

Gekaapte Brieven application manual

Table of contents

Introduction	3
Information about the corpus	3
GiGaNT Lexicon service	4
Metadata categories	4
Basic	4
Language	4
Text type	4
Year (range)	4
Sender	5
Name	5
Role	5
Gender	5
Occupation	5
Country	5
Region	5
City	5
Ship	5
Recipient	5
Name	5
Role	5
Gender	5
Occupation	5
Country	6
Region	6
City	6
Ship	6
Application user manual	7
Getting started	7
Searching the corpus	8
Simple search	8
Search	8
Wildcards	8
Reset	9
History	9
Global settings	9
Extended search	10

Starting a new search	12
Filter search by	12
Basic	12
Filter by year	13
Filter by Sender / Recipient	13
Expert search	14
Import query	14
Gap filling	14
Viewing results	16
Per Hit view	16
Sorting results	17
Grouping results	17
Per Document view	19
Sorting results	19
Grouping results	19
Exporting results	19
Information about a document	20
Content	20
Metadata	21
Statistics	21
Page image	21
Exploring the corpus	21
Documents	21
N-grams	22
Options	22
Example	23
Statistics (frequency lists)	24
Options	24
Example	24
Appendix: Corpus Query Language	25
CQL support	25
Supported features	25
Differences from CWB	26
(Currently) unsupported features	27
Using Corpus Query Language	27
Matching tokens	27
Sequences	28
Regular expression operators on tokens	28
Case- and diacritics-sensitivity	29
Matching XML elements	29
Labeling tokens, capturing groups	30
Global constraints	30

Introduction

This manual describes the corpus exploitation environment for the *Gekaapte Brieven*. The corpus application is developed by the Dutch Language Institute (Instituut voor de Nederlandse Taal or INT). The backend of the application is the BlackLab Lucene based search engine developed for corpora with token-based annotation (<http://inl.github.io/BlackLab/>). The web-based frontend is a further development of the corpus-frontend application developed by INT (<https://github.com/INL/corpus-frontend>) in CLARIN and CLARIAH projects. Its design is inspired by the first version of the OpenSoNaR user interface by Tilburg and Radboud University (<https://github.com/Taalmonsters/WhiteLab2.0>).

Information about the corpus

The *Gekaapte Brieven* comprises the transcriptions of 5862 letters and other documents, such as bills, that were written in the 17th and 18th centuries to and from sailors and others from abroad. These letters and documents were present on Dutch ships that were hijacked by the English during one of the four wars that were fought by Britain and the Republic of the Seven United Provinces in this period. These letters ended up in the archives of the High Court of Admiralty (now part of The National Archives in Kew), where Dutch historian S. Braunius rediscovered them in 1980. Within *Metamorfoze*, the contents of 7 boxes were photographed, yielding around 9000 scans available at the Dutch [National Archive](#) with the inventory numbers [HCA30-223](#), [HCA30-226](#), [HCA30-227](#), [HCA30-336](#), [HCA30-379](#) and [HCA30-749](#).

In 2011 a crowdsourcing project was launched at the Meertens Institute by Nicoline van der Sijs to add metadata (such as date, sender, addressee, addresses, genre) and transcriptions to the 9000 available scans in order to make the *Gekaapte Brieven* accessible for research. This project was made possible by support from the Culture Fund. Volunteers made so-called diplomatic transcriptions, meaning they followed the original as faithfully as possible. For more information (in Dutch) see [this pdf](#). Thanks to the crowdsourcing project, it became clear that about 6000 of the 9000 scans contain handwritten text (the rest are blank pages or pages from printed books), with about half of the documents being bills, and the other half being personal letters. Most of the documents are written in Dutch, but some are in other languages, such as French, Spanish, Portuguese or German. The letters date from the 17th and 18th centuries. Two-thirds of the 17th-century letters were written between 1664-1672. Most of the 18th-century letters come from the period 1773-1790.

In two tranches, in 2012 and 2014, the transcriptions and scans were placed on a website at the Meertens Institute, developed by Rob Zeeman. Since then the metadata have been enriched and corrected with a grant from the [Time Capsule](#) project. The technology of the website was outdated in 2019, which meant the transcriptions were no longer available. The data were then transferred to the Institute of the Dutch Language (INT), where the metadata were cleaned up and enriched, and a new website and search engine for the data were developed.

The title of each document contains information about the inventory number under which it can be found in the National Archives. Texts can be searched and filtered by metadata and by words in the texts. The page and a thumbnail of the original photograph are displayed. That photo can also be

viewed separately in high resolution. If a letter or document consists of more than one page, you can consult the corresponding pages by clicking on the thumbnails displayed in the left-hand column.

Of a small number of pages, we could not identify which document they belonged to; this was caused by the way the originals were photographed. In those cases, we displayed the documents in the order of the inventory number. To find all letters from one writer or one recipient, the spelling of the names in the Sender field and in the Recipient field have been normalised. Some documents are undated. To these, we automatically added an indication of the period from which they date, e.g. 1642-1675 or 1773-1790.

The *Gekaapte Brieven* website supplements the website [Brieven als Buit](#), also hosted by the INT, which contains 1033 letters with additional metadata and was developed under the supervision of Marijke van der Wal.

This first online accessible version of the *Gekaapte Brieven* was released in December 2023.

GiGaNT Lexicon service

To make the *Gekaapte Brieven* more accessible, suggestions for query expansion are given, using the INT lexicon service with the historical computational lexicon [GiGaNT-HILEX](#).

The current version of GiGaNT-HILEX in the lexicon service contains the lexicon modules based on the *Dictionary of the Dutch Language (Woordenboek der Nederlandsche Taal, WNT)* and the *Dictionary of Middle Dutch (Middelnederlandsch Woordenboek, MNW)*.

If you want to make use of this service, please contact Katrien Depuydt (katrien.depuydt@ivdnt.org).

Metadata categories

The *Gekaapte Brieven* have been enriched with an elaborate set of metadata categories. These metadata are described below. In the corpus application it is possible to limit a search by filtering on metadata categories. The metadata have of course only been added in those cases that the information could be inferred from the text.

Basic

Language

The language or languages in which a text is written.

Text type

The type of text to which the letter or document belongs (e.g. *brief* ‘letter’, *rekening* ‘bill’, *vrachtlijst* ‘cargo list’)

Year (range)

The year or years in which a letter or document was written.

Sender

Name

The name of the sender.

Role

The role played by the person writing the letter (e.g. *ontvanger* ‘recipient’, *getuige* ‘witness’, *koper* ‘buyer’)

Gender

The gender of the sender (*[onbekend]* ‘unknown’, *Man* ‘male’, *Vrouw* ‘female’).

Occupation

The profession of the person who wrote the text.

Country

The country from which a letter has been sent.

Region

The region from which a letter has been sent.

City

The city from which a letter has been sent.

Ship

The name of the ship from which a letter has been sent or on which a letter has been written.

Recipient

Name

The name of the recipient.

Role

The role played by the person receiving the letter (e.g. *ontvanger* ‘recipient’, *getuige* ‘witness’, *koper* ‘buyer’)

Gender

The gender of the recipient (*[onbekend]* ‘unknown’, *Man* ‘male’, *Vrouw* ‘female’).

Occupation

The profession of the person who received the letter.

Country

The country to which a letter has been sent.

Region

The region to which a letter has been sent.

City

The city to which a letter has been sent.

Ship

The name of the ship to which a letter has been sent.

Application user manual

Getting started

Here are a few examples of what you can do with the corpus application (the links will take you to the application):

- To search for a word literally in the form you specify, use Simple Search:
 - Simple Search for Word [schip](#)
- To search for different spellings and inflections/conversions of a given word, use Extended Search:
 - Extended Search for word forms of [schip](#)
- To search for words satisfying a certain pattern, use *wildcards* in Simple Search or Extended Search, or *regular expressions* in Expert Search
 - words starting with *ver* and ending with *len* in [Simple Search](#)
 - words starting with *ver* and ending with *len* in [Extended Search](#)
 - words starting with *ver* and ending in *eren* with one syllable in between in [Expert Search](#)
- To see which unique forms occur as a result of your search, use the Group hits by feature.
 - example Group by Context (advanced): [all words following lieve](#)
 - example Group by Word before: [different words preceding the word huis](#)
- To explore the distribution of document properties in the corpus, use the Explore feature
 - example: [characteristics about the text types in Gekaapte Brieven](#)
 - example: [sender](#)

Searching the corpus

Simple search

Search

The Simple Search allows you to quickly search for specific word forms (e.g. *huys*). After entering a search term, a spinner briefly appears on the right side of the search bar. Based on the keyed in word, suggestions are given of possible variants of spelling and/or form from the GiGaNT-lexicon.

Based on the information in this lexicon all spelling variants of the search term found are suggested (see screenshot below). You can then choose from the presented suggestions or select all at the same time (Select all). To make your search even more targeted, it is also possible to limit the search to the parts of speech that were found in GiGaNT-HILEX in connection to the search term.

The screenshot shows a search interface with a red header bar containing 'Search' and 'Explore'. Below the header, there is a search bar with the text 'Search for ...'. Underneath the search bar, there are three tabs: 'Simple' (selected), 'Extended', and 'Expert'. A 'Word' input field contains 'huis'. Below the input field, there are 'Select all' and 'Deselect all' buttons. A list of word variants is displayed with checkboxes: hues, hus, huysen, huis, huse, husen, huise, huisen, huys, huizen, and huise. Below this list, there is a section 'Limit to Part of Speech' with a checked checkbox for 'huis (NOU-C)'. At the bottom, there are four buttons: 'Search' (red), 'Reset', 'History', and a settings gear icon.

If you know exactly which word you are looking for, you can also – while the wheel is spinning – press Enter directly. The search will then start immediately.

It is also possible to enter a phrase: *groet mijn* or *als God mij*. You will then find all occurrences of that exact phrase.

Note that in Simple Search the patterns will be matched case-insensitively: *capitein* for instance will deliver the same results as *Capitein* or *CAPITEIN*. See the paragraph [Grouping results](#) in Per Hit view to see how it is nevertheless possible to distinguish between uppercase and lowercase letters.

Wildcards

In Simple Search, the use of wildcards can prove good service to search for specific word forms. A wildcard is a symbol used to replace or represent one or more characters. The following two wildcards are supported:

- * The asterisk matches any character zero or more times. Therefore, *a*n* matches all values that start with an *a* and end with a *n*, e.g. *Aan*, *alleen*, *Amen* and *Andriessen*.

? The question mark matches a single character once. Therefore, searching for *ann?* matches *only* four-letter values starting with *ann* and ending with a random character, e.g. *Anna*, *Anne*, *anni* and *anno*.

This wildcard can be used more than once. Thus *ar???n* matches words like *Ariaen*, *arijan* en *Arztin*. Note that searching with wildcards is limited to Simple Search and Extended Search. [In Expert Search you can use so-called regular expressions instead of wildcards.]

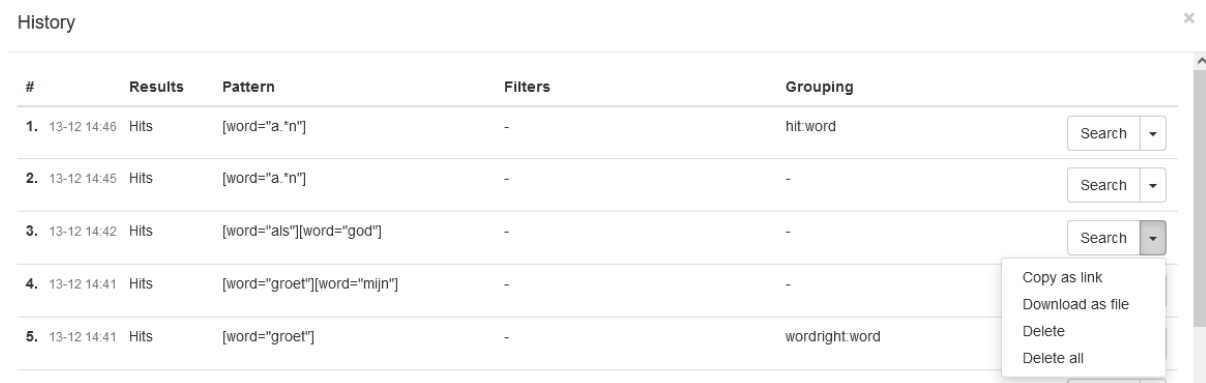
Reset

You can start a new search by pressing the Reset button. By doing so, both the search query and the hits found will be cleared. Your search history, however, will remain unchanged.

Note that it is also possible to start a new search by entering a new word or phrase in the search field Word.

History

The History button will display your query history. Per search query there are several possibilities (as shown in the screenshot below): you can perform the search query again (Search), you can copy the search query as a link (Copy as link), you can download the search query as a file (Download as file), you can delete a single search query (Delete) or delete all search queries (Delete all).



#	Results	Pattern	Filters	Grouping	
1.	13-12 14:46 Hits	[word="a.*n"]	-	hit:word	Search
2.	13-12 14:45 Hits	[word="a.*n"]	-	-	Search
3.	13-12 14:42 Hits	[word="als"][word="god"]	-	-	Search
4.	13-12 14:41 Hits	[word="groet"][word="mijn"]	-	-	Copy as link Download as file Delete Delete all
5.	13-12 14:41 Hits	[word="groet"]	-	wordright:word	

Every search query has its own url. If you copy this url via History (Copy as link) or directly from the address bar of your browser, you can send it to someone else who can import this link via Import from a link. It offers that person the possibility to run the search on his or her own computer.

Global settings

The Global settings dialogue, activated by pressing the wheel button, allows you to configure five settings: Results per page, Sample size, Seed, Context size and Wide View.

- *Results per page*: you can choose whether you want 20, 50, 100 or 200 results to be shown;
- *Sample size*: selecting a value here will instruct the search engine to return a random sample drawn from the complete result set. (Pressing the Reset button does not change the sample size. It must be changed manually.) The sample size can be limited by
 - a percentage of the total number of search results (percentage)
 - the number of results displayed (count);

- *Seed*: a ‘random seed’ is a number used to initialize a so-called pseudo-random number generator. Keeping the same seed will ensure that two samples drawn from the same result set are identical. A new seed will most likely result in a different sample;
- *Context size*: by entering a number you can determine the number of words Before hit and After hit;
- *Wide View*: the default setting is ‘small view’; you can change to Wide View by ticking the checkbox.

The screenshot shows a 'Global settings' dialog box with the following configuration:

- Results per page:** 20 results
- Sample size:** percentage, 1
- Seed:** 1287260416360736
- Context size:** Context size
- Wide View**

A red 'Close' button is located in the bottom right corner.

Extended search

Like in Simple search, Extended Search allows you to quickly search for specific word forms. The search is performed in the same way as described for Simple Search.

After entering a search term, a spinner briefly appears on the right side of the search bar. Based on the keyed in word, suggestions are given of possible variants of spelling and/or forms from the GiGANT-lexicon.

Based on the information in this lexicon all spelling variants of the search term found are suggested (see screenshot below).

Search **Explore**

Search for ...

Simple Extended **Expert**

Word

Select all Deselect all

vlees vleesch vleeschjes
 vleijs vleis vlesche
 vleys

Limit to Part of Speech

vlees (NOU-C)
 Case- and diacritics-sensitive

You can then choose from the presented suggestions or select all at the same time (Select all). To make your search even more targeted, it is also possible to limit the search to certain parts of speech that were found in GiGANT-HILEX in connection to the search term. It is also possible to enter a phrase: *wij ordonneren den schipper* or *gelieft ons altijd te aduijseren*.

In Extended Search it is also possible to search case- and diacritics-sensitive. Note that the default setting for search is case- and diacritics-insensitive. For example, searching for the Word *willem* (& Willem (NOU-P NOU-C)) will result in 167 occurrences of this name (*Willem*, *willem*). By ticking the box Case- and diacritics-sensitive you will only find 35 occurrences of the Word *willem*, but 132 of *Willem*. In order to directly find only occurrences of the Word (form) *Willem*, use the search term *Willem* and tick the box Case- and diacritics-sensitive under the search field Word (as shown below).

Search **Explore**

Search for ...

Simple Extended **Expert**

Word

Select all Deselect all

willem willems

Limit to Part of Speech

Willem (NOU-P NOU-C)
 Case- and diacritics-sensitive

If you know exactly which word you're looking for, you can also – while the wheel is spinning – press Enter directly. The search will then start immediately.

Like in Simple Search, wildcards are supported in Extended Search. (See for a short explanation of wildcards [Simple Search](#))

In the search field Word it is possible to search for different values simultaneously by separating them without spaces by a vertical line, e.g. *god|man|lief* or – with the use of wildcards – *god|aan*|hond*.

For the search field Word it is possible to search for a series of tokens by entering multiple values – including wildcards – separated by a space, e.g. *Spaans groen, Spaans ** or ** groen*. It will be obvious that these three searches give different results.

Starting a new search

You can start a new search by pressing the Reset button. By doing so, both the search query and the hits found will disappear. Your search history, however, will remain unchanged.

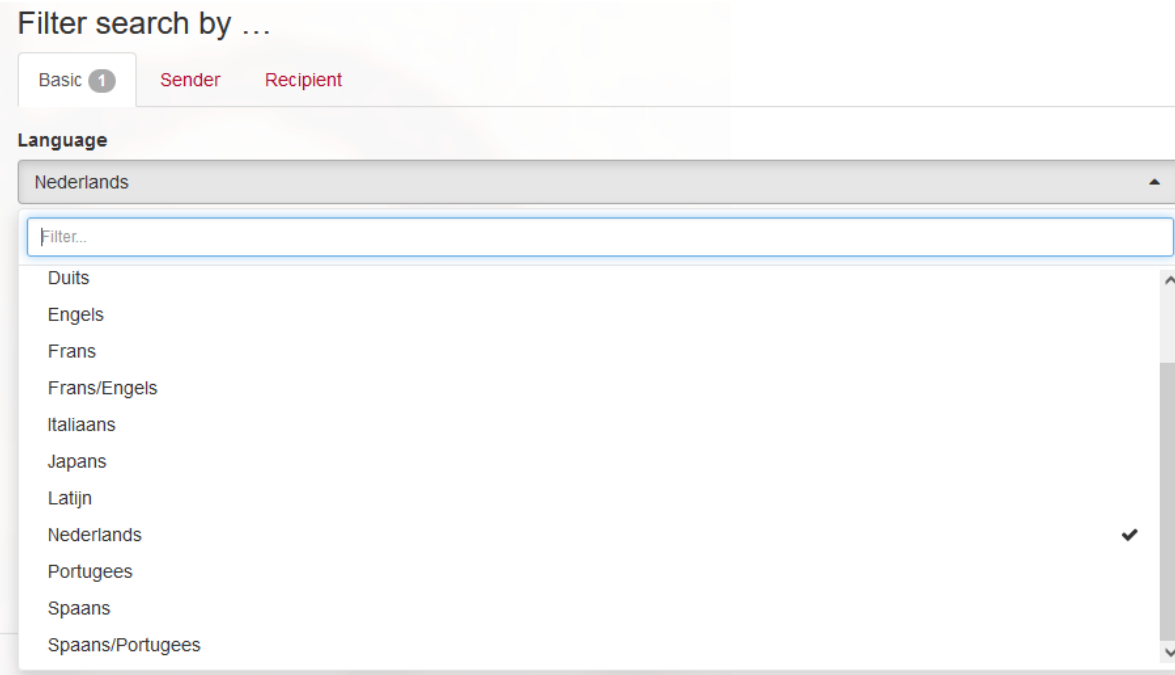
There are two possibilities to start a search: fill in the desired value and press enter or fill in the desired value and then click the Search button.

Filter search by

At the right side you will find the option to limit your query to a subset of documents with specific metadata values. You can apply different filters for Basic (*Language, Text type, Year (range)*), Sender (*Name, Role, Gender, Occupation, Country, Region, City, Ship*) and Recipient (*Name, Role, Gender, Occupation, Country, Region, City, Ship*).

Basic

There are two different ways to specify a filter, depending on the field type. You can either fill in a value yourself or choose one or more values from a drop-down list. The drop-down list has been applied especially when the number of values to choose from is relatively small. *Language* for instance has only thirteen possibilities. You can pick one of these values by clicking on it; your choice will be marked with a tick.



The screenshot shows a web interface for filtering search results. At the top, it says "Filter search by ...". Below this, there are three tabs: "Basic" (with a small circle containing the number 1), "Sender", and "Recipient". The "Basic" tab is selected. Underneath, the "Language" filter is expanded. The current selection is "Nederlands". Below the selection is a search box labeled "Filter...". A list of language options is shown below the search box, including "Duits", "Engels", "Frans", "Frans/Engels", "Italiaans", "Japans", "Latijn", "Nederlands" (which has a checkmark to its right), "Portugees", "Spaans", and "Spaans/Portugees".

It is possible to choose several values. If you want to delete a selection, you can click on the corresponding line again. To close the drop-down list, you can either press the upward pointing arrow in the upper right corner or simply press escape.

Filter by year

The documents in this corpus were written or printed in the period between 1625 and 1798. You can find documents from a specific year by entering the same year in the "from" row as in the "to" row (see screenshot below). If you do not enter a date, the entire corpus is searched. If you want to filter by another year or another period, please press the "reset" button.

Filter search by ...

Basic 1 Sender Recipient

Language

Text type

Year (range)

1672 1672

Permissive Strict

Filter by Sender / Recipient

When on the other hand the set of possible values is rather large (e.g. Sender or Recipient), you have to type a specific value in the search field. After entering a single character, a list of possible values is suggested. Clicking on an auto-completed value will paste that value in the field. Note that this only works with a single word, like *backer*.

By means of a number at the top of 'Filter search by', the number of values used to filter on, is displayed as can be seen in the above screenshot.

For a detailed description of the metadata, see the section [Metadata categories](#).

Expert search

The Corpus Query Language (CQL) editor allows you to type your own CQL query, to import a previously downloaded query and to upload a tab separated list of values to substitute for gap values (see below for further explanation).

CQL queries are expressions built up with the help of a few sequence operators and brackets from basic blocks enclosed by square brackets, in each of which one or more token attributes are specified.

In CQL, spaces only affect a search if they are included in quotes. Whether the search command is `[word="schip"]` or `[word = "schip"]` (or just “schip”) does not make any difference to the result. However, there is a difference between the queries `[word="schip"]` and `[word=" schip"]`. The first search results in exactly 27 hits, but the second one in zero!

Some examples:

- Simple: `[word="schip"]`, e.g. the attribute word matches the regular expression *schip*; `[word!="schip"]`, e.g. the attribute word does **not** match the regular expression *schip*; `[word="*.man"]` matches all words ending with *man*, including *man* itself. (Note that `[word="*man"]` will not give any results, because in Expert Search an asterisk is not a wildcard but a repetition operator.)
- Combination of attributes (combining operators are `&`, `|`, `!`), e.g. `[word="hoop|geloof|liefde"]` matches either the word *geloof*, the word *hoop* or the word *liefde*.
- The empty `[]` matches any token, e.g. `[word="man"][]{}[word="ik"]` matches a sequence of *man* followed by *ik* with three arbitrary tokens in between.
- Operators `|`, `&` and parentheses `()` and the repetition operators `(+)`, `*`, `?` and `{}` can be used to build complex sequence queries. Example: `"deese" "goederen" | "mijn" "moeder"`, matching any sequence of *deese goederen* or *mijn moeder*.

This short list does not cover all CQL features. For more detailed information on how to write CQL, please consult the short [Appendix: Corpus Query Language](#), which contains further pointers.

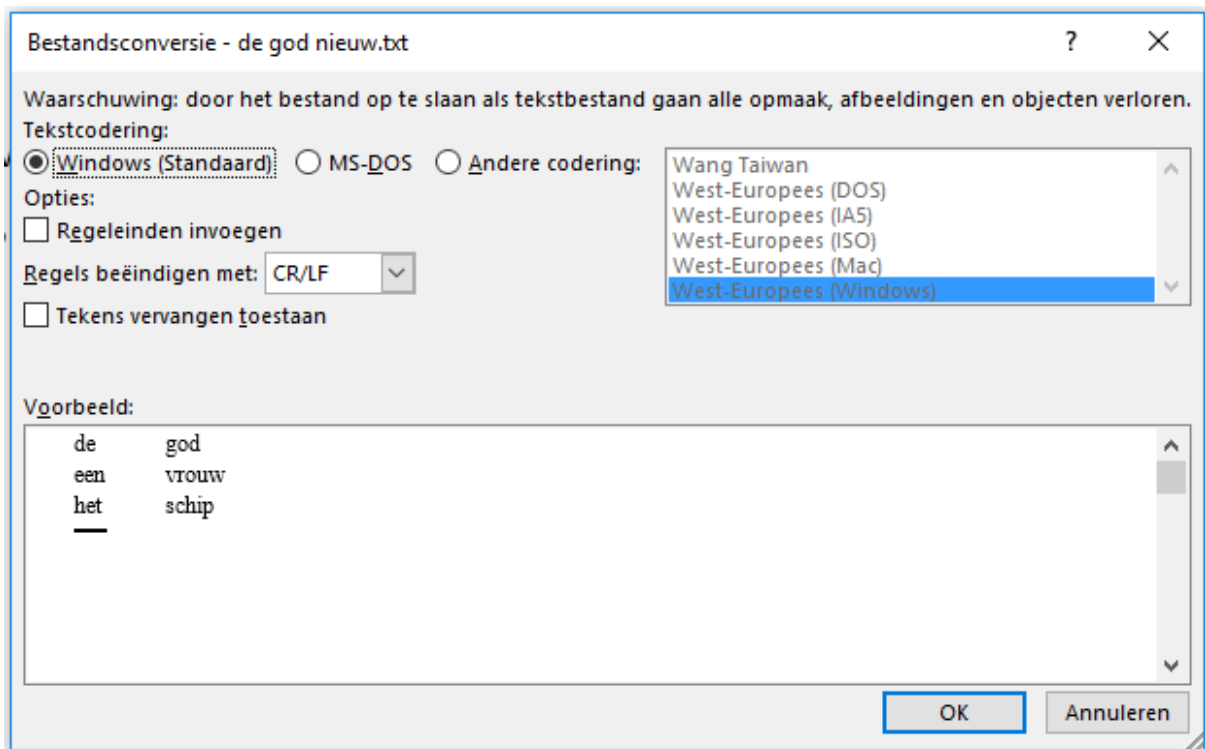
Import query

If you have entered a search query, you can find it back by clicking the History button. On the right hand side you can select Download as file in the drop-down menu (default value is Search) and save the file. (For a more elaborate description of the History button see [Simple Search](#))

Previously saved queries can be used again by uploading them through the Import query button.

Gap filling

Use this button to upload a Tab Separated Values (TSV) file, which is a simple text format for storing data in a tabular structure. Each record in the table is one line of the text file. Each field value of a record is separated from the next by a tab character. It is also possible to upload a plain text file (.txt) that has the same properties, as is shown in the following screenshot:

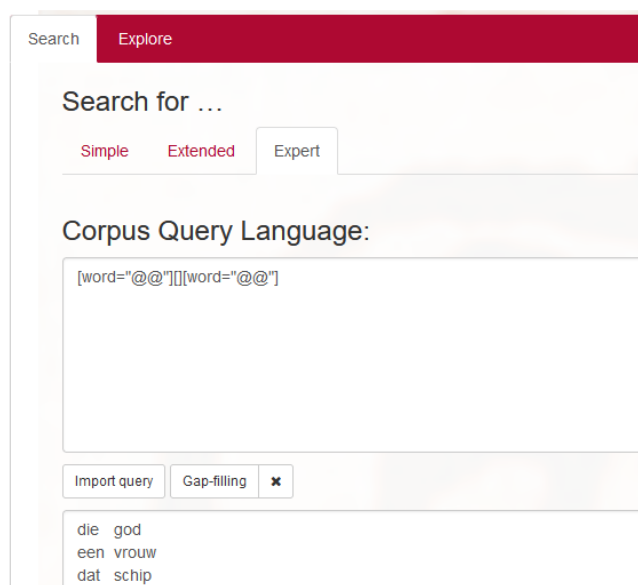


A .tsv file or a comparable .txt file enables you to complete a query with marked gaps.

If, for instance, you are interested in the distribution of words that can be placed between two specific words you can create this query in the Corpus Query Language field:

```
[word="@@"][word="@@"]
```

By clicking Gap-filling you can upload a file with a tab-separated list of values from your computer to substitute them for the gap values, i.e. the at signs (@@) in your query. After the upload your values will appear in a separate box:



The values in the first column – *die, een, dat* – will be entered at the position of the first gap (@@) and the values in the second column – *god, vrouw, schip* – at the position of the second gap. With these values, gap-filling yields the following results (titles are hidden):

Per Hit		Per Document		
Hits				Total hits: 39 (0.00395%)
Group hits by...				Search time: 0.003s
<input type="text" value="1"/> <input type="text" value="2"/>				
Before hit	Hit	After hit	Language	
...goede negotie daermede incontreeren, alsmede	dat ue schip	Sint Salvator soo schoene compagnien...	Nederlands	
...solemnelen eede ver- klaert heeft /	dat het schip	genaemt St. Maria van Conceptie...	Nederlands	
...moeder en suster lena belangt	die zijn god	lof noch gesont hope dat...	Nederlands	
...over sien hebben ende oock	dat het schip	soo haest verrock doch sal...	Nederlands	
... hebbe jck alhier verstaen als	dat ons schip	al thujs was daer jck...	Nederlands	
...eijghe wil is Geresolveer geweest,	dat het schip	niet langer bij mij soude...	Nederlands	
...agteren op de vleet afstoot, &	dat 't schip	tot ber- stens toe daardoor...	Nederlands	
... 't roer verkeert leggen tot	dat 't schip	eenigzins afdraayt, & wanneer de wind...	Nederlands	
...wagten, & met 't roer stutten,	dat 't schip	, als de zeylen vol raakten...	Nederlands	
...aan boord, zo meen ik	dat 't schip	voor de wind om moet...	Nederlands	
...nieuws hier reeds bekend is,	dat dit schip	't welk den 16. December...	Nederlands	
...vrouw, ach! 't was zo	een goede vrouw	! ik heb 'er veelmaals lekkertjes...	Nederlands	
...eer UWelEd. te bedeeelen, als	dat het schip	de Ganges behouden ter deeser...	Nederlands	
...my daarvan afsien, so om	dat dit schip	naer de saidanhabaay ten anker...	Nederlands	
...mij daarvan afzien, so om	dat dit schip	naer de Daldanhabaay ten anker...	Nederlands	
...zal koomen, en wij Sustineeren	dat 't schip	de Zeeploeg waarmede die brief...	Nederlands	
...Praemie. Terwyl wij hier dugten,	dat 't Schip	de Zeeploeg, waarmede myn brief...	Nederlands	
...Wal gehad, en wil hoopen,	dat zijn schip	met hem meede gelukkig bij...	Nederlands	
...al toe gemaakt, en om	dat het schip	vertrekt, ken ik hem niet...	Nederlands	
...bemaent synde wel te weten	dat tselve schip	ghemackt is te finteete ? bij...	Nederlands	

Please note that for this to work, you do need to enter @@ in the field where you want the substitution to take place. An empty field ([]) will match any term.

Viewing results

Results can be viewed in two ways: Per hit (hit is defined as one token or a group of tokens that matched the query), or Per document (each document listed contains at least one hit).

Per Hit view

Click a hit – i.e. a line with the bold word(s) in the column Hit – to display the properties and values of the hit (in the following example **dat tselve schip**). Click the hit again to close.

Document id: nl-hana_hca30-226.1_5_0076	Before hit	Hit	After hit	Language
nl-hana_hca30-226.1_5_0076: Brief, 1652-1673				
...bemaent synde wel te weten	dat tselve schip	ghemackt is te finteete ? bij...	Nederlands	
<p>stijlle ende poorters ende inwoonders der voornomde stede, de wel opt versouck van Joos de vos ende pieter clajnsen, jeghenwoordich opgevoert tot douners/douers ? van een schip van oorloghe competeerende synne Counclickicke Majesteit van Enghelant, hebben sij beyde comparanten verclaert onder solemneelen eede daertoe bij ons bemaent synde wel te weten dat tselve schip ghemackt is te finteete ? bij t voornomde veurne waarmede sij opgevoert sijn naer t voorenn. douers twelcke is ghedaen maecken bij Jaecquis van der balcke fg pieters oock schipper porter ende inwoonder van t voorseijde veurne dat hij vercocht heeft aen voorseijde de vos ende clajnsen ende vant den voorseijde</p>				
Property		Value		
Word		dat	tselve	schip
Main part of speech				
Part of speech (with features)				
Lemma				

Hit rows are always preceded by a row containing the document title in which those hits occurred, in this case *nl-hana_hca30-226.1_5_0076: Brief, 1652-1673*. The document titles can be toggled on or off by using the Hide Titles (or Show Titles when titles are hidden) button at the bottom of the page. If

you hover the mouse over the title, the identification number of the document appears, in this case: *nl-hana_hca30-226.1_5_0076*.

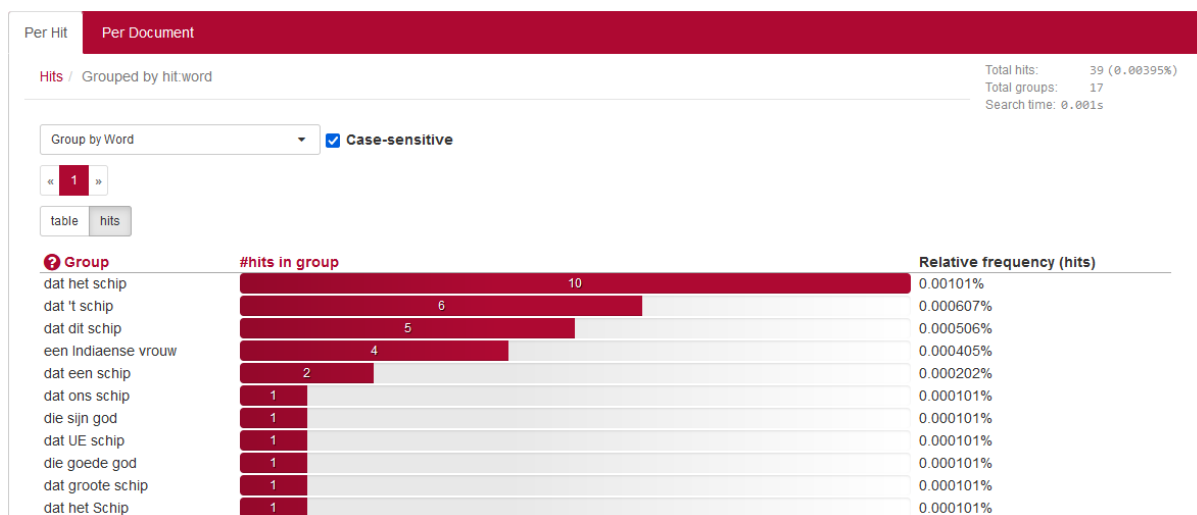
Sorting results

Click on any of the column headings to sort the hits on Words within that column, clicking again inverts the sorting.

You can also sort the results by means of the drop-down menu at the bottom of the page (Sort by ...), which offers you the possibility to sort by various attributes as Hit, Before hit, After hit, Basic, Sender and Recipient.

Grouping results

Results Per Hit can be grouped by properties of Hit, Before hit, After hit, Basic, Sender and Recipient. Grouping is facilitated by the drop-down menu Group hits by. By selecting one of the properties a tick box appears that makes it possible to distinguish between case sensitive and case insensitive.



Advanced grouping options are available by selecting the option Context (advanced). It allows you to group the results by up to 5 tokens before or after the hits. It also allows you to group the results based on (parts of) the hits. By pressing the New context group you can group the results by another property or another range.

We will work that out using an example. A noun phrase consisting of a pronoun/determiner *die*, the adjective *lieve* or the adjective or noun *liebe* and an arbitrary word may be found in Expert Search with the following query: [\[word="die"\]\[word="lie.?e"\]\[word=".*"\]](#). This produces the following hits (Titles are hidden):

Per Hit Per Document

Hits Total hits: 9 (0.000911%)
Search time: 0.002s

Group hits by...

« 1 »

Before hit ▾	Hit ▾	After hit ▾	Language
...lieb achte diese Plausereij nichtes,	die liebe Jungffr.	thuet dem H. hinwieder umb...	Duits
...ik ben ook seer verlangend	die liebe jonge	te omhelsen de berigten van...	Nederlands
...mein Sohn thu uns doch	die Liebe und	schreib unß sehnlich von deinen...	Duits
...es Glück mich gedienet mit	die liebe Jungfr	. Lesson etzliche Mahl zur Hochzeit...	Duits
...dan zijt hij tegen mijn	die lieve swalker	van een vader komt die...	Nederlands
...de Negootzie gegeve dus kunne	die lieve luijtjes	het zeer wel stellen . hy...	Nederlands
...hübsch munter und sie und	die liebe Schwester	Flore haben ein ander recht...	Duits
...Lust bey ihnen zu wohnen.	Die liebe schwarze	Gemeine hat mir auch einen...	Duits
...mirs bis diese Stunde geblieben.	Die Liebe, mit	welcher wir wie ein Hausgemeinlein...	Duits

Sort by... [Show Titles](#) [Export](#) [Export for Excel](#)

It is now possible to group the hits by the second and third tokens of those hits. See below.

Per Hit Per Document

Hits / Grouped by context.word.i:H2-3 Total hits: 9 (0.000911%)
Total groups: 9
Search time: 0.003s

Context (advanced) [Apply](#)

Word ▾ Before Hit After Case-sensitive

From end of hit

New context group

« 1 »

table hits

Group	#hits in group	Relative frequency (hits)
liebe Jungffr.	1	0.000101%
liebe Jungfr	1	0.000101%
Liebe mit	1	0.000101%
liebe schwarze	1	0.000101%
Liebe und	1	0.000101%
lieve jonge	1	0.000101%
liebe Schwester	1	0.000101%
lieve luijtjes	1	0.000101%
lieve swalker	1	0.000101%

Sort by... [Export](#) [Export for Excel](#)

Click a group to show or hide hits within that group, as shown below. Click once more on the group to close it again.

Group	#hits in group	Relative frequency (hits)
liebe Jungffr.	1	0.000101%
liebe Jungfr	1	0.000101%
Liebe mit	1	0.000101%
liebe schwarze	1	0.000101%
Liebe und	1	0.000101%
lieve jonge	1	0.000101%

« View detailed concordances »

Before	Hit	After
... ik ben ook seer verlangend	die liebe jonge	te omhelsen de berigten van ...

liebe Schwester	1	0.000101%
lieve luijtjes	1	0.000101%
lieve swalker	1	0.000101%

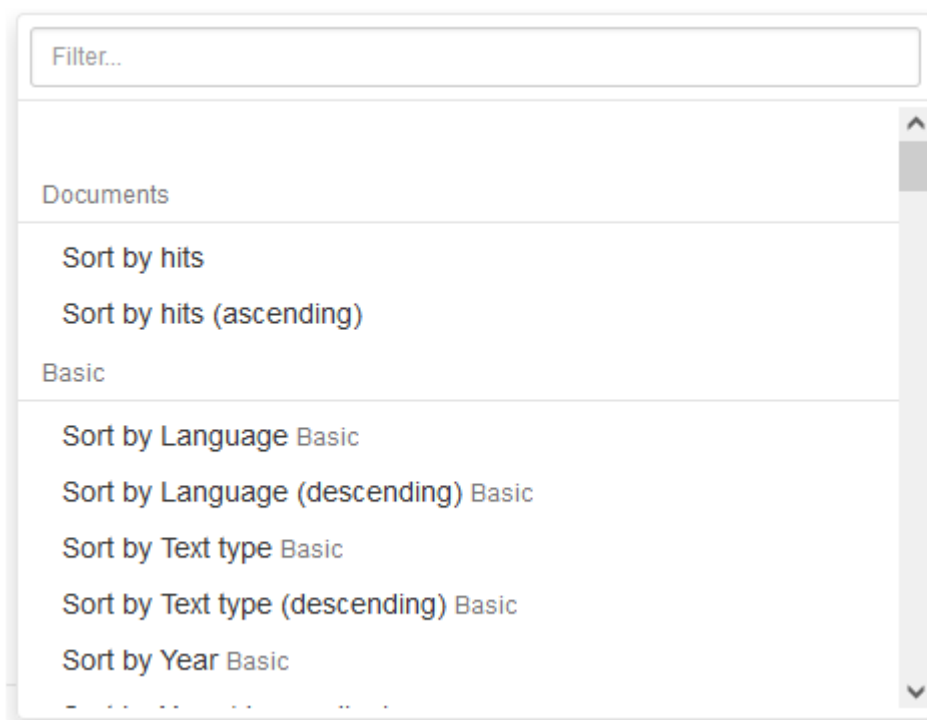
If more than twenty hits are found in a document, you can make them appear by clicking on Load more concordances; this button will appear right to the button View detailed concordances.

Click on View detailed concordances to go back to the normal hits view to see more detailed information for the hits in this group. The button Go back to grouped view brings you back to the list of groups.

Per Document view

Sorting results

Results can be sorted by means of the drop-down menu at the bottom of the page, which enables you to sort by Documents (e.g. the number of hits for your search query) and by Basic (e.g. Language, Text type, Year), Sender or Recipient.



Click on a Document title to show the Content of the document in a new window. Hits from the current query will be highlighted in bold in the opened document. In the case of several hits the first hit will also appear in shadow. You can go to the next (or previous) hit within the same document by pressing the Hits (and – in the case of many hits – Pages) button.

Grouping results

Results Per Document can be grouped by the metadata of the documents in which those hits occur (Basic (e.g. Language, Text type, Year), Sender or Recipient). Here, grouping is facilitated by the drop-down menu Group docs by.

Exporting results

The search results – both Per hit as Per document – can be exported by using the Export or the Export for Excel button at the bottom right of the page. The first button transfers the search results – including all metadata – to a Comma-Separated Values-file. These CSV-files consist only of text data,

which makes it easy to implement (read and/or write) them into a spreadsheet or database program. The second button offers the possibility to export the results – including all metadata – to a CSV-file for use with Excel.

Grouped results can be exported in the same way. However, if you would like to have the metadata with each concordance of a group, you must first click on the red bar of a specific group and then on View detailed concordances (see screenshot below). The results you then see can be exported by the use of the Export buttons. This operation must be carried out for each individual group you wish to export.

Document	Date	Text type	Language	Containing document id	Hits
nl-hana_hca30-223_5_0101: Cognossement, 1672-09-03	1672-09-03	Cognossement	Nederlands	doc_3534	4
nl-hana_hca30-223_5_0103: Cognossement, 1672-09-13	1672-09-13	Cognossement	Nederlands	doc_1535	4
nl-hana_hca30-223_5_0105: Brief, 1672-03-30	1672-03-30	Brief	Nederlands	doc_2813	1
nl-hana_hca30-223_5_0111: Brief, 1672-09-05	1672-09-05	Brief	Nederlands	doc_1158	1
nl-hana_hca30-223_5_0123: Brief, 1672-11-05	1672-11-05	Brief	Nederlands	doc_2047	1
nl-hana_hca30-223_6_0002: Brief, 1672-09-14	1672-09-14	Brief	Nederlands	doc_3256	1
nl-hana_hca30-223_6_0001: Brief, 1672-08-31	1672-08-31	Brief	Nederlands	doc_3256	1
nl-hana_hca30-223_6_0005: Cognossement, 1672-08-23	1672-08-23	Cognossement	Nederlands	doc_0564	4
nl-hana_hca30-223_6_0007: Brief, 1672-09-03	1672-09-03	Brief	Nederlands	doc_2310	1
nl-hana_hca30-223_6_0009: Brief, 1672-08-06	1672-08-06	Brief	Nederlands	doc_0005	2
nl-hana_hca30-223_6_0011: Brief, 1672-09-12	1672-09-12	Brief	Nederlands	doc_1035	1
nl-hana_hca30-223_6_0016: Brief, 1672-09-14	1672-09-14	Brief	Nederlands	doc_1562	1
nl-hana_hca30-223_6_0015: Brief, 1672-09-03	1672-09-03	Brief	Nederlands	doc_1562	3
nl-hana_hca30-223_6_0017: Brief, 1672-08-31	1672-08-31	Brief	Nederlands	doc_2206	1

Information about a document

Click on a document title to open this document in a new window: the Content window.

Content

On the left are thumbnails of the original pages, on the right is a transcription of the text in that picture. The photo of the one shown has a red bar at the top and bottom. Comments about the transcription and uncertain transcriptions are highlighted in yellow:

nl-hana_hca30-223_6_0009: Brief, 1672-08-06

Pages Current page content

Seigneur J. Van der poelen
 Ady **VI te?** augusto 1672 zerunaame

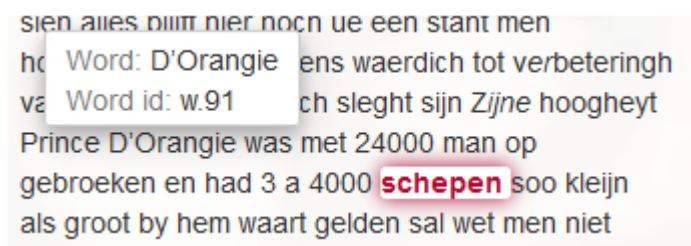
zeven regels lager links in de marge: Copia

Monsieur In dato 26 majj per Capteijn J. dimmens van Castij was aan UE mijnen laasten waer Aen mij ref **?** en seedert Is hier op den 4 Junio gearriveert t **schip** d'Anna per Schipper pieter doncker waermeede de droeuiege Tijdinge becoomen, dat de engelsse en franssen, teegen onssen staat den oorlogh Gepubliceert en al reets vijandtijck Geattacqueert hebbe. godt heere wil ons beschermen en bij staan want het 2 magtige

Hits from the current query will be highlighted in bold in the opened document. In the case of several hits only the current hit will also appear in shadow. You can navigate from one hit to another by using the arrows at the Hits button:



When you hover with your mouse over a specific word in the document a pop-up will appear with the word form and its word id:



Metadata

In the Metadata tab all metadata properties of the document are displayed. They provide information about Basic (Title, Language, Text type, Year, Date), Sender, Recipient and Document length (tokens).

Statistics

The Statistics tab shows several document statistics: the number of tokens, the number of unique word types and the type/token ratio. It is possible to print or to download these statistics via the menu symbol right of the title Vocabulary Growth.

Page image

You can click on 'Page Image' to look at the original document itself. If you hover the mouse over the photo, a navigation menu with six active buttons appears at the top left. The plus and minus sign allow you to zoom the photo in and out, respectively. The home button returns the photo to its original size. To view the photo in full screen, you can press the adjacent button. The arrow buttons allow you to rotate the photo to the left and right, respectively.

Exploring the corpus

The Explore tab has three subdivisions: Documents, N-grams and Statistics.

Documents

This subtab allows you to investigate the documents. It consists of two drop-down menus to specify the grouping of the metadata and to specify the way the groups are to be shown.

A simple example: suppose we want to know which documents are written in German (in Dutch: *Duits*) in the *Gekaapte Brieven*.

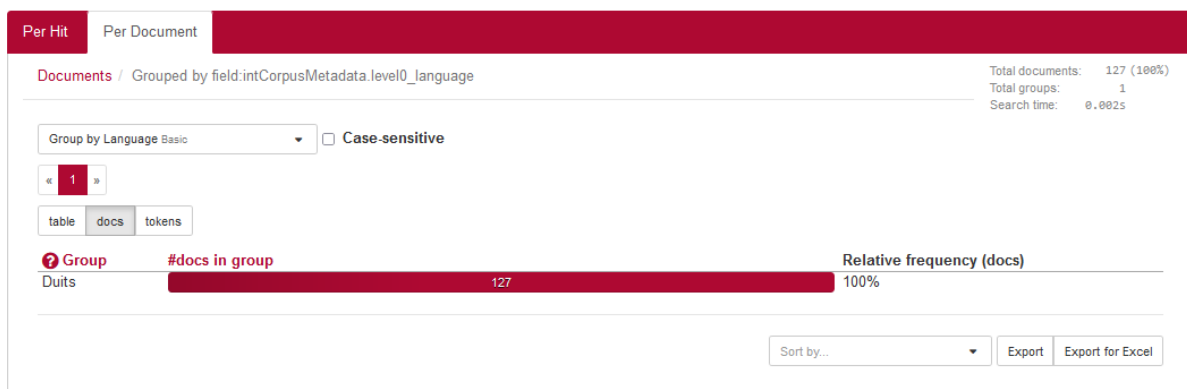
- In the Group documents by metadata drop-down menu, choose Group by Language

- In Show groups as, select *docs*
- In the metadata search form (Filter search by), select in Basic Language *Duits*
- Press Search

The screenshot shows the 'Explore' interface with the following settings:

- Navigation: Search | Explore
- Explore ...: Documents | N-grams | Statistics
- Group documents by metadata: Group by Language Basic
- Show groups as: docs
- Filter search by ...: Basic (1) | Sender | Recipient
- Language: Duits

You will get this result:



N-grams

An *N-gram* is a sequence of *N* items. This option will list the frequency of different N-grams in a (sub-)corpus.

Options

- *N-gram size*: the length of the sequence (a number from 1 to 5; default setting is 5)
- *N-gram-type*: a sequence of five consecutive Words (i.e. word forms). If you do not specify the search term a series of five arbitrary words will be searched for.
- It is also possible to restrict to, for instance, 5-grams with some slots already specified, as is shown in the following example. After entering a search term, a spinner briefly appears on the right side of the search bar. Based on the keyed in word, suggestions are given of possible variants of spelling and/or form from the GiGaNt-lexicon and of parts of speech.
- By using the Filter search by ... you can create a subcorpus within the *Gekaapte Brieven* for specific metadata.

Example

Search Explore

Explore ...

Documents N-grams Statistics

N-gram size: 5

N-gram type: Word

Word Word Word Word Word

dat|de|deen|den|des|die|dier Word Word frouw|varen|ver|vere|veren|vor Word

Select all Deselect all Select all Deselect all

- dat
- de
- deen
- den
- des
- die
- dien
- diens
- dier
- dies

Limit to Part of Speech

- die (PD)
- de (PD ART)
- dij (NOU-C)
- dijen (VRB)
- du (PD)

- frouw
- varen
- ver
- vere
- veren
- voren
- vrou
- vrouwen
- vrous
- vrouw
- vrouwe
- vrouwen
- vrouwen
- vrouws

Limit to Part of Speech

- vrouw (NOU-C)

Within all the documents of the *Gekaapte Brieven*, you will find 11 occurrences of this so-called 5-gram in letters sent from Surinam (Filter search by ..., Sender, Country: Suriname).

Per Hit Per Document

Hits / Grouped by hit:word

Total hits: 11 (0.00735%)
Total groups: 11
Search time: 0.01s

Group by Word Case-sensitive

« 1 »

table hits

Group	#hits in group	Relative frequency (hits)
den met een vrouw over	1	0.000668%
den Heer Gouverneur ver loff	1	0.000668%
dat ons weder varen is	1	0.000668%
den Geinsinuerde bewoog ver mitshij	1	0.000668%
den almachtigen ende ver blijve	1	0.000668%
die alle dinge ver koop	1	0.000668%
dat mij liefve vrouwe en	1	0.000668%
de heer ter vere gij	1	0.000668%
dat Zijn Ed. ver geeten	1	0.000668%
dat ghij mijn vrouwe doch	1	0.000668%
die ons alle ver willecomde	1	0.000668%

Sort by... Export Export for Excel

Statistics (frequency lists)

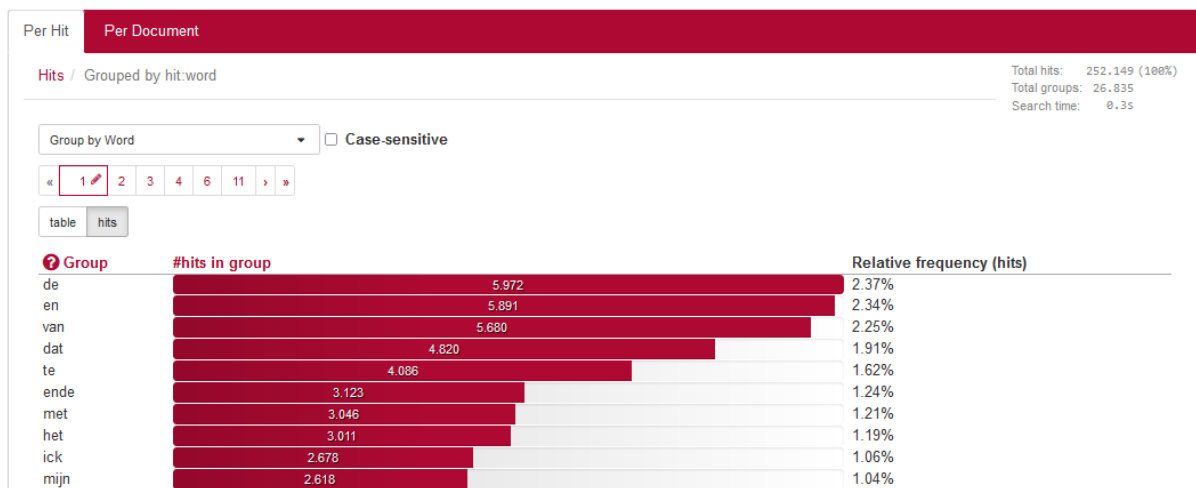
Here, you can produce frequency lists for the corpus. It is rather similar to the previous option, but restricted to 1-grams.

Options

- *Frequency list type*: in this corpus, it is only possible to create frequency lists of Words (i.e. word forms)
- By using the Filter search by... you can create a subcorpus within the *Gekaapte Brieven* for specific metadata.

Example

It is possible to determine the use of the ten most frequently used words in Dutch letters, written between 1650 and 1675 by searching for Frequency list type Word and by filtering search by Language and Text type. This results in:



Appendix: Corpus Query Language

BlackLab supports Corpus Query Language, a full-featured query language introduced by the IMS Corpus WorkBench (CWB) and also supported by the Lexicom Sketch Engine. It is a standard and powerful way of searching corpus.

The basics of Corpus Query Language is the same in all three projects, but there are a few minor differences in some of the more advanced features, as well as some features that are exclusive to some projects. For most queries however, this will not be an issue.

This page will introduce the query language and show all features that BlackLab supports. If you want to learn even more about CQL, see [CWB CQP Query Language Tutorial](#) and [Sketch Engine Corpus Query Language](#).

CQL support

For those who already know CQL, here's a quick overview of the extent of BlackLab's support for this query language. If there is a feature we don't support, yet is important to you, please let us know. If it's quick to add, we may be able to help you out.

Supported features

BlackLab currently supports (arguably) most of the important features of Corpus Query Language:

- Matching on token annotations (also called properties or attributes), using regular expressions and =, !=, !. Example: [word="bank"] (or just "bank")
- Case/accent-sensitive matching. Note that, unlike in CWB, case-INsensitive matching is currently the default. To explicitly match case/accent-insensitivity, use "(?i)...". Example: "(?i)Mr\." "(?i)Banks"
- Combining criteria using &, | and !. Parentheses can also be used for grouping. Example: [lemma="bank" & pos="V"]
- Match-all pattern [] matches any token. Example: "a" [] "day"
- Regular expression operators +, *, ?, {n}, {n,m} at the token level. Example: [pos="AA"]+
- Sequences of token constraints. Example: [pos="AA"] "cow"
- Operators |, & and parentheses can be used to build complex sequence queries. Example: "happy" "dog" | "sad" cat"
- Querying with tag positions using e.g. <s> (start of sentence), </s> (end of sentence), <s/> (whole sentence) or <s> ... </s> (equivalent to <s/> containing ...). Example: <s> "The" . XML attribute values may be used as well, e.g. <ne type="PERS"/> ("named entities that are persons").
- Using within and containing operators to find hits inside another set of hits. Example: "you" "are" within <s/>
- Using an anchor to capture a token position. Example: "big" A:[]. Captured matches can be used in global constraints (see next item) or processed separately later (using the Java interface; capture information is not yet returned by BlackLab Server). Note that BlackLab can actually capture entire groups of tokens as well, similarly to regular expression engines.

- Global constraints on captured tokens, such as requiring them to contain the same word.

Example: "big" A:[] "or" "small" B:[] :: A.word = B.word

See below for features not in this list that may be added soon, and let us know if you want a particular feature to be added.

Differences from CWB

BlackLab's CQL syntax and behaviour differs in a few small ways from CWBs. In future, we'll aim towards greater compliance with CWB's de-facto standard (with some extra features and conveniences).

For now, here's what you should know:

- Case-insensitive search is currently the default in BlackLab, although you can change this if you wish. CWB and Sketch Engine use case-sensitive search as the default. We may change our default in a future major version.
If you want to switch case-/diacritics-sensitivity, use "(?i).." (case-sensitive) or "(?i).." (case-insensitive, usually the default). CWBs %cd flags for setting case/diacritics-sensitivity are not (yet) supported, but will be added.
- If you want to match a string literally, not as a regular expression, use backslash escaping: "e.g.". %l for literal matching is not yet supported, but will be added.
- BlackLab supports result set manipulation such as: sorting (including on specific context words), grouping/frequency distribution, subsets, sampling, setting context size, etc. However, these are supported through the REST and Java APIs, not through a command interface like in CWB. See [BlackLab Server overview](#).
- Querying XML elements and attributes looks natural in BlackLab: <s/> means "sentences", <s> means "starts of sentences", <s type='A'> means "sentence tags with a type attribute with value A". This natural syntax differs from CWBs in some places, however, particularly when matching XML attributes. While we believe our syntax is the superior one, we may add support for the CWB syntax as an alternative.
We only support literal matching of XML attributes at the moment, but this will be expanded to full regex matching.
- In global constraints (expressions occurring after ::), only literal matching (no regex matching) is currently supported. Regex matching will be added soon. For now, instead of A:[] "dog" :: A.word = "happy|sad", use "happy|sad" "dog".
- To expand your query to return whole sentences, use <s/> containing (...). We don't yet support CWBs expand to, expand left to, etc., but may add this in the future.
- The implication operator -> is currently only supported in global constraints (expressions after the :: operator), not in regular token constraints. We may add this if there's demand for it.
- We don't support the @ anchor and corresponding target label; use a named anchor instead. If someone makes a good case for it, we will consider adding this feature.
- backreferences to anchors only work in global constraints, so this doesn't work: A:[] [] [word = A.word]. Instead, use something like: A:[] [] B:[] :: A.word = B.word. We hope to add support for these in the near future, but our matching approach may not allow full support for this in all cases.

(Currently) unsupported features

The following features are not (yet) supported:

- intersection, union and difference operators. These three operators will be added in the future. For now, the first two can be achieved using & and | at the sequence level, e.g. "double" [] & [] "trouble" to match the intersection of these queries, i.e. "double trouble" and "happy" "dog" | "sad "cat" to match the union of "happy dog" and "sad cat".
- _ meaning "the current token" in token constraints. We will add this soon.
- lbound, rbound functions to get the edge of a region. We will probably add these.
- distance, distabs functions and match, matchend anchor points (sometimes used in global constraints). We will see about adding these.
- using an XML element name to mean 'token is contained within', like [(pos = "N") & !np] meaning "noun NOT inside in an tag". We will see about adding these.
- a number of less well-known features. If people ask, we will consider adding them.

Using Corpus Query Language

Matching tokens

Corpus Query Language is a way to specify a "pattern" of tokens (i.e. words) you're looking for. A simple pattern is this one:

```
[word="man"]
```

This simply searches for all occurrences of the word "man". If your corpus includes the per-word properties lemma (i.e. headword) and pos (part-of-speech, i.e. noun, verb, etc.), you can query those as well. For example, to find a form of word "search" used as a noun, use this query:

```
[lemma="search" & pos="NOU-C"]
```

This query would match "search" and "searches" where used as a noun. (Of course, your data may contain slightly different part-of-speech tags.)

The first query could be written even simpler without brackets, because "word" is the default property:

```
"man"
```

You can use the "does not equal" operator (!=) to search for all words except nouns:

```
[pos != "NOU-C"]
```

The strings between quotes can also contain wildcards, of sorts. To be precise, they are [regular expressions](#), which provide a flexible way of matching strings of text. For example, to find "man" or "woman", use:

```
"(wo)?man"
```

And to find lemmata starting with "under", use:

```
[lemma="under.*"]
```

Explaining regular expression syntax is beyond the scope of this document, but for a complete overview, see [here](#).

Sequences

Corpus Query Language allows you to search for sequences of words as well (i.e. phrase searches, but with many more possibilities). To search for the phrase "the tall man", use this query:

```
"the" "tall" "man"
```

It might seem a bit clunky to separately quote each word, but this allows us the flexibility to specify exactly what kinds of words we're looking for. For example, if you want to know all single adjectives used with man (not just "tall"), use this:

```
"an? | the" [pos="AA"] "man"
```

This would also match "a wise man", "an important man", "the foolish man", etc.

Regular expression operators on tokens

Corpus Query Language really starts to shine when you use the regular expression operators on whole tokens as well. If we want to see not just single adjectives applied to "man", but multiple as well:

```
"an? | the" [pos="AA"]+ "man"
```

This query matches "a little green man", for example. The plus sign after [pos="AA"] says that the preceding part should occur one or more times (similarly, * means "zero or more times", and ? means "zero or one time").

If you only want matches with two or three adjectives, you can specify that too:

```
"an? | the" [pos="AA"] {2,3} "man"
```

Or, for two or more adjectives:

```
"an? | the" [pos="AA"] {2,} "man"
```

You can group sequences of tokens with parentheses and apply operators to the whole group as well.

To search for a sequence of nouns, each optionally preceded by an article:

```
("an? | the"? [pos="NOU-C"])+
```

This would, for example, match the well-known palindrome "a man, a plan, a canal: Panama!" (A note about punctuation: in BlackLab, punctuation tends to not be indexed as a separate token, but as a property of a word token - CWB and Sketch Engine on the other hand tend to index punctuation as a separate token instead. You certainly could choose to index punctuation as a separate token in BlackLab, by the way -- it's just not commonly done. Both approaches have their advantages and disadvantages, and of course the choice affects how you write your queries.)

Case- and diacritics-sensitivity

CWB and Sketch Engine both default to (case- and diacritics-)sensitive search. That is, they exactly match upper- and lowercase letters in your query, plus any accented letters in the query as well.

BlackLab, on the contrary, defaults to *IN*sensitive search (although this default can be changed if you like). To match a pattern sensitively, prefix it with "(?-i)":

```
" (?-i) Panama "
```

If you've changed the default search to sensitive, but you wish to match a pattern in your query insensitively, prefix it with "(?i)":

```
[pos=" (?i) NOU-C "]
```

Although BlackLab is capable of setting case- and diacritics-sensitivity separately, it is not yet possible from Corpus Query Language. We may add this capability if requested.

Matching XML elements

Corpus Query Language allows you to find text in relation to XML elements that occur in it. For example, if your data contains sentence tags, you could look for sentences starting with "the":

```
<s>"the"
```

Similarly, to find sentences ending in "that", you would use:

```
"that" </s>
```

You can also search for words occurring inside a specific element. Say you've run named entity recognition on your data and all person names are surrounded with <person>...</person> tags. To find the word "baker" as part of a person's name, use:

```
"baker" within <person/>
```

Note the forward slash at the end of the tag. This way of referring to the element means "the whole element". Compare this to <person>, which means "the element's open tag", and </person>, which means "the element's close tag".

The above query will just match the word "baker" as part of a person's name. But you're likely more interested in the entire name that contains the word "baker". So, to find those full names, use:

```
<person/> containing "baker"
```

Or, if you simply want to find all persons, use:

```
<person/>
```

As you can see, the XML element reference is just another query that yields a number of matches. So as you might have guessed, you can use "within" and "containing" with any other query as well. For example:

```
( [pos="AA"]+ containing "tall" ) "man"
```

will find adjectives applied to man, where one of those adjectives is "tall".

Labeling tokens, capturing groups

Just like in regular expressions, it is possible to "capture" part of the match for your query in a "group".

CWB and Sketch Engine offer similar functionality, but instead of capturing part of the query, they label a single token. BlackLab's functionality is very similar but can capture a number of tokens as well. For example:

```
"an?|the" Adjectives: [pos="AA"]+ "man"
```

This will capture the adjectives found for each match in a captured group named "Adjectives".

BlackLab also supports numbered groups:

```
"an?|the" 1: [pos="AA"]+ "man"
```

Global constraints

If you tag certain tokens with labels, you can also apply "global constraints" on these tokens. This is a way of relating different tokens to one another, for example requiring that they correspond to the same word:

```
A: [] "by" B: [] :: A.word = B.word
```

This would match "day by day", "step by step", etc.