20th ACS/IEEE International Conference on Computer Systems and Applications
Giza, Egypt 4 December, 2023

# 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


جامعة بيرزيت
BIRZEIT UNIVERSITY
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

## 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 90s% Intent 88.4% Offensive 88.4% .....

## Big Linguistic Data Graph

https://ontology.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** | |

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

Mustafa Jarrar
*Birzeit University*
Birzeit, Palestine
mjarrar@birzeit.edu

Fadi A Zaraket
*American University of Beirut*
Beirut, Lebanon
fz11@aub.edu.lb

Tymaa Hammouda
*Birzeit University*
Birzeit, Palestine
1171779@student.birzeit.edu

Daanish Masood Alavi
*UN Department of Peace-building and Political Affairs*
New York, USA
masoodd@un.org

Martin Wählisch
*UN Department of Peace-building and Political Affairs*
New York, USA
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 كلب/*tšlb* (dog).

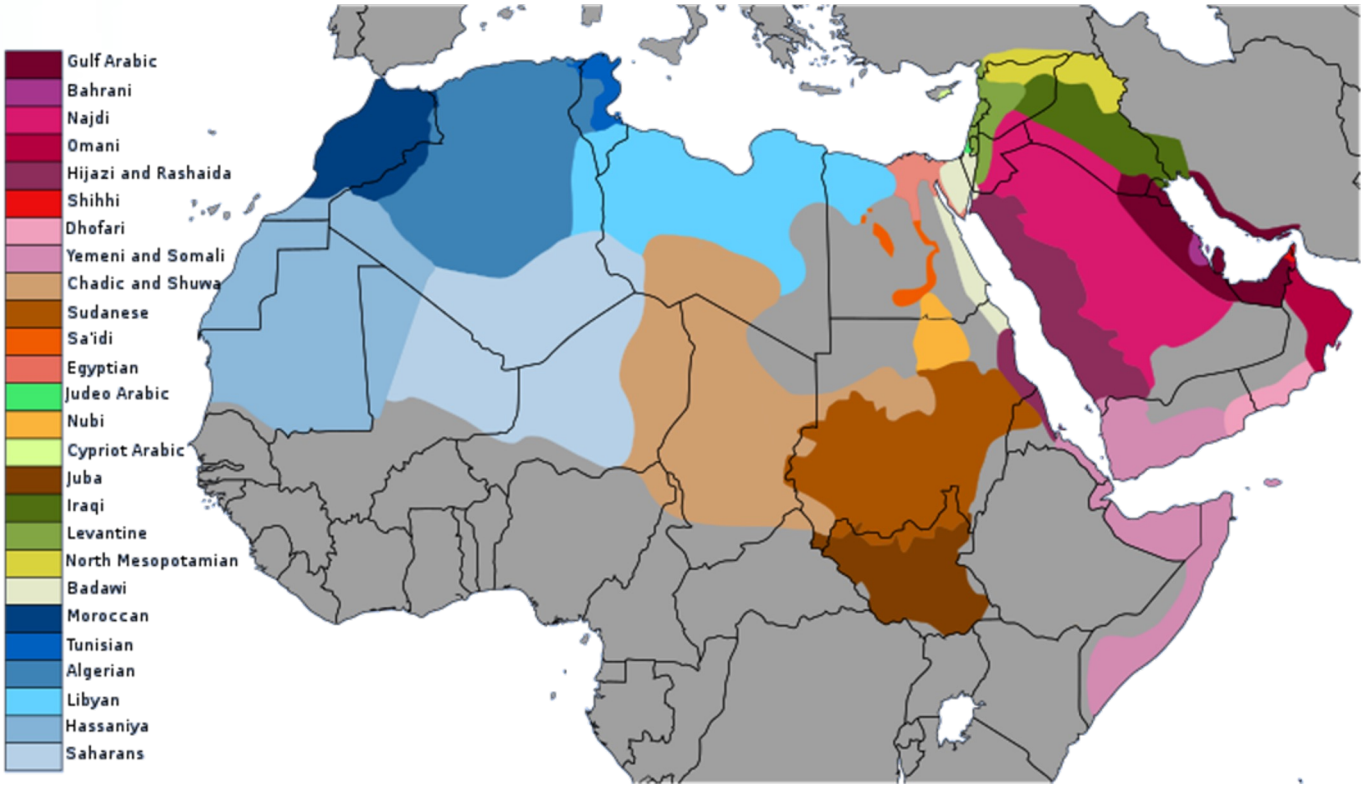
(ii) **Morphology**: Arabic dialects are similar to MSA

Jarrar, M., Zaraket, F., Hammouda, T., Masood, D., Waehlisch, M.: **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

The problem

# Arabic is a low resources language

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



Arabic Dialects Corpora

حوسبة اللهجات العربية

# Currasat كراسات

Arabic Dialects Corpora

مدونة اللهجات العربية

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

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

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

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

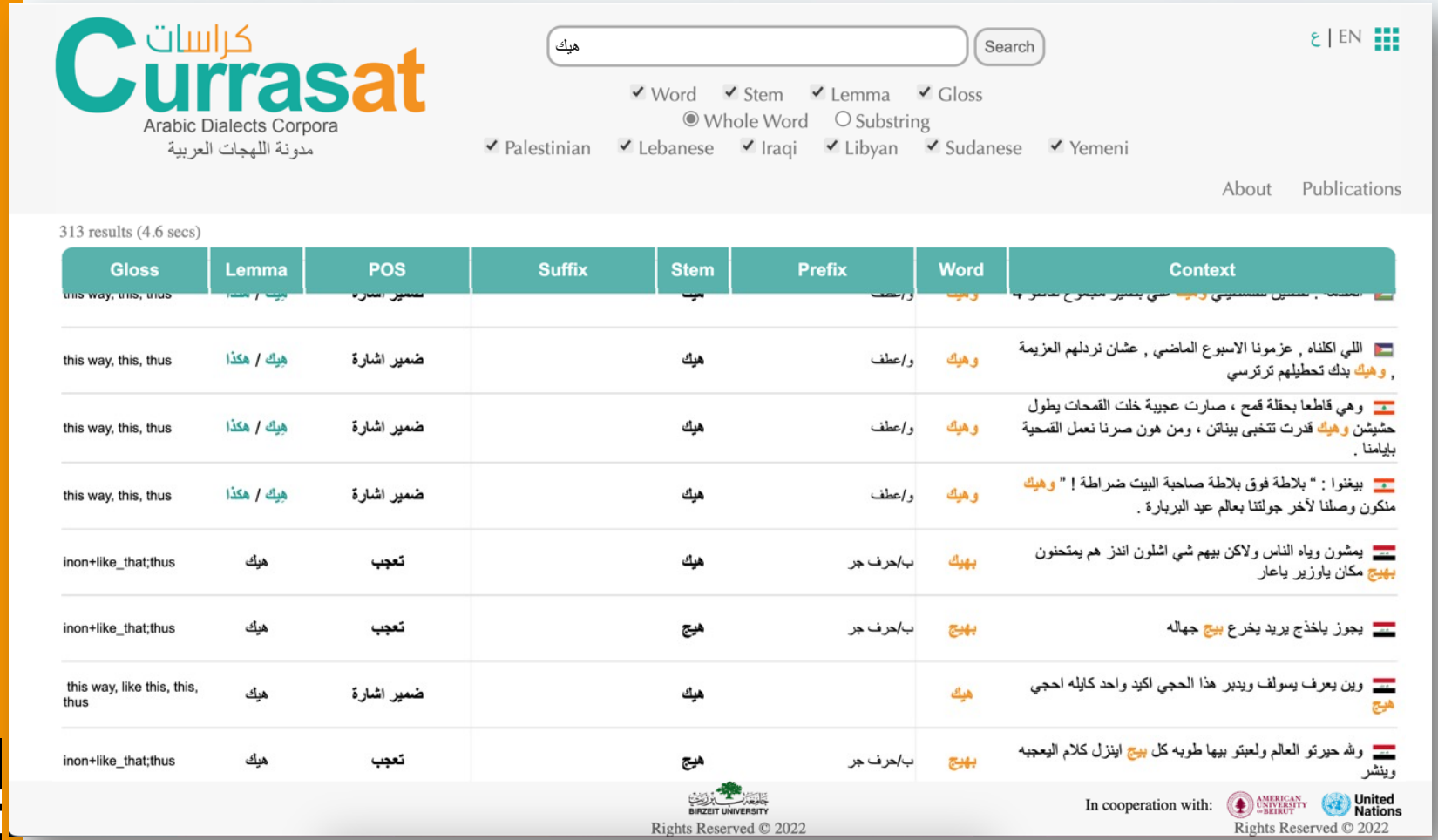
**Lisan لِسان:** Yemeni, Iraqi, Libyan, and Sudanese dialects corpora

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

DPPA AMERICAN UNIVERSITY OF BEIRUT BIRZEIT UNIVERSITY

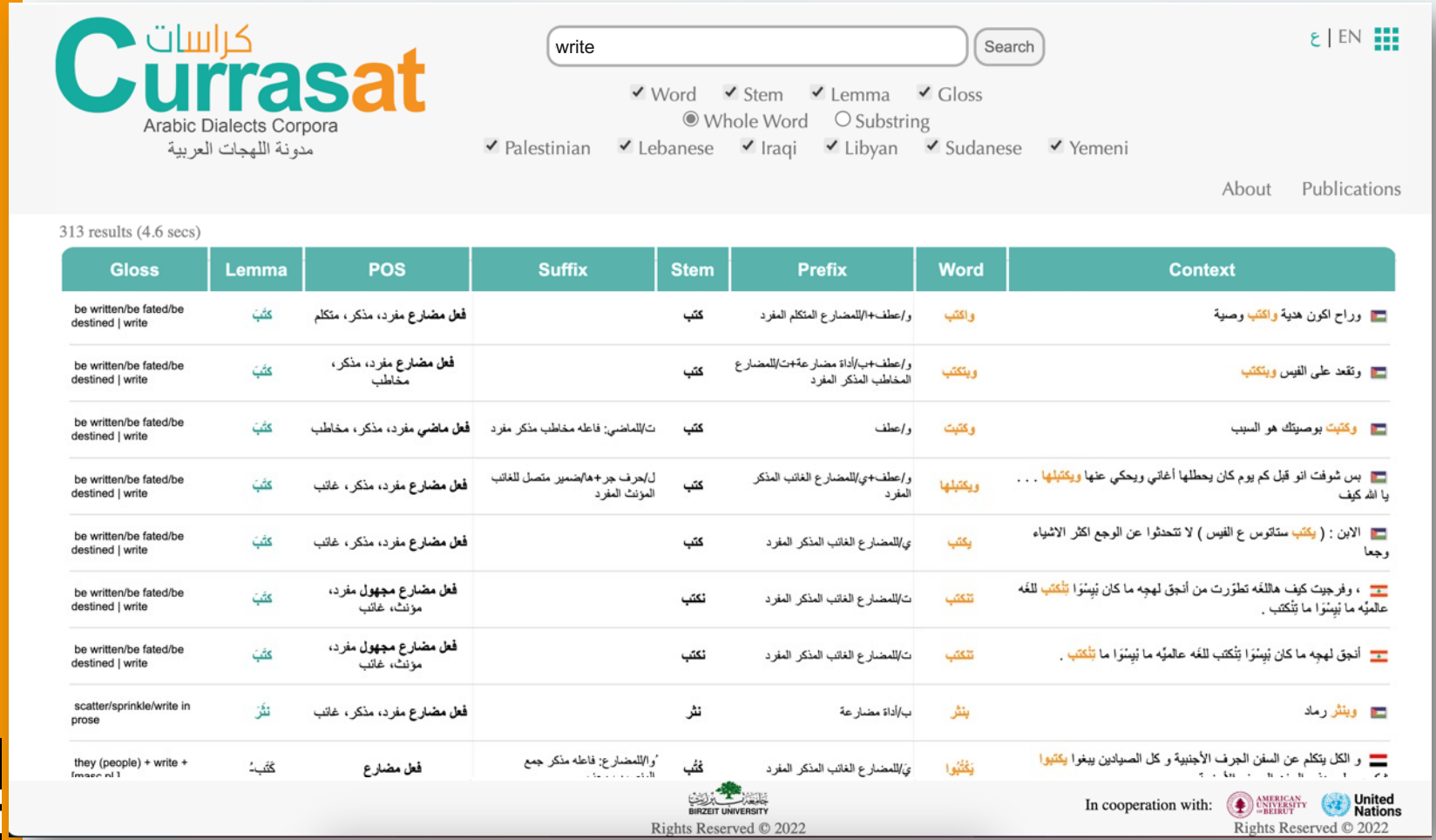

- El Haff, K., Jarrar, M., Hammouda, T., Zaraket, F., (2022). **Curras + Baladi: Towards a Levantine Corpus**. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.
- Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam 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




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

Try and download




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

Annotation Tools

SinaLab

# Tawseem Portal

بوابة سينا لتوسيم المدونات

Tymaa | Rol | دخول

Loggin 14:26:35 (118)

Stat Log

Lemmas | Annotations

بصير

اسم فعل وظيفية

Count: 2

ب+صير صار صارَ
فعل مضارع مفرد مذكر غائب
ب/أداة مضارعة
become/begin to

ب+صير صار صارَ
فعل مضارع مفرد مذكر متكلم
ب/أداة مضارعة
become/begin to

| Dialect | Context | Token | Prefix | Stem | Suffix | POS | Number | Gender | Person | DA Lemma | MSA Lemma | Gloss |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | شلون | | شلون | | أداة استفهام | | | | شلُوْن | كَيْفَ 1 | how |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | دا | | دا | | أداة استقبال | | | | رَحْ \| رَاحْ3 | سَوْفَ 1 | will/shall |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | تدخلي | ت/للمضارع المخاطب المؤنث المفرد | دخل | ي/للمضارع: فاعله مخاطب مؤنث مفرد | فعل مضارع | مفرد | مؤنث | مخاطب | دَخَلْ 1 | دَخَلْ 1 | enter |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | تسلمي | ت/للمضارع المخاطب المؤنث المفرد | سلم | ي/للمضارع: فاعله مخاطب مؤنث مفرد | فعل مضارع | مفرد | مؤنث | مخاطب | سَلَّم 1 | سَلَّم 1 | hand over/surren |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | عالنفسا | ع/حرف جر +ال/أداة تعريف | نفسا | | اسم | مفرد | مؤنث | | نَفْساء 1 | نَفْساء 1 | in childbed |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | .. | | .. | | علامة ترقيم | | | | .. | .. | .. |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | ما | | ما | | أداة نفي | | | | ما 2 | ما 2 | not |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | بصير | ب/أداة مضارعة | صير | | فعل مضارع | مفرد | مذكر | غائب | صارَ | صارَ | become/be gin to |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | يالله | ي/أداة نداء | الله | | اسم | مفرد | مذكر | | الله | اَللهْ | to/for + Allah/God + |
| حلبية | شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو | خيتو | | خيتو | | اسم | | مؤنث | - | أُخْت 1 | أُخْت 1 | sisters \| sister/count |

# 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:

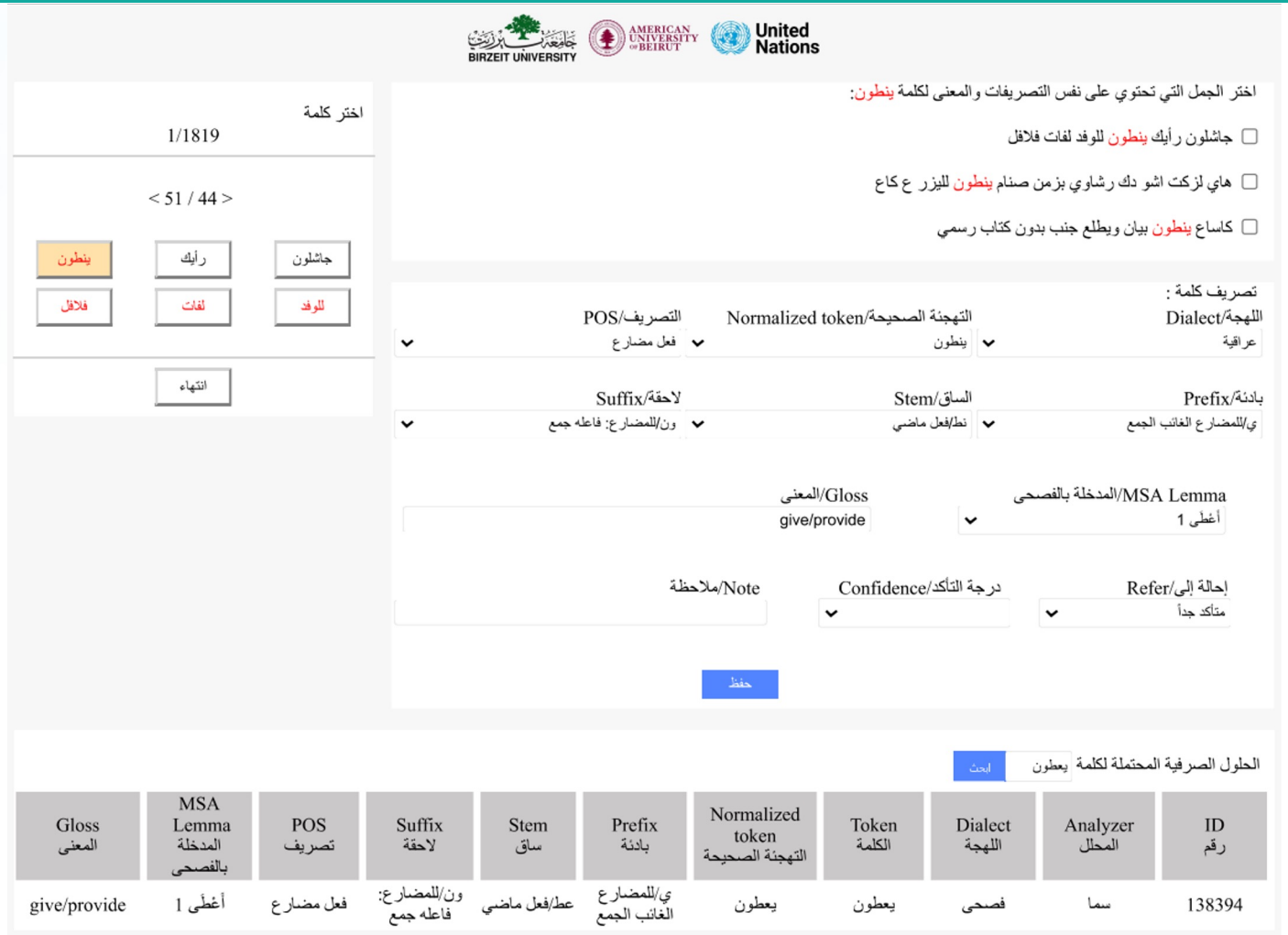
Yemeni, Iraqi, Sudanese, and Libyan.

## Annotation Methodology:

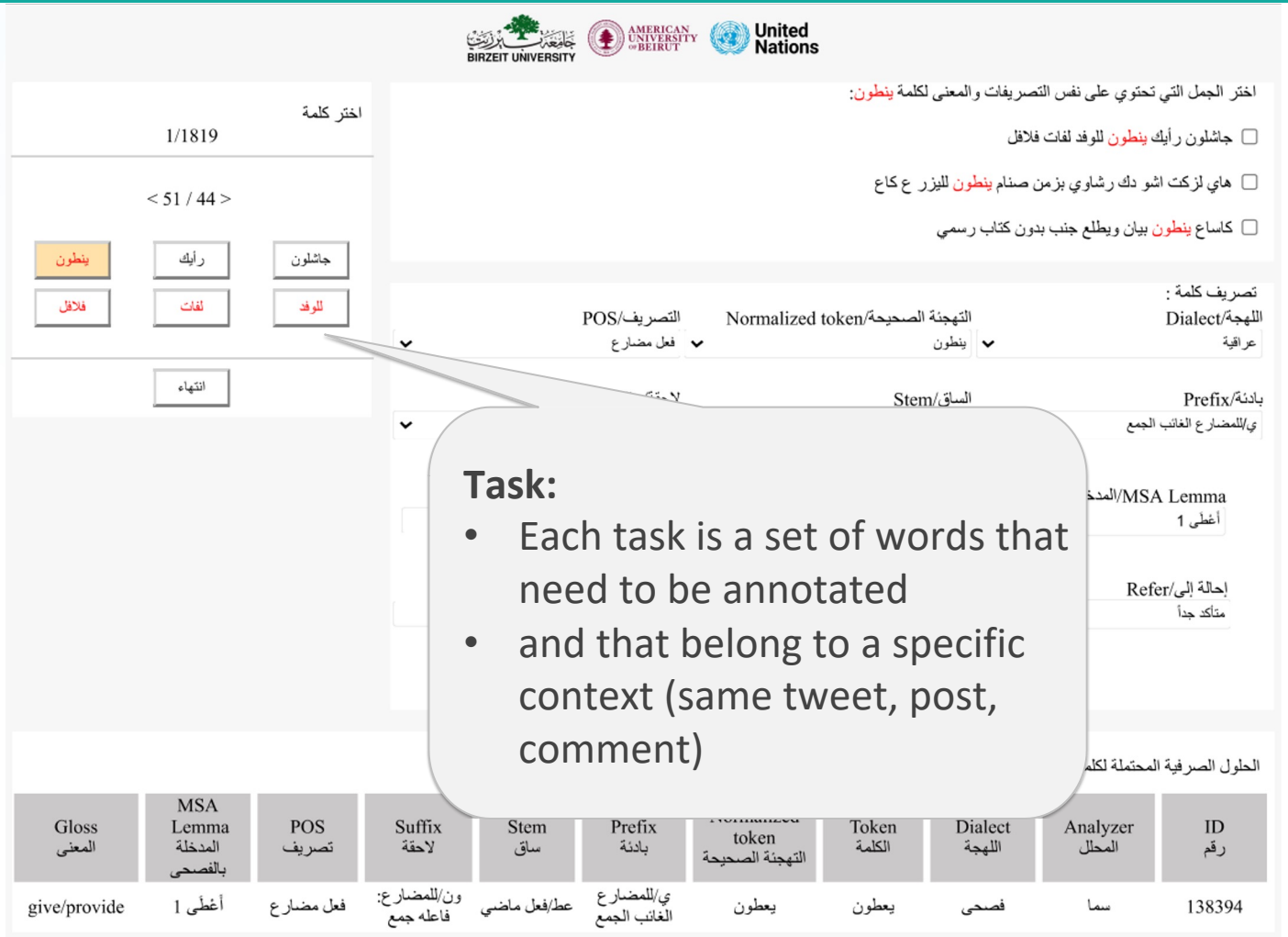
Tool: Arabic Dialect Annotation Toolkit (ADAT), over two years

| 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



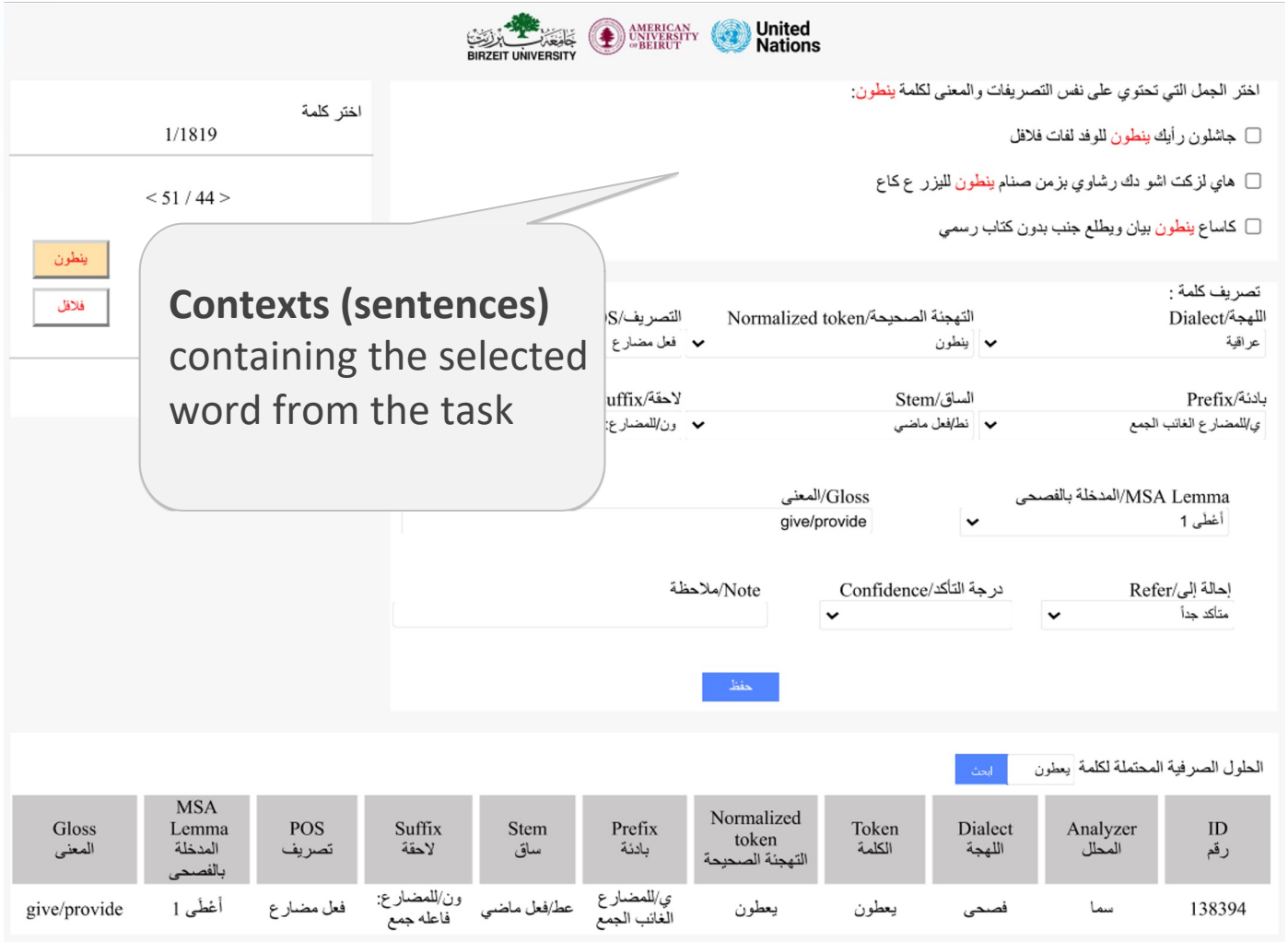
# Annotation Methodology



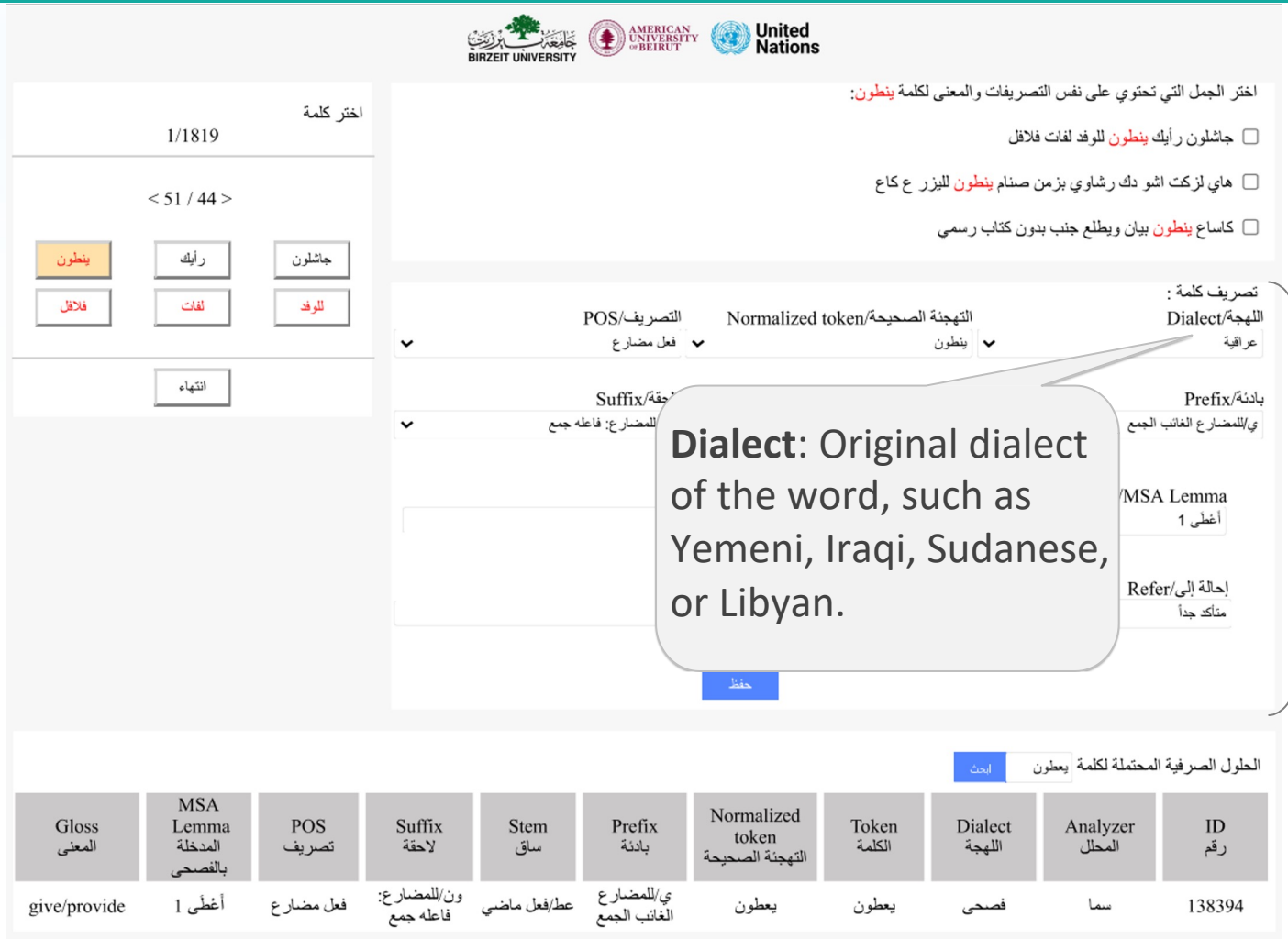

**Task:**

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

# Annotation Methodology

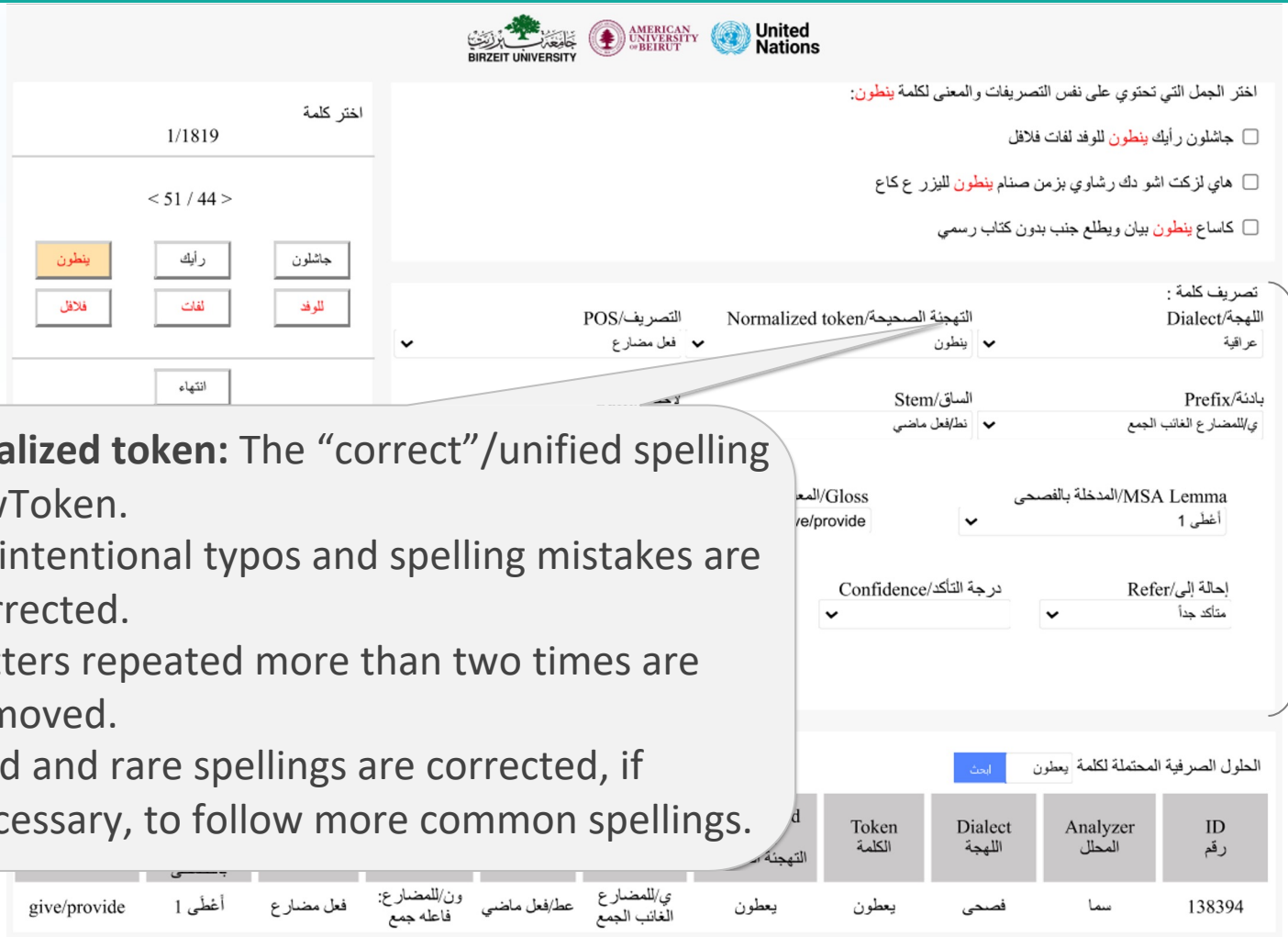


# Annotation Methodology




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

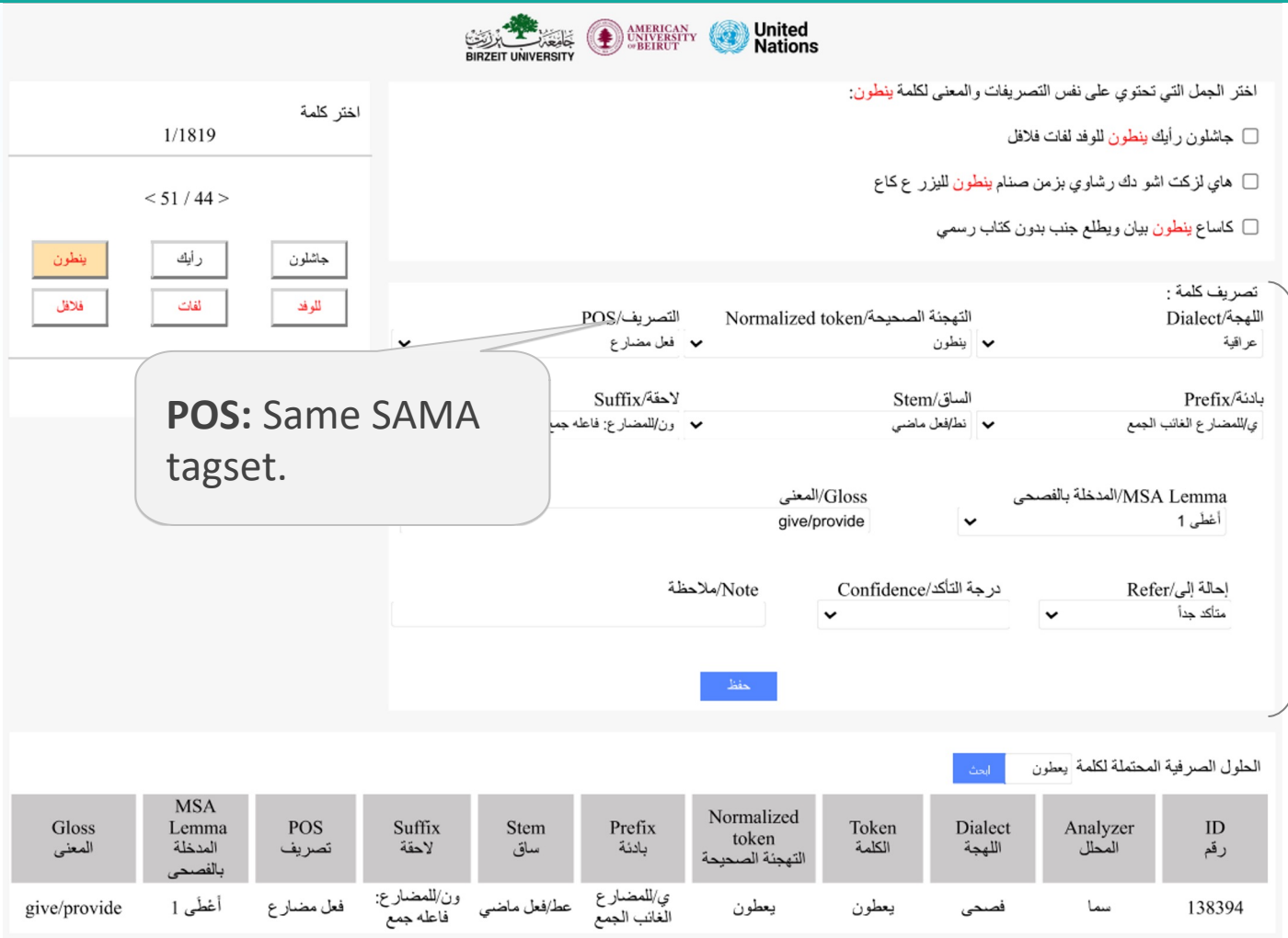
# Annotation Methodology



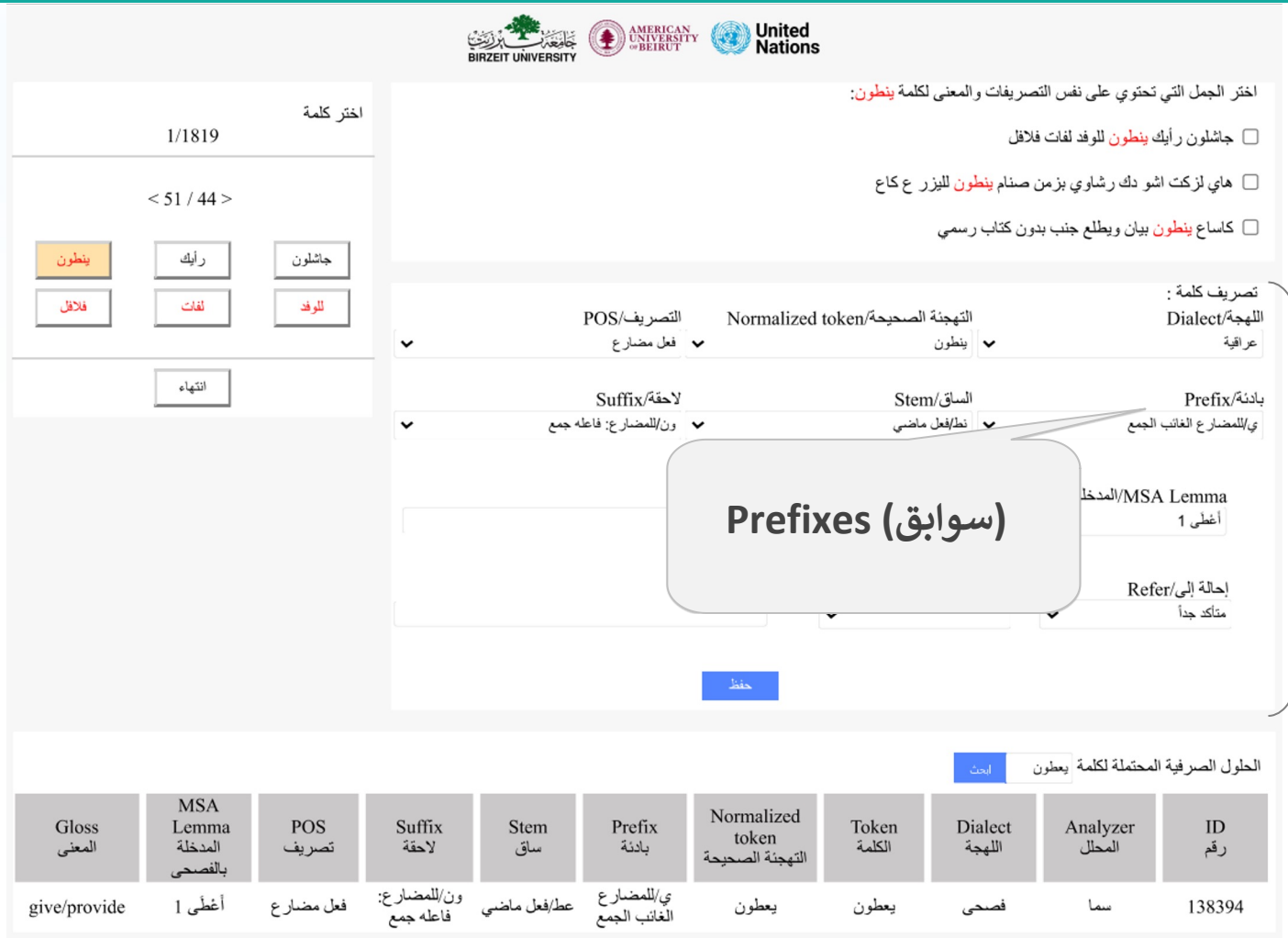

**Normalized token:** The “correct”/unified spelling of rowToken.

- 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.

# Annotation Methodology

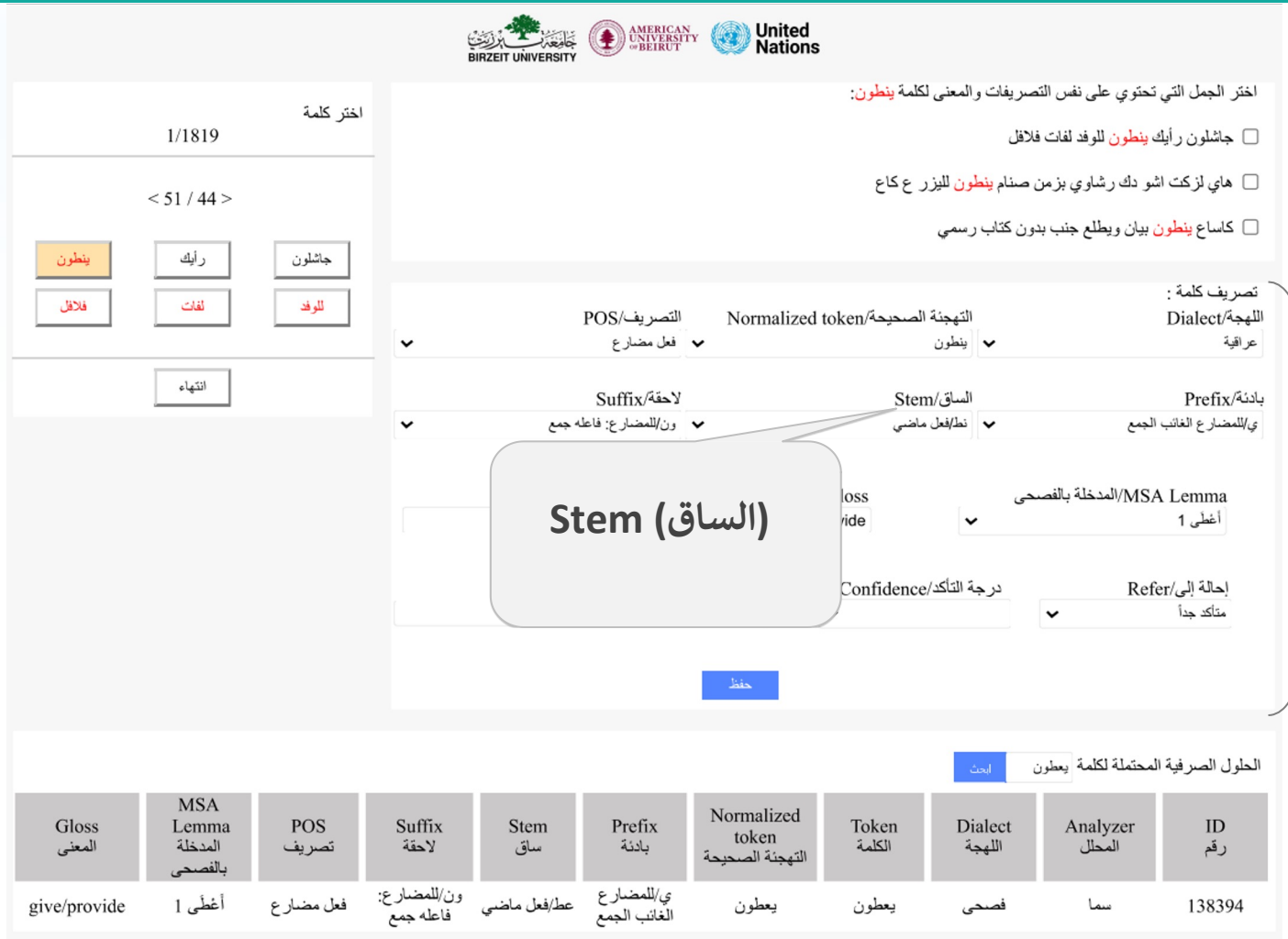


# Annotation Methodology

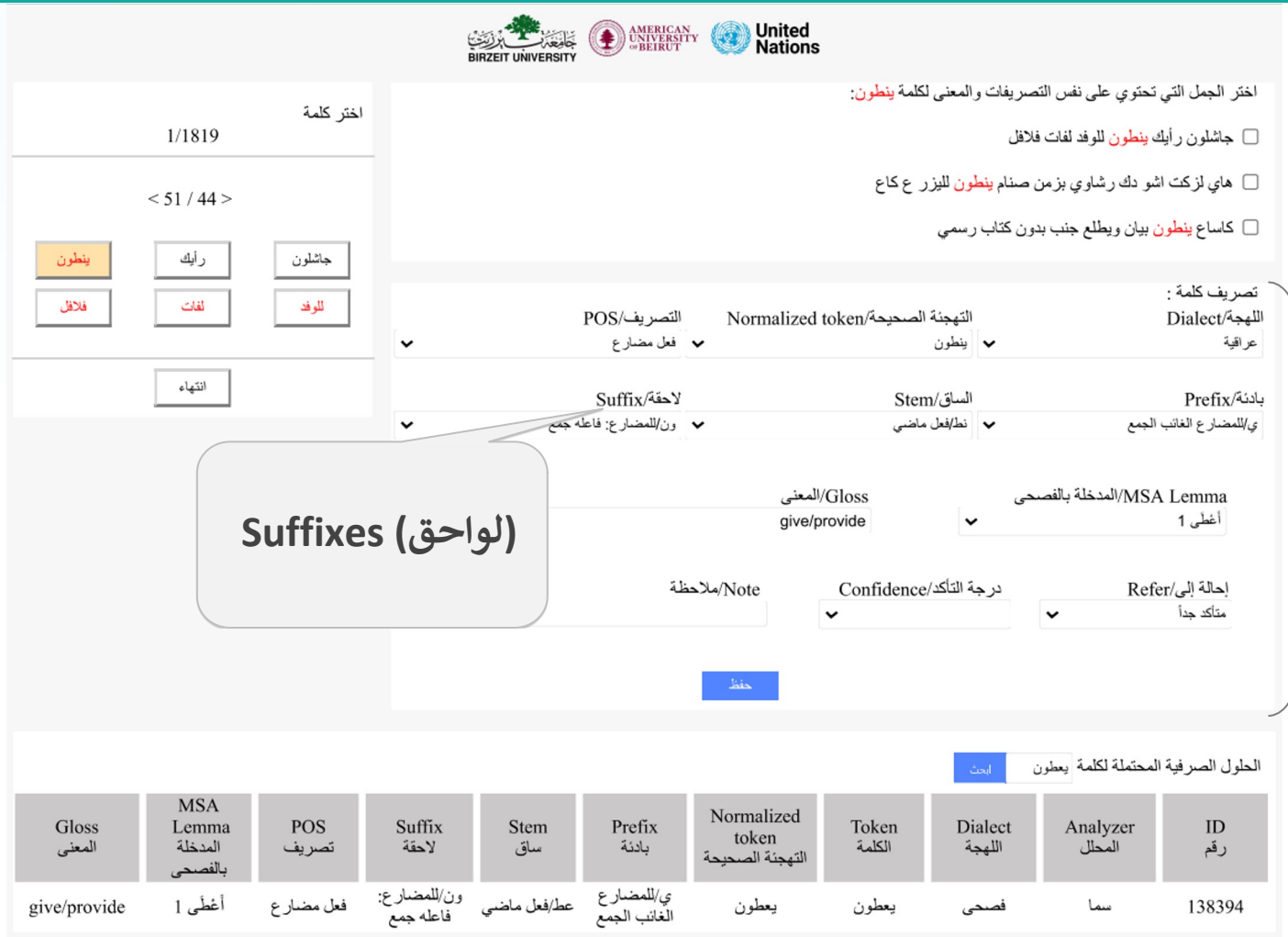



| ID رقم | Analyzer المحلل | Dialect اللهجة | Token الكلمة | Normalized token التهجئة الصحيحة | Prefix بادئة | Stem ساق | Suffix لاحقة | POS تصريف | MSA Lemma المدخلة بالفصحى | Gloss المعنى |
|---|---|---|---|---|---|---|---|---|---|---|
| 138394 | سما | فصحى | يعطون | يعطون | ي/للمضارع الغائب الجمع | عط/فعل ماضي | ون/للمضارع: فاعله جمع | فعل مضارع | أغطى 1 | give/provide |

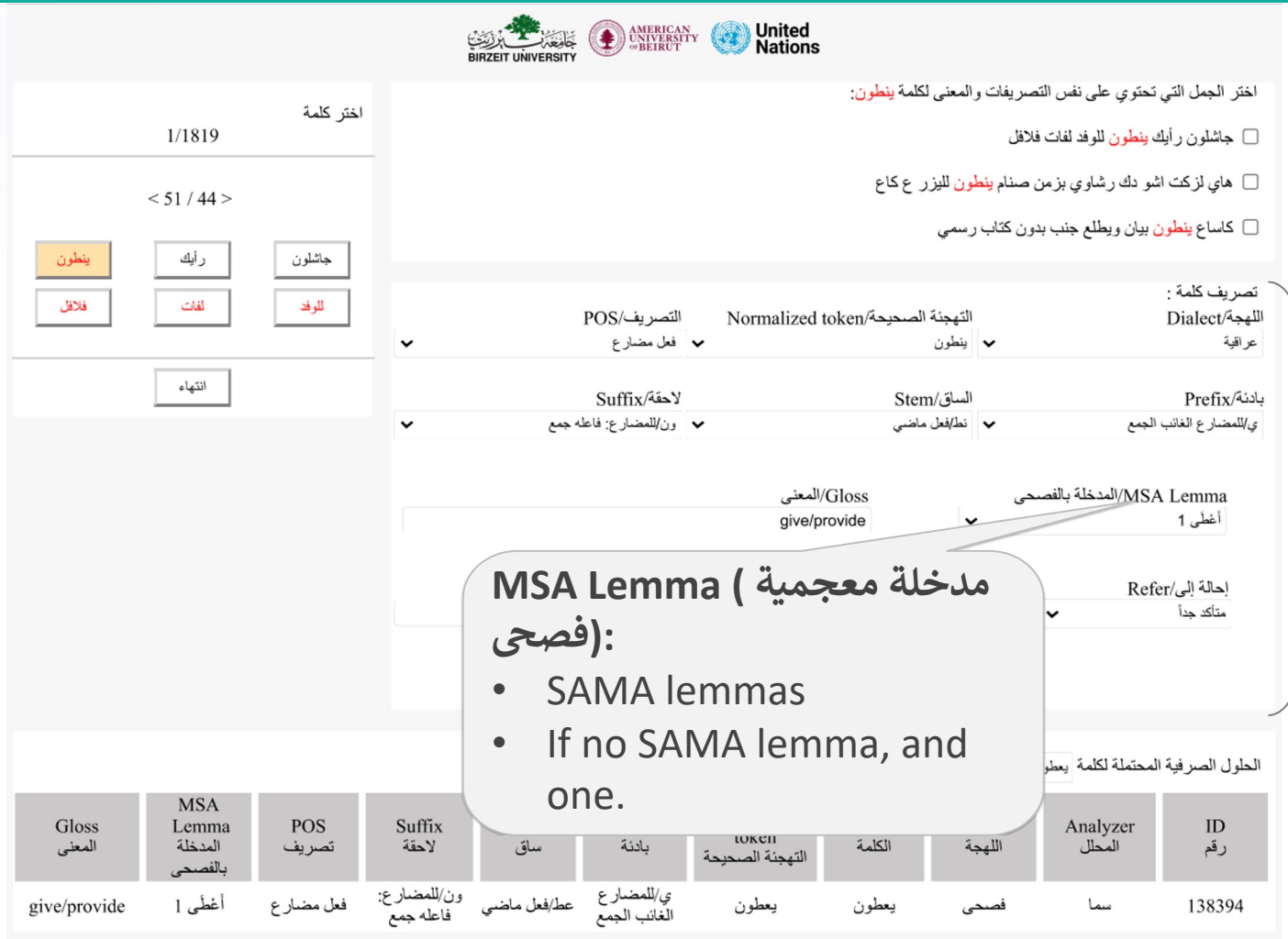
# Annotation Methodology



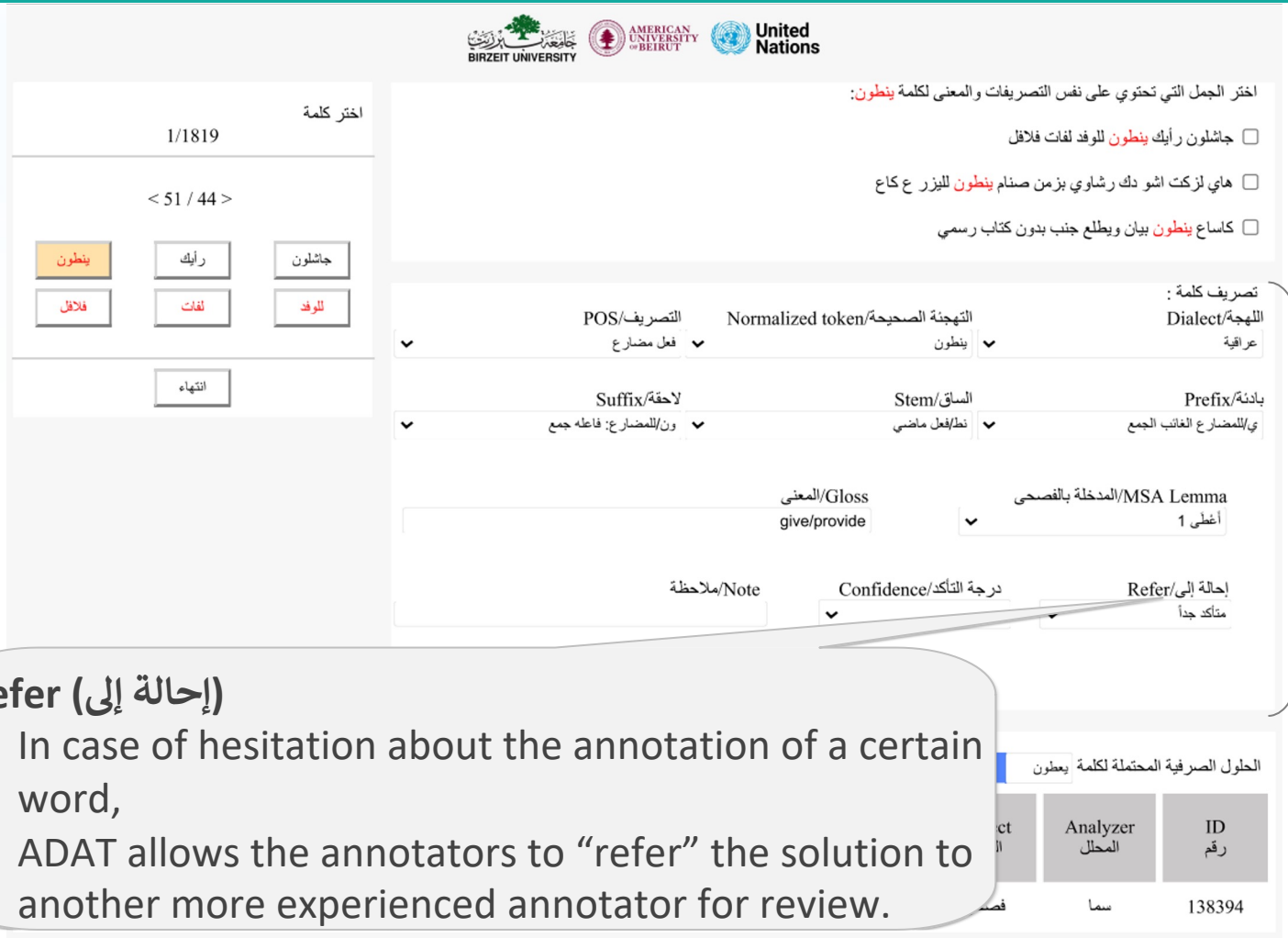
# Annotation Methodology



# Annotation Methodology



# Annotation Methodology

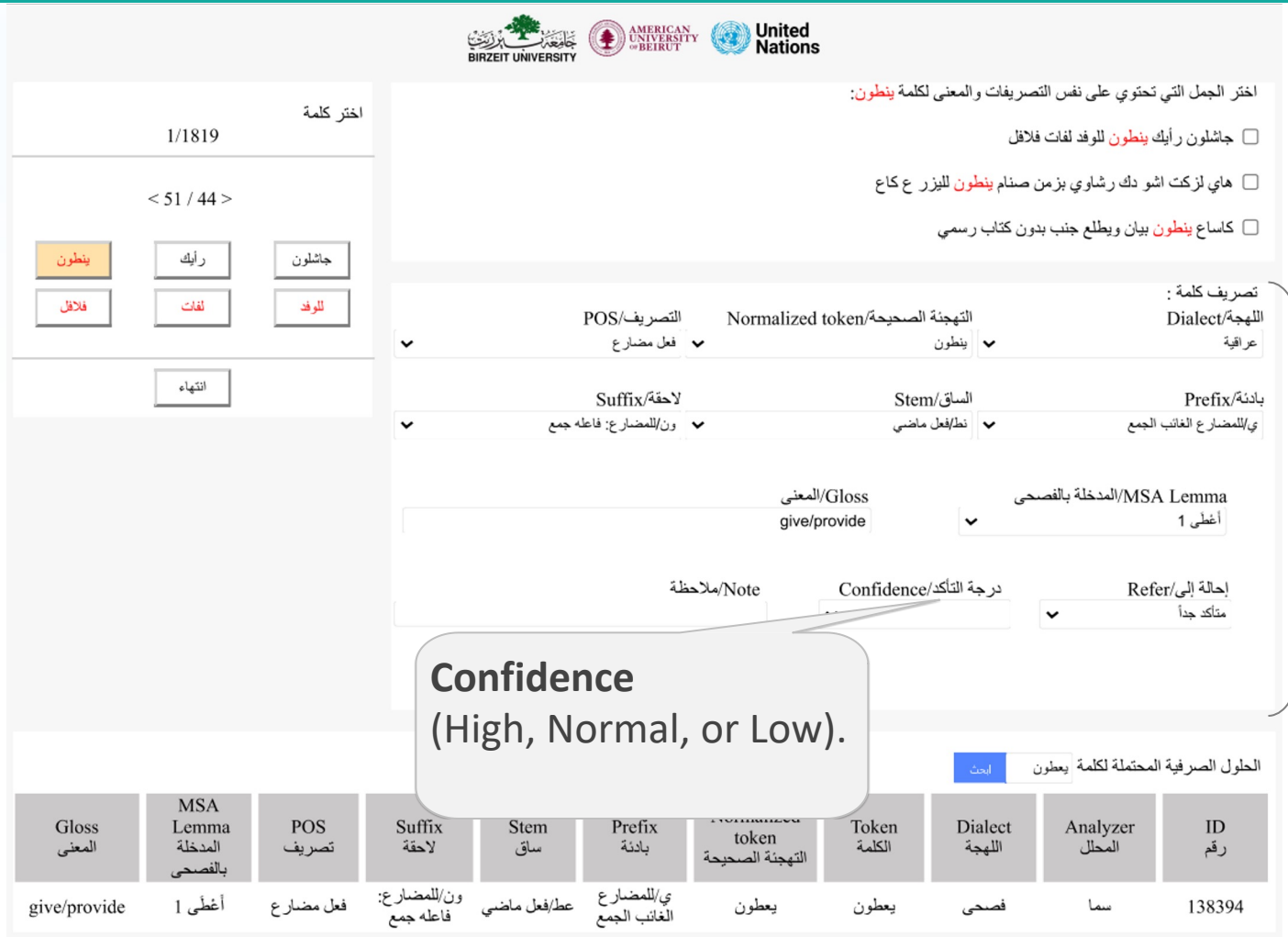



Morphological Features

**Refer (إحالة 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.

# Annotation Methodology



# Lisan Evaluation

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

| | Iraqi | | | Libyan | | | Sudanese | | | Yemeni | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature | 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/

# 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 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.
5. 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.
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: 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
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, 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).