معجم قبس الحاسوبي

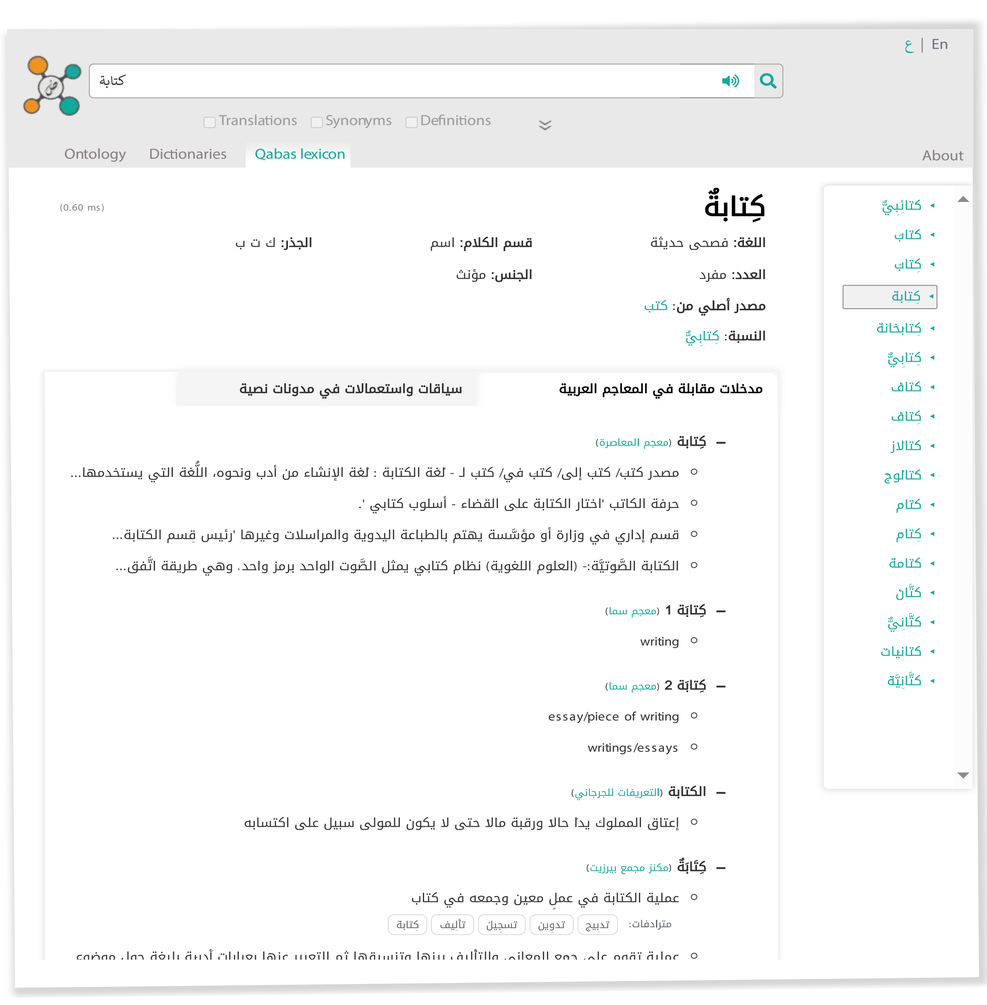
Qabas : An Open -Source Arabic Lexicographic Database

Mustafa Jarrar, Tymaa Hammouda
Birzeit University
Palestine

Contributions

- **Novel and open-source Lexicon** (58K lemmas) linked with many NLP resources.
- **Lexicographic data graph** interlinking 110 lexicons (256K lemmas) and 12 corpora (2M tokens) in MSA and dialects.

Demo and Mappings



110 Lexicons mapped with 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	28,435	24, 864 87 %
ArabicWordNet	10,929	9,578 88 %
ALECSO 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** $^{84\%}$

Mapping Framework

Mapping correspondence between lemmas l_1 and l_2:

$$<l_1, l_2, R_i>$$

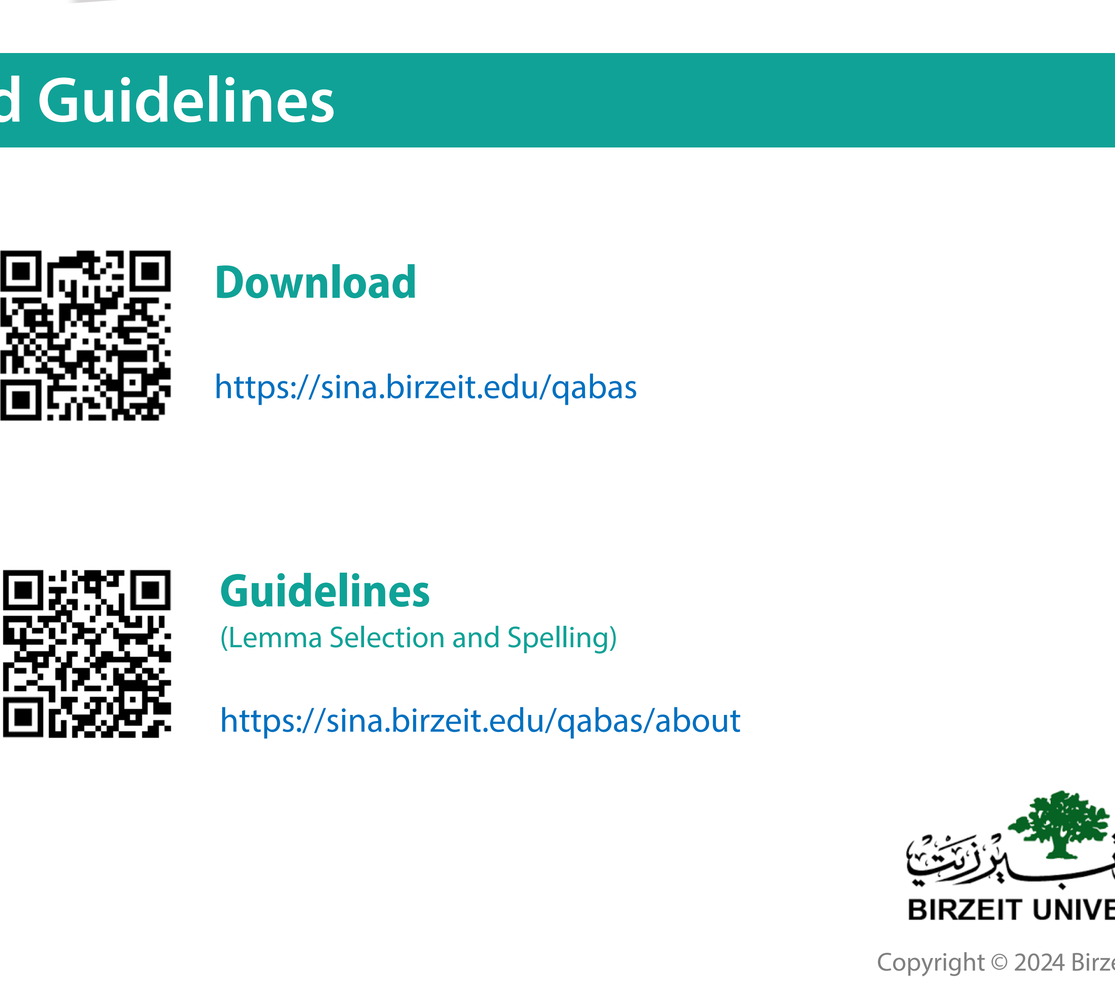
Where:
l_1 and l_2 are lemmas to be mapped
R_i is the mapping relation between l_1 and l_2, $R_i \in \{R_1 ... R_2\}$

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**

Count of the mapping correspondences for each relation

Corpora mapped with Qabas so far

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%



Coverage and Guidelines

Coverage Evaluation

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**

Download

https://sina.birzeit.edu/qabas

Guidelines

(Lemma Selection and Spelling)

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

BIRZEIT UNIVERSITY

Copyright © 2024 Birzeit University