

Clear-cut methodology for Arabic OCR and post-correction with low technical skilled annotators



Alicia González, Tillmann Feige, Thomas Eich

Outline

- 1 The COBHUNI project
- 2 Requirements
- 3 The OCR phase
- 4 Post-correction
- 5 Technical workflow
- 6 Results
- 7 Conclusions

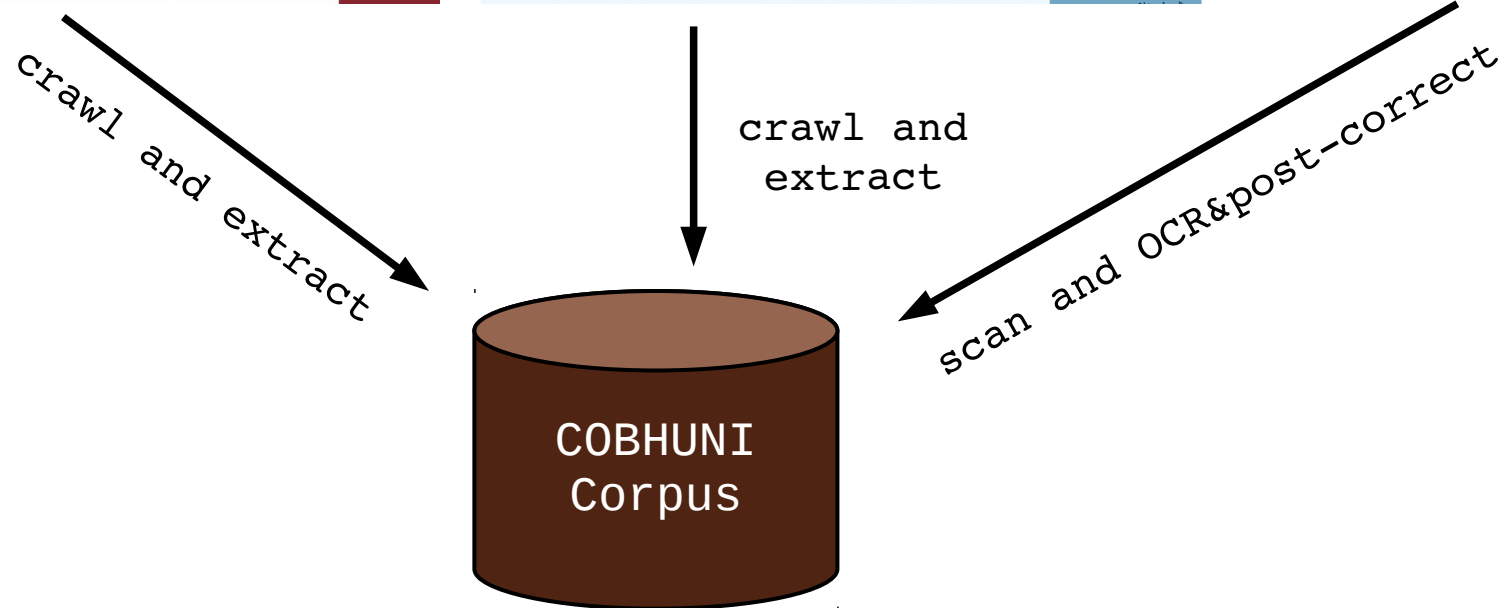
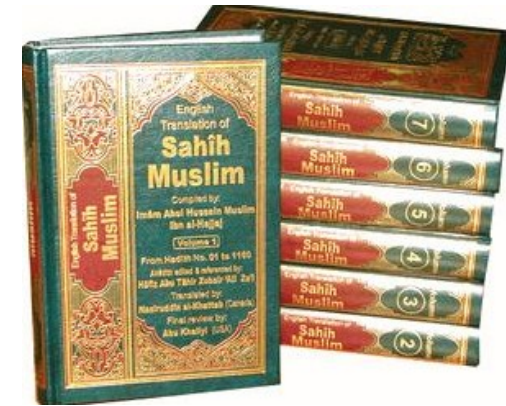
The COBHUNI Project

The **COBHUNI project** aims at diversifying our understanding of how pre-natal life is conceptualized in texts of Islamic normativity.

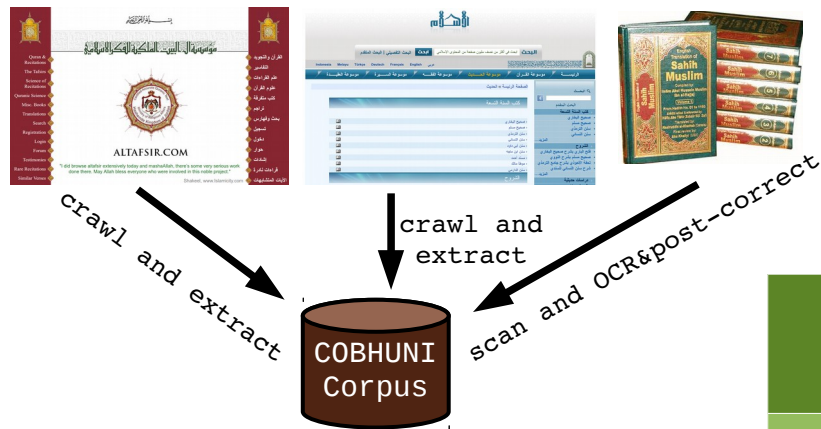
ثُمَّ جَعَلْنَاهُ لُطْفَةً فِي قَرَارٍ مَّكِينٍ ﴿٥٠﴾
ثُمَّ خَلَقْنَا النُّطْفَةَ عَلَقَةً فَخَلَقْنَا
الْعَلَقَةَ مُضْغَةً فَخَلَقْنَا الْمُضْغَةَ
عِظْمًا فَنَسُوجًا فَأَنشَأْنَا الْبَشَرِ
النَّاسُ خَلْقًا آخَرَ فَتَبَرَكَ اللَّهُ أَحْسَنُ
الْمَخْلُوقِينَ ﴿٥١﴾



The COBHUNI Project

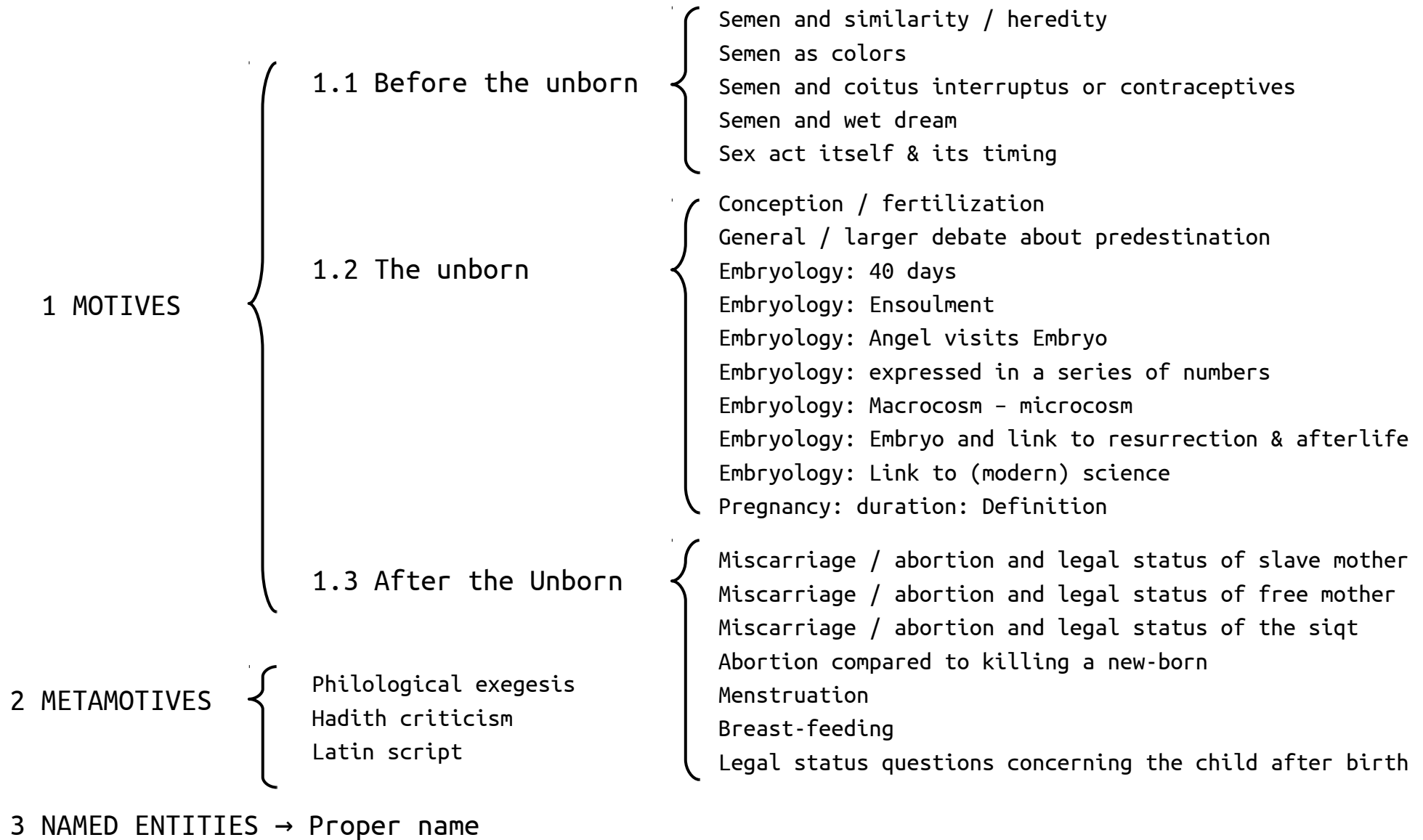


The COBHUNI Project



Source material	No. tokens
<i>altafsir.com</i> (Quran exegesis)	1,306,233
<i>hadith.al-islam.com</i>	144,122
Scan and OCR texts	36,132
TOTAL	1,486,487

The COBHUNI Project



The COBHUNI Project

Great ideas

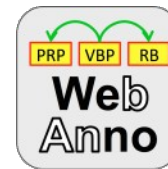
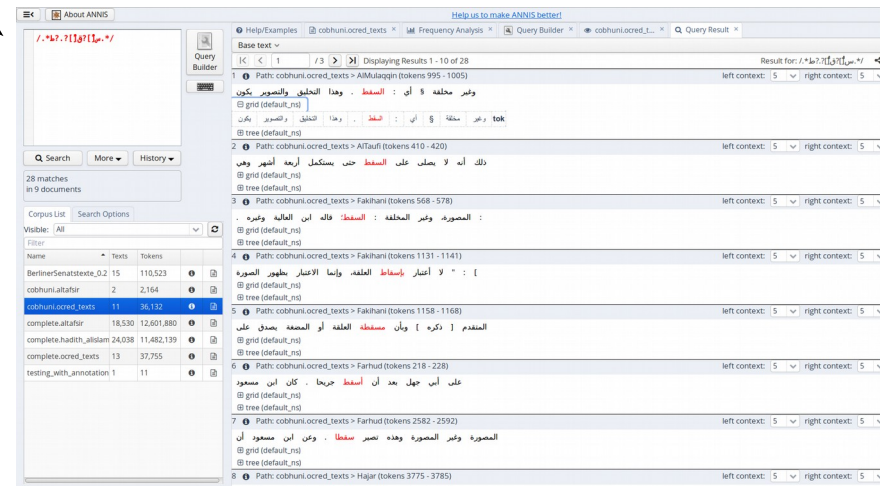


query

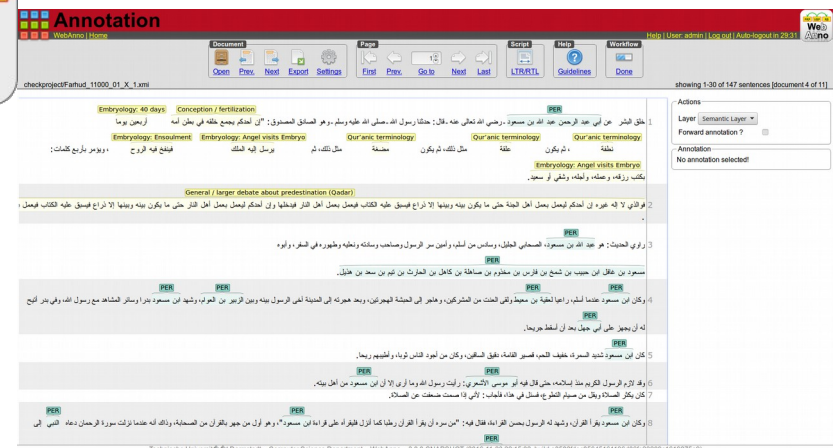
processed
information



Visualization Tool: Annis



Annotation Tool: WebAnno



insert

insert

insert

import

import

export

COBHUNI
Corpus

The COBHUNI Project

Great ideas

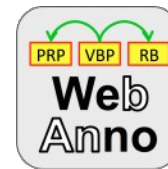
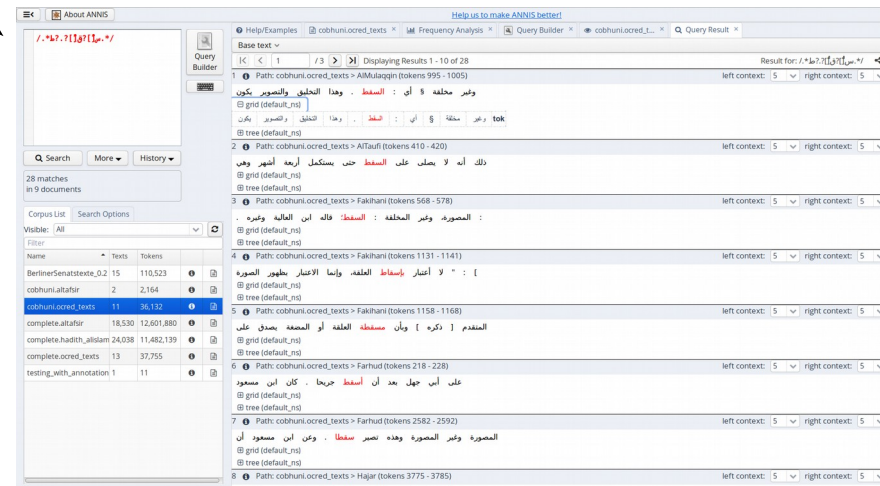


query

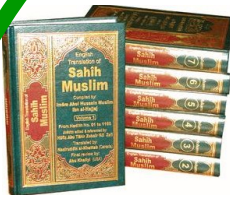
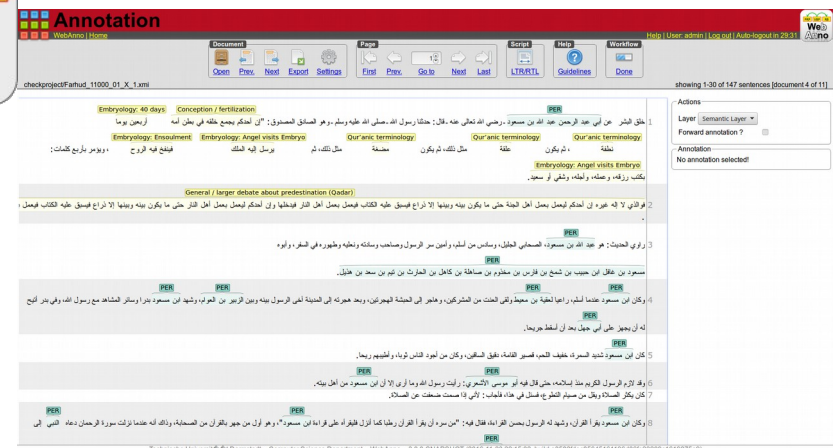
processed
information



Visualization Tool: Annis



Annotation Tool: WebAnno



insert

insert

insert

import

COBHUNI
Corpus

import

export

Requirements

- ✓ OCR engine for Arabic
- ✓ 4+ skilled annotators in Classical Arabic and religious literature
- ✓ Easy-to-use software for post-correction
- ✓ Unicode and RTL support
- ✓ Quality control system

The OCR phase



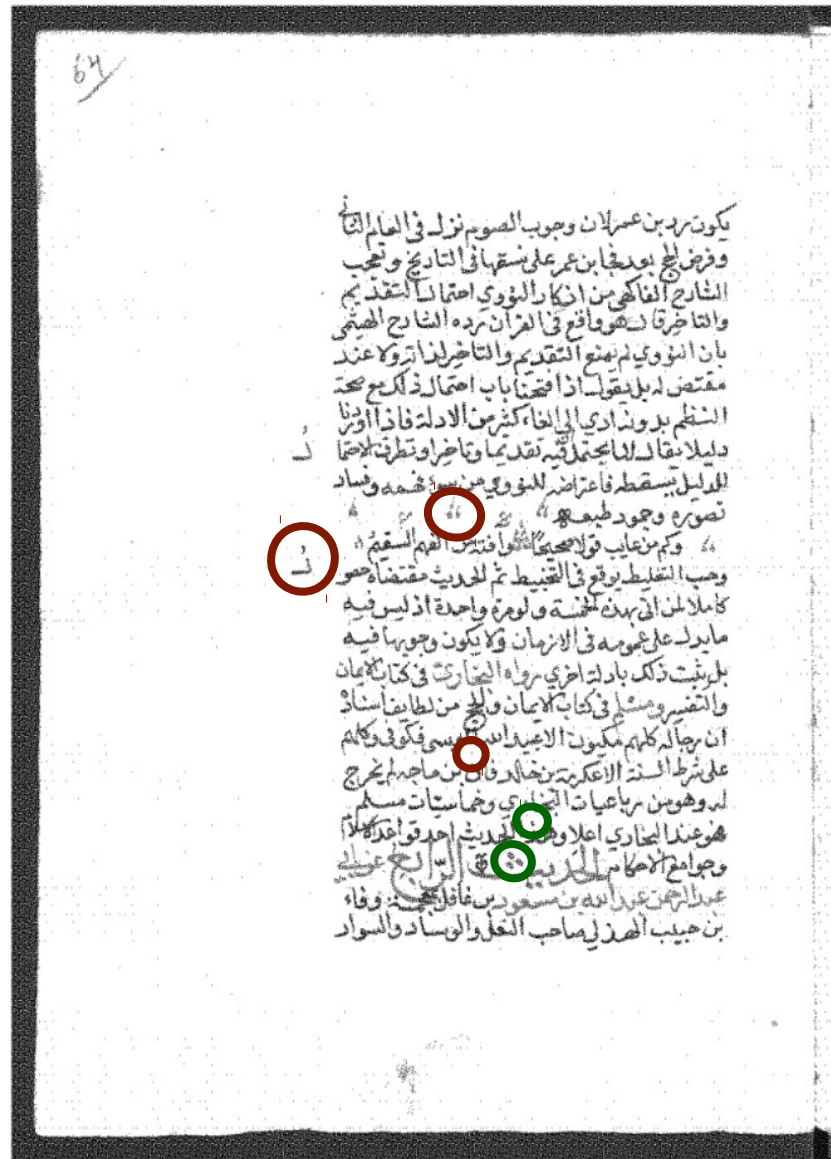
tesseract-ocr

An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.

- ✓ Open Source
- ✓ State-of-the-art levels of accuracy



The OCR phase



Post-correction

Human resources

Corrector	Native	Group
A	yes	1
B	no	1
C	yes	2
D	no	2

- ✓ Highly skilled in classical Arabic
- ✓ Master students in related fields

Post-correction MediaWiki Proofread extension

[English](#) [Alicia](#) [Talk](#) [Preferences](#) [Watchlist](#) [Contributions](#) [Log out](#)

[Read](#) [Edit](#) [View history](#) [More](#)

Page:Al-Taufi.djvu/1

This page has been proofread

صفحة ٨٣

الحديث الرابع

عن أبي عبد الرحمن عبد الله بن مسعود رضي الله عنه قال: "حدثنا رسول الله صلى الله عليه وسلم وهو الصادق المصدق إن أحدكم يجمع خلقه في بطن أمه أربعين يوماً

كتاب التعيين في شرم الأربعين للطنوفي

الحديث الرابع :

عن أبي عبد الرحمن عبد الله بن مسعود رضي الله عنه قال : حدثنا رسول الله ﷺ وهو الصادق المصدق إن أحدكم يجمع خلقه في بطن أمه أربعين يوماً نطفة ، ثم يكون علقه مثل ذلك ، ثم يكون مضغاً مثل ذلك ، ثم يرسل الله إليه الملك فينفخ فيه الروح ويؤمر بأربع كلمات بكب رزقه وأجله وعمله وشقي أو سعيد . فوالذي لا إله غيره إن أحدكم ليعمل بعمل أهل الجنة حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل النار حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل الجنة فيدخلها . رواه البخاري ومسلم .

الكلام على هذا الحديث في لفظه ومعناه .

أما لفظه فمنه / قوله : حدثنا رسول الله ﷺ ، وهو أصل فيما يستعمله المحدثون من قولهم حدثنا ، وآخرنا ، وأنبأنا . ومعنى حدثنا أنشأ لنا خيراً حادثاً .

ومنه الصادق المصدق فالصادق الآتي بالصدق ، وهو الخير المطابق ، المصدق الذي ، بأنه غيره بالصدق ، علم هذا القياس . الكاذب ، المكذب ،

✓ Unicode and RTL support

✓ Extremely easy to use

Post-correction MediaWiki Proofread extension

[English](#) [Alicia](#) [Talk](#) [Preferences](#) [Watchlist](#) [Contributions](#) [Log out](#)



[Page](#) [Discussion](#) [Image](#) [Edit](#) [View history](#) [More](#)

[Search](#)

Editing Page:Al-Taufi.djvu/1

[B](#) [I](#) [Advanced](#) [Special characters](#) [Help](#) [Proofread tools](#)

صفحة ٨٣

== الحديث الرابع ==

عن أبي عبدالرحمن عبدالله بن مسعود رضي
الله عنه قال: "حدثنا

رسول الله صلى الله عليه وسلم وهو
الصادق المصدوق إن أحدكم يجمع خلقه في
بطن أمه أربعين يوماً نطفة، ثم يكون
علقة مثل ذلك، ثم يكون مضغة مثل ذلك ثم
يرسل الله إليه الملك فينفخ فيه الروح

ويؤمر بأربع كلمات يكتب رزقه
وأجله وعمله وشقي أو سعيد. فوالذي لا

إله غيره إن أحدكم ليعمل يوماً

كتاب التعيين في شرم الأربعين للطوفي

الحديث الرابع :

عن أبي عبدالرحمن عبدالله بن مسعود رضي الله عنه قال : حدثنا
رسول الله ﷺ وهو الصادق المصدوق إن أحدكم يجمع خلقه في بطن أمه
أربعين يوماً نطفة ، ثم يكون علقه مثل ذلك ، ثم يكون مضغة مثل ذلك ،
ثم يرسل الله إليه الملك فينفخ فيه الروح ويؤمر بأربع كلمات يكتب رزقه
وأجله وعمله وشقي أو سعيد . فوالذي لا إله غيره إن أحدكم ليعمل بعمل
أهل الجنة حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل الجنة فيدخلها .
وبعمل أهل النار فيدخلها . وإن أحدكم ليعمل بعمل أهل النار حتى ما يكون
بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل النار فيدخلها .
رواه البخاري ومسلم .

الكلام على هذا الحديث في لفظه ومعناه .

أما لفظه فممنه / قوله : حدثنا رسول الله ﷺ ، وهو أصل فيما يستعمله
المحدثون من قولهم حدثنا ، وأخبرنا ، وأنبأنا . ومعنى حدثنا أنشأ لنا خبراً
حدثنا .

Post-correction

Quality control

- ✓ Revision carried out by annotators
- ✓ Error checker

Warnings

- Character absent in the Arabic charset
- Token too long (8 chars excluding diacritics)

Errors

- Ta marboota found at the beginning or in the middle of a word
- More than one short vowels together

https://gitlab.com/alrazi/ini_xmiconverter/blob/master/src/main/java/ini_xmiconverter/XmiConverterOcred.java

Post-correction

Quality control

- ✓ Revision carried out by annotators
- ✓ Error checker

Warnings

- Character absent in the Arabic charset
- Token too long (8 chars excluding diacritics)

Errors

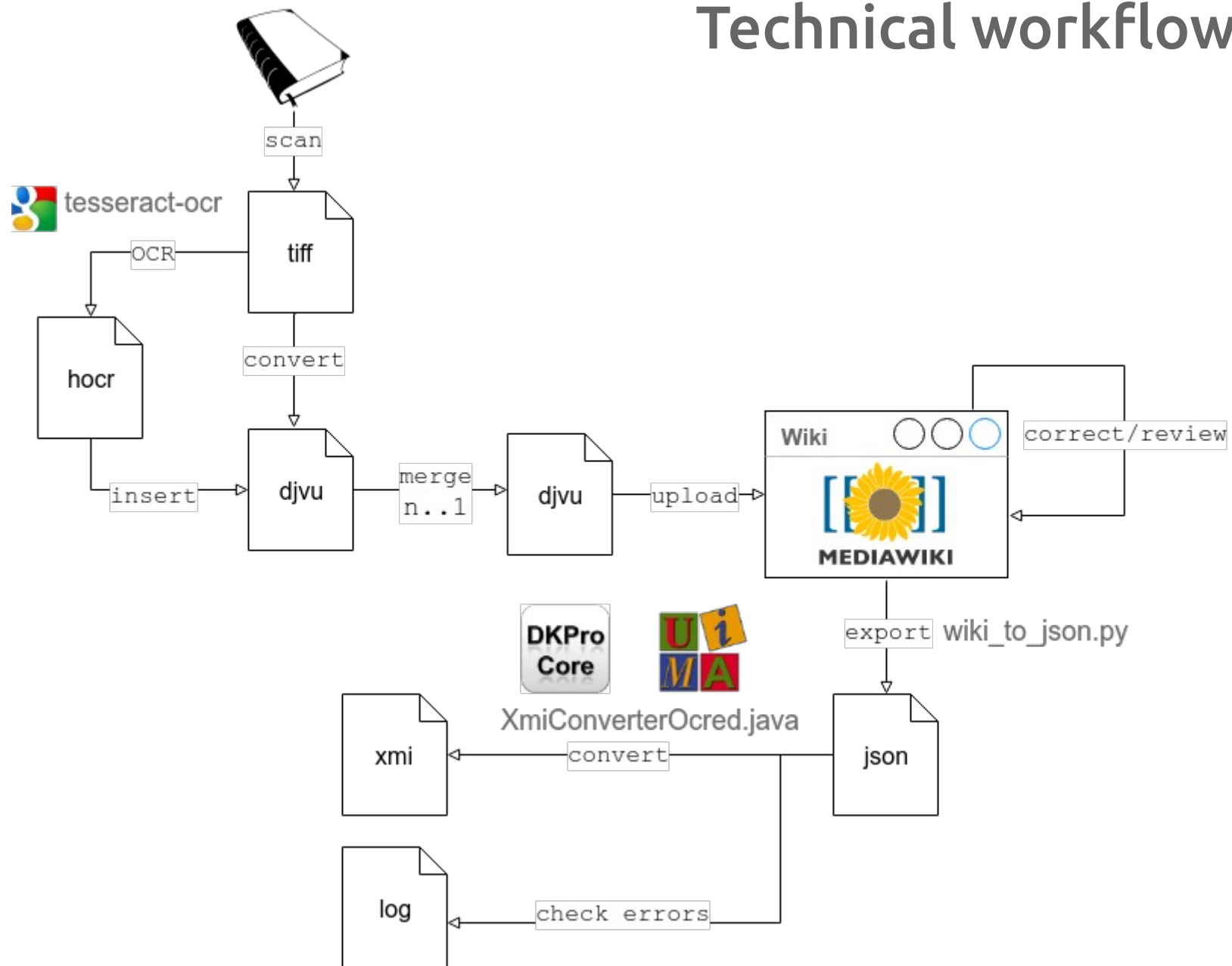
- Ta marboota found at the beginning or in the middle of a word
- More than one short vowels together

https://gitlab.com/alrazi/ini_xmiconverter/blob/master/src/main/java/ini_xmiconverter/XmiConverterOcred.java

Post-correction

```
~/Desktop/error_checker.log - Sublime Text
File Edit Selection Find View Goto Tools Project Preferences Help
error_checker.log
1 Warning in page 1 of scan Attaiyin.djvu: opening character missing for char »
2 Warning in page 2 of scan Attaiyin.djvu: Arabic zero "." may be in position of a dot "."
3 Warning in page 2 of scan Al-mulqin.djvu: Arabic zero "." may be in position of a dot "."
4 Modification in page 2 of scan Al-mulqin.djvu: waw separated from quoted word.
5 Modification in page 2 of scan Al-mulqin.djvu: waw separated from quoted word.
6 Warning in page 3 of scan Al-mulqin.djvu: Arabic zero "." may be in position of a dot "."
7 Modification in page 3 of scan Al-mulqin.djvu: waw separated from quoted word.
8 Warning in page 3 of scan Al-mulqin.djvu: Arabic zero "." may be in position of a dot "."
9 Warning in page 5 of scan Al-mulqin.djvu: Arabic zero "." may be in position of a dot "."
10 Warning in page 1 of scan Al-mulqin.djvu: closing character missing for char «
11 Warning in page 1 of scan Al-mulqin.djvu: closing character missing for char "
12 Warning in page 1 of scan Al-mulqin.djvu: closing character missing for char «
13 Warning in page 10 of scan Fakihani.djvu: Arabic zero "." may be in position of a dot "."
14 Warning in page 10 of scan Fakihani.djvu: closing character missing for char $
15 Modification in page 6 of scan Farhud.djvu: All Quotation marks except "«»" normalised to (")
16 Warning in page 4 of scan Farhud.djvu: closing character missing for char $
17 Warning in page 15 of scan Farhud.djvu: closing character missing for char "
18 Warning in page 1 of scan Hajar.djvu: closing character missing for char "
19 Warning in page 12 of scan Rajab.djvu: Arabic zero "." may be in position of a dot "."
20 Warning in page 16 of scan Rajab.djvu: Arabic zero "." may be in position of a dot "."
21 Warning in page 16 of scan Rajab.djvu: Arabic zero "." may be in position of a dot "."
22 Warning in page 14 of scan Rajab.djvu: closing character missing for char (
23 Modification in page 1 of scan Iyad.djvu: Double prime character (U+2033) found and changed to tanwin hamza (U+06b4).
24 Modification in page 2 of scan Iyad.djvu: All Quotation marks except "«»" normalised to (")
25 Modification in page 3 of scan Iyad.djvu: All Quotation marks except "«»" normalised to (")
26 Warning in page 3 of scan Iyad.djvu: closing character missing for char $
27 Warning in page 4 of scan Iyad.djvu: closing character missing for char "
28 Modification in page 2 of scan Qurtubi.djvu: waw separated from quoted word.
29 Warning in page 3 of scan Qurtubi.djvu: opening character missing for char "
30 Warning in page 4 of scan Qurtubi.djvu: opening character missing for char "
31 Warning in page 5 of scan Qurtubi.djvu: Arabic zero "." may be in position of a dot "."
32 Warning in page 3 of scan Qurtubi.djvu: closing character missing for char (
33 Warning in page 4 of scan Qurtubi.djvu: closing character missing for char (
34 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "اعم:" may contain a typo.
35 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "ايحل:" may contain a typo.
36 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "ايحل:" may contain a typo.
37 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "دلو:" may contain a typo.
38 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "ايرسلاب:" may contain a typo.
39 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "ايرسلاب:" may contain a typo.
40 Warning in section "عبارل شي دحل" of scan Attaiyin.djvu: word "اموي:" may contain a typo.
41 Warning in section "عبارل شي دحل" of scan Al-mulqin.djvu: word "ملسرمو:" may contain a typo.
42 Warning in section "عبارل شي دحل" of scan Al-mulqin.djvu: word "شيدكعنسلال" may contain a typo.
43 Warning in section "عبارل شي دحل" of scan Al-mulqin.djvu: word "مدآعاجم" may contain a typo.
44 Warning in section "هولع ماللل م" of scan Al-mulqin.djvu: word "ثالث:" may contain a typo.
45 Warning in section "عبارل شي دحل" of scan Al-mulqin.djvu: word "لل:" may contain a typo.
Line 1, Column 1 Spaces: 4 Plain Text
```

Technical workflow



Results

Document	No. tokens	Group	Corrector	Reviewer
Al-Taufi	744	1	A	B
Al-Mulaqqin	1,525	1	A	B
Fakihani	2,557	1	A	B
Farhud	4,012	1	A	B
Fashni	2,143	1	B	A
Ibn Hajar	8,965	1	B	A
Ibn Rajab	3,739	2	C	D
Munawi	5,538	2	C	D
Qadi Iyad	1,398	2	C	D
Qurtubi	1,666	2	D	C
Nabrawi	3,845	2	D	C

Conclusions

- ✓ Lack of easy-to-use software for OCR post-correction
- ✓ MediaWiki Proofread extension is a suitable solution due to easy usability and RTL support
- ✓ We managed to successfully implement an efficient and simple workflow for the tasks of OCRing post-correcting and quality control

https://github.com/cobhuni/wiki_export

https://github.com/cobhuni/ini_xmiconverter

شکرا جزىلا

Clear-cut methodology for Arabic OCR and post-correction with low technical skilled annotators



Alicia González, Tillmann Feige, Thomas Eich

Good afternoon. I'm Alicia Gonzalez and I'm going to present the paper Clear-cut methodology for Arabic OCR and post-correction with low technical skilled annotators.

Outline

- 1 The COBHUNI project
- 2 Requirements
- 3 The OCR phase
- 4 Post-correction
- 5 Technical workflow
- 6 Results
- 7 Conclusions

I divided the presentation in 7 parts.
First, I will present the COBHUNI Project, in which we are working on, to give you some context.
Second, I will list our requirements for developing an OCR and post-correction workflow.
Third, I will show how we did the OCR.
Fourth, I will describe the post-correction.
Then, I will sum-up showing the complete technical workflow.
Then, the results we achieved,
And in the end, I will present the conclusions.

The COBHUNI Project

The **COBHUNI project** aims at diversifying our understanding of how pre-natal life is conceptualized in texts of Islamic normativity.

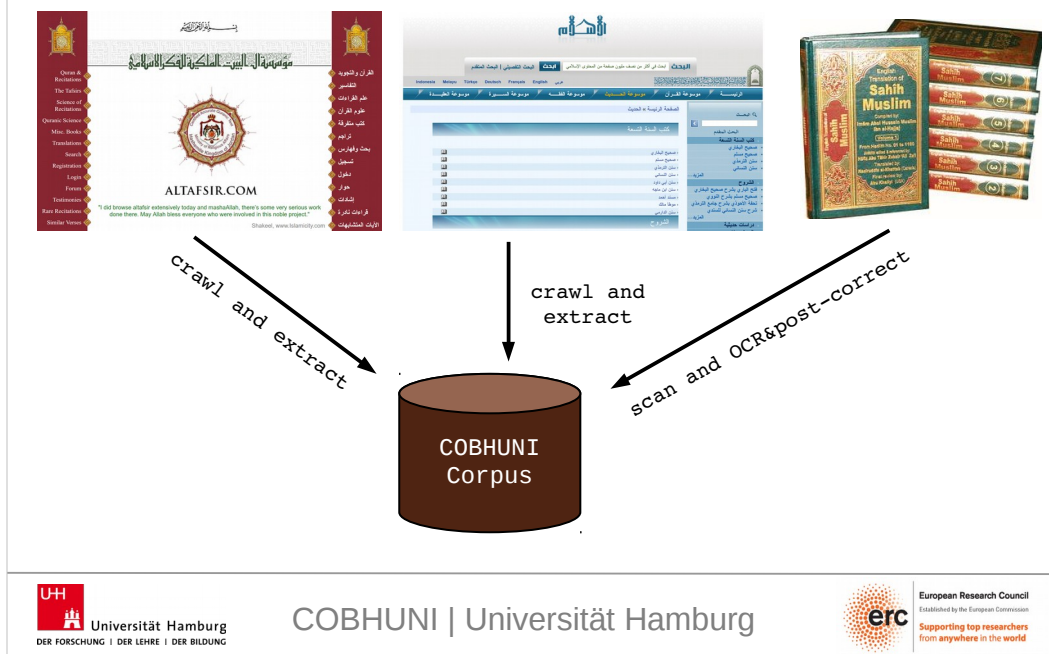
ثُمَّ جَعَلْنَاهُ نَظْفَةً فِي فَئِزٍ مَّكِينٍ
ثُمَّ خَلَقْنَا النُّطْفَةَ عَلَقَةً فَخَلَقْنَا
الْعَلَقَةَ مُضْغَةً فَخَلَقْنَا الْمُضْغَةَ
عِظَامًا فَكَسَّوْنَا الْعِظَامَ لَحْمًا ثُمَّ
أَنشَأْنَاهُ خَلْقًا آخَرَ فَتَبَرَكَ ذَاكَ اللَّهُ أَحْسَنُ
الْخَالِقِينَ



The COBHUNI project aims at diversifying our understanding of how pre-natal life is conceptualized in texts of Islamic normativity.

So our researchers need to analyse a large amount of texts related to Islamic embryology and reach a deep understanding of the ideas and conceptions present in those texts.

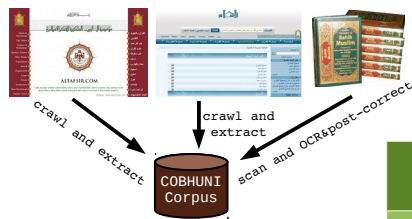
The COBHUNI Project



We have basically three sources of text we want to analyse:

- * altafsir.com, a webpage that contains quranic exegesis.
- * hadith-al-islam, a webpage that includes a large amount of hadith material along with its commentaries.
- * and then we have physical books with other hadith material that we couldn't find in electronic format. So we needed to scan and OCR those texts.

The COBHUNI Project



Source material	No. tokens
<i>altafsir.com</i> (Quran exegesis)	1,306,233
<i>hadith.al-islam.com</i>	144,122
Scan and OCR texts	36,132
TOTAL	1,486,487

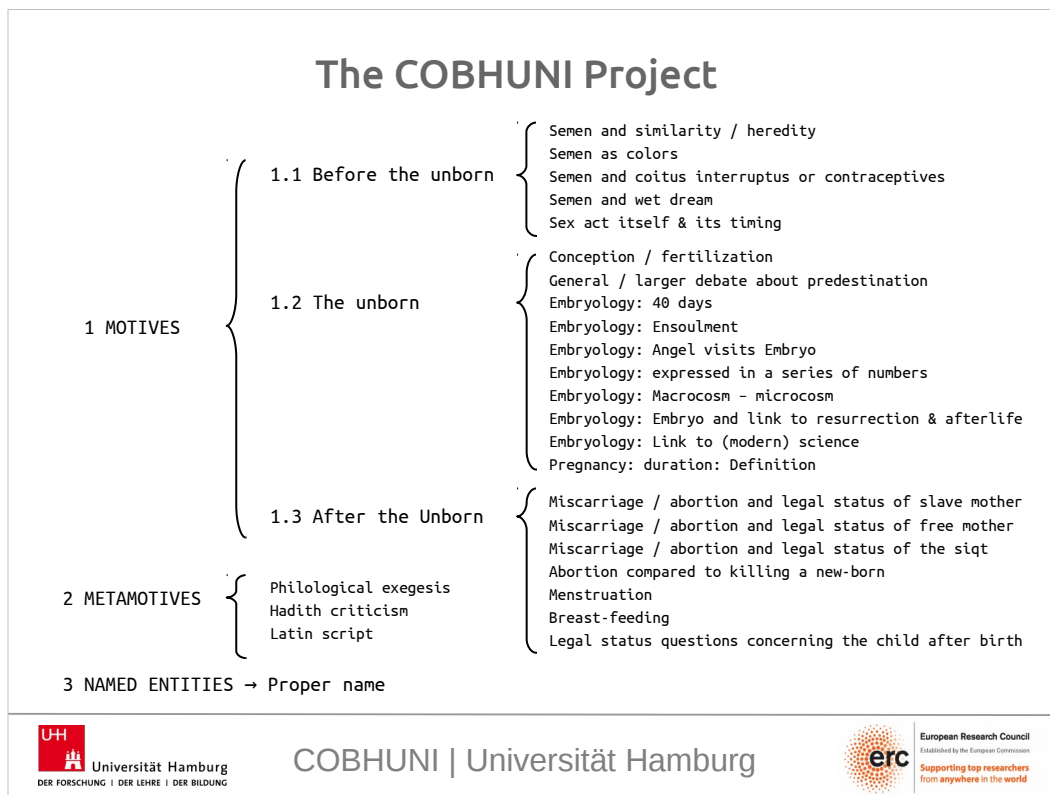
And here we have the number of tokens of the selected material from each source.

Up to now, from *altafsir.com* we have 1.3 million tokens.

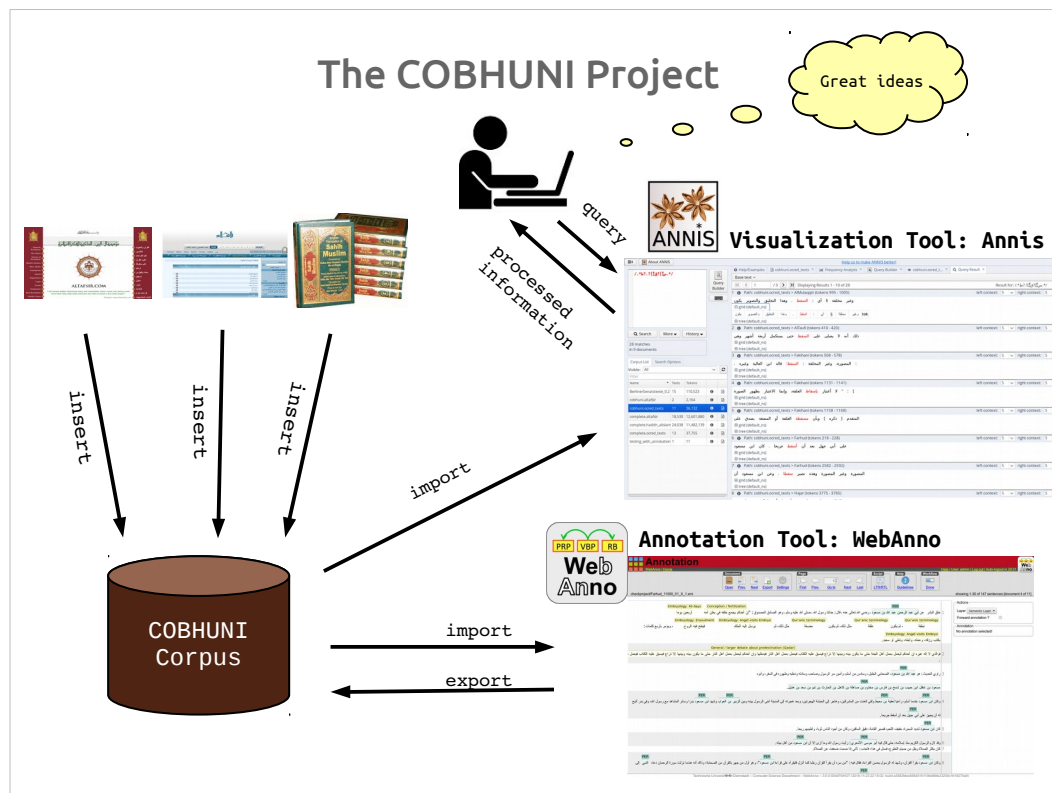
From *hadith.al-islam* we have 144 thousand tokens.

And from the scanned and OCR-ed material we have 36 thousand tokens.

This makes a total of almost 1.5 million tokens.



Now, what do we want to find in these texts? This is the tagset of semantic concepts developed by our group of researchers containing the ideas that we want to identify in the texts. So we are right now annotating the texts with this tagset.

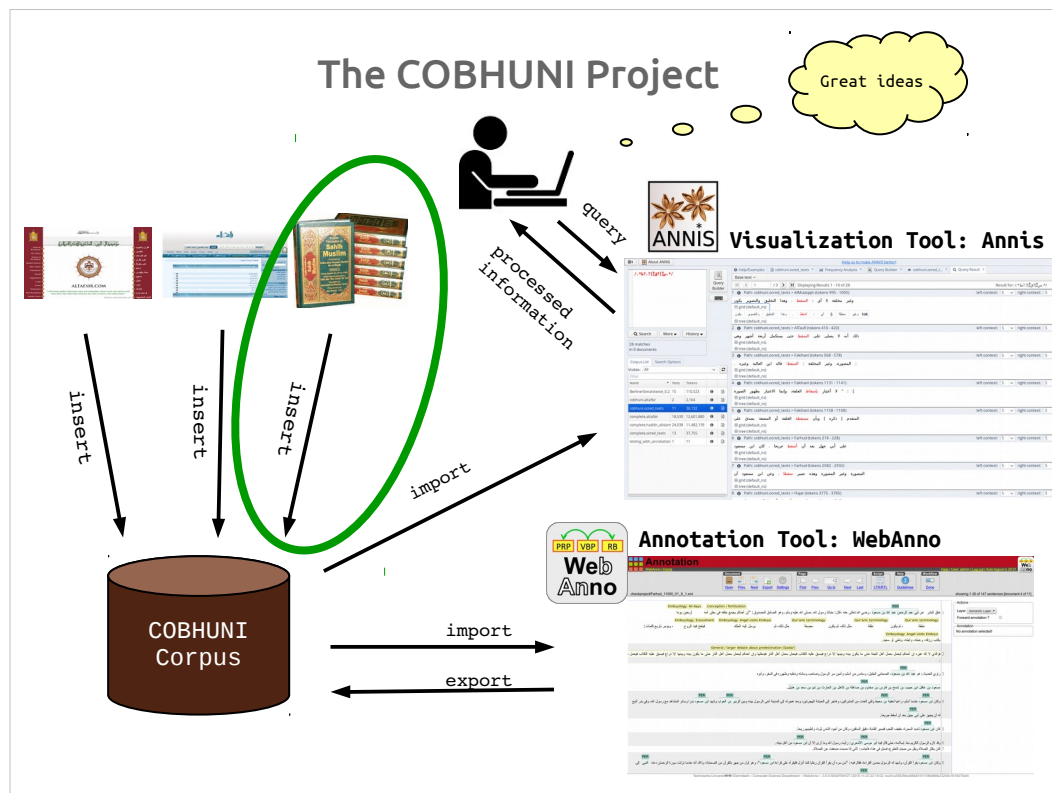


And this is the general workflow of the COBHUNI project.

First, we collect all the text material for the corpus.

Then we annotate this corpus with semantic information. For this, we are using a tool called WebAnno, developed in the University of Darmstadt.

And finally, we ingest this enriched data into a visualization tool so that researchers can query the texts and find relevant information for their analyses. For the visualization, we chose Annis, a software developed in Humboldt University.



So the aim of this presentation is to show how we developed a methodology for OCR-ing and post-correcting the text material taken from physical books.

Requirements

- ✓ OCR engine for Arabic
- ✓ 4+ skilled annotators in Classical Arabic and religious literature
- ✓ Easy-to-use software for post-correction
- ✓ Unicode and RTL support
- ✓ Quality control system

So, what are our requirements?

Of course, we need an OCR engine for Arabic.

We need at least 4 annotators highly skilled in Classical Arabic and knowledgeable in religious literature.

Unfortunately, it's quite uncommon to find highly skilled annotators in the subject we are working on that also have high technical knowledge. So, it is crucial that we have an easy-to-use software for post-correction. Also, the software we use must have Unicode and RTL support. And this is something many software still lacks.

And at last, we need to have some sort of quality control system.

The OCR phase



tesseract-ocr

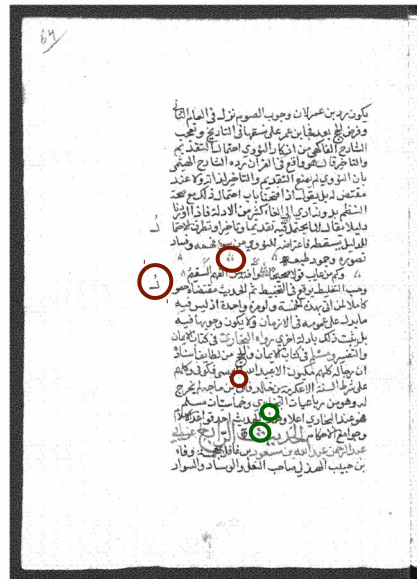
An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.

- ✓ Open Source
- ✓ State-of-the-art levels of accuracy

So, for the OCR, we tried different systems and ended up choosing the Tesseract engine, which is currently developed by google.

An advantage of Tesseract compared to other systems is that it is open source, and, although it is still not very good for Arabic, it achieves state-of-the-art levels of accuracy, around 80%. This is because Arabic cursive script is still a challenge for OCR engines.

The OCR phase



Another problem is that Arabic script contains many meaningful dots that are sometimes confused with ornamental strokes, dirt in the paper or noise in the quality of the scans.

And the other way round. Ornaments and noise may be mistakenly interpreted as part of a letter of the Arabic alphabet.

Here we have an example of an especially complicated document. In manuscripts, OCR engines achieve very low results and typically turned out to be worse than typing the text from scratch.

The red circles in the image show strokes that must be excluded from the text, whereas the green circles are part of letters.

Starting from above, the first red one in the middle is an ornament, the one on the left side is a gloss, the small red one is just some noise. And the green ones are part of letters.

Post-correction


Human resources

Corrector	Native	Group
A	yes	1
B	no	1
C	yes	2
D	no	2

- ✓ Highly skilled in classical Arabic
- ✓ Master students in related fields

So, for the post-correction we chose 4 master students fluent in Arabic, 2 native and 2 non-native. And we separated them in 2 groups, each having a native and a non-native person. The idea is that both a native and a non-native annotator end up checking the same texts.

And all annotators were master students in fields relevant to the project.



Post-correction

MediaWiki Proofread extension

English Alicia Talk Preferences Watchlist Contributions Log out

Page Discussion Image Read Edit View history More Search

Page:Al-Taufi.djvu/1

This page has been proofread

Main page

Recent changes

Random page

Help

Tools

What links here

Related changes

Upload file

Special pages

Printable version

Permanent link

Page information

صفحة ٨٣

الحديث الرابع

عن أبي عبد الرحمن عبد الله بن مسعود رضي الله عنه قال : حدثنا رسول الله ﷺ وهو الصادق المصدوق إن أحدكم يجمع خلقه في بطن أمه أربعين يوما نطفة ، ثم يكون علقة مثل ذلك ، ثم يكون مضغة مثل ذلك ، ثم يرسل الله إليه الملاك فينفخ فيه الروح ويؤمر بأربع كلمات يكتب رزقه وأجله وعمله وشقي أو سعيد . فوالذي لا إله غيره إن أحدكم ليعمل بعمل أهل الجنة حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل النار فيدخلها . وإن أحدكم ليعمل بعمل أهل النار حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل الجنة فيدخلها . » رواه البخاري ومسلم .


الكلام على هذا الحديث في لفظه ومعناه .

أما لفظه فانه / قوله : حدثنا رسول الله ﷺ ، وهو أصل فيما يستعمله المحدثون من قولهم حدثنا ، وأخبرنا ، وأتانا . ومعنى حدثنا أنشأ لنا حديثا جادا .

ومنه الصادق المصدوق فالصادق الآتي بالصدق ، وهو الخبر المطابق ، والمصدوق ، لأنه قد أتته به بالصدق ، غير هذا القام . الكاذب ، المكذب ،


Unicode and RTL support

Extremely easy to use



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

COBHUNI | Universität Hamburg



European Research Council
Established by the European Commission
Supporting top researchers from anywhere in the world

Then, the difficult part was to find a software that was easy to use for the annotators, easy to configure, compatible with the Arabic writing system, and server-client based, so that the annotators don't have to share data and can work easily on a collaborative way.

We decided to use the MediaWiki Proofread extension. It was compatible with RTL scripts, extremely easy to use, and we were already using the wiki for documentation purposes, so the proofread extension fitted very well in our workflow.

The interface is really simple—on the right part of the window we have the scanned image and, on the left, we copied the text resulted from the OCR.



Post-correction

MediaWiki Proofread extension

[English](#)
[Alicia](#)
[Talk](#)
[Preferences](#)
[Watchlist](#)
[Contributions](#)
[Log out](#)

[Page](#)
[Discussion](#)
[Image](#)
[Recent changes](#)
[New history](#)
[More](#)

[Edit](#)

Editing Page:Al-Taufi.djvu/1

B I

Advanced Special characters Help Proofread tools

صفحة ٨٣

== الحديث الرابع ==

عن أبي عبد الرحمن عبد الله بن مسعود رضي الله عنه قال : حدثنا رسول الله صلى الله عليه وسلم وهو الصادق المصدوق إن أحدكم يجمع خلقه في بطن أمه أربعين يوما نطفة ، ثم يكون علقة مثل ذلك ، ثم يكون مضغة مثل ذلك ، ثم يرسل الله إليه الملك فينفخ فيه الروح ويؤمر بأربع كلمات بكتب رزقه وأجله وعمله وشقي أو سعيد . فوالذي لا إله غيره إن أحدكم ليعمل بعمل أهل الجنة حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل الجنة فيدخلها . وإن أحدكم ليعمل بعمل أهل النار حتى ما يكون بينه وبينها إلا ذراع فيسبق عليه الكتاب فيعمل بعمل أهل النار فيدخلها . رواه البخاري ومسلم .

الكلام على هذا الحديث في لفظه ومعناه .

أما لفظه فانه / قوله : حدثنا رسول الله صلى الله عليه وسلم ، وهو أصل فيما يستعمله المحدثون من قولهم حدثنا ، وأخبرنا ، وأنبأنا . ومعنى حدثنا أنشأ لنا شيئا حدثنا .

Main page

Recent changes

Random page

Help

Tools

What links here

Related changes

Upload file

Special pages

Page information



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

COBHUNI | Universität Hamburg



European Research Council

Established by the European Commission

Supporting top researchers from anywhere in the world

To work on the text, the annotators just have to enter the edit mode and start correcting.

We had to made some configuration changes to be able to work with Arabic. The Proofread extension relies on the main language of the wiki, in this case English. Therefore we had to adjust the RTL alignment just for the editor and increase the font size for readability.


Post-correction

Quality control

- ✓ Revision carried out by annotators
- ✓ Error checker


<h4>Warnings</h4> <ul style="list-style-type: none"> ➤ Character absent in the Arabic charset ➤ Token too long (8 chars excluding diacritics) 	<h4>Errors</h4> <ul style="list-style-type: none"> ➤ Ta marboota found at the beginning or in the middle of a word ➤ More than one short vowels together
---	--

https://gitlab.com/alrazi/ini_xmiconverter/blob/master/src/main/java/ini_xmiconverter/XmiConverterOcred.java



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

COBHUNI | Universität Hamburg



European Research Council
Established by the European Commission
Supporting top researchers from anywhere in the world

We included two quality control measures.

First, for each text, we had two people working on it.

One doing the correction itself and a second one checking the corrected text.

Second, we developed an ad-hoc error checker to scan the text after the annotators had finished both the correction and the review of the correction.

This tool prints a warning, for example, when it finds a character absent in the Arabic charset or a word is considered too long. We set the maximum length at 8 characters, excluding diacritics, because we found that limit to be a good balance between false negatives and false positives. And this checking was very successful for finding words written together because of the lack of a space in between them.

Post-correction

Quality control

- ✓ Revision carried out by annotators
- ✓ Error checker

Warnings

- Character absent in the Arabic charset
- Token too long (8 chars excluding diacritics)

Errors

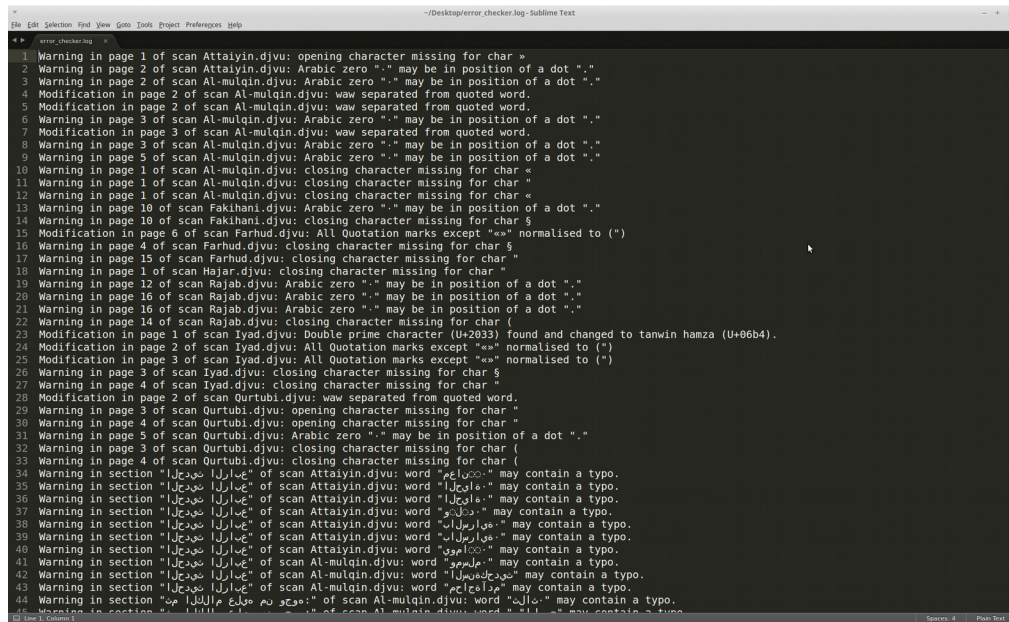
- Ta marboota found at the beginning or in the middle of a word
- More than one short vowels together

https://gitlab.com/alrazi/ini_xmiconverter/blob/master/src/main/java/ini_xmiconverter/XmiConverterOcred.java

On the other hand, the tool prints an error for example if it finds an obvious mistake in orthography. In the first error checking, the ta marboota is a consonant letter that can only appear at the end of a word. So if it is found in another position, we can be sure that there is an error in the text. In the second checking, we alert when two short vowels are written together, because there's a special unicode character for that or maybe the annotators meant to write only one vowel.

The error checker is integrated in another module that exports the texts from the MediaWiki into json and xmi files. If a warning appears, the conversion continues as expected. But if an error is found, the conversion is aborted. So the annotators must fix the error in the MediaWiki and then we execute the conversion again, until there are no errors.

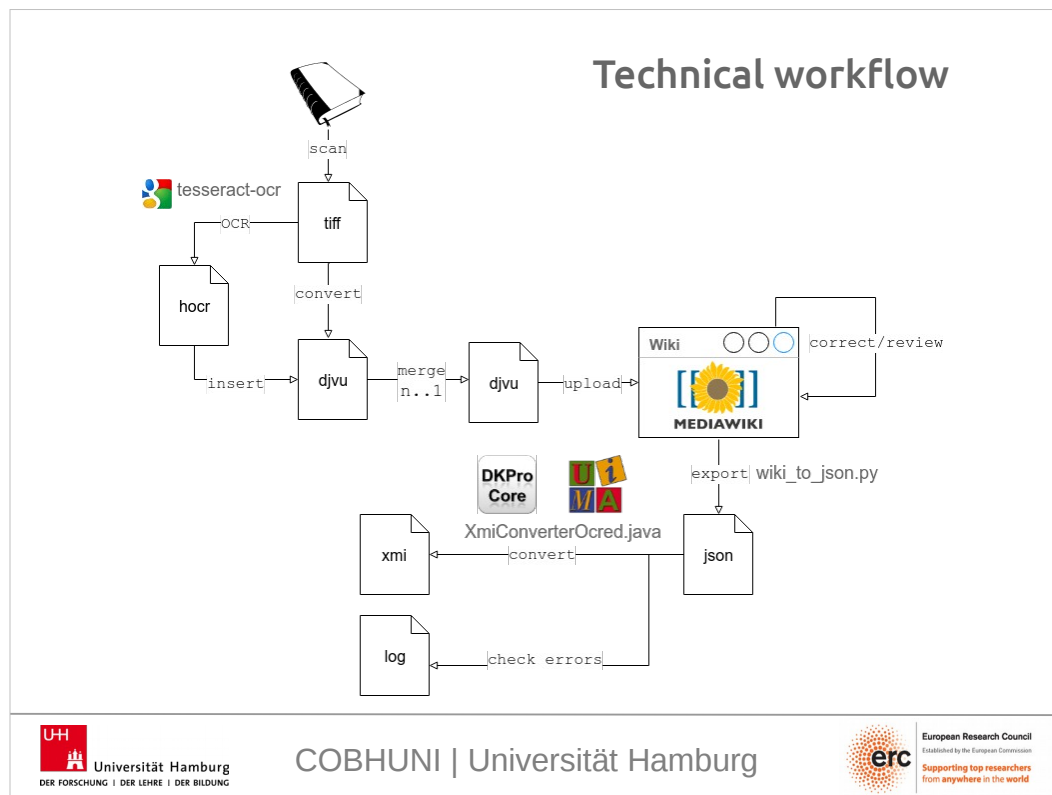
Post-correction



```
File Edit Selection Find View Goto Tools Project Preferences Help
--Desktop/error checker.log-Sublime Text
error_checker.log
1 Warning in page 1 of scan Attaiyin.djvu: opening character missing for char «
2 Warning in page 2 of scan Attaiyin.djvu: Arabic zero "٠" may be in position of a dot "."
3 Warning in page 2 of scan Al-mulqin.djvu: Arabic zero "٠" may be in position of a dot "."
4 Modification in page 2 of scan Al-mulqin.djvu: waw separated from quoted word.
5 Modification in page 2 of scan Al-mulqin.djvu: waw separated from quoted word.
6 Warning in page 3 of scan Al-mulqin.djvu: Arabic zero "٠" may be in position of a dot "."
7 Modification in page 3 of scan Al-mulqin.djvu: waw separated from quoted word.
8 Warning in page 3 of scan Al-mulqin.djvu: Arabic zero "٠" may be in position of a dot "."
9 Warning in page 5 of scan Al-mulqin.djvu: Arabic zero "٠" may be in position of a dot "."
10 Warning in page 1 of scan Al-mulqin.djvu: closing character missing for char «
11 Warning in page 1 of scan Al-mulqin.djvu: closing character missing for char "
12 Warning in page 1 of scan Al-mulqin.djvu: closing character missing for char «
13 Warning in page 10 of scan Fakihani.djvu: Arabic zero "٠" may be in position of a dot "."
14 Warning in page 10 of scan Fakihani.djvu: closing character missing for char §
15 Modification in page 6 of scan Farhud.djvu: All Quotation marks except "«" normalised to (")
16 Warning in page 4 of scan Farhud.djvu: closing character missing for char §
17 Warning in page 15 of scan Farhud.djvu: closing character missing for char "
18 Warning in page 1 of scan Hajar.djvu: closing character missing for char "
19 Warning in page 12 of scan Rajab.djvu: Arabic zero "٠" may be in position of a dot "."
20 Warning in page 16 of scan Rajab.djvu: Arabic zero "٠" may be in position of a dot "."
21 Warning in page 16 of scan Rajab.djvu: Arabic zero "٠" may be in position of a dot "."
22 Warning in page 14 of scan Rajab.djvu: closing character missing for char (
23 Modification in page 1 of scan Iyad.djvu: Double prime character (U+2033) found and changed to tanwin hamza (U+06b4).
24 Modification in page 2 of scan Iyad.djvu: All Quotation marks except "«" normalised to (")
25 Modification in page 3 of scan Iyad.djvu: All Quotation marks except "«" normalised to (")
26 Warning in page 3 of scan Iyad.djvu: closing character missing for char §
27 Warning in page 4 of scan Iyad.djvu: closing character missing for char "
28 Modification in page 2 of scan Qurtubi.djvu: waw separated from quoted word.
29 Warning in page 3 of scan Qurtubi.djvu: opening character missing for char "
30 Warning in page 4 of scan Qurtubi.djvu: opening character missing for char "
31 Warning in page 5 of scan Qurtubi.djvu: Arabic zero "٠" may be in position of a dot "."
32 Warning in page 3 of scan Qurtubi.djvu: closing character missing for char (
33 Warning in page 4 of scan Qurtubi.djvu: closing character missing for char (
34 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
35 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
36 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
37 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
38 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
39 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
40 Warning in section "فهارس الجدل" of scan Attaiyin.djvu: word "تدافع" may contain a typo.
41 Warning in section "فهارس الجدل" of scan Al-mulqin.djvu: word "تدافع" may contain a typo.
42 Warning in section "فهارس الجدل" of scan Al-mulqin.djvu: word "تدافع" may contain a typo.
43 Warning in section "فهارس الجدل" of scan Al-mulqin.djvu: word "تدافع" may contain a typo.
44 Warning in section "فهارس الجدل" of scan Al-mulqin.djvu: word "تدافع" may contain a typo.
45 Warning in section "فهارس الجدل" of scan Al-mulqin.djvu: word "تدافع" may contain a typo.
```

Here we have an example of a log file. In this case we only have warnings. So for example a typical warning is the one in the second line indicating that an Arabic zero is found in a context that is more likely to be a dot.

And then, in the end, we have many words that look suspicious, for example because they are too long. To make the warning and error messages more meaningful, we indicate the name of the document where they are found and the exact page.



So this is the technical workflow.

We scan the texts and store them into tiff files. We OCR them and create hocr files. Then we merge the image and the OCR text into djvu. And we group all djvu files belonging to the same text into one.

When we have all the files prepared, we upload them into the MediaWiki, where the annotators correct and review them. And when this is finished we export them into json, and then we convert them into xmi, which is the format WebAnno uses internally. The conversion module calls the error checker automatically and creates a log file with warnings and errors. And if an error appears the conversion into xmi is aborted.

Results

Document	No. tokens	Group	Corrector	Reviewer
Al-Taufi	744	1	A	B
Al-Mulaqqin	1,525	1	A	B
Fakihani	2,557	1	A	B
Farhud	4,012	1	A	B
Fashni	2,143	1	B	A
Ibn Hajar	8,965	1	B	A
Ibn Rajab	3,739	2	C	D
Munawi	5,538	2	C	D
Qadi Iyad	1,398	2	C	D
Qurtubi	1,666	2	D	C
Nabrawi	3,845	2	D	C

So to sum-up, this is the list of texts OCR-ed and post-corrected, with the number of tokens each have and the work assigned to each annotator as a corrector and as a reviewer.

Conclusions

- ✓ Lack of easy-to-use software for OCR post-correction
- ✓ MediaWiki Proofread extension is a suitable solution due to easy usability and RTL support
- ✓ We managed to successfully implement an efficient and simple workflow for the tasks of OCRing post-correcting and quality control

As a conclusion, there's still a lack of many available easy-to-use software for OCR and post-correction, especially when dealing with RTL languages.

But, we have shown that the MediaWiki Proofread extension turned out to be a very suitable solution for performing Arabic post-correction having a team of low-skilled annotators. And it supports Arabic quite well.

So, we have managed to successfully implement an efficient and simple workflow for the tasks of OCRing, post-correcting and quality control.

https://github.com/cobhuni/wiki_export

https://github.com/cobhuni/ini_xmiconverter

شكرا جزىلا

And this is all. Thank you so much. Here I included the links to the projects containing the code.