TexT – Text Extractor Tool for Handwritten Document Transcription and Annotations

Anders Hast^, Per Cullhed* & Ekta Vats^  
^Department of Information Technology, Division of Visual Information and Interaction  
*Uppsala University Library

14:th Italian Research Conference on Digital Libraries, Udine, Italy
Outline

• Transcription is a tedious task that can be made simple by handwritten text recognition techniques, such as word spotting
• It will be shown how transcription can be done in a new way, but also how researchers can get easy access, using similar techniques, to manuscripts that are otherwise not searchable
• Semi-Automatic transcription
  • In arbitrary word order instead of line by line, word by word
  • (Problem solving through a visualisation of the problem)
Copying books is an old trade
Tres digití scribunt toium corpusque laborat. Scribere qui nescit nullum putat esse laborem.

[Three fingers write and the whole body labours. He who does not know how to write thinks it is no work.]
• The monks tried to brighten up their day by scribbling in the margins

• If copying the text was boring, transcribing it will be likewise boring…
Even modern transcribers make mistakes...

https://transkribus.eu/Transkribus/
• Now, this sounds interesting, but it starts to get really exciting if you consider that
  • once you have properly transcribed e.g. 100 images you may inform us and we will train an HTR engine from the Computational Intelligence Technology Lab (CITlab) of the University of Rostock on your documents and
  • you will be able to transcribe further pages of your documents with the support of automatically produced handwritten text.
Veele handen

Is a Dutch initiative where volunteers can sign in and start transcribing in a simple manner.
From the university of St. Louis in the United States is much used for the transcriptions of medieval manuscripts.
The Oxford based Zooniverse is a general crowdsourcing platform used for many purposes, not only transcription. It has one million users-mostly from English speaking countries.
(1) Transcribe
Volunteer(s) work together to transcribe document. When finished, can request review.

(2) Review
A new volunteer reviews transcription. Can approve or send back to (1) for more edits.

(3) Approve
Smithsonian staff approves transcription or sends back to (1) for more edits.
Every transcriber knows exactly where to start contributing as this is obvious from the GUI
The Purpose of Transcription

• Our goal with this project is to make our collections:
  • more accessible and useful to curators, researchers, and anyone with a curious spirit
• Because computers have a hard time understanding handwriting, many of our collections still hold many secrets and hidden knowledge inside their pages
  • With your help, we can bring that knowledge to life.
• Searchability & Readability - Usefulness over Perfection
TexT-project - Uppsala

(1) Transcribe
Volunteer(s) work together to transcribe document. When finished, can request review.

(2) Review
A new volunteer reviews transcription. Can approve or send back to (1) for more edits.

(3) Approve
Smithsonian staff approves transcription or sends back to (1) for more edits.

HTR

Machine learning
How can we make it fun?

• Why do people love Sudoku?!

• Is it really a game?
  • Or rather about problem solving?

• And the satisfaction of reaching a goal?

• You learn NOTHING?!

• From crosswords you might at least learn something…
Can Transcription be “fun”?

• Can we make it a game?
  • How and why? Should we even??
• Or is it rather about problem solving?
• Can you learn something while doing it?
• This is a challenging task!
• **Gamification** is *the application of game-design elements and game principles in non-game contexts*

• Probably no first person shooter game…
  • not even medieval fighting games… or?!

• What about problem solving?
  • If Sudoku can be fun..
• The user marks up a word
• …and the HTR engine finds it in the text
• Can be used to make large text document collections searchable
• Ex: Monk
  • L. Schomaker, Rijksuniversiteit Groningen
  • Search & annotation tools for handwritten manuscripts
• Monk learns by user annotation
- We have developed an HTR engine for Word Spotting (WS)
- Segmentation Free
  - Based on stitching principles from Computer Vision
  - i.e pattern matching
- Learning free

A Segmentation-free Handwritten Word Spotting Approach by Relaxed Feature Matching
Anders Hast and Alicia Fornés
• We plan to use the WS approach for transcription as well!

• The user marks the red rectangle

• The system finds the “perfect” bounding box

• Principle: “all the word and nothing but the word”!

On-the-fly Historical Handwritten Text Annotation
Ekta Vats and Anders Hast
1st International Workshop on Human-Document Interaction (HDI 2017), in conjunction with ICDAR 2017
Semi Automatic Transcription

- User should mark a word and transcribe it ONCE
- Then the HTR-Engine should find all occurrences
- Annotation possibilities!
  - person, places, time, etc...

TexT - Text Extractor Tool for Handwritten Document Transcription and Annotation
Anders Hast, Per Cullhed and Ekta Vats
14:th Italian Research Conference on Digital Libraries, IRCDL 2017
Problems

• Only whole words are found
• Note how “d’Audenarde” and “a Audenarde” are contracted to one word
  • This is usually a small problem
• A bigger problem are the split words
• Solvable by letter spotting?
Problems with Initials

Close to searched word (Audenarde)

Or if it is the word we search (Quant)
An Example of problems with Orthography

- A section of Marco Polo’s account dealing with Fansur
- Here spelled three different ways:
  - <fanur>
  - <sansur>
  - <fausur(y)>

https://medium.com/@siwaratrikalpa/barus-sumatra-in-a-medieval-egyptian-text-edc7be9517
### Abbreviations is not a new thing

The most common texting abbreviations, **Italian-style**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>mi (I, me)</td>
</tr>
<tr>
<td>t</td>
<td>ti (you)</td>
</tr>
<tr>
<td>xke</td>
<td>perché (why, because)</td>
</tr>
<tr>
<td>cmq</td>
<td>comunque (anyway)</td>
</tr>
<tr>
<td>bc</td>
<td>baci (kisses)</td>
</tr>
<tr>
<td>midi</td>
<td>mi dispiace (I’m sorry)</td>
</tr>
<tr>
<td>pfv</td>
<td>per favore (please)</td>
</tr>
<tr>
<td>d</td>
<td>da (from, since, of)</td>
</tr>
<tr>
<td>grz</td>
<td>grazie (thanks)</td>
</tr>
<tr>
<td>tn</td>
<td>tanto (a lot, much, long time)</td>
</tr>
<tr>
<td>k</td>
<td>chi (who, what)</td>
</tr>
<tr>
<td>c6</td>
<td>Ci sei? (Are you there?)</td>
</tr>
<tr>
<td>qls</td>
<td>qualcosa (something)</td>
</tr>
<tr>
<td>+</td>
<td>più (more)</td>
</tr>
<tr>
<td>risp</td>
<td>rispondi (answer)</td>
</tr>
<tr>
<td>nn</td>
<td>non (no, not)</td>
</tr>
<tr>
<td>prox</td>
<td>prossima (next)</td>
</tr>
<tr>
<td>gg</td>
<td>giorno (day)</td>
</tr>
<tr>
<td>tvb</td>
<td>ti voglio bene (I love you)</td>
</tr>
<tr>
<td>ta</td>
<td>ti amo (I love you)</td>
</tr>
</tbody>
</table>

Common abbreviations, **Monk-style**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>b,b</td>
<td>brev., bar..., ub</td>
</tr>
<tr>
<td>p,p</td>
<td>praev., prec...</td>
</tr>
<tr>
<td>c</td>
<td>cum, con, cen...</td>
</tr>
<tr>
<td>pp,pp</td>
<td>propt., papa...</td>
</tr>
<tr>
<td>s</td>
<td>condam (quondam)</td>
</tr>
<tr>
<td>q</td>
<td>qui...</td>
</tr>
<tr>
<td>t</td>
<td>de..., der..., ud</td>
</tr>
<tr>
<td>q</td>
<td>quae</td>
</tr>
<tr>
<td>h</td>
<td>haec, hoc, her...</td>
</tr>
<tr>
<td>t.t</td>
<td>vel, ul..., el</td>
</tr>
<tr>
<td>m</td>
<td>men..., mun...</td>
</tr>
<tr>
<td>n</td>
<td>non, nun...</td>
</tr>
<tr>
<td>θ</td>
<td>obit</td>
</tr>
<tr>
<td>p</td>
<td>per, par..., por...</td>
</tr>
</tbody>
</table>

https://kuscholarworks.ku.edu/bitstream/handle/1808/1821/47cappelli.pdf
Actually it’s not a new thing at all


Marcus Agrippa, son of Lucius, made [this building] when consul for the third time
Text and Writing Variations
Query Expansion

• Search for a word
• Repeat the search for each and every word found
• Old thing… but we do not have a feature for each word to use in order to improve there result
• Use the Visualisation tool: Beatrice
‘Human in the loop’
Results
Why?

• Use the tool to find a cut between inliers and outliers
• Examine further:
  • the worst ‘correct’ and
  • the best ‘incorrect’
  • in doubt: look at the sentence where the word occurs
• Let the system LEARN from this
• The advantage is that it is MUCH easier than looking through the 50 pages to check…
• Non found words will be “left over” after complete transcription
Conclusion

- WS can be used to make non transcribed collections searchable
- It can also be an efficient tool for making transcriptions easier
- Use visualisations to **solve** the problem:
  - correct vs. incorrect words
- Still many interesting challenges!