البلاغة **The BALAGHA Score**
Arabic Rhetoric Scoring System

BALAGHAScore.com Arabic Word Tokenisation Scheme v0.1.0

1) Introduction

This document presents the **BALAGHAScore.com Arabic Word Tokenisation Scheme v0.1.0**, a modality-agnostic, script-agnostic and dialect-agnostic morphological tokenisation framework for tokenising Arabic words into constituent *word units* for rhetorical density calculations.

The Scheme was originally developed for use with the BALAGHA Score, an implementation of rhetorical density measurement for Arabic-language texts. However, it has been published on Github and Zenodo with version control for use in any other digital humanities applications in Arabic, and as an exemplar for tokenisation in other languages.

2) Why is tokenisation required?

Rhetorical density is defined as the number of rhetorical devices per 100 orthographic words (for isolating languages such as English and Chinese), or per 100 word units (for fusional languages such as Arabic and Spanish, and agglutinative languages such as Turkish and Finnish.) Arabic is a fusional language where compound words are formed by the addition of meaningful components to other words.

For example:

A house (2 words)	بيت (1 word)
My house (2 words)	ي + بيت → بيتي (1 word)
His house (2 words)	ه + بيت →بيته (1 word)

These examples demonstrate that a meaning which needs 2 words in English – such as “my house” would be represented by only one word in Arabic – “بيتي”. This difference affects the “word count” denominator in the rhetorical density calculation, making it impossible to compare the rhetorical densities of texts from different languages.

More importantly, this behaviour can affect “word count” comparison between Arabic texts as well, because the “word count” is affected by the syntax of words being joined together, and is no longer a reflection of how much information is contained within the text.

As an example, the word فأسقيناكموه consists of 4 individual units of meaning. A text with lots of words like this would have fewer overall words and an artificially high rhetorical density, compared to a text with the same amount of information but with simpler – and more numerous – words. The rhetorical density is therefore altered due to stylistic factors within the text – which become a confounding variable - rather than the density of rhetorical devices itself.

The way to solve this problem is to tokenise the Arabic text – break the words down to expose smaller units of meaning – and to count these word units instead of the words. It is crucial to tokenise different texts in a consistent and reproducible way so that rhetorical densities for different texts can be compared.

3) The tokenisation strategy

Arabic tokenisation has no universally correct or official standard. The appropriate level of tokenisation depends entirely on the intended analytical use case such as machine translation, sentiment analysis or morphological analysis. Existing tokenisation schemes such as lemmatisation or morphological parsing create either overly large tokens which conceal meaningful rhetorical structure, or overly granular tokens which inflate word unit counts and distort density measures. These schemes were not intended for rhetorical density analysis.

The **BALAGHAScore.com Arabic Word Tokenisation Scheme** therefore defines the “word unit” in a way optimised specifically for rhetorical density calculation. The underlying philosophy reflects the need to consistently enumerate the amount of semantic and linguistic information in the text, as a denominator for the rhetorical density calculation. The philosophy consists of three pillars.

Pillar 1: Syntactic layer – compound words will be segmented into their syntactically independent components for the reasons explained above. For example, “بيتي”is broken into 2 word units: “بيت” and “ي+”.

Pillar 2: Morphologic layer – morphologically discrete words will be preserved because further tokenisation increases word unit counts for every text, across the whole spectrum of texts, without yielding useful information about rhetorical density. While more granular tokenisation is required for other use cases, the additional computational complexity associated with sub-atomic splitting does not benefit rhetorical density calculation.

- Root and pattern decomposition is not performed. For example, ذهبنا is not broken into ذهب نا+, and أكبر is not broken into كبير + أفعل.
- Etymological decomposition is not performed. For example, words such as لماذا are retained as a single word unit, despite their etymological origin ل + ماذا.

Pillar 3: Orthographic layer - orthographic variations – which do not represent variation in the amount of information conveyed – will be normalised into the simplest and most efficient forms. For example:

- Writing a number in digits conveys the same amount of information as writing it in long form. However, the former is one word, while the latter may be more than five words.
- A text with diacritics can be tokenised into more words than a text without diacritics. This is merely a reflection of orthographic style rather than the amount of information conveyed, and is hence a confounding factor when comparing rhetorical densities.

4) Scope of this tokenisation scheme

This tokenisation scheme can be used to tokenise Arabic text in any orthographic form (Arabic script, Romanized, Arabizi) and in any modality (written text, transcribed speech). The scheme handles Modern Standard Arabic, regional dialects, and mixed varieties. Code-switching with non-Arabic languages is handled through the rules for non-Arabic words.

The primary requirement is that the annotator or computer script applying the tokenisation scheme can identify morphological boundaries - where clitics attach, and how conjugation is encoded, for example. The surface form (written, spoken, or in any script) does not affect the tokenization rules. The only limitation is the annotator's or computer script's ability to recognize the underlying structure. An annotator or computer script untrained with Levantine Arabic, for example, may not correctly tokenise Levantine-specific forms.

5) Publication

The tokenisation scheme has been published with versioning in external, independent repositories – Github and Zenodo – so that it can be adopted for the tokenisation of Arabic words for rhetorical density measurement and other use cases.[1] This ensures transparency, reproducibility, and consistency across datasets and analyses.

The tokenisation scheme can be implemented manually or computationally. An online version is available on the BALAGHAScore.com website.

6) Version history

- 2025-12-09 - **v0.1.0**: Initial release on Zenodo and Github including:
  - Core tokenisation scheme for Arabic.
  - Reference example.

[1] Cite as: Marathe, Mandar. *BALAGHAScore.com Arabic Word Tokenisation Scheme*. V0.1.0. Zenodo. 10 December 2025. https://doi.org/10.64393/balagha-score.tokenisation-v0.1.0

7) Tokenisation rules

Pillar 1: Syntactic Layer: Segmentation

Linguistic element	**Type**	**Action**	**Example**
Proclitics	Conjunctional e.g. و, ف	Clitic segmentation	و+ ذهب → وذهب 1 word → 2 word units
	Modal e.g. س		س+ أذهب → سأذهب 1 word → 2 word units
	Interrogative e.g. أ		أ+ تذهب → أتذهب 1 word → 2 word units
	Prepositional e.g. ب, لِ		ب+ السيارة → بالسيارة 1 word → 2 word units
	Comparative e.g. ك		أحمد ك+ الأسد → أحمد كالأسد 2 words → 3 word units
Enclitics	Possessive e.g. ي, نا	Clitic segmentation	بيت +ي → بيتي 1 word → 2 word units
	Object pronoun e.g. هم, ها	Clitic segmentation	رأيت + ه → رأيته 1 word → 2 word units
Multi-clitic clusters	All	Clitic segmentation	و+ س+ تذهب → وسنذهب 1 word → 3 word units

Pillar 2: Morphologic Layer: Preservation

Linguistic element	Type	Action	Example
Definiteness markers	Alif & lām (*al*-)	None	البيت No change, 1 word unit
	Nunation		بيتاً No change, 1 word unit
Verb conjugations	Imperfect tense	None	نذهب No change, 1 word unit
	Perfect tense	None	ذهبنا No change, 1 word unit
Active and passive participles	All	None	كتبت الكاتبة المكتوب No change, 3 word units
Plurals	All	None	معلمون No change, 1 word unit
Elative adjectives	All	None	أكبر No change, 1 word unit
Nisba adjectives	All	None	مصري No change, 1 word unit
Nouns of place	All	None	مكتب No change, 1 word unit
Tool nouns	All	None	مفتاح No change, 1 word unit
Colour or defect adjectives	All	None	أسود No change, 1 word unit

Genitive construct	All	Each component counted individually	مكتب المدير 2 word units: مكتب & المدير
Multi-word fixed expressions	All	Each component counted individually	“على أي حال” 3 words → 3 word units
Fused particles	All e.g. لماذا, ربما, لقد	None	لماذا ذهبت No change, 2 word units

Pillar 3: Orthographic Layer: Normalisation

Linguistic element	Type		Action	Example
Punctuation marks	All		Removed	هل تذهب؟ → هل تذهب
Emojis	All		Removed	مبروك 😂😂😂 → مبروك
Diacritics	All		Removed	كُتِبَ → كتب
Long-form numbers	All		Converted to numeral	ألف وثلاثمائة وسبعة وعشرون → “1327” 4 words → 1 word unit
Dates	Year	Long-form year	Converted to numeral	عام ألف وتسعمائة وخمسة وسبعين → “1975” 5 words → 1 word unit
		Year qualifiers	Only number year retained	2015 عام → 2015 2 words → 1 word ٢٠٢٥م (2025CE) → “2025”

	Month	Only name of month retained	شهر أبريل → أبريل 2 words → 1 word
	Day	Converted to numeral	الخامس من أبريل → 5 أبريل 3 words → 2 word units
Numbers with currency signs and percentages	All	Currency /percentage signs removed	١٠٠$ → ١٠٠ ١٠٠٪ → ١٠٠ 1 word → 1 word
Named entities	All	Each component counted individually	أبو الحارث نعمان بن عبد الرحمن التميمي No change, 7 word units
Duplicated letters for emphasis	All	Counted as normal word	"مبرووووك" 1 word → 1 word unit
Tatwīl / Kashīda	All	Removed	كتـــاب → كتاب No change, 1 word unit
Non-Arabic words	All	Each component counted individually	أقرأ الموقع BBC News كل يوم. No change, 6 word units
URLs	All	Counted as one word	أقرأ الموقع https://news.bbc.co.uk كل يوم. No change, 5 word units
Hashtags	All	Each component counted individually	#أبحاث_اللغة_العربية_والحوسبةِ_AI → أبحاث اللغة العربية و+ الحسوبة AI 1 hashtag → 6 word units
Unicode ligatures	All	Removed	الرسول ﷺ → الرسول 1 word → 1 word unit
Mentions & user handles	All	Counted as one word	@arabic_AI_lab → arabicAIlab 3-word handle → 1 word unit

Abbreviations	Single	Each abbreviation counted individually	د. خالد → د. خالد No change, 2 word units
	Abbreviations used as a word	Counted as one word	مكتب أرامكو No change, 2 word units
	Multi-letter abbreviations referring to one entity	Counted as one word	قناة إم بي سي → قناة إمبيسي 4 words → 2 word units

Note: Given the infinite variety of real-world text, users may encounter cases not explicitly listed here. In this case, the overarching philosophies stated in Section 3 should guide the final tokenisation decision.

8) Example

Before tokenisation: 105 words

> والحقيقةُ أنّ التكنولوجيا ستُغيّر حياتنا بشكلٍ كبيرٍ، فهي كالنهر الذي لا يتوقّف عن الجريان.
> أسمعتَ بالذكاء الاصطناعي؟ إنّه يساعدنا في كلّ شيء، ويُسهّل علينا أعمالنا اليومية. لقد
> بدأنا نستخدمهُ في مدارسنا ومستشفياتنا منذ الخامس عشر من شهر يناير عام ألفين وخمسة
> وعشرين، وربما سيصبح جزءاً لا يتجزّأ من مستقبلنا. وسنراهُ قريباً في تطبيقاتٍ جديدةٍ أيضًا
> 🙂 ؛ فقد قرأتُ أمس مقالاً على موقع بي بي سي، ثم شاهدتُ تقريراً قصيراً على MBC
> يتحدّث عن دراسة تضمّ 1327 مشاركاً في مجال الـ AI والـ ML في منطقة الشرق الأوسط،
> ورابطُ الدراسة موجودٌ هنا .https://www.ai-research.edu.sa: لماذا نخافُ منه إذاً؟ بالعلمِ
> والمعرفةِ نستطيع أن نفهمهُ ونوظّفهُ لخدمة الإنسانية. #التكنولوجيا_والمستقبل.

After tokenisation: 130 word units

و+ الحقيقة أن التكنولوجيا س+ تغير حياة +نا ب+ شكل كبير ف+ هي ك+ النهر الذي لا يتوقف عن الجريان أ+ سمعت ب+ الذكاء الاصطناعي إن +ه يساعد +نا في كل شيء و+ يسهل علي +نا أعمال +نا اليومية لقد بدأ +نا نستخدم +ه في مدارس +نا و+ مستشفيات +نا منذ 15 يناير 2025 و+ ربما س+ يصبح جزء لا يتجزأ من مستقبل +نا و+ سنرا +ه قريب في تطبيقات جديدة أيضا ف+ قد قرأت أمس مقال على موقع بيبيسي ثم شاهدت تقرير قصير على MBC يتحدث عن دراسة تضم 1327 مشارك في مجال ال AI و+ ال ML في منطقة الشرق الأوسط و+ رابط الدراسة موجود هنا https://www.ai-research.edu.sa لماذا نخاف من +ه إذا ب+ العلم و+ المعرفة نستطيع أن نفهم +ه و+ نوظف +ه ل+ خدمة الإنسانية التكنولوجيا و+ المستقبل.

If this text contained 10 rhetorical devices, the rhetorical density would be:

- (10/105)*100 = 9.52 rhetorical devices per 100 *words*, but
- (10/130)*100 = 7.69 rhetorical devices per 100 *word units*,

which is almost a 20% discrepancy.