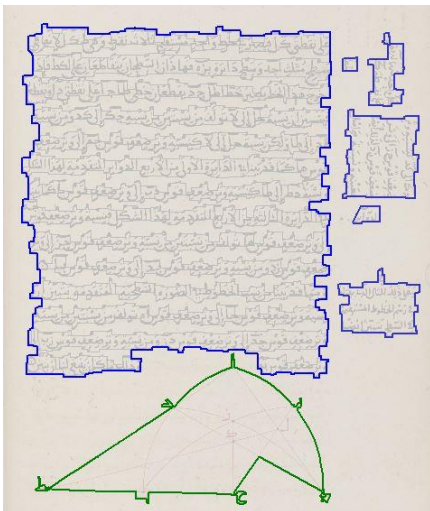


About the Project

Cultural heritage institutions around the world are digitising vast archives containing hundreds of thousands of pages of historical handwritten Arabic manuscript collections. The ability to make these texts fully text searchable has the potential to truly transform research.

Computer scientists are working on this challenge, building systems which can automatically transcribe images of handwritten text, but for handwritten Arabic script a solution remains just out of reach.

Your transcriptions will be turned into much needed [ground truth](#) resources, ensuring historical Arabic collections benefit from state-of-the-art developments in handwritten text recognition, transforming accessibility of this rich content through enabling full-text search and large-scale text analysis.



What is Ground Truth?

[Optical Character Recognition](#) (OCR) systems essentially turn a *picture* of text into text itself—in other words, producing something like a .TXT or .DOC file from a scanned .JPG of a printed or handwritten page. Most OCR systems require [ground truth](#), a set of files which represent the truthful record of elements of an image, for training and evaluation purposes.

The ground truth of an image's text content, for instance, is the complete and accurate record of every character and word in the image.

By knowing what the system is supposed to recognise on a page of handwritten text, researchers can both train their system to recognise the characters as well as test how well the system does once trained.

Helping Transform Research

Our aim is to build an open image and ground truth dataset of historical handwritten Arabic texts to support continued OCR developments in this area. This project is a proof of concept exploring how it might be done at scale. Any data produced in this pilot will be hosted by the British Library and made freely available, without rights restriction, for anyone wishing to advance the state-of-the-art in optical character recognition technology. Specifically, resources created will be used to evaluate entrant systems in the [RASM2018 ICFHR2018 Competition on Recognition of Historical Arabic Scientific Manuscripts](#) and will be contributed to ground-breaking projects such as [Transkribus](#), the [Open Islamic Texts Initiative](#), and the [IMPACT Centre of Competence Image and Ground Truth Resources](#).

A Collaborative Approach

This project is utilising a free and open-source platform, [From the Page](#), which allows volunteers from around the world to experience the manuscripts up close, many for the first time, to discuss, learn and share expertise in their transcription.

The platform allows volunteers to transcribe *as much or little* as they like, *whenever* they like, from *wherever* they like! All contributions are valuable and very much appreciated.

Version control is built-in so there's no need to worry about making a mistake, writing over someone else's transcription or providing an incomplete or less than perfect transcription. All changes are saved within the system, so everything can be restored.

Get in touch

This project is a collaboration between British Library and QDL with special support from Turing Institute and the PRImA Research Lab.

Please contact digitalresearch@bl.uk with any questions.

Getting Started: Transcription Guidelines for Ground Truth

Basic Principle

When transcribing works for ground truth creation the aim is to present as accurate a representation of what is precisely written on the page as possible.

For scholars this might feel awkward and counter-intuitive but for training pattern recognition software this is essential (think diplomatic transcription rather than a scholarly edition).

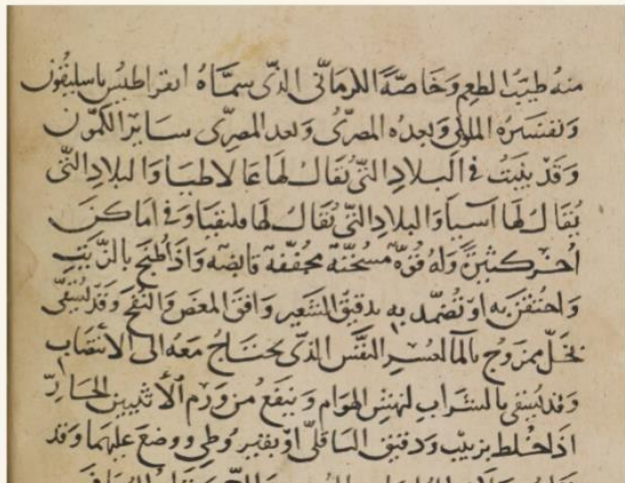
So if you see misspellings, or words are left out, do not correct this, always aim to transcribe exactly what is seen on the page.

When in doubt just have a go, insert [?] in the transcription when anything is unclear, and leave the community a note below the page and we'll all have a look!

General Guidelines

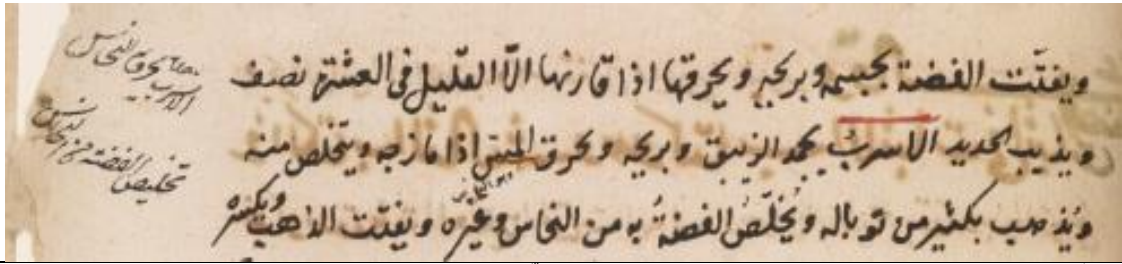
- **Please transcribe the main block of text and main text lines only.**

We've included numbers for each line to make this easier.

Facsimile	Transcription
	<p>[1] منه طيب الطعم وخاصة الكرمانى الذي سماه ابقراطيس ماسليقون [2] وتفسره الملوكي وبعده المصري وبعده المصري ساير الكمون [3] وقد بنيت في البلاد التي يقال لها حالاطيا والبلاد التي [4] يقال لها اسيا والبلاد التي يقال لها ملقيا وفي اماكن [5] اخر كثيرة وله قوة مسخنة مجففة قابضة واذا طبخ بالزيت [6] واحتقن به او تضمد به يدقيق الشعير وافق المغص والنفخ وقد يسقى [7] بخل ممزوج بالماء لفسر النفس الذي يحتاج معه الى الانتصاب [8] وقد يسقى بالشراب لهش الهوام وينفع من ورم الاثنيين [9] الحار [9] اذا خلط بزبيب ودقيق الباقلي او بغيره ووضعه عليهما وقد [10] يقطع سيلان الرطوبات المزمن من الرحم ويقطع الرحم [11] واذا قرب [9] من الالف وهو مسحوق وقد خلط بخل ويصفر البن [12] اذا شرب او تلتخ [فيمن 9] وهو الكمون البري</p>

[British Library, Or. 3306, f. 48v](https://www.bl.uk/manuscripts/Or.3306/f.48v)

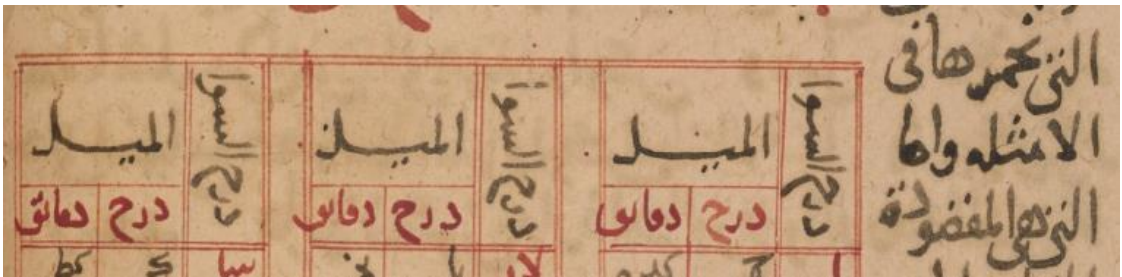
- **Marginalia & interline text:** Do not transcribe at this time.



ويفتت الفضة بجسمه وبريحه ويحرقها اذا قارنها الا القليل في العشرة نصف
ويذيب الحديد الاسرب بجمد الزبيق وبريحه ويحرق المس اذا مزجه ويتخلص منه
ويذهب بكثير من توباله ويخلص الفضة به من النحاس وغيره ويفتت الذهب ويكسره

[British Library, Or. 13006, f. 78r](#)

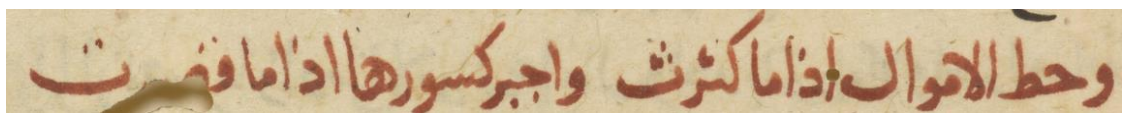
- **Text within diagrams:** Do not transcribe at this time.



التي نحمرها في
الامثله واما
التي هي المقصودة

[British Library, Or 5593, f. 10v](#)

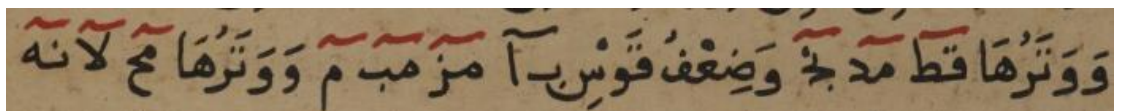
- **Illegible text:** Indicate illegible readings (such as a hole in the page or a character that is not understood) in single square brackets with a [?]



وحط الاموال اذا ما كثرت واجبر كسورها اذا ما قويت
وحط الاموال اذا ما كثرت واجبر كسورها اذا ما قويت [?]

[British Library, Delhi Arabic 1910, f. 25r](#)

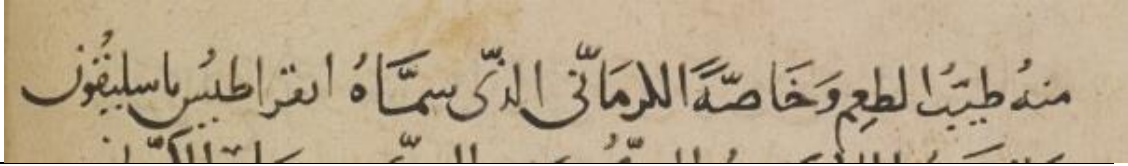
- **Abjad numbers:** Please use initial form of the character



ووترها قط مد نج وضعف قوس ب امز مب م ووترها مح لانه

[British Library, Add MS 7474, f. 18v](#)

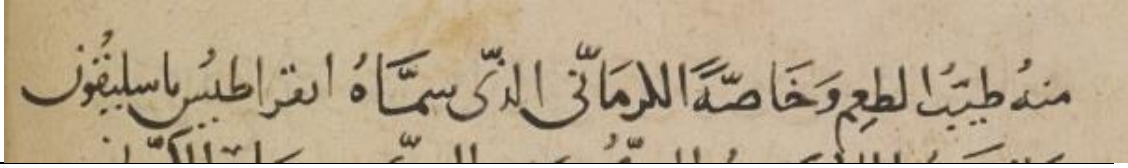
- **Diacritics:** Tashkeel signs should be transcribed as (and only if!) they appear on the page.



منه طيب الطعم وخاصة الكرمانى الذى سماه انقراطيس باسليقون

[British Library, Or 3366, f. 48v](#)

- **Missing punctuation (dots):** Leave them out even if it doesn't feel/read right!



منه طيب الطعم وخاصة الكرمانى الذى سماه انقراطيس باسليقون

[British Library, Or 3366, f. 48v](#)

Typical non-standard letter forms

If you encounter these non-standard letter forms, copy and paste these into your transcription from below. More can be found here but the below are most typical:

<https://www.compart.com/en/unicode/bidiclass/AL>

- ب U+066E Arabic Letter Dotless Beh
- ق U+066F Arabic Letter Dotless Qaf
- ف U+06A1 Arabic Letter Dotless Feh
- ڤ U+06A2 Arabic Letter Feh With Dot Moved Below