BADAM: A Public Dataset for Baseline Detection in Arabic-Script Manuscripts

Benjamin Kiessling [1]
Daniel Stökl Ben Ezra [2]
Matthew Thomas Miller [3]

[1]Université PSL, Leipzig University
[2]EPHE, Université PSL
[3]University of Maryland

Motivation

Layout Analysis is a major preprocessing step in any digitization pipeline.

A community of researchers is working on historical document LA and regular competitions display improvements in the state of the art.

Publicly available datasets and competitions are almost exclusively on Western texts written in the Latin script.

Motivation - Arabic

Arabic and Persian manuscript culture is several times larger than the European one.

A large number of humanities scholars is working on these manuscripts and there is immense interest in digitizing them.

There exists a wide range of topic and styles, often with a complexity of layout rarely encountered in Latin writing.

Bootstrap the research into machine learning-driven methods for Arabic LA.

Baseline Detection

A variety of data models for describing text lines have been developed with rectangular bounding boxes being the most widely used in practice.

These are generally unsuitable for Arabic handwriting or require substantial effort in training data generation.

The baseline detection paradigm as defined by the cBAD dataset is both easy to annotate and requires only minimal adaptation to Arabic handwriting.

Baseline Detection

The annotations in the dataset are baselines, virtual lines on which most characters rest.

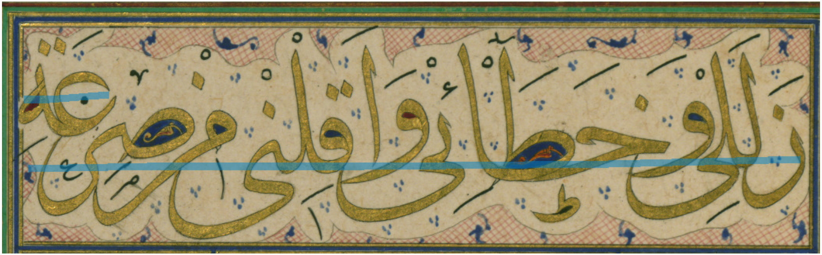
These polylines can generally be arbitrarily shaped.

Currently these are not orientated, i.e. there is no annotation on the direction the line should be read.

Baseline Detection - Arabic

Arabic writing has a number of differences to Western alphabetic scripts.

Naskh and Kufic writing generally has a single baseline per logical text line. Thuluth and Nastaliq display per word slanted baselines.



Walters W579

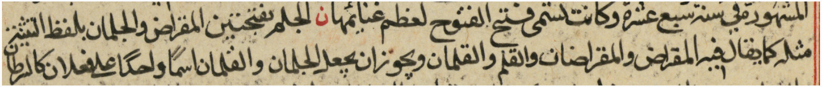
We adapt the baseline definition by drawing a single baseline through an imaginary rotation point at each word.

Baseline Detection - Arabic

Hyphenation is considered unacceptable in Arabic writing.

A number of alternatives which complicate annotation have been developed.

The most common feature is warping of the baseline to extend the available space and placing fragments above the text line.

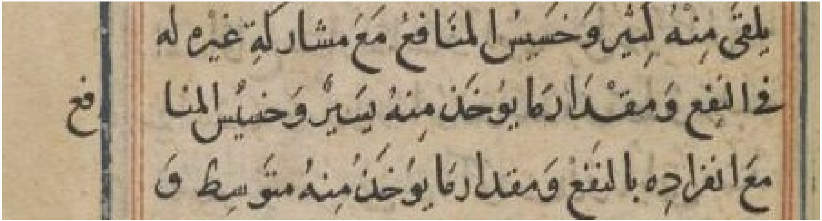


المشهور في سنة سبع عشرة وكانت تسمى فتح الفتوح لعظم غنايمها ان الجلم بفتحتين المقراض والجلمان بلفظ التثنية
مثله كما يقال لخبر المقراض والمقراضان والقلم والقلمان ويجوز ان يجعل الجلمان والقلمان اسما واحدا على فعلان كالسرطان

Walters W590

Baseline Detection - Arabic

Rarely part of the text is expulsed into the margin, instead.



يلقى منه كثير وخسيس المنافع مع مشاركة غيره له

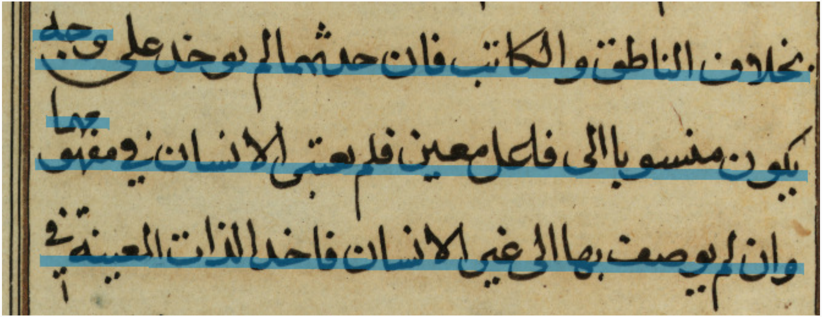
في النفع ومقدار ما يوخذ منه يسير وخسيس المنا فع

مع انفراده بالنفع ومقدار ما يوخذ منه متوسط و

QDL Or 9452

Baseline Detection - Arabic

In cases of majority overlap between main text line and fragment we annotate the fragment as a separate baseline.



بخلاف الناطق والكاتب فان حدثهما لم يوخذ على وجه
يكون منسوبا الى فاعل معين فلم يعتبر الانسان في مفهومها
وان لم يوصف بها الى غير الانسان فاخذ الذات المعينة في

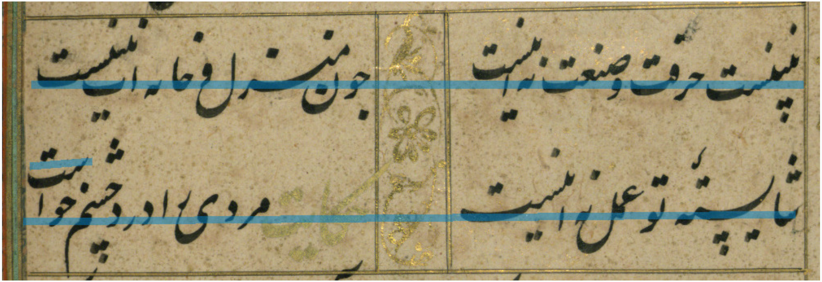
Walters W591

Otherwise it is deemed a warped baseline and annotated as such.

Expulsed fragments are always annotated as separate baselines.

Baseline Detection - Poetry

Verses in Arabic poetry consist almost exclusively of two hemistichs with the half-verse break forming pseudo-columns.

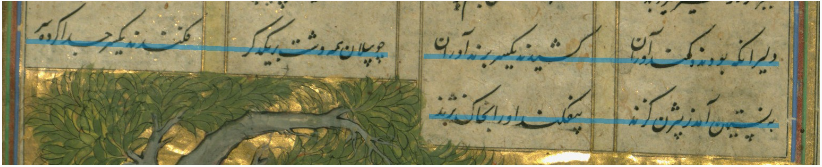


Walter W619

The pseudo-columns are annotated jointly.

Baseline Detection - Poetry

In some cases there is a combination of these pseudo-columns and true multi-column text.

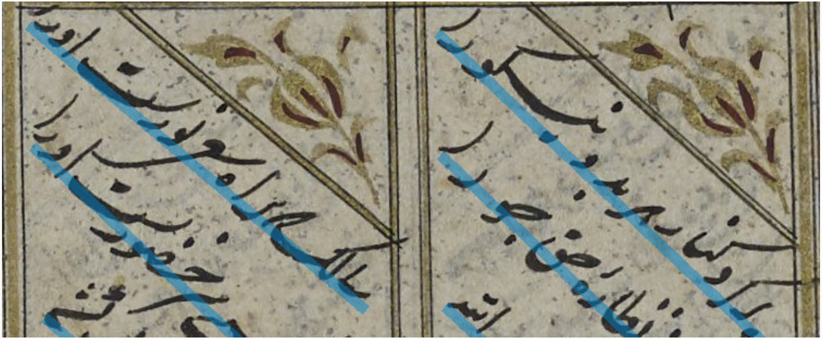


Walter W602

The pseudo-columns are annotated jointly while true columns are not.

Baseline Detection - Poetry

As an additional complication somtimes manuscripts contain 45 degree slanted half-verses.

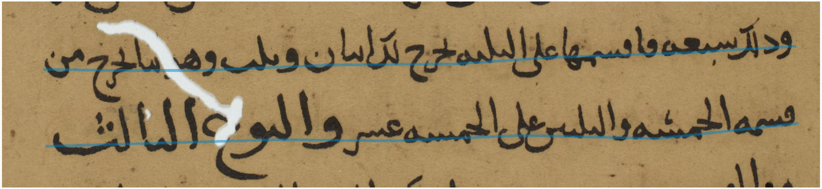


UPenn LJS 44

These are annotated as separate baselines.

Baseline Detection - Fragments and Damage

For manuscripts with faded ink and holes the baselines are continued.



UPenn LJS 293

Dataset

The dataset consists of *400* pages sampled from *42* manuscripts of *4* digital collections:

15	Qatar Digital Library
13	Walters Art Museum
6	Beinecke Rare Book and Manuscript Library
8	University of Pennsylvania Libraries manuscript collection

All images are in the public domain. The annotation are licensed under CC-BY-SA.

Dataset

10 representative pages from each manuscript were annotated with the labelme[1] tool.

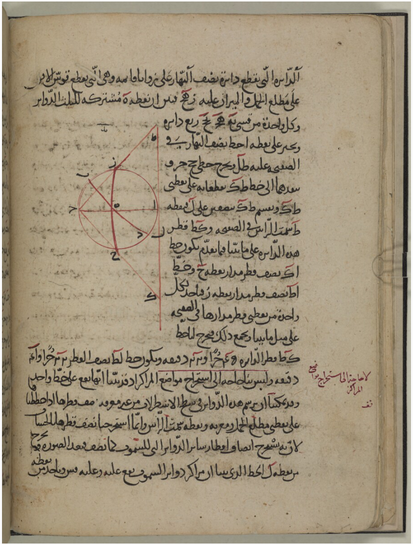
There are *107700* lines in the corpus with a range of 3 to 176 lines per manuscript page.

The majority of the corpus is written in Naskh with the remainder being split between Thuluth, Nastaliq, and Kufic.

[1] https://github.com/wkentaro/labelme

Dataset

Medical, astronomical, and other scientific treatises



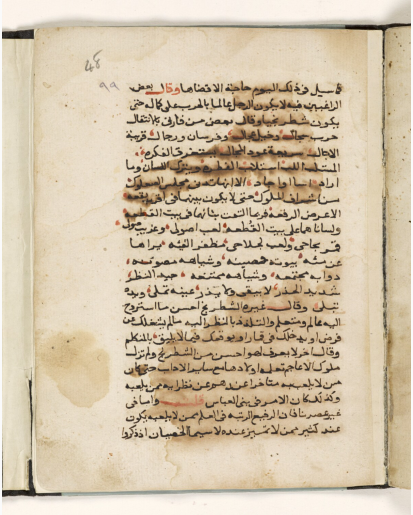
الدّايره التي تقطع دايرة نصف النّهار على زوايا قايمه وهي التي تقطع قوس الار
على مطلع الحمل والميزان عليه زهح ببن ان نقطه ه مشتركه للثلث الدّواير
وكل واحدة من قسيّه هح زح ربع دايره
ونخرج على نقطه ا خط نصف النّهار في
الصفيحه عليه طل ونخرج خطي حح حز و
نفذهما الى خط طك فنقطعانه على نقطتي
طك ونقسم طك بنصفين على ل فنقطه
ل سمت الرّاس في الصفيحه وكط قطر
هذه الدّايره على ما بيّنا فيما تقدّم ويكون خط
اك نصف قطر مدار نقطه ح وخطّ
اط نصف قطر مدار نقطه ز فلناخذ لكل
واحدة من نقطتي قطر مدارها في الصفيحه
على مثل ما بينا ونجمع ذلك فنخرج لنا خط
كط قطر الدّايره ه عح دًا و٥٣ دقيقه ويكون خط لط نصف القطر ٢ مح دًا و٥٦
دقيقه وليس بنا حاجه الى استخراج مواضع المراكز اذ قد بيّنا انها تقع على خطّ واحد
وقد يمكننا ان نرسم هذه الدّواير في سطح الاسطرلاب من غير معرفه نصف قطرها اذا اخططناها
على نقطتي مطلع الحمل ومغربه ونقطه سمت الرّاس وانما استخرجنا نصف قطرها للحساب
لانّه يستخرج انصاف اقطار ساير الدّواير التي للسموت كما نصف بعد الصوره ونخرج
من نقطه ك الخط الذي يسار مراكز دواير السموت ونقع عليه ونعلم عليه فسرد لاحد من نقطه

لا حاجتنا الى استخراج مواضع المراكز
نف

QDL Or 14270

Dataset

Legal texts



48

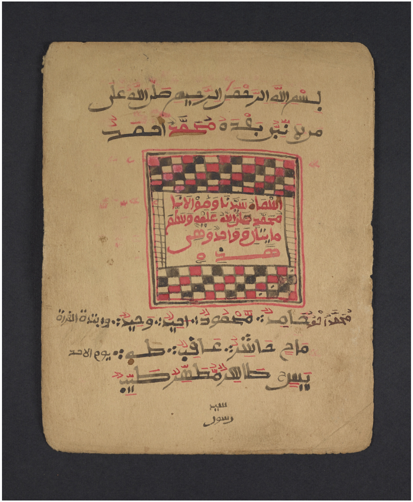
٩٩

فاسيل وذلك اليوم حاجته الا اقتضاها وقال بعض
الراغبين فيه لا يكون الرجل عالما بالحرب على كماله حتى
يكون شطرنجيا وقال بعض من فارق بلا انتقال
حرب سجال وخيل عجال وفرسان ورجال قريبة
الاجال سريعة عود المجال يستخرج الفكر
المستلب اللب استلاب القطع ويترك اللسان وما
اراد اسا او اجاد الا انها تدني مجلس الصعلوك
من اشراف الملوك حتى لا يكون بينهما في اقرب بقعة
الاعرض الرقعة فربما التعت بنا نهما في بيت القطعة
ولسانا هما على بيت القطعة لعب اصولي وعزيب صولي
قمر بجاحي ولعب لجلاحي مظهر الفيه يراها
عن مثه بيوته حصينه وشباهه مصونه
دوابه مجتمعه وشباهه ممتنعه جيد النظر
شديد الحذر لا يبقى ولا يذر عينه تملى ويده
تبلى وقال غيره الشطرنج احسن ما استروح
اليه عالم ومتعلم والتلذذ بالنظر اليه سالم يشغلك عن
فرض او يدخلك في قمار او يوقعك فيما لا يليق بالمتكلم
وقال اخر لا يعرف لهوا احسن من الشطرنج ولم تزل
ملوك الاعاجم تعلمه اولادها مع ساير الاداب حتى كان
من لا يلعبه متاخرا عندهم عن نظرايه ممن يلعبه
وكذلك كان الامر في بني العباس قلت واما في
غير عصرنا فان الرفيع الرتبه في العلم ممن لا يلعبه يكون
عند كثير ممن لا تمييز عنده لا سيما الخصيان اذ ذكروا

QDL Or 9227

Dataset

Prayer books



بسم الله الرحمن الرحيم صلى الله على

من لا نبي بعده محمد احمد

اسماء سيدنا ومولانا
محمد صلى الله عليه وسلم
مائتان وواحد وهي
هذه

محمد احمد حامد محمود احيد وحيد وبشارة القرة

ماح حاشر عاقب طه يوم الاحد

يس طاهر مطهر طيب

سيد
رسول

UPenn CAJS Rar Ms 132

Dataset

Illuminated poetic works

بربط با بانگ دهل بر نیاید و بوی عبیر در برابر سیر چندان نپاید

بلند آواز نادان گردن افراخت
که دانا را به بی‌شرمی بینداخت

نمیداند که آهنگ حجازی
فرو ماند ز بانگ طبل غازی

مشک آنست که ببوید نه آنکه عطار بگوید دانا چون طبلهٔ عطار است
خاموش و نادان چون طبل غازی بلند آوازه و میان تهی

دوستی را که بعمری بچنگ آرند نشاید که بیکدم بیازارند

سنگی که چند سال شود لعل پاره
زنهار تا بیک نفسش نشکنی بسنگ

عقل در دست نفس همچنان گرفتار است که مرد عاجز در دست زن گربز

در خرمی بر سرایی ببند
که بانگ زن از وی برآید بلند

رای بی قوت مکر است

Walters W619

BLLA

Table 1: Results for the cBAD 2017 dataset and BADAM

	P-val	**R-val**	**F-val**
cBAD Simple Track			
BYU	0.878	0.907	0.892
dhSegment	0.943	0.939	0.941
ARU-Net	0.977	0.980	0.978
C-BLLA	0.944	0.966	0.954
BADAM			
C-BLLA	0.941	0.901	0.924

Conclusion and Outlook

BADAM seems to be significantly more difficult than comparable datasets.

Nevertheless, the current data model is insufficient for an operational layout analysis system.

We hope to further annotate logical text lines and separate main text and notes.

Link: `https://doi.org/10.5281/zenodo.3274428`