BiLand
End-user Manual
web-application for bilingual historical analysis

Fons Laan
Informatics Institute
University of Amsterdam
Science Park 904
1098 XH Amsterdam

version 1·1

6 Nov 2013

Contents

1 Introduction

In this document we will describe how to use the web application of the Clarin BiLand project. With your browser[1] you can find the application at:
`http://fietstas2.science.uva.nl:8008/biland/`. [2]

BiLand is a research tools for historians that uses the newspaper data of the Koninklijke Bibliotheek as input material. One can search with single query terms or with combinations thereof. Apart from showing the articles that match the query, the results can be visualized by word clouds of single articles together with sentiment words highlighted, or by a word cloud of the whole query result set.

Additional information about the project can be obtained from the Biland cms site `http://biland.nl`. Biland is the successor of wahsp and both are described in the cms.

Apart from a multitude of smaller differences, there are several main differences between wahsp and Biland:

- Biland uses two corpora: in addition to the KB newspapers, Biland uses a small corpus of old German newspapers from the Staatsbibliothek zu Berlin.
- All article data (ocr plus metadata) is stored in a local ElasticSearch instance, which is a Lucene-based search engine plus datastore.

2 User interface

In this section we will give an overview of the components of the user interface. After accessing the Biland url you will see the login window, see fig. 1. Type your username and password

[1] Internet Explorer may not work with BiLand. Please use Google Chrome, a recent Firefox, Opera, ...
[2] The url will become `http://texcavator.nl/`.

and press Enter or click the Login button.

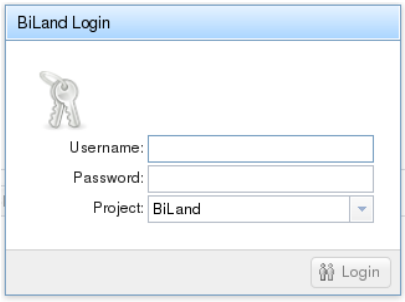


Figure 1: Login window.

The BiLand opening window is shown in fig. 2.

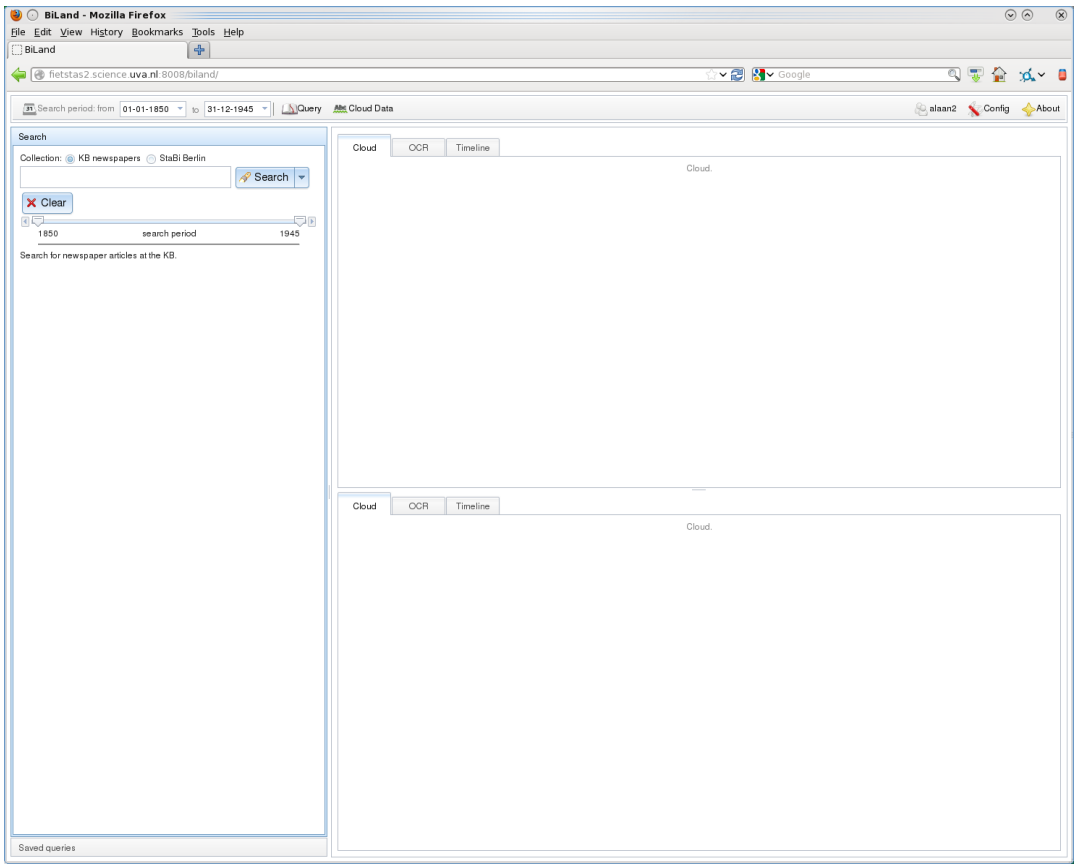


Figure 2: BiLand opening screen.

The window consists of the following screen regions:

- The toolbar at the top
- An accordion widget at the left
- A panel with tabs at the top-right for the Dutch KB data
- A panel with tabs at the bottom-right for the German StaBi data

The *toolbar* consists of the following widgets:

- Two date widgets to limit the search period. The BiLand date range for the KB newspapers has been set to 1850–1945. The available date range for the German data is much smaller.
- A query widget, used to *i*) combine saved queries into a new query, *ii*) modify a given query, and *iii*) export the OCR plus metadata of those articles that correspond to the query.
- A cloud data widget, to view the cloud words and their frequencies, which can be exported as a csv-file.
- A logout widget.
- A configuration widget, to select an increasing number of options.
- An about widget, showing the collaborators of the project, and a link to this document.

The query *accordion* on the left has the two divisions:

- Search
  Here one selects the corpus, and creates queries to be sent to the search engine.
- Saved queries
  This shows the list of your saved queries, which are used to retrieve the OCR data of the articles, create word clouds, timelines.

The screen area to the right of the accordion is for displaying the OCR, clouds and timelines, and will be discussed together with searching.

3 Searching

A trivial way to search is by using a single query word. Say, we type eugenetiek in the textline area in the accordion, and then click the Seach button; see fig. 3a. It shows that 410 articles are found. The first chunk is displayed with their titles in bold and underlined. Underneath the title is some additional information: the newspaper title, article date, and newspaper 'type' (country-wide, or regional). The score number in brackets indicates how well the article matches the query. Clicking next gives the next chunk of articles. Clicking Clear clears the search area.

Underneath the Seach and Clear buttons is a date range slider to set the search period in years. It operates in conjunction with the two date widgets on the toolbar, which can be used to adjust the range regarding month and day.

At the top of the search accordion one can select the collection to be searched, either the big KB collection (dutch), or the small StaBi one (german).

When you click one of the article titles, its OCR text is shown in the Text tab, see fig. 4. Clicking the Original tab shows the scan image of the newspaper article (fig. 5).

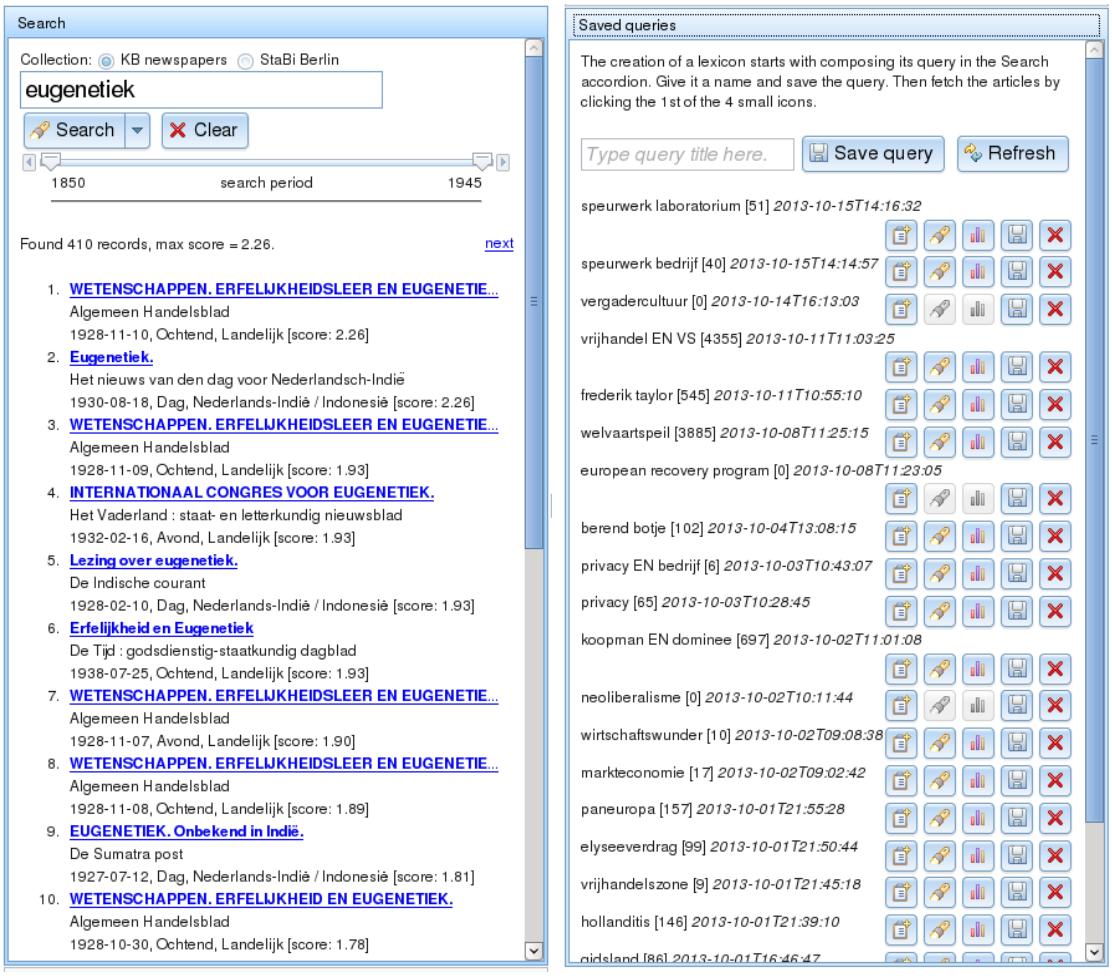
The tab View at KB opens the KB search engine page in a new browser window (or tab).

The corresponding word cloud of the article is shown in fig. 6. The used font size of the words is the graphical equivalent of their frequency in the document. Words of too low frequency may not be shown, and in general 'noise' (i.e. stopwords) is also suppressed.

When you have chosen the SVG (Scalable Vector Graphics) cloud option in the config menu, you can see the word frequencies at the top of the cloud when you hover the words with your mouse. And when you click on a word, you can add that word to your personal stopword list, to be removed the next time you generate a cloud.

3.1 Query editor

Creating queries that consist of more than a single word is done with the built-in query editor. The editor is easiest to explain by creating an example query. Let say that we create a new query that we will later save with the name 'luminal'. Proceed with the following steps:



(a) Search panel (b) Saved queries

Figure 3: Search and Saved queries in the accordion.

Figure 4: OCR text of a KB article in the OCR tab.

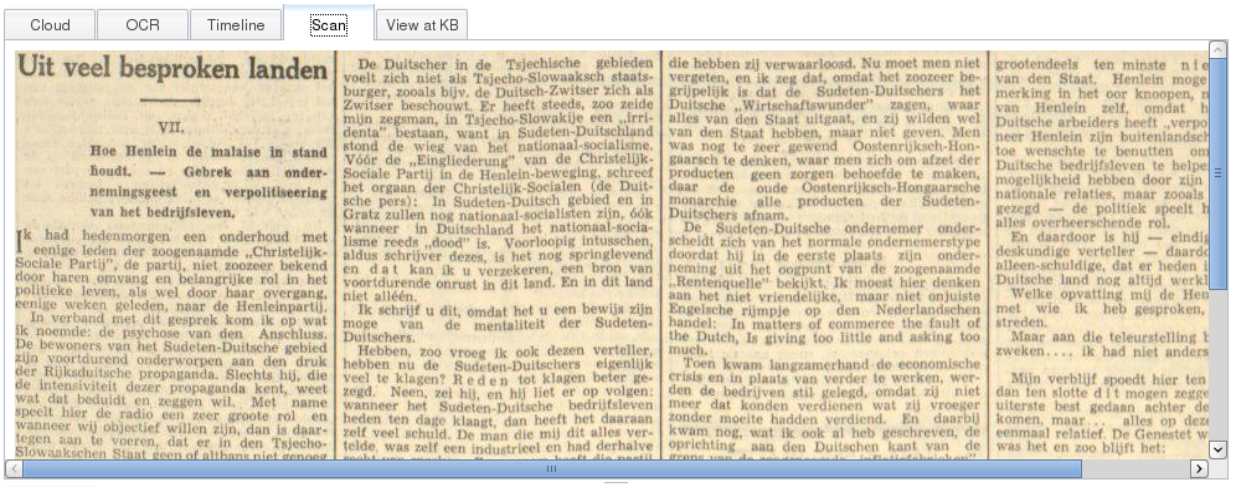


Figure 5: Scan image of a KB article in the Scan tab.

Figure 6: Word cloud of a single KB article.

- In the Search panel of the accordion, type luminal as search term.
- Click on the tiny arrow on the right half of the Search button.
- Click on the button Start search that appeared underneath the Search button.
- Below the text widget (that now contains (cql.serverChoice exact "luminal")) there is a new button with text luminal. Click on its arrow at the right side.
- You will see a new frame with several buttons and other widgets. Click on the button Create word list.
- Next to Word list: luminal there is a tiny icon of the inline editbox, click on it.
- Type chloral in the text region (see fig. 7) en then press Enter. Next to luminal we now also see chloral in the word list.
- Once more press the icon.
- Type wekaminen and press Enter.
- Click the Search button, which shows the found records.
- Then go to Saved queries in the accordion and at Type query title here. type luminal, and click the Save query button. Then luminal is displayed as the new saved query (unless that name is already taken).
- Click its first icon (with hover text Create basis lexicon: luminal). This loads the OCR data of all the luminal articles from the KB.
- When the loading is done, click the second icon Apply query: luminal. That creates the cloud of the luminal articles, plus some statistics.

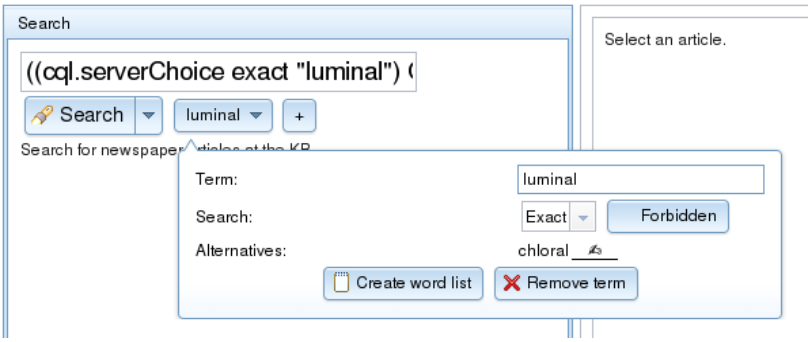


Figure 7: Query editor.

The Saved queries panel (fig. 3(b)) shows the query titles, their article count, and the creation date of the queries that you have saved. If the article count is zero your query did not yield a single hit (using the specified date range). Clicking the Refresh button updates the article counts of the queries. In doing so it uses the corpus that has been selected by the radio buttons.

To the right of each query are five small icons. When you move your mouse over them, you will see their hover text:

- Re-Search
- Apply query
- Timeline
- Modify
- Delete

When you click the third icon, a 'timeline' is generated: a histogram of article frequencies over time, see fig. 8. High-frequency 'bursts' are flagged in red.

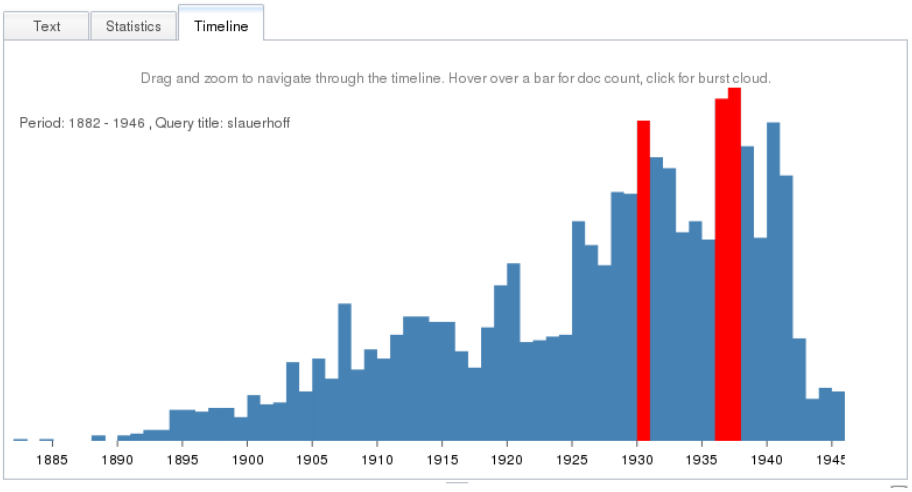


Figure 8: Histogram of article frequencies over time.

3.2 Queries

3.2.1 Query combiner

With the query widget (see fig. 9, reachable from the toolbar) one can combine two existing (i.e. saved) queries into a new query. First select the desired boolean combination operator (AND, OR or NOT), and then select the first and second query from the available list. The widget will suggest a name for the combined query, but you can change that before clicking OK.

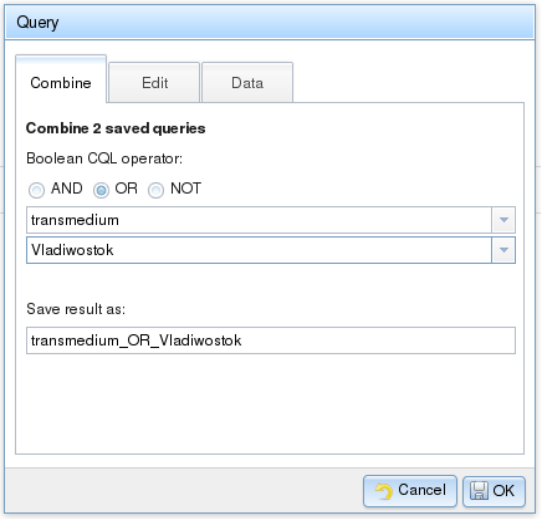


Figure 9: Combining existing queries.

3.2.2 Query editor

The second tab of the query widget shows the beginning of a new query editor, see fig. 10. This never got finished. You can edit a given query, but as there is no validation of your edits the resulting query may be syntactically wrong.

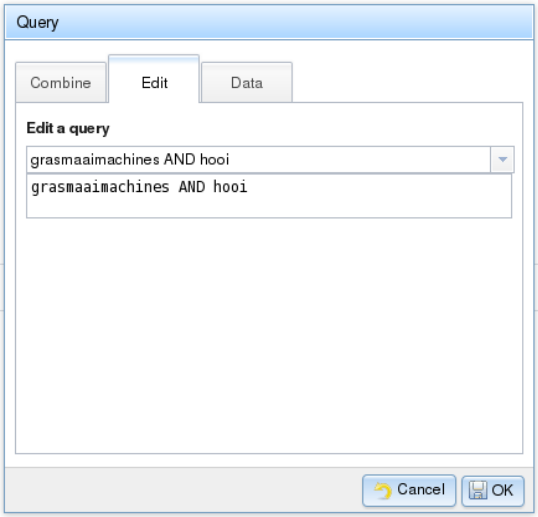


Figure 10: Editing a query.

3.2.3 Query data export

The third tab of the query widget (fig. 11) shows the export option of query data.

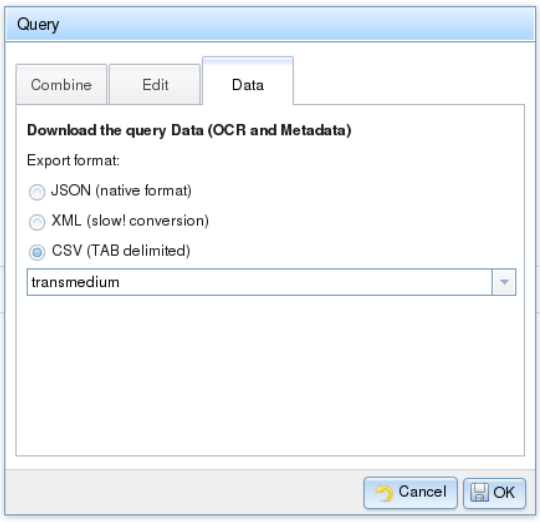


Figure 11: Editing a query.

4 Word cloud configuration

The word cloud in fig. 6 was made with default cloud parameters, but there are several options to tune the result according to your wishes. Fig. 12 shows the word cloud options. This configuration widget can be opened from the toolbar.

The word cloud options have the following effect:

- Require fresh cloud. This adds a dummy variable with random value to the cloud request. This should convince your browser not to return a cached result.
- Reduce font size differences. When the word sizes decline too fast at the cloud edge, this option should improve the result.

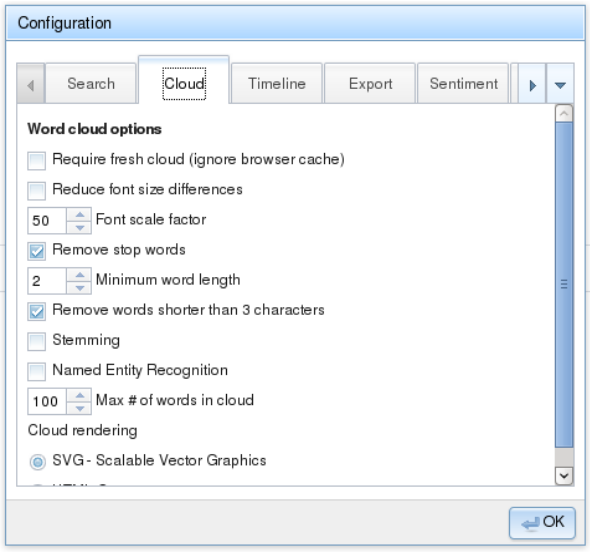


Figure 12: Word cloud configuration.

- Font scale factor. This scale factor determines the maximum font size.
- Remove stop words. This removes short words, as specified by a pre-defined list, plus the user-defined stopwords from the cloud.
- Remove words shorter that 3 characters. When the stop word list does not block enough noise, this will filter more.
- Stemming. This applies stemming to the words before computing the cloud. A consequence of stemming is that all words become lower-case. Currently stemming is applied before NER, which makes NER helpless because upper-case letters are crucial. In the course of 2013 the xTAS pipeline will change, circumventing this problem.
- Named-Entity Recognition. This applies NER, currently a bit slow.
- Max. # of words in cloud. The number of words returned by the server can be very big. Truncating the list before generating the cloud speeds it up.
- Cloud rendering. Choose between the original HTML canvas cloud, or the newer interactive SVG cloud.

Fig. 13 shows the word cloud of the query wekaminen, which yielded (only) 11 articles. Often, as in this case, the cloud does not properly occupy the available space. One can increase the maximum number of words displayed to remedy this, assuming more words are indeed available. But when the words at the border of the cloud are already small, that does not help much, because words that are too small become invisible anyway. Then it is better to reduce the font size differences, see fig. 14 for the result.

Finally, fig. 15 shows the same word cloud with Named-Entity Recognition. Used colors: locations, persons, organizations, and miscellaneous. The latter means that the NER algorithm 'thinks' these are entities, but cannot be more specific about it. The NER we used is Stanford, trained for Dutch. It is not perfect, but it is better than several alternatives. Notice that the figure only shows the recognized entities, the remaining words are left out.

Figure 13: Word cloud of the query wekaminen.

Figure 14: Word cloud of the query wekaminen with reduced font size differences.

Figure 15: Word cloud of the query wekaminen with *Stanford* Named-Entity Recognition. Used colors: locations, persons, organizations and miscellaneous.

5 Automatic translation

Automatic translation of the German OCR is done with the free version of Google Translate. With Google Translate active, the GUI gets two additional widgets, see figs. 16 and 17.

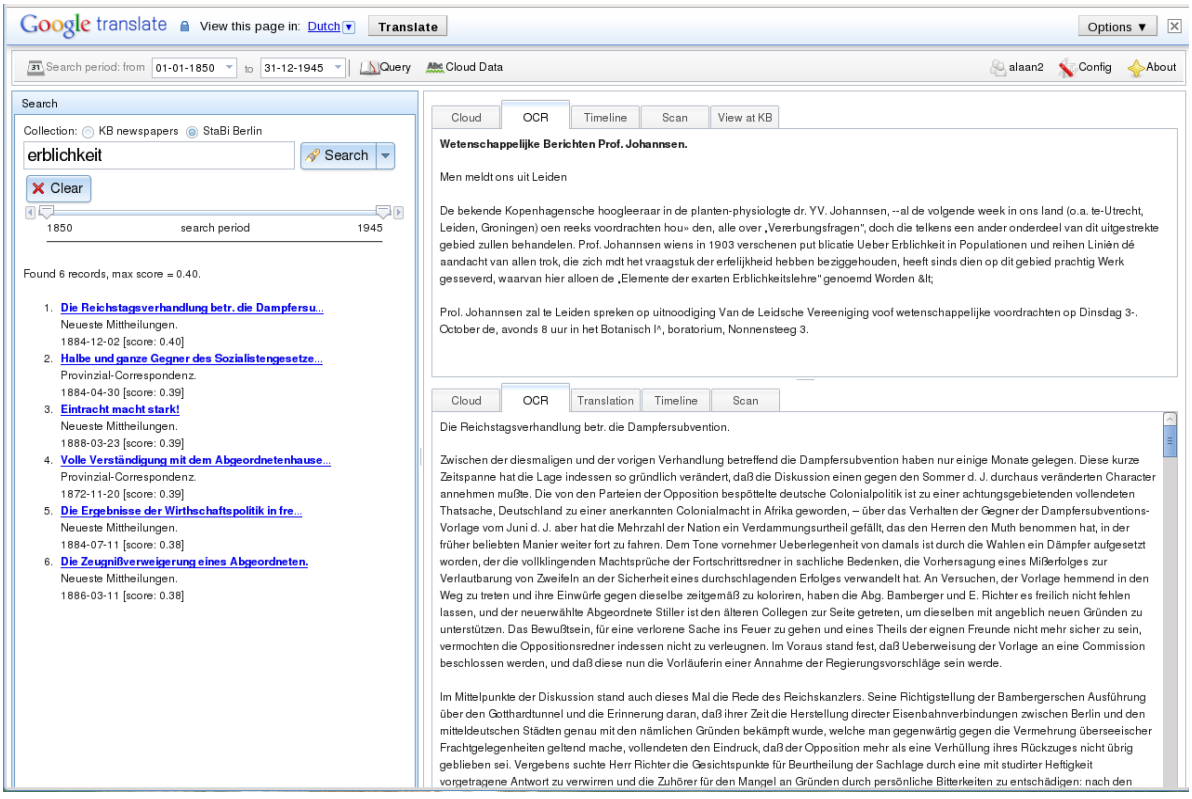


Figure 16: Dutch and German text in their OCR panes (of two different articles), with a Google translate toolbar at the top.

The first Google widget is an extra toolbar at the top. When you do not need it anymore, you can remove the toolbar by pressing its ‘close’ icon on the far right. The second Google

witdget is an second language selection widget (there is also one on the toolbar). We put that widget in a *Translate* tab next to the German OCR tab. As soon as one selects a destination language, the whole GUI text is translated into that language. (And the toolbar re-appears if it was closed.)

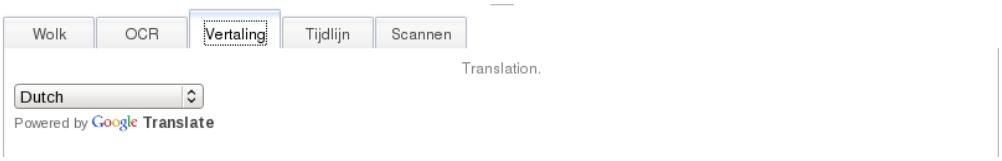


Figure 17: Google translate language selection. Everything is translated, including our English widget texts.

The result of the translation can be seen in fig. 18. The toolbar has a Show original button to go back to the original text. With the current free version of the toolbar it is no longer possible to only translate a certain part of the GUI text.

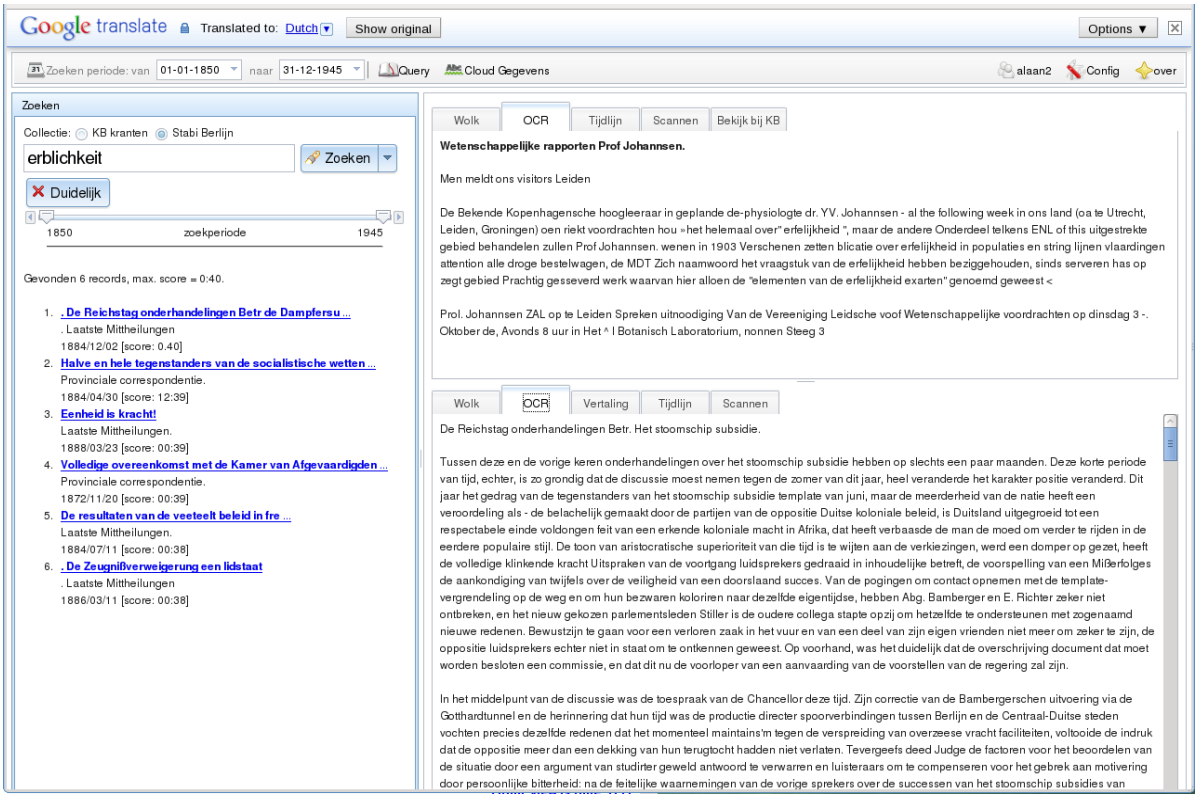


Figure 18: Google translate result.

6 Sentiment highlighting

In fig. 4 we showed the plain OCR text of an article. After turning on the sentiment option in the configuration widget (see fig. 19), the article OCR looks as depicted in fig. 20, with positive and negative sentiment words highlighted[3].

[3]The sentiment highlighting only applies to the dutch text, as we have no sentiment word list for other languages.

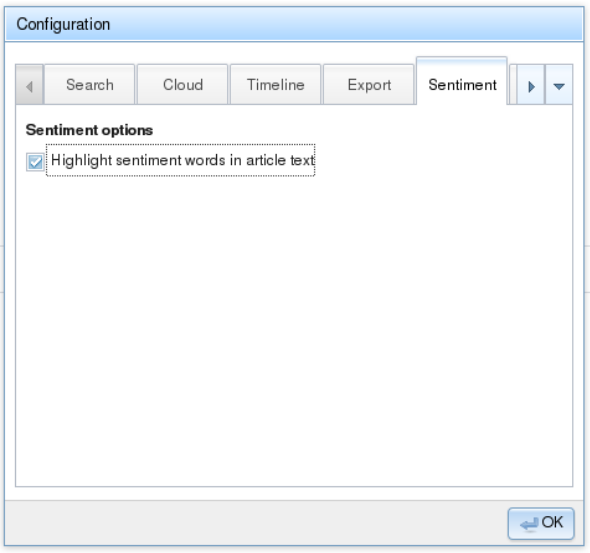


Figure 19: Sentiment option in configuration widget.

The figure also shows that what is highlighted are not whole words, but substrings, which may lead to curious mistakes. And the OCR will never be perfect, which clearly affects the results[4].

Figure 20: OCR text of a KB article with positive and negative sentiment highlighting.

[4]Apart from OCR mistakes, there is a second shortcoming in the data. The semi-automatic segmentation of the newspaper scans into individual articles is not perfect either, leading to numerous 'oversegmentation': 'articles' consisting of just their title, their body text having been delegated to the next article. The current settings of the KB search engine imply that short articles come first in the result list.

7 Miscellaneous options

Fig. 21 shows some search options. The KB documents consist of four different types, which can be selected here. One can also set the number of documents to be returned in each 'chunk'.

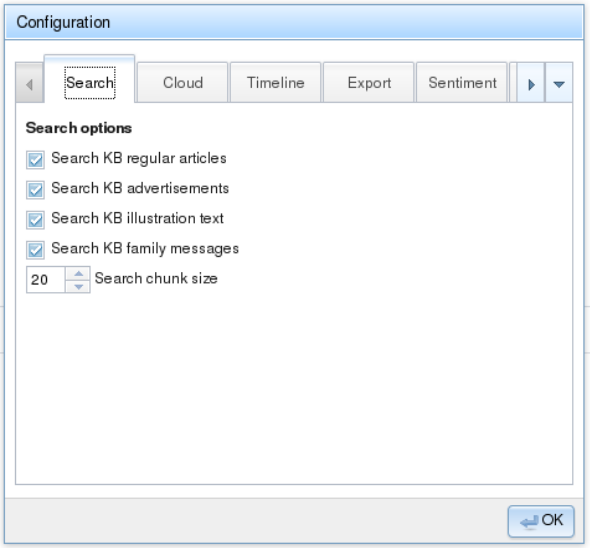


Figure 21: Filter option to select the article types of the KB corpus.

And finally, some timeline optins in fig. 22. The number of KB documents available per time unit (day, month, year) can vary considerably. With normalization the relative counts are used.

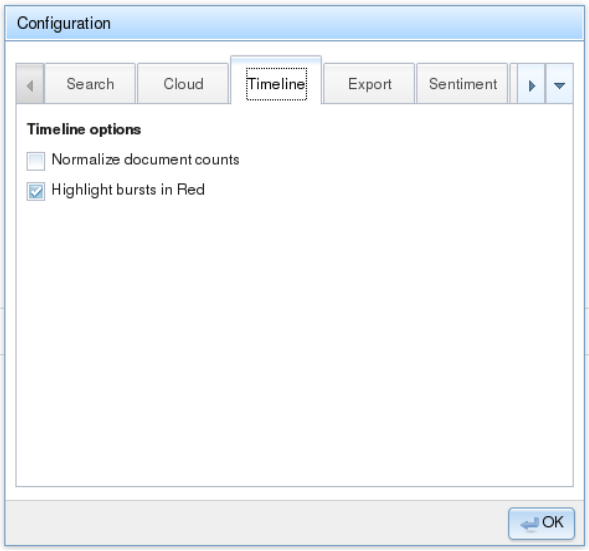


Figure 22: Timeline options.

8 Some abbreviations

Abbr.	*Meaning*
CQL	Contextual Query Language
GUI	Graphical User Interface
KB	Koninklijke Bibliotheek
NER	Named-Entity Recognition
OCR	Optical Character Recognition
SRU	Search/Retrieval via URL
XML	eXtensible Markup Language
xTAS	Text Analysis Service
URL	Uniform Resource Locator
WAHSP	Web Application for Historical Sentiment mining in Public media

Table 1: Abbreviations.

9 Acknowledgments

Apart from having received comments from my WAHSP colleagues (Daan Odijk, Stephen Snelders & Toine Pieters), I also got contributions from José de Kruijf and Jaap Verheul of Utrecht University, and my new Biland colleague Pim Huijnen.