Lit, Cornelis van, and Dirk Roorda. "Neither Corpus Nor Edition: Building a Pipeline to Make Data Analysis Possible on Medieval Arabic Commentary Traditions." *Journal of Cultural Analytics*, vol. 9, no. 3, June 2024, https://doi.org/10.22148/001c.116372.

ARTICLE

Neither Corpus Nor Edition: Building a Pipeline to Make Data Analysis Possible on Medieval Arabic Commentary Traditions

Cornelis van Lit[1], Dirk Roorda[2]

[1] Utrecht University, [2] Royal Netherlands Academy of Arts and Sciences

Keywords: intertextuality, commentaries, digitization, OCR, text as data

https://doi.org/10.22148/001c.116372

Journal of Cultural Analytics

Vol. 9, Issue 3, 2024

We have built a suite of tools in Python to proficiently analyze text reuse and intertextuality for a specific kind of set of medieval Arabic texts (commentaries) available in print. We take these printed editions, scan them, pre-process the images, give it to an OCR engine, clean the results, and store it in a data structure that mimics the explicit intertextual relation the texts have, and continue to perform data analysis on it. Digital approaches to medieval Arabic texts have either been at the micro-level in what has become known as a 'digital edition', i.e. the digital representation of one text, densely annotated, most commonly in TEI-XML, or it has been done at the macro-level in what is called a 'digital corpus', consisting of thousands of loosely encoded and sparsely annotated plain text files, accompanied by an entire infrastructure and high-performing software to perform broadly scoped queries. The micro-level generally is at the level of tens of thousands of words while the macro-level can be at the level of over a billion words. The micro-level is explicitly designed to be human readable first, while the macro-level is built to be machine readable first. At the micro-level, every little detail needs to be correct and in order, while at the macro-level a fairly large margin of error is still negligible as a mere rounding error. Amidst these levels we have been seeking a meso-level of digital analysis: neither edition nor corpus, but rather a group of texts at the level of hundreds of thousands to millions of words, with a small but perceptible margin of error, and a light but noticeable level of annotations, principally geared towards machine readability, but with ample opportunity for visual inspection and manual correction. In this paper we explain the rationale for our approach, the technical achievements it has led us to, and the results we so far obtained.

In this article we describe a new methodology and a newly developed toolkit to computationally analyze text reuse and intertextuality for a specific kind of medieval Arabic texts available in print. By taking advantage of the explicit (and perhaps at a later stage implicit) intertextuality of post-classical commentaries, we make it exceedingly easier and beneficial to traverse these texts which are otherwise very hard to pierce through. After a methodological introduction, the main emphasis of this paper is the technological side of our inventions. We have chosen to highlight the technology as it is a driving factor of a large part of our methodological advances. We also think that scholars with equal skills in using Python will be able to take advantage of our toolset right now, and perhaps continue to develop it for purposes we have not foreseen, thereby also methodologically developing it in new directions. The tools for the individual steps of the workflow we developed, are useful products in and of themselves.

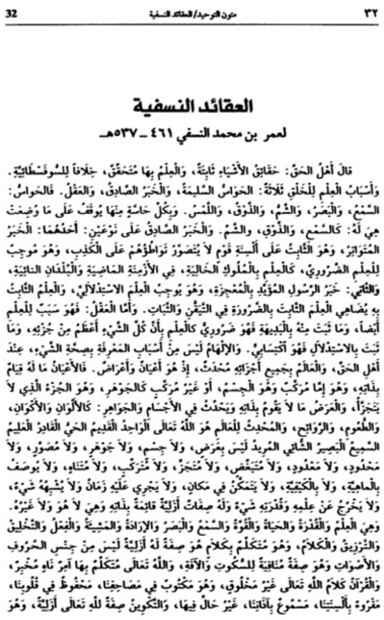
Nonetheless, we think the underlying methodology is equally of interest. It decisively pushes the boundaries of studying the medieval Islamic literary heritage by introducing a new scope of analysis which we call the 'meso-level' of academically producing and studying digital texts: neither corpus nor edition. Digital approaches to medieval Arabic texts have either been at the micro-level in what has become known as a 'digital edition', i.e. the digital representation of one text, densely annotated, most commonly in TEI-XML (see e.g. Samarqandī), or it has been done at the macro-level in what is called a 'digital corpus', consisting of thousands of loosely encoded and sparsely annotated plain text files, accompanied by an entire infrastructure and high-performing software to perform broadly scoped queries (see e.g. Nigst et al.). The micro-level generally is at the level of tens of thousands of words while the macro-level can be at the level of over a billion words. The micro-level is explicitly designed to be human readable first, while the macro-level is built to be machine readable first. At the micro-level, every little detail needs to be correct and in order, while at the macro-level a fairly large margin of error is still negligible as a mere rounding error. Amidst these levels we have been seeking a meso-level of digital analysis: neither edition nor corpus, but rather a group of texts at the level of hundreds of thousands to millions of words, with a small but perceptible margin of error, and a light but noticeable level of annotations, principally geared towards machine readability, but with ample opportunity for visual inspection and manual correction. In this article we explain the rationale for our approach, the technical achievements it has led us to, and the results we so far obtained.

Post-classical commentaries: a genre most suitable for a meso-level approach

Our methodological-technological development grew in answer to a specific research question, namely, to study commentary traditions more effectively. Commentary writing is a staple of the Islamic literary heritage and is in fact the dominant writing style for the post-classical period (which we define as the 12th to 19th centuries CE). Scholars have come to understand that commentary-writing is first and foremost a form-aspect and does not say anything about the ideas contained in them. A commentary is not necessarily redundant or even exegetical towards the text it expands upon (the *matn*), nor is a commentator necessarily an adherent or follower of the earlier author. The form-aspect can take different concrete forms and may therefore best be defined in a loose, abstract sense of having structural textual correspondence. This is the case if a commentary not only evidently relies in structure on a base text, but shows intentional textual correspondence exactly in those places of the base text that characterize its composition (van Lit, "Commentary and Commentary Tradition" 15). One may be eager to divide up such texts into the actor categories such as *sharḥ* ('commentary'), *ḥāshiya* ('gloss'), or *manẓūma* ('versification') but this is fairly inconsequential in terms of understanding what a commentary is about and what a commentator is trying to say. It is more important to track which text a commentary is showing structural textual correspondence to, as it is not exceptional to come across commentaries upon commentaries, sometimes up to five degrees removed from the base text.

Reading higher-degree commentaries is a strange affair: A few words of a previous commentary are cited and concluded by *ilā ākharihi* ('etcetera') and the commentator continues with words of his own, and this is repeated every few words, sentences, or sometimes a paragraph. In other words, a base text, commentary, or super-commentary is divided up into individual topics which a later author uses as a table of contents, or perhaps a better comparison is a buffet, from which he may take whatever he likes. This means that a higher-order commentary can usually only be understood correctly if it is compared with the previous texts it relates to. This explicit intertextual relationship, therefore, should be exploited. Rather than finding all texts by one author, it makes more sense to collect all authors of one text. Closer inspection invariably reveals a much more complex reality: the explicit intertextuality is interspersed with implicit intertextuality as authors frequently cite other authors (for example, fellow commentators) without mentioning them by name or even warning that it is a citation. Readers of such texts were expected to be equally steeped in the literature as the author, so that subtle invocations of other authors would be understood.

To comprehend this sophisticated intertextuality, let us look at the example of the commentary tradition on Nasafī's creed. Abū Ḥafṣ 'Umar al-Nasafī (d. 1142) wrote a short creed which in print runs to four pages. The first page looks like this:



32 متون التوحيد/العقائد النسفية ٣٢

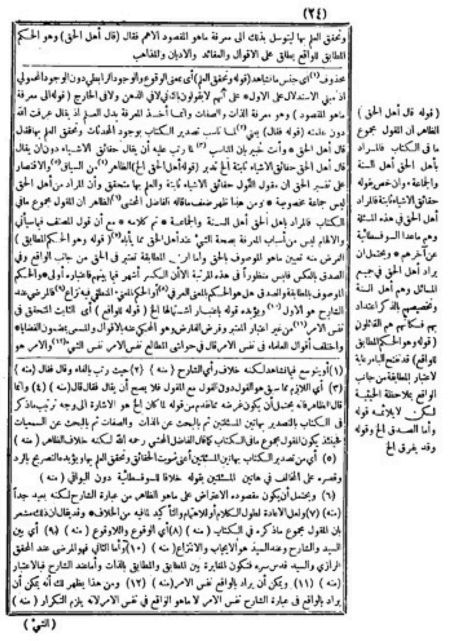
العقائد النسفية

لعمر بن محمد النسفي ٤٦١ ـ ٥٣٧هـ

قَالَ أَهْلُ الحَقِّ: حَقَائِقُ الأَشْيَاءِ ثَابِتَةٌ، وَالْعِلْمُ بِهَا مُتَحَقِّقٌ، خِلَافاً لِلسُّوفَسْطَائِيَّةِ.
وَأَسْبَابُ الْعِلْمِ لِلْخَلْقِ ثَلَاثَةٌ: الحَوَاسُّ السَّلِيمَةُ، وَالْخَبَرُ الصَّادِقُ، وَالعَقْلُ. فَالحَوَاسُّ:
السَّمْعُ، وَالْبَصَرُ، وَالشَّمُّ، وَالذَّوْقُ، وَاللَّمْسُ. وَبِكُلِّ حَاسَّةٍ مِنْهَا يُوقَفُ عَلَى مَا وُضِعَتْ
هِيَ لَهُ: كَالسَّمْعِ، وَالذَّوْقِ، والشَّمِّ. وَالْخَبَرُ الصَّادِقُ عَلَى نَوْعَيْنِ: أَحَدُهُمَا: الْخَبَرُ
المُتَوَاتِرُ، وَهُوَ الثَّابِتُ عَلَى أَلْسِنَةِ قَوْمٍ لاَ يُتَصَوَّرُ تَوَاطُؤُهُمْ عَلَى الْكَذِبِ، وَهُوَ مُوجِبٌ
لِلْعِلْمِ الضَّرُورِيِّ، كَالْعِلْمِ بِالْمُلُوكِ الخَالِيَةِ، فِي الأَزْمِنَةِ المَاضِيَةِ وَالْبُلْدَانِ النَائِيَةِ،
وَالثَّانِي: خَبَرُ الرَّسُولِ المُؤَيَّدِ بِالْمُعْجِزَةِ، وَهُوَ يُوجِبُ الْعِلْمَ الاسْتِدْلَالِيَّ، وَالْعِلْمُ الثَّابِتُ
بِهِ يُضَاهِي الْعِلْمَ الثَّابِتَ بِالضَّرُورَةِ فِي التَّيَقُّنِ وَالثَّبَاتِ. وَأَمَّا الْعَقْلُ: فَهُوَ سَبَبٌ لِلْعِلْمِ
أَيْضاً، وَمَا ثَبَتَ مِنْهُ بِالْبَدِيهَةِ فَهُوَ ضَرُورِيٌّ كَالْعِلْمِ بِأَنَّ كُلَّ الشَّيْءِ أَعْظَمُ مِنْ جُزْئِهِ، وَمَا
ثَبَتَ بِالاسْتِدْلَالِ فَهُوَ اكْتِسَابِيٌّ. وَالإِلْهَامُ لَيْسَ مِنْ أَسْبَابِ المَعْرِفَةِ بِصِحَّةِ الشَّيْءِ، عِنْدَ
أَهْلِ الحَقِّ، وَالْعَالَمُ بِجَمِيعِ أَجْزَائِهِ مُحْدَثٌ، إِذْ هُوَ أَعْيَانٌ وَأَعْرَاضٌ. فَالأَعْيَانُ مَا لَهُ قِيَامٌ
بِذَاتِهِ، وَهُوَ إِمَّا مُرَكَّبٌ وَهُوَ الجِسْمُ، أَوْ غَيْرُ مُرَكَّبٍ كَالجَوْهَرِ، وَهُوَ الجُزْءُ الَّذِي لاَ
يَتَجَزَّأُ، وَالْعَرَضُ مَا لاَ يَقُومُ بِذَاتِهِ وَيَحْدُثُ فِي الأَجْسَامِ وَالجَوَاهِرِ: كَالأَلْوَانِ وَالأَكْوَانِ،
وَالطُّعُومِ، وَالرَّوَائِحِ، وَالمُحْدِثُ لِلْعَالَمِ هُوَ اللَّهُ تَعَالَى الْوَاحِدُ الْقَدِيمُ الحَيُّ الْقَادِرُ الْعَلِيمُ
السَّمِيعُ الْبَصِيرُ الشَّائِي المُرِيدُ لَيْسَ بِعَرَضٍ، وَلاَ جِسْمٍ، وَلاَ جَوْهَرٍ، وَلاَ مُصَوَّرٍ، وَلاَ
مَحْدُودٍ، وَلاَ مَعْدُودٍ، وَلاَ مُتَبَعِّضٍ، وَلاَ مُتَجَزٍّ، وَلاَ مُتَرَكِّبٍ، وَلاَ مُتَنَاهٍ، وَلاَ يُوصَفُ
بِالْمَاهِيَّةِ، وَلاَ بِالْكَيْفِيَّةِ، وَلاَ يَتَمَكَّنُ فِي مَكَانٍ، وَلاَ يَجْرِي عَلَيْهِ زَمَانٌ وَلاَ يُشْبِهُهُ شَيْءٌ،
وَلاَ يَخْرُجُ عَنْ عِلْمِهِ وَقُدْرَتِهِ شَيْءٌ وَلَهُ صِفَاتٌ أَزَلِيَّةٌ قَائِمَةٌ بِذَاتِهِ وَهِيَ لاَ هُوَ وَلاَ غَيْرُهُ.
وَهِيَ الْعِلْمُ وَالْقُدْرَةُ وَالحَيَاةُ وَالْقُوَّةُ وَالسَّمْعُ وَالْبَصَرُ وَالإِرَادَةُ وَالمَشِيئَةُ وَالْفِعْلُ وَالتَّخْلِيقُ
وَالتَّرْزِيقُ وَالْكَلَامُ، وَهُوَ مُتَكَلِّمٌ بِكَلَامٍ هُوَ صِفَةٌ لَهُ أَزَلِيَّةٌ لَيْسَ مِنْ جِنْسِ الحُرُوفِ
وَالأَصْوَاتِ وَهُوَ صِفَةٌ مُنَافِيَةٌ لِلسُّكُوتِ وَالآفَةِ، وَاللَّهُ تَعَالَى مُتَكَلِّمٌ بِهَا آمِرٌ نَاهٍ مُخْبِرٌ،
وَالْقُرْآنُ كَلَامُ اللَّهِ تَعَالَى غَيْرُ مَخْلُوقٍ، وَهُوَ مَكْتُوبٌ فِي مَصَاحِفِنَا، مَحْفُوظٌ فِي قُلُوبِنَا،
مَقْرُوءٌ بِأَلْسِنَتِنَا، مَسْمُوعٌ بِآذَانِنَا، غَيْرُ حَالٍّ فِيهَا، وَالتَّكْوِينُ صِفَةٌ لِلَّهِ تَعَالَى أَزَلِيَّةٌ، وَهُوَ

Figure 1. Example of base text

On this creed, many commentaries were written. One commentary stood out: the one by Saʿd al-Dīn al-Taftāzānī (d. 1390). This commentary received itself a great many super-commentaries. A typical printed representation of these texts is as shown below.



(٢٤)

وتحقق العلم بها ليتوسل بذلك الى معرفة ماهو المقصود الاهم فقال (قال أهل الحق) وهو الحكم
المطابق للواقع يطلق على الاقوال والعقائد والاديان والمذاهب

محذوف⁽¹⁾ اي جنس ماشاهد (قوله وتحقق العلم) أي بمعنى الوقوع والوجود الرابطي دون الوجود المحمولي
اذ مبنى الاستدلال على الاول • على أنهم لايقولون بذلك في لافي الذهن ولافي الخارج (قوله الى معرفة
ماهو المقصود) وهو معرفة الذات والصفات وانما أخذ المعرفة بدل العلم اذ يقال عرفت الله
دون علمته (قوله فقال) يعني⁽²⁾ لما ناسب تصدير الكتاب بوجود المحدثات وتحقق العلم بها فقال
قال أهل الحق • وأنت خبير بأن التناسب⁽³⁾ لا رتب عليه أن يقال حقائق الاشياء دون ان يقال
قال أهل الحق حقائق الاشياء ثابتة الخ تدبر (قوله أهل الحق الخ) الظاهر⁽⁴⁾ من السياق⁽⁵⁾ والاقتصار
على تفسير الحق ان مقول القول حقائق الاشياء ثابتة والعلم بها متحقق وأن المراد من أهل الحق
ليس جماعة مخصوصة • ومن هذا ظهر ضعف ماقاله الفاضل الخيالي⁽⁶⁾ الظاهر ان المقول مجموع مافي
الكتاب فالمراد بأهل الحق أهل السنة والجماعة • تم كلامه • مع أن قول المصنف فيما يأتي
والالهام ليس من أسباب المعرفة بصحة الشيء عند أهل الحق مما يأباه⁽⁷⁾ (قوله وهو الحكم المطابق)
الغرض منه تعيين ماهو الموصوف بالحق وأما ان المطابقة تعتبر في الحق من جانب الواقع وفي
الصدق بالعكس فليس منظوراً في هذه المرتبة الأان الكسر أشهر فيما بينهم فاعتباره أولى • والحكم
الموصوف بالمطابقة والصدق هل هو الحكم بالمعنى العرفي⁽⁸⁾ أو الحكم المنطقي المتعلق به نزاع⁽⁹⁾ فالمرضي عند
الشارح هو الاول⁽¹⁰⁾ ويؤيده قوله باعتبار اشتمالها الخ (قوله للواقع) أي الثابت المتحقق في
نفس الامر⁽¹¹⁾ من غير اعتبار المعتبر وفرض الفارض وهو المحكي عنه بالاقوال والمسمى بمضمون القضايا •
واختلف أقوال العلماء في نفس الامر قال في حواشي المطالع نفس الامر نفس الشيء⁽¹²⁾ والامر هو

(١) أو يتوسع فيما شاهد لكنه خلاف رأي الشارح (منه) (٢) حيث رتب بالفاء وقال فقال (منه)
(٣) أي اللازم مما سبق هو القول دون المقول مع المقول فلا يصح أن يقال فقال قال (منه) (٤) وانما
قال الظاهر فانه يحتمل أن يكون غرضه مما تقدم من قوله لما كان الخ هو الاشارة الى وجه ترتيب ماذكر
في الكتاب بالتصدير بهاتين المسئلتين ثم بالبحث عن الذات والصفات ثم بالبحث عن السمعيات
فحينئذ يكون المقول مجموع مافي الكتاب كما قال الفاضل الخيالي رحمه الله لكنه خلاف الظاهر (منه)
(٥) أي من تصدير الكتاب بهاتين المسئلتين أعني ثبوت الحقائق وتحقق العلم بها ويؤيده التصريح بالرد
وقصره على المخالف في هاتين المسئلتين بقوله خلافا للسوفسطائية دون البواقي (منه)
(٦) ويحتمل أن يكون مقصوده الاعتراض على ماهو الظاهر من عبارة الشارح لكنه بعيد جداً
(منه) (٧) ولعل الاعادة لطول الكلام أو للاهتمام والتأكيد لما فيه من الخلاف • وقد قال ان ذلك مشعر
بان المقول مجموع ماذكره في الكتاب (منه) (٨) أي الوقوع واللاوقوع (منه) (٩) أي بين
السيد والشارح وعند السيد هو الايجاب والانتزاع (منه) (١٠) وأما الثاني فهو المرضي عند المحقق
الرازي والسيد قدس سره فتكون المغايرة بين المطابق والمطابق بالذات وأما عند الشارح فبالاعتبار
(منه) (١١) ويمكن أن يراد بالواقع نفس الامر (منه) (١٢) ومن هذا يظهر لك أنه يمكن أن
يراد بالواقع في عبارة الشارح نفس الامر لا ماهو الواقع في نفس الامر لانه يلزم التكرار (منه)

(الشيء)

(قوله قال أهل الحق)
الظاهر ان المقول مجموع
ما في الكتاب فالمراد
بأهل الحق أهل السنة
والجماعة • وان خص قوله
حقائق الاشياء ثابتة فالمراد
أهل الحق في هذه المسئلة
وهم ماعدا السوفسطائية
عن آخرهم • ويحتمل ان
يراد أهل الحق في جميع
المسائل وهم أهل السنة
وتخصيصهم بالذكر اعتداد
بهم فكأنهم هم القائلون
(قوله وهو الحكم المطابق
للواقع) قد تمنع الباء رعاية
لاعتبار المطابقة من جانب
الواقع بملاحظة الحيثية
لكن لايلائمه قوله
وأما الصدق الخ وقوله
وقد يفرق الخ

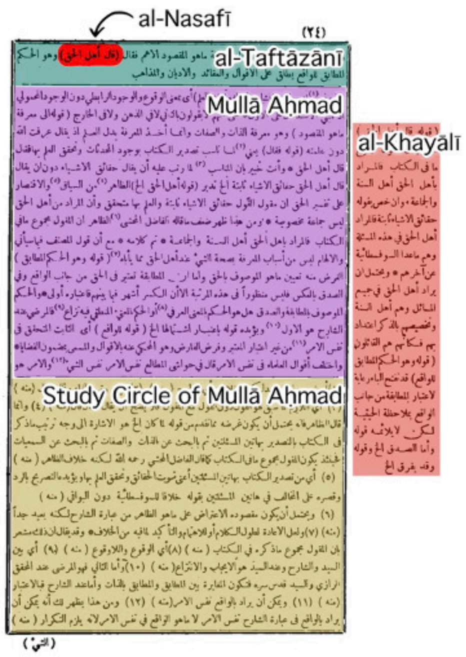

al-Nasafī
al-Taftāzānī
Mullā Aḥmad
al-Khayālī
Study Circle of Mullā Aḥmad

Figure 2. Example of typesetting four texts on one page. On the left a blank example, on the right the different parts of the page highlighted according to their source.

At the top, in teal, we see a snippet of al-Taftāzānī's text, which expands on the first three words of Nasafī's text which are cited in-text and highlighted in bright red. This passage, in turn, spawns the three other texts that are typeset on the page: Mullā Aḥmad's commentary on al-Taftāzānī in purple, notes from him or his students in yellow, and in red, on the side, the commentary on al-Taftāzānī by al-Khayālī.

Al-Khayālī's super-commentary in turn spawned many super-super-commentaries. The most famous one is by al-Siyalkūtī. A printed rendition of al-Siyalkūtī's text is given below.

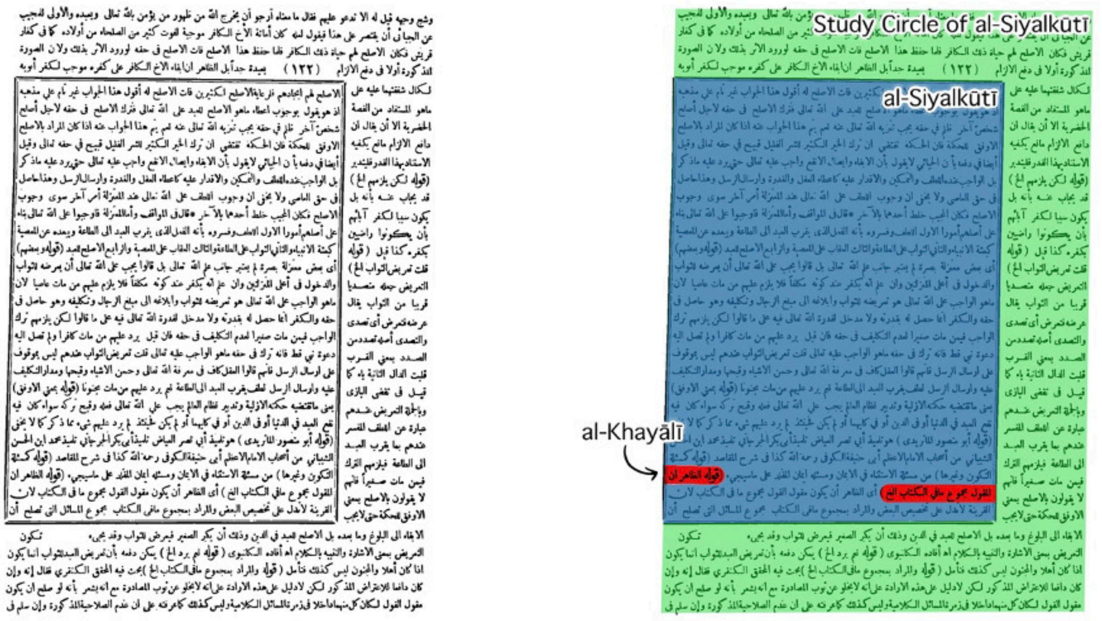


Figure 3. Example of typesetting of higher-order commentaries. On the left a blank example, on the right the different parts of the page highlighted according to their source.

In bright red are a few words from al-Khayālī's text, cited in-text by al-Siyalkūtī (given in blue), which is surrounded by additional notes from later readers of al-Siyalkūtī. By the time al-Siyalkūtī was expanding upon al-Khayālī, the original words by al-Nasafī were even left out. In fact, even al-Khayālī's text is not cited in full by al-Siyalkūtī. He merely cites the first few words of the passage he wants to expand upon and then continues with his own deliberations. Thus, al-Khayālī's passage is divided up into several sub-passages, with each a deliberation.

As one can surmise: from just the first three words of al-Nasafī's creed, an entire ocean of literature evolved, each related to the others. Al-Taftāzānī turned it into 56 words. This passage of 56 words was commented upon by al-Khayālī in 206 words. Eventually, al-Siyalkūtī expanded upon this passage making it grow to 1678 words. Obviously, having only al-Siyalkūtī's text at one's disposal leaves a reader perplexed. The density and references to texts contained in other books is simply too much to overcome by close reading only.

The relationship that each text has with its predecessor can be given in the diagram of Figure 4, which should be read from right to left. On the right of the page, we see the first three words by al-Nasafī (*qāla ahl al-ḥaqq*, 'the people of truth say'). In the next column, in green, we see what al-Taftāzānī added to it. This passage as a whole 'hangs' onto the three words by al-Nasafī. It itself has been divided by al-Khayālī, who made comments about six parts of al-Taftāzānī's passage (given in purple). In other words, al-Khayālī made six comments each 'hanging' onto a different part of al-Taftāzānī's passage. The entire passage by al-Khayālī was grounds for al-Siyalkūtī to write seventeen comments. With this schema it becomes much easier to understand the meaning of any of these texts.

Teasing out the explicit and implicit intertextuality in order to correctly contextualize the commentary and fruitfully understand its contents is possible to do by hand if the commentary is not too big and the commentary tradition not too large. The example given above of the commentary tradition on al-Nasafī's creed shows that doing so for the entire text by al-Nasafī borders at the impossible. The microlevel of a highly detailed and annotated digital edition is out of the question, when the first three words alone balloon up to 1678 words three commentary-levels later. But the macro level is not satisfactory either. As we noticed, these 1678 words have a precise (seventeen-fold) relationship with al-Khayālī's text which in turn has a (sixfold) relationship with al-Taftāzānī's text which itself is directly related to the three words written by al-Nasafī. Thus, simply dumping all four texts into a database is a gross misunderstanding of their inner structure and will invariably result in shallow analysis (if at all). What such midsize sets of texts with intricate relationships need, is a meso-level approach: neither corpus nor edition. Indeed, if all these intertextual relations are saved in the digital domain but without taking over the critical apparatus of the editions on which the digital rendering of the text is based, or even the accuracy of the text itself would be somewhat lacking, it would probably still be just fine for most analytical purposes.

The specifics of Ibn ʿArabī's commentary tradition

Our intention was, and is, to apply this meso-level approach to the commentary tradition on the twelfth century muslim mystic-philosopher Ibn ʿArabī's *Fuṣūṣ al-ḥikam*. This comes with its own peculiar pros and cons. The most important downside is that the total number of commentaries is well above a hundred (from all centuries up until today and all parts of the Muslim world), and even if we only concern ourselves with the most popular ones this still amounts to a work load too great for manual inspection

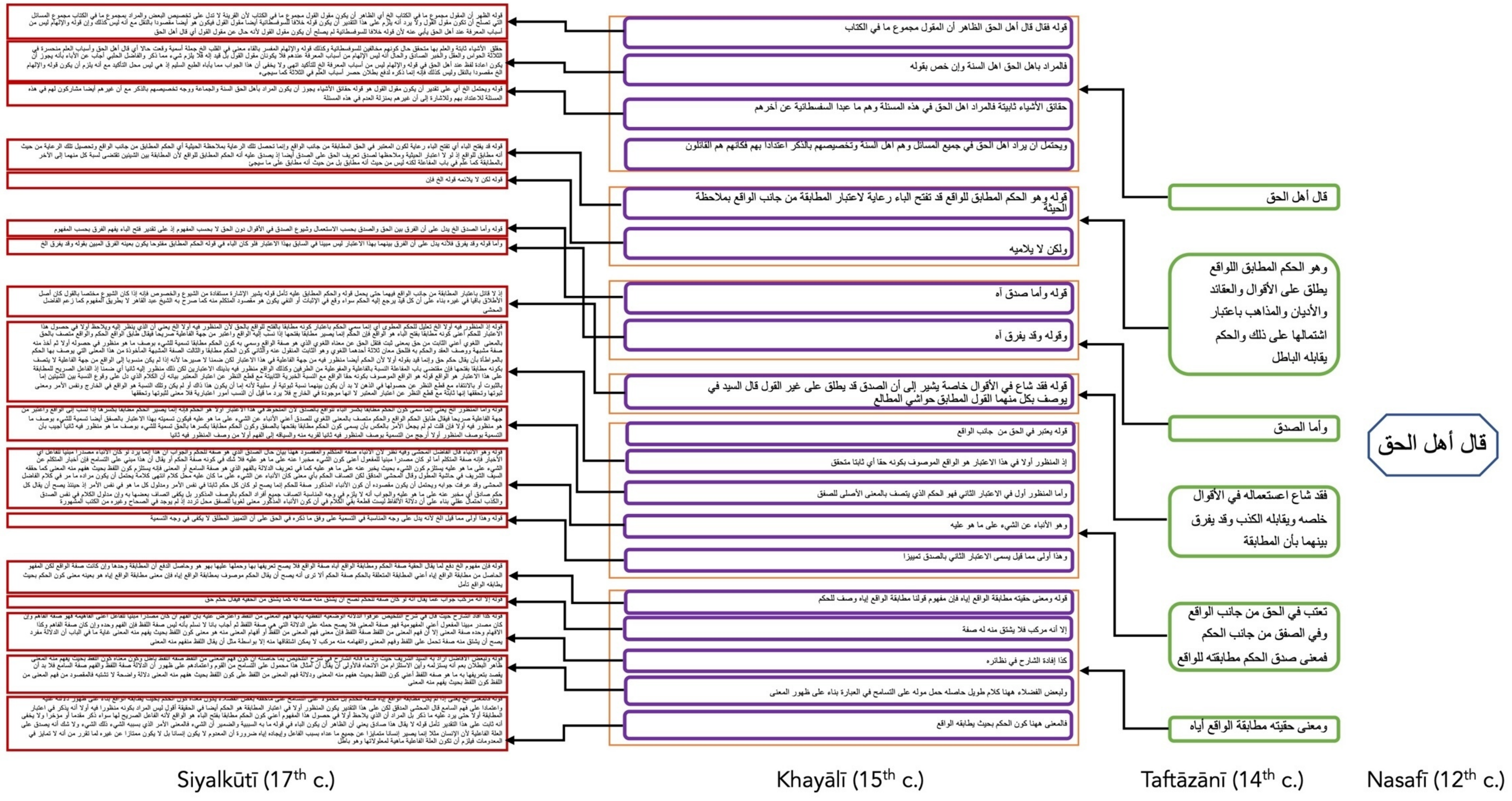
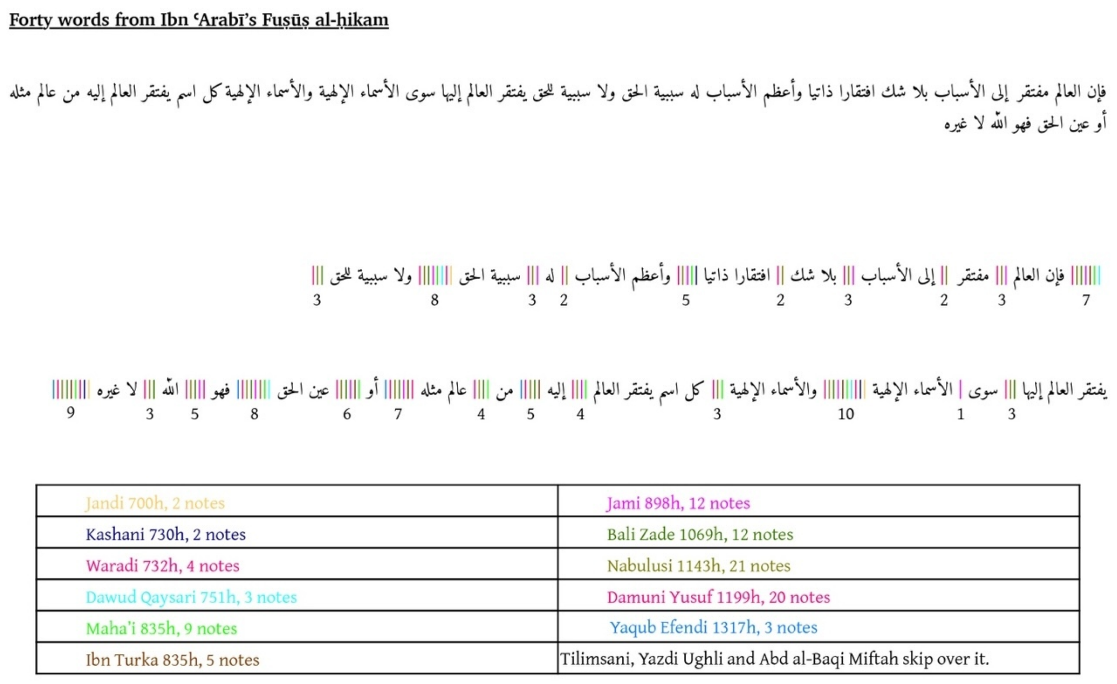

قال أهل الحق
قال أهل الحق
وهو الحكم المطابق اللواقع يطلق على الأقوال والعقائد والأديان والمذاهب باعتبار اشتمالها على ذلك والحكم يقابله الباطل
وأما الصدق
فقد شاع اعستعماله في الأقوال خلصه ويقابله الكذب وقد يفرق بينهما بأن المطابقة
تعتب في الحق من جانب الواقع وفي الصفق من جانب الحكم فمعنى صدق الحكم مطابقته للواقع
ومعنى حقيته مطابقة الواقع أياه
قوله فقال قال أهل الحق الظاهر أن المقول مجموع ما في الكتاب
فالمراد باهل الحق اهل السنة وإن خص بقوله
حقائق الأشياء ثابيتة فالمراد اهل الحق في هذه المسئلة وهم ما عبدا السفسطائية عن آخرهم
ويحتمل ان يراد اهل الحق في جميع المسائل وهم اهل السنة وتخصيصهم بالذكر اعتدادا بهم فكانهم هم القائلون
قوله وهو الحكم المطابق للواقع قد تفتح الباء رعاية لاعتبار المطابقة من جانب الواقع بملاحظة الحيثة
ولكن لا يلاميه
قوله وأما صدق آه
وقوله وقد يفرق آه
قوله فقد شاع في الأقوال خاصة يشير إلى أن الصدق قد يطلق على غير القول قال السيد في يوصف بكل منهما القول المطابق حواشي المطالع
قوله يعتبر في الحق من جانب الواقع
إذ المنظور أولا في هذا الاعتبار هو الواقع الموصوف بكونه حقا أي ثابتا متحقق
وأما المنظور أول في الاعتبار الثاني فهو الحكم الذي يتصف بالمعنى الأصلى للصفق
وهو الأنباء عن الشيء على ما هو عليه
وهذا أولى مما قيل يسمى الاعتبار الثاني بالصدق تمييزا
قوله ومعنى حقيته مطابقة الواقع إياه فإن مفهوم قولنا مطابقة الواقع إياه وصف للحكم
إلا أنه مركب فلا يشتق منه له صفة
كذا إفادة الشارح في نظائره
ولبعض الفضلاء ههنا كلام طويل حاصله حمل موله على التسامح في العبارة بناء على ظهور المعنى
فالمعنى ههنا كون الحكم بحيث يطابقه الواقع
Siyalkūtī (17th c.)
Khayālī (15th c.)
Taftāzānī (14th c.)
Nasafī (12th c.)

Figure 4. Example of three levels of commentarial intertextual relations

(van Lit, "Ibn ʿArabī's School of Thought: Philosophical Commentaries, Not a Sufi Order"). The most important upside is that all commentaries are (or purport to be) direct, first-level commentaries on Ibn ʿArabī's *Fuṣūṣ al-ḥikam*. This means that the ever-more intricate intertextual relationships of normal commentary traditions, which produce higher level commentaries with each century, need not to be captured. Instead, to develop a meso-level methodology and a concomitant technological toolkit, we only need to solve the simpler case of the relationship between a base text (*matn*) and a commentary (*sharḥ*).



Forty words from Ibn ʿArabī's Fuṣūṣ al-ḥikam

فإن العالم مفتقر إلى الأسباب بلا شك افتقارا ذاتيا وأعظم الأسباب له سببية الحق ولا سببية للحق يفتقر العالم إليها سوى الأسماء الإلهية والأسماء الإلهية كل اسم يفتقر العالم إليه من عالم مثله أو عين الحق فهو الله لا غيره

||||||| فإن العالم ||| مفتقر || إلى الأسباب ||| بلا شك || افتقارا ذاتيا ||||| وأعظم الأسباب || له ||| سببية الحق |||||||| ولا سببية للحق |||

3 8 3 2 5 2 3 2 3 7

يفتقر العالم إليها ||| سوى | الأسماء الإلهية |||||||||| والأسماء الإلهية ||| كل اسم يفتقر العالم |||| إليه ||||| من |||| عالم مثله ||||||| أو |||||| عين الحق |||||||| فهو ||||| الله ||| لا غيره |||||||||

9 3 5 8 6 7 4 5 4 3 10 1 3

Jandi 700h, 2 notes	Jami 898h, 12 notes
Kashani 730h, 2 notes	Bali Zade 1069h, 12 notes
Waradi 732h, 4 notes	Nabulusi 1143h, 21 notes
Dawud Qaysari 751h, 3 notes	Damuni Yusuf 1199h, 20 notes
Maha'i 835h, 9 notes	Yaqub Efendi 1317h, 3 notes
Ibn Turka 835h, 5 notes	Tilimsani, Yazdi Ughli and Abd al-Baqi Miftah skip over it.

Figure 5. Interjections by commentators on a forty-word passage from Ibn ʿArabī

In this sense, every text from the commentary tradition on Ibn ʿArabī's *Fuṣūṣ al-ḥikam* 'hangs' onto the base text itself. Thus, we can take a commentary, and divide it up according to its interventions upon the *Fuṣūṣ*. We thereby use the *Fuṣūṣ* as an index, and whenever a commentary cites the *Fuṣūṣ* and then interjects with additional words, we attach those words to the index of the last word cited from the *Fuṣūṣ*. Based on the previous last word cited and the current one, we can compute the passage in the *Fuṣūṣ* which a commentator expands upon. This divides a commentary into semantic units according to where it breaks into the base text. Doing this for several commentaries will allow a few immediate gains: from this data structure it will become clear how the base text was used as a platform to develop the discourse, as we can now study the break-in structure and see how different commentators break into the base text similarly or differently. If commentators contributed to the discourse in the same place (snippets of commentaries are attached to the same index), we can easily compare the commentaries for text reuse, similarity in vocabulary (topic modeling), and the like. Observe in Figure 5 a manually created example to illustrate the approach.

At the top we see a forty-word passage from *Fuṣūṣ al-ḥikam* and below we see it again, but with vertical lines whenever a commentator interjects Ibn ʿArabī's text with words of his own, each commentator using a different color. The number under the lines tell us how many commentators interject at that point. We may notice that the fourteen commentators used, are not once in agreement on where to interject the text. At the same time, we do see groupings of interjections, for example Jāmī and Nābulūsī are almost always in agreement. If we make the length of the vertical interjection lines correspond with the number of words of the interjection, and we abstract away the forty words into a line, we can get the following figure which should be read from right to left.

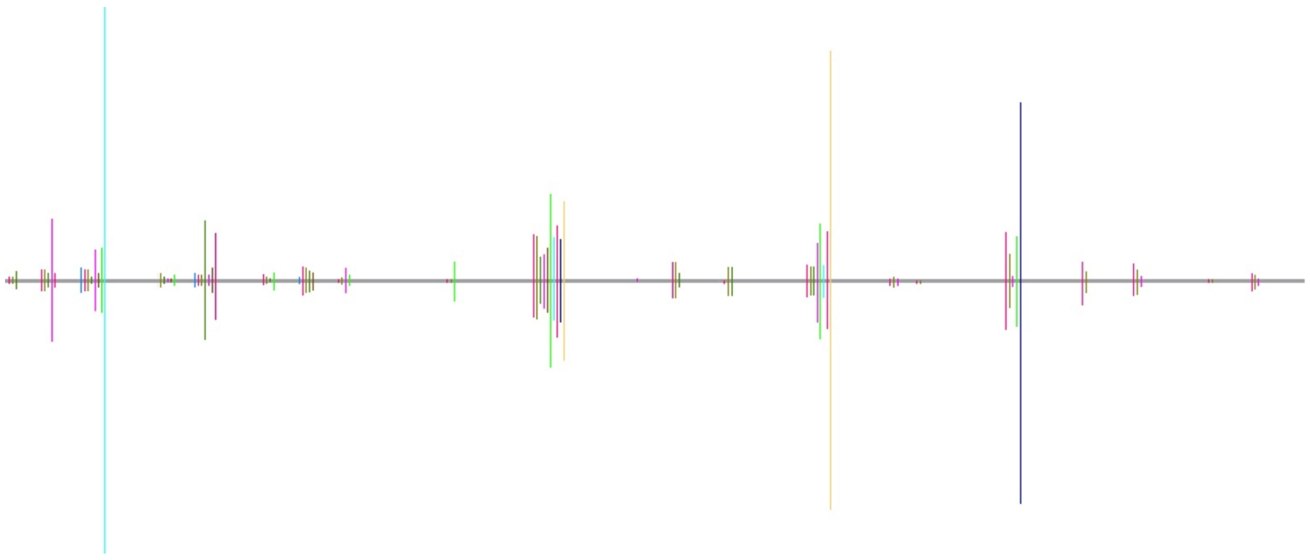


Figure 6. A 'heartbeat' of a commentary tradition

In this figure we see the forty words represented by the grey line, from right to left. Whenever there is an interjection by a commentator, a vertical line appears on it. The visualization looks somewhat like an electrocardiogram as we know it from measuring electrical activity from our heart's contraction and thus I coin this visualization the heartbeat of a commentary tradition. In one visualization, we see what kind of effect the source text had on generations of commentators and how together they make up a unique discourse. Doing this for forty words only works as a proof of concept and is fairly meaningless in itself. But at scale, the benefits are hopefully obvious: we can immediately measure on which parts of the source text the most commentator-activity was, we can easily discern groupings of commentators and obvious partitions of the source text, and it will be possible to see how work on the source text may have evolved over the centuries. At this meso-scale, between corpus and edition, such computational methods go hand in hand with human readability. If we can model our data in this fashion, it will be easy to only read parts of a commentary we deem relevant (for one reason or another). Even more interesting, we can easily call up all commentaries on a particular passage to concurrently read them, for example in search of implicit intertextuality. This can be aided by semi-automated analyses such as sentiment or vocabulary analysis on the separate interjections.

This 'heartbeat' is not the specific subject of this paper, but only its pole star towards which our research is guided. Instead, we are interested in setting up the machinery to make it generally possible to get to a place to do this kind of work, and a specific focus on how a source text can meaningfully function as an index onto which commentaries can place interjections. This is complicated by the historical attrition that any text faces, falling apart into variants, additions, sometimes even different versions. To combat this, we approach the source text as a commentary onto a synthetic self: the true index which we will use is a product of our own making, onto which we can apply annotations to ensure that all necessary information (such as a critical apparatus) is enclosed in our data set. This means our first result is a calculation of the purity of the source text itself, but to this we shall return in the conclusion. It also means that we are, in this article, only concerned with two texts and they are actually two renderings of the same text: the 1946 edition of *Fuṣūṣ al-ḥikam* by Abu al-Ala al-Afifi and the 2013 edition of *Fuṣūṣ al-ḥikam* by Nizam al-Din Ahmad al-Husayni al-Lakhnawi. With these two texts we are ready to 1) construct an index onto which commentaries can hang up their comments, 2) render the base text, and 3) annotate it with meaningful information such as chapter divisions.

From methodology to technological workflow

The abstract idea of what we wanted to achieve turned into a multi-headed monster once we wanted to make things concrete. For better or for worse, the two renditions of the *Fuṣūṣ* were obtained in two different formats. We had a paper copy of Afifi's edition, and a digitally typeset PDF of Lakhnawi's edition. This means we had to invent two primary stages to get each of their text into a plain text-format ready for comparison. For the paper copy, we developed a pipeline which prepares scans of printed Arabic texts for OCR. For the digital PDF, we developed a toolkit to extract Arabic texts from digitally typeset PDFs. For both, we developed a converter to deliver the output from either pipeline or toolkit into a unified data structure. Together they make up our *fusus* library for Python (see Lit and Roorda).

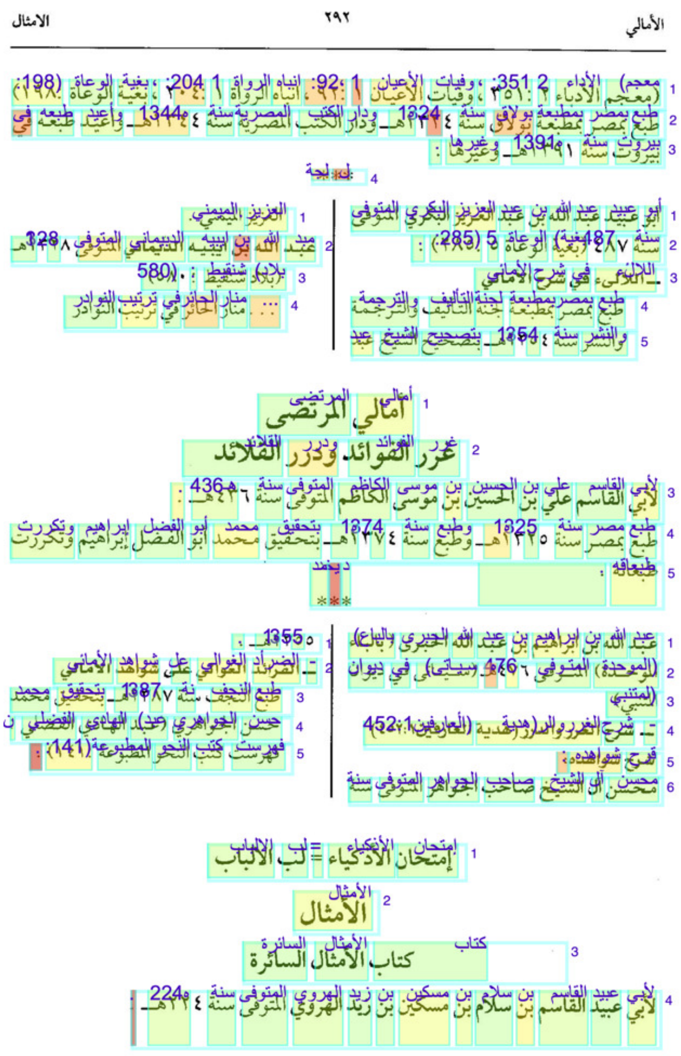
Pipeline for OCR-preparation

Our main source of information are JPGs of scans or photos of printed books. These can be obtained in various ways but become quite uniform by running everything through the image processing software ScanTailor. This ensures that every image 1) shows only one page, 2) is to some extent straightened (some skewing is allowed as it is double checked and fixed later on), 3) is black and white, 4) has reduced or removed noise, and 5) is 300ppi (higher than this causes performance issues). The variety that is left is almost solely due to differences in typesetting.

As we think that encoding is an editorial practice, we perceived our end goal as a digital edition and thus asked ourselves which information we wanted to capture and encode. We decided the text in as much a pure and simple form is what we wanted, with our greatest critical editorial practice being to model and structure the text of each commentary according to the relationship it has with the base text. In other words, we are after the abstract form of the 'work' (Riva et al. 19–23). Of course, we can only access the work through an expression (a specific edition), but since we are aiming at the work, we will leave the critical apparatus out of our scope. Additionally, only for the *Fuṣūṣ al-ḥikam* itself do we consider more than one edition.

Next to a critical apparatus, the layout of an edition is a significant editorial effort and similar to typesetting it is a major hurdle to overcome. In our experience, print editions are advanced systems of information that can combine a great number of information sources and parameters. The base unit this is done on is the page. We can speak, indeed, of the hegemony of the page (van Lit, *Among Digitized Manuscripts* 14–15). This means that the space of a page is (often times) maximally utilized. By placing key pieces of information in specific ways, the editor can signal all kinds of information to the reader. Interestingly, as regular as this may sound, it is exceedingly hard to automate the capture of its dynamics. It is, essentially, fine-tuned for human readability and not for computer readability. We decided not to 'reverse engineer' this human readability into computer readability, and instead we focused our efforts on bypassing the layout as much as possible. This means we would drop a significant amount of information but this would be perfectly in line with our principle of going after the 'work', not the 'expression' of the text.

We further wished to make use of existing OCR-technology and put our own efforts into building a pipeline around it. Experimentation led us to choose Kraken as the best choice 'out of the box' (see Kraken), especially in conjunction with an existing model trained on old printed Arabic books (see OpenITI). Since these experiments were somewhat haphazard, we cannot report in full on it. However, next to our choice for Kraken we noticed performance gains if we would offer a very clean image with ample white space. Since Kraken can give back X,Y coordinates and a confidence level, it was easy to transpose the OCR results onto the original scan, color coding those words which Kraken itself was not sure about, to allow for manual inspection (see figure below).


الأمالي ٢٩٢ الامثال

(معجم الأدباء ٢ : ٣٥١ ، وفيات الأعيان ١ : ٩٢ ، انباه الرواة ١ : ٢٠٤ ، بغية الوعاة ١٩٨)
طبع بمصر بمطبعة بولاق سنة ١٣٢٤ هـ ودار الكتب المصرية سنة ١٣٤٤ هـ وأعيد طبعه في
بيروت سنة ١٣٩١ هـ وغيرها :

أبو عبيد عبد الله بن عبد العزيز البكري المتوفى
سنة ٤٨٧ (بغية الوعاة ٥ : ٢٨٥) :
ـ اللآلىء في شرح الأمالي
طبع بمصر بمطبعة لجنة التأليف والترجمة
والنشر سنة ١٣٥٤ هـ بتصحيح الشيخ عبد

العزيز الميمني.
عبد الله بن ابيبه الديماني المتوفى ١٣٢٨ هـ
(بلاد شنقيط : ٥٨٠).
... منار الحائر في ترتيب النوادر

أمالي المرتضى
غرر الفوائد ودرر القلائد
لأبي القاسم علي بن الحسين بن موسى الكاظم المتوفى سنة ٤٣٦ هـ :
طبع بمصر سنة ١٣٢٥ هـ وطبع سنة ١٣٧٤ هـ بتحقيق محمد أبو الفضل إبراهيم وتكررت
طبعاته :

***

عبد الله بن ابراهيم بن عبد الله الحبري والباع
(الموحدة) المتوفى ٤٧٦ هـ (سيأتي) في ديوان
(المتنبي)
ـ شرح الغرر والدرر (هدية العارفين ١ : ٤٥٢)
فرح شواهده :
محسن آل الشيخ صاحب الجواهر المتوفى سنة

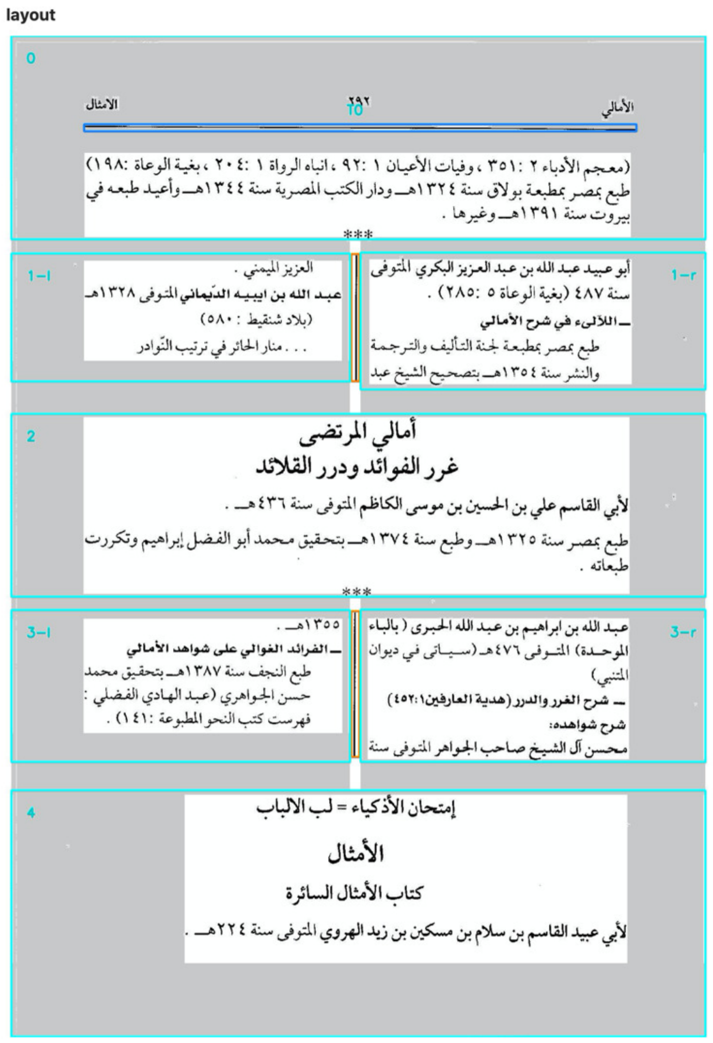
١٣٥٥ هـ :
ـ الفرائد الغوالي على شواهد الأمالي
طبع النجف سنة ١٣٨٧ هـ بتحقيق محمد
حسن الجواهري (عبد الهادي الفضلي
فهرست كتب النحو المطبوعة ١٤١) :

إمتحان الأذكياء = لب الألباب
الأمثال
كتاب الأمثال السائرة
لأبي عبيد القاسم بن سلام بن مسكين بن زيد الهروي المتوفى سنة ٢٢٤ هـ :

Figure 7. Interface for checking OCR results

With this information, we were ready to build out the intermediate steps between scans and OCR engine, by offering the OCR engine the text and just the text, cleaned as much as possible from anything else. We built two pieces of machinery for this, one for establishing the intended text blocks, and one for cleaning the text blocks line by line.

In order to bypass the layout and get to the text only, we still needed to know where, on each page, the text we needed was located. At time of writing, no satisfactory third-party library existed for this purpose. Fully automated approaches tend to be developed with newspapers in mind, while manual approaches take manuscripts as their primary use case. In order to retain maximum flexibility, we developed our own. The main idea driving our approach is to cast weighted histograms along the X and Y axis of the image, along with educated guesses as to what a typical layout for modern editions of medieval Arabic texts should look like. On one hand, this does not make our pipeline fully generalized, on the other hand, it does give very exact and useful results in a fully automated fashion. We first divide pages in horizontal blocks, possibly subdivided into columns, and assigning *header, middle,* or *footer* to each of them.



الأمالي ٢٩٢ الامثال

(معجم الأدباء ٢ :٣٥١ ، وفيات الأعيان ١ :٩٢ ، انباه الرواة ١ :٢٠٤ ، بغية الوعاة :١٩٨)
طبع بمصر بمطبعة بولاق سنة ١٣٢٤هـ ودار الكتب المصرية سنة ١٣٤٤هـ وأعيد طبعه في
بيروت سنة ١٣٩١هـ وغيرها .

***

أبو عبيد عبد الله بن عبد العزيز البكري المتوفى
سنة ٤٨٧ (بغية الوعاة ٥ :٢٨٥) .
ــ اللآلىء في شرح الأمالي
طبع بمصر بمطبعة لجنة التأليف والترجمة
والنشر سنة ١٣٥٤هـ بتصحيح الشيخ عبد

العزيز الميمني .
عبد الله بن ابيبه الدّيماني المتوفى ١٣٢٨هـ
(بلاد شنقيط :٥٨٠)
. . . منار الحائر في ترتيب النّوادر

أمالي المرتضى
غرر الفوائد ودرر القلائد

لأبي القاسم علي بن الحسين بن موسى الكاظم المتوفى سنة ٤٣٦هـ .
طبع بمصر سنة ١٣٢٥هـ وطبع سنة ١٣٧٤هـ بتحقيق محمد أبو الفضل إبراهيم وتكررت
طبعاته .

***

عبد الله بن ابراهيم بن عبد الله الحبرى (بالباء
الموحدة) المتوفى ٤٧٦هـ (سياتى في ديوان
المتنبي)
ــ شرح الغرر والدرر (هدية العارفين١:٤٥٢)
شرح شواهده:
محسن آل الشيخ صاحب الجواهر المتوفى سنة

١٣٥٥هـ .
ــ الفرائد الغوالي على شواهد الأمالي
طبع النجف سنة ١٣٨٧هـ بتحقيق محمد
حسن الجواهري (عبد الهادي الفضلي :
فهرست كتب النحو المطبوعة :١٤١) .

إمتحان الأذكياء = لب الالباب

الأمثال

كتاب الأمثال السائرة

لأبي عبيد القاسم بن سلام بن مسكين بن زيد الهروي المتوفى سنة ٢٢٤هـ .

Figure 8. Segmentation of the layout elements

We disregard headers, footers, and dividers and are left with blocks of interest, within which we identify lines by further subdivision into horizontal bands. We determine the regular line height by analyzing the peaks and the distances between them, based on the histogram of black pixels. If we have just one peak, there will be no distances, so instead we take the last line height that has been calculated. There is also a problem with short lines, which may have peaks that are lower than the detection threshold. We estimated the actual line lengths and calibrated the peaks by that information. There is a mild circularity there, because in order to determine the line lengths, we need to know already where the lines are.

We can, already, feed these lines to the OCR engine. However, within these lines there is still a lot of editorial intervention that stands in between the text as 'work' and 'expression.' For medieval Arabic texts, this is chiefly punctuation which is entirely added by the editor and has no historical reality. As much as it may be appreciated for human reading, it will be of no use, or rather, it will be an impediment to computational reading. Moreover, in our experiments we noticed an increase in accuracy if an OCR engine is given a

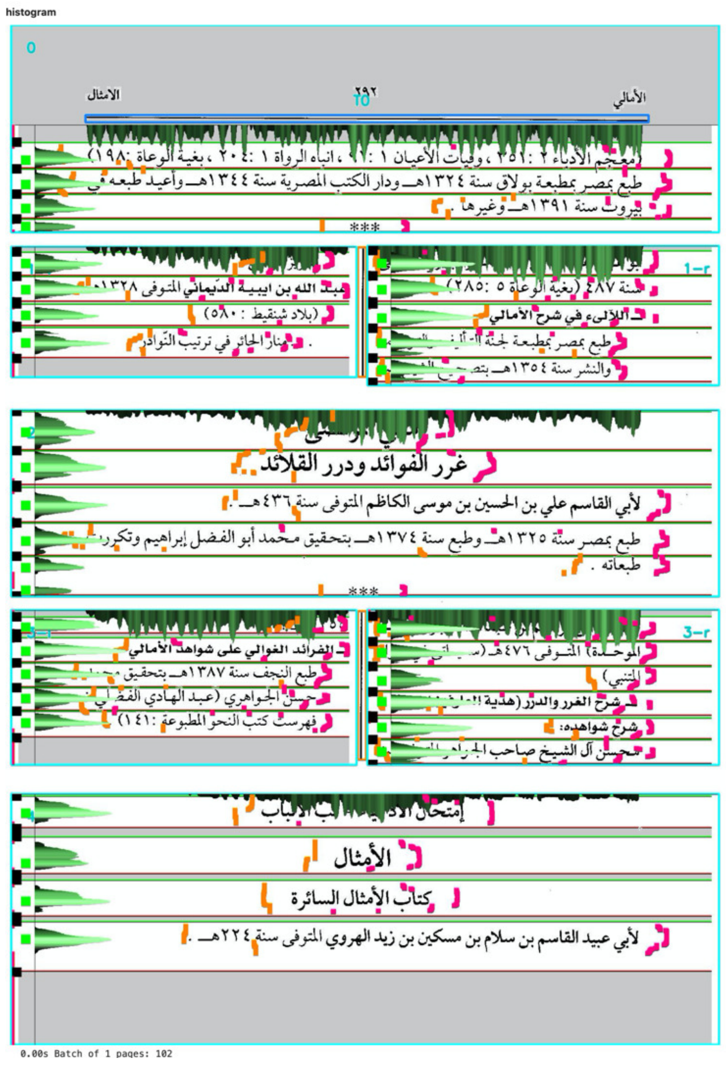
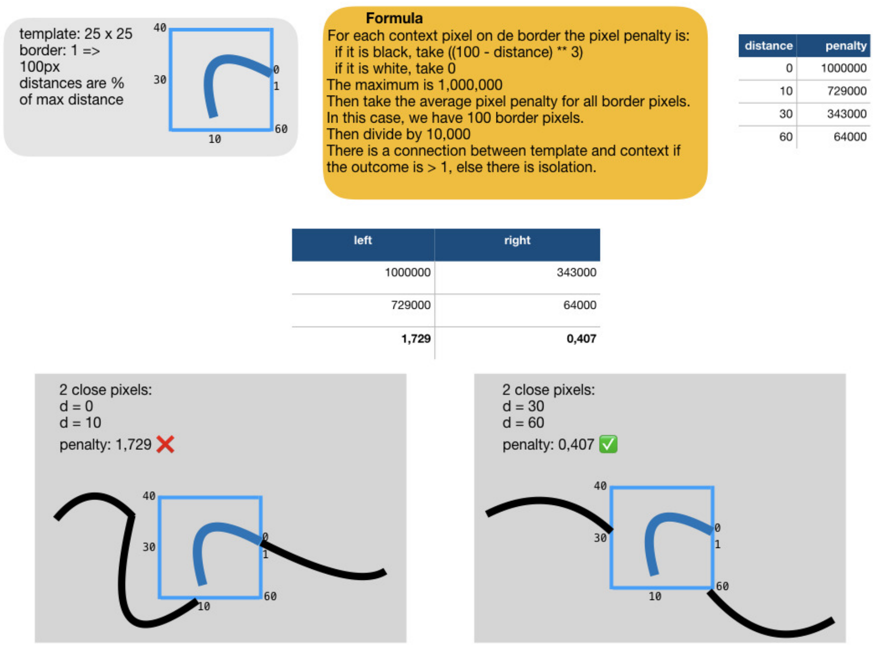


Figure 9. Line detection based on black pixels

snippet without punctuation. This could be because the standard training set contained no or little punctuation or simply because there is less information for the OCR engine to process, giving more space for letters and words to be identified. We found it therefore worth our while to invest in a process to remove such marks from the page before giving it to the OCR engine.

Marks can consist of anything, but typically include punctuation, vocalization, and footnote numbers. A set of exemplars needs to be prepared and saved as .png. This is done by looking for them on the processed scan images and cutting them out. Of each mark, more than one may be saved as exemplar, to attain a higher catch rate. These exemplars are used to pattern match using a computer vision library. These marks are typically small and could have a generic shape. Two things are installed to counter false positives.

One is that each line is yet again subdivided in horizontal bands. We defined the following: high is the upper band of a line, useful for catching footnotes numbers and vocalization; low as the lower band of a line, useful for punctuation and vocalization; mid as the central, narrow band of a line,



Formula
For each context pixel on de border the pixel penalty is:
  if it is black, take ((100 - distance) ** 3)
  if it is white, take 0
The maximum is 1,000,000
Then take the average pixel penalty for all border pixels.
In this case, we have 100 border pixels.
Then divide by 10,000
There is a connection between template and context if the outcome is > 1, else there is isolation.

distance	penalty
0	1000000
10	729000
30	343000
60	64000

left	right
1000000	343000
729000	64000
1,729	**0,407**

2 close pixels:
d = 0
d = 10
penalty: 1,729 ❌

2 close pixels:
d = 30
d = 60
penalty: 0,407 ✅

Figure 10. Disambiguating vocalization and punctuation from writing

where the majority of letter shapes are found; main is a wider version of mid, to find nearly any letter shape, also useful for some punctuation marks such as brackets; broad is broader than main, for very large punctuation and the like; inter attempts to capture the white space in between lines, which may be needed to find special glyphs such as a *dagger alif*.

This allows us to search for certain marks in only one or some bands, by design excluding false positives from other parts of the page. The marks can be placed in subfolders carrying the name of the bands and our software automatically picks them up. When it reads the exemplars, it will crop all white borders from it and surround the result with a fixed small white border, meaning that you do not have to be precise in trimming the exemplars.

The other counter measure is installed in evaluating the pattern matching based on three parameters. *Accuracy*, which is a generic measure to tell the computer vision library how closely an area should match the exemplar in order for it to be a hit (setting this to 1 would mean only a pixel-for-pixel exact match would be returned). *Connect-border-width* and *connect-ratio*, controls how many black pixels outside the area of a possible match is allowed to be adjacent to black pixels inside the area, i.e. it evaluates the mark is distinctly a separate thing and part of the text, as illustrated below. The border width sets the width of the inner and outer border around the area of a match which is inspected. The ratio sets the allowable adjacency between these inner and outer borders.

Once a positive match has been identified, all pixels in the area of the match are set to white, effectively removing the mark. For fine-tuning the parameters, we created several visual outputs, including one that displays matches of marks. We have set up a walkthrough on fine tuning these parameters in the Jupyter Notebook *comma.ipynb* in the example folder.

https://nbviewer.org/github/among/fusus/blob/master/notebooks/example/comma.ipynb

If accuracy needs to be adapted, it has proven to be better to include other exemplars. The result of this process of pre-OCR cleaning is illustrated below.

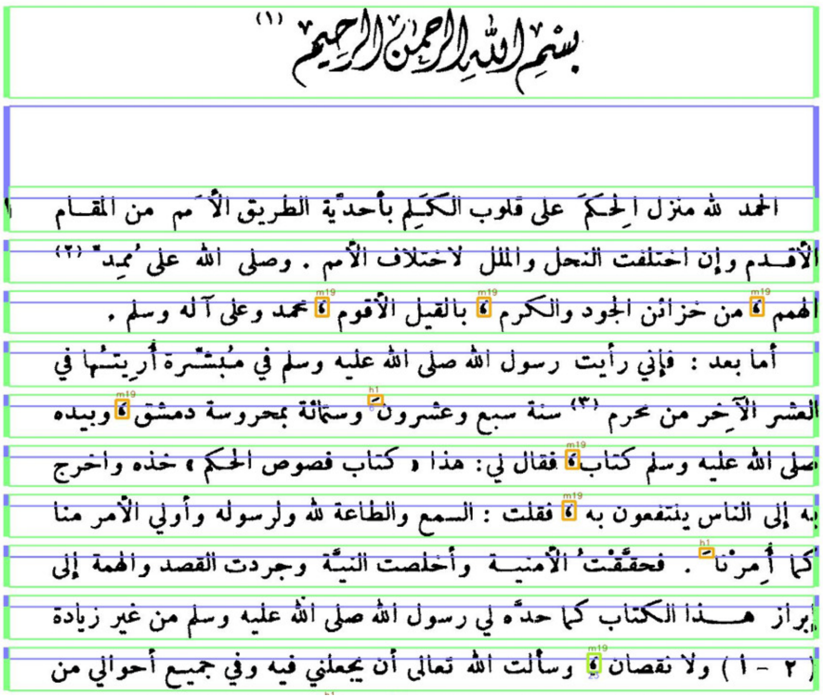

بسم الله الرحمن الرحيم (١)
الحمد لله منزل الحِكَم على قلوب الكَلِم بأحديّة الطريق الأَمَم من المقام
الأقدم وإن اختلفت النحل والملل لاختلاف الأمم . وصلى الله على مُحمد (٢)
الهمم ، من خزائن الجود والكرم ، بالقيل الأقوم ، محمد وعلى آله وسلم .
أما بعد : فإني رأيت رسول الله صلى الله عليه وسلم في مُبشّرة أُريتُها في
العشر الآخِر من محرم (٣) سنة سبع وعشرون وستمائة بمحروسة دمشق ، وبيده
صلى الله عليه وسلم كتاب ، فقال لي: هذا ، كتاب فصوص الحكم ، خذه واخرج
به إلى الناس ينتفعون به ، فقلت : السمع والطاعة لله ولرسوله وأولي الأمر منا
كما أُمِرنا . فحققتُ الأمنيــة وأخلصت النيّة وجردت القصد والهمة إلى
إبراز هــــذا الكتاب كما حدّه لي رسول الله صلى الله عليه وسلم من غير زيادة
(٢ - ١) ولا نقصان ، وسألت الله تعالى أن يجعلني فيه وفي جميع أحوالي من

Figure 11. Lines and punctuation detected, making the text ready for best possible OCR

The pipeline outputs a TSV file with each word on a new row and columns annotating the word to exactly describe its origin on the scanned image, as well as the OCR engine's accuracy. Out of the box a function is supplied to convert this format to Text-Fabric format, a specialized library for analysis of ancient texts plus annotations.

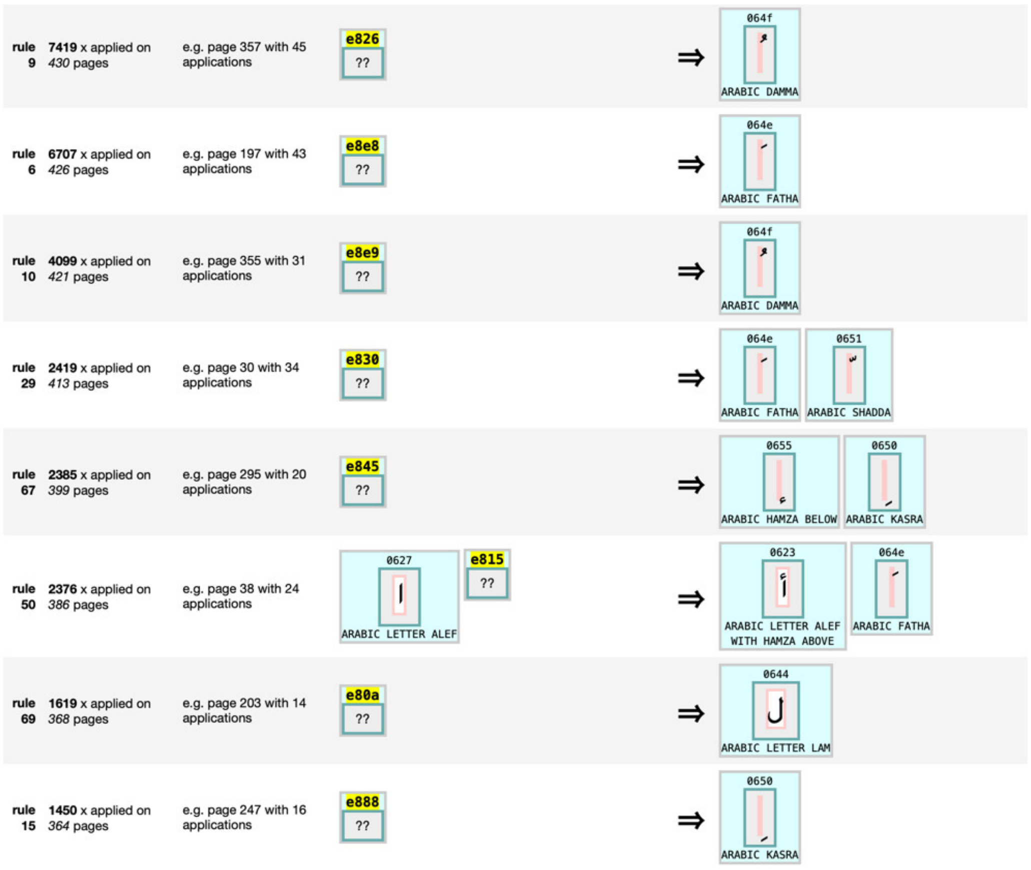
The entire *fusus* library is divided into submodules, which are described in our documentation. The steps we described before are captured in submodules *layout*, *lines* and *clean*. Especial note should be made of two submodules that supply the 'command and control' classes *book* and *page*. The entire pipeline can even be initiated with standard settings from the command line using *python 3 -m fusus.book* while in the directory of the book, with subfolders *in* and *marks* already prepared. *Book* also supplies information about the availability of a variety of things that one may wish to manually control through the pipeline, such as targeting only a certain range of pages. This can only be done if the *book.process* method is first run.

Such manual intervention is based done through a Jupyter Notebook. For convenience, a *htmlPages* method is available to produce a human readable file of the OCR output, if exploration within Jupyter Notebook is not preferred, which can be the case if a more general picture of the OCR output needs to be obtained, as scrolling through such a HTML-page can be done much more efficiently than looking through the output in a Jupyter Notebook. The *page* class is the one most often used in exploring the fine-tuning of the parameters, as it has the extensive *page.show* method to give visual feedback on all stages of the pipeline.

Toolkit for reverse-engineering PDFs

Within *fusus*, there is an entirely separate toolkit that allows for high performing Arabic PDF reverse engineering. We decided to keep this within *fusus* as it has the same function: a pipeline from original edition with a highly elaborate layout and many editorial additions towards a plain text structured data set. The reason for its development is quite simple: we found a digitally typeset PDF of a brand-new edition of *Fuṣūṣ al-ḥikam*, which based itself on the earliest manuscript witness penned by Ibn ʿArabī's student Ṣadr al-Dīn Qūnawī (Hirstenstein and Clark 6–7; Ibn ʿArabī, *Fuṣūṣ Al-Ḥikam* 10). We felt that using this edition was mandatory, and extracting the 'work' from this 'expression' seemed best done through this PDF, as this obviously gives absolute recognition of the characters. The extraction proved to be not straightforward at all. Two problems needed to be overcome: one is that the text was not set in a linear sequence and we needed to abstract away from it to the extent of reaching to letters individually, and reconstructing words, lines and the text block based on position on the page. Concomitant to this we noticed that horizontal whitespace (indicating a space between words) is hard to detect because of oversized bounding boxes of many characters. The other problem was that the encoding of the text used a great number of fonts, dual Unicode points, and private use characters.

We did initial exploration using FontReporter (see PDFLib) and PyMuPDF (see Artifex). Relaying every unique encoding with visual inspection gave us information which glyph was meant by what encoding, from which we built a comparison table. With the comparison table, we were able to transform every encoded character into its proper Unicode representation. An example of this work is the following figure:

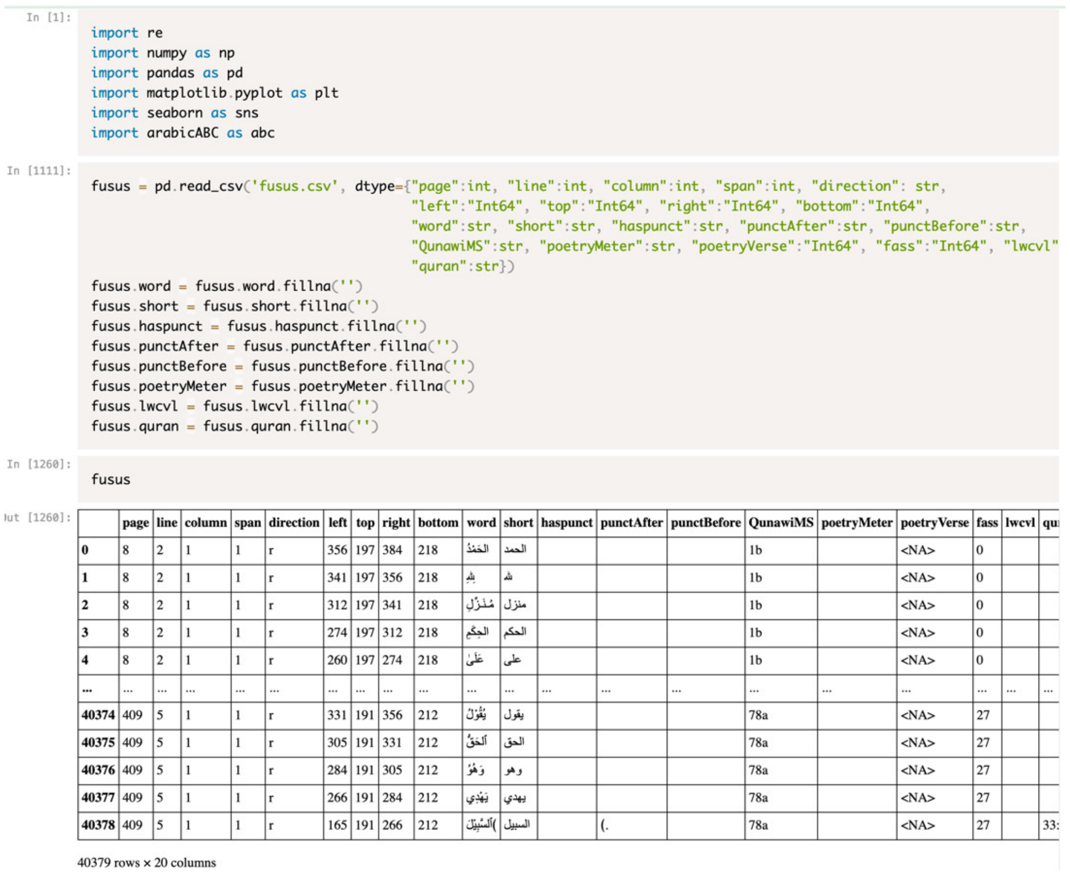


rule 9	7419 x applied on 430 pages	e.g. page 357 with 45 applications	e826 ??	⇒	064f ARABIC DAMMA
rule 6	6707 x applied on 426 pages	e.g. page 197 with 43 applications	e8e8 ??	⇒	064e ARABIC FATHA
rule 10	4099 x applied on 421 pages	e.g. page 355 with 31 applications	e8e9 ??	⇒	064f ARABIC DAMMA
rule 29	2419 x applied on 413 pages	e.g. page 30 with 34 applications	e830 ??	⇒	064e ARABIC FATHA; 0651 ARABIC SHADDA
rule 67	2385 x applied on 399 pages	e.g. page 295 with 20 applications	e845 ??	⇒	0655 ARABIC HAMZA BELOW; 0650 ARABIC KASRA
rule 50	2376 x applied on 386 pages	e.g. page 38 with 24 applications	0627 ARABIC LETTER ALEF; e815 ??	⇒	0623 ARABIC LETTER ALEF WITH HAMZA ABOVE; 064e ARABIC FATHA
rule 69	1619 x applied on 368 pages	e.g. page 203 with 14 applications	e80a ??	⇒	0644 ARABIC LETTER LAM
rule 15	1450 x applied on 364 pages	e.g. page 247 with 16 applications	e888 ??	⇒	0650 ARABIC KASRA

Figure 12. Translating exotic encodings to their uniform representation

A similar approach to the preprocessing for OCR led us to divide the page into horizontal bands, to find the lines. Ordering all characters within a band according to their placement coordinates, from right to left, gave us the characters on a line in the right order. These characters, however, did not include spaces. The spaces were only visually perceivable, but not actually encoded into the PDF. It thus required us to insert spaces ourselves, based on contextual information. The task of placing spaces between words proved to be exceedingly difficult. Some general rules could be applied related to distance, to which we added frequently occurring exceptions, such as detecting the end of a word when a final form is used (e.g. *tāʾ marbūṭa*). These exceptions are perhaps not that 'exceptional', as already final form detection provided about 60% of all word splits. The output was again provided in a TSV file.

We do not actually store spaces - we simply place every word on a new row and consider the jump from one row to the next a space. This meant that many rows contained two or more words written without spaces. Since our modeling of commentaries uses this sequence of rows of the *Fuṣūṣ* as an index, it is necessary to get a correct separation of words. Additionally, the extraction from PDF included full vocalization and punctuation, including signs to indicate Quran citations and poetry. While these are worthy editorial contributions, they are part of the 'expression' and not of the 'work', and we wanted to separate them out. As this cleaning is text- and PDF-specific, we do not offer specifics for it within the *fusus* library. We found Pandas was a useful tool for this, as it gave flexible tabular visuals and a high-order query capacity. An example of our setup is given below, which can be further explored in our Jupyter Notebook *SplittingWords.ipynb* which is not a fully worked out tutorial but should provide enough examples and edge cases to learn from and adapt to your specific needs.



```
In [1]:
import re
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import arabicABC as abc
```

```
In [1111]:
fusus = pd.read_csv('fusus.csv', dtype={"page":int, "line":int, "column":int, "span":int, "direction": str,
                                        "left":"Int64", "top":"Int64", "right":"Int64", "bottom":"Int64",
                                        "word":str, "short":str, "haspunct":str, "punctAfter":str, "punctBefore":str,
                                        "QunawiMS":str, "poetryMeter":str, "poetryVerse":"Int64", "fass":"Int64", "lwcvl"
                                        "quran":str})

fusus.word = fusus.word.fillna('')
fusus.short = fusus.short.fillna('')
fusus.haspunct = fusus.haspunct.fillna('')
fusus.punctAfter = fusus.punctAfter.fillna('')
fusus.punctBefore = fusus.punctBefore.fillna('')
fusus.poetryMeter = fusus.poetryMeter.fillna('')
fusus.lwcvl = fusus.lwcvl.fillna('')
fusus.quran = fusus.quran.fillna('')
```

```
In [1260]:
fusus
```

Out [1260]:

	page	line	column	span	direction	left	top	right	bottom	word	short	haspunct	punctAfter	punctBefore	QunawiMS	poetryMeter	poetryVerse	fass	lwcvl	qu
0	8	2	1	1	r	356	197	384	218	الحَمْدُ	الحمد				1b		<NA>	0		
1	8	2	1	1	r	341	197	356	218	لِلهِ	لله				1b		<NA>	0		
2	8	2	1	1	r	312	197	341	218	مُنَزِّلِ	منزل				1b		<NA>	0		
3	8	2	1	1	r	274	197	312	218	الحِكَمِ	الحكم				1b		<NA>	0		
4	8	2	1	1	r	260	197	274	218	عَلَىٰ	على				1b		<NA>	0		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
40374	409	5	1	1	r	331	191	356	212	يَقُوْلُ	يقول				78a		<NA>	27		
40375	409	5	1	1	r	305	191	331	212	ٱلْحَقَّ	الحق				78a		<NA>	27		
40376	409	5	1	1	r	284	191	305	212	وَهُوَ	وهو				78a		<NA>	27		
40377	409	5	1	1	r	266	191	284	212	يَهْدِي	يهدي				78a		<NA>	27		
40378	409	5	1	1	r	165	191	266	212	ٱلسَّبِيْلَ)	السبيل		(.		78a		<NA>	27		33:

40379 rows × 20 columns

Figure 13. Data cleaning and enriching in Pandas

Next to a handful of standard libraries, we wrote a lightweight module called *arabicABC*, based on an early version of what is now the mature PyArabic (see Zerrouki). It's primary purpose is to find or insert the correct Unicode value for Arabic characters using a more reader-friendly format (for example *abc.JEEM* instead of `62c`). It also groups letters (for example, *abc.ALEFAT* gives all characters whose base is an *alif*) and gives a quick way to switch between sura numbers and names for the Quran. We added functions for normalization and rasmization (for example, *zayn* is changed to *ra*). Normalization is especially useful when obtaining text from a digital source, such as a digitally typeset PDF, rasmization may be a useful instrument when obtaining text from an image, especially when it needs to be compared or aligned with other text.

Note that in Pandas, for a file this size, dtypes should be defined for all columns in order for the file to be loaded fast. Other optimizations are hardly needed. We have found that even with hundreds of thousands of words, a text with stand-off annotation like this can be loaded in full in memory. Recourse to a database file type is not necessary.

In succession, the columns define the following data:

Index The left column without heading gives a successive number to every word in the text. During cleaning this regularly needs to be updated but remains fixed afterwards.

Page Indicates on which page number the word appears. We ensure this corresponds to the page number as shown/typeset on the page, and not the number of the page as it appears in the PDF/image set.

Line Gives line number of the word on the page.

Column/span If the page has a complicated layout, this information helps identifying the correct line.

Direction If available, text-direction information is stored here. *r* means right to left.

left/top X,Y position on the page of the top left corner of the word.

right/bottom X,Y position on the page of the bottom right corner of the word. Note that in cleaning we did not update this information.

Word Provides the raw input from either digital PDF or OCR.

Short Provides the cleaned text, in the state in which we like to have it as much as possible. This means to us: an unvocalized, unpunctuated, un-isolated, non-private-use, form of the Arabic characters.

haspunct A temporary column which we store information if we detect punctuation marks in the *word* column. This is useful since the bi-directionality of Arabic (RTL) and most punctuation marks (LTR) makes a very messy representation in which visual inspection cannot be trusted.

punctAfter For those rows which have their *haspunct* column filled, we move most of it to this column.

punctBefore Some punctuation should come before a word, e.g. bracket-open.

QunawiMS The edition we used had in-line information about the place of the text in the oldest extant manuscript, in between square brackets. We deleted the rows which contained this information but stored it in this column.

poetryMeter/poetryVerse Poetry was also indicated with in-line information, including verse numbering and meter. We deleted those rows but saved the information in these column.

fass Provides chapter number (27 in total).

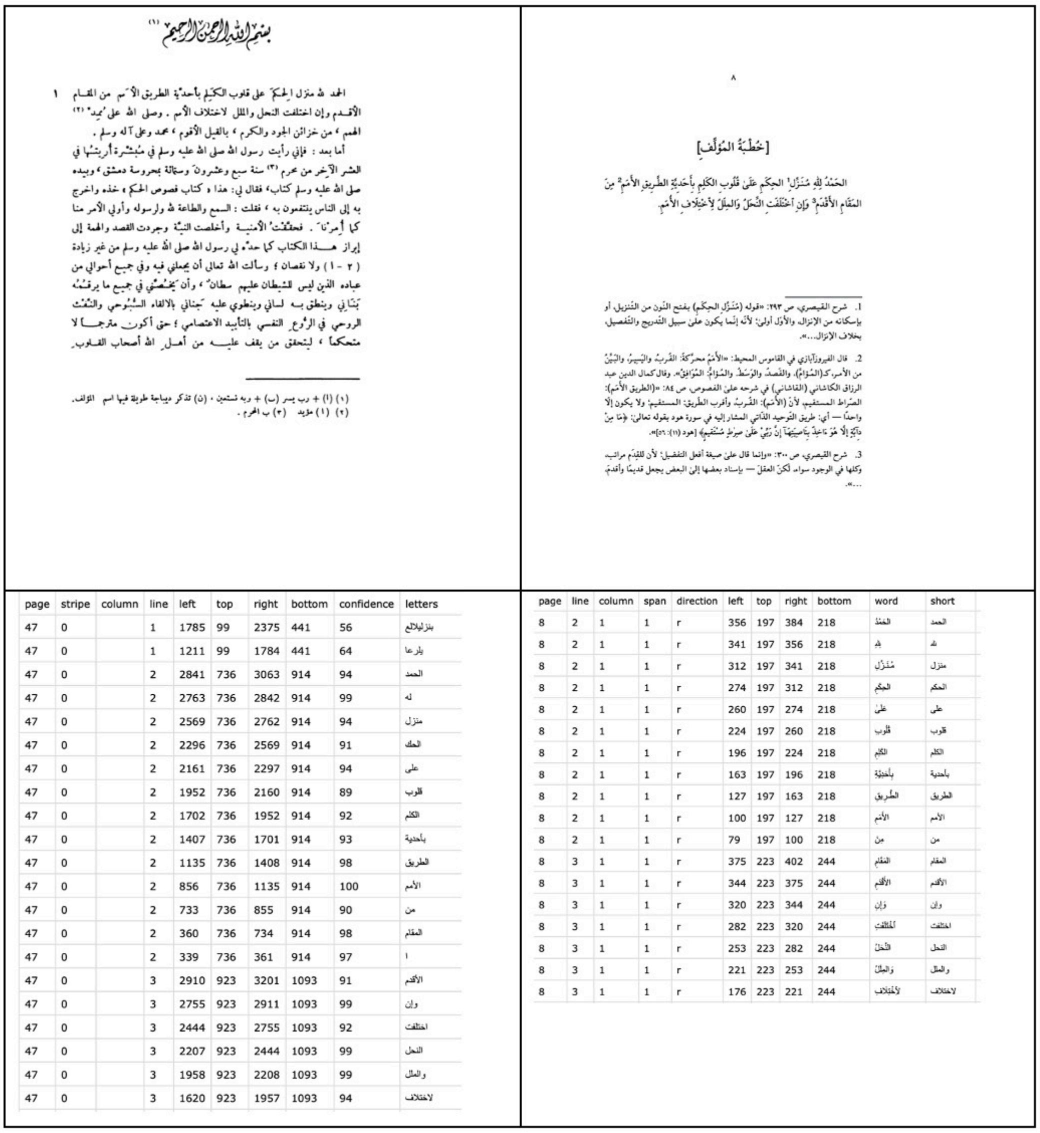
lwcvl Contains our own comments for future reference.

quran Gives sura/aya numbers for those parts of the text that seemed to be cited from the Quran.

The long tail of cleaning the reverse-engineered PDF took a very long time. To finalize it and establish a version of the *Fuṣūṣ* we could confidently call the 'work', we decided to compare the data we got with another edition which we could parse through our original pipeline.

A converter to unify text input and perform basic analysis

Two 'expressions' of the same 'work': the editions of *Fuṣūṣ al-ḥikam* by Afifi (1946, see Ibn ʿArabī and Afifi) and Lakhnawi (2013) were respectively processed through the OCR pipeline and through the PDF pipeline, resulting in the following two files (excerpt):



بسم الله الرحمن الرحيم (١)

١ الحمد لله منزل الحكَم على قلوب الكَلِم بأحديّة الطريق الأَمَم من المقام الأقدم وإن اختلفت النحل والملل لاختلاف الأمم . وصلى الله على مُمِد (٢) الهمم ، من خزائن الجود والكرم ، بالقيل الأقوم ، محمد وعلى آله وسلم .
أما بعد : فإني رأيت رسول الله صلى الله عليه وسلم في مُبشّرة أُريتُها في العشر الآخِر من محرم (٣) سنة سبع وعشرون وستمائة بمحروسة دمشق ، وبيده صلى الله عليه وسلم كتاب ، فقال لي : هذا ، كتاب فصوص الحكم ، خذه واخرج به إلى الناس ينتفعون به ، فقلت : السمع والطاعة لله ولرسوله وأولي الأمر منا كما أُمِرنا . فحقّقتُ الأمنية وأخلصت النيّة وجردت القصد والهمة إلى إبراز هـذا الكتاب كما حدّه لي رسول الله صلى الله عليه وسلم من غير زيادة (٢ - ١) ولا نقصان ؛ وسألت الله تعالى أن يجعلني فيه وفي جميع أحوالي من عباده الذين ليس للشيطان عليهم سلطان ، وأن يَخُصَّني في جميع ما يرقُمُه بَنَاني وينطق بـه لساني وينطوي عليه جَناني بالالقاء السُّبُّوحي والنَّفْث الروحي في الرُّوعِ النفسي بالتأييد الاعتصامي ؛ حتى أكون مترجماً لا متحكماً ، ليتحقق من يقف عليـه من أهل الله أصحاب القلوب

(١) (ا) + رب يسر (ب) + وبه نستعين ، (ن) تذكر ديباجة طويلة فيها اسم المؤلف.
(٢) (١) مؤيد (٣) ب المحرم .

٨

[خُطْبَةُ المُؤلِّفِ]

الحَمْدُ لِلهِ مُنَزِّلِ¹ الحِكَمِ عَلَىٰ قُلُوبِ الكَلِمِ بِأَحَدِيَّةِ الطَّرِيقِ الأَمَمِ² مِنَ المَقَامِ الأَقْدَمِ³ وَإِنِ اخْتَلَفَتِ النِّحَلُ وَالمِلَلُ لِاخْتِلَافِ الأُمَمِ.

1. شرح القيصري، ص ٢٩٣: «قوله (مُنَزِّلِ الحِكَم) بفتح النون من التنزيل، أو بإسكانه من الإنزال، والأوّل أولى؛ لأنّه إنّما يكون علىٰ سبيل التدريج والتفصيل، بخلاف الإنزال...».

2. قال الفيروزآبازي في القاموس المحيط: «الأَمَمُ محرّكةً: القُربُ، واليَسِيرُ، والبَيِّنُ من الأمر، كـ(المُؤامّ)، والقَصدُ والوَسَطُ. والمُؤامُّ: المُوَافِقُ». وقال كمال الدين عبد الرزاق الكاشاني (القاشاني) في شرحه علىٰ الفصوص، ص ٨٤: «(الطريق الأَمَم): الصِّراط المستقيم، لأنّ (الأَمَم): القُربُ، وأقرب الطّريق: المستقيم؛ ولا يكون إلّا واحدًا — أي: طريق التّوحيد الذّاتي المشار إليه في سورة هود بقوله تعالى: ﴿مَا مِنْ دَابَّةٍ إِلَّا هُوَ آخِذٌ بِنَاصِيَتِهَا إِنَّ رَبِّي عَلَىٰ صِرَاطٍ مُسْتَقِيمٍ﴾ [هود (١١): ٥٦]».

3. شرح القيصري، ص ٣٠٠: «وإنما قال علىٰ صيغة أفعل التفضيل؛ لأن للقِدَم مراتب، وكلها في الوجود سواء. لكنّ العقلَ — بإسناد بعضها إلىٰ البعض يجعل قديمًا وأقدمَ...».

page	stripe	column	line	left	top	right	bottom	confidence	letters
47	0		1	1785	99	2375	441	56	بتزليلالع
47	0		1	1211	99	1784	441	64	يلرعا
47	0		2	2841	736	3063	914	94	الحمد
47	0		2	2763	736	2842	914	99	له
47	0		2	2569	736	2762	914	94	منزل
47	0		2	2296	736	2569	914	91	الحك
47	0		2	2161	736	2297	914	94	على
47	0		2	1952	736	2160	914	89	قلوب
47	0		2	1702	736	1952	914	92	الكلم
47	0		2	1407	736	1701	914	93	بأحدية
47	0		2	1135	736	1408	914	98	الطريق
47	0		2	856	736	1135	914	100	الأمم
47	0		2	733	736	855	914	90	من
47	0		2	360	736	734	914	98	المقام
47	0		2	339	736	361	914	97	ا
47	0		3	2910	923	3201	1093	91	الأقدم
47	0		3	2755	923	2911	1093	99	وإن
47	0		3	2444	923	2755	1093	92	اختلفت
47	0		3	2207	923	2444	1093	99	النحل
47	0		3	1958	923	2208	1093	99	والملل
47	0		3	1620	923	1957	1093	94	لاختلاف

page	line	column	span	direction	left	top	right	bottom	word	short
8	2	1	1	r	356	197	384	218	الحَمْدُ	الحمد
8	2	1	1	r	341	197	356	218	لِلهِ	لله
8	2	1	1	r	312	197	341	218	مُنَزِّلِ	منزل
8	2	1	1	r	274	197	312	218	الحِكَمِ	الحكم
8	2	1	1	r	260	197	274	218	عَلَىٰ	على
8	2	1	1	r	224	197	260	218	قُلُوبِ	قلوب
8	2	1	1	r	196	197	224	218	الكَلِمِ	الكلم
8	2	1	1	r	163	197	196	218	بِأَحَدِيَّةِ	باحدية
8	2	1	1	r	127	197	163	218	الطَّرِيقِ	الطريق
8	2	1	1	r	100	197	127	218	الأَمَمِ	الأمم
8	2	1	1	r	79	197	100	218	مِنَ	من
8	3	1	1	r	375	223	402	244	المَقَامِ	المقام
8	3	1	1	r	344	223	375	244	الأَقْدَمِ	الأقدم
8	3	1	1	r	320	223	344	244	وَإِنِ	وان
8	3	1	1	r	282	223	320	244	اخْتَلَفَتِ	اختلفت
8	3	1	1	r	253	223	282	244	النِّحَلُ	النحل
8	3	1	1	r	221	223	253	244	وَالمِلَلُ	والملل
8	3	1	1	r	176	223	221	244	لِاخْتِلَافِ	لاختلاف

Figure 14. Above: sample pages from two editions. Below: End result of digital encoding

We set up a final part of our *fusus* library to process the two results to achieve a pragmatic mapping of the two editions. The result of Lakhnawi's edition was cleaner and arguably closer to the original work given its reliance on the oldest extant manuscript, and therefore we used it as benchmark. Thus, if the result of Afifi's edition was imperfect, this did not dirty our results. Imperfections consist chiefly of three things: wrong character recognition, accidental merging of two words, and accidental separation of one word.

In case of wrong character recognition, we would not want this to result in the flagging of a discrepancy between the two versions. Instead, we want to confirm a mapping of the two rows and subsequently use the character values of the result of Lakhnawi's edition (the *short* column).

In case of accidental merging, we are content with a mapping of multiple rows of the result of Lakhnawi's edition to the one row in the result of Afifi's edition. By using the same *short* column again to actually store the text, we ensure that we do not accidentally flag this as an addition in Lakhnawi's edition.

For accidental separation we do the same, ensuring we do not have a false positive for an addition in Afifi's edition.

This means that in terms of actual text, we rely on the output of Lakhnawi's edition for all words that are either present in both editions or are only in Lakhnawi's edition. In cases were Afifi's edition has a unique contribution, we then rely on the output of Afifi's edition.

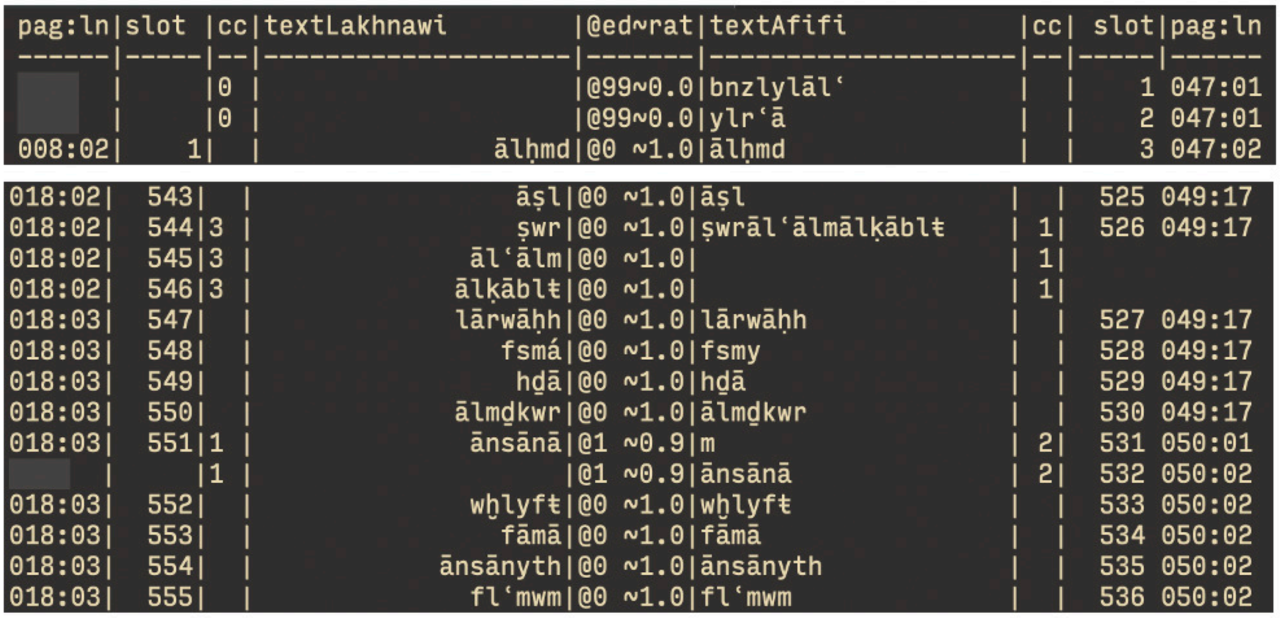
We implemented the combining of the two editions in two steps: alignment and merging.

For the alignment step we initially tried the Python port of the generic collation tool Collatex (see Interedition). It would need 5 hours to run it on these two editions so we abandoned this. Instead, we exploited the fact that the two editions are very close variants of each other. The logic we were after is that we walk through both sources and make comparisons of single words and short sequences of words and, if that fails, increasingly big jumps. If that fails too, we rely on one or two dozen explicit exceptions that we found by trial and error. At the heart of the comparisons lies the concept of edit distance, in the Levenshtein sense, for which we use a Python library (see Python-Levenshtein). The resulting alignment function takes about 1 second to complete.

We have checked the results of the alignment by analyzing the output. We checked for sanity (the alignment preserves all material of both editions in the right order), we analyzed the closeness of the paired words, we detected suspect pairings where the algorithm may have missed the most obvious matches. The results of the analysis can be seen in the notebook that ran the alignment, *compareAfLk.ipnyb*. We also produced a full table with rich information in ascii form of the complete alignment, available at https://github.com/among/fusus/blob/master/notebooks/zipLK-AF-complete.txt

We adapted and tweaked the parameters and added special cases until there are no longer any suspect pairings in the end result.

We found that 87% of the words were immediately correct. We can take this to mean that our OCR pipeline has a success-rate of about this percentage. The remaining percentage was narrowed down to 306 discrepancies, thanks to our algorithm. These discrepancies mean that one or more words were only found in one edition, not the other. Manual inspection revealed that the discrepancies break down in the following:



```
pag:ln|slot |cc|textLakhnawi        |@ed~rat|textAfifi            |cc| slot|pag:ln
------|-----|--|--------------------|-------|---------------------|--|-----|------
      |     |0 |                    |@99~0.0|bnzlylālʿ            |  |    1 047:01
      |     |0 |                    |@99~0.0|ylrʿā                |  |    2 047:01
008:02|    1|  |               ālḥmd|@0 ~1.0|ālḥmd                |  |    3 047:02

018:02|  543|  |                 āṣl|@0 ~1.0|āṣl                 |  |  525 049:17
018:02|  544|3 |                 ṣwr|@0 ~1.0|ṣwrālʿālmālḳāblŧ     | 1|  526 049:17
018:02|  545|3 |               ālʿālm|@0 ~1.0|                    | 1|
018:02|  546|3 |              ālḳāblŧ|@0 ~1.0|                    | 1|
018:03|  547|  |              lārwāḥh|@0 ~1.0|lārwāḥh             |  |  527 049:17
018:03|  548|  |                 fsmá|@0 ~1.0|fsmy                |  |  528 049:17
018:03|  549|  |                  hḏā|@0 ~1.0|hḏā                 |  |  529 049:17
018:03|  550|  |              ālmḏkwr|@0 ~1.0|ālmḏkwr             |  |  530 049:17
018:03|  551|1 |              ānsānā|@1 ~0.9|m                    | 2|  531 050:01
      |     |1 |                    |@1 ~0.9|ānsānā               | 2|  532 050:02
018:03|  552|  |              wẖlyfŧ|@0 ~1.0|wẖlyfŧ               |  |  533 050:02
018:03|  553|  |                 fāmā|@0 ~1.0|fāmā                 |  |  534 050:02
018:03|  554|  |            ānsānyth|@0 ~1.0|ānsānyth             |  |  535 050:02
018:03|  555|  |               flʿmwm|@0 ~1.0|flʿmwm               |  |  536 050:02
```

Figure 15. Algorithm at work for aligning the two editions

Number	Case
164	actual additions - one or more words are attested in only one edition
100	honorifics - one or the other edition added a standard praise after a name
25	substitutes - the two texts switched the order of one or more words
17	false positives - algorithm could be amended with another special case

Of the actual editions, a vast majority are truly insignificant, functioning as a synonym. The rest can be classified as very minor, without changing the meaning of the text.

With the alignment table in hand, we then merged the cleaned and enriched csv files of the Lakhnawi and Affifi editions into a new TSV-file. When working with CSV/TSV-files of different sources in different stages of completion, it becomes more and more difficult to keep track of which columns each file has, what they stand for, and how they code the concepts of the text, not only the words, but also the lines, pages, sentences and pieces. When information is added, new columns are introduced, and existing code that deals with the table must be updated. This will quickly get very cumbersome. That is one of the reasons why we convert the corpus into a Text-Fabric dataset. Text-Fabric is a file format and a library and an interface for working with text and annotations in a uniform and performant way (see Roorda et al.). The tool makes it easy to preprocess the corpus in any format that subsequent analysis might require. At the same time, it offers a browsing interface to see the corpus and its annotations, even annotations contributed by others. It makes it easy for others to download the corpus, compute with it, generate new annotations, and share them so that everybody can use them.

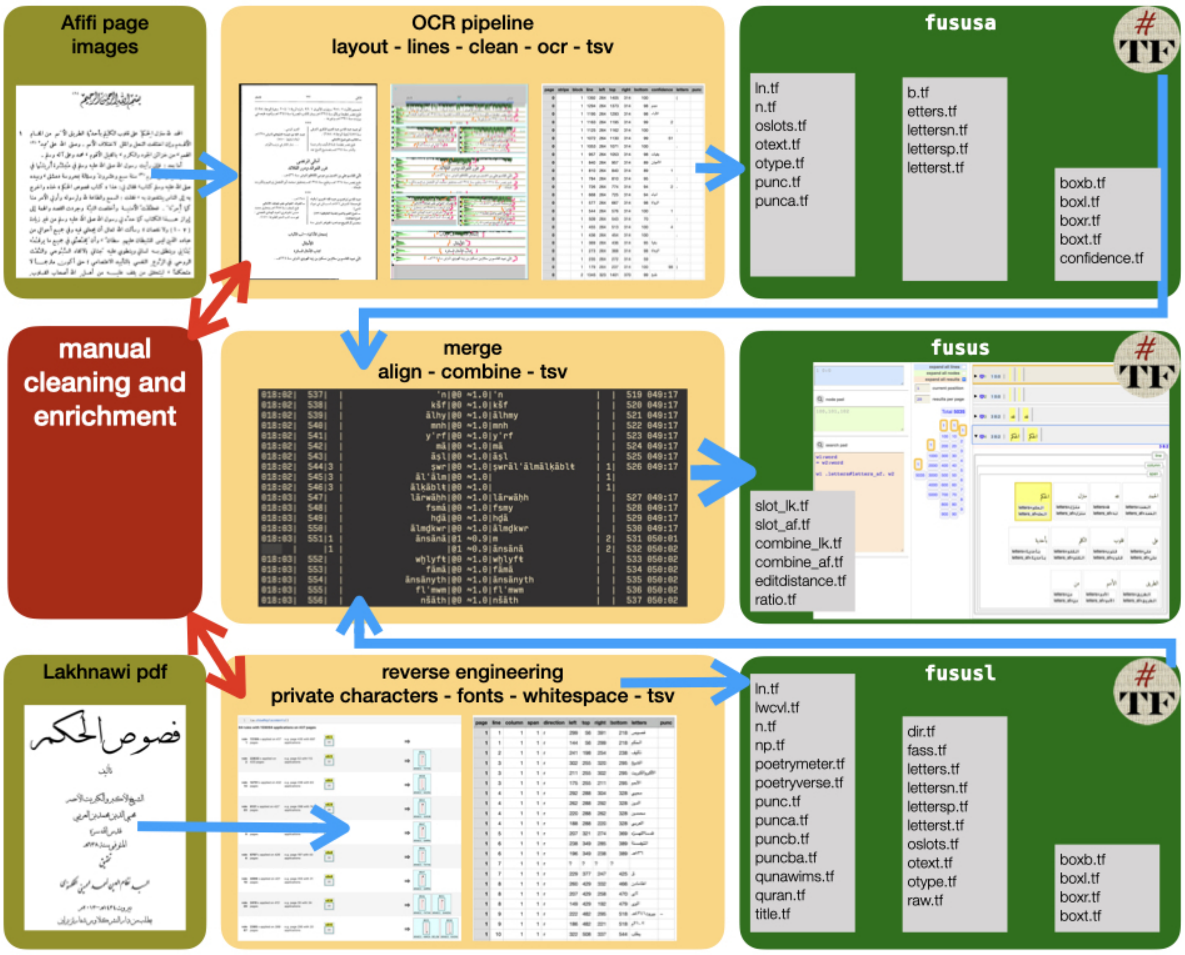
Technically, Text-Fabric sees a corpus as a graph of textual elements, linked by edges that denote containment or any other relationship. Nodes and edges can be annotated with features, which are mappings from nodes or node pairs to values. In a Text-Fabric dataset, each feature corresponds to exactly one file. A feature file looks like a CSV-file, but with only a single column, and with optimizations to deal with sparsity in the column. When new features are added, existing code that was not aware of them, remains valid. Text-Fabric is additionally a useful tool when we have processed commentaries, to access all texts computationally and visually in a unified way.

Conclusion

We set out to make methodological headway in studying postclassical commentaries from the Islamic literary heritage. Two characteristics of this genre stood out to us: the large size of texts and their complicated intertextuality. These characteristics called out to us to use a meso-level digital approach. To turn a base text and several (higher-level) commentaries into digital text, resulting in a data set containing hundreds of thousands to millions of words, will have the benefit of being able to computationally analyzing them which solves the issue of not being able to read through such a large quantity. Mimicking the intertextual links within a commentary tradition through a data structure will solve the difficulty of barely being able to keep track of the correct context in which a commentator is saying something. It allows ease of representation of that context, as a result to a computation or by computing which disparate passages from all kinds of texts need to be displayed together. One such computational result is a 'heartbeat' of a commentary tradition: an analysis where multiple commentaries on the same base text break into the text. Such an analysis will shed light on which parts of the base text were most discussed and how commentaries can be grouped together.

In order to make progress along these lines, we started out to consider the simplified case of Ibn ʿArabī's commentary tradition which has no higher-order commentaries, that is, all commentaries go directly back to Ibn ʿArabī's base text called *Fuṣūṣ al-ḥikam*. Within this particular example, we simplified our work even further by first dealing with the base text only. For this, we used the *IFLA Library Reference Model* to differentiate between the specific 'expression' of the *Fuṣūṣ al-ḥikam*, that is, the original Arabic text as Ibn ʿArabī envisioned it, and the 'manifestations' of this work, that is, the different editions that exist (in our case: editions by Afifi and Lakhnawi). Simplifying the problem twice, and first ironing out all problems within this much narrower context seemed, in hindsight to have been a good choice. Not only did it allow us to develop a fairly mature toolset, it also gave us significant insight in the inner workings of the intertextual relations we wish to tease out.

The following diagram illustrates our entire workflow.



```
018:02|  537|  |             'n|@0 ~1.0|'n                    |  |  519 049:17
018:02|  538|  |            kšf|@0 ~1.0|kšf                   |  |  520 049:17
018:02|  539|  |           ālhy|@0 ~1.0|ālhmy                 |  |  521 049:17
018:02|  540|  |            mnh|@0 ~1.0|mnh                   |  |  522 049:17
018:02|  541|  |           y'rf|@0 ~1.0|y'rf                  |  |  523 049:17
018:02|  542|  |             mā|@0 ~1.0|mā                    |  |  524 049:17
018:02|  543|  |            āṣl|@0 ~1.0|āṣl                   |  |  525 049:17
018:02|  544|3 |            ṣwr|@0 ~1.0|ṣwrāl'ālmālķāblŧ      | 1|  526 049:17
018:02|  545|3 |         āl'ālm|@0 ~1.0|                      | 1|
018:02|  546|3 |        ālķāblŧ|@0 ~1.0|                      | 1|
018:03|  547|  |        lārwāḥh|@0 ~1.0|lārwāḥh               |  |  527 049:17
018:03|  548|  |           fsmā|@0 ~1.0|fsmy                  |  |  528 049:17
018:03|  549|  |            hḏā|@0 ~1.0|hḏā                   |  |  529 049:17
018:03|  550|  |        ālmḏkwr|@0 ~1.0|ālmḏkwr               |  |  530 049:17
018:03|  551|1 |         ānsānā|@1 ~0.9|m                     | 2|  531 050:01
      |     |1 |               |@1 ~0.9|ānsānā                | 2|  532 050:02
018:03|  552|  |         wḫlyfŧ|@0 ~1.0|wḫlyfŧ                |  |  533 050:02
018:03|  553|  |           fāmā|@0 ~1.0|fāmā                  |  |  534 050:02
018:03|  554|  |       ānsānyth|@0 ~1.0|ānsānyth              |  |  535 050:02
018:03|  555|  |          fl'mwm|@0 ~1.0|fl'mwm                |  |  536 050:02
018:03|  556|  |          nšāth|@0 ~1.0|nšāth                 |  |  537 050:02
```

Figure 16. Diagram of the Fusus workflow

It should be pointed out that the OCR pipeline is not a silver bullet, and each new commentary will require the personal attention of an operator to identify the marks that need to be cleaned, to spot unusual page layouts, and to monitor the quality of the output data. The workflow is a machine, to be operated by a skilled person, in order to get meaningful results out of it. But, looking back at the page images of the Afifi edition and the warped pdf of the Lakhnawi edition, we do think we have come a long way. Instead of scans of printed pages, we now look at CSV/TSV-files, which we can study in Python with Pandas, and we have the 'work' of the *Fuṣūṣ* in Text-Fabric, both 'expressions' separately and a merged version with the differences still intact. In short, we have data in hand that we can compute with, in an organized and efficient manner. This is neither an edition in the classic sense: we have abstracted away from existing critical editions as much as possible and our own results are not instantly human-readable. But neither is it a corpus, for which little care is given exactly what the shape and size is of the text as the emphasis is on scaling up. We are striking at a level in between the two, creating a new path of inquiry into texts. As this path is truly new, it took us considerable time finding and building the right tools to move us forward. Our ambitions to include commentaries and expand this dataset are therefore not over, they are just beginning.

Data repository: https://doi.org/10.7910/DVN/3KCFRT

Submitted: November 13, 2023 EDT, Accepted: December 12, 2023 EDT

WORKS CITED

Artifex. *PyMuPDF*. https://pymupdf.readthedocs.io/en/latest/index.html.

Hirstenstein, S., and J. Clark. "Establishing Ibn ʿArabī's Heritage: First Findings from the MIAS Archiving Project." *Journal of the Muhyiddin Ibn ʿArabi Society*, vol. 52, 2012, pp. 1–32.

Ibn ʿArabī. *Fuṣūṣ Al-Ḥikam*. Edited by S.A. al-Lakhnawi, Klaus Schwarz, 2013.

---. *Fuṣūṣ Al-Ḥikam*. 1946. Edited by A.A. ʿAfifi, Dār al-kitāb al-ʿarabī, 2002.

Interedition. *Collatext*. http://interedition.github.io/collatex/pythonport.html.

Kraken. https://kraken.re/main/index.html.

Lit, C. van, and D. Roorda. *Fusus, Version 0.6*. 2 Nov. 2021, https://github.com/among/fusus/.

Nigst, L., et al. *OpenITI: A Machine-Readable Corpus of Islamicate Texts (2021.2.5) [Data Set]*. Zenodo, 5 Feb. 2021, https://doi.org/10.5281/ZENODO.5550338.

OpenITI. *arabic_generalized.mlmodel*. https://github.com/OpenITI/OCR_GS_Data/tree/master/ara/abhathz.

PDFLib. "FontReporter." *PDF Association*, https://pdfa.org/product/pdflib-fontreporter/.

Python-Levenshtein. https://github.com/ztane/python-Levenshtein.

Riva, P., et al. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. IFLA, 2017, pp. 19–23.

Roorda, D., et al. *Annotation/Text-Fabric: For Tool Registry*. 2022, https://doi.org/10.5281/ZENODO.7067373.

van Lit, L. W. C. *Among Digitized Manuscripts: Philology, Codicology, Paleography in a Digital World*. Brill, 2020.

---. "Commentary and Commentary Tradition: The Basic Terms for Understanding Islamic Intellectual History." *MIDÉO*, vol. 32, 2017, pp. 3–26.

---. "Ibn ʿArabī's School of Thought: Philosophical Commentaries, Not a Sufi Order." *Journal of Islamic Philosophy*, vol. 14, 2023, pp. 162–87.

Young, W., and F. Elwert, editors. *Samarqandī, Kitāb ʿAyn al-Naẓar Fī ʿIlm al-Jadal*. https://pages.ceres.rub.de/ayn-al-nazar/.

Zerrouki, T. "PyArabic: A Python Package for Arabic Text." *Journal of Open Source Software*, vol. 8, no. 84, Apr. 2023, p. 4886, https://doi.org/10.21105/joss.04886.