



# Lisan

## **Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations**

**Mustafa Jarrar**

Birzeit University  
Palestine

**Fadi Zaraket**

AUB  
Lebanon

**Tymaa Hammouda**

Birzeit University  
Palestine

**Daanish Alavi**

United Nations  
USA

**Martin Wählisch**

United Nations  
USA



We build  
tools and  
resources  
for NLU

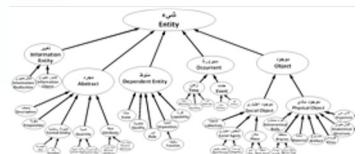
# Lexical Resources at SinaLab - Birzeit University

## Lexicographic Database



150 lexicons  
Largest Arabic lexicographic database

## Arabic Ontology/Wordnet



Formal Arabic Wordnet with ontologically clean content

Synonyms 90s%

WSD 84%

NER 90s%

Intent 88.4%

Offensive 88.4% .....

## Big Linguistic Data Graph

<https://ontology.birzeit.edu>

## Annotated Corpora



Dialects,  
NER, WSD, synonyms  
Intents, hate  
....

## NLP library



APIs  
Linguistic Data, synonyms, Nested NER, intents, ...



# Natural Language Understanding Tools and Datasets



Open Source

<https://sina.birzeit.edu/resources>



## Resources

Download and try NLP/NLU datasets, corpora, tools and services



+ Lexicographic Database (150 lexicons)

حوسبة المعاجم

+ Arabic Ontology

الأنطولوجيا العربية

+ Dialect Corpora (Currasat)

كراسات مدونة العاميات

+ Arabic Synonyms

استخراج متزدقات

+ Named Entity Recognition (Wojood)

وجود - لاستخراج أسماء الاعلام

+ Word Sense Disambiguation (Salma)

سلمى - محلل دلالي

+ ArBanking77 Intent Detection

تحديد المقصود في المساعدات الآلية

+ Offensive Language Detection

خطاب الكراهية بالعبرية

+ Lemmatizer

+ NLP Tools

# Lîsañ: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations

Mustafa Jarrar

*Birzeit University*

Birzeit, Palestine

[mjarrar@birzeit.edu](mailto:mjarrar@birzeit.edu)

Fadi A Zaraket

*American University of Beirut*

Beirut, Lebanon

[fz11@aub.edu.lb](mailto:fz11@aub.edu.lb)

Tymaa Hammouda

*Birzeit University*

Birzeit, Palestine

[1171779@student.birzeit.edu](mailto:1171779@student.birzeit.edu)

Daanish Masood Alavi

*UN Department of Peace-building and Political Affairs*

New York, USA

[masoodd@un.org](mailto:masoodd@un.org)

Martin Wählisch

*UN Department of Peace-building and Political Affairs*

New York, USA

[waehlisch@un.org](mailto:waehlisch@un.org)

**Abstract**—This article presents morphologically-annotated Yemeni, Sudanese, Iraqi, and Libyan Arabic dialects (Lîsañ) corpora. Lîsañ features around 1.2 million tokens. We collected the content of the corpora from several social media platforms.

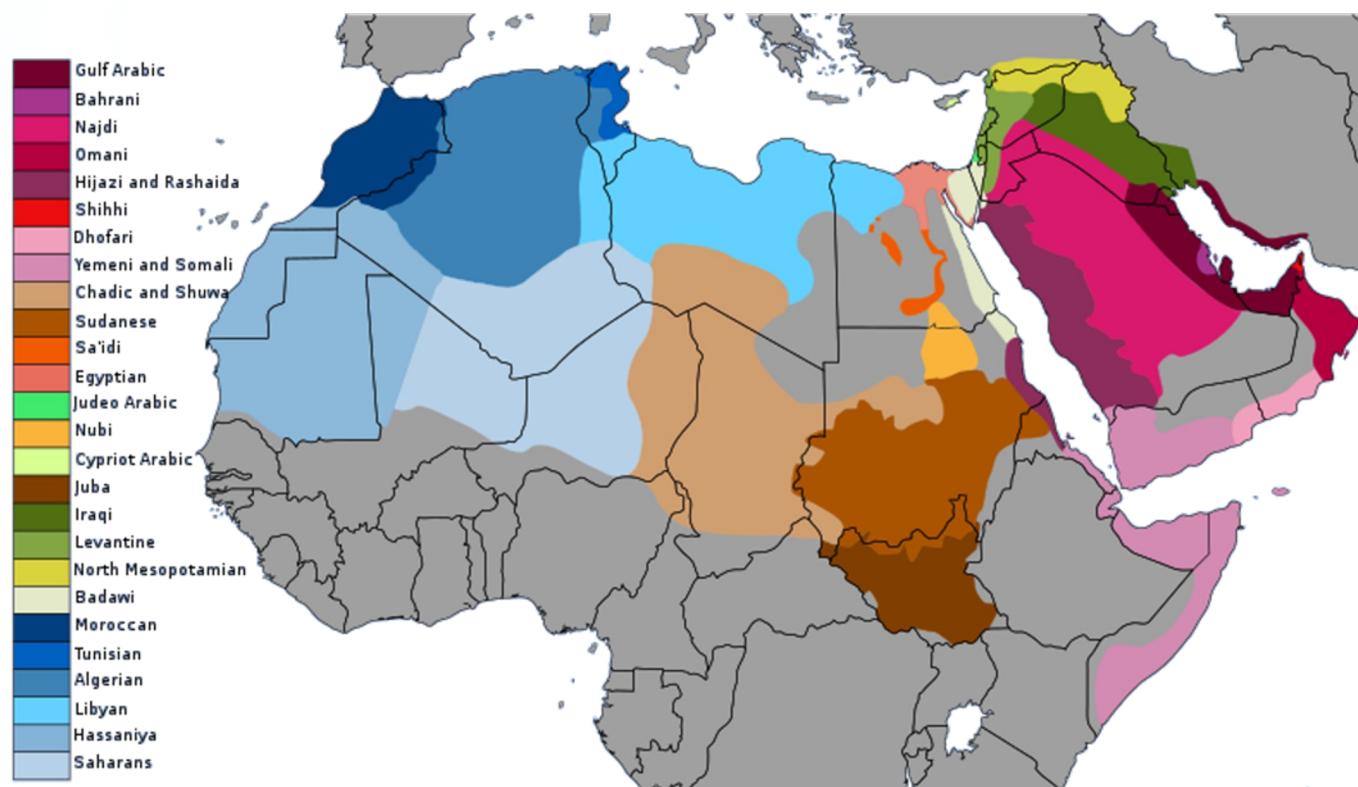
MSA, is pronounced *tš* 'tS' in Iraqi as in the word كَبْلَتْشِلْبَ (dog).

(ii) **Morphology**: Arabic dialects are similar to MSA in inheriting templatic morphology where affixes play an

# The problem

## Arabic is a low resources language

- Classical Arabic
- Modern Standard Arabic (more resources)
- Arabic Dialects





Seven morphologically-annotated Arabic dialect corpora (**1.35 million tokens**)

**Curras2** : گراس Palestinian dialect corpus (56K tokens)

**Baladi** : بلدي Lebanese dialect corpus (10K tokens)

**Nabra** : نبرة Syrian dialect corpus (60K tokens)

**Lisan** : لسان Yemeni, Iraqi, Libyan, and Sudanese dialects corpora

Yemeni (1.2 million), Iraqi (46K tokens), Libyan(52K tokens), Sudanese(53K tokens)



- El Haff, K., Jarrar, M., Hammouda, T., Zaraket, F., (2022). **Curras + Baladi: Towards a Levantine Corpus**. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.
- Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: **Nâbra: Syrian Arabic Dialects with Morphological Annotations**. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
- Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: **Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations**. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(-). IEEE. Egypt. 2023

Try and download



هيك

Search

EN | ع

Word  Stem  Lemma  Gloss

Whole Word  Substring

Palestinian  Lebanese  Iraqi  Libyan  Sudanese  Yemeni

About Publications

313 results (4.6 secs)

Gloss	Lemma	POS	Suffix	stem	Prefix	Word	Context
this way, this, thus	هيك / هكذا	ضمير اشارة		هيك		و/عطف	اللي اكلناه ، عزمنا الاسبوع الماضي ، عشان نردهم العزيمة ، <b>وهيك</b> بذلك تحطيلهم ترترسي
this way, this, thus	هيك / هكذا	ضمير اشارة		هيك		و/عطف	وهي قاطعا بحفلة قمة ، صارت عجيبة خللت الفحفات يطول حشيشن <b>وهيك</b> قدرت تخبي بيناتن ، ومن هون مصربنا نعمل الفحصة بيايامنا .
this way, this, thus	هيك / هكذا	ضمير اشارة		هيك		و/عطف	بيغنو : " بلاطة فوق بلاطة صاحبة البيت ضراطة ! " <b>وهيك</b> منكون وصلنا لآخر جولتنا بعالم عبد البربارا .
inon+like_that;thus	هيك	تعجب		هيك		ب/حرف جر	يمشون وياه الناس ولكن بيهم شي اشنون انذر هم يمتحنون <b>بهيج</b> مكان ياوزير ياعار
inon+like_that;thus	هيك	تعجب		هيج		ب/حرف جر	يجوز ياخذ بريد يخرع <b>بيج</b> جهاله
this way, like this, this, thus	هيك	ضمير اشارة		هيك		هيك	وين يعرف يسولف ويدبر هذا الحجي اكيد واحد كليله احجي <b>هيج</b>
inon+like_that;thus	هيك	تعجب		هيج		ب/حرف جر	و ش حيرتو العالم ولعيتو ببها طوبه كل <b>بيج</b> اينزل كلام اليعجبه <b>ويشر</b>



Rights Reserved © 2022

In cooperation with:



Rights Reserved © 2022

<https://sina.birzeit.edu/currasat/>

Try and download



write

Search

EN ع

Word  Stem  Lemma  Gloss  
 Whole Word  Substring  
 Palestinian  Lebanese  Iraqi  Libyan  Sudanese  Yemeni

About Publications

313 results (4.6 secs)

Gloss	Lemma	POS	Suffix	Stem	Prefix	Word	Context
be written/be fated/be destined   write	كتب	فعل مضارع مفرد، منذكر، متكلّم		كتب	و/عطف+المضارع المتكلّم المفرد	واكتب	وراح تكون هية <b>واكتب</b> وصيّة
be written/be fated/be destined   write	كتب	فعل مضارع مفرد، منذكر، مخاطب		كتب	و/عطف+إادة مضارع+ات+المضارع المخاطب المنذكر المفرد	وبنكتب	ونتعد على ليس <b>وبنكتب</b>
be written/be fated/be destined   write	كتب	فعل مضارع مفرد، منذكر، مخاطب	ت+للماضي: فاعله مخاطب منكر مفرد	كتب	و/عطف	وكتب	وكتب بوصيتك هو السبب
be written/be fated/be destined   write	كتب	فعل مضارع مفرد، منذكر، غائب	لحرف جر+هـ/اضمير متصل للغائب	كتب	و/عطف+ي/المضارع الغائب المنذكر الغدر	ويكتبهما	بس ثوفت انو قبل كم يوم كان يحطلها أغاني ويحكى عنها <b>ويكتبهما</b> ...
be written/be fated/be destined   write	كتب	فعل مضارع مفرد، منذكر، غائب		كتب	ي/المضارع الغائب المنذكر المفرد	يكتب	الاين : ( <b>يكتب</b> سباتوس ع ليس) لا تتحدثوا عن الواقع اكتر الاشياء وجما
be written/be fated/be destined   write	كتب	فعل مضارع مفرد، منذكر، غائب		كتب	ت/المضارع الغائب المنذكر المفرد	تنكتب	، وفجيئت كيف هالله تطزورت من أتحقق لهجة ما كان <b>ينسوا تكتب</b> للغة عالميه ما <b>ينسوا</b> ما <b>تكتب</b> .
be written/be fated/be destined   write	كتب	فعل مضارع مجهول مفرد، مؤنث، غائب		كتب	ت/المضارع الغائب المنذكر المفرد	تنكتب	أنجح لهجة ما كان <b>ينسوا تكتب</b> للغة عالميه ما <b>ينسوا</b> ما <b>تكتب</b> .
scatter/sprinkle/write in prose	تشر	فعل مضارع مفرد، منذكر، غائب		تشر	ب/إادة مضارعة	بنثر	وبنثر رماد
they (people) + write + اسماء معرفة	كتبوا	فعل مضارع	و/المضارع: فاعله منذكر جمع	كتبوا	ي/المضارع الغائب المنذكر المفرد	يكتبوا	و الكل يتكلّم عن السنن الجرف الأجنبية وكل الصيادين يبغوا <b>يكتبوا</b>



Rights Reserved © 2022

In cooperation with:  
   
 AMERICAN UNIVERSITY OF BEIRUT United Nations  
 Rights Reserved © 2022

<https://sina.birzeit.edu/currasat/>

# Annotation Tools



SinaLab

# Tawseem Portal

بوابة سينا لتوسيم المدونات

Tymaa Rol دخول  
Login 14:26:35 (118)

Stat Log

Lemmas	Annotations	Gloss	MSA Lemma	DA Lemma	Person	Gender	Number	POS	Suffix	Stem	Prefix	Token	Context	Dialect
بصیر	اسم فعل وظيفية صار صار بصیر	how	كيف 1	شلون				أداة استفهام		شلون		شلون	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
بصیر	Count: 2 صار صار بصیر فعل مضارع مفرد ذکر غائب ب/أداة مضارعة become/begin to	will/shall	سوف 1	رَجَحُوا				أداة استقبال		دا		دا	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
بصیر	صار صار بصیر فعل مضارع مفرد ذکر منكلم ب/أداة مضارعة become/begin to	enter	دخل 1	دخل	مفرد	مؤنث	مفرد	فعل مضارع	ي/المضارع: فاعله مخاطب مؤنث مفرد	دخل	ت/المضارع: المخاطب المؤنث المفرد	تدخلی	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
بصیر	صار صار بصیر فعل مضارع مفرد ذکر منكلم ب/أداة مضارعة become/begin to	hand over/surren	سلم 1	سلم	مفرد	مؤنث	مفرد	فعل مضارع	ي/المضارع: فاعله مخاطب مؤنث مفرد	سلم	ت/المضارع: المخاطب المؤنث المفرد	تسلمي	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
بصیر	صار صار بصیر فعل مضارع مفرد ذکر منكلم ب/أداة مضارعة become/begin to	in childbed	نفساء 1	نفساء	مفرد	مؤنث	مفرد	اسم		نفسا	ع/حرف جر+ال/أداة تعريف	عالنفسا	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
			..	..				علامة ترقيم		..		..	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
		not	ما 2	ما				أداة نفي		ما		ما	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
		become/be gin to	صار 2	صار	غائب	ذکر	مفرد	فعل مضارع			ب/أداة مضارعة	صیر	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
		to/for + Allah/God +	الله ..	الله		ذکر	مفرد	اسم		الله	ي/أداة نداء	يالله	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة
		sisters   sister/count	أخت 1	أخت	-	مؤنث		اسم		خيتو		خيتو	شلون دا تدخلی تسلمي عالنفسا ... ما بصیر يالله خیتو	حلبیة

# Lisan Corpora

## Corpus Collection:

From: social media networks, mainly from Twitter, Facebook, and YouTube.

Size: 1.2M tokens, represented by 48K documents.

## Lisan Dialect Corpora Covered:

Yemeni, Iraqi, Sudanese, and Libyan.

## Annotation Methodology:

Tool: Arabic Dialect Annotation Toolkit (ADAT), over two years

Corpus name	Yemeni	Iraqi	Sudanese	Libyan
Tweets	38,819	3,326	3,000	3,053
Token	1,098,222	45,881	52,616	51,686
Unique Tokens	136,801	17,812	18,242	18,556
Unique Lemma	43,320	9,086	10,251	9,924
Nouns	627,907	26,550	28,557	27,761
Verbs	178,381	8,371	9,249	9,827
Functional words	260,655	10,097	13,347	12,954
Digit	3,962	128	7	177
Others (e.g., Foreign words)	27,317	735	1,456	967

# Annotation Methodology



اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **يقطون**:

جاشلون رايك **يقطون** للوقد لغات فلافل

هاي لزكت اشو دك رشاوي بزمن صنام **يقطون** للبزير ع كاع

كاساع **يقطون** بيان ويطلع جنب بدون كتاب رسمي

تصريف كلمة:	التجهنة الصحيحة/Normalized token:	اللهمجة/Dialect:	العربية/Arabic
POS/تصريف فعل مضارع	يقطون	يقطون	
Suffix/لاحقة:	ون/لل مضارع: فاعله جمع	الساق/Stem:	ي/لل مضارع الغائب الجمع
Gloss/المعنى:	give/provide	MSA Lemma/المدخلة بالفصحي:	أعطي 1
Note/ملاحظة:		Confidence/درجة التأكيد:	
		Refer/إحالات إلى:	متناك جداً
<input type="button" value="حفظ"/>			

الحلول الصرفية المحتملة لكلمة يقطون										الحلول الصرفية المحتملة لكلمة يقطون
Gloss المعنى	MSA Lemma المدخلة بالفصحي	POS تصريف	Suffix لاحقة	Stem ساق	Prefix باذنة	Normalized token التجهنة الصحيحة	Token الكلمة	Dialect اللهمجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريحات والمعنى لكلمة **ينطون**:

جاشلون رايك **ينطون** للوقد لغات فلابل

هاي لزكت اشو دك رشاوي بزمن صنام **ينطون** للبزر ع كاع

كاساع **ينطون** بيان ويطلع جنب بدون كتاب رسمي

تصريف كلمة : Dialect/اللهجة عراقيه

POS التصريف/ فعل مضارع Normalized token/التهجنة الصحيحة/ Stem/الساق Prefix/بادنة/ي/لل مضارع الغائب الجمع

MSA Lemma/المدخلة أخطى 1 Refer إلى/إحاله إلى من تلك جداً

الحلول الصرفية المحتملة لكلم

**Task:**

- Each task is a set of words that need to be annotated
- and that belong to a specific context (same tweet, post, comment)

Gloss المعنى	MSA Lemma المدخلة بالفصحى	POS تصريف	Suffix لاحقة	Stem ساق	Prefix بادنة	Normalized token التهجنة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أخطى 1	فعل مضارع	ون/لل مضارع: فعل مضارع past/فاعله جمع	عط/فعل مضارع	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **يقطون**:

جاشلون رايك **يقطون** للوقد لغات فلافل

هاي لزكت اشو دك رشاوي بزمن صنام **يقطون** للبizer ع كاع

كاساع **يقطون** بيان ويطلع جنب بدون كتاب رسمي

تصريف كلمة : **اللهم** / Dialect / عراقي

Normalized token/التهجنة الصحيحة : **يقطون**

S/التصريف :  **فعل مضارع**

Prefix/بادنة : **ي** /لل مضارع الغائب الجمع

Suffix/لاحقة : **ون** /لل مضارع

Stem/الساق : **قطن** /نط/ فعل مضارع

Gloss/المعنى : **give/provide**

MSA Lemma/المدخلة بالفصحي : **أعطي**

Note/ملاحظة :

Confidence/درجة التأكيد :

Refer to/إحاله إلى : **متناك جداً**

حفظ

الحلول الصرفية المحتملة لكلمة **يقطون**

Gloss المعنى	MSA Lemma المدخلة بالفصحي	POS تصريف	Suffix لاحقة	Stem ساق	Prefix بادنة	Normalized token التهجنة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع:	فاعله جمع	عط/ فعل مضارع	ي/لل مضارع	يعطون	فصحي	سما	138394

Contexts (sentences) containing the selected word from the task

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **ينطون**:

- جاشلون رأيك **ينطون** للوقد لغات فلابل
- هاي لزكت اشو دك رشاوي بزمن صنام **ينطون** للبزر ع كاع
- كاساع **ينطون** بيان وبطلع جنب بدون كتاب رسمي

تصريف كلمة : **ينطون**

اللهجة/ Dialect: **عربي**

تصريف الكلمة: **ينطون**

التهجنة الصحيحة/ Normalized token: **ينطون**

التصريف/ POS: فعل مضارع

النهاية/ End:

العنوان/ MSA Lemma: **أعطي**

ي/ للمضارع الغائب الجمع

يادنة/ Prefix: **ين-**

ي/ للمضارع الغائب الجمع

حالات إلى/ Refer: **متناك جداً**

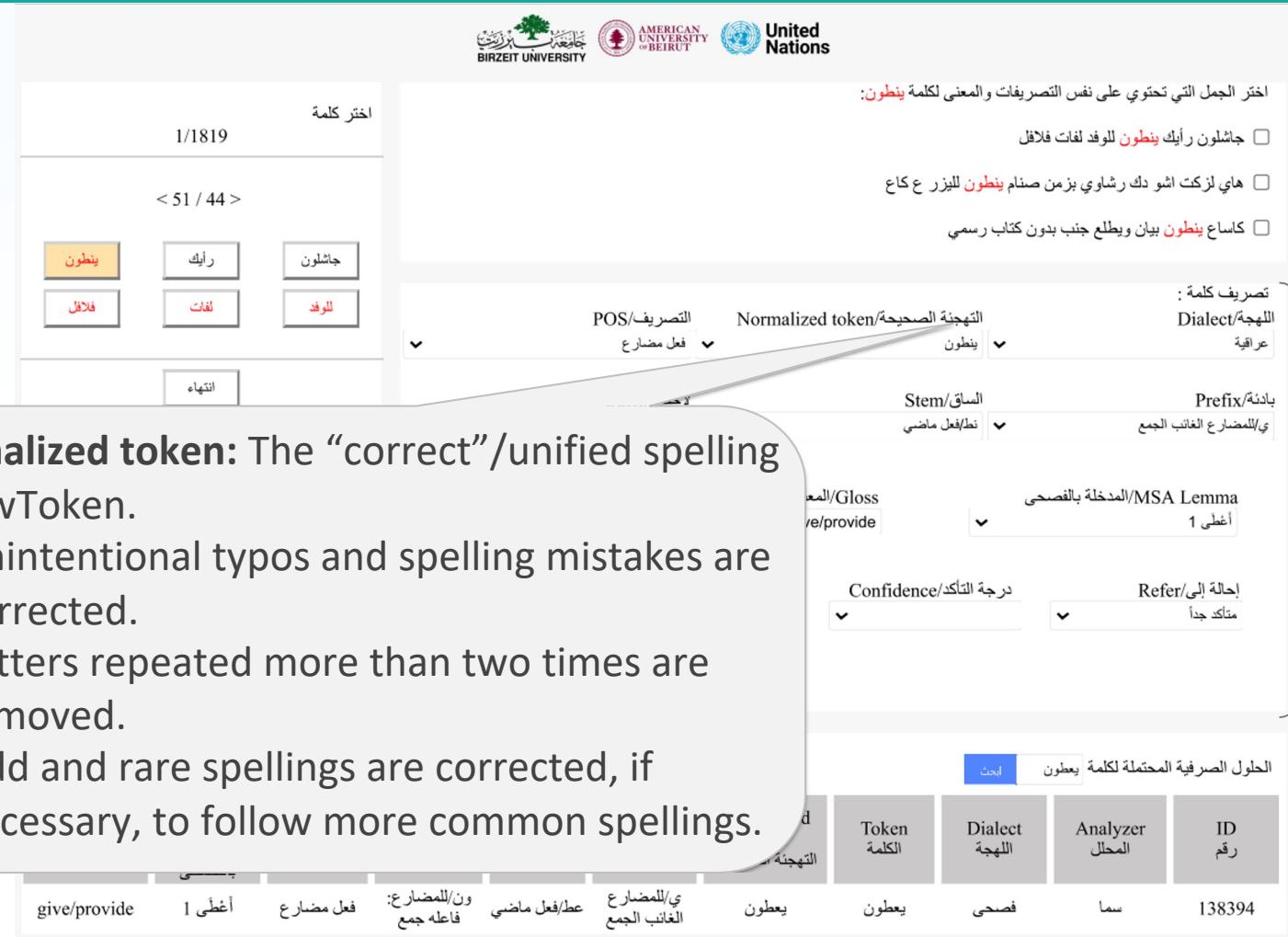
مorfologische Features

**Dialect: Original dialect of the word, such as Yemeni, Iraqi, Sudanese, or Libyan.**

الحلول الصرفية المحتملة لكلمة **يعطون**

Gloss المعنى	MSA Lemma المدخلة بالنصجي	POS تصريف	Suffix لاحقة	Stem ساق	Prefix يادنة	Normalized token التهجنة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/للمضارع: فاعله جمع	عط/فعل مضارع ماضي	ي/ للمضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

# Annotation Methodology



**Normalized token:** The “correct”/unified spelling of rowToken.

- Unintentional typos and spelling mistakes are corrected.
  - Letters repeated more than two times are removed.
  - Odd and rare spellings are corrected, if necessary, to follow more common spellings.

# Morphological Features

# Annotation Methodology



اختر الجمل التي تحتوي على نفس التصريفات والمعنى للكلمة **يُنطون**:

جاشلون رأيك **يُنطون** للوقد لغات فلابل

هاي لزكت اشو دك رشاوي بزمن صنام **يُنطون** للبزير ع كاع

كاساع **يُنطون** بيان وبطع جنب بدون كتاب رسمي

اختر كلمة

1/1819

< 51 / 44 >

يُنطون
رأيك
جاشلون

فلابل
لغات
اللوقد

تصريف الكلمة:
اللهجة/ Dialect
عرقانية

التصريف/POS
التهجنة الصحيحة/Normalized token
يُنطون

فعل مضارع
الساق/ Stem
ي/لل مضارع الغائب الجمع

لا حقة/ Suffix
نط/ فعل مضارع
ي/لل مضارع الغائب الجمع

المدخلة بالفصحي/Gloss
ماضي
المعنى

الكلمة/ MSA Lemma
أعطي 1
المعنى

ملاحظة/ Note
درجة التأكيد/ Confidence
إحالات إلى/ Refer

حفظ

**POS: Same SAMA tagset.**

Morphological Features

الحلول الصرفية المحتملة لكلمة <b>يُعطون</b>										
Gloss المعنى	MSA المدخلة بالفصحي	POS تصريف	Suffix لا حقة	Stem ساق	Prefix باذنة	Normalized token التهجنة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع ماضي	ي/لل مضارع الغائب الجمع	يُعطون	يُعطون	فصحي	سما	138394

# Annotation Methodology

اختر الكلمة

1/1819

< 51 / 44 >

<span style="background-color: #f0e68c; padding: 2px 5px;">بنطون</span>	<span style="border: 1px solid #ccc; padding: 2px 5px;">رأيك</span>	<span style="border: 1px solid #ccc; padding: 2px 5px;">جاشلون</span>
<span style="border: 1px solid #ccc; padding: 2px 5px;">فلافل</span>	<span style="border: 1px solid #ccc; padding: 2px 5px;">لغات</span>	<span style="border: 1px solid #ccc; padding: 2px 5px;">اللوفد</span>

انتهاء

اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **بنطون**:

جاشلون رأيك **بنطون** اللوفد لغات فلافل

هاي لزكت اشو دك رشاوي بزمن صنام **بنطون** الليزر ع كاع

كاساع **بنطون** بيان وبطلع جنب بدون كتاب رسمي

تصريف الكلمة:

اللهجة/  
Dialect/  
عرقية

Normalized token/  
التهجنة الصحيحة/  
 فعل مضارع

POS/  
Suffix/  
ون/لل مضارع: فاعله جمع

Prefix/  
ي/لل مضارع الغائب الجمع

Stem/  
الساق/  
نطا/ فعل مضارع

MSA Lemma/  
المدخلة/  
أعطي 1

Refer/  
إحاله إلى/  
متناكب جداً

**Prefixes (سوابق)**

حفظ

الحلول الصرفية المحتملة لكلمة **يعطون**

بحث

Gloss	MSA Lemma	POS	Suffix	Stem	Prefix	Normalized token	Token	Dialect	Analyzer	ID
المعنى	المدخلة بالنصحي	تصريف	لاحقة	ساق	بادنة	التهجنة الصحيحة	الكلمة	اللهجة	المحل	رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

Morphological Features

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **يعطون**:

- جاشلون رأيك **ينطون** للوقد لغات فلابل
- هاي لزكت اشو دك رشاوي بزمن صنام **ينطون** للبزير ع كاع
- كاساع **ينطون** بيان وبطع جنب بدون كتاب رسمي

تصريف كلمة : **اللهجة/ Dialect** **عرقية/ Arabic**

POS التصريف/ Normalized token/ التهجة الصحيحة/ Stem/ الباقي

فعل مضارع ينطون

Suffix/ لاحقة ون/لل مضارع: فاعله جمع Stem/ الباقي نطا/ فعل مضارع

Prefix/ بادئة ي/لل مضارع الغائب الجمع MSA Lemma/ المدخلة بالفصحي

الـ **الساقا** (الساقا) gloss/ide

Confidence/ درجة التأكيد Refer/ إحالـة إلى

حفظ

الحلول الصرافية المحتملة لكلمة **يعطون**

Gloss المعنى	MSA Lemma المدخلة بالفصحي	POS تصريف	Suffix لاحقة	Stem ساق	Prefix بادئة	Normalized token التهجة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

Morphological Features

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريفات والمعنى للكلمة **يقطون**:

جاشلون رأيك **يقطون** للوفد لغات فلابل

هاي لزكت اشو دك رشاوي بزمن صنام **يقطون** للبزير ع كاع

كاساع **يقطون** بيان وبطعن جنب بدون كتاب رسمي

تصريف الكلمة:

اللهجة/Dialect: عراقيّة

تصريف/POS	التهجنة الصحيحة/Normalized token	الлярدة/Prefix
فعل مضارع	يقطون	ي
Suffix/لاحقة	ون/لل مضارع: فاعله جمع	الساق/Stem
		نطا/ فعل مضارع
		ي
		لل مضارع الغائب الجمع

**Suffixes (واحد)**

المدخلة بالفصحي/MSA Lemma: أعطي 1

المعنى/Gloss: give/provide

ملاحظة/Note: ملاحظة

درجة التأكيد/Confidence: حفظ

إحالات إلى/Refer: مناكد جداً

حفظ

Morphological Features

Gloss المعنى	MSA Lemma المدخلة بالفصحي	POS تصريف	Suffix لاحقة	Stem ساق	Prefix بادئة	Normalized token التهجنة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع ماضي	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **ينطون**:

جاشلون رأيك **ينطون** للوفد لغات فلابل

هاي لزكت اشو دك رشاوي بزمن صنام **ينطون** للبزار ع كاع

كاساع **ينطون** بيان وبطعن جنب بدون كتاب رسمي

تصريف كلمة:	اللهجة/Dialect:	عرقية
ينطون	ينطون	
فلابل		
لغات		
الوفد		

انتهاء

POS التصريف/ فعل مضارع      Normalized token/ التهجنة الصحيحة/ الباقي

Suffix/ لاحقة ون/لل مضارع: فاعله جمع      Stem/ الساق نطا/ فعل مضارع      Prefix/ باذنة ي/لل مضارع الغائب الجمع

Gloss/ المعنى give/provide      MSA Lemma/ المدخلة بالفصحي أغطي 1

الحلول الصرفية المحتملة لكلمة بعطا

Morphological Features

**مدخلة معجمية (فصحي):**

- SAMA lemmas
- If no SAMA lemma, and one.

Gloss المعنى	MSA Lemma المدخلة بالفصحي	POS تصريف	Suffix لاحقة	ساق	باذنة	TOKEN التهجنة الصحيحة	الكلمة	اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

# Annotation Methodology

The screenshot shows the ADAT annotation interface with various morphological features and a 'Refer' feature.

**Morphological Features:**

- POS التصريف/ فعل مضارع:** Normalized token/الهجة الصحيحة/اللهم
- Suffix/لاحقة:** ون/لل مضارع: فاعله جمع Stem/الساق: نطا/ فعل مضارع Past Prefix/بادنة/ي لل مضارع الغائب الجمع
- Gloss/المعنى:** give/provide MSA Lemma/المدخلة بالفصحي: أغلى 1
- Note/ملاحظة:** Confidence/درجة التأكيد Refer/إحالات إلى من تأكد جداً

**Refer (إحالات إلى):**

- In case of hesitation about the annotation of a certain word,
- ADAT allows the annotators to “refer” the solution to another more experienced annotator for review.

**Annotations:**

- Chosen sentence: اختيار الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **ينطون**:
- Options: جاشلون رأيك **ينطون** للوقد لغات فلابل
- Options: هاي لزكت اشو دك رشاوي بزمن صنام **ينطون** للبزر ع كاع
- Options: كاساع **ينطون** بيان وبطعن جنب بدون كتاب رسمي

**Bottom Panel:**

- الحلول الصرفية المحتملة لكلمة **يعطون**
- Buttons: Select, Analyzer, ID
- Text: فصل, سما, 138394

Morphological  
Features

# Annotation Methodology

اختر الجمل التي تحتوي على نفس التصريفات والمعنى لكلمة **ينطون**:

جاشلون رأيك **ينطون** للوفد لغات فلابل

هاي لزكت اشو دك رشاوي بزمن صنام **ينطون** للبزير ع كاع

كاساع **ينطون** بيان وبطعن جنب بدون كتاب رسمي

تصريف كلمة:	اللهجة/Dialect:	عرقية/Arabic
POS التصريف/ فعل مضارع	التهجنة الصحيحة/Normalized token	ينطون
Suffix/لاحقة ون/لل مضارع: فاعله جمع	الساق/Stem	نط/ فعل مضارع
	Prefix/بادنة	ي/لل مضارع الغائب الجمع
	Gloss/المعنى	give/provide
	MSA Lemma/المدخلة بالفصحي	أعطي 1
Note/ملاحظة	Confidence/درجة التأكيد	Refer/إحالات إلى
		متناكب جداً

**Confidence**  
(High, Normal, or Low).

الحلول الصرفية المحتملة لكلمة **يعطون**

البحث

Gloss المعنى	MSA Lemma المدخلة بالفصحي	POS تصريف	Suffix لاحقة	Stem ساق	Prefix بادنة	Normalized token التهجنة الصحيحة	Token الكلمة	Dialect اللهجة	Analyzer المحلل	ID رقم
give/provide	أعطي 1	فعل مضارع	ون/لل مضارع: فاعله جمع	عط/ فعل مضارع	ي/لل مضارع الغائب الجمع	يعطون	يعطون	فصحي	سما	138394

Morphological Features

- ❖ Inter-annotation agreement (IAA), number of normalized unique categories (UNQ), and total number of overlaps per feature (OVP)

Feature	Iraqi			Libyan			Sudanese			Yemeni		
	IAA	UNQ	OVP	IAA	UNQ	OVP	IAA	UNQ	OVP	IAA	UNQ	OVP
Stem	.972	6,764	61,829	.975	7,072	65,670	.989	6,914	64,307	.981	25,237	1,366,425
Lemma	.933	8529	61,828	.930	9,194	65,670	.944	9,562	64,307	.948	35,503	1,366,482
LemmaD	.894	7	60,120	.904	7	65,282	.926	7	63,818	.899	18	1,335,496
POS	.950	147	61,610	.953	165	65,468	.970	117	64,222	.956	447	1,362,675
Prefix	.975	188	128,233	.976	280	135,725	.981	133	133,048	.982	788	2,827,159
Suffix	.920	265	128,257	.941	338	136,394	.938	189	132,558	.921	1,397	2,839,170
POS-P	.802	496	61,609	.785	648	65,521	.795	620	64,174	.813	3,063	1363800
POS-X	.874	806	61,090	.871	1,041	65,128	.877	881	63,689	.870	3,973	1,350,781
Voc	.978	15,400	61,796	.984	15,940	65,668	.989	15,777	64,306	.990	106,878	1,366,469

Try Online



<https://sina.birzeit.edu/currasat/>

ع | EN

# كراسات Currasat

Arabic Dialects Corpora

مدونة اللهجات العربية

Search

انبجت

Word Stem Lemma Gloss

Whole Word Substring

Palestinian Lebanese Syrian Iraqi Libyan Sudanese Yemeni

About Publications

Gloss	Lemma	POS	Suffix	Stem	Prefix	Word	Context
become angry	غَيْبَ ١	فعل مضارِي مفرد، مؤنث، غائب	ت/للماضي: فاعله غائب مؤنث	انبجت		انبجت	انبجت مثل الأباء
become angry	غَيْبَ ١	فعل مضارِي مفرد، متكلّم	ت/للماضي: فاعله متكلّم مفرد	انبجت		انبجت	يوم انا ما انبجت .... بس انتي كل مادا تسرحلي تقليلي قاعي ابركي والا ليهيك

# Summary

- **Baladi:** Lebanese morphologically annotated corpus (9.6K)
- **Curras:** Palestinian morphologically annotated corpus (56K)
- **Nabra:** Syrian morphologically annotated corpora (60K)

**Baladi + Curras + Nabra = a more Levantine Corpus (125.6K)**

- **Lisan:** Iraqi, Yemeni, Sudanese, and Libyan morphologically annotated corpora (1.2M)

Total Tokens: **Currasat** contains (1.325.6M tokens)

# References

1. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammad Khalilia: SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
2. Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi Zaraket, Mohamad-Bassam Kurdy: Nâbra: Syrian Arabic Dialects with Morphological Annotations. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
3. Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem: ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
4. Haneen Lqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, Muhammad AbdulMageed: Arabic Fine-Grained Entity Eecognition. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
5. Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim El-madany, Nagham Hamad, Alaa' Omar: WojooodNER 2023: The First Arabic Named Entity Recognition Shared Task. In Proceedings of the 1st Arabic Natural Language Processing Conference (Arabic- NLP), Part of the EMNLP 2023. ACL.
6. Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, Tymaa Hammouda, Mustafa Jarrar: Open-source thesaurus development for under-resourced languages: a Welsh case study. The 4th LDK Conference on Language, Data and Knowledge, Vienna, Austria, 12-15 September 2023
7. Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, Nadim Nashif: Offensive Hebrew Corpus and Detection using BERT. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). IEEE. Egypt. 2023
8. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp. ). San Sebastian, Spain, 2023
9. Sanad Malaysha, Mustafa Jarrar, Mohammad Khalilia: Context-Gloss Augmentation for Improving Arabic Target Sense Verification. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp. ). San Sebastian, Spain, 2023
10. Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem: Wojoood: Nested Arabic Named Entity Corpus and Recognition using BERT. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
11. Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(-). IEEE. Egypt. 2023 arXiv, DOI 10.48550/ARXIV.2212.06468. 2023
12. Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, Fadi Zaraket: Curras + Baladi: Towards a Levantine Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
13. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
14. Moustafa Al-Hajj, Mustafa Jarrar: ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40-48, 2021
15. Moustafa Al-Hajj, Mustafa Jarrar: LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation. In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). PP 748–755, Association for Computational Linguistics. 2021
16. Eman Naser-Karajah, Nabil Arman, Mustafa Jarrar: Current Trends and Approaches in Synonyms Extraction: Potential Adaptation to Arabic. In Proceedings of the 2021 International Conference on Information Technology (ICIT). PP 748–755, Association for Computational Linguistics. pp. 428-434, IEEE. 2021
17. Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: Extracting Synonyms from Bilingual Dictionaries. The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021
18. Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, Hamdy Mubarak: A Panoramic Survey of Natural Language Processing in the Arab World. Communications of the ACM, April 2021, Vol. 64 No. 4, Pages 72-81
19. Mustafa Jarrar: Digitization of Arabic Lexicons. Arabic Language Status Report. UAE Ministry of Culture and Youth. Pages 214-2017. Dec 2020
20. Mustafa Jarrar, Hamzeh Amayreh: An Arabic-Multilingual Database with a Lexicographic Search Engine. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Pages(234-246). LNCS 11608, Springer. 2019
21. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: Representing Arabic Lexicons in Lemon - a Preliminary Study. The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
22. Diana Alhafi, Anton Deik, Mustafa Jarrar: Usability Evaluation of Lexicographic e-Services. The 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Abu Dhabi, UAE. 2019
23. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: Diacritic-Based Matching of Arabic Words. ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
24. Mustafa Jarrar: Search Engine for Arabic Lexicons. The 5th Conference on Translation and the Problematics of Cross-cultural Understanding. The Forum for Arab and International Relations. Doha, Qatar. December, 2018
25. Diab Abuaiadah, Dileep Rajendran, Mustafa Jarrar: Clustering Arabic Tweets for Sentiment Analysis. The 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications. Pages(499-506). IEEE Computer Society. ISBN:9781538635810. (doi.10.1109/AICCSA.2017.162). Hammamet, Tunisia. 2017
26. Mustafa Jarrar, Nizar Habash, Faeg Alrimawi, Divam Akra, Nasser Zalmout: Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Journal Language Resources and Evaluation. Pages(745-775). Volume(51).