READ

Recognition and Enrichment of Archival Documents

How To recognise and search documents automatically - Keyword Spotting

Version v1.9.1

Last update of this guide: 28/11/2019

This guide explains how to use the Keyword Spotting function of Transkibus. Keyword Spotting is a new and powerful searching tool that will help you search for distinct words in your document collection.

Download the Transkribus Expert Client, or make sure you are using the latest version:

- https://transkribus.eu/

Consult the Transkribus Wiki for further information and other How to Guides:

- https://transkribus.eu/wiki/

Transkribus and the technology behind it are made available via the following projects and sites:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact:

- The Transkribus Team: email@transkribus.eu

Contents

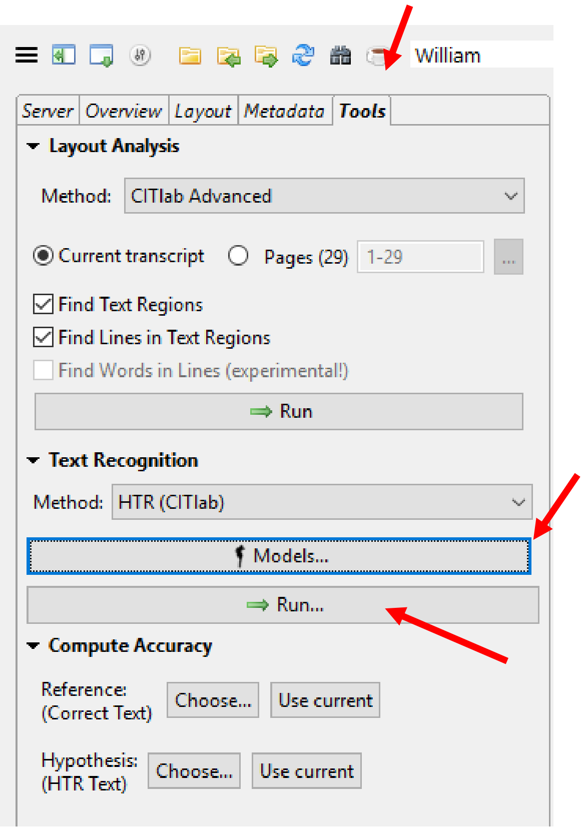
The READ project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 674943.

Introduction

- Keyword Spotting (KWS) with Transkribus makes it possible for you to **search for distinct words in your documents**.
- **The main advantage: there is no need for you to transcribe your documents before you search them.** Simply run a Handwritten Text Recognition (HTR) model to produce a transcript and then search your documents immediately. Even if the automatically generated transcript contains errors, KWS will reliably find words, phrases and even parts of words and regular expressions in your documents.
- The program will show **you on which pages your keyword has been found**. Moreover, it will give you a figure between 0 and 1 to rate the confidence of the results.
- Note: KWS is an intensive computing task. It is implemented as a "job" in Transkribus: you can start the search and perform other tasks while you wait for the results. All results are stored in Transkribus and can be (re-)opened and examined at any time. We are convinced that working with search results from KWS will become a standard task for many historians and philologists in the future.

Preparation – recognising text

- Before you commence KWS, you need to apply a HTR model to your documents and produce a first transcription.
- First, **upload** your documents to Transkribus.
- Second, **segment** your documents into text regions, lines and baselines.
- For more information on **uploading** and **segmentation**, please consult How To Transcribe Documents with Transkribus – Introduction.
- Next, you need to run HTR on your documents.
- To access your model, click on the "Tools" tab and go to the "Text Recognition" section.
- Click "Run", then click "Configure". Choose your HTR model from the list on the left-hand side of the screen and click OK.
- Note: KWS works best when you use a specific HTR model that has been trained on your documents. But useable results can also be generated with general models. If you do not have your own HTR model, you can experiment with public models available in Transkribus. Transkribus currently provides access to public models trained on English and German writing from the eighteenth and nineteenth centuries. More public models will be made available soon.
- Press "Run" to start the text recognition process.



Method: CITlab Advanced

● Current transcript ○ Pages (29) 1-29

☑ Find Text Regions

☑ Find Lines in Text Regions

□ Find Words in Lines (experimental!)

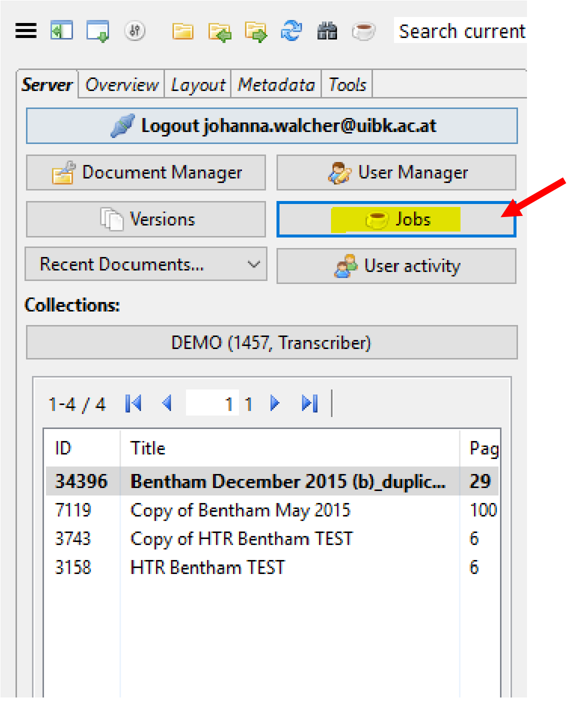
Method: HTR (CITlab)

Reference: (Correct Text)

Hypothesis: (HTR Text)

Figure 1 Run model

- You can check on the progress of the recognition by clicking the “Jobs” button in the “Server” tab

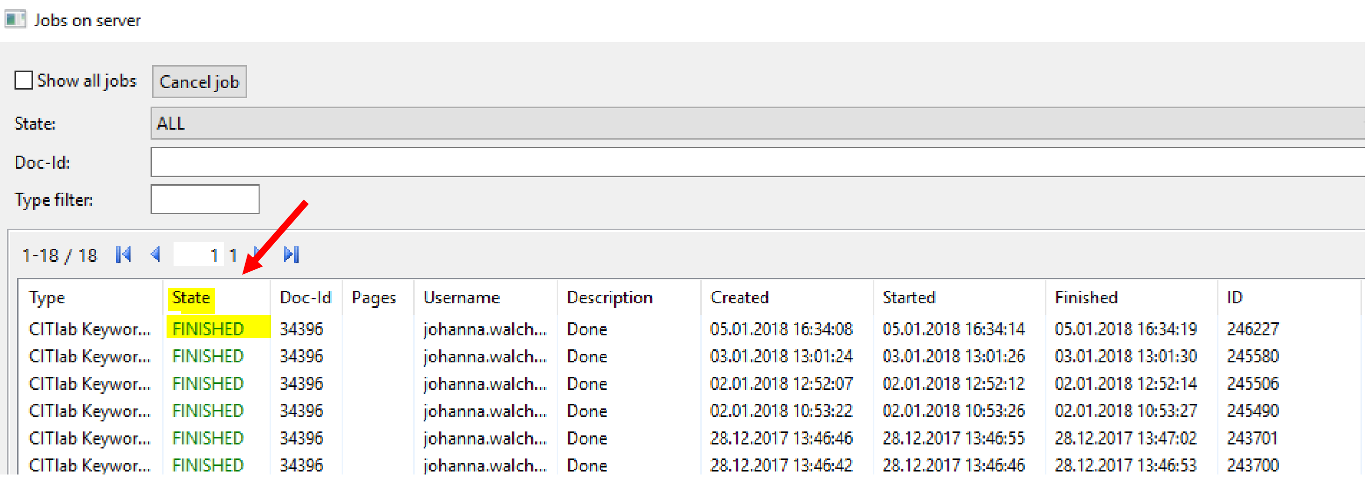


Collections:

DEMO (1457, Transcriber)

ID	Title	Pag
34396	Bentham December 2015 (b)_duplic...	29
7119	Copy of Bentham May 2015	100
3743	Copy of HTR Bentham TEST	6
3158	HTR Bentham TEST	6

Figure 2 Click the “Jobs” button to check progress of HTR.



Jobs on server

☐ Show all jobs Cancel job

State: ALL

Doc-Id:

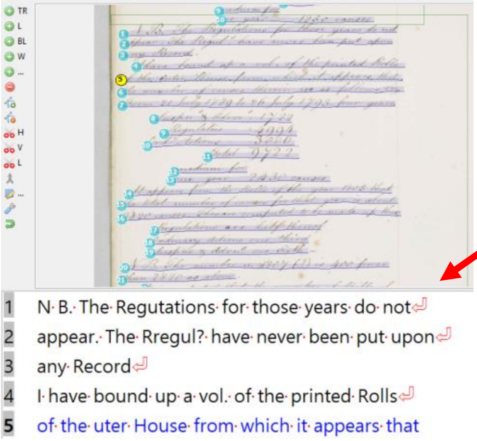
Type filter:

1-18 / 18 1 1

Type	State	Doc-Id	Pages	Username	Description	Created	Started	Finished	ID
CITlab Keywor...	FINISHED	34396		johanna.walch...	Done	05.01.2018 16:34:08	05.01.2018 16:34:14	05.01.2018 16:34:19	246227
CITlab Keywor...	FINISHED	34396		johanna.walch...	Done	03.01.2018 13:01:24	03.01.2018 13:01:26	03.01.2018 13:01:30	245580
CITlab Keywor...	FINISHED	34396		johanna.walch...	Done	02.01.2018 12:52:07	02.01.2018 12:52:12	02.01.2018 12:52:14	245506
CITlab Keywor...	FINISHED	34396		johanna.walch...	Done	02.01.2018 10:53:22	02.01.2018 10:53:26	02.01.2018 10:53:27	245490
CITlab Keywor...	FINISHED	34396		johanna.walch...	Done	28.12.2017 13:46:46	28.12.2017 13:46:55	28.12.2017 13:47:02	243701
CITlab Keywor...	FINISHED	34396		johanna.walch...	Done	28.12.2017 13:46:42	28.12.2017 13:46:46	28.12.2017 13:46:53	243700

Figure 3 "Jobs on server" window

- Once the recognition is finished, the automated transcription will appear in the text editor field.



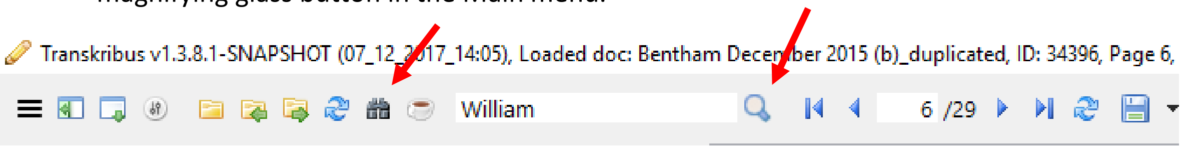
1 N B. The Regutations for those years do not
2 appear. The Rregul? have never been put upon
3 any Record
4 I have bound up a vol. of the printed Rolls
5 of the uter House from which it appears that

Figure 4 Automated transcription appears in the text editor field.

Keyword Spotting function

Where to find it

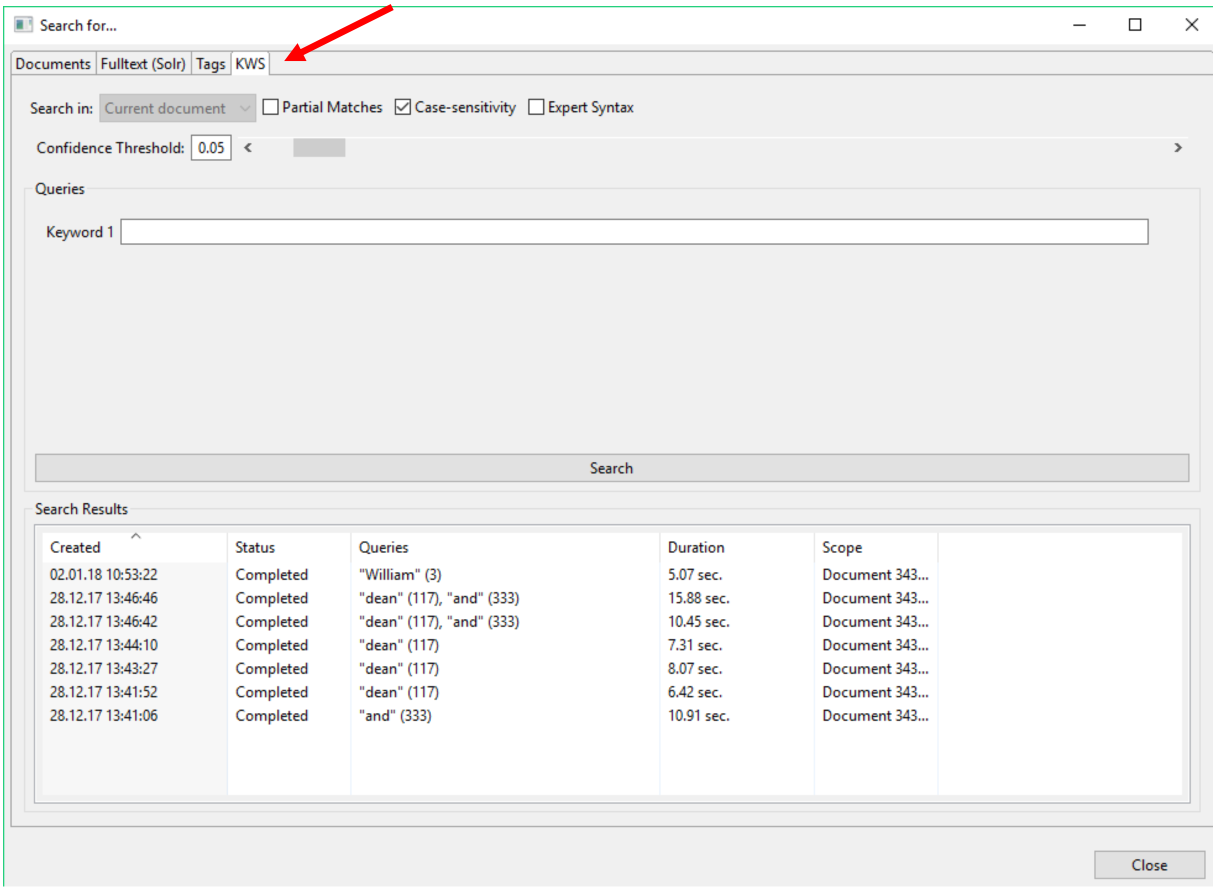
- You can open the Keyword Spotting feature by clicking the binoculars button or the magnifying glass button in the Main menu.



Transkribus v1.3.8.1-SNAPSHOT (07_12_2017_14:05), Loaded doc: Bentham December 2015 (b)_duplicated, ID: 34396, Page 6,

Figure 5 How to open Keyword Spotting

- In the search window, click the “KWS” tab



Search in: Current document □ Partial Matches ☑ Case-sensitivity □ Expert Syntax

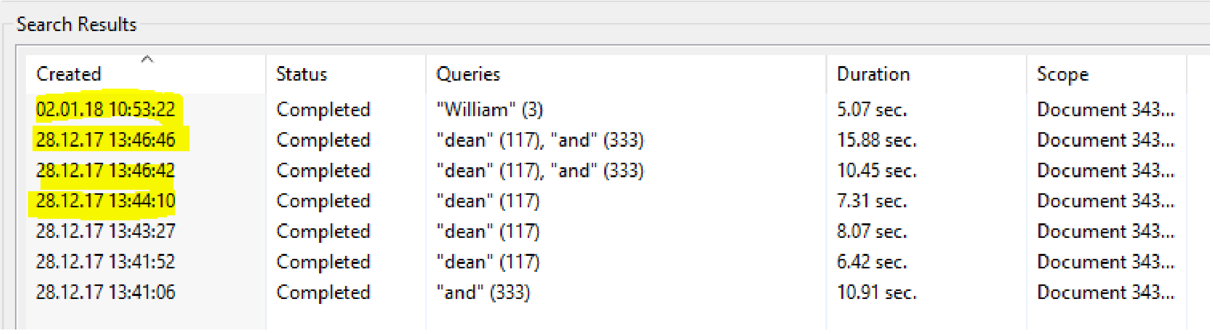
Confidence Threshold: 0.05

Created	Status	Queries	Duration	Scope
02.01.18 10:53:22	Completed	"William" (3)	5.07 sec.	Document 343...
28.12.17 13:46:46	Completed	"dean" (117), "and" (333)	15.88 sec.	Document 343...
28.12.17 13:46:42	Completed	"dean" (117), "and" (333)	10.45 sec.	Document 343...
28.12.17 13:44:10	Completed	"dean" (117)	7.31 sec.	Document 343...
28.12.17 13:43:27	Completed	"dean" (117)	8.07 sec.	Document 343...
28.12.17 13:41:52	Completed	"dean" (117)	6.42 sec.	Document 343...
28.12.17 13:41:06	Completed	"and" (333)	10.91 sec.	Document 343...

Figure 6 “KWS” tab

Latest and previous search results

- Your previous and current Keyword Spotting queries will appear at the bottom of the “KWS” tab.

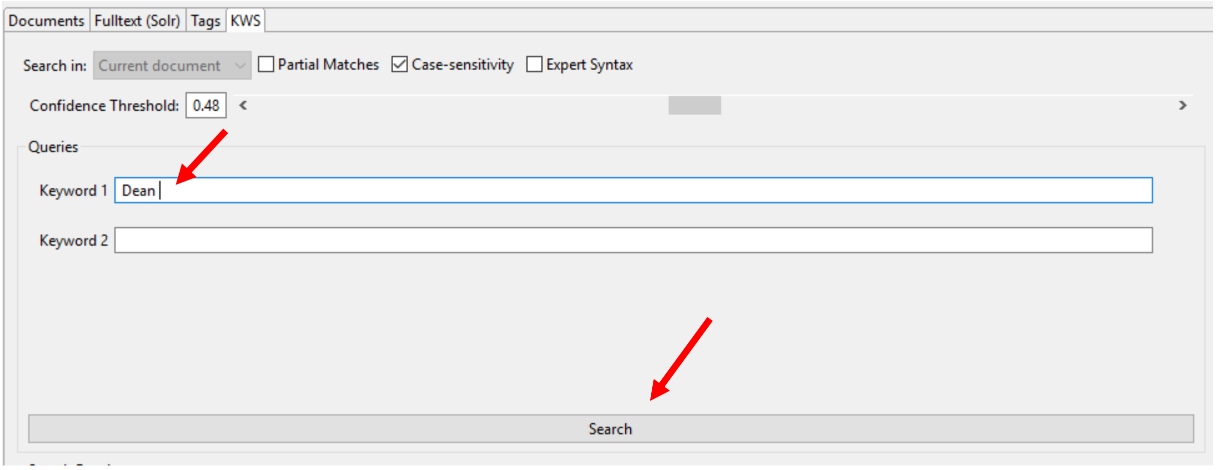


Created	Status	Queries	Duration	Scope
02.01.18 10:53:22	Completed	"William" (3)	5.07 sec.	Document 343...
28.12.17 13:46:46	Completed	"dean" (117), "and" (333)	15.88 sec.	Document 343...
28.12.17 13:46:42	Completed	"dean" (117), "and" (333)	10.45 sec.	Document 343...
28.12.17 13:44:10	Completed	"dean" (117)	7.31 sec.	Document 343...
28.12.17 13:43:27	Completed	"dean" (117)	8.07 sec.	Document 343...
28.12.17 13:41:52	Completed	"dean" (117)	6.42 sec.	Document 343...
28.12.17 13:41:06	Completed	"and" (333)	10.91 sec.	Document 343...

Figure 7 Current and past search results

Using Keyword Spotting

- To use the Keyword Spotting function simply type the word you would like to search for in the “Keyword 1” box and press the “Search” button.



Search in: Current document □ Partial Matches ☑ Case-sensitivity □ Expert Syntax

Confidence Threshold: 0.48

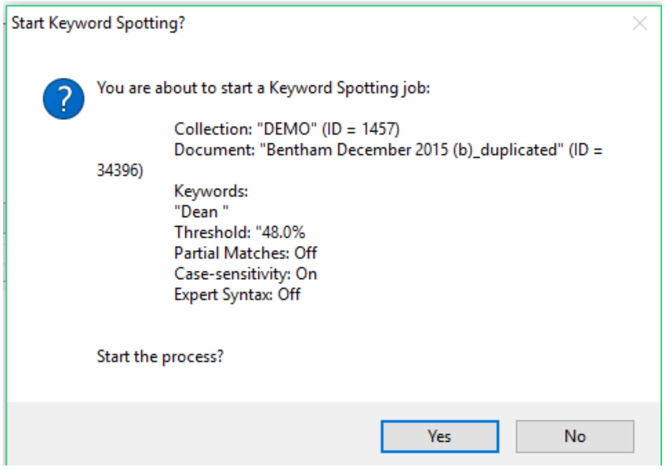
Queries

Keyword 1 Dean

Keyword 2

Figure 8 Searching for a keyword

- A confirmation window will pop-up. Click “Yes” to start your Keyword Spotting query



Start Keyword Spotting?

You are about to start a Keyword Spotting job:

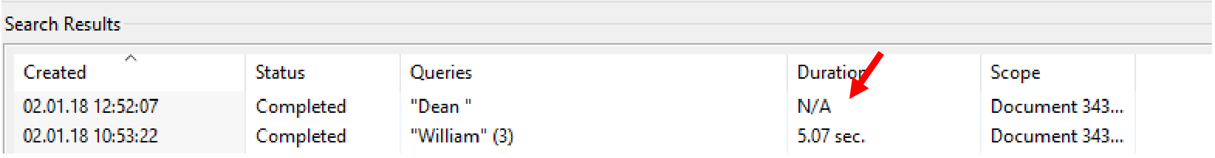
Collection: "DEMO" (ID = 1457)
Document: "Bentham December 2015 (b)_duplicated" (ID = 34396)

Keywords:
"Dean "
Threshold: "48.0%
Partial Matches: Off
Case-sensitivity: On
Expert Syntax: Off

Start the process?

Figure 9 Confirmation window

- Keyword Spotting queries take at least a few seconds to complete.
- When “N/A” is shown in the “Duration” column of the “KWS” tab, this means that the program is still searching.
- Once the process is finished, the “Duration” value will change into the amount of time that the query took to complete.

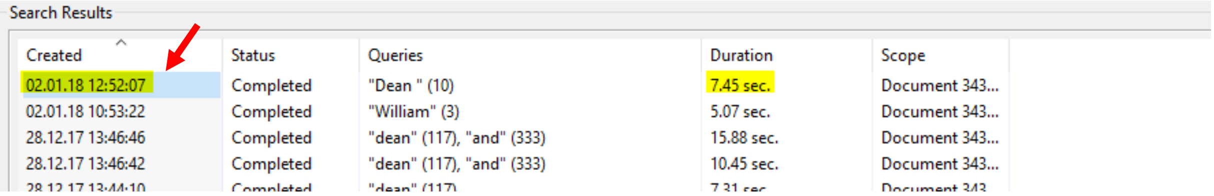


Search Results

Created	Status	Queries	Duration	Scope
02.01.18 12:52:07	Completed	"Dean "	N/A	Document 343...
02.01.18 10:53:22	Completed	"William" (3)	5.07 sec.	Document 343...

Figure 10 Keyword Spotting in progress

- Double-click the date and numerical value in the “Created” column to access your search results

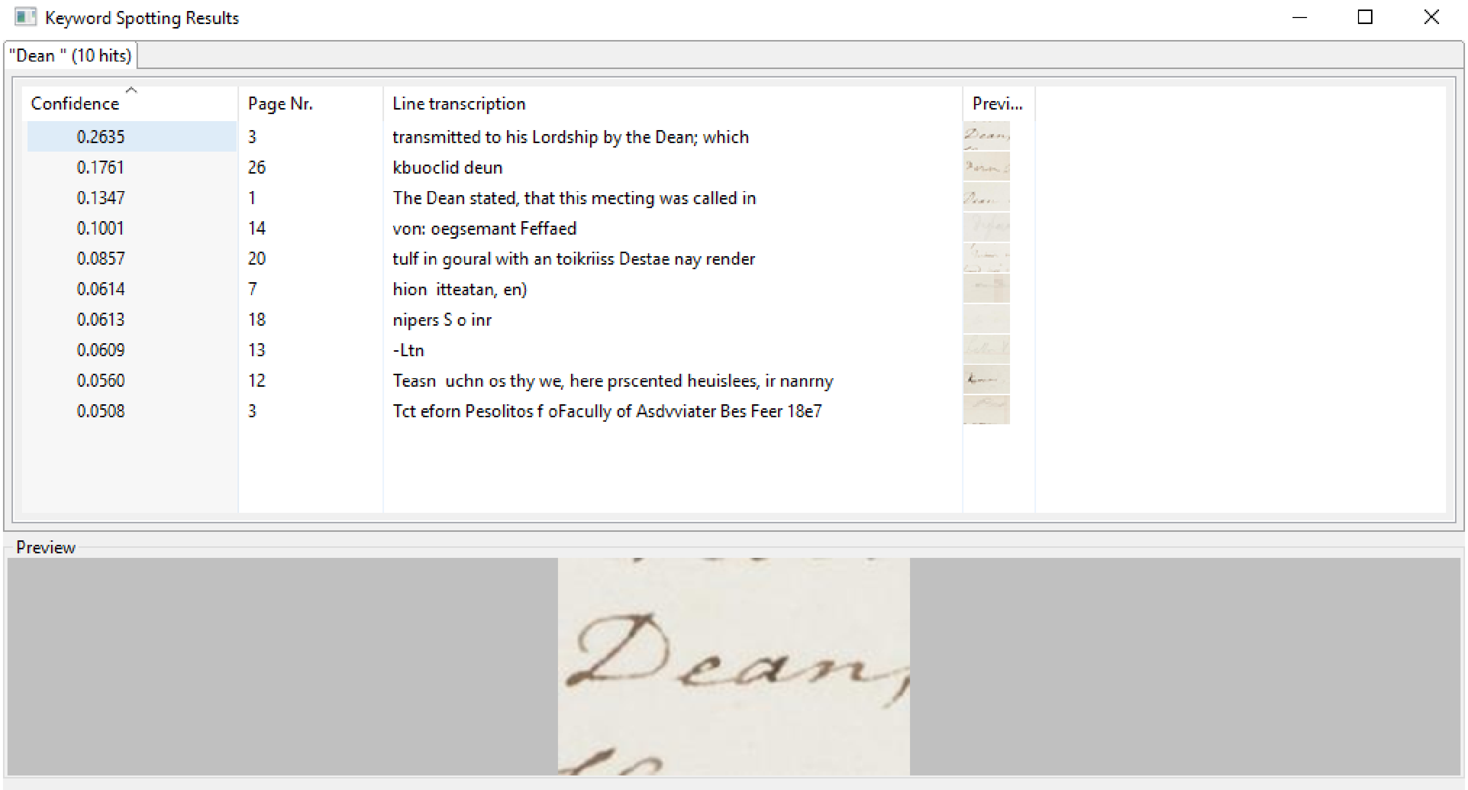


Search Results

Created	Status	Queries	Duration	Scope
02.01.18 12:52:07	Completed	"Dean " (10)	7.45 sec.	Document 343...
02.01.18 10:53:22	Completed	"William" (3)	5.07 sec.	Document 343...
28.12.17 13:46:46	Completed	"dean" (117), "and" (333)	15.88 sec.	Document 343...
28.12.17 13:46:42	Completed	"dean" (117), "and" (333)	10.45 sec.	Document 343...
28.12.17 13:44:10	Completed	"dean" (117)	7.31 sec.	Document 343

Figure 11 Keyword Spotting results

- The "Keyword Spotting Results" window will show you a list of places where that keyword appears, with the following information:
  - The confidence level of the individual results (between 0 and 1).
  - The number of the page of your document where the word had been found.
  - The automated transcription in which the word is embedded.
  - An extract of the page image. When you hover the cursor over this image a bigger preview appears at the bottom of the window.
  - Double-click on the image in the "Preview" column to go straight to the page where your keyword appears.



Keyword Spotting Results

"Dean " (10 hits)

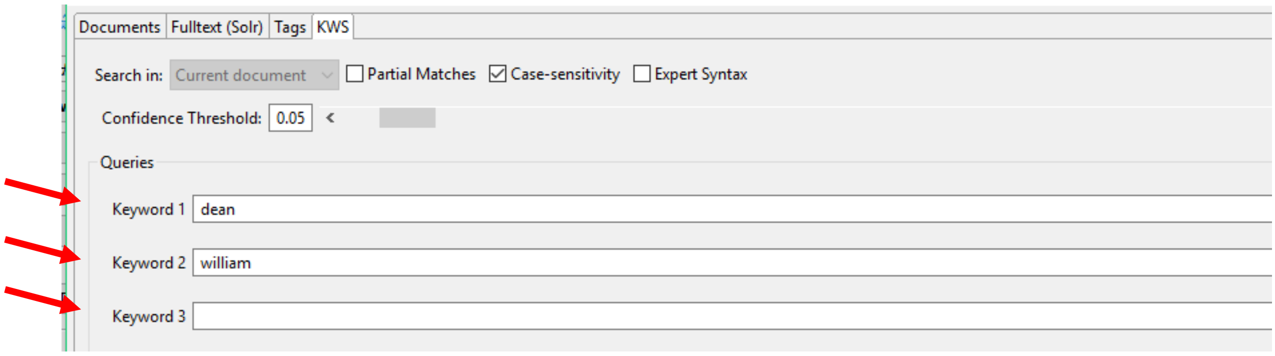
Confidence	Page Nr.	Line transcription	Previ...
0.2635	3	transmitted to his Lordship by the Dean; which	
0.1761	26	kbuoclid deun	
0.1347	1	The Dean stated, that this mecting was called in	
0.1001	14	von: oegsemant Feffaed	
0.0857	20	tulf in goural with an toikriiss Destae nay render	
0.0614	7	hion itteatan, en)	
0.0613	18	nipers S o inr	
0.0609	13	-Ltn	
0.0560	12	Teasn uchn os thy we, here prscented heuislees, ir nanrny	
0.0508	3	Tct eforn Pesolitos f oFacully of Asdvviater Bes Feer 18e7	

Preview

Figure 12 Information about your Keyword Spotting results

Searching for two keywords at the same time

- It is also possible to search for two keywords at the same time. Simply add extra keywords into the corresponding spaces in the "KWS" tab.



Search in: Current document □ Partial Matches ☑ Case-sensitivity □ Expert Syntax

Confidence Threshold: 0.05

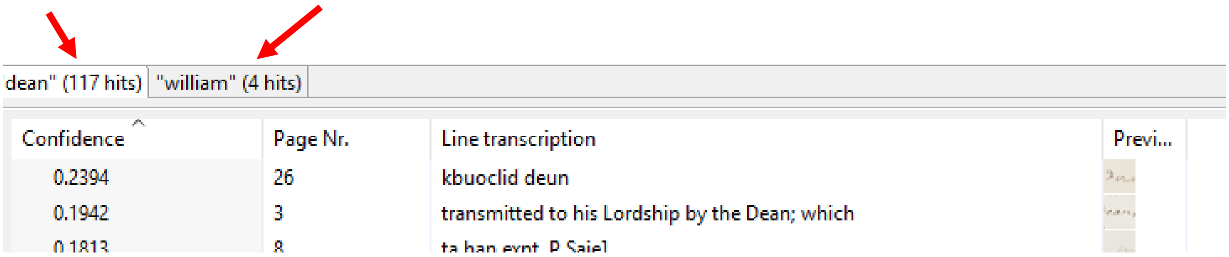
Keyword 1 dean

Keyword 2 william

Keyword 3

Figure 13 Searching for multiple keywords at the same time

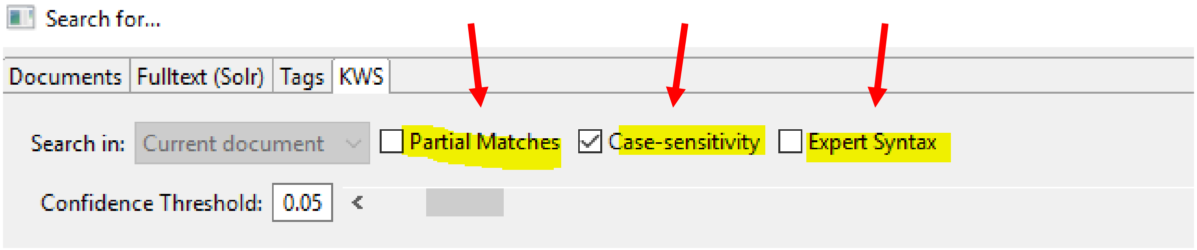
- The results for each keyword will be displayed in separate tabs.



Confidence	Page Nr.	Line transcription	Previ...
0.2394	26	kbuoclid deun	
0.1942	3	transmitted to his Lordship by the Dean; which	
0.1813	8	ta han exnt P Saiel	

Figure 14 Search results for multiple keywords

More searching options



Search in: Current document □ Partial Matches ☑ Case-sensitivity □ Expert Syntax

Confidence Threshold: 0.05

Figure 15 Searching options

Partial Matches

- If you choose this option, the program will search for all words which contain the text you entered in the search field. E.g. if you search for “**ity**”, the program will return results like “conventiona**lity**”, “proportiona**lity**”, “indefensibi**lity**” etc.

Case Sensitivity

- With this function the system will take upper and lower case into account. E.g. if you search for “Kingdom” results where this word is written with a capital “K” will have a higher confidence level.

Expert Syntax

- Instead of searching for words, you can search for regular expressions.
- Some examples of searches for regular expressions would be:
  - **date:** .*(?<KW>[0-3][0-9]\.[0-1][0-9]\.[0-9]{4}).* matches any line containing a date of the form TT.MM.JJJJ
  - **abbreviations:** .*(?<KW>Dr\.|Doctor).* matches any line containing Doctor and its abbreviation Dr.

  - **uncertainties:** .*(?<KW>(k|c|che|chh)rist?).* matches any line containing Old High German spellings for Christ: e.g. kris, krist, crist, cherist, chhrist

- In contrast to standard usage of regular expressions, the search patterns have to match the whole line, e.g. .*[0-9]{4,6} will match only lines which end with a number of at least 4 digits. To allow arbitrary characters after the 4 digits, one has to add .* at the end: .*[0-9]{4,6}.* Similarly, [0-9]{4,6}.* matches only lines which begin with 4 digits.

- Standard regular expression features which are supported in KWS in Transkribus:

. any character

+ one or more repetitions of the previous literal

* zero or more repetitions of the previous literal

[] class of characters, e.g. [0-9] any digit between 0 and 9; [aeiou] any vowel; [A-Z] any capital letter

? the previous literal is optional

{X} repeat previous literal X times

{X,Y} repeat previous literal between X and Y times

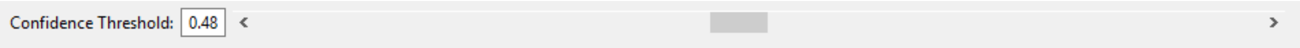
| or operation, e.g. a|b means either a or b

() parentheses are used to group the regular expression: (a|b)c matches ac or bc while a|bc matches a or bc

\ escape operator: to match e.g. a + or . one needs to escape it by \+ or \.

Confidence Threshold

- It is also possible to adjust the confidence threshold of your Keyword Spotting query. **Note:** this function is not yet activated but will be available soon.
- The confidence threshold is a number between 0 and 1.
- If the confidence threshold is 0.5 or above, this means that the program will be very confident in finding keywords which match your search query.
- If the confidence threshold is 0.1 or below, this means that the program will uncover more possible matches for your keyword but it will be less sure about these matches. Many “false alarms” will appear in your search results and it is up to you to check their accuracy.



Confidence Threshold: 0.48

Figure 16 Confidence Threshold

Outlook

The Transkribus Team is currently working to update the KWS tool. A future version will allow users to validate the results of their searches and export search results as tabular data.

Credits

We would like to thank the many users who have contributed their feedback to help improve the Transkribus software.

Transkribus is made available to the public as part of H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission.