كلية دار العلوم

قسم علم اللغة والدراسات السامية والشرقية

مُدَوَّنة مُعجَمٍ عَرَبِيٍّ مُعاصِرٍ

مُعالَجة لُغَوِيَّة حاسوبِيَّة

أطروحة مُقدَّمة للحصول على درجة الماجستير في عِلم اللُغة

إعداد

المُعتَزّ باللّه السَّعيد طه

المعيد بالقسم

1428هـ/ 2007م

إشراف

الأستاذ الدكتور / محمد حسن عبد العزيز

أستاذ عِلم اللُّغة – بجامعة القاهرة

وعضو مجمع اللُّغة العربيَّة بالقاهرة

وإشراف

الأستاذة الدكتورة / سَلوى السَّيِّد حمادة

أستاذ الإلكترونِيَّات والاتِّصالات بالمركز القومِيّ للبُحوث


بسم الله الرحمن الرحيم

شكر وتقدير

يتقدَّم الباحِث بخالِص الشُّكر والتَّقدير إلى:

الأستاذ الدكتور/ مُحَمَّد حسن عبد العزيز

الَّذي تحمل مشقة الإشراف على هذه الدِّراسة، وتَبَنَّى موضُوعَها فِكرةً، فَخُطَّةً، فبحثًا علمِيًّا، و ما لبث يغدق على صاحِبِها من علمه ووقته ومكتبته الزاخرة بكل نفيس وجليل، فكان الأب والمعلم والمربي؛ فله خالِص الشُّكر وعظيم التَّقدير.

والأستاذة الدكتورة/ سلوى السَّيِّد حمادة

الَّتي تحملت مشقة الإشراف على الجانب الحاسُوبِيّ من الدِّراسة، فلم تَضِنَّ على صاحِبِها بوقتٍ أو جهدٍ، أو مراجِعَ وأدواتٍ حاسُوبِيَّةٍ تُيَسِّر عملَه؛ فلها أجزل الشكر وأوفاه.

كما يتقدَّم بالشُّكر العميق إلى أستاذَيه الكريمين:

الأستاذ الدكتور/ مُحَمَّد يونُس الحلاوي

والأستاذ الدكتور/ مُصطفى صلاح قُطب

على تفضلهما بقبول مناقشة هذه الرسالة؛ فقد أنفقا من وقتهما وجهدهما في قراءتها، لتقييمِها وتقويمِها، وإقامةِ مُعوَجِّها، وتهذيب حوشِيِّها؛ فلسيادتِهما جزيل الشُّكر ووافر الامتنان.

الفهرس الموضوعيّ

مُقَدِّمة

الحمدُ لله ربِّ العالَمين، والصَّلاة والسَّلام على مُحَمَّدٍ خاتَم النَّبيِّين، وبعد...

فتشتمِل مُقَدِّمة هذه الدِّراسة على العناصِر التَّالِيَة:

- موضـوع الدِّراسة.
- أهــداف الدِّراسة.
- منهـــج الدِّراسة.
- الدِّراسات السَّابِقة.
- محـــاوِر الدِّراسة.

أوَّلا: موضُوعُ الدِّراسة:

موضُوعُ هذه الدِّراسة (مُدَوَّنة مُعجَمٍ عربِيٍّ مُعاصِر، مُعالجَة لُغَوِيَّة حاسوبِيَّة)، وهو مُحاوَلةٌ لتصميمِ مُدَوَّنةٍ لُغَوِيَّة Corpus Linguistics للعربِيَّة المُعاصِرة، ومُعالجَتِها آلِيًّا للإفادة منها في صَناعة مُعجَمٍ عربِيٍّ مُعاصِرٍ.

ثانِيًا: أهـداف الدِّراسة:

وتهدف الدِّراسة إلى:

1. التَّعريف بمنهجٍ لُغَوِيٍّ، حديثٍ نِسبِيًّا، هو منهج دِراسة المُدَوَّنات اللُّغَـوِيَّة الإلكترونِيَّة Electronic Corpus Linguistics.
2. تصميم مُدَوَّنةٍ لُغَوِيَّةٍ إلكترونِيَّةٍ تعكِس الواقِعَ اللُّغَوِيَّ للعربِيَّة في الوقت الرَّاهن.
3. تقديم نموذَجٍ لمُعجَمٍ عربِيٍّ مُعاصِرٍ، يستَمِدُّ مادَّته بأكملِها مِن مُدَوَّنةٍ لُغَوِيَّة إلكترونِيَّة، ويفي بحاجة أرباب العربِيَّة ومُتَعلِّميها، في مُحاوَلةٍ لسدِّ بعضِ الفجوات المُعجَمِيَّة الَّتي تُعاني منها المعاجِم العربِيَّة المُعاصِرة.

ثالِثًا: منهج الدِّراسة:

صُمِّمَت المُدَوَّنة اللُّغَوِيَّة – موضُوع الدِّراسة – وفقَ المنهج الوَصفِيّ، ويتنَوَّع مجالُ البحث بينَ علم اللُّغة الحاسُوبِيّ Computational Linguistics الَّذي يقوم على إخضاع الآلة للتَّحليل اللُّغوِيّ بِكافَّة مُستَوياتِه باستِخدام أدواتٍ حاسُوبِيَّةٍ مُعَيَّنة، وهندَسة اللُّغة Language Engineering الَّتي تُعنى بتصميم هذه الأدوات وتطويرِها وَفقًا لطبيعة اللُّغة موضُوعِ الدِّراسة. ووُضِعَ النَّموذَج المُعجَمِيّ المُلحَق بالدِّراسة – كذلك – وفقَ المنهج الوَصفِيّ، إذ استَمَدَّ مادَّته من مُدَوَّنة لُغَوِيَّة للعربِيَّة الفُصحى المُعاصِرة، تَضُمُّ نُصوصًا تنتمي إلى الفترة من 1950 م، إلى 2007 م.

رابِعًا: الدِّراسات السَّابِقة:

لم يتَمكَّن الباحِث مِن العُثور على دراساتٍ عربِيَّةٍ عن المُدوَّناتِ اللُّغوِيَّة العربِيَّة، ورُبَّما يرجِع ذلك إلى جِدَّة المنهج نسبِيًّا؛ لكنَّ هناك عددًا من المُدَوَّنات اللُّغَوِيَّة الموضُوعة ضِمنَ دراساتٍ غير عربِيَّةٍ لباحِثين عَرَب، ومن هذه المُدَوَّنات:

1. المُدَوَّنة مُتَعَدِّدة اللُّغات Multilingual Corpus، صُمِّمَت ضِمنَ الأطروحة العِلمِيَّة الَّتي تقدَّمَ بها الباحِث سَتَّار عِزويني إلى جامِعة مانشستر للحُصُول على درجة الدُّكتوراه في (2003 م).
2. المُدَوَّنة العربِيَّة العِلمِيَّة العامَّة General Scientific Arabic Corpus، صُمِّمَت ضِمنَ الأطروحة العِلمِيَّة الَّتي تقدَّمَ بها الباحِث أمين المُهَنَّى إلى جامِعة مانشستر للحُصُول على درجة الدُّكتوراه في (2003 م).
3. مُدَوَّنة العربِيَّة الفُصحى Classical Arabic Corpus، صُمِّمَت ضِمنَ الأطروحة العِلمِيَّة الَّتي تقدَّمَ بها الباحِث عبد الحامِد عليوة إلى جامِعة مانشستر للحُصُول على درجة الدُّكتوراه في (2004 م).

4. مُدَوَّنة العربيَّة المُعاصِرة Contemporary Arabic Corpus، صُمِّمَت ضِمنَ الأطروحة العِلمِيَّة الَّتي تقدَّمَت بها الباحِثة القطرِيَّة لطيفة السُّليطي – أستاذ الدِّراسات اللُّغَوِيَّة بجامِعة قطَر – إلى جامِعة ليدز للحُصُول على درجة الماجستير في العُلوم في (2004 م).
5. مُدَوَّنة تطوير تقنِية مُحاكاة الأصوات Chatbot Corpus، صُمِّمَت ضِمنَ الأطروحة العِلمِيَّة الَّتي تقدَّمَت بها الباحِثة الأردنِيَّة بيان عارف أبو شاور إلى جامِعة ليدز للحُصُول على درجة الدُّكتوراه في (2005 م).

وقد تَعرَّضَ الباحِث لهذه المشروعات الفردِيَّة وغيرِها مِن المشروعات المُنجَزة بواسِطة مؤسَّساتٍ أو هيئاتٍ عِلمِيَّة، بشيءٍ من التَّفصِيل في الباب الأوَّل من هذه الدِّراسة.

خامِسًا: محاوِر الدِّراسة:

وسعيًا وراء بُلوغ الهدَف المنشود، فقد قَسَّمَ الباحِث الدِّراسة إلى بابَين، تعقبهما خاتمة، على النَّحو التَّالي:

- **الباب الأوَّل: الإطار العامّ لِمُدَوَّنة المُعجَم العربِيّ المُعاصِر**. ويشتمِل على ثلاثة فُصُول:
  - الفَصل الأوَّل: مباحِث تمهيدِيَّة. ويناقِش بعضَ القضايا النَّظَرِيَّة الَّتي تتعلَّق بموضوع الدِّراسة، ويَضُمُّ أربعة محاوِرَ أساسِيَّة:
    1. **منهج دِراسة المُدَوَّنة اللُّغَوِيَّة (Corpus Linguistics).**
    2. **إرهاصات المنهج، ومراحِل تطوُّر دراسة المُدَوَّنات اللُّغَوِيَّة.**
       - المرحلة الأولى: ما قبل الحاسب الآلي.
       - المرحلة الثانية: المدوَّنات الإلكترونيَّة في أطوارها الأولى.
       - المرحلة الثالثة: مرحلة المشروعات اللُّغَوِيَّة الكُبْرَى.

3. أنـــواع المُـــــدَوَّنات اللُّغَـــــــوِيَّة الإلكتـــــــرونيَّة.

4. المُـــــدَوَّنات العـــــــــربية المُنْجَـــــزَة إلكتـــــرونيًّا.

- الفصل الثَّاني: الخُطوات المنهجِيَّة لِصَناعة مُدَوَّنة مُعجَمٍ عربِيٍّ مُعاصِر، ويَضُمُّ سِتَّة محاوِرَ أساسِيَّة:

1. تحـديد مـادَّة المُـدَوَّنة اللُّغَـــــوِيَّة وتعيين مَصادِرِها .

2. إدخال نصـــــــوص المُـــــدَوَّنة اللُّغَـــــــوِيَّة.

- الإدخال الآليّ للنُّصـــــــــــــــوص.
- الإدخال اليدويّ للنُّصـــــــــــــــوص.
- النُّصـــــــــوص المُتَاحة إلكتـــــــرونيًّا.

3. المُراجَعـــة الإمـلائِيَّـــة لِنُصُـــــوص المُدَوَّنة اللُّغَـوِيَّة.

4. تنسيق المـــــــادَّة المُدخَلة تمهيـــدًا لِمُعالجَتِهـــا آلِيًّا.

5. ترميز النُّصُـــــــوص المُدخَلة Text Encoding.

6. تصميم قاعِدة بيانات المُدَوَّنة اللُّغَوِيَّة Corpus DB.

وتأتي تَتِمَّة الفَصل لتَضُمَّ نماذِجَ من المُدَوَّنة اللُّغَوِيَّة – موضوع الدِّراسة.

- الفصل الثَّالث: المُعالجَة الآلِيَّة لِنُصُوص المُدَوَّنة اللُّغَوِيَّة، ويَضُمُّ أربعةَ محاوِرَ أساسِيَّة:

1. مفهوم المُعـــــالجَة الآلِيَّـــــة للُّغـــــــات الطَّبيعِيَّة.

2. مفهوم الخوارِزمِيَّة وعلاقتها بالمُعالجَة الآلِيَّة لِلُّغة العربِيَّة.

3. أدوات التَّحليـــل الآليّ لِنُصُـــــوص المُدَوَّنات اللُّغَوِيَّة.

- المُفَهرِس الآليّ للنُّصُـــوص Concordance.
- المُحلِّل الصَّـــــرفيّ العربيّ Arab Morpho.
- مُعَنوِن التَّراكيـــب العربيَّة Arab Tagger.

- المعاجم العربيَّة المُتاحة على الشبكة العنكبوتيَّة.

4. المُعـــالجَة الآلِيَّـــة لِمُدَوَّنة المُعجَـــم المنشـــود.

- **الباب الثَّاني: المُعجَم العربِيّ المُعاصِر (الخطوات التَّنفِيذِيَّة، والنَّموذَج المُعجَمِيّ)**. ويشتَمِل على فَصلَين:
  - الفَصل الأوَّل: الخُطوات التَّنفِيذِيَّة لِصَناعة مُعجَمٍ عربِيٍّ مُعاصِر، ويَضُمُّ سبعةَ محاوِرَ أساسِيَّة:

1. تعييـــن المداخل المُعجَمِيَّة، وتحـــديد الكلمـات الرَّأسِيَّـــة.

2. تعيين معلومات التَّكرار Frequency information.

3. تعيين المعـــاني المُعجَمِيَّـــة لمداخِل النَّمـــوذَج المُعجَمِي.

4. إدراج الأمثِلة التَّوضِيحِيَّـــة لمداخِل النَّمـــوذَج المُعجَمِي.

5. تعيين الوَصـــف النَّحْـــوِيّ بِجانِبَيه البِنَـــوِيّ والتَّركيبيّ.

6. تعيين الجـوانِب البراجمـــاتِيَّة لمداخِل النَّمـــوذَج المُعجَمِي.

7. الكتـــابة الصَّـــوتيَّة لمداخِل النَّمـــوذَج المُعجَمِي.

وتأتي تَتِمَّة الفَصل لتعرِضَ نموذَجًا يوازِن بينَ المُخرَج الحاسُوبِيّ والمُخرَج المُعجَمِيّ.

  - الفَصل الثَّانِي: النَّموذَج المُعجَمِيّ، وقد اشتمل على المداخل المُعجَمِيَّة لباب (الباء) بكامل كلماتِه الرَّأسِيَّة الواردة في نصوص المُدَوَّنة اللغوية موضوع الدِّراسة، وقد وَصَل عدد المداخِل المُعجَمِيَّة للنَّموذَج المُعجَمِيّ إلى مائتين واثنين وخمسين (252) مدخلاً مُعجَمِيًّا، تبدأ بمدخل (ب)، وتنتَهي بِمدخل (بيي)؛ وتشتمل هذه المداخِلُ على سَبعمائة وأحد عشر (711) كلمة رأسِيَّة.

- **الخاتمة**، وتشتمل على أهمّ النَّتائِج والتَّوصيات.

والله من وراء القصد وهو يهدي السبيل...

رموز الكتابة الصَّوتِيَّة المُستَخدَمة

الحرف	الرَّمز الصَّوتِيّ	الحرف	الرَّمز الصَّوتِيّ
أ	ʔ	ع	ʕ
ب	b	غ	ɣ
ت	t	ف	f
ث	ŧ	ق	ķ
ج	ǧ	ك	k
ح	ħ	ل	l
خ	ĥ	م	m
د	d	ن	n
ذ	đ	هـ	h
ر	r	و	w
ز	z	ي	y
س	s	فتحة قصيرة	a
ش	š	فتحة طويلة	ā
ص	ṣ	كسرة قصيرة	e
ض	d’	كسرة طويلة	ē
ط	ţ	ضمَّة قصيرة	u
ظ	ż	ضمَّة طويلة	ū

الاختصارات المُستَخدَمة

الاختصار	المصطلَح
BNC	**British National Corpus**
CAC1	**Contemporary Arabic Corpus**
CAC2	**Classical Arabic Corpus**
COBUILD	**Collins Birmingham University Language Database**
EAC	**English – Arabic Corpus**
ELRA	**European Language Resources Association**
GSAC	**General Scientific Arabic Corpus**
HTML	**Hyper Text Markup Language**
LDC	**Linguistic Data Consortium**
LOB	**Lancaster-Oslo/ Bergen Corpus**
MLC	**Multilingual Corpus**
OCR	**Optical Character Readers**
OED	**Oxford English Dictionary**
SDC	**System Development Corporation**
SSE	**Survey Of Spoken English**
SUE	**Survey of English Usage**
XML	**Extensible Markup Language**
CLARA	**Corpus Linguae Arabicae**
PDF	**Portable Document Format**
TXT	**Text Document**
RTF	**Rich Text Format**
DOC	**Word Document**
ASP	**Active server pages**
PHP	**Personal Home Page**
JSP	**Java Server Pages**

الباب الأول

الإطار العامّ لِمُدَوَّنة المُعجَم العربيّ المُعاصِر

الفصل الأول: مَباحِث تمهِيديَّة

1. منهج دِراسة المُدَوَّنة اللُّغَوِيَّة (Corpus Linguistics).

2. إرهاصات المنهج، ومراحِل تطوُّر دراسة المُدَوَّنات اللُّغَوِيَّة.

- المرحلة الأولى: ما قبل الحاسب الآلي.
- المرحلة الثانية: المدوَّنات الإلكترونيَّة في أطوارها الأولى.
- المرحلة الثالثة: مرحلة المشروعات اللُّغَوِيَّة الكُبْرَى.

3. أنواع المُدَوَّنات اللُّغَوِيَّة الإلكترونيَّة.

4. المُدَوَّنات العربية المُنْجَزَة إلكترونيًّا.

1. منهج دِراسة المُدَوَّنة اللُّغَوِيَّة

Corpus Linguistics Study

يمكن تعريف المُدَوَّنة اللُّغَـــوِيَّة بأنها كتلـة غـير منتظمـة مـن النصـوص المكتوبـة أو المنطوقة التي تُستخدَم لدراسة جوانب اللُّغة، يمكن قراءتها والتعامل معها آلِيًّا بعـد إدخالهـا على الحاسب الآليّ (1)، كما يمكن الـتَّحَكُّمُ في بياناتهـا ومُـدخَلاتها، بالإضـافة أو الحـذف أو التعديل من خلال قواعدَ بيانـاتٍ (Databases) صُـمِّمَت خُصُوصًـا للتعامـل مـع هـذه النصوص. وتُعتَبَرُ قاعدةُ البياناتِ الحاويةُ لنصـوص المُدَوَّنة اللُّغَوِيَّة مخزنًا كبيرًا للُّغة، يُرجَـع إليه وقتَ الحاجـة، ويَتحمَّـل أيَّ قـدرٍ مـن النُّصـــوص التـي تُضـــافُ إلى المـادَّة الأساسـيَّة مُستَقبلاً.

ومادَّة المُدَوَّنة اللُّغَوِيَّة ليست نصوصًا تقييديَّة أو عشوائية؛ إنها كتلـةٌ غـيرُ منتظمـةٍ مـن النُّصوص الَّتي تخضع لمجموعةٍ من الأسُس والمعايير، يُحَـدِّدها الهـدفُ المنشـودُ مـن المُدَوَّنـة اللُّغَوِيَّة؛ فالمُدَوَّنة التـي يُعتَمَـد عليهـا في صَـناعة مُعجَـمٍ لُغَـويٍّ، تختلـف مادَّتُهـا عـن تلـك المُستَخدَمةِ في حَصر مجموعةٍ من الأنماط التركيبيَّةِ أو البِنَوِيَّةِ للُّغـة، كـما تختلِـفُ مـادَّةُ المُدَوَّنـةِ المُستخدَمةِ في صَناعة مُعجَمٍ تكراريّ عن تلك الَّتي يُعتَمَد عليها في صَناعة المَعاجمِ التاريخيَّة. كذلك.. فإنَّ المُعالجَةَ الآليَّةَ للنُّصوص تتَّفِق وطبيعةَ المُدَوَّنة؛ فالبرامجُ الحاسوبيَّة المُستَخدَمة، وطريقةُ مُعالجَةِ النصوص، وطرائقُ إدارةِ قواعدِ البيانات، كـلُّ هـذا يخضَـعُ لتلـك الأسُـس والمعاييرِ الَّتي تُحدِّدها طبيعةُ المُدَوَّنة اللُّغَوِيَّة.

(1) **Hartmann, R. R. K.** and **Stork, F. C.** (1972). Dictionary Of Language and Linguistics. London. P.55, **Kennedy, G.** (1998). An Introduction to Corpus Linguistics. Longman. P.1, and For More Information About Corpus Linguistics Definition, See: **McEnery, T.** and **Wilson, A.** (1997). Corpus Linguistics Edinburgh: Edinburgh University Press. P. 21. **Kübler, S.** (2005). Introduction to Corpus Linguistics.University of Tübingen. P.2. It is Available from: http://www.sfs.uni-tuebingen.de/~kuebler/rocoli/intro_corp_ling.pdf.