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, 2004), is a stem database, designed only for morphological modeling. It contains stems and their lemmas and compatible affixes.

This article proposes *Qabas*, a novel open-source Arabic lexicon designed for NLP applications. The novelty of *Qabas* lies in its synthesis of many lexical resources. Each lexical entry (i.e., lemma) in *Qabas* is linked with equivalent lemmas in 110 lexicons, and with 12 morphologically-annotated corpora (about $2M$ tokens). This linking was done through $256K$ mappings correspondences (as shown in Table 3). That is, the philosophy of *Qabas* is to construct a large lexicographic data graph by linking existing Arabic lexicons and annotated corpora. This enables the integration and reuse of these resources for NLP tasks. For example, by linking the lemma (كريم٢/*kariym2*) in SAMA with (كريم/*kariym*) in the Modern lexicon, one would 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 methodology, and Section 4 presents lemma mapping. In Section 5 we evaluate the coverage; and in Section 6 we summarize our conclusions.

2. Related Work

In recent years, many standardization efforts have been proposed for representing, publishing, and linking linguistic resources. For example, the W3C's Lemon RDF model (Philipp Cimiano, 2016) enables employing lexicons in ontologies and various NLP applications. Moreover, the Linguistic Linked Open Data Cloud (LLOD) (McCrae et al., 2016) used Lemon to interlink the lexical entries of several linguistic resources. The ISO's Lexical Markup Framework (LMF) standard aims at representing lexicons in a machine-readable format (Francopoulo et al., 2006).

LREC-COLING 2024, pages 13363–13370

Different encyclopedic dictionaries integrated WordNets with other resources, such as BabelNet (Navigli et al., 2012) and ConceptNet 5.5 (Speer et al., 2017). Compared with our work, we provide an interlinking of many lexicons and corpora, forming a lexicographic, rather than an encyclopedic graph. Given that digitized and available Arabic lexicons are limited, there are several attempts to digitize and represent them in the standard formats. The first attempt to represent Arabic lexicons in ISO LMF standard can be found in (Salmon-Alt et al., 2005; Maks et al., 2009; Khemakhem et al., 2016). Other attempts suggested using the W3C Lemon RDF model (Khalfi et al., 2016; Jarrar et al., 2019). While several online portals for Arabic lexicographic search exist (e.g., lisaan.net, almaany.com, almougem.com), each portal contains a limited number of lexicons, and their content is partially structured (i.e., available in flat text). *Qabas* is developed as a synthesis of 110 lexicons that we digitized earlier (Jarrar and Amayreh, 2019).

3. Methodology

3.1. Scope and Objectives

The objective of *Qabas* is to link existing Arabic lexicons and corpora and enable them to be integrated and re-used in NLP tasks (Darwish et al., 2021). In other words, *Qabas* lemmas are used as a proxy to link between different resources, forming a large Arabic lexicographic data graph. Thus, all *Qabas* lemmas are collected mainly from these resources (Section 3.2). As such, *Qabas* is designed to be an open-source and open-ended project, targeting all forms of Arabic: Classical Arabic, Modern Standard Arabic, Arabic dialects, and foreign words that are transliterated and commonly used in Arabic.

In this paper, we focus on including the morphological features for each lemma, such as the spelling(s) of the lemma, its root(s), POS, gender, number, person, and voice. Including semantic information (e.g., glosses, synonyms, relations, and translations) is not discussed in this article due to space limitations. Nevertheless, it is worth noting that based on *Qabas* mappings, (i) we developed a synonym extraction tool[1] (Ghanem et al., 2023); (ii) we extracted glosses and contexts from these mapped lexicons to build a large set of context-gloss pairs for Word-Sense Disambiguation (Al-Hajj and Jarrar, 2021; Malaysha et al., 2023); and (iii) a graph representing morpho-semantic relationships in Arabic was extracted based on Arabic derivational morphology, see Figure 4 in (Jarrar, 2021).

3.2. Data Sources

Among the 150 lexicons that we previously digitized (Jarrar and Amayreh, 2019), 110 lexicons and 12 morphologically annotated corpora were prepared to be linked and to construct *Qabas*. See our copyright notice in section 6.2 regarding the collected resources and the sharing of *Qabas*.

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**$^{84\%}$

Table 1: List of lexicons mapped with *Qabas* so far.

Table 1 categorizes the 110 lexicons into: the LDC's SAMA (Maamouri et al., 2010), Modern lexicon (Omar, 2008), Ghani lexicon (Abul-Azm, 2014), the Al-Waseet lexicon (Cairo, 2004), the Al-Waseet Madrasi lexicon, the Arabic Ontology and two lexicons (Jarrar, 2021, 2011), the Arabic WordNet (Black et al., 2006), 40 of the ALECSO's Unified dictionaries. We also collected 16 lexicons produced by the Arabic Language Academies in Cairo and in Damascus (Cairo; Damascus), the Arabic Wikdata, in addition to 7 thesauri and 37 Other lexicons.

As we are concerned with linking the lexical entries (i.e., lemmas) in these resources, each distinct lemma is given a unique identifier. In addition, we are only concerned with linking single-word lemmas, thus multi-word lemmas are ignored at this phase, such as (ثاني أكسيد الكربون، سرعة الضوء). The total number of single-word lemmas in the 110 lexicons is about $297K$ lemmas, about $255K$ (84%) of which are mapped (See Table 1).

As shown in Table 2, we collected 12 Arabic corpora, especially those that are annotated with morphological features: the MSA LDC's Arabic Treebank (Maamouri et al.), the MSA SALMA corpus (Jarrar et al., 2023a), the Quran corpus (Dukes and Habash, 2010), the Palestinian Curras and the Lebanese Baladi corpora (Haff et al., 2022), the Lisan (Iraqi, Lybian, Sudanese, and Yemeni) corpora (Jarrar et al., 2023b), The Emirati Gummar corpus (Khalifa et al., 2018), the Syrian Nabra corpus(Nayouf et al., 2023), and the LDC's Egyptian Treebank (Maamouri et al., 2021). These corpora compass $2.4M$ tokens annotated with about $144.5K$ lemmas, 84% of which are mapped with *Qabas*; i.e., *Qabas* is linked with about $2M$ tokens.

3.3. Lexicon Construction Phases

Qabas was constructed semi-automatically over different phases, and using a web-based tool (illustrated in Figure 1).

[1] https://sina.birzeit.edu/synonyms. It can be also used to evaluate synonyms (Khallaf et al., 2023).

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%**

Table 2: List of corpora linked with *Qabas* so far.

To bootstrap *Qabas*, we first adopted all lemmas from the Modern lexicon and uploaded them to the tool. Three lexicographers then reviewed and manually revised and enriched these lemmas with morphological features (described in Section 3.4) and linked them with lemmas in other lexicons. This methodology allowed the lexicographers to construct *Qabas* based on the information in other lexicons while linking *Qabas* to those lexicons at the same time (see guidelines in Section 3.4). To accelerate the linking process, we used heuristic rules to automatically discover candidate mappings for the lexicographers to verify (see Section 4.2).

To cover the remaining lemmas in lexicons other than Modern (i.e., that are not linked in the previous phase), we collected these lemmas and prioritized them. Higher priority is given to those lemmas that are more frequent across the 110 lexicons and 12 corpora. This prioritized list of candidate lemmas was uploaded to the tool, for the lexicographers to review and make the necessary edits. This approach allowed us to efficiently expand the lemma coverage of *Qabas*. The expansion is an ongoing and open-ended endeavor, as there is no limit to the number of lemmas that could potentially be added to *Qabas*. As will be discussed in section 5, our progress indicates that we have covered most of the lemmas in the 110 lexicons and 12 corpora.

Mapping *Qabas* with the 12 corpora (in table 2) was straightforward. As most of the lemmas in these corpora are SAMA lemmas, which we manually linked with *Qabas*, we only replaced SAMA lemmaIDs with *Qabas* lemmaIDs. For the non-SAMA lemmas, we selected the most frequent lemmas in the 12 corpora and added them to *Qabas* manually.

3.4. Guidelines

Each lemma in *Qabas* is tagged with the following eight morphological features: (1) the 41 POS tagset shown in Table 4, (2) the gender tags {$Masculine$, $Feminine$, N/A}, (3) Number tags {$Singular$, $Dual$, $Plural$}, (4) the Aspect tags {PV, IV, CV, PV_PASS, IV_PASS}, (5) and Person tags {1^{st}, 2^{nd}, 3^{rd}}. We additionally tag each lemma with its (6) root(s), (7) augmentation {$Augmented$, $Unaugmented$}, and (8) transitivity {$Transitive$, $Intransitive$}.

Lemma selection and spelling, our full list of guidelines not included in this article for space limitation but can be found online[2]. Our guidelines are similar to those described in the introduction of the Modern (Omar, 2008). However, we introduced additional guidelines, such as: the lemma should be fully diacritized including the last letter; the POS of a lemma can be $Noun\_Prop$ only if all of its meanings refer to proper nouns; additional spellings of the same lemma are separated by "|" and ordered by frequency, such as (تِلِيفُون|تِلِفُون); dialectal lemmas are spelled according to the CODA rules used in Curras (Jarrar et al., 2017, 2014), hence we write (قَزاز/*qazāz*) rather than (أزاز/*ʾazāz*); each dialect lemma is mapped with an MSA lemma, e.g. (قَزاز/*qazāz*) and its MSA (زُجاج/*zuǧāǧ*); a lemma is considered $adjective$ if all of its meanings are either $ActiveParticiple$ اسم فاعل, $PassiveParticiple$ اسم مفعول, $Relative$ نسبة, $AdjectivalPropriety$ صفه مشبهة, $Exaggeration$ صيغة مبالغة, or $Diminutive$ تصغير ; among other guidelines.

4. Lemma Linking

This section presents the framework and methods we used to map between lemmas across lexicons.

4.1. Mapping Framework

This framework aims to enable lemmas to be interlinked through a mapping correspondence.

[2]Guidelines https://sina.birzeit.edu/qabas/about

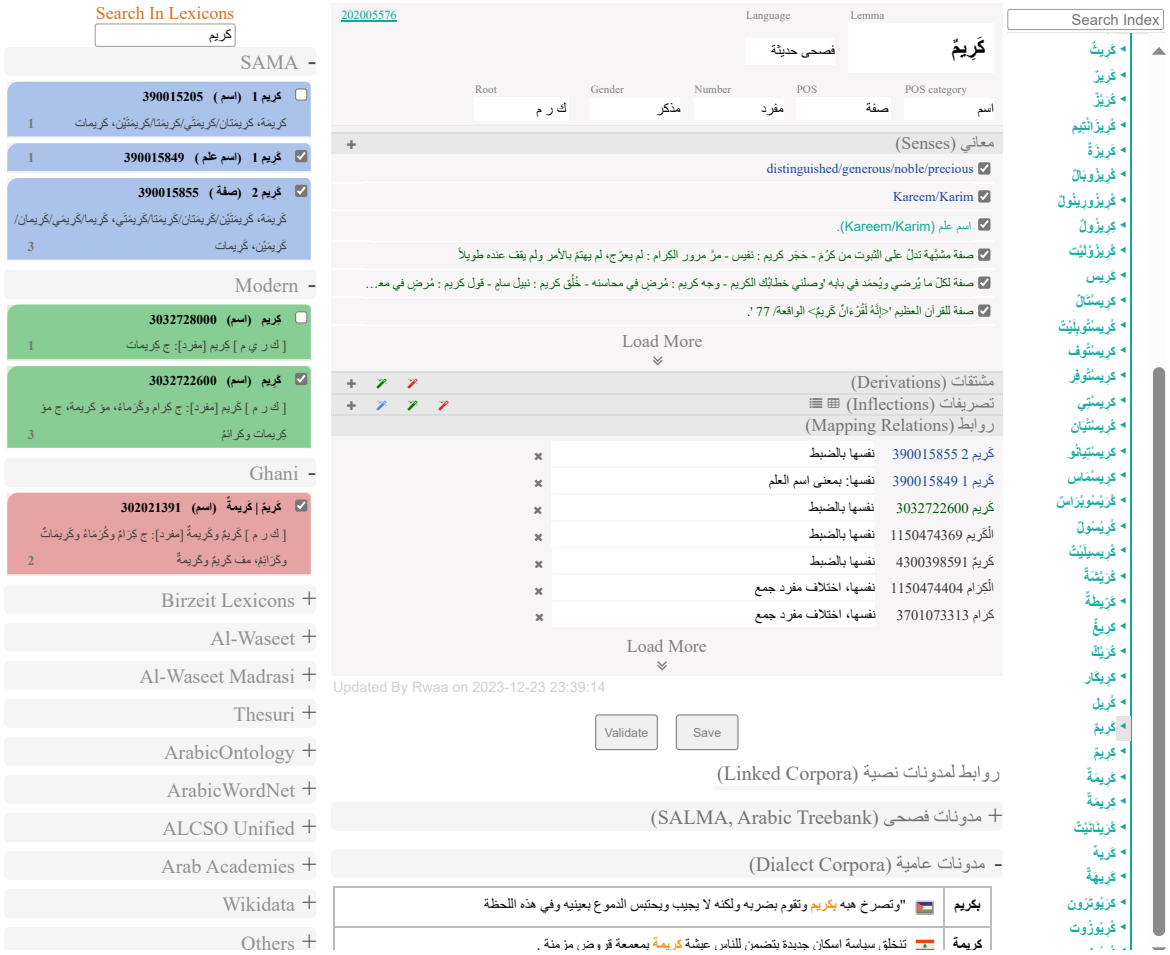

Figure 1: Screenshot of our web-based tool, which we developed for constructing *Qabas*

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**

Table 3: The six mapping relations and their counts

Definition 1**:** Given two lemmas l_1 and l_2, a *mapping correspondence* between them is defined as:

$$<l_1, l_2, R_i>$$

Where:

- l_1, l_2 are lemmas to be mapped.
- R_i is the mapping relation between l_1 and l_2, $R_i \in \{R_1 ... R_6\}$ shown in Table 3.

This mapping framework was implemented in our tool (See Figure 1) and used by our lexicographers. Table 3 presents the count of the mapping correspondences for each relation, which are about $256K$ correspondences in total.

4.2. Automatic Mapping

To speed up the mapping process, this section proposes a set of heuristic rules to discover candidate mappings. Before presenting these rules, we discuss how Arabic word forms can be compared.

Comparing words in Arabic is not trivial. First, Arabic is diacritic-sensitive, thus we cannot compare words using equality. For example, the same lemma in one lexicon might be spelled as كَلِمَة/*kalmah* and in another as كَلِمَةٌ/*klamtun*. Second, lexicons are not always self-consistent or follow the same guidelines in structuring or writing word forms (Amayreh et al., 2019). For example, some lexicons provide the feminine and masculine forms of a perfect verb {يَكتب/*yaktb*, تَكتُب/*taktub*}, while others provide one {يكتُب/*yktub*} or none {}. To overcome these challenges, when comparing word forms, we implemented the following definitions of *compatibility* - as explained in (Jarrar et al., 2018).

Definition 2: Given two words w_1 and w_2, we consider them *diacritic-compatible*, *iff*: (1) both words have the same letters, and (2) no contradictions between the diacritics of the same, pair-wise, letters of these words.

Definition 3: Given two sets of words W_1 and W_2, we consider these sets *compatible*, *iff* there exists a diacritic-compatible word w in both sets, $w \in W_1$

and $w \in W_2$, i.e., their intersection is not empty. The mapping heuristic rules are:

- h_1**:** A mapping correspondence is established between two verb lemmas if the following two conditions are true: (i) each lemma has a perfective form(s) PV and these forms are compatible, and (ii) if each lemma has root(s), imperfect form(s) IV and command form(s) CV, and these roots, IVs, and CVs are compatible. **Example**: (i) PV_1={كَتَبَ} and PV_2={كَتَبَ} which are compatible, and (ii) IV_1={أكتب, يكتب} and IV_2={يَكْتُبُ}, CV_1={} and CV_2={أكتب}, and $root_1$={ك ت ب} and $root_2$={}, which are all compatible.
- h_2**:** A mapping correspondence is established between two noun lemmas if the following two conditions are true: (i) each lemma has a singular form(s) and these forms are compatible, and (ii) if each lemma has root(s), dual(s) and plural(s), and these root(s), dual(s), and plural(s) are compatible.

With these heuristics, we were able to discover $179K$ candidate mapping correspondences. We then uploaded these mapping relations to the tool and labeled them with "Auto-mapped". Lexicographers were given these mappings to confirm and assign them one of the six relations (See the relations division at the bottom of Figure 1). Lexicographers can edit these relations and search the lexicons to include more mappings if needed.

5. Evaluation and Discussion

We evaluate the coverage of *Qabas* by comparing it with two resources: SAMA and Modern, which are well-developed resources for Arabic. SAMA is designed for morphological modeling, while Modern is a typical MSA lexicon focusing on semantics. Table 4 shows that *Qabas*'s coverage is almost double of Modern and is 40% larger than SAMA. Table 1 also shows that *Qabas* contains all Modern lemmas and 99% of SAMA lemmas. We did not add the 1% as we found them to be typos or with redundant spellings. Another critical issue in SAMA is that it treats each proper noun as a separate lemma (e.g., كَرِيم ١/*kariym1* as a proper noun and كَرِيم ٢/*kariym2* as adjective). We believe that this is problematic because most Arabic words can be used as proper nouns (Jarrar et al., 2022). Proper nouns in *Qabas* are considered as such only if all meanings denote proper nouns. Thus, the lemma كَرِيم/*kariym* would be tagged with an adjective, and one of its meanings is a proper noun. Hence, most of the $5,540$ proper nouns in SAMA are merged and mapped with *Qabas* lemmas through the R_6 relations.

An Inter-Annotator Agreement (IAA) evaluation was conducted to evaluate the lemma mappings. We randomly selected 2850 lemmas (5% of *Qabas*) and asked each of the three lexicographers (A_1, A_2, A_3) to map them. The IAAs using the Kappa coefficient κ are: A_1-A_2 is 85%, A_2-A_3 is 88%, and A_1-A_3 is 86%, which are "almost perfect" (Viera and Garrett, 2005).

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**

Table 4: Coverage Evaluation of *Qabas*, per POS

6. Conclusion

We presented *Qabas*, a novel and open-source Arabic lexicon linked with 110 lexicons and 12 morphologically annotated corpora. Additionally, the $256k$ mappings correspondences between *Qabas* and each of the 110 lexicons can be also downloaded from Qabas Page. As such, *Qabas* is a large lexicographic data graph, linking existing Arabic lexicons and annotated corpora.

6.1. Limitations and Future Work

One of the major challenges faced during the construction of *Qabas* was convincing the owners of the lexicons to publish their lexicons as open-source. While we agreed with the owners of the lexicons to only publish the mapping links between *Qabas* and their lexicons, we hope that our work will encourage others to publish their lexicons as open-source in the future. Adding dialect lemmas to *Qabas* is another challenge. Since our three lexicographers are familiar with Levantine dialects, adding lemmas from other dialects requires knowledge of these dialects. *Qabas* is currently limited to the frequently used dialect lemmas or those that are known to most Arabs. We plan to recruit more lexicographers from other dialects to extend *Qabas*. Last but not least, we plan to represent *Qabas* and publish the mapping correspondences using the W3C RDF Lemon model.

6.2. Ethical and copyright Considerations

We obtained permission to use the lexicons and corpora listed in this article, and since our lexicon will be open-source, we will not share any copyrighted data. We will share: (1) *Qabas* itself (all lemmas and their full morphological features), and (2) the mapping links (i.e., correspondences) between *Qabas* and the other external resources. Obtaining licenses for these external resources is the responsibility of the users.

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.

7. References

Abdul-Ghani Abul-Azm. 2014. *Al-Ghani Al-Zaher Dictionary*. Rabat: Al-Ghani Publishing Institution.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. 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)*, pages 40–48, Online. INCOMA Ltd.

Arabization Coordination Bureau (Rabat) ALECSO. http://www.arabization.org.ma/.

Diana Alhafi, Anton Deik, and Mustafa Jarrar. 2019. Usability evaluation of lexicographic e-services. In *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEE.

Hamzeh Amayreh, Mohammad Dwaikat, and Mustafa Jarrar. 2019. Lexicon digitization-a framework for structuring, normalizing and cleaning lexical entries. *Technical Report, Birzeit University*.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Christiane Fellbaum, et al. 2006. Introducing the arabic wordnet project. In *Proceedings of the third international WordNet conference*, pages 295–300. Jeju Korea.

Tim Buckwalter. 2004. Buckwalter arabic morphological analyzer (bama) version 2.0. linguistic data consortium (ldc) catalogue number ldc2004l02. Technical report, ISBN1-58563-324-0.

Arabic Language Academy Cairo. 2004. *Al-Waseet Dictionary*. Shorouk International Bookshop.

Arabic Language Academy in Cairo. https://www.arabicacademy.gov.eg/.

Arabic Language Academy in Damascus. http://arabacademy.gov.sy/.

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, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of quranic arabic. In *Lrec*, pages 2530–2536.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pages 215–222. Global Wordnet Association.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.

Mustafa Jarrar. 2020. *Digitization of Arabic Lexicons*, pages 214–217. UAE Ministry of Culture and Youth.

Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.

Mustafa Jarrar, Hamzeh Amayreh, and John P. McCrae. 2019. Representing arabic lexicons in lemon - a preliminary study. In *The 2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014, Workshop on Arabic Natural Language*, pages 18–27. Association For Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. 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.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammad Khalilia. 2023a. 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.

Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. *ACM Asian and Low-Resource Language Information Processing*, 18(2):10:1–10:21.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023b. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Mustapha Khalfi, Ouafae Nahli, and Arsalane Zarghili. 2016. Classical dictionary al-qamus in lemon. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 325–330. IEEE.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, and Tymaa Hammoudaand Mustafa Jarrar. 2023. Open-source thesaurus development for under-resourced languages: a welsh case study.

Aïda Khemakhem, Bilel Gargouri, Abdelmajid Ben Hamadou, and Gil Francopoulo. 2016. Iso standard modeling of a large arabic dictionary. *Natural Language Engineering*, 22(6):849–879.

Mohamed Maamouri, Ann Bies, Seth Kulick, Sondos Krouna, Dalila Tabassi, and Michael Ciul. 2021. Bolt egyptian arabic treebank - sms/chat.

Mohamed Maamouri, Ann Bies, Sondos Krouna Seth Kulick, Fatma Gaddeche, and Wajdi Zaghouani. Arabic treebank: Part 3 v 3.2.

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick. 2010. Ldc standard arabic morphological analyzer (sama) version 3.1. *LDC2010L01*.

Isa Maks, Carole Tiberius, and Remco van Veenendaal. 2009. Standardising bilingual lexical resources according to the lexicon markup framework. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sanad Malaysha, Mustafa Jarrar, and Mohammad Khalilia. 2023. Context-gloss augmentation for improving arabic target sense verification. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*. Global Wordnet Association.

John P McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard De Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, et al. 2016. The open linguistics working group: Developing the linguistic linked open data cloud.

Navigli, Roberto, Ponzetto, and Simone Paolo. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian arabic dialects with morphological annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Ahmed Mukhtar Omar. 2008. *Contemporary Arabic Dictionary.*, volume 14. World of Books, Cairo, Egypt.

Paul Buitelaar Philipp Cimiano, John P. McCrae. 2016. Lexicon model for ontologies. final community group report, 10 may 2016.

Susanne Salmon-Alt, Amine Akrout, and Laurent Romary. 2005. Proposals for a normalized representation of standard arabic full form lexica. In *International Conference on Machine Intelligence*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Anthony Viera and Joanne Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.