



الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة المعلوماتية

قسم الذكاء الصناعي

معالجة اللغات الطبيعية

Text to phonemes in Arabic

Desktop Application

إعداد الطلاب :

آلاء محمد بيرقدار

غرام صايل الحلقي

ارجوان عبد الفتاح طيب

إيمان موسى البلخي

إشراف مهندسي العملي

4/1/2023

تجريد

الكتابة الصوتية هي مجموعة من الرموز الكتابية غايتها تسجيل الأصوات اللغوية تسجيلًا دقيقًا، فالصوت الواحد لا تمثله إلا علامة واحدة والعلامة الواحدة لا يمثلها إلا رمز واحد.

تغير في الفونيم الصوتي يعطي معنى آخر للكلمة ومن هنا دعت الحاجة إلى بناء نظام دقته عالية

تم إجراء دراسة مرجعية للأنظمة الموجودة حاليًا لمعرفة أيهما يعطي دقة أعلى وجدنا أن هنالك مكاتب وأنظمة عدة تعطي نتائج ممتازة للغة الانكليزية بينما للغة العربية كانت النتائج قليلة وانتهت الدراسة المرجعية بالحصول على ثلاث أنظمة

تم بناء عدة حالات اختبار من عدد قواعد قراءة الكلمات في اللغة العربية التي تم ذكرها في الدراسة النظرية وذلك من أجل حساب الدقة لكل نظام وتم استخدام مكتبة لحساب نسبة التشابه بين `two string` phonemes

ومن ثم حساب الدقة لكامل حالات الاختبار .

وتبين أن نظام Toph يعطي الدقة الأفضل بين هذه الأنظمة حيث أعطى دقة 90.003%

وتم الاعتماد على نظام Toph وتعديل بعض القواعد الموجودة ضمنه لتناسب مع الآليات القرآنية لأنه يراعي القراءة المتسلسلة وهي من أهم متطلبات قراءة الآيات القرآنية وكذلك حالات خاصة مثل بعض الكلمات القرآنية وحالات الوقوف على ساكن أو ألف تبعاً لنهاية الكلمة. وبعد التعديل تم الحصول على دقة تقريبا 94% بالإضافة إلى أن هذا النظام مبني بلغة برمجية ++C فتم ربطته مع بايثون وبناء واجهة مناسبة.

الكلمات المفتاحية : تجويد , صوتيمات , نص مشكول , نسبة تشابه , قواعد قراءة الكلمات في اللغة العربية , النطق , السياق الصرفي والتركيب , المقاطع الصوتية , الصوائت والصوامت , الرموز الصوتية , الثنائيات الصوتية

مقدمة:

قسم اللغويين اللغة لمستويات مختلفة لتسهيل دراسة الظواهر اللغوية, منها المستوى النحوي والصرفي والدلالي والصوتي.

فالكتابة الصوتية هي مجموعة من الرموز الكتابية غايتها تسجيل الأصوات اللغوية تسجيلًا دقيقًا, فالصوت الواحد لا تمثله إلا علامة واحدة والعلامة الواحدة لا يمثلها إلا رمز واحد.

يختلف المظهر الصوتي في اللغة عن المظهر الكتابي المعروف لأختلاف طبيعتها, مع الملاحظة أن رمز الصوت لا يحتوي كل الخصائص الصوتية لهذا الصوت .

هدف المشروع

تحويل نص من رموز كتابية إلى رموز صوتية وذلك بغية تمثيل النص العربي كما يقرأ للحصول على النص المقروء

على اعتبار أن دخل البرنامج مُشكّل كلياً (كما في جميع الدراسات التي تجري في إطار تركيب الكلام في اللغة العربية)، وذلك باعتبار أن عملية تشكيل النصوص في العربية لا تتطلب محلاً صرفياً-تركيبياً فقط بل محلاً دلالياً أيضاً، وأن عملية التشكيل يمكن اعتبارها مرحلة مستقلة مستقبلية.

الطريقة التي يتم فيها اللفظ وفق كيفية معينة بلغة ما، ويعد اللبنة الأساسية في بناء المعنى اللغوي، لان اختلاف لفظ الأصوات يؤدي لإحداث تمايز دلالي في المعنى، فكلمتي "صار" و"سار" لكل منهما معنى معين مختلف عن الآخر لاختلاف لفظ الصوتيم (ص) عن (س) فيهما.

تعد الأخطاء الصوتية الأهم والأكثر حساسية فهي تشمل حالات استبدال صوتيم بأخر أو حذفه أو إضافة صوتيم على الكلمة المنطوقة. وعلى نطاق أصغر نجد الأخطاء المتعلقة بنطق الصوتيم نفسه بطريقة مختلفة عن النطق الصحيح وذلك بتغيير أحد سماته الصوتية .

الأصوات في اللغة العربية القياسية المنطوقة

مجموعة وحدات صوتية تسمى الصوتيمات **phonemes**، حيث يشير مصطلح الصوتيم في علم الصوت الى أصغر وحدة صوتية في اللغة تميز كلمة عن كلمة أخرى. وتختلف الصفات المميزة للصوتيم من لغة الى أخرى على الرغم من اشتغال كل منها على صوتيمات متماثلة، يتولد عن اجتماع الصوتيمات المقاطع الصوتية **Syllables**.

تسمى هذه الرموز الأبجدية الصوتية العالمية (International Phonetic Alphabet) أو اختصاراً IPA

إن اللغة العربية هي لغة **الصوامت consonants**، وفيها ما لا يقل عن ثمانية وعشرين صامتاً وهو عدد حروف الأبجدية بأخذ الهمزة مكان الألف، والواو والياء غير المديتين . يُضاف إلى هذه الصوامت اللام والراء المفخمتين مما يشكل ثلاثين صامتاً.

أما **الصوائت vowels** (الحركات والأحرف المدية)، فيصنفها اللغويون في ستة صوائت عادية (الفتحة والضمة والكسرة وهي صوائت قصيرة، و الألف المدية والواو المدية والياء المدية وهي صوائت طويلة موافقة للسابقة على التوالي)، إضافة إلى ستة صوائت موافقة مفخمة **emphatics** وهي الحالات التي يأتي فيها الصائت بعد حرف مفخم. وتُعرّف الصوائت بأنها عناصر صوتية تميز بعدد من الخصائص (الميزات) الواضحة التي تميزها عن الصوائت الأخرى في اللغة.

أما الصوامت المفخمة فهي التالية:

- اللام في لفظ الجلالة
- الحروف (ص، ض، ط، ظ)، يُضاف إليها الحروف (خ، ع، ق) في تلاوة القرآن فقط.
- الراء إذا أتت في إحدى الحالات التالية:
 - إذا تلاها الفتحة أو الضمة أو الألف المدية أو الواو المدية.
 - إذا سبقها كسرة وتلاها أحد الحروف (ص، ض، ط، ظ).

- إذا سبقها فتحة أو ضمة أو ألف مدية أو واو مدية وتلاها صمت أو وقف.
- والجدير بالذكر، أنَّ هناك صائتين إضافيين هما الواو المُمالَة و الياء المُمالَة ، وتُمال الضمة والكسرة في الحالات التالية:
- إذا تلاهما حرف ساكن كما في أفعال الأمر مثل: (قُمْ، انتصر، قِف).
- في لفظ تنوين الضم أو تنوين الكسر إذا لم يأت بعد التنوين حرف ساكن يجبرنا على تحريك نون التنوين الملفوظة بالكسر وحذف الإمالة. وهناك أيضاً الواو المُمالَة المفخمة والياء المُمالَة المفخمة. وبهذا يصبح العدد الكلي للصوائت المعتمدة عشرة.

قواعد القراءة في اللغة العربية [البواب، ميرعلم، والطيان 84]

قواعد قراءة الكلمات في اللغة العربية

تُقرأ الكلمات في اللغة العربية غالباً كما تُكتب، إلا أن هناك بعض الحالات التي يختلف فيها المنطوق عن المكتوب. وفيما يلي القواعد التي تحدد النطق [البواب، ميرعلم، والطيان 84].

اللام القمرية واللام الشمسية

تظهر لام (ال) التعريف إذا تليت بأحد الحروف الأربعة عشر المجموعة في { ابغ حك وخف عقيمه } وتسمى اللام القمرية (ومثالها القمر).
وتدغم هذه اللام بالحرف الذي يليها إذا تليت بأحد الحروف الأربعة عشر التالية { ط، ث، ص، ر، ت، ض، ذ، ن، د، س، ظ، ز، ش، ل } وتسمى اللام الشمسية (ومثالها الشمس).

التفخيم والترقيق

التفخيم هو تعظيم الحرف في النطق حتى يمتلئ الفم بصداه. وتُفخَّم الحركات والراء واللام في مواضع معينة وضوابط نوجزها فيما يلي:

1. تفخيم الحركات وحروف المد

تُفخَّم الحركات وحروف المد التي تلي حروف التفخيم (التي تُسمَّى أيضاً حروف الاستعلاء) والمجموعة في عبارة { خُصَّ ضَغَطٌ قِطْ }، كما تُفخَّم الحركات وحروف المد بعد الراء واللام المفخمتين، ونظراً لأن اللام المفخمة لا تَرِدُ إلا في لفظ الجلالة فتُفخَّم الألف المدية بعد اللام المفخمة في كلمة الله.

2. تفخيم الراء

تُفخَّم الراء إذا كانت مفتوحة أو مضمومة، أو كانت ساكنة ومسبوقة بضم أو فتح، أو إذا كانت ساكنة وسُبقَت بكسر وتُبعت بحرف استعلاء غير مكسور.
وترُقَّق فيما عدا ذلك.

3. تفخيم اللام

لا تُفخَّم اللام إلا في لفظ الجلالة إذا سُبِقَ بفتح أو ضم. أما إذا سُبِقَ بكسر فترقق اللام فيه، وكذا في جميع حالاتها الأخرى.

همزة الوصل

همزة الوصل همزةٌ يُتوصَّلُ بها إلى النطق بالحرف الساكن، وتظهر بالنطق في أول الكلام وتختفي في درجته.

ومواضعها تكون في ماضي الخماسي والسداسي، وفي بعض الأسماء (مثل ابن وابنة واسم..)، وفي (أل) التعريف.

حروف المد

وهي ثلاثة الألف والواو والياء.

1. إسقاط حرف المد

تُحذف حروف المد الساكنة إذا تبعها حرف ساكن (ومثاله: واستغفر تصبح وَسْتَغْفِرُ)، كما يسقط حرف المد من آخر الكلمة إذا وليه حرف ساكن في أول الكلمة التالية له إلا إذا وَقَفَ عليه (ومثاله: في البيت تصبح فليبت)، كما يسقط إذا وليه تنوين (مثل فتى).

2. حروف مد تُكتب ولا تُقرأ

هناك طائفة من الكلمات يكتب فيها حرف المد ولا يُقرأ، وهي: عمرو، أولئك، أولو، مائة، وألف التفريق التي تلي واو الجماعة مثل: جلسوا.

3. حروف مد تُقرأ ولا تُكتب

هناك طائفة من الكلمات يُقرأ فيها حرف المد دون أن يكتب، وهي: الله، اللهم، لكن، هذا (ومشتقاتها)، الرحمن، السموات، إله، أولئك.

تاء التأنيث (أو التاء المربوطة)

تُقرأ هاء إذا وَقَفَ عليها وتاء إذا وُصِلَتْ بما بعدها.

التقاء الساكنين

كل حرف ساكن صحيح في آخر الكلمة يحرك بالكسر إذا تلاه ساكن آخر (مثل: خذ الكتاب تصبح خُذْكِتاب). وينطبق ذلك على التنوين (مثل: عميقُ استمر تصبح عميقُنِستمر)، ويستثنى من ذلك موضعان:

■ إذا كانت الكلمة الأولى (مِنْ) والكلمة الثانية تبدأ بـ (ال) فإن الساكن الأول يُحرَكُ حينئذ بالفتح (مثل: مِنْ المَكْتَبِ تصبح مِنْمَلِكْتَب).

■ إذا كانت الكلمة الأولى منتهية بميم الجمع والثانية تبدأ بـ (ال)، فإن الساكن الأول يُحرَكُ حينئذ بالضم (مثل: لَهُمُ الغُرْفَةُ تصبح لَهُمُلْغُرْفَةُ).

الوقف

لا يُوقف في العربية إلا على ساكن مهما كانت حركة الموقوف عليه، ويستثنى من ذلك حالة التنوين

المنصوب فيوقف عليها بالألف، أما إذا كان التنوين المنصوب على الحرف مباشرة دون ألف فالوقف يكون بالسكون.

التنوين

هو نون ساكنة تلي ضمة أو فتحة أو كسرة (مثل: قلمٌ أحمر تصبح قلمُنُ أحمر).
ويوقف على التنوين بالسكون عدا تنوين الفتح فيوقف عليه بألف ساكنة.

تشديد الحروف

يوضع رمز الشدة على الحروف في اللغة العربية للدلالة على وجوب تشديدها (أي لفظها مرتين متتاليتين).

قراءة الكلمات على نحو مستمر

تُقرأ الكلمات على نحو مستمر، ونقف في الكلام عند نهاية جملة، أو شبه جملة. فلا نلفظ مثلاً جملة "وجدته في البيت" كل كلمة على حدة، بل نصل الكلمات إلى نهاية الجملة.

مرحلة تحويل الرموز الكتابية إلى رموز صوتية :

يجري في هذه المرحلة تحويل الرموز الكتابية إلى رموز صوتية وفق قواعد تخزين في الحاسب، ويتم فيها تحويل النص من شكله الإملائي أو المكتوب إلى مكافئة الصوتي أو المنطوق وذلك باستخدام خوارزميات متعلقة باللغة.

على اعتبار أن دخل البرنامج مُشكّل كلياً (كما في جميع الدراسات التي تجري في إطار تركيب الكلام في اللغة العربية)، وذلك باعتبار أن عملية تشكيل النصوص في العربية لا تتطلب محلاً صرفياً-تركيبياً فقط بل محلاً دلالياً أيضاً، وأن عملية التشكيل يمكن اعتبارها مرحلة مستقلة مستقبلية.

الدراسة المرجعية

سنتحدث عن أهم أنظمة تحويل النص لـ فونيمات صوتية الموجودة حالياً، ونشرح طبيعة كل منها والآلية التي تعمل بها

إن دراسة هذه الأنظمة يساعد في جمع متطلبات المشروع من خلال معرفة نقاط القوة والضعف الموجودة في كل نظام، والاطلاع على التقنيات و الخوارزميات المستخدمة ضمن النظام ،لنتمكن من الاستفادة منها للحصول على نظام متكامل وجيد.

الاعمال المشابهة

تصميم نظام TOPH

يقوم هذا النظام بتحويل النصوص العربية من رموز كتابية إلى رموز صوتية، مصاغاً بلغة TOPH على شكل مجموعة من القواعد الحتمية determinist.

وذلك بإعادة بناء النظام TOPH (Orthographic-Phonetic Transcription)، الذي يقوم بإجراء التحويل من رموز كتابية إلى صوتية للغة الفرنسية [Ghneim, Aubergé 85] ، من أجل اللغة العربية.

بنیان النظام TOPH [Ghneim, Auberge 85]

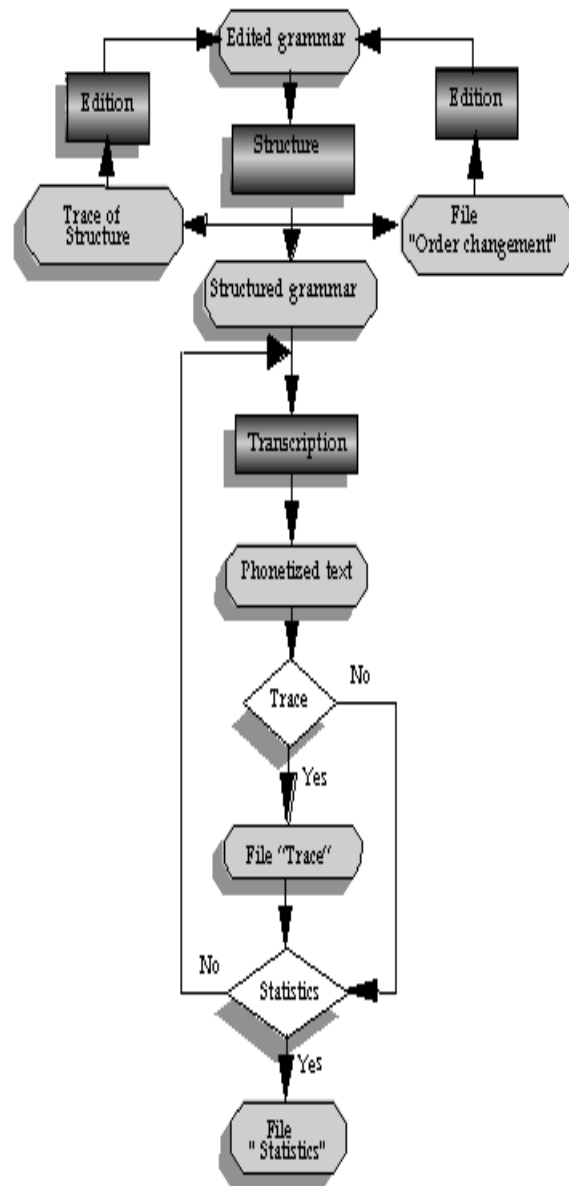
يتألف النموذج القواعدي grammar الخاص بالنظام TOPH من جزأين، يتضمن الجزء الأول عدة مجموعات؛ حيث يجري في هذه المجموعات تجميع العناصر التي لها نفس المهمة كمجموعة الفواصل بين الكلمات، أو التي لها نفس الصفة كالأرقام والحروف وهكذا، في حين يتضمن الجزء الثاني مجموعة القواعد rules، التي يكتبها الخبير اللغوي ضمن ملف نصي (باستخدام محرر نصوص)، يعتمد بناء هذه القواعد على النموذج التركيبي للقواعد ذات السياق grammar contextual ضمن تصنيف Chomsky. فكان الشكل العام لقواعد TOPH كالتالي:

$$[phon_Ch] = \{Ctg_Set\} (SCtxt) + graph_Ch + (PCtxt)$$

حيث PCtxt و SCtxt هما السياق السابق واللاحق للسلسلة الكتابية والضروريان لتعيين القاعدة نسبة لما يتبعها من قواعد. graph_Ch هي سلسلة الرموز الكتابية التي يجري تحويلها. Ctg_Set هي صنف أو مجموعة من الأصناف (تذكير، تأنيث،...) التي يجب أن تطبق معها القاعدة (لا توضع إلا في حال كون القاعدة تعالج إبهاماً متعلقاً بالصنف) و phon_Ch هي سلسلة الرموز الصوتية الموافقة للسلسلة الكتابية.

تتألف البرمجية TOPH من كتلتين: الأولى كتلة البناء structure وتقوم بالاعتماد على القواعد التي كتبها الخبير اللغوي بإجراء التحليل الصرفي (lexical) والتركيبي (syntactic) للقواعد، وإعطاء رسائل الأخطاء (الصرفية والتركيبية)، إن وُجِدَت. وتولد ثلاث ملفات تُستخدم كدخل للكتلة الثانية transcription التي تقوم بترجمة نص ما إلى رموز صوتية. الهدف من هذا التقسيم إلى كتلتين هو فصل عملية إنشاء ملف قواعد التصويت للغة المدروسة، والتي تتطلب عمليات ترجمة للملف

وتعديل فيه عند وجود أخطاء، عن عملية التصويت نفسها. يبين الشكل (1) ببيان النظام TOPH.



ويبين الجدول التالي مجموعة الأحرف المقابلة للأصوات في اللغة العربية التي اعتمدها النظام والترميز الموافق لها:

رمز الأسكي ASCII code	الحرف الأجنبي (Phoneme)	الحرف العربي (Grapheme)
98	b	ب
116	t	ت
120	x	ث
106	j	ج
72	H	ح
88	X	خ
100	d	د
118	v	ذ
114	r	ر
122	z	ز
115	s	س
74	J	ش
83	S	ص
68	D	ض
84	T	ط
90	Z	ظ
67	C	ع
71	G	غ
102	f	ف
113	q	ق
107	k	ك
108	l	ل
109	m	م

110	n	ن
104	h	هـ
99	c	ء
76	l	ل (مفخمة)
82	R	ر (مفخمة)
65	A	أ (مفخمة)
85	U	و (مفخمة)
73	l	و (مفخمة)
101	e	ي (مُثَالَة)
111	o	ي (مُثَالَة)
119	w	و (غير مدية)
121	y	ي (غير مدية)
97	a	أ
117	u	و
105	i	ي

قواعد التصويت في اللغة العربية مكتوبة بلغة TOPH

وقد اعتمد النظام في كتابة هذه القواعد على الدراسة المفصلة [البواب، ميرعلم، والطيان 84]، وعلى دراسة للأصوات في اللغة العربية القياسية المنطوقة [Dakkak, Ghneim 99]. نتج عن ذلك حوالي (180) قاعدة، شملت (حذف اللام الشمسية، التفخيم والترقيق، همزة الوصل، حروف المد، تاء التأنيث أو التاء المربوطة، التقاء الساكنين، الوقف، التنوين، تشديد الحروف، قراءة الكلمات على نحو مستمر، القواعد العامة للفظ الحروف)

تجريب النظام

ولمعرفة مدى إمكانية النظام في تمثيل الفونيمات للكلمة بشكل صحيح يجب تجربتها على قواعد قراءة الكلمات في اللغة العربية التي تم ذكرها سابقاً.

نظام TOPH

حالة الإختبار	الجملة	خرج النظام	الحالة
اللام الشمسية واللام القمرية	الشَّمْسُ والقَمَرُ	waalqqmaR cJJamsu	راعى اللام الشمسية والقمرية
تفخيم الحركات و حروف المد	طَارَ وَ سَارَ	saaRA wa TaaRA	راعى الحالات المفخمة لأحرف المد والحركات
تفخيم الحركات و حروف المد	طُورَ وَ سُورَ	suur Wa Tuur	راعى الحالات المفخمة لأحرف المد والحركات
تفخيم الحركات و حروف المد	قِيسِي	qiisii	راعى الحالات المفخمة لأحرف المد والحركات
تفخيم الراء	القَمَرُ وَ القمرِ	lqamari wa clqamaRU	راعى حالات تفخيم الراء
تفخيم اللام	مِنَ الله مِنْ الله	llah mina llah mani	لم يراع حالات تفخيم اللام
همزة الوصل	استِعارَةٌ وَ استِعارَةٌ	wa aastiCaaRAtu aastiCaaRAh	لم يراع حذف همزة الوصل في درج الكلام
إسقاط حروف المد	واستَغْفِرَ	waastaGfaR	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
إسقاط حروف المد	وَ استَغْفِرَ	aastaGfaR u	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
إسقاط حروف المد	فِي الْبَيْتِ فَتَى	fata lbayti fi	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
حالات تكتب حروف المد فيها ولا تُقرأ	عَمَرُو	CamRu	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تكتب حروف المد فيها ولا تُقرأ	أُولَئِكَ	cuulaacik	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تقرأ حروف المد فيها ولا تكتب	اللَّهُم	Cllahmu	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تقرأ حروف المد فيها ولا تكتب	لَكِنْ	Laaken	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تقرأ حروف المد فيها ولا تكتب	الرَّحْمَنُ وَ السَّمَوَاتِ	wa aalrrHmanu ssamaawaat	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تقرأ حروف المد فيها ولا تكتب	هَذَانِ	haavaan	راعى بعض الكلمات و فشل في كلمات أخرى
تاء التانيث	قِمَّةُ الْجَبَلِ	ljabal Qimmatu	راعى النظام حالات التاء المربوطة
التقاء ساكنين	مِنَ الْمَكْتَبِ	lmaktabi mina	راعى النظام تغيير الحركة عند التقاء ساكنين

التقاء ساكنين	لَهُمُ الْغُرْفَةُ	lGuRfah Lahumu	راعى النظام تغيير الحركة عند التقاء ساكنين
الوقف	نَشِيطاً	naJiTaa	راعى الوقف على ألف في حال تنوين النصب و على سكون في حال الحركات
قراءة الكلمات على نحو مستمر	وَجَدْتَهُ فِي الْبَيْتِ	lbayt wəjadtuHu fi	راعى النظام القراءة المستمرة
قراءة الأعداد	..1,2	cixnaʔin wəHid	راعى النظام قراءة الأرقام

إن هذا النظام راعى اغلب الحالات ولكنه مكتوبة بلغات برمجية غير البايثون ولتحسين النظام نحتاج الى بناء قاموس بالكلمات الشاذة ونعطيها ترميزها المناسب او كتابة قواعد خاصة بها .

اللغات التي تدعمها هذه المكتبة مع قواعد هذه اللغة: الأمريكية _ الألمانية _ الهولندية _ الانكليزية _ التركية _ أوزبكي

اللغات التي تدعمها بشكل محدد (بسبب غموض قواعد هذه اللغات): العربية - الكاتالونية - سوراني - البرتغالية - الفرنسية - الفارسية

تمثل الأحرف العربية بالشكل:

IPA	Latin	Name	Final	Medial	Initial	Isolated	IPA	Latin	Name	Final	Medial	Initial	Isolated
[tʰ]	t	tʰo:	طاء	ط	ط	ط	[ʔ]	()	'ʔaləf	ألف	ا	—	—
[z; dʰ]	z	zʰo:	ظاء	ظ	ظ	ظ	[b]	b	be:	باء	ب	ب	ب
[ʕ]		ʕe:m	عين	ع	ع	ع	[t]	t	te:	تاء	ت	ت	ت
[ɣ~ɣ]	g, gh	ɣe:m	غين	غ	غ	غ	[t; s]	t, th	se:	ثاء	ث	ث	ث
[f]	f	fe:	فاء	ف	ف	ف	[ʒ; ʒ̄]	ʒ	ʒim	جيم	ج	ج	ج
[ʔ, q; g]	q	ʔo:f	قاف	ق	ق	ق	[h]	h	ha:	حاء	ح	ح	ح
[k]	k	ka:f	كاف	ك	ك	ك	[x]	h, kh	xa:	خاء	خ	خ	خ
[l, l̥]	l	lam	لام	ل	ل	ل	[d]	d	da:l	دال	د	—	—
[m]	m	mim	ميم	م	م	م	[d; z]	d, dh	za:l	ذال	ذ	—	—
[n]	n	num	نون	ن	ن	ن	[r]	r	re:	راء	ر	—	—
[h]	h	he:	هاء	ه	ه	ه	[z]	z	ze:	زاي	ز	—	—
[w]	w	wa:w	واو	و	—	و	[s]	s	sim	سين	س	س	س
[j]	y	je:	ياء	ي	ي	ي	[ʃ]	ʃ, sh	ʃim	شين	ش	ش	ش
		'hamze	همزة	ء	ء	ء	[sʰ]	sʰ	sʰo:d	صاد	ص	ص	ص
							[dʰ]	dʰ	dʰo:d	ضاد	ض	ض	ض

تمثل الحروف المتحركة بالشكل :

IPA	Latin	Name	Final	Medial	Initial	Isolated	IPA	Latin	Name	Final	Medial	Initial	Isolated
[o; ɔː]	aw/ō		—	—	—	و	[ʔ]	()	'ʔaləf	ا	—	—	ا
[aʊ; ɔː]	aw/au		—	—	—	و	[a; ɔː]	ā		ا	—	—	ا
[uːw]	uww / ūw	wa:w 'jad:e	—	—	—	و	[ʔa; ʔaː]	ā-	'ʔaləf mad:e	آ	—	—	آ
[i; iː]	ī		—	—	—	ي	[ʔi, ʔiː]	-ā'	'ʔaləf 'hamze	اء	—	—	اء
[e; ɛː]	ay/ē		—	—	—	ي	[el~el]	al/el	la:m 'ʔaləf	ا	ا	—	ا
[ai; aɪ]	ay/ai		—	—	—	ي	[hu]	hu		ه	—	—	ه
[ij]	iy/yi	je: 'jad:e	—	—	—	ي	[-e; -eː]	-at -e -et -et	te: mar'bu:t'e	ة	—	—	ة
[e]	-ay/ē	moʔ's'ure	ي	—	—	ي	[u; uː~oː]	ū		و	—	—	و

تجريب المكتبة

أن مطابقة الأحرف مع الفونيمات الصوتية على حسب جدول الفونيمات لا تفني بالغرض لأن هنالك العديد من القواعد يجب معالجتها

Epitran			
حالة الاختبار	الجملة	خرج النظام	التمثيل الصحيح
الام الشمسية واللام القمرية	القَمَرُ وَ الشَّمْسُ	alqmr̥ uː aːlʃms̥	alqmr̥uːaːlʃms̥
تفخيم الحركات و حروف المد	طَارَ وَ سَارَ	t̥aːr̥ uː s̥aːr̥	t̥aːruːs̥aːr̥
تفخيم الحركات و حروف المد	طُورَ وَ سُورَ	t̥uːr̥ uː s̥uːr̥	t̥uːruːsuːr̥
تفخيم الحركات و حروف المد	طِيرِي وَ سِيرِي	t̥iːr̥iː uː siːr̥iː	t̥iːriːuːsiːriː
تفخيم الراء	القَمَرُ وَ القَمَرُ	alqmr̥ uː aːlqmr̥	alqmr̥uːalqmr̥
تفخيم اللام	مِنَ اللّهِ وَ مِنَ اللّهِ	m̥n̥ aːll̥h̥ uː m̥n̥ aːll̥h̥	m̥n̥aːll̥huːm̥n̥aːll̥h̥
همزة الوصل	اسْتِعَارَةٌ وَ اسْتِعَارَةٌ	ast̥ʔaːr̥o uː aːst̥ʔaːr̥o	ast̥ʔaːrauːast̥ʔaːra
إسقاط حروف المد	وَ اسْتَغْفِرْ	uː aːst̥ɣfr̥	uːaːst̥ɣfr̥
إسقاط حروف المد	وَاسْتَغْفِرْ	uː aːst̥ɣfr̥	uːaːst̥ɣfr̥
إسقاط حروف المد	فِي الْبَيْتِ	fiː aːlb̥iːt̥	fiːalbiːt̥
إسقاط حروف المد	فَتَى	ft̥i	fteː
حالات تكتب حروف المد فيها ولا تُقرأ	عَمَرُو	ʔmr̥uː	ʔmr̥uː
حالات تكتب حروف المد فيها ولا تُقرأ	أُولَئِكَ	auːliːk̥	aːuːliːk̥
حروف مد تكتب ولا تُقرأ	أُولُو	auːluː	aːuːluː
حروف مد تكتب ولا تُقرأ	مَائَةٌ	m̥aːiːq̥	m̥aːiːa
حروف مد تُقرأ ولا تكتب	اللَّهُ	all̥h̥	all̥h̥
حالات تُقرأ حروف المد فيها ولا تكتب	اللَّهُمَّ	all̥h̥m̥	all̥hm̥
حالات تُقرأ حروف المد فيها ولا تكتب	لَكِنْ	l̥kn̥	lkn̥
حالات تُقرأ حروف المد فيها ولا تكتب	الرَّحْمَنُ وَ السَّمَوَاتُ	alr̥hm̥n̥ uː aːls̥m̥uːaːt̥	alhman̥uːals̥muːaːt̥
حروف مد تُقرأ ولا تكتب	إِلَهُ	al̥h̥	alh̥
حروف مد تُقرأ ولا تكتب	هَذِهِ	h̥ðh̥	hðh̥
حالات تُقرأ حروف المد فيها ولا تكتب	هَذَانِ	h̥ðaːn̥	hðaːn̥
تاء التأنيث	قِمَّةُ الْجَبَلِ	qm̥o̥ aːld̥ʒbl̥	qmetald̥ʒbl̥

mnałmktb	mna:lmktb	مِنَ الْمَكْتَبِ	التقاء ساكنين
lhmalyrfa	lhm a:lyrfö	لَهُمُ الْغُرْفَةُ	التقاء ساكنين
nfi:ta	nfi:t'a:	نَشِيطاً	الوقف
u:dʒdthfiabli:t	u:dʒdth fi: a:lbi:t	وَجَدْتُهُ فِي الْبَيْتِ	قراءة الكلمات على نحو مستمر

بعد التجريب تبين أن المكتبة لا تقوم بالتفريق بين اللام القمرية واللام الشمسية وايضا لا يتم التفريق بين الأحرف المرققة والمفخمة ولا يقوم بتمثيل أي من الحركات وبالتالي فهو لا يراعي الوقوف على التنوين أو الحركات ولا يراعي الحروف المشددة ولم يراعي إلقاء ساكنين او تاء التأنيث بالاضافة الى تمثيل غير منطقي لبعض الكلمات ومن هذا كله نتوصل إلى أن هذه المكتبة لا يمكننا الاستناد عليها.

من خلال عملية البحث توصلنا إلى عدة أبحاث لدكتور نوار الحلبي الذي عمل على بناء عدة أنظمة منها تحويل النص العربي إلى فونيمات صوتية للاستخدام في توليد مركب كلام باللغة العربية واعتمد على مجموعة معطيات النطق بالعربية Arabic speech Corpus وهي مجموعة معطيات صوتية تم بناءها كجزء من أطروحة الدكتوراة

حيث قام الباحثون ببناء هذه المجموعة عن طريق تجميع النص من موقع الجزيرة لتعليم اللغة العربية إضافة لنصوص ليس لها معنى تم توليدها ألياً للحصول على تشكيلات مختلفة للشائيات الصوتية diphones, بعد ذلك تم توليد التمثيل الصوتي phonetic Transcript ألياً , يحوي هذا التمثيل 82 صوتياً (الجدول 1) (بالنسبة للصوامت المضعفة لم يتم ذكرها في الجدول اختصاراً لكن تم تمثيلها برمز مضاعف مماثل لرمز الصوتيم بدون تضعيف مثل الباء المشددة (بّ) تم تمثيلها BB). بعد التمثيل جرى تقطيع النص بداية على مستوى الكلمة ثم على مستوى الصوتيم وتذييلها بوسم صوتي محدد (اسم الصوتيم)

الجدول 1 مجموعة صوتيات مجموعة معطيات النطق العربي

il	[ʔ]	u0	ʔ	y	ي	g	غ	r	ر	<	أ
uu1	[و]	i0	َ	v	ف	f	ف	z	ز	B	ب
ii1	[ي]	AA	ʾ	p	پ	Q	ق	s	س	T	ت
UI	([ʔ])	UU0	و	G	ج	k	ك	\$	ش	^	ث
II	([ʔ])	II0	ي	J	ج (d̥ʒ)	l	ل	S	ص	J	ج (3)
UU1	([و])	A	(ّ)	aa	ا	m	م	D	ض	H	ح
III	([ي])	U0	(ّ)	uu0	و	n	ن	T	ط	X	خ
sil	pause	I0	(ّ)	ii0	ي	h	ه	Z	ظ	D	د
Dist	distortion	u1	[ʔ]	a	ا	w	و	E	ع	*	ذ

في الجدول 1 يشير العمود الأيمن في كل قسم للصوتيم بالخط العربي, بينما يشير العمود الأيسر لترميز Buckwalter (تمثيل يستخدم للمقابلة بين الحروف العربية والرموز اللاتينية) (تم استخدام النسخة المعدلة منه). تشير الرموز التي تحتها خط إلى الحروف التي يتم نطقها في الكلمات الأجنبية مثل فيديو, أما الرموز المائلة فتشير إلى الصوائت , وتشير باقي الرموز إلى الصوامت .

وتشير الحروف ضمن قوسين "()" لحروف جوازها مفخم , بينما تشير الحروف ضمن قوسين "[]" لحروف ممالة عند اللفظ

عند العمل على تصنيف الصوتيات السابقة جرى تقليص الصوتيات في مجموعة تضم 38 صوتياً وذلك بعد دمج الصوت والصوت المشدد منه، ودمج الحركة وحرف المد التابع لها (مثل الفتحة والألف)، إضافة لاعتبار الاصوات الاجنبية مثل (P,v,G,) مع التشويش (dist) صوتياً واحداً. يوضح الجدول 2 كيفية تقليص الصوتيات. يمثل العمود الأيمن الصوتيم المقابل بعد التقليص (الصمت لم يتم ذكره باعتبار لا مقابل له) يوضح الجدول 3 الصوتيات بعد التقليص.

الجدول 2 التقابل بين 82 إلى 83 صوتيم في مجموعة صوتيات معطيات النطق بالعربية

الصوتيم	الصوتيات المقابلة	الصوتيم	الصوتيات المقابلة
<	<, <<, AH, Ah	g	g, gg
b	b, bb	f	f, ff
t	t, tt	q	q, qq
^	^, ^^	k	k, kk
j	j, jj	l	l, ll
H	H, HH	m	m, mm
x	x, xx	n	n, nn
d	d, dd	h	h, hh
*	*, **	w	w, ww
r	r, rr	y	y, yy
z	z, zz	aa	a, a', aa, aa'
s	s, ss	AA	A, A', AA, AA'
\$	\$, \$\$	uu0	u0, u0', uu0, uu0', u
S	S, SS	UU0	U0, U0', UU0, UU0'
D	D, DD	uu1	u1, u1', uu1, uu1', UU1', U1
T	T, TT	ii0	i0, i0', ii0, ii0'
Z	Z, ZZ	II0	I0, I0', II0, II0'
E	E, EE	ii1	i1, i1', I1, I1', ii1, ii1'
		dist	p, pp, v, j, G, dist, -

الشكل 3 صوتيات مجموعة معطيات النطق بالعربية بعد التقليص إلى 38 صوتيم

الصوتيم	الترميز	الصوتيم	الترميز	الصوتيم	الترميز
ء	>	ض	D	ألف مدية	a
ب	b	ط	T	ألف مدية جوارها مقخم	A
ت	t	ظ	Z	واو مدية	u0
ث	^	ع	E	واو مدية جوارها مقخم	U0
ج	j	غ	g	واو مدية ممالة	u1
ح	H	ف	f	ياء مدية	i0
خ	x	ق	q	ياء مدية جوارها مقخم	I0
د	d	ك	k	ياء مدية ممالة	i1
ذ	*	ل	l	تشويش	dist
ر	r	م	m	صمت	sil
ز	z	ن	n		
س	s	هـ	h		
ش	\$	و	w		
ص	S	ي	y		

Arabic Phonetiser			
حالة الإختبار	الجملة	خرج النظام	الحالة
الام الشمسية واللام القمرية	القَمَرُ وَ الشَّمْسُ	<alqAmarwa\$\$amsu0	راعى اللام الشمسية والقمرية
تفخيم الحركات و حروف المد	طَارَ وَ سَارَ	TAArawasaaara	راعى الحالات المفخمة لأحرف المد والحركات
تفخيم الحركات و حروف المد	طُورَ وَ سُورَ	TUUOrwasuuOr	راعى الحالات المفخمة لأحرف المد والحركات
تفخيم الحركات و حروف المد	طِيرِي وَ سِيرِي	TIIOriiOwasiiOriiO	راعى الحالات المفخمة لأحرف المد والحركات
تفخيم الراء	القَمَرُ وَ القَمَرُ	<alqAmanru0wanlqAnmanr	راعى حالات تفخيم الراء
تفخيم اللام	مِنَ اللّهِ وَ مِنَ اللّهِ	mi0nallahi0wamani0llahu0	لم يراع حالات تفخيم اللام
همزة الوصل	اسْتِعَارَةٌ وَ اسْتِعَارَةٌ	sti0Eaaratawasti0Eaarati0	لم يراع حذف همزة الوصل في درج الكلام
إسقاط حروف المد	وَ اسْتَغْفِرْ	wastagfi0ri0	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
إسقاط حروف المد	وَاسْتَغْفِرْ	waastagfi1r	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
إسقاط حروف المد	فِي الْبَيْتِ	fii0lbayti0	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
إسقاط حروف المد	فَتَى	fataa'an	لم يسقط النظام حروف المد في بعض حالات اسقاطها و نجح في حالات أخرى
حالات تكتب حروف المد فيها ولا تقرأ	عَمَرُو	Eamruu0	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تكتب حروف المد فيها ولا تقرأ	أُولَئِكَ	uu0la<i0ka	راعى بعض الكلمات و فشل في كلمات أخرى
حروف مد تكتب ولا تقرأ	أُولُو	<uu0luu0	
حروف مد تكتب ولا تقرأ	مَائَةٌ	maa<a	
حروف مد تقرأ ولا تكتب	اللَّهُ	<allluhu0	
حالات تقرأ حروف المد فيها ولا تكتب	اللَّهُمَّ	<allluhu0mmau0	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تقرأ حروف المد فيها ولا تكتب	لَكِنْ	laaki0nna	راعى بعض الكلمات و فشل في كلمات أخرى
حالات تقرأ حروف المد فيها ولا تكتب	الرَّحْمَنُ وَ السَّمَوَاتِ	<alrrHmanu0wassamawaattau0	راعى بعض الكلمات و فشل في كلمات أخرى
حروف مد تقرأ ولا تكتب	إِلَهُ	<i0lahu0	
حروف مد تقرأ ولا تكتب	هَذِهِ	haa*i0hi0	
حالات تقرأ حروف المد فيها ولا تكتب	هَذَانِ	haa*aan	راعى بعض الكلمات و فشل في كلمات أخرى
تاء التانيث	قِمَّةُ الْجَبَلِ	qi0mmtu0ljabali0	

	mi0nlmaktab	مِنَ الْمَكْتَبِ	التقاء ساكنين
	lahu1mlgU0rfa	لَهُمُ الْغُرْفَةُ	التقاء ساكنين
راعى الوقف على ألف في حال تنوين النصب و على سكون في حال الحركات	na\$ii0TAn	نَشِيطاً	الوقف
راعى النظام القراءة المستمرة	wajadtu0hu0fii0lbayti0an	وَجَدْتُهُ فِي الْبَيْتِ	قراءة الكلمات على نحو مستمر

إن هذا النظام راعى اغلب الحالات ولتحسين النظام نحتاج الى بناء قاموس بالكلمات الشاذة ونعطيها ترميزها المناسب او كتابة قواعد خاصة بها .

منهجية العمل المتبعة

بعد إنهاء الدراسة المرجعية وجدنا ان هنالك عدة أنظمة تعمل على تحويل النص العربي إلى فونيمات صوتية مع اختلاف ترميز الفونيمات من نظام لآخر.

بالإضافة إلى أن هنالك أنظمة راعت حالات معينة من قواعد القراءة في اللغة العربية وأخرى لم تراعيها ولمعرفة النظام الأفضل بينها تم بناء تقريبا 30 حالة إختبار تشمل كل حالات قواعد القراءة في اللغة العربية حيث كل حالة إختبار تم تمثيلها بطريقتين (الاولى: هي خرج النظام , والثانية هي التمثيل الصحيح وفق فونيمات النظام)

وبناء على ذلك تم بناء (3 Dictionary)

حيث كل نظام له (Dictionary) : هو تمثيل للفونيمات الصحيحة المقابلة للنص العربي مع فونيمات (خرج النظام)

وبعدها تم حساب الدقة لكل نظام وذلك عن طريق مكتبة jellyfish التي تساعد في حساب نسبة التشابه بين (two string phonemes)

حيث يتم حساب نسبة التشابه لكل زوج فونيمات على حدا , ومن ثم حساب المتوسط الحسابي لكل النسب في كل نظام بغية الحصول على دقة النظام

كانت النتائج على الشكل التالي :

	system_name	similarity
0	Toph	0.900379
1	Arabic Phonetiser	0.890814
2	epitran lib	0.719056

ومن خلال النتائج نجد أن نظام Toph أفضل نظام من بين الأنظمة المقترحة

وتم الاعتماد عليه وتعديل بعض القواعد الموجودة ضمنه لتناسب مع الايات القرآنية لانه يراعي القراءة المتسلسلة وهي من أهم متطلبات قراءة الآيات القرآنية وكذلك حالات خاصة مثل بعض الكلمات القرآنية وحالات الوقوف على ساكن أو ألف تبعاً لنهاية الكلمة. وبالإضافة الى أن هذا النظام مبني بلغة برمجة c++ فتم ربطته مع بايثون وبناء واجهة مناسبة.

القواعد المعدلة :

تم مراعاة حالة همزة الوصل مضمومة

تم الاستغناء عن الحركات (الضمة , الكسرة) الممالة لمخالفتها لقواعد التجويد

تم مراعاة تفخيم حرف الراء مع الشدة

تم الاستغناء عن حالة تفخيم الواو والياء المديتان طبقا للمعتمد من قواعد التجويد مع الحفاظ على تفخيم الحركات في مكانها المناسب

تم الاستغناء عن تصويت الاعداد لعدم حاجتنا لها أثناء التلاوة

تم إضافة الحروف (خ,غ,ق) لاحرف الاستعلاء

تم مراعاة تفخيم الألف والحركات بعد الراء المشددة المفخمة

بعد تعديل النظام وتم الحصول على دقة تقريبا 94% أي تم تحسين النظام بنسبة 4%

ونستطيع رفع الدقة أكثر عند مراعاة كامل القواعد وذلك بعد تجربتها على الايات القرآنية وإيجاد الأخطاء وإصلاحها

لمحة مختصرة عن المكتبة المستخدمة في المقارنة jellyfish

jellyfish is a library for approximate & phonetic matching of strings.

• Included Algorithms

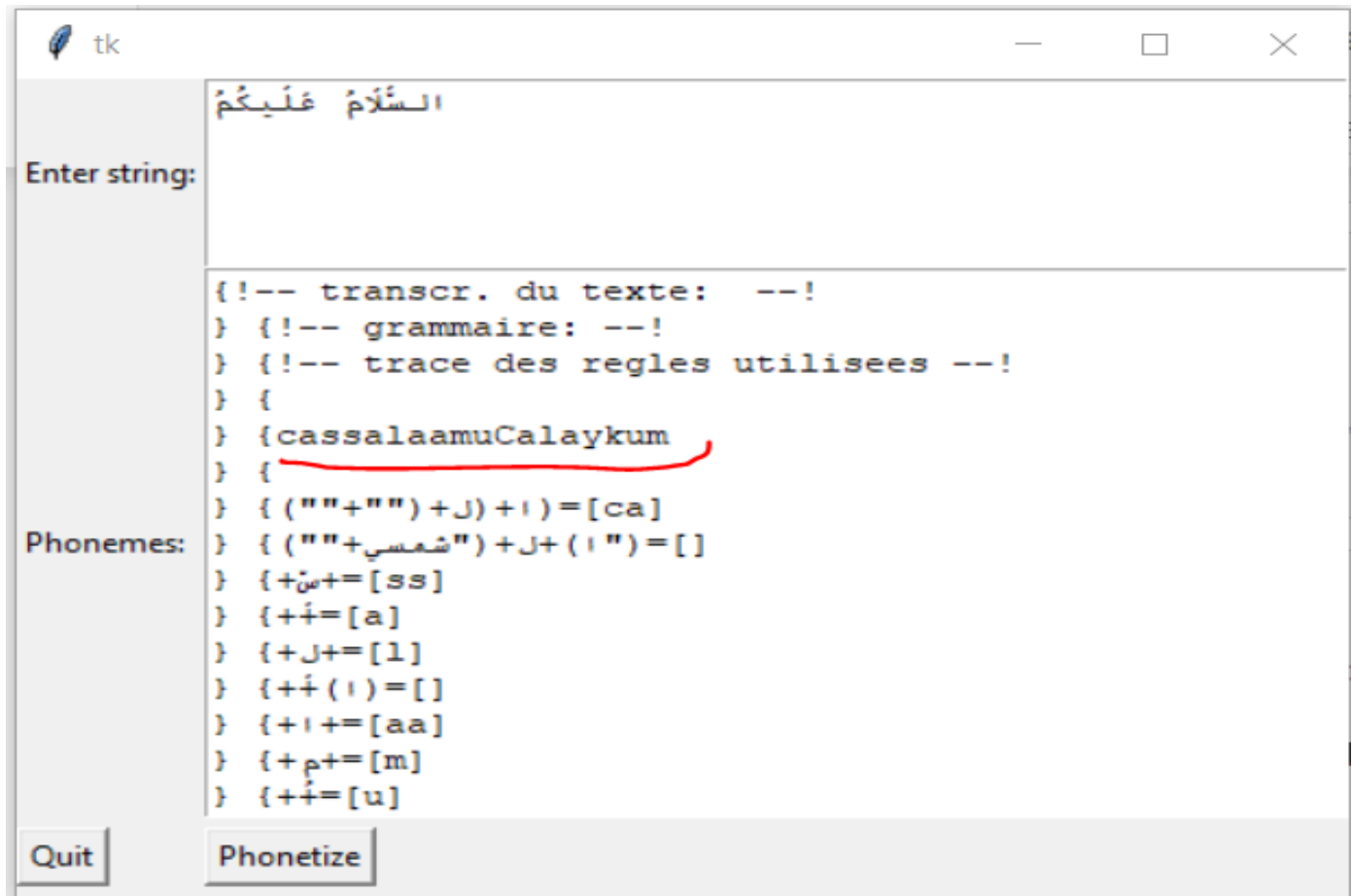
String comparison:

- Levenshtein Distance
- Damerau-Levenshtein Distance
- Jaro Distance
- Jaro-Winkler Distance
- Match Rating Approach Comparison
- Hamming Distance

Phonetic encoding:

- American Soundex
- Metaphone
- NYSIIS (New York State Identification and Intelligence System)
- Match Rating Codex

واجهة التنفيذ



- [Ghneim, Aubergé 85] Nada Ghneim, Veronique Aubergé, "Optimizing Tools for the French Letter-to-Phone Grammar TOPH With a View to Phonographic Spelling Correction", Institute de la Communication Parlée, INPG/Université Stendhal, 1985
- [Dakkak,Ghneim 99] Oumayma Aldakkak, Nada Ghneim, "Towards Man- Machine Communication in Arabic", Syria-Lebanese Conference, Damascus University, 1999
- مروان البواب، يحيى ميرعلم، محمد حسان الطيان، إشراف محمد نشرة داخلية، مركز [البواب، ميرعلم، والطيان ٨٤] الدراسات والبحوث العلمية، دمشق، « الكتابة الصوتية العربية » مرياتي . سورية، ١٩٨٤
- M. Alghamdi, Arabic Phonetics and Phonology. Riyadh: Altawbah Bookshop, 2015.
- Modern Standard Arabic Phonetics for Speech Synthesis. Nawar Halabi , July 2016
- F. K. Fahmy, M. I. Khalil, and H. M. Abbas, "A Transfer Learning End-to-End Arabic Text-To-Speech (TTS) Deep Architecture," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12294 LNAI, 2020, pp. 266–277.
-] L. Zhang, Z. Zhao, C. Ma, L. Shan, H. Sun, and L. Jiang, "End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC / Attention Architecture," 2020
- Paeper Automatically Determining Correct Application of Basic Quranic Recitation Rules
- Paeper Automatic Spoken Qur’anic Phonemes Classification I
- Paeper Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features
- C. Lopes and F. Perdigao, "Phoneme Recognition on the TIMIT Database," in Speech Technologies, vol. 1, InTech, 2011, pp. 285–302.
- K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 11. pp. 1641–1648, 1989, doi: 10.1109/29.46546