Morphological Annotation of Quranic Arabic

Kais Dukes

Leeds University, UK

Nizar Habash

Columbia University, New York, USA

Presented by

Abdul Baquee Sharaf

Leeds University, UK

The Quranic Arabic Corpus
http://corpus.quran.com

University of Leeds, United Kingdom
Kais Dukes, Dr. Eric Atwell, A. M. Sharaf

Columbia University, New York, USA
Professor Nizar Habash

University of Maryland, USA
Professor Tim Buckwalter

University of Montreal, Canada
Wajdi Zaghouani (Linguistic Data Consortium)

Publications

RANLP 2009 INFOS 2010 LREC 2010

The Quran

- 1,400 years old
- The central religious text of Islam
- Written in Quranic Arabic, the direct ancestor language of Modern Arabic
- Highly studied linguistically for over 1,000 years
- Large body of existing published analyses of the Quran, but these are not machine-readable

The Quranic Arabic Corpus

- An international project involving researchers from several institutions
- Aim is to enable further understanding of the Quran through annotation
- Produce highly accurate machine-readable datasets of linguistic analysis

Multi-Level Annotation

- Word-by-word English to Arabic interlinear translation
- Part-of-speech tagging
- Morphological segmentation and inflection features
- Syntactic analysis using dependency grammar

http://corpus.quran.com

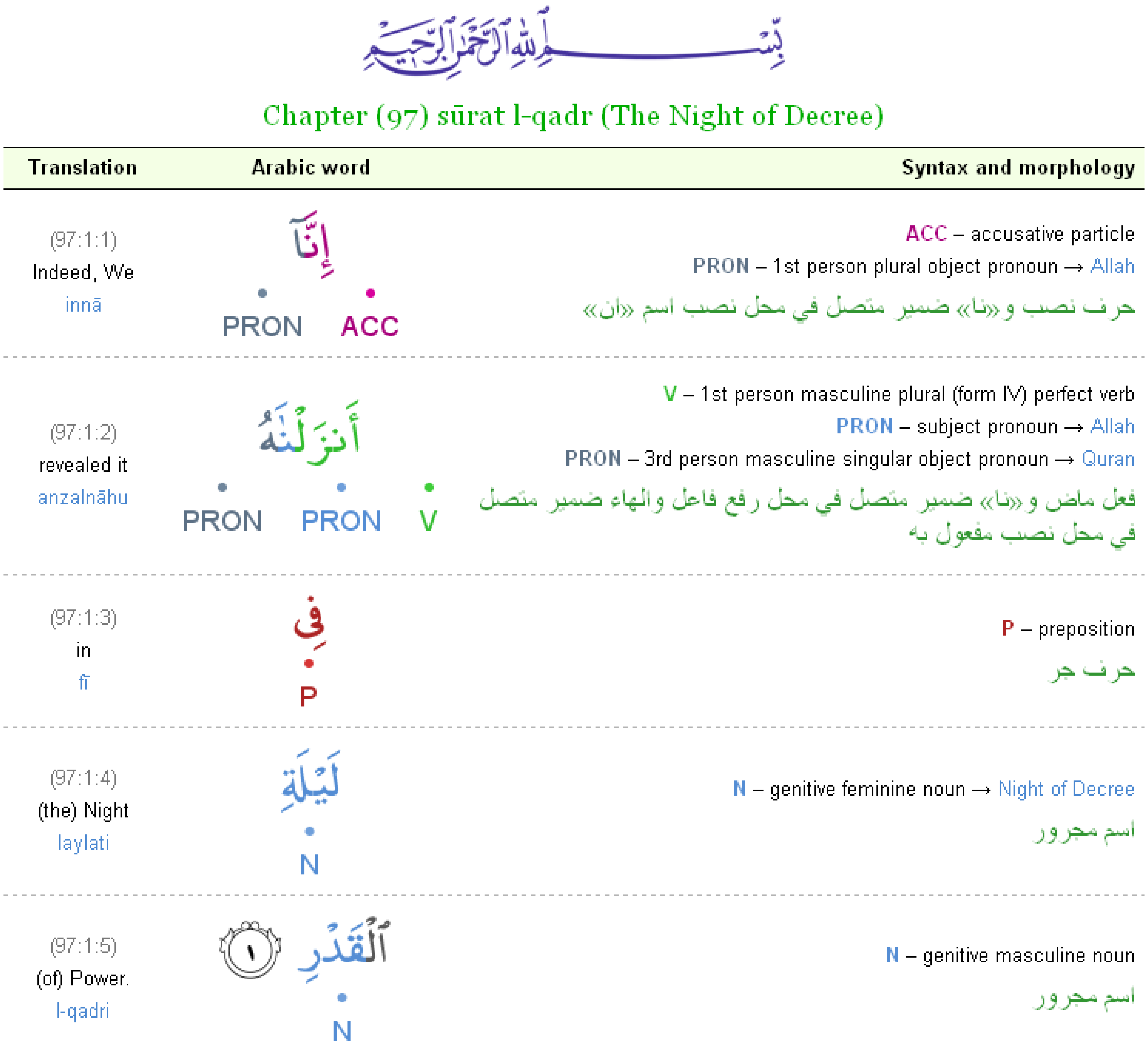
- A popular website (50,000 users per month)
- Used by researchers, scholars and students of the Quran and Arabic

Online Collaborative Annotation

- Anyone can view existing annotation
- Registered users can suggest corrections through a message board

Web-based Tools for Annotators and Researchers

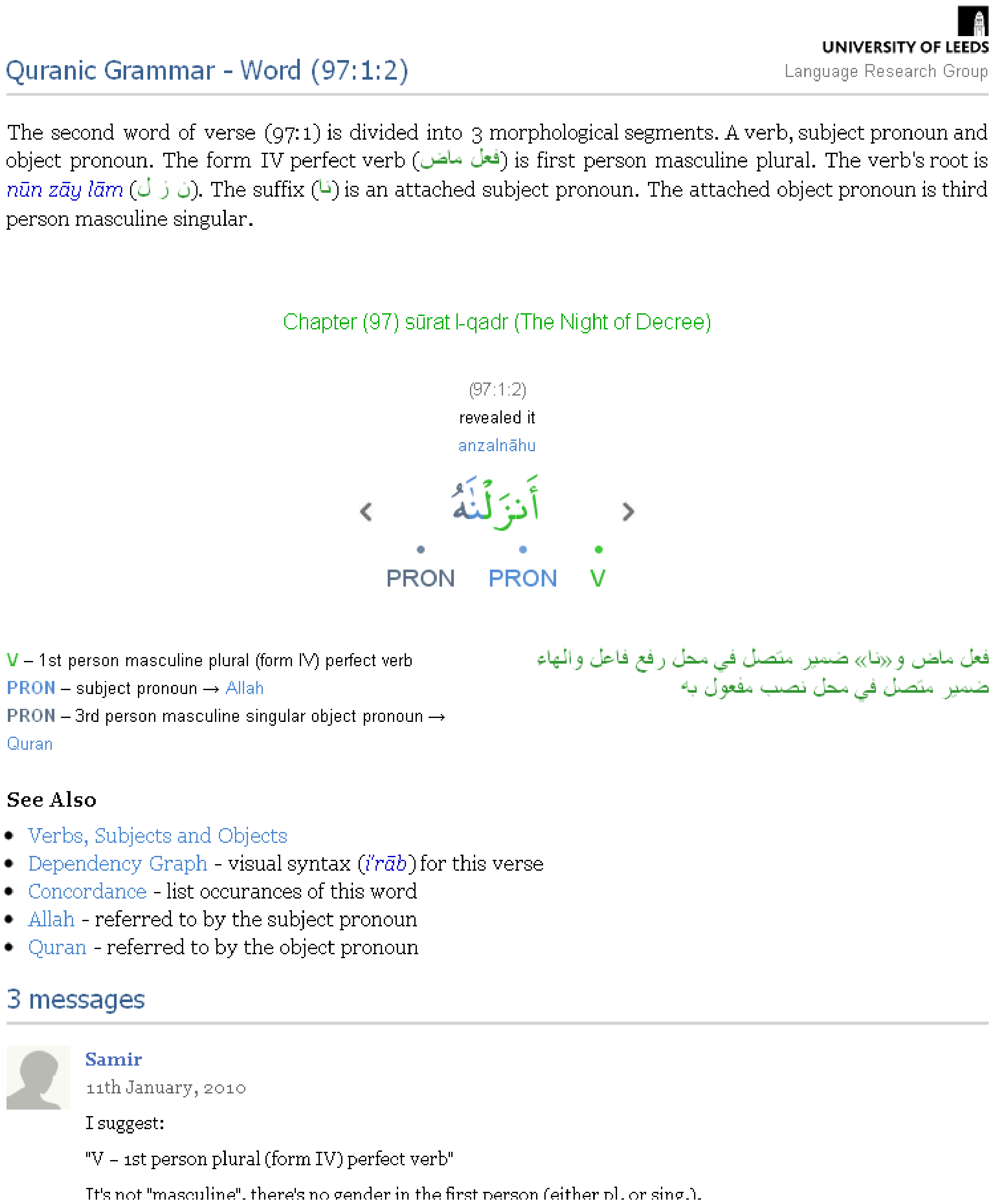
- Concordance of the Quran (to see how related words have been tagged)
- Morphological search by root, lemma or stem
- 7 parallel translations into English for each verse
- Automatically generated phonetic transcription
- Natural language generation used to create grammatical summaries
- Audio recitation in Arabic



بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Chapter (97) sūrat l-qadr (The Night of Decree)

Translation	Arabic word	Syntax and morphology
(97:1:1) Indeed, We innā	إِنَّآ PRON ACC	ACC – accusative particle PRON – 1st person plural object pronoun → Allah حرف نصب و «نا» ضمير متصل في محل نصب اسم «ان»
(97:1:2) revealed it anzalnāhu	أَنزَلْنَٰهُ PRON PRON V	V – 1st person masculine plural (form IV) perfect verb PRON – subject pronoun → Allah PRON – 3rd person masculine singular object pronoun → Quran فعل ماض و «نا» ضمير متصل في محل رفع فاعل والهاء ضمير متصل في محل نصب مفعول به
(97:1:3) in fī	فِى P	P – preposition حرف جر
(97:1:4) (the) Night laylati	لَيْلَةِ N	N – genitive feminine noun → Night of Decree اسم مجرور
(97:1:5) (of) Power. l-qadri	ٱلْقَدْرِ ١ N	N – genitive masculine noun اسم مجرور



UNIVERSITY OF LEEDS
Language Research Group

Quranic Grammar - Word (97:1:2)

The second word of verse (97:1) is divided into 3 morphological segments. A verb, subject pronoun and object pronoun. The form IV perfect verb (فعل ماض) is first person masculine plural. The verb's root is *nūn zāy lām* (ن ز ل). The suffix (نا) is an attached subject pronoun. The attached object pronoun is third person masculine singular.

Chapter (97) sūrat l-qadr (The Night of Decree)

(97:1:2)
revealed it
anzalnāhu

‹ أَنزَلْنَٰهُ ›

PRON PRON V

V – 1st person masculine plural (form IV) perfect verb
PRON – subject pronoun → Allah
PRON – 3rd person masculine singular object pronoun → Quran

فعل ماض و «نا» ضمير متصل في محل رفع فاعل والهاء ضمير متصل في محل نصب مفعول به

See Also

- Verbs, Subjects and Objects
- Dependency Graph - visual syntax (*i'rāb*) for this verse
- Concordance - list occurances of this word
- Allah - referred to by the subject pronoun
- Quran - referred to by the object pronoun

3 messages

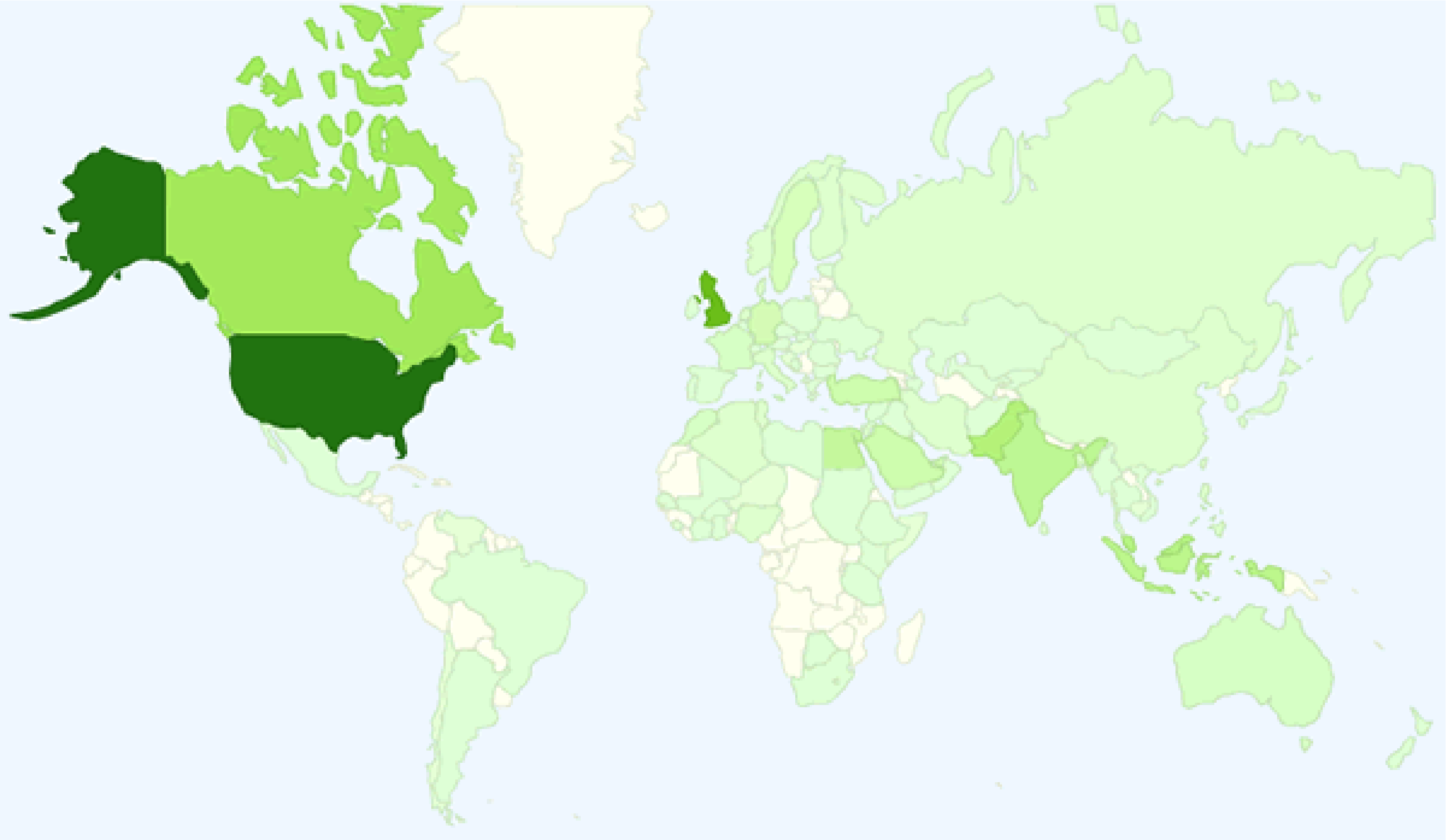
Samir
11th January, 2010

I suggest:

"V – 1st person plural (form IV) perfect verb"

It's not "masculine". there's no gender in the first person (either pl. or sing.).

Online Volunteer Collaboration (يد الله مع الجماعة)



- 50,000 Users per month
- 150 Researchers on the comp-quran mailing list
- Hundreds of online expert volunteers (Quranic Scholars)

Annotating the Quran

- 77,430 words in the Quran
- Each word is part-of-speech tagged, with morphological analysis
- This is initially done off-line, using a morphological analyzer
- Adapted BAMA (Buckwalter Morphological Analyzer)
- Approx 80% accurate

Difficulties in Adapting BAMA

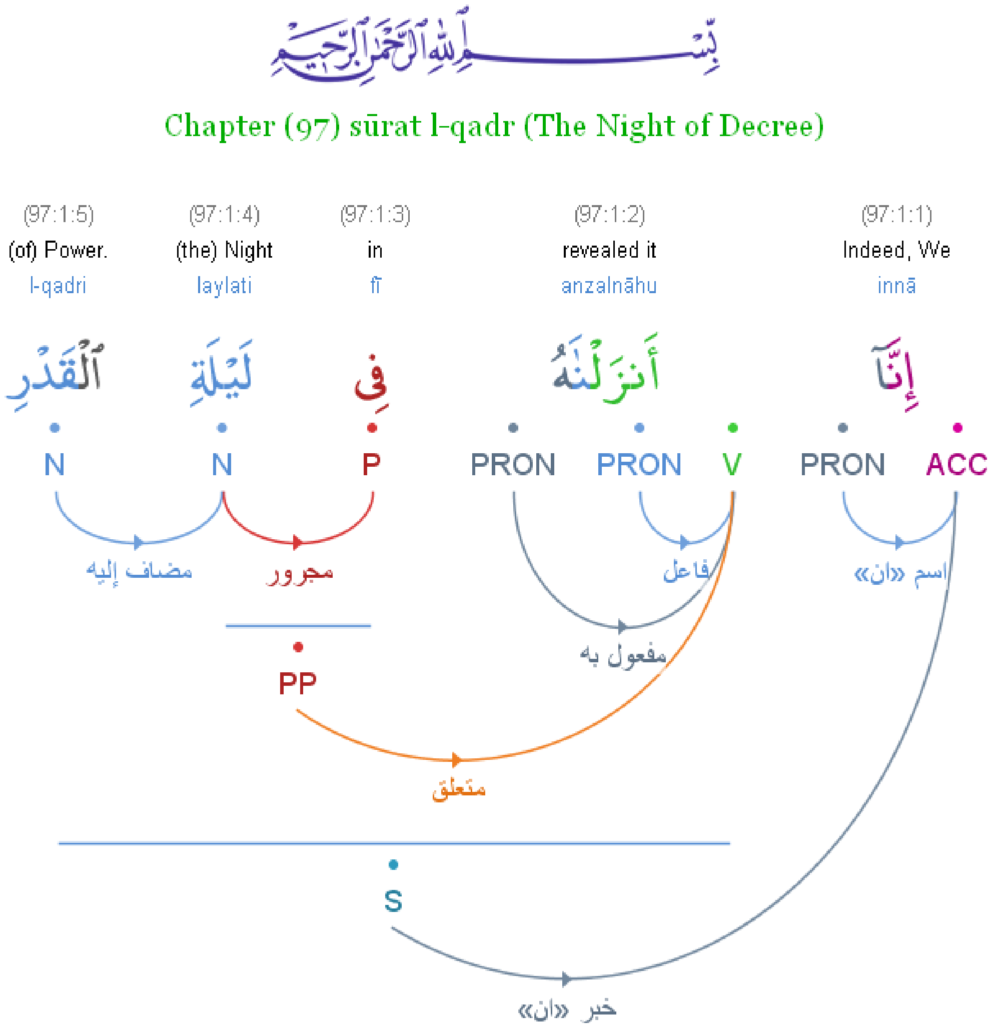
- Quran uses different spelling compared to Modern Arabic
- Out of vocabulary errors
- BAMA does not use context – multiple possible analyses

Solutions

- Diacritic edit-distance, to find best match
- Initial off-line manual verification
- Filtering of BAMA results based on contextual grammatical rules

Syntactic Treebank

- Dependency Grammar (11,000 words completed so far)



Chapter (97) sūrat l-qadr (The Night of Decree)

Novel Contributions of this Research

- NLP and Corpus Linguistics applied to Classical Arabic
- Arabic language processing tools
- Datasets for Further Research
- Methodology - Large-scale online collaborative annotation

Future Work

- Pronoun resolution (in progress, 11,000 words annotated so far)
- Semantic ontology (in progress, 300 concepts defined with relations)
- Quranic PropBank (planned)

Applications

- A popular and unique free online Quranic Arabic study tool
- Datasets enable further automatic computational analysis of the Quran
- Training data for morphological analyzers and parsers for Classical Arabic

Thank You

Question:
kais@kaisdukes.com