق محمد صالح والتي كان لها عددٌ من الملاحظات التقنية الهامة.

4. د. صديق عمر صديق الأستاذ بقسم اللغة العربية بجامعة الخرطوم

ومعهد البروفيسور عبدالله الطيب.

# المستخلص

مع تزايد الإهتمام بمعالجة اللغة العربية الطبيعية -أي تصميم برامج تجعل الحاسوب قادراً على فهم نصوصها المكتوبة والمنطوقة والتعامل معها- برزت الحاجة لأنظمة التشكيل الآلي كمكونات أساسية في أنظمة معالجة اللغة العربية كالترجمة الآلية وتوليد الكلام من النصوص المكتوبة والتلخيص الآلي وغيرها، وتخلو غالبية النصوص العربية المعاصرة من علامات التشكيل ولكي تتمكن أنظمة معالجة اللغة العربية من إتمام مهامها بصورة جيدة فهذا يقتضي أن تكون مدخلاتها من النصوص مُشكَّلةً بالكامل وبصورة صحيحة لأنها لاتستطيع لوحدها اكتشاف التشكيل الصحيح على العكس من الإنسان الذي يعتمد في النطق الصحيح للكلمات على ذخيرته اللغوية التي اكتسبها في سنين حياته بل تعتمد في ذلك على أنظمة التشكيل الآلي والتي تستخدم عدة خوارزميات لاختيار الكلمة ذات التشكيل الأنسب. ومع بداية ثمانينيات القرن الماضي ظهرت العديد من أنظمة التشكيل الآلي التي طورتها شركات عربية وعالمية كصخر، RDI وجوجل واستُخدم بعضها كبنية تحتية لعددٍ من أنظمة معالجة اللغة العربية.

على الرغم من الدقة العالية لهذه الأنظمة فإنها تخطئ في تشكيل بعض الكلمات مما يستدعي وجود نظام مكمل لها لتصحيح هذه الأخطاء.

يساعد النظام المقترح في تشكيل الكلمات غير المُشكَّلة أو التي شكلت بصورة خاطئة وهو بذلك يتميز على بقية الأنظمة والتي يتم فيها مراجعة وتصحيح أخطاء التشكيل فيها يدوياً.

ويعتمد النظام المقترح على تقنية ربط الجافا بقواعد البيانات وهو نظام مكمل لأنظمة التشكيل الآلي حيث تمثل مخرجاتها مدخلات له فيقوم بتحديد الكلمات التي يشتبه في أن تكون قد شكّلت بصورة خاطئة -إعتمادا على قاعدة بياناته أو مايسمى بالذخيرة اللغوية- ويقدم للمستخدم عدة خيارات للتشكيل ليختار إحداها.

وقد أثبتت الإختبارات التي أجريت على النظام قدرته على تقليل نسبة الخطأ في تشكيل الكلمات والتي تحصل عند استخدام الانظمة الموجودة بمفردها.

# Abstract

With incremental care of the processing of natural Arabic language –i.e. enabling computer to understand its written and spoken texts and dealing with them- the need appears for automatic diacritization systems as basic components of natural Arabic language processing systems such as automatic translation, generating speech from written texts, automatic summarization etc. The majority of Arabic texts are written without diacritic marks; to enable Arabic language processing systems from achieving their tasks in a good manner inputs have to be a truly fully diacritized texts because these systems cannot discover right diacritization by themselves in opposite to human who depends on his linguistic asset which he gained through his life, but they depend on automatic diacritization systems which use several algorithms to select most appropriate diacritization. At the beginning of eighties of the last century several automatic diacritization systems emerged; these systems has been developed by Arabic and international companies such as Sakhr, RDI and Google; some of them was used as an infrastructure for a number of Arabic language processing systems.

Although of high accuracy of these systems but they diacritize some words wrongly, so there is a need for a complementary system to correct those errors. The proposed system helps in diacritizing wrongly or none diacritized words, this distinguish it from other systems which need manual revision and correction for diacritization errors.

The proposed system depends on Java Database Connectivity technology; it is a complement for automatic diacritization system, so their output acts as its input, then it specifies the words which may be diacritized in wrong manner –

depending on its database or what we call it linguistic asset- and presents multiple diacritization choices for user to select one of them.

Tests on system approved his ability to decrease percentage of diacritization errors which happens when we use existing systems alone.



# المصطلحات

Affixes.....	السوابق واللواحق.....
Natural language Processing (NLP).....	معالجة اللغات الطبيعية.....
American Standard Code for Information Interchange (ASCII)	الشفرة الأمريكية القياسية لتبادل المعلومات ..
Automatic hyphenation.....	وضع الشروط آلياً.....
Automatic translation .....	الترجمة الآلية .....
Context Free Grammar (CFG) .....	القواعد غير المقيدة بالسياق.....
Corpus .....	الذخيرة اللغوية/ المتن/ المكنز اللغوي .....
Computational Linguistics .....	اللسانيات الحاسوبية.....
Diacritization.....	التشكيل.....
Diacritics .....	علامات التشكيل .....
Encoding .....	التشفير.....
Glottal stops .....	وقفات مزمارية.....
Grammar Checking.....	التدقيق النحوي.....
Graphemes.....	الجرافيمات .....
Hidden Markov Models (HMM) .....	نماذج ماركوف الخفية .....

International Phonetic Alphabet (IPA) .....	التمثيل الصوتي العالمي للحروف
Java Database Connectivity .....	ربط الجافا بقواعد البيانات
Morphemes .....	وحدات صرفية
Morphological Analysis.....	التحليل الصرفي
Optical Character Recognition(OCR) .....	التعرف الضوئي على الحروف
Syntactic Analysis.....	التحليل النحوي
Semantic Analysis.....	التحليل الدلالي
Part of Speech tagging(POS).....	تبويب أقسام الكلام
Phonetic Analysis.....	التحليل الصوتي
Prefix.....	سابقة
Root.....	جذر
Suffix.....	لاحقة
Speech Recognition .....	التعرف على الكلام
Spell Checking.....	التدقيق الإملائي
Text Preparation\Editing.....	تجهيز/تحرير النصوص
Voice Recognition.....	التعرف على الصوت

# قائمة بالأشكال

الصفحة	الشكل	الباب. رقم الشكل
8	..... بعض مراحل عملية معالجة اللغات الطبيعية	1.2
12	..... نص طبقت عليه خاصية وضع الشروط آلياً	3.2
13	..... هيكل نظام تطبيقي يستخدم اللغات الطبيعية كواجهة	3.2
19	..... مثال لشكل أبي الأسود الدولي	1.3
19	..... علامة الشدة في أول ظهور لها	2.3
20	..... علامة الشدة بعد تعديلها	3.3
25	..... علاقة التشكيل بمراحل معالجة اللغات الطبيعية	4.3
26	..... شاشة اختبرت فيها خدمة تشكيل Google	5.3
28	..... شاشة اختبرت فيها خدمة المصحح الآلي في موقع عجيب	6.3
	..... شاشة إختبار نظام ArabDiac	7.3
29	.....	
	..... شاشة إختبار نظام ArabDiac وتظهر فيها رسالة تفيد بعدم قبول نص	8.3
29	..... يتجاوز العشر كلمات	
34	..... خوارزمية النظام المقترح	1.4
35	..... مخطط الكينونة العلائقي لنظام التشكيل حسب الطلب	2.4

38	واجهة خدمة تشكيل Google بعد إدخال النص وتشكيله .....	3.4
39	النص بعد تشكيله الياً وتظهر الكلمات التي شكلت بصور خاطئة مظلمة..	4.4
39	نافذة فتح ملف نصي مُشكَّل لإختبار النظام عليه .....	5.4
40	محرر نصوص النظام وتظهر الكلمات التي فيها لبس بلون مختلف .....	6.4
41	خيارات التشكيل لكلمة شكَّلت بصورة خاطئة .....	7.4
41	كلمة تركت على تشكيلها الأصلي الصحيح باختيار الخيار الأول .....	8.4

# قائمة بالجدول

الصفحة	الجدول	الباب.رقم الجدول
22	علامات التشكيل العربية، تمثيلها الصوتي العالمي (IPA) وأسمائها....	1.3
	مقارنة بين مخرجات خدمة تشكيل Google قبل وبعد استخدام النظام	1.4
42	المقترح .....	

# الفهرس

الصفحة	الموضوع	الباب. الفصل. الفقرة
ب	آية .....	
ج	إهداء .....	
د	شكر و عرفان .....	
هـ	المستخلص .....	
ز	Abstract .....	
ط	المصطلحات .....	
ك	فهرس الأشكال .....	
م	فهرس الجداول .....	
	المقدمة	الباب الأول
2	مقدمة .....	1.1
3	مشكلة البحث .....	2.1
4	هدف البحث .....	3.1
4	أهمية البحث .....	4.1
4	منهج الحل .....	5.1
5	نطاق البحث .....	6.1

5	ترتيب أبواب البحث.....	7.1
<b>المعالجة الآلية للغات الطبيعية</b>		<b>الباب الثاني</b>
7	مقدمة.....	1.2
7	مستويات تحليل اللغات الطبيعية.....	2.2
8	التحليل الصرفي.....	1.2.2
9	التحليل النحوي/ الإعرابي.....	2.2.2
10	التحليل الدلالي.....	3.2.2
10	تطبيقات ومنتجات المعالجة الآلية للغات الطبيعية.....	2.2
10	تصنيفات أنظمة معالجة اللغات الطبيعية.....	3.2
11	تجهيز/ تحرير النصوص.....	1.3.2.1
	استخدام اللغة الطبيعية كوسيط بين الانسان من جهة	2.1.3.2
13	وقواعد البيانات وأنظمة أخرى من جهة أخرى.....	
14	الترجمة الآلية من لغة طبيعية إلى أخرى.....	3.1.3.2
15	التعرف الضوئي على الحروف.....	4.1.3.2
16	التعرف على الكلام/ فهم الكلام.....	5.1.3.2
<b>التشكيل الآلي</b>		<b>الباب الثالث</b>
18	لمحة تاريخية عن التشكيل.....	1.3
18	الشكل بطريقة النقط.....	1.1.3

20	الشكل بطريقة الحروف الصغيرة.....	2.1.3
22	طرق التشكيل.....	2.3
25	أمثلة لبرمجيات ودراسات في مجال التشكيل.....	3.3
26	خدمة تشكُّيل Google.....	1.3.3
27	برنامج التشكيل الآلي من شركة صخر.....	2.3.3
27	المشكُّل الآلي للنص العربي من شركة RDI.....	3.3.3
30	نظام التشكيل الآلي ( Automatic Arabic Text ) .....(Diacritizer)	4.3.3
	<b>النظام المقترح</b>	<b>الباب الرابع</b>
32	مقدمة.....	1.4
32	الخوارزمية.....	2.4
33	مراحل العمل.....	3.4
35	بناء قاعدة البيانات.....	1.3.4
36	تخزين البيانات.....	2.3.4
36	بناء محرر النصوص.....	3.3.4
36	الصعوبات التقنية.....	4.4
37	وصف النظام.....	5.4
37	إختبار النظام.....	6.4



41	..... النتائج ومناقشتها	7.4
	<b>الخاتمة والتوصيات</b>	<b>الباب الخامس</b>
45	..... الخاتمة	1.5
45	..... التوصيات	2.5
		<b>المصادر والمراجع</b>
48	..... المراجع	
		<b>الملاحق</b>
51	..... شفرة برنامج التشكيل حسب الطلب	ملحق أ.أ
56	..... شفرة برنامج بناء جداول النظام	ملحق أ.ب
57	..... شفرة برنامج لتخزين جذور الكلمات	ملحق أ.ج
58	..... شفرة برنامج لتخزين كل التشكيلات المحتملة للكلمة	ملحق أ.د
59	..... شفرة برنامج تخزين السوابق	ملحق أ.ذ
60	..... شفرة برنامج تخزين اللواحق	ملحق أ.ر
61	..... النص الذي تم استخدامه كدراسة حالة	ملحق ب