



دور قواعد البيانات في معالجة اللغات الطبيعية

أ. د. منصور بن محمد الغامدي

www.mghamdi.me

ندوة
معالجة اللغات الطبيعية

14 شعبان 1442 (27 مارس 2021)

مجمع اللغة العربية بمكة المكرمة

مكة ، المملكة العربية السعودية



قواعد بيانات

محرك

مخرجات

- نص مترجم
- نصوص مولدة
- كلام
- تقويم
- معاجم
- ...



- برمجيات
- خوارزميات
- ذكاء صناعي
- ...



كلام

- التعرف على المتحدث
- التعرف على اللغة
- التعرف على اللهجة
- التعرف على المزاج
- التعرف على الجنس
- التعرف على المخاطب
- التعرف على العمر
- التعرف على الكلام
- توليد الكلام
-

نص

- التعرف على اللغة
- التعرف على الموضوع
- التعرف على الكاتب
- التحليل الصرفي
- التحليل النحوي
- التحليل الدلالي
- فهم النص
- تلخيص النص
- توليد الجمل
- توليد المعاجم
- ...

صور

- لغة الإشارة
- نظام برايل
- تعابير الوجه
- تعابير حركية
- المخطوطات
- ...

The database contains:

- > 1,000 speakers
 - 51% male
 - 70% mobile
 - 50% aged 16-30
- > 60,000 PCM files
- > 60,000 wave files
- > 60,000 transcription files
- > 60,000 text files

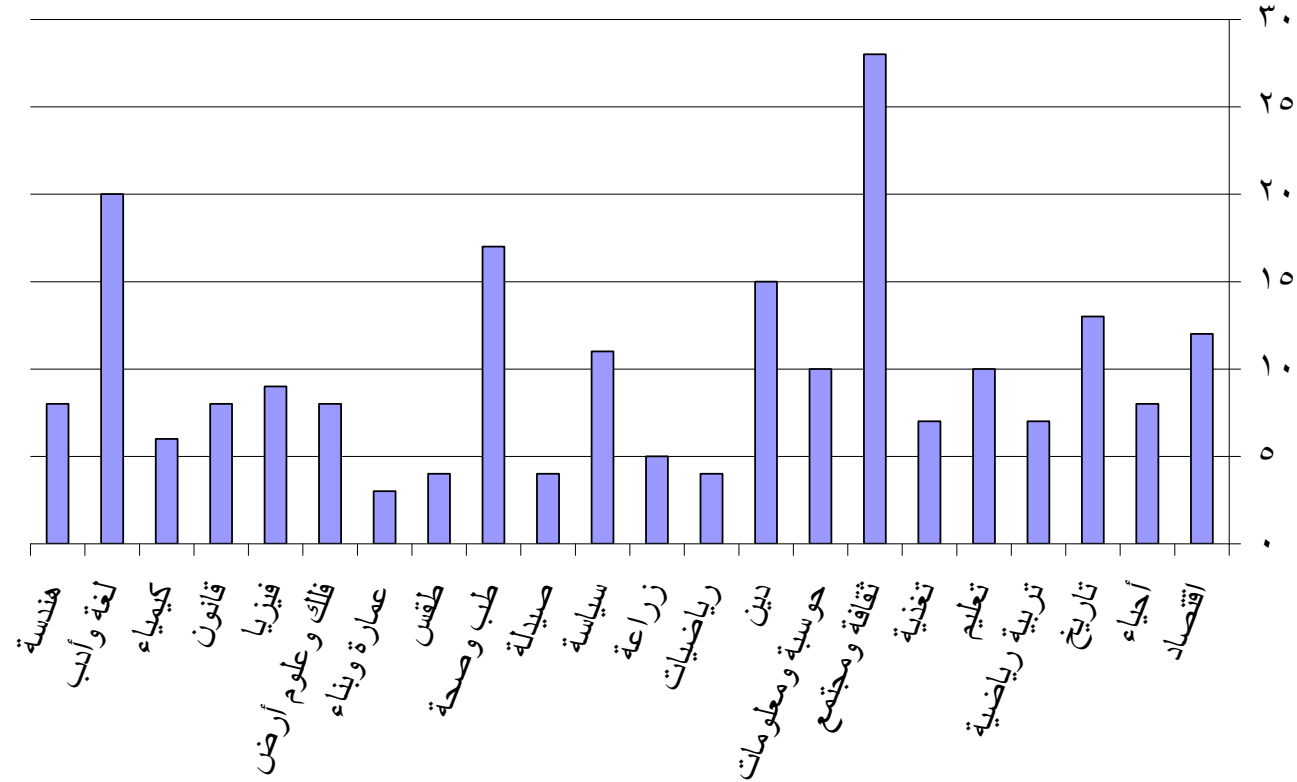


Quran database

Code	Nbr	MiD	MaD	AvD	ToD	Code	Nbr	MiD	MaD	AvD	ToD
as10	30296	0.011	1.939	0.124	3743	ns20	1199	0.096	1.434	0.774	928
as20	10100	0.053	3.343	0.459	4634	qs10	1810	0.008	0.490	0.145	263
bs10	3130	0.013	0.412	0.114	355	qs20	60	0.236	0.792	0.345	21
bs20	609	0.183	0.654	0.305	186	rs10	4322	0.014	0.352	0.085	368
cs10	2660	0.033	0.299	0.122	325	rs20	298	0.075	0.404	0.261	78
cs20	40	0.175	0.420	0.277	11	sb10	628	0.079	0.260	0.148	93
Db10	403	0.021	0.311	0.135	54	sb20	210	0.197	0.388	0.289	61
Db20	40	0.209	0.629	0.362	14	sil	11876	0.022	8.576	0.315	3745
ds10	1932	0.039	0.361	0.143	276	ss10	1933	0.071	0.384	0.156	301
ds20	258	0.193	0.433	0.307	79	ss20	242	0.176	0.371	0.285	69
fs10	2930	0.016	0.307	0.104	304	tb10	500	0.073	0.310	0.144	72
fs20	90	0.236	0.369	0.295	27	tb20	60	0.253	0.464	0.327	20
gs10	410	0.049	0.387	0.138	57	Ts10	3343	0.019	0.959	0.135	450
Hb10	1457	0.050	0.335	0.128	186	Ts20	140	0.161	0.395	0.306	43
hs10	4509	0.029	0.336	0.111	500	us10	6800	0.029	1.374	0.127	861
hs20	50	0.207	0.376	0.304	15	us20	2310	0.032	1.739	0.473	1093
Hz10	6281	0.008	0.295	0.079	494	vb10	1841	0.031	0.246	0.087	161
is10	10189	0.030	1.103	0.130	1322	vb20	250	0.176	0.371	0.279	70
is20	2566	0.052	1.833	0.518	1330	vs10	590	0.032	0.272	0.115	68
jb10	1160	0.015	0.361	0.121	140	vs20	10	0.221	0.311	0.276	3
jb20	80	0.225	0.428	0.316	25	ws10	4520	0.016	0.814	0.119	537
js10	800	0.080	0.478	0.143	114	ws20	127	0.042	1.021	0.341	43
js20	67	0.191	0.371	0.271	18	xs10	750	0.072	0.348	0.147	110
Ks10	2850	0.028	0.398	0.125	356	xs20	20	0.222	0.420	0.308	6
Ks20	190	0.220	0.480	0.314	60	ys10	3423	0.019	0.957	0.137	469
ls10	7662	0.015	0.350	0.118	904	ys20	254	0.058	1.392	0.329	84
ls20	1404	0.077	0.767	0.300	421	zb10	166	0.054	0.293	0.103	17
Ms10	6393	0.027	1.221	0.129	825	zb20	10	0.214	0.360	0.305	3
Ms20	432	0.084	1.640	0.777	336	zs10	380	0.060	0.316	0.117	45
ns10	6961	0.021	1.458	0.294	2045	zs20	60	0.177	0.352	0.276	17

Phonemes and their frequencies and durations in seconds

KACST Diacritized Text





مدينة الملك عبدالعزيز
للعلم والتكنولوجيا KACST

المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية



المدونة

بإذنا الزائر، مرحبا بك في موقع المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية
مدونة العربية). إحدى المشاريع الاستراتيجية لمبادرة الملك عبدالله للمحتوى العربي. يهدف المشروع
إلى بناء مدونة لغوية عربية تحوي مليار -بليون- كلمة مما دون بالعربية ابتداء من العصر الجاهلي وحتى
العصر الحديث ومن مختلف المناطق والبلدان مع الأخذ في الحسبان طبيعة وحجم النشاط الفكري لكل
منطقة وتنوع أوعية النشر فيها (مخطوطات، وصحف، وكتب، ومجلات، ودوريات علمية، ...) والساند من
جالات العلمية والفكرية المختلفة (المعتقدات، وعلوم العربية، والعلوم الطبيعية، والأدب، ...). كما
يمل المشروع بالإضافة إلى المادة اللغوية المصنفة المزيد....

إحصائيات

عدد النصوص الكلي: 1,323,185
عدد الكلمات الكلي: 1,182,515,633
عدد الكلمات بدون تكرار: 9,006,585



قوائم الكلمات الأكثر تكراراً

كامل المدونة
المخطوطات المحققة
الكتب
الصحف
المجلات
الدوريات المحكمة
الرسائل الجامعية
المناهج الدراسية
الإصدارات الرسمية
وكالات الأنباء
الإنترنت

الكلمات الأكثر بحثاً

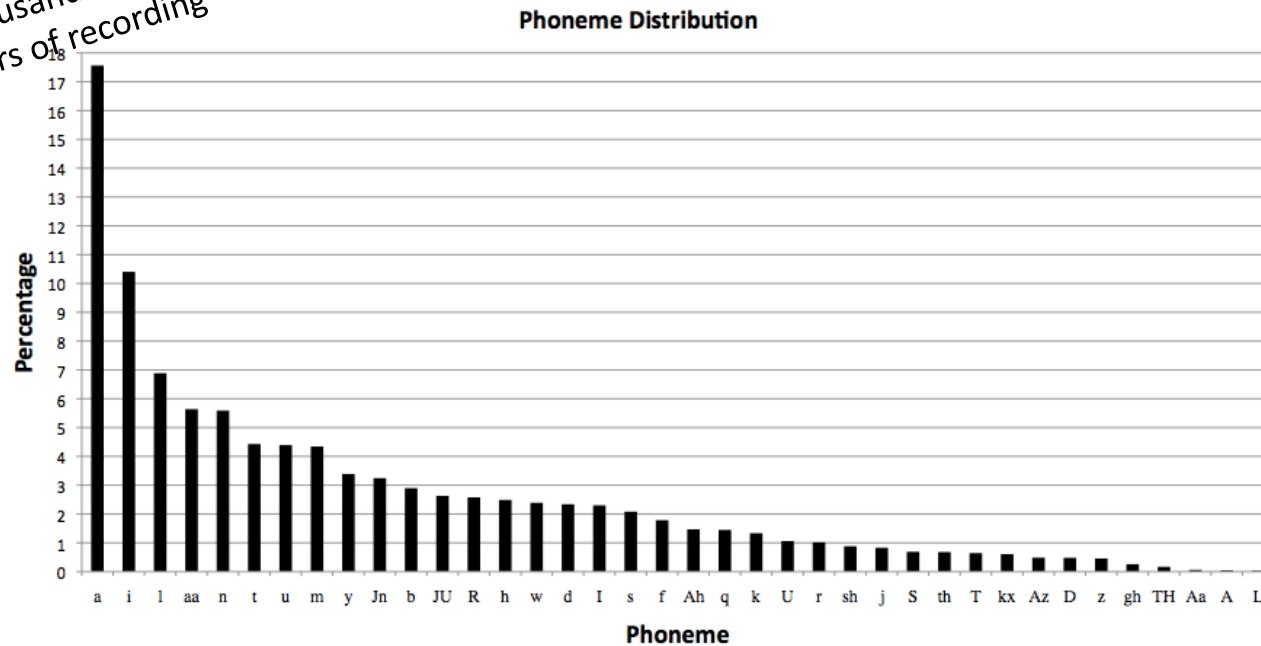
عربية	149	قرشي	33
ال	77	المملكة	32
ب	65	الكتابة	32
ب	54	هي	32
ل	52	تظاهر	31
برأة	45	من	28
بادف	44	ملك	25
ريعة	38	الجنات	25
ن خلال	34	حوسبة	24
وة	34	تاريخ	22



Standard Arabic Single Speaker Corpus

For text-to-speech systems:

- EGG
- 51 thousand words
- 7 hours of recording



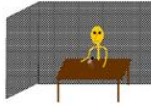
KACST Romanized Arabic Names

72000 Arabic Names:

- Diacritized
- Undiacritized
- Romanized



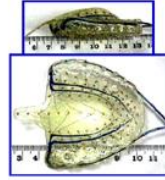
Kacst Arabic Phonetic Database (KAPD)



vocal folds vibration



facial image



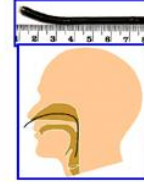
tongue movement



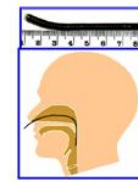
oral and nasal sounds



velo-pharyngeal port



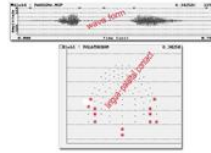
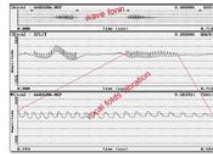
eegottiss



vocal folds



air flow & air pressor



مقالات تعبير

- 3400 مقالة مكتوبة على ورق
- أدخلت كنص على الحاسب
- صححها عضوا هيئة تدريس

شكرا على المتابعة