Qabas - Arabic Lexicographic Database

Online: https://sina.birzeit.edu/qabas
Download, **Guidelines**, **and Statistics**: https://sina.birzeit.edu/qabas/about

Contact: SinaLab for Computational Linguistics and Artificial Intelligence
Prof. Mustafa Jarrar (sina@birzeit.edu)

Version

Version 1.1 (Last modification: 10-07-2024)

License:

See the attached (license.pdf) file

Data description:

1. **Qabas dictionary data:**
   The data is stored as a CSV file (Qabas-dataset.csv), with the following columns:
   - **lemma_id**: unique lemma public identifier. For example, for the lemma 202001621, its public page is: https://sina.birzeit.edu/qabas/lemma/202001621
   - **lemma**: the exact spelling(s) of lemma. Noticed that this column may contain spelling variations of the lemmas, separated by "|". The spelling is the most common (i.e., default).
   - **language**: the language of the lemma, which could be: [MSA (فصحى حديثة), Foreign (أجنبية), Dialect (عامية)]
   - **pos_cat**: The part-of-speech category, which could be: [اسم (Noun), فعل (Verb), كلمة وظيفية (Functional Word)].
   - **pos**: The part-of-speech tag, one of the 41 tags. Examples: ADJ, NOUN_PROP, PV, IV, DEM_PRON, etc.
   - **root**: The root(s) of the lemma. multiple are separated by "|".
   - **augmentation**: a morphological feature for verbs, indicating whether it is augmented or unaugmented [, مجرد مزيد].
   - **number**: the grammatical number, which can be singular, dual, plural [جمع , مثنى , مفرد].
   - **person**: the grammatical person, which can be 1st, 2nd, or 3rd person [غائب , مخاطب , متكلم].
   - **gender**: the gender, which can be Masculine, Feminine [مذكر , مؤنث].
   - **voice**: Indicates whether the lemma is in the active or passive voice [معلوم , مجهول].
   - **transitivity**: Indicates whether the verb lemma is transitive or intransitive [متعد , لازم].
   - **uninflected**: only for lemmas that are uninflected (ممنوعة من الصرف).

2. **Lemma Mappings** (Qabas-SAMA_mapping.csv)**:**
   A table that maps our Qabas lemmas to SAMA lemmas. It contains four columns: **qabas_lemma_id**, **qabas_lemma**, **sama_lemma_id**, and **sama_lemma**. This allows users to map the Qabas dictionary to SAMA lemmas using the CSV file.

Citation:

Qabas was developed based on the following articles, which should be acknowledged and cited properly each time Qabas is used:

[1] Mustafa Jarrar, Tymaa Hammouda: Qabas: An Open-Source Arabic Lexicographic Database. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 13363–13370, Torino, Italia. ELRA and ICCL.

[2] 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

[3] Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021