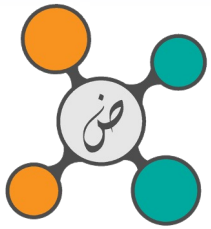
LREC-COLING 2024,
Torino, Italy.
20-25 May 2024.



Qabas : An Open-Source Arabic Lexicographic Database

Mustafa Jarrar **Tymaa Hammouda**

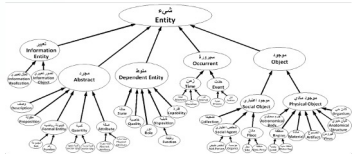
Birzeit University
Palestine

Lexical Resources at SinaLab - Birzeit University

Lexicographic Database

150 lexicons
Largest Arabic lexicographic database

Arabic Ontology/ Wordnet



Formal Arabic Wordnet
with ontologically clean content

Annotated Corpora

Currasat
كراسات
Arabic Dialects Corpora
مدونة اللهجات العربية

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

NLP library

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

Synonyms 90s% WSD 84% NER 92s% Intent 88.4% Offensive 87% .....

Big Linguistic Data Graph

https://sina.birzeit.edu

developed in

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

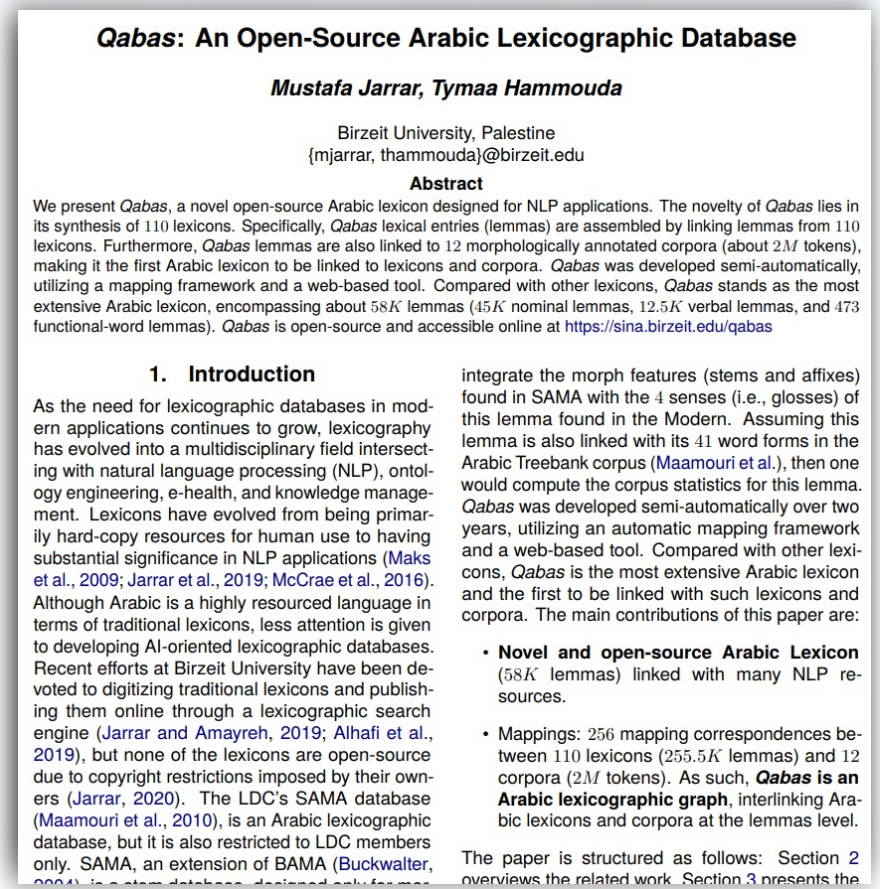
News Team Resources

Resources

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

+ Lexicographic Database (150 lexicons)	حوسبة المعاجم
+ Arabic Ontology	الأنطولوجيا العربية
+ Dialect Corpora (Currasat)	كراسات مدونة العاميات
+ Arabic Synonyms	استخراج مترادفات
+ 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	

BIRZEIT UNIVERSITY



Qabas: An Open-Source Arabic Lexicographic Database

Mustafa Jarrar, Tymaa Hammouda

Birzeit University, Palestine
{mjarrar, thammouda}@birzeit.edu

Abstract

We present *Qabas*, a novel open-source Arabic lexicon designed for NLP applications. The novelty of *Qabas* lies in its synthesis of 110 lexicons. Specifically, *Qabas* lexical entries (lemmas) are assembled by linking lemmas from 110 lexicons. Furthermore, *Qabas* lemmas are also linked to 12 morphologically annotated corpora (about $2M$ tokens), making it the first Arabic lexicon to be linked to lexicons and corpora. *Qabas* was developed semi-automatically, utilizing a mapping framework and a web-based tool. Compared with other lexicons, *Qabas* stands as the most extensive Arabic lexicon, encompassing about $58K$ lemmas ($45K$ nominal lemmas, $12.5K$ verbal lemmas, and 473 functional-word lemmas). *Qabas* is open-source and accessible online at https://sina.birzeit.edu/qabas

1. Introduction

As the need for lexicographic databases in modern applications continues to grow, lexicography has evolved into a multidisciplinary field intersecting with natural language processing (NLP), ontology engineering, e-health, and knowledge management. Lexicons have evolved from being primarily hard-copy resources for human use to having substantial significance in NLP applications (Maks et al., 2009; Jarrar et al., 2019; McCrae et al., 2016). Although Arabic is a highly resourced language in terms of traditional lexicons, less attention is given to developing AI-oriented lexicographic databases. Recent efforts at Birzeit University have been devoted to digitizing traditional lexicons and publishing them online through a lexicographic search engine (Jarrar and Amayreh, 2019; Alhafi et al., 2019), but none of the lexicons are open-source due to copyright restrictions imposed by their owners (Jarrar, 2020). The LDC's SAMA database (Maamouri et al., 2010), is an Arabic lexicographic database, but it is also restricted to LDC members only. SAMA, an extension of BAMA (Buckwalter,

integrate the morph features (stems and affixes) found in SAMA with the 4 senses (i.e., glosses) of this lemma found in the Modern. Assuming this lemma is also linked with its 41 word forms in the Arabic Treebank corpus (Maamouri et al.), then one would compute the corpus statistics for this lemma. *Qabas* was developed semi-automatically over two years, utilizing an automatic mapping framework and a web-based tool. Compared with other lexicons, *Qabas* is the most extensive Arabic lexicon and the first to be linked with such lexicons and corpora. The main contributions of this paper are:

- **Novel and open-source Arabic Lexicon** ($58K$ lemmas) linked with many NLP resources.
- Mappings: 256 mapping correspondences between 110 lexicons ($255.5K$ lemmas) and 12 corpora ($2M$ tokens). As such, ***Qabas* is an Arabic lexicographic graph**, interlinking Arabic lexicons and corpora at the lemmas level.

The paper is structured as follows: Section 2 overviews the related work. Section 3 presents the

Jarrar, M., Hammouda, T.: Qabas: An Open-Source Arabic Lexicographic Database. In Proceedings of the International Conference on Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy. 2024.

The problem

No open-source lexicographic databases for Arabic NLP.

Our Contributions

- The first Open-Source Arabic Lexicographic Database (58K lemmas).

  أول معجم **حاسوبي** للغة العربية (مفتوح المصدر)

- Linked with 110 lexicons (255.5K lemmas)

  جميع مدخلاته مربوطة مع مداخلات ١١٠ معاجم عربية

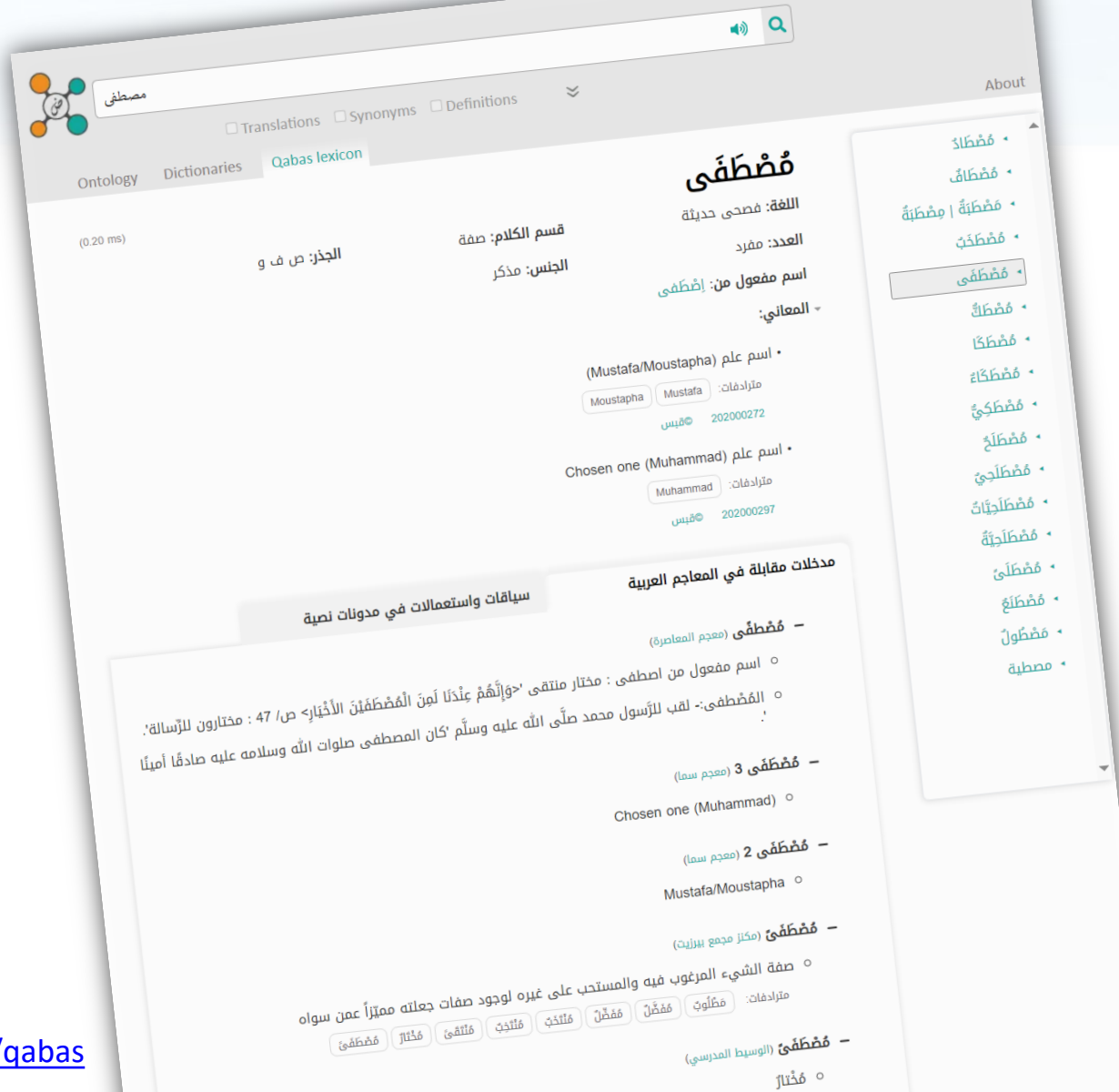
- Linked with 12 corpora (2M tokens).

  تم ربطه مع مدونات نصية (٢ مليون كلمة)

Big Linguistic Data Graph

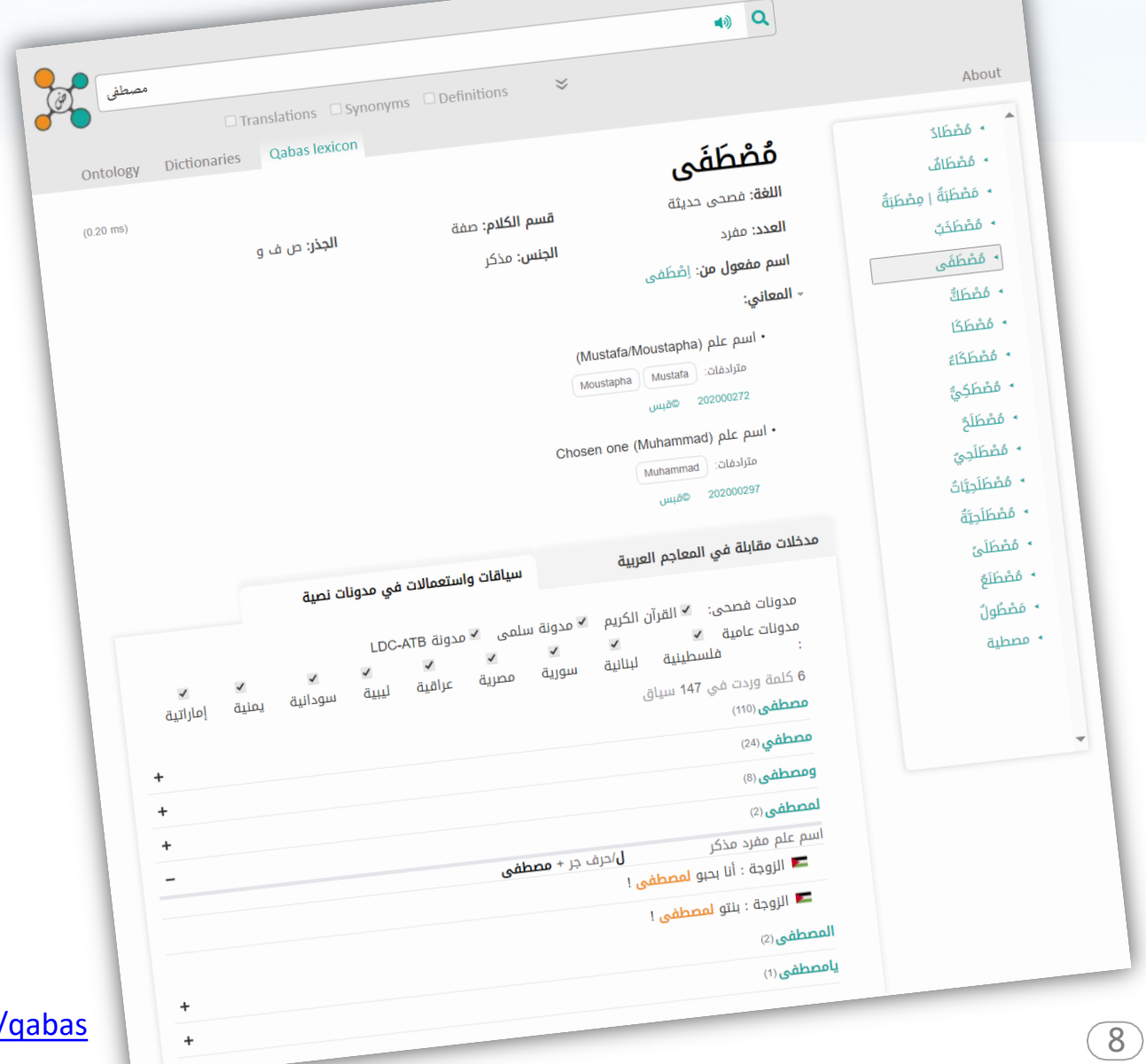
شبكة بيانات لغوية ضخمة للغة العربية

Qabas Webpage



https://sina.birzeit.edu/qabas

Qabas Webpage



https://sina.birzeit.edu/qabas

Qabas Statistics (so far)

POS category	POS	*Qabas*
Nominal	NOUN اسم	29,053
	NOUN_PROP اسم علم	4,319
	ADJ صفة	11,067
	ADJ_COMP صفة مقارنة	295
	ADJ_NUM صفة عدد	12
	NOUN_NUM اسم عدد	44
	NOUN_QUANT اسم كم	19
	DIGIT عدد	10
	NOUN_VOICE صوت	16
	ABBREV حرف اختصار	106
	Total	**44,941**
Verb	PV ماضي	12,679
	IV مضارع	9
	CV أمر	6
	PV_PASS ماضي مجهول	63
	IV_PASS مضارع مجهول	
	Total	**12,757**
Functional words	PRON, DEM_PRON, EMOJI REL_PRON, REL_ADV, ADV, INTERROG_PART, INTERROG_ADV,PREP,CONJ, INTERROG_PRON, PART RESTRIC_PART,PUNC,INTERJ, FOCUS_PART, DET, VERB VOC_PART, PROG_PART, SUB_CONJ, VERB_PART, FUT_PART,EXCLAM_PRON PSEUDO_VERB,NEG_PART	473
	Total	**58,171**

Qabas Lexicon

Data Sources used to build Qabas:

From: 110 Arabic lexicons, and 12 corpora.

110 Arabic lexicons

Lexicon	Unique Lemmas	Lemmas mapped
SAMA	40,639	$40,330^{99\%}$
MODERN	32,300	$32,276^{100\%}$
Ghani	29,854	$24,452^{82\%}$
Al-Waseet	36,632	$17,829^{49\%}$
Al-Waseet Madrasi	7,649	$7,384^{97\%}$
Thesuri$_{(7)}$	15,236	$12,892^{85\%}$
ArabicOntology&Lexicons	28,435	$24,864^{87\%}$
ArabicWordNet	10,929	$9,578^{88\%}$
ALCSO Unified$_{(40)}$	40,861	$38,876^{95\%}$
Arab Academies$_{(16)}$	9,675	$7,597^{79\%}$
Others$_{(37)}$	45,398	$34,785^{77\%}$
Wikidata	–	4665^{--}
Total^{110}	**297,608**	**$255,528^{84\%}$**

12 corpora

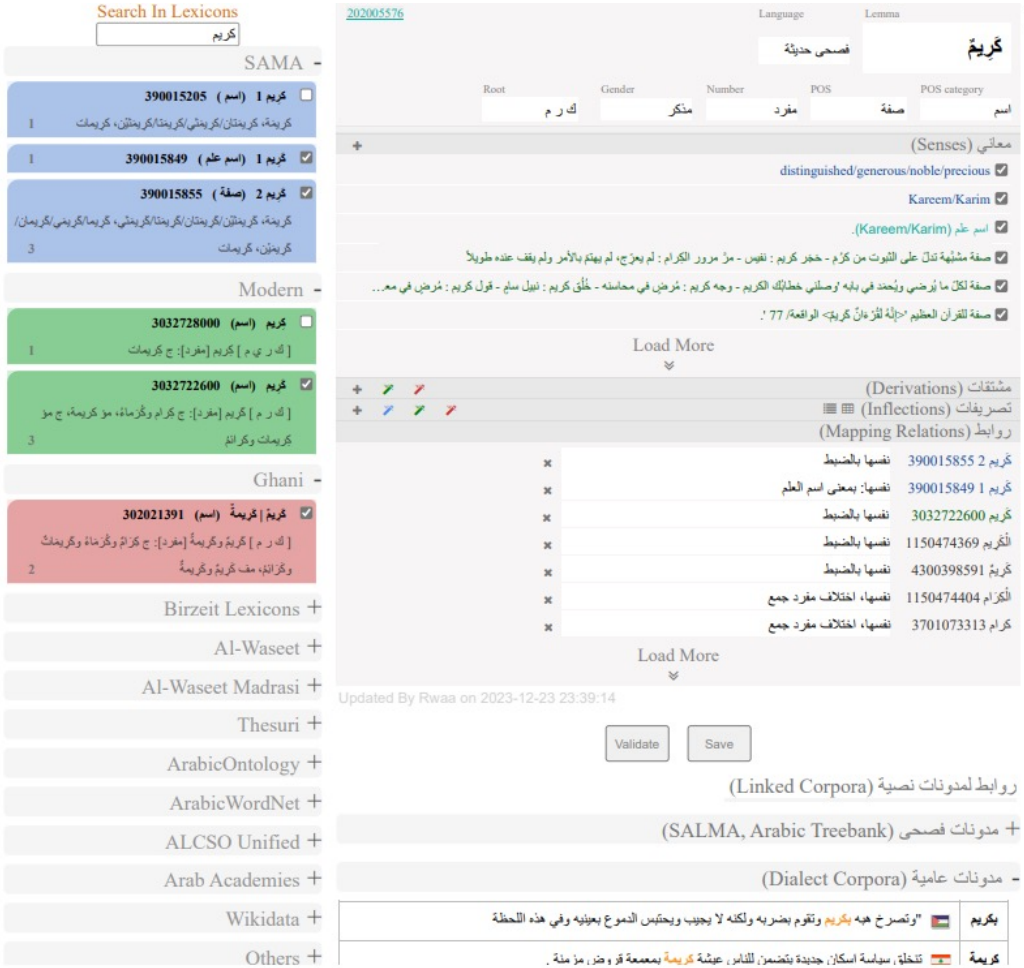
Corpus	Tokens	Tokens mapped	Unique lemmas	Lemmas mapped
Arabic Treebank (MSA)	339,710	$282,155^{83\%}$	13,078	$12,948^{99\%}$
SALMA (MSA)	34,253	$34,253^{100\%}$	3,875	$3,875^{100\%}$
Quran (Classical)	77,469	$62,123^{80\%}$	4,830	$4,100^{84\%}$
Curras (Palestinian)	56,169	$56,010^{100\%}$	6,033	$5,966^{99\%}$
Baladi(Lebanese)	9,561	$9,493^{99\%}$	2,406	$2,365^{98\%}$
Lisan (Iraqi)	45,881	$40,615^{89\%}$	9,306	$7,520^{81\%}$
Lisan (Lybian)	51,686	$39,508^{76\%}$	10,174	$7,550^{74\%}$
Lisan (Sudanese)	52,616	$44,136^{84\%}$	10,455	$8,709^{83\%}$
Lisan (Yemeni)	1,098,222	$901,335^{82\%}$	44,331	$33,244^{75\%}$
Gummar (Emirati)	202,329	$182,155^{90\%}$	7,590	$6,800^{90\%}$
Nabra (Syrian)	60,021	$60,021^{100\%}$	10,191	$10,191^{100\%}$
Egyptian Treebank	400,448	$297,188^{74\%}$	22,258	$18,626^{83\%}$
Total	**2,428,365**	**2,008,992$^{83\%}$**	**144,527**	**121,894 $^{84\%}$**

Lexicon Construction methodology (phases)

Mapping framework:

Annotate the lemma “كريم”

Using a web-based tool



Qabas Guidelines

- Each lemma in Qabas is tagged with several morphological features: (Language, root(s), 41 POS tags, gender, number, aspect, person, augmentation, transitivity).

- Guidelines for lemma selection and spelling can be found online (الدليل المعياري وقواعد ضبط المدخلات) :

  https://sina.birzeit.edu/qabas/about

Mapping Relations Guidelines

Mapping Framework:

$<l_1, l_2, R_i>$

Where:

l_1, l_2 are lemmas to be mapped.

R_i is the mapping relation between l_1 and l_2

Mapping Relations:

The six mapping relations used by annotators

Relations			count
R_1	نفسها بالضبط	Same Exactly	248,882
R_2	نفسها، اختلاف مفرد جمع	Same, Singular-Plural difference	3,010
R_3	نفسها، اختلاف مفرد مثنى	Same, Singular-Dual difference	74
R_4	نفسها، اختلاف مذكر مؤنث	Same, Male-Female difference	1,784
R_5	نفسها، اختلاف حالة إعرابية	Same, Case difference	372
R_6	نفسها، بمعنى اسم العلم	Same, but Proper Noun	1,918
Total (mapping correspondences)			**256,040**

Qabas Evaluation

Coverage:

Coverage Evaluation of Qabas, per POS compared with SAMA and Modern

POS category	POS	Modern	SAMA	*Qabas*
Nominal	NOUN اسم	21,456	19,705	29,053
	NOUN_PROP اسم علم		5,540	4,319
	ADJ صفة		5,500	11,067
	ADJ_COMP صفة مقارنة		204	295
	ADJ_NUM صفة عدد		12	12
	NOUN_NUM اسم عدد		33	44
	NOUN_QUANT اسم كم		23	19
	DIGIT عدد			10
	NOUN_VOICE صوت			16
	ABBREV حرف اختصار		60	106
	Total	**21,456**	**31,077**	**44,941**
Verb	PV ماضي	10,475	8,133	12,679
	IV مضارع		990	9
	CV أمر		16	6
	PV_PASS ماضي مجهول		32	63
	IV_PASS مضارع مجهول		78	
	Total	**10,475**	**9,249**	**12,757**
Functional words	PRON, DEM_PRON, EMOJI REL_PRON, REL_ADV, ADV, INTERROG_PART, INTERROG_ADV,PREP,CONJ, INTERROG_PRON, PART RESTRIC_PART,PUNC,INTERJ, FOCUS_PART, DET, VERB VOC_PART, PROG_PART, SUB_CONJ, VERB_PART, FUT_PART,EXCLAM_PRON PSEUDO_VERB,NEG_PART	369	313	473
	Total	**32,300**	**40,639**	**58,171**

Qabas Evaluation

Inter-Annotator Agreement:

- Randomly selected 2850 lemmas (5% of Qabas) and asked each of the three lexicographers (A1, A2, A3) to map them manually.
- Kappa coefficient κ =

  A1-A2 is 85%

  A2-A3 is 88%

  A1-A3 is 86%

  which are “almost perfect”

Acknowledgment

We would like to thank the main lexicographers who contributed to this project, especially **Shimaa Hamayel**, **Hiba Zayed**, and **Rwaa Assi**;

as well as Diyam Akra, Sanad Malaysha, Sondus Hamad, Asmaa Motan, Yaqout Abu Allia, Nour Dana, who also contributed to various lexicographic and technical aspects.

Thank You

References

1. Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia: NLU-STR at SemEval-2024 Task 1: Generative-based Augmentation and Encoder-based Scoring for Semantic Textual Relatedness In Proceedings of the SemEval 2024 Shared Task 1 (Semantic Relatedness). 2024.
2. 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.
3. 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.
4. 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.
5. Haneen Liqreina, 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.
6. Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim El-madany, Nagham Hamad, Alaa' Omar: WojoodNER 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.
7. 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
8. 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
9. 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
10. 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
11. Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem: Wojood: 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
12. 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
13. 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
14. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
15. 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
16. 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
17. 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
18. 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
19. 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
20. Mustafa Jarrar: Digitization of Arabic Lexicons. Arabic Language Status Report. UAE Ministry of Culture and Youth. Pages 214-2017. Dec 2020
21. 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
22. 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
23. 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
24. 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
25. 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
26. 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
27. Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, Nasser Zalmout: Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Journal Language Resources and Evaluation. Pages(745-775). Volume(51), Issue(3). Springer (doi.org/10.1007/s10579-016-9370-7). 2017
28. Mamoun Abu Helou, Matteo Palmonari, Mustafa Jarrar: Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping. Journal of Artificial Intelligence Research, Special Track on Cross-language Algorithms and Applications. Pages(165-208). Volume(55), Number(1). AI Access Foundation. 2016
29. Mustafa Jarrar, Nizar Habash, Diyam Akra, Nasser Zalmout: Building a Corpus for Palestinian Arabic: a Preliminary Study. Arabic Natural Language Processing Workshop, at the Conference on Empirical Methods in Natural Language Processing