# *Gekaapte Brieven* application manual

## Table of contents

# 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 (https://blacklab.ivdnt.org/). The web-based frontend is a further development of the corpus-frontend application developed by INT (https://github.com/instituutnederlandsetaal/blacklab-frontend). Its design is inspired by the first version of the OpenSoNaR user interface by Tilburg University 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* has 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.

#### *Permissive / Strict*

It is possible to do a permissive or a strict search for Year.

## 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

The language of the corpus application is set to Dutch by default. Press the globe icon in the top right corner to select English.

## 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 or Extended search.
  - Simple Search for Word *schip*
  - Extended Search for Word *schip*
- To get an approximation for different spellings and inflections/conversions of a given word, use Extended Search (the option Select all corresponding to a lemma in the lexicon).
  - 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 Advanced Search and 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 (mostly) one syllable in between in Advanced Search
  - words starting with *ver* and ending in *eren* with (mostly) one syllable in between in Expert Search
- To see which unique forms occur as a result of your search, use Group Results.
  - example Group by Annotation: different words following *lieve*
  - example Group by Annotation: different words preceding the word *huis*
- To explore the distribution of document properties in the corpus, use the Explore feature.
  - example: characteristics of the text types
  - example: sender

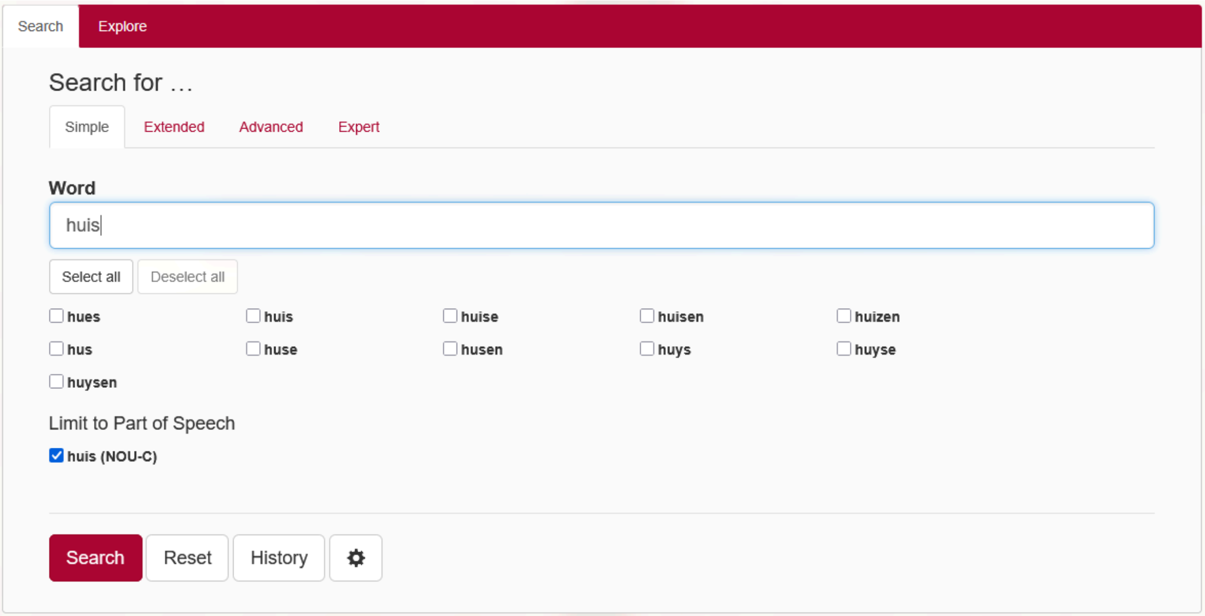
# Searching the corpus

## Simple search

### Search

The Simple Search allows you to quickly search for specific word forms (e.g. *huis*). 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. 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.

Based on the information in this lexicon all spelling variants of the search term found are suggested (see the 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 the historic component of the GiGaNT-lexicon in connection to the search term.



Word: huis

| | | | | |
|---|---|---|---|---|
| □ hues | □ huis | □ huise | □ huisen | □ huizen |
| □ hus | □ huse | □ husen | □ huys | □ huyse |
| □ huysen | | | | |

Limit to Part of Speech

☑ huis (NOU-C)

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

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, or go to Extended Search.

### 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, searching for *a*n* matches all word forms 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* and *Arztin*.

Note that searching with wildcards is limited to Simple Search and Extended Search. [In Advanced Search and Expert Search you can use so-called regular expressions instead of wildcards.]

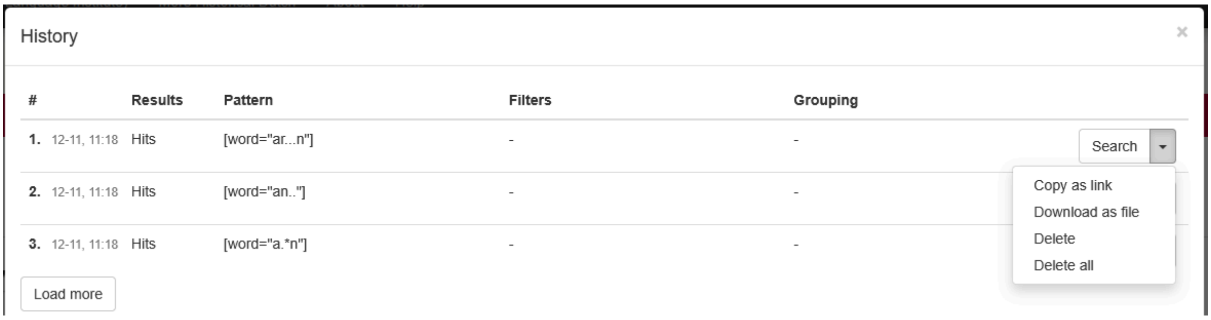
### Reset

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.

### 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).



History

| # | Results | Pattern | Filters | Grouping |
|---|---|---|---|---|
| 1. 12-11, 11:18 | Hits | [word="ar...n"] | - | - |
| 2. 12-11, 11:18 | Hits | [word="an.."] | - | - |
| 3. 12-11, 11:18 | Hits | [word="a.*n"] | - | - |

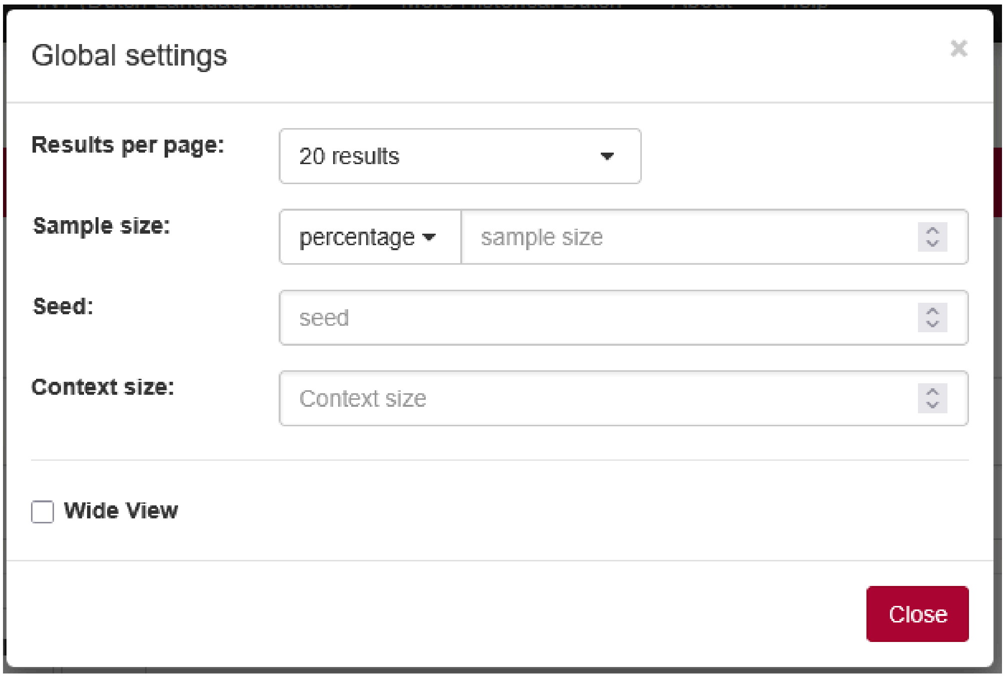
Every search query has its own url. If you copy this url via History (Copy as link) or directly from the address bar of your browser, you can send it to someone else who can import this link via Import from a link. It offers that person the possibility to run the search on his or her own computer.

### 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. 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.



Global settings

Results per page: 20 results

Sample size: percentage sample size

Seed: seed

Context size: Context size

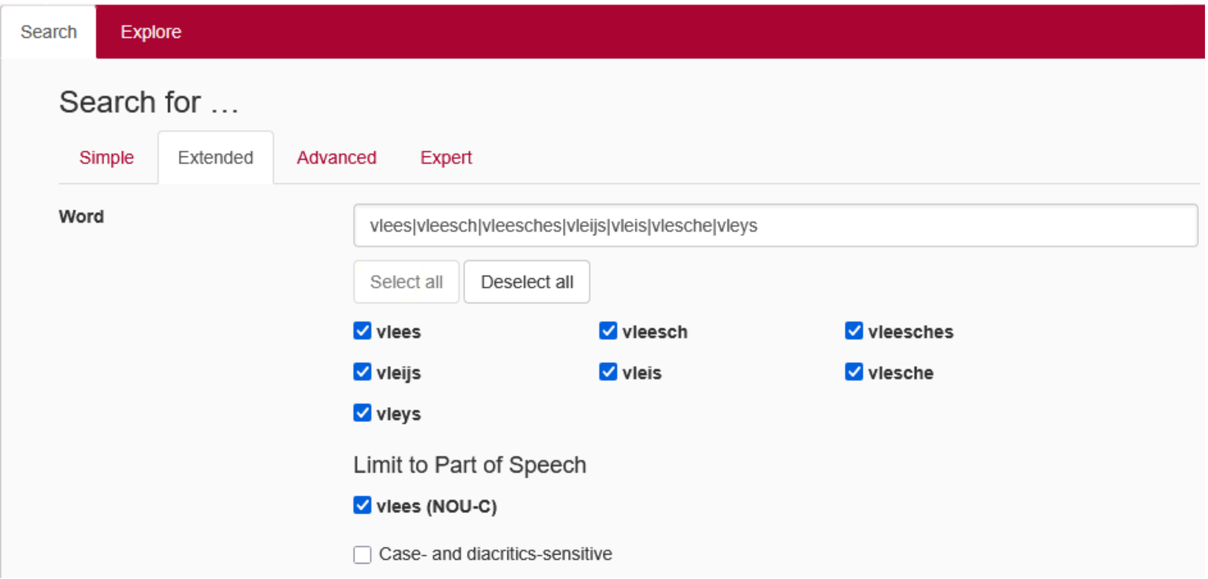
□ Wide View

## Extended search

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

After entering a search term in the search field Word, 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. 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, you can also enter a phrase here.

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



Word: vlees|vleesch|vleesches|vleijs|vleis|vlesche|vleys

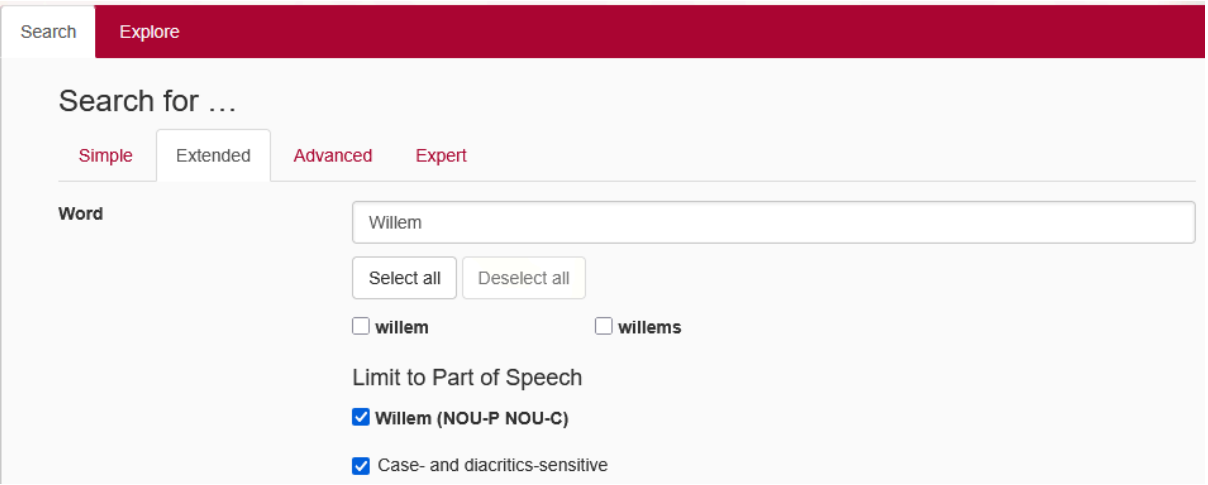
☑ vlees ☑ vleesch ☑ vleesches
☑ vleijs ☑ vleis ☑ vlesche
☑ vleys

Limit to Part of Speech

☑ vlees (NOU-C)

☐ Case- and diacritics-sensitive

Extended Search allows 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 none of *Willem*. In order to directly find only occurrences of the Word (form) *Willem* (132x), use the search term *Willem* and tick the box Case- and diacritics-sensitive under the search field Word (as shown below).



Word: Willem

☐ willem ☐ willems

Limit to Part of Speech

☑ Willem (NOU-P NOU-C)

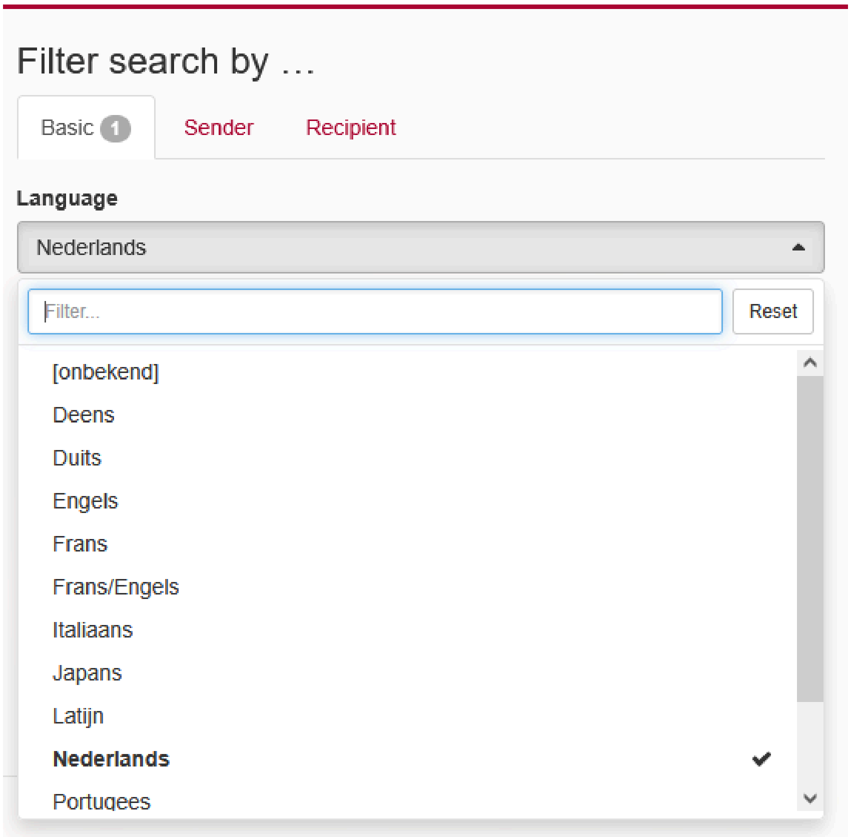
☑ Case- and diacritics-sensitive

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

## 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*). To view the results for all documents, simply leave the attributes in the filtering form empty.

There are two different ways to specify a filter, depending on the field type. Most fields allow you to choose one or more values from a drop-down list, while Year (range) (see below) and Name allow you to fill in the value(s) yourself. 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. 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 search by …

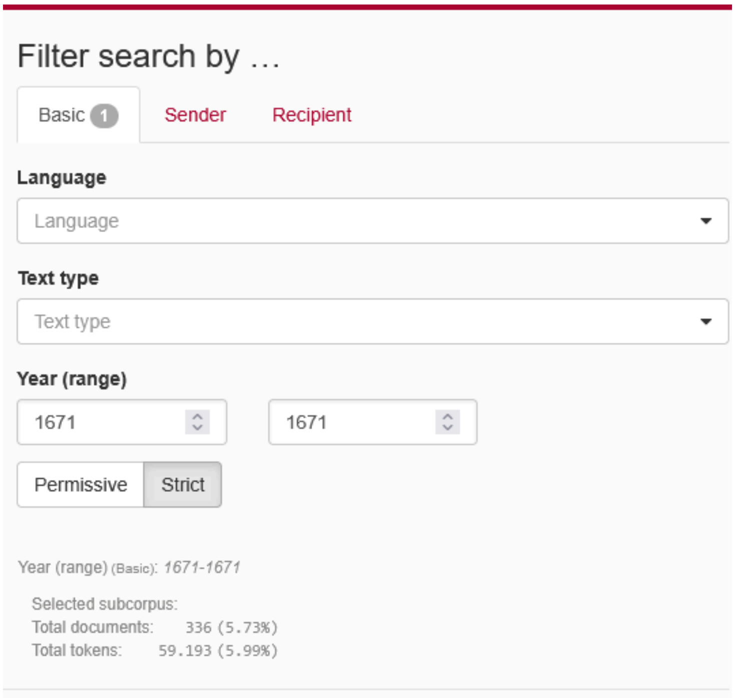
Language: Nederlands

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.

When on the other hand the set of possible values is rather large (e.g. Name), 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*.

### 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 specific year, the entire corpus is searched. If you want to filter by another year or another period, please press the reset button.



Year (range): 1671 – 1671

Year (range) (Basic): *1671-1671*

Selected subcorpus:
Total documents: 336 (5.73%)
Total tokens: 59.193 (5.99%)

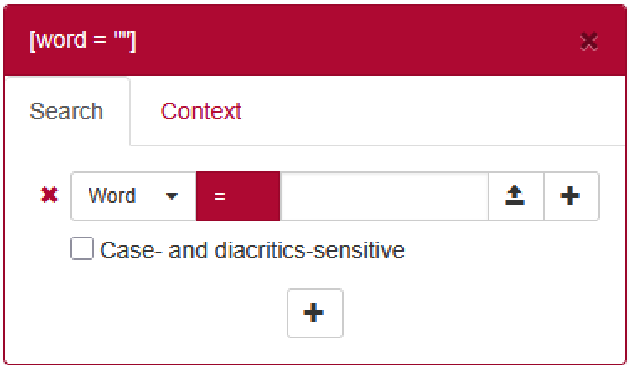
For a detailed description of the metadata, see the section Metadata categories.

## Advanced search

### The query builder

The basic building block in the query builder is the *token box* (see below). Each box represents a token – usually just a single word – or a simple repetition of tokens; when multiple tokens are used, they are matched in order from left to right.

You can use the query builder to create complex queries without writing CQL (here: Corpus Query Language). Therefore, it is easy to use.



[word = ""]

☐ Case- and diacritics-sensitive

A token box in the querybuilder has two tabs: Search and Context.

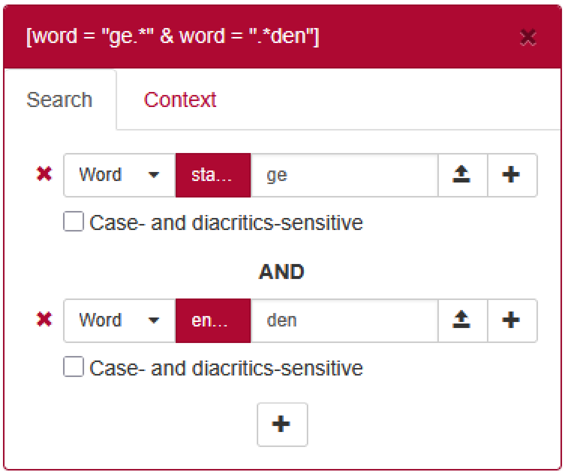
### The tab Search

The tab Search contains a set of attributes a token in the corpus must have to be matched by the query. By clicking the +-button on the right hand side of this token, you can add new attributes (see below).

Then enter a value that the attribute must have for the token to be found. The search command Word 'starts with' *ge* and Word 'ends with' *den* for example results in both verbal forms (*gesonden*, *gesonden*, *gehouden*) and plural nouns (*gesontheden*, *geliefden*).

It is only possible to search by word forms. However, you can specify whether that word form should be equal or not equal to the entered search term. You can also specify whether or not a word should begin or end with a particular character combination.

The CQL query generated to match this token (the *token query*) in the corpus is displayed in the top bar of the box, to help you understand what is happening internally. The following applies to our example:



[word = "ge.*" & word = ".*den"]

✖ Word ▾ | sta… | ge

□ Case- and diacritics-sensitive

AND

✖ Word ▾ | en… | den

□ Case- and diacritics-sensitive

*Token attributes*

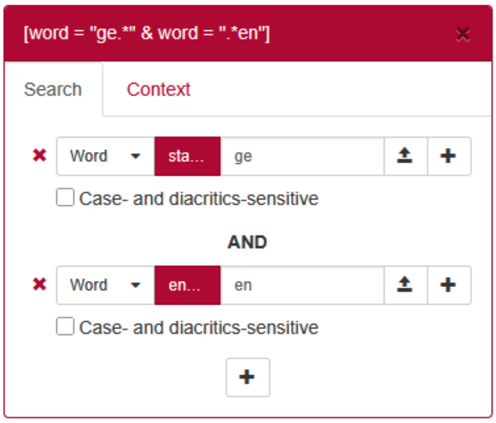
Specifying token attributes is similar to the Extended Search form. Select which attribute a token should have, and enter the value that the attribute must have for the token to be matched. Attributes in the query builder are interpreted as *regular expressions*. Note that this is different from the Extended Search, where token patterns use wildcards.

Going beyond single-attribute token queries, a token box also allows you to combine several attributes and to specify repetition options.

*Adding attributes to a token box*

Using the +-button, new attributes can be added. Two options exist: *AND* and *OR*.

The *AND* option creates a new attribute restriction that a token must match in addition to the ones which were already there. As an example: suppose we want to match past participles of strong verbs. First, fill in the attribute Word 'starts with' *ge*, then click +, choose *AND*, and choose Word 'ends with' *en*.



[word = "ge.*" & word = ".*en"]

✖ Word ▾ | sta… | ge

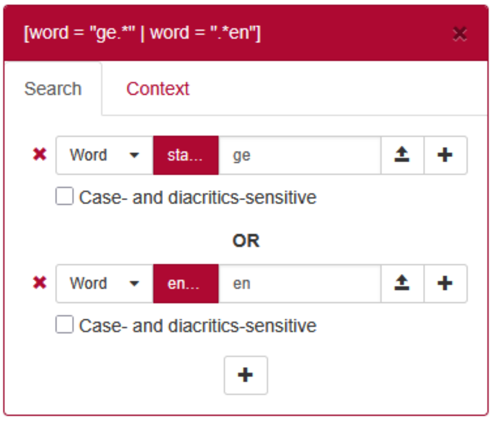
□ Case- and diacritics-sensitive

AND

✖ Word ▾ | en… | en

□ Case- and diacritics-sensitive

Similarly, creating a new attribute using *OR* will create a token query matching tokens that have the original attribute *or* the new attribute. For instance, enter Word 'starts with' *ge*, add a new attribute with the *OR* option and enter Word 'ends with' *en* to match tokens as *gelieven, geschreven* and *laten, mulen*.


[word = "ge.*" | word = ".*en"]

Word sta... ge

OR

Word en... en

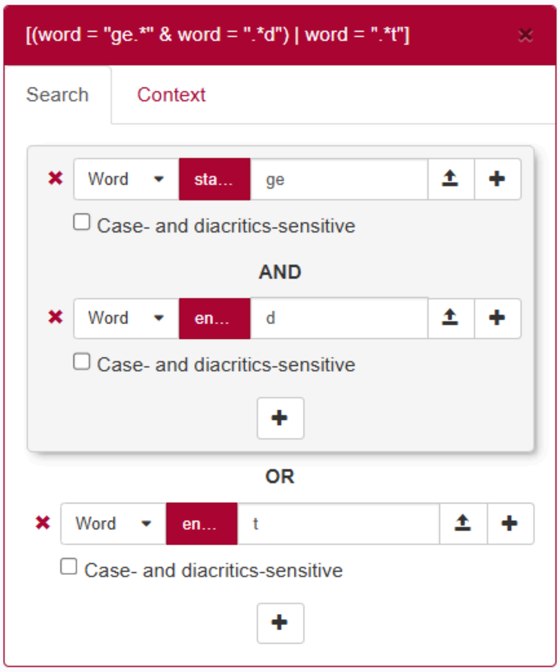
Search
Context
□ Case- and diacritics-sensitive
□ Case- and diacritics-sensitive

*Function of the two +-buttons in a token box*

The difference between the +-sign on the right of an attribute and the one below it, is that the +-sign on the right keeps the newly added attribute 'within a subclause'. This is most easily explained by means of an example.

Suppose we want to look for past participles of weak verbs, i.e. verbs that end in an *-d* or a *-t*. If we add the attributes in the order Word 'starts with' *ge* AND Word 'ends with' *d*, OR Words 'ends with' *t* using the +-signs **below** the attributes, as in the left screenshot below, we get the token query [(word = "ge.*" & word = ".*d") | word = ".*t"]. This will also match forms such as *stadt, niet*, so this is not what we were after.

If, on the other hand, we add OR Word 'ends with' *t* with the +-sign to the **right** of the attribute Word 'ends with' *d*, it will be inserted in a subclause, thus resulting in the correct query [word = "ge.*" & (word = ".*d" | word = ".*t")], as shown in the right screenshot below.


[(word = "ge.*" & word = ".*d") | word = ".*t"]

Word sta... ge

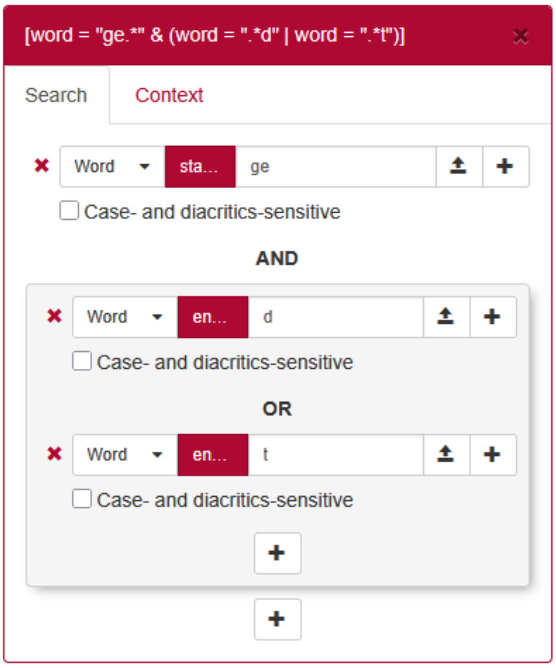
AND

Word en... d

OR

Word en... t

Search
Context
□ Case- and diacritics-sensitive
□ Case- and diacritics-sensitive
□ Case- and diacritics-sensitive


[word = "ge.*" & (word = ".*d" | word = ".*t")]

Word sta... ge

AND

Word en... d

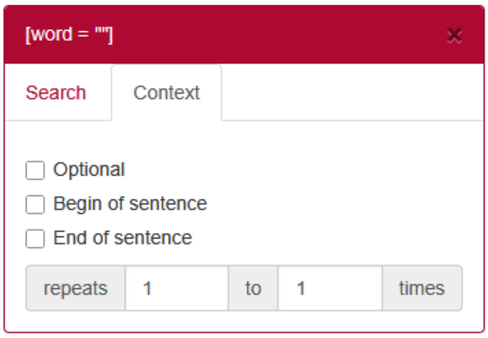
OR

Word en... t

Search
Context
□ Case- and diacritics-sensitive
□ Case- and diacritics-sensitive
□ Case- and diacritics-sensitive

### The tab Context

The tab Context specifies the contextual properties, such as whether the token occurs at the end of a sentence, and the repetition pattern:

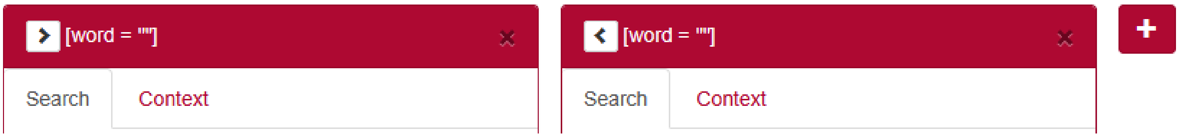


### Managing sequences of token boxes

There are three ways to manage the sequence and the number of token boxes:

- *Rearrange* a token by clicking on the arrow in the top-left corner of a box (1). This arrow only appears if there are multiple token boxes.
- *Delete* a token by clicking the x in the top-right corner of a box (2).
- *Create a new token box* by clicking the +-button next to the upper right corner of the utmost right token box (3).

↓ (1) ↓ (2) ↓ (1) ↓ (2) ↓ (3)



### Uploading value lists in the query builder

It is also possible to upload a list of values, separated by a white space. To do so, click the upload button (with the arrow pointing upwards) and select a text file. Tokens will then be matched for any of the values from the file.

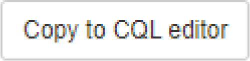
Note that this function only works for *.txt-files. If you are using a text editor like Word, you have to save your file as a *.txt file or you can copy and paste the values into a *.txt file first.

After uploading a file, the text can be edited by clicking the yellow marked file name in the text field. Editing the text is temporary and will not modify your original file.

To remove an uploaded file and go back to typing a value, click on the cross (x) next to the yellow text box. Another possibility to clear the uploaded values is by clicking the yellow marked text field and then pressing the Clear button on the bottom left corner of the Edit box. Using the Reset button will start a complete new search.

### Copy to CQL editor

You can use the query builder to create complex queries without writing CQL. Any time a query is created in the querybuilder, it can be copied to the CQL editor, where you can further edit the query by hand. This will take you automatically to the Expert Search screen, after which you can start the search or adjust the query if desired.



## Expert search

The Corpus Query Language (CQL) editor allows you to type your own CQL query, to copy your query into the query builder (in Advanced Search), 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"][]{3}[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*. Note that, while most queries up to this point could also have been constructed with the query builder, we really need the power of CQL from here on.

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.

### Copy to query builder

When the query is relatively simple – like [word="schip"][word="den"] – it can also be imported into the querybuilder using the *Copy to query builder* button. This will take you automatically to the Advanced Search screen, after which you can start the search or adjust the query if desired.

A message will be displayed next to the button if the query couldn't be parsed.

### 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

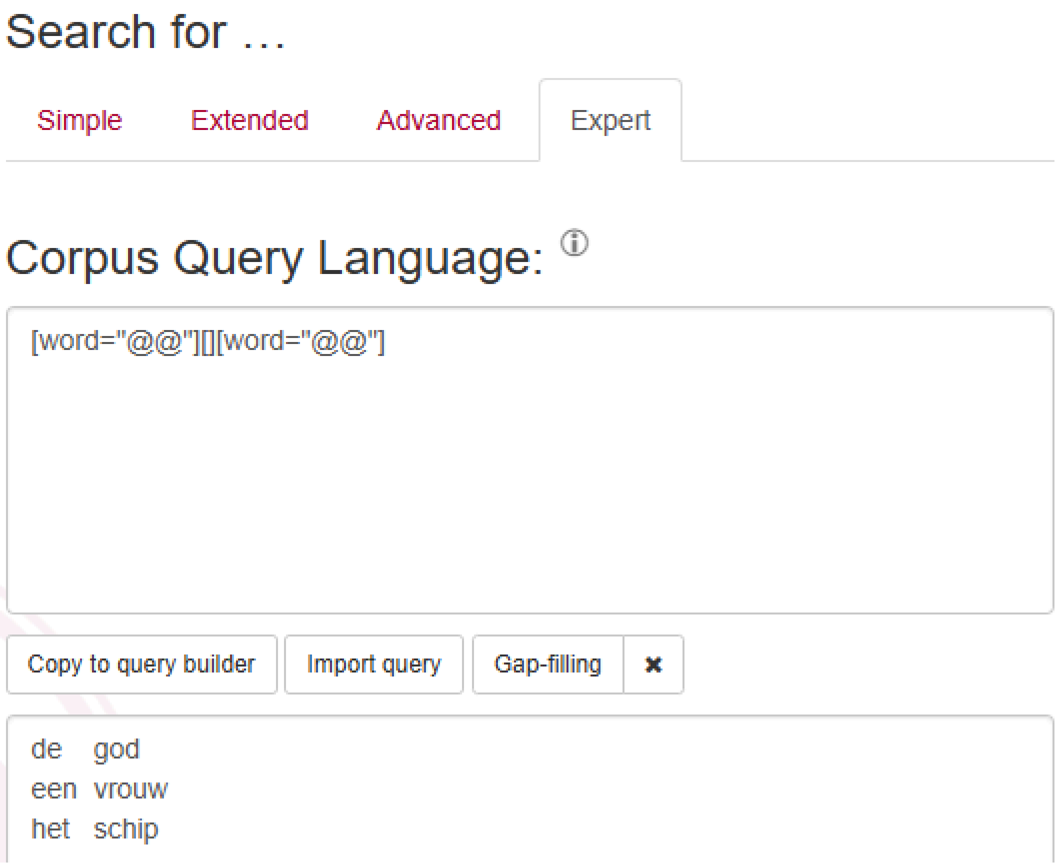
Use this button to upload a Tab Separated Values (TSV) file, which is a simple text format for storing data in a tabular structure. Each record in the table is one line of the text file. Each field value of a record is separated from the next by a tab character. It is also possible to upload a plain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

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="@@"]
```

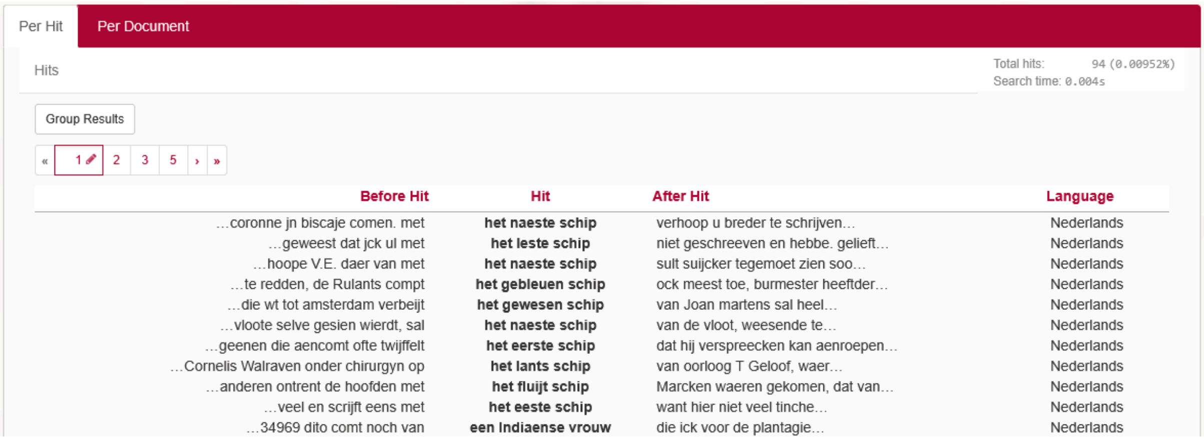
By clicking Gap-filling you can upload a file with a tab-separated list of values from your computer to substitute them for the gap values, i.e. the at signs (@@) in your query. After the upload your values will appear in a separate box:



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

|  |  |
|---|---|
| de | god |
| een | vrouw |
| het | schip |

The values in the first column – *de, een, het* – 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):



Total hits: 94 (0.00952%)
Search time: 0.004s

| Before Hit | Hit | After Hit | Language |
|---|---|---|---|
| …coronne jn biscaje comen. met | **het naeste schip** | verhoop u breder te schrijven… | Nederlands |
| …geweest dat jck ul met | **het leste schip** | niet geschreeven en hebbe. gelieft… | Nederlands |
| …hoope V.E. daer van met | **het naeste schip** | sult suijcker tegemoet zien soo… | Nederlands |
| …te redden, de Rulants compt | **het gebleuen schip** | ock meest toe, burmester heeftder… | Nederlands |
| …die wt tot amsterdam verbeijt | **het gewesen schip** | van Joan martens sal heel… | Nederlands |
| …vloote selve gesien wierdt, sal | **het naeste schip** | van de vloot, weesende te… | Nederlands |
| …geenen die aencomt ofte twijffelt | **het eerste schip** | dat hij verspreecken kan aenroepen… | Nederlands |
| …Cornelis Walraven onder chirurgyn op | **het lants schip** | van oorloog T Geloof, waer… | Nederlands |
| …anderen ontrent de hoofden met | **het fluijt schip** | Marcken waeren gekomen, dat van… | Nederlands |
| …veel en scrijft eens met | **het eeste schip** | want hier niet veel tinche… | Nederlands |
| …34969 dito comt noch van | **een Indiaense vrouw** | die ick voor de plantagie… | Nederlands |

This mimics the functionality to upload a list of values in the Advanced Search interface.

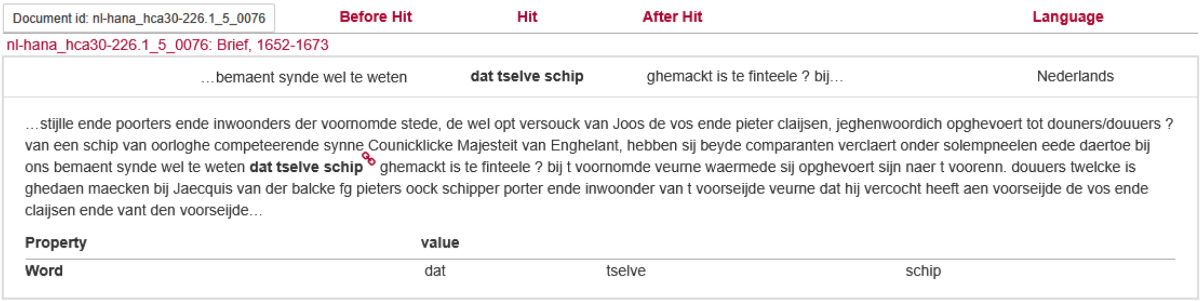
Please note that for this to work, you do need to enter @@ in the field where you want the substitution to take place. An empty field ([]) will match any term.

## 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 finteele ? bij… | Nederlands |

…stijlle ende poorters ende inwoonders der voornomde stede, de wel opt versouck van Joos de vos ende pieter claijsen, jeghenwoordich opghevoert tot douners/douuers ? van een schip van oorloghe competeerende synne Counicklicke Majesteit van Enghelant, hebben sij beyde comparanten verclaert onder solempneelen eede daertoe bij ons bemaent synde wel te weten **dat tselve schip** ghemackt is te finteele ? bij t voornomde veurne waermede sij opghevoert sijn naer t voorenn. douuers 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 claijsen ende vant den voorseijde…

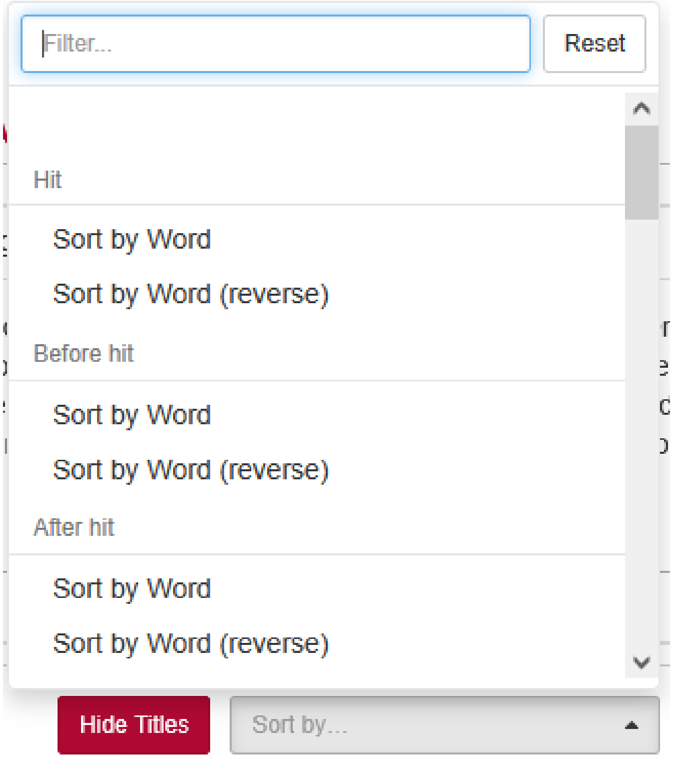
| Property | value | | |
|---|---|---|---|
| Word | dat | tselve | schip |

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

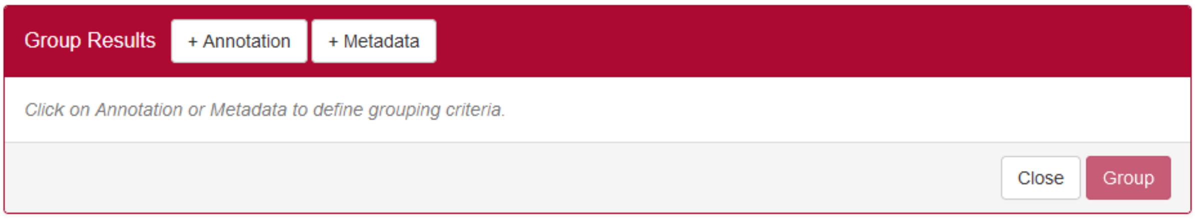
Click on any of the column headings to sort the hits on Words within that column, clicking again inverts the sorting.

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 for Hit, Before hit, After hit, Basic, Sender and Recipient.



### Grouping results

It is possible to group the results by clicking on the button Group Results, after which the following menu appears:



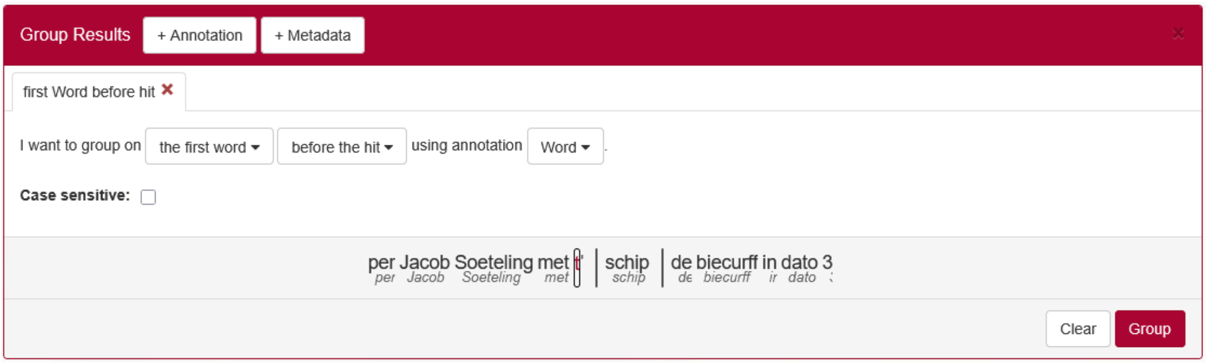
*Click on Annotation or Metadata to define grouping criteria.*

Results can be grouped by Annotation and by Metadata.

By clicking +Annotation you can group by the first word, by all words or by specific words, whether before the hit, within the hit or after the hit, and based on the annotation Word. When grouping by the first word or specific words, you can also group from the end of the hit. The default grouping is grouping all words within the hit using annotation Word. Clicking +Metadata allows you to group by metadata assigned to the document (Basic, Sender, Recipient).

By clicking the Case sensitive box it is possible to distinguish between case sensitive and case insensitive.

The example below is grouped by the first word before the hit. The example dynamically updates when the grouping options are changed.

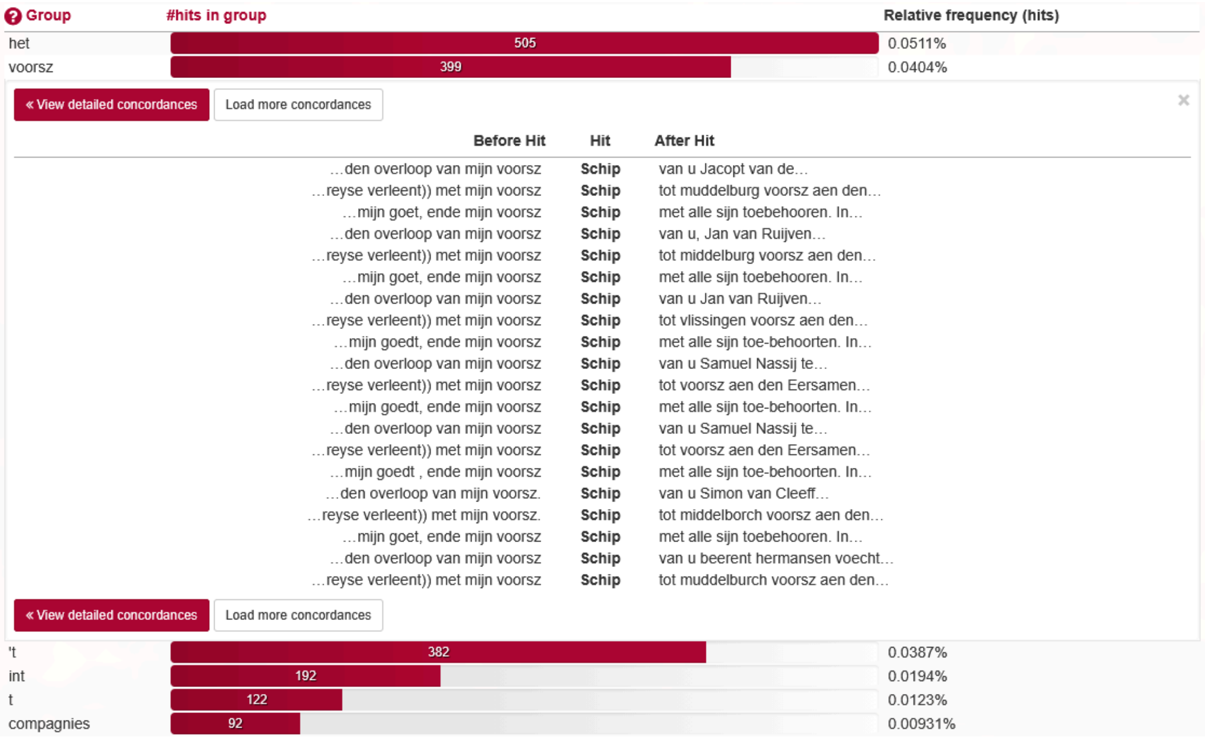


I want to group on the first word ▾ before the hit ▾ using annotation Word ▾ .

**Case sensitive:** □

per Jacob Soeteling met t' | schip | de biecurff in dato 3

Click a group to show or hide hits within that group, as shown below. Click once more on the group to close it again. If more than twenty hits are found in a document, you can make them appear by clicking on Load more concordances.



| Group | #hits in group | Relative frequency (hits) |
|---|---|---|
| het | 505 | 0.0511% |
| voorsz | 399 | 0.0404% |

| Before Hit | Hit | After Hit |
|---|---|---|
| …den overloop van mijn voorsz | Schip | van u Jacopt van de… |
| …reyse verleent)) met mijn voorsz | Schip | tot muddelburg voorsz aen den… |
| …mijn goet, ende mijn voorsz | Schip | met alle sijn toebehooren. In… |
| …den overloop van mijn voorsz | Schip | van u, Jan van Ruijven… |
| …reyse verleent)) met mijn voorsz | Schip | tot middelburg voorsz aen den… |
| …mijn goet, ende mijn voorsz | Schip | met alle sijn toebehooren. In… |
| …den overloop van mijn voorsz | Schip | van u Jan van Ruijven… |
| …reyse verleent)) met mijn voorsz | Schip | tot vlissingen voorsz aen den… |
| …mijn goedt, ende mijn voorsz | Schip | met alle sijn toe-behoorten. In… |
| …den overloop van mijn voorsz | Schip | van u Samuel Nassij te… |
| …reyse verleent)) met mijn voorsz | Schip | tot voorsz aen den Eersamen… |
| …mijn goedt, ende mijn voorsz | Schip | met alle sijn toe-behoorten. In… |
| …den overloop van mijn voorsz | Schip | van u Samuel Nassij te… |
| …reyse verleent)) met mijn voorsz | Schip | tot voorsz aen den Eersamen… |
| …mijn goedt , ende mijn voorsz | Schip | met alle sijn toe-behoorten. In… |
| …den overloop van mijn voorsz. | Schip | van u Simon van Cleeff… |
| …reyse verleent)) met mijn voorsz. | Schip | tot middelborch voorsz aen den… |
| …mijn goet, ende mijn voorsz | Schip | met alle sijn toebehooren. In… |
| …den overloop van mijn voorsz | Schip | van u beerent hermansen voecht… |
| …reyse verleent)) met mijn voorsz | Schip | tot muddelburch voorsz aen den… |

| Group | #hits in group | Relative frequency (hits) |
|---|---|---|
| 't | 382 | 0.0387% |
| int | 192 | 0.0194% |
| t | 122 | 0.0123% |
| compagnies | 92 | 0.00931% |

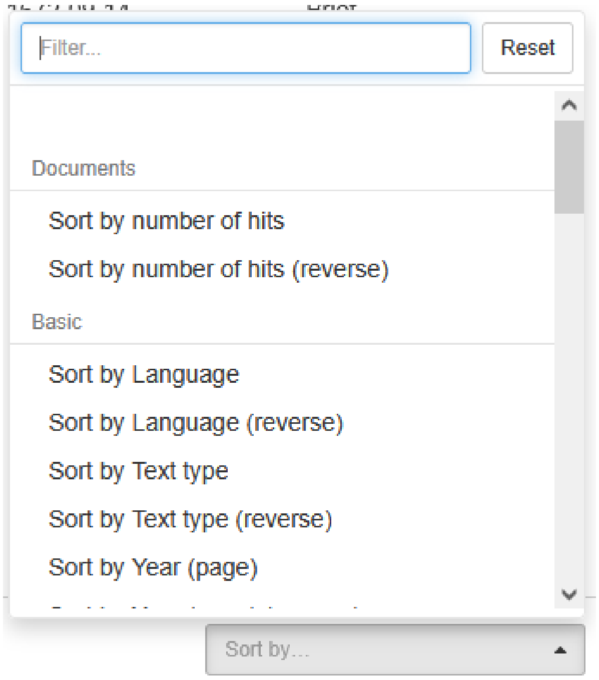
Click on View detailed concordances to go back to the normal hits view to see more detailed information for the hits in this group. The button Go back to grouped results brings you back to the list of groups.

## Per Document view

### Sorting results

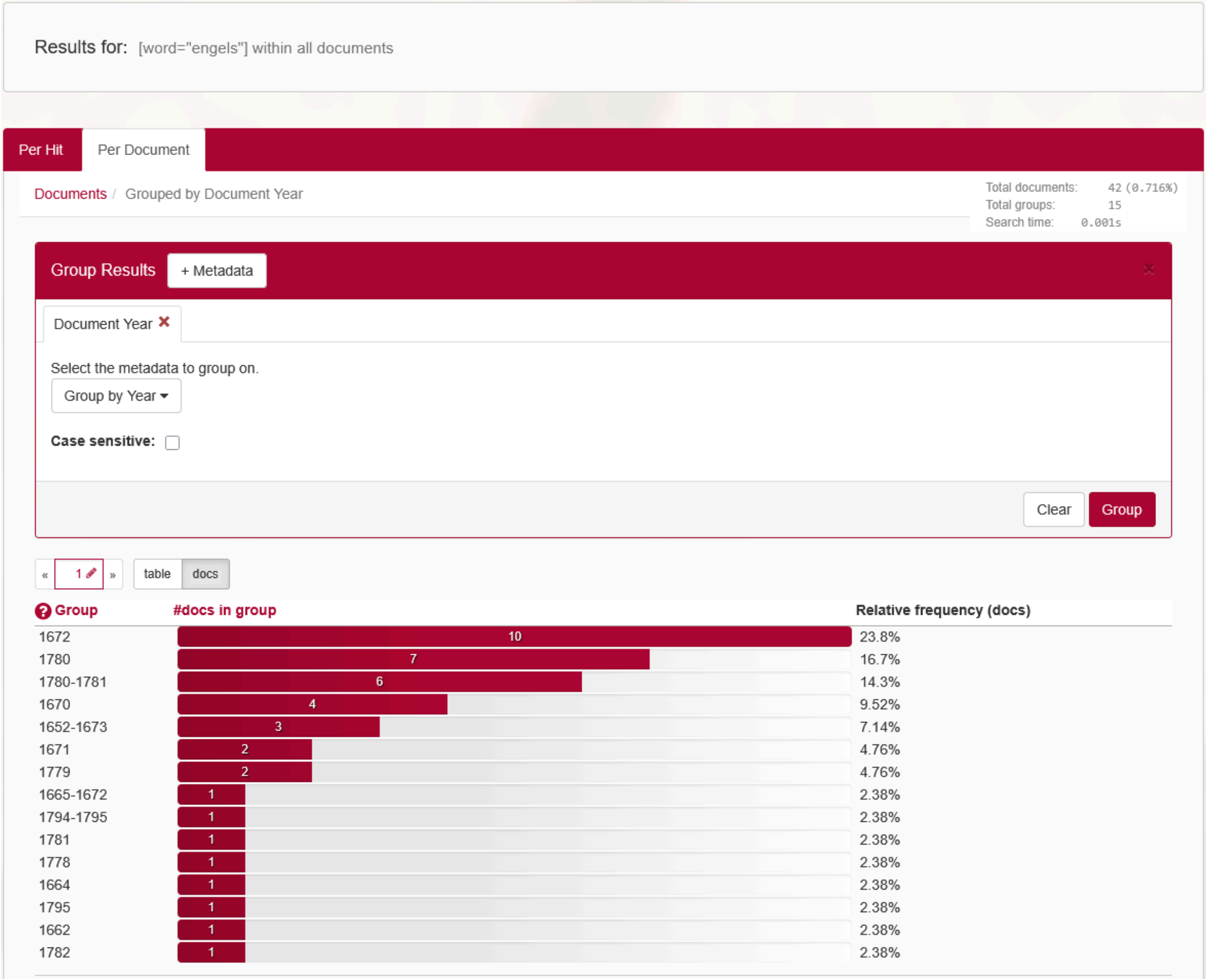
Click on any of the column headings to sort the documents by Document (name), Date, Text Type, Language or Hits within that column, clicking again inverts the sorting.

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 such as Hit (Documents), Language (Basic) and Name (Sender or Recipient).



### Grouping results

Results Per Document can be grouped by metadata assigned to the document (Basic (e.g. Language, Text type, Year), Sender or Recipient). The example below shows all documents in which the Word *engels* occurs grouped by year.



Results for: [word="engels"] within all documents

Total documents: 42 (0.716%)
Total groups: 15
Search time: 0.001s

Select the metadata to group on.
Group by Year ▾

Case sensitive: □

| Group | #docs in group | Relative frequency (docs) |
|---|---|---|
| 1672 | 10 | 23.8% |
| 1780 | 7 | 16.7% |
| 1780-1781 | 6 | 14.3% |
| 1670 | 4 | 9.52% |
| 1652-1673 | 3 | 7.14% |
| 1671 | 2 | 4.76% |
| 1779 | 2 | 4.76% |
| 1665-1672 | 1 | 2.38% |
| 1794-1795 | 1 | 2.38% |
| 1781 | 1 | 2.38% |
| 1778 | 1 | 2.38% |
| 1664 | 1 | 2.38% |
| 1795 | 1 | 2.38% |
| 1662 | 1 | 2.38% |
| 1782 | 1 | 2.38% |

## 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. 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.

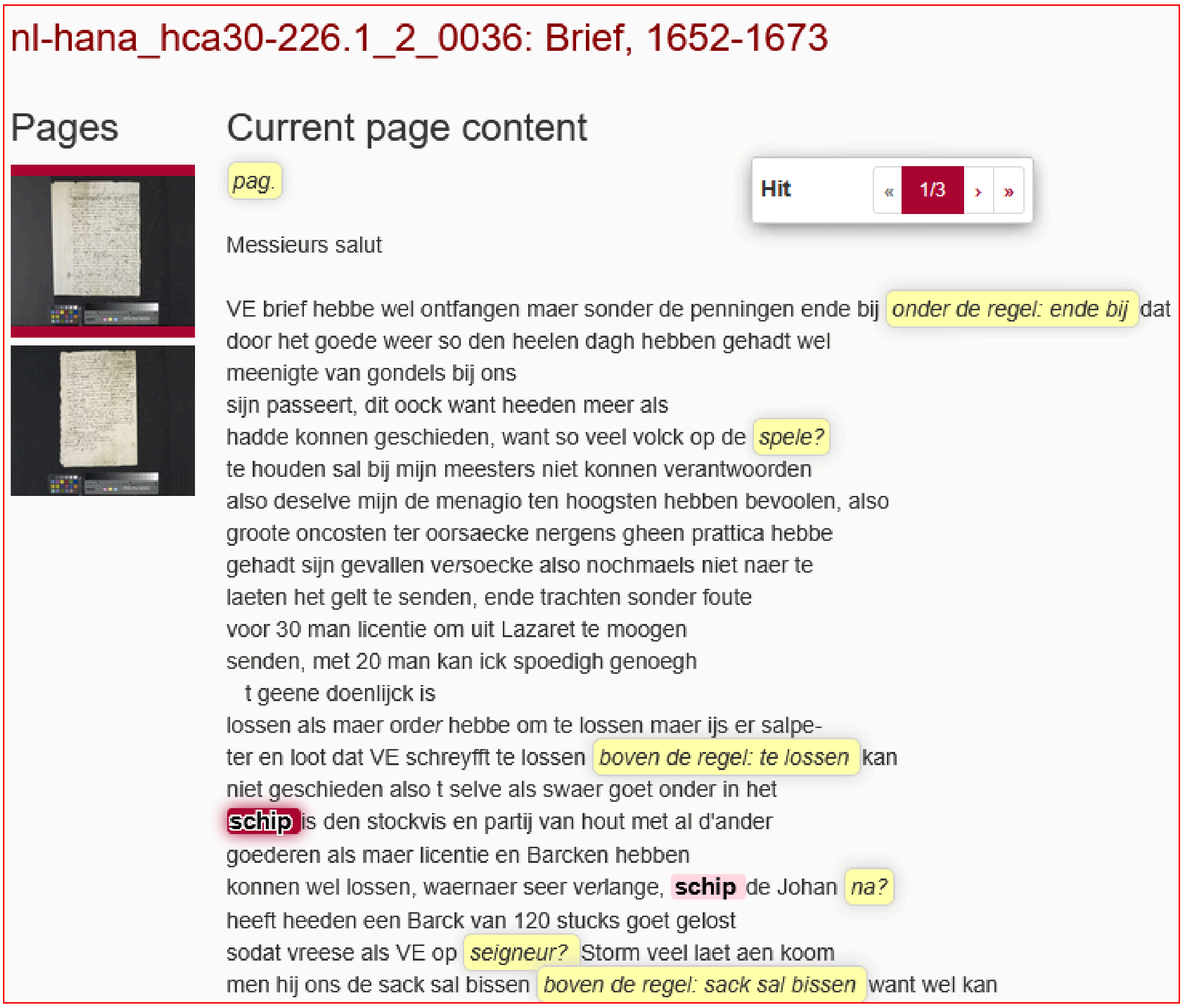
## Information about a document

Click on a document title or the chain icon in the per hit view 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.

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 (such as *schip* in the example below). You can navigate from one hit to another by using the arrows at the Hits button (this button can be dragged around):



nl-hana_hca30-226.1_2_0036: Brief, 1652-1673

Pages

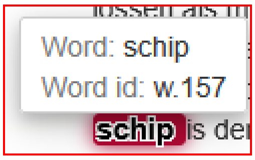
Current page content

*pag.*

Messieurs salut

VE brief hebbe wel ontfangen maer sonder de penningen ende bij *onder de regel: ende bij* dat
door het goede weer so den heelen dagh hebben gehadt wel
meenigte van gondels bij ons
sijn passeert, dit oock want heeden meer als
hadde konnen geschieden, want so veel volck op de *spele?*
te houden sal bij mijn meesters niet konnen verantwoorden
also deselve mijn de menagio ten hoogsten hebben bevoolen, also
groote oncosten ter oorsaecke nergens gheen prattica hebbe
gehadt sijn gevallen versoecke also nochmaels niet naer te
laeten het gelt te senden, ende trachten sonder foute
voor 30 man licentie om uit Lazaret te moogen
senden, met 20 man kan ick spoedigh genoegh
t geene doenlijck is
lossen als maer order hebbe om te lossen maer ijs er salpe-
ter en loot dat VE schreyfft te lossen *boven de regel: te lossen* kan
niet geschieden also t selve als swaer goet onder in het
**schip** is den stockvis en partij van hout met al d'ander
goederen als maer licentie en Barcken hebben
konnen wel lossen, waernaer seer verlange, **schip** de Johan *na?*
heeft heeden een Barck van 120 stucks goet gelost
sodat vreese als VE op *seigneur?* Storm veel laet aen koom
men hij ons de sack sal bissen *boven de regel: sack sal bissen* want wel kan

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:



Word: schip
Word id: w.157

**schip** is de

### Metadata

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

### Statistics

The Statistics tab shows several document statistics: the number of Tokens, Types (unique word forms) 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. You can also zoom with the scroll button on your mouse. 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

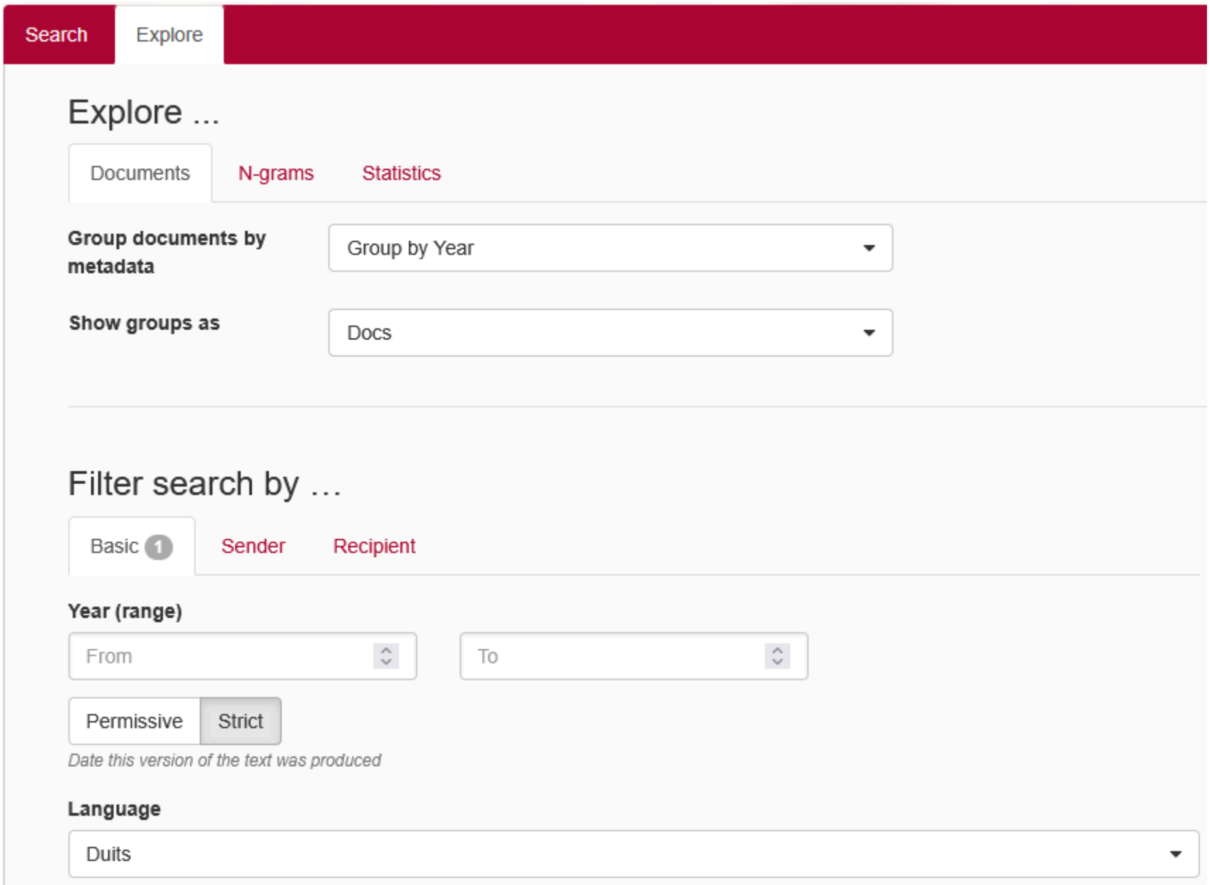
The Explore tab has three subdivisions: Documents, N-grams and Statistics.

## 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 the year distribution of texts written in German (in Dutch: *Duits*) in the *Gekaapte Brieven*.

- In the Group documents by metadata drop-down menu, choose Group by Year
- In Show groups as, select *Docs*
- In the metadata search form (Filter search by), select in Basic Language *Duits*
- Press Search



Explore …

Group documents by metadata: Group by Year

Show groups as: Docs

Filter search by …

Year (range)

From

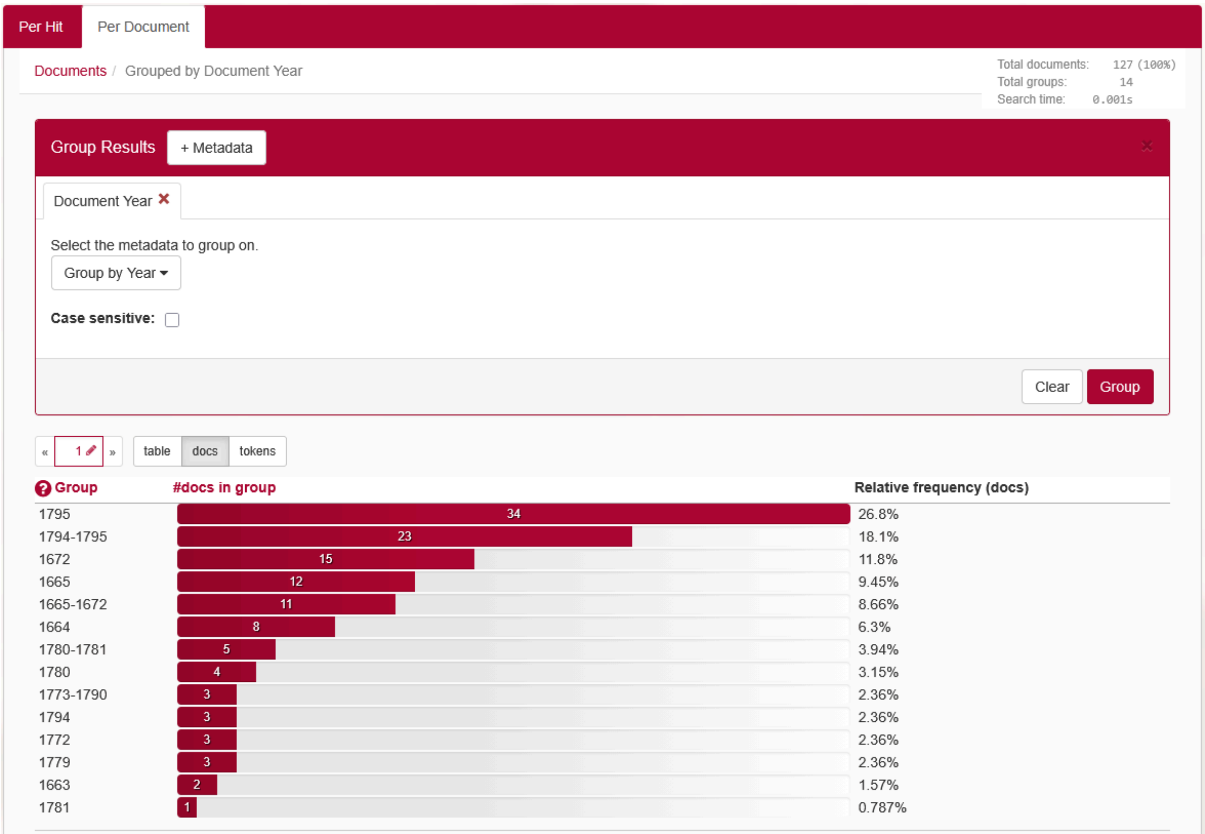
To

Permissive Strict

*Date this version of the text was produced*

Language: Duits

You will get this result:



Documents / Grouped by Document Year

Total documents: 127 (100%)
Total groups: 14
Search time: 0.001s

Select the metadata to group on.

Group by Year

Case sensitive: □

| Group | #docs in group | Relative frequency (docs) |
|---|---|---|
| 1795 | 34 | 26.8% |
| 1794-1795 | 23 | 18.1% |
| 1672 | 15 | 11.8% |
| 1665 | 12 | 9.45% |
| 1665-1672 | 11 | 8.66% |
| 1664 | 8 | 6.3% |
| 1780-1781 | 5 | 3.94% |
| 1780 | 4 | 3.15% |
| 1773-1790 | 3 | 2.36% |
| 1794 | 3 | 2.36% |
| 1772 | 3 | 2.36% |
| 1779 | 3 | 2.36% |
| 1663 | 2 | 1.57% |
| 1781 | 1 | 0.787% |

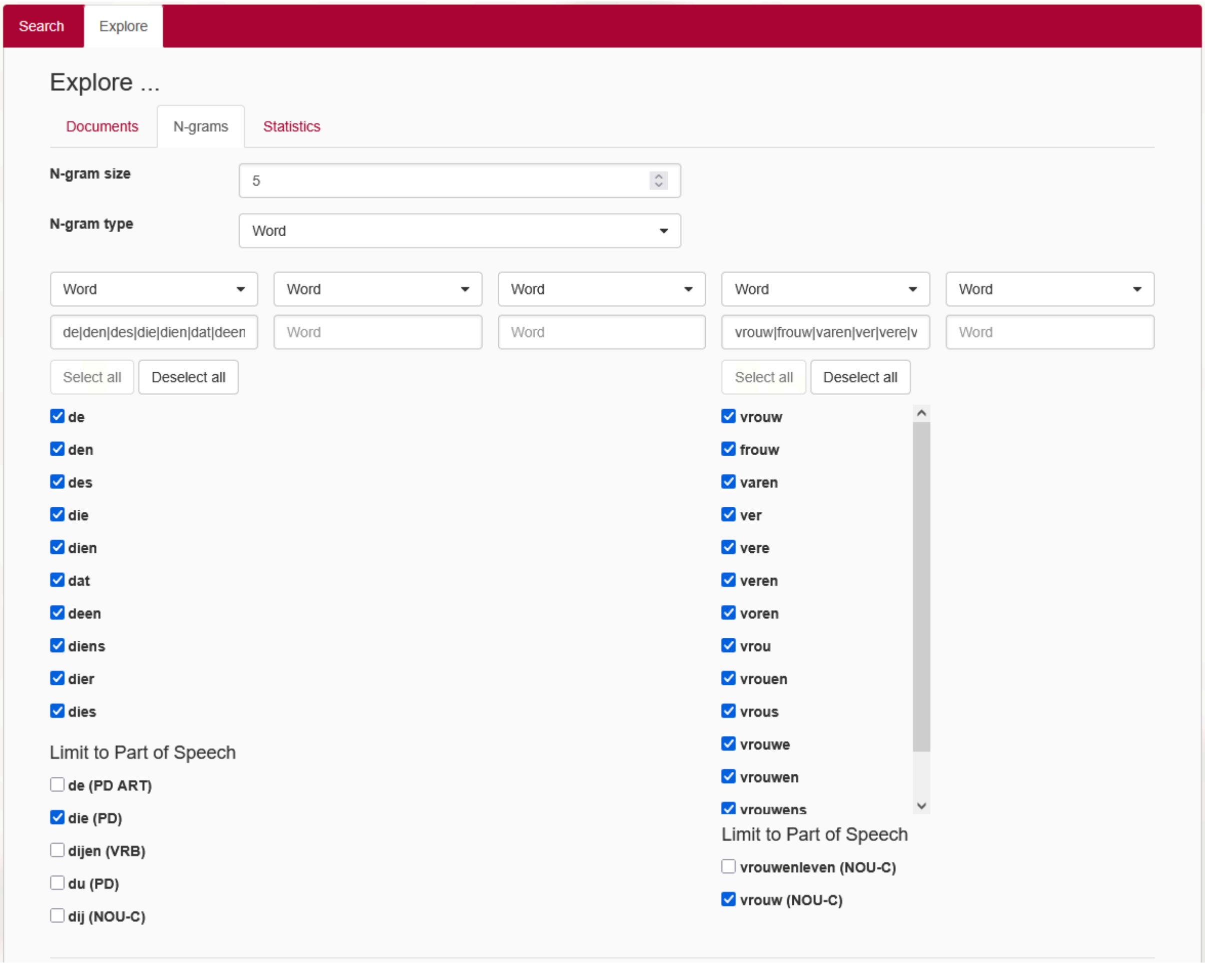
## N-grams

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*: the attribute to search for. You can choose: Word (i.e. word form). If you do not specify the search term a series of arbitrary words equal to the n-gram size will be searched for.
- It is also possible to restrict to, for instance, n-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 clicking on ‘Select all’ all forms belonging to a GiGaNT lemma are added.
- By using the Filter search by … you can create a subcorpus within the *Gekaapte Brieven* for specific metadata.

### Example



Explore ...

N-gram size: 5

N-gram type: Word

| Word | Word | Word | Word | Word |
|---|---|---|---|---|
| de\|den\|des\|die\|dien\|dat\|deen | Word | Word | vrouw\|frouw\|varen\|ver\|vere\|v | Word |

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

Limit to Part of Speech

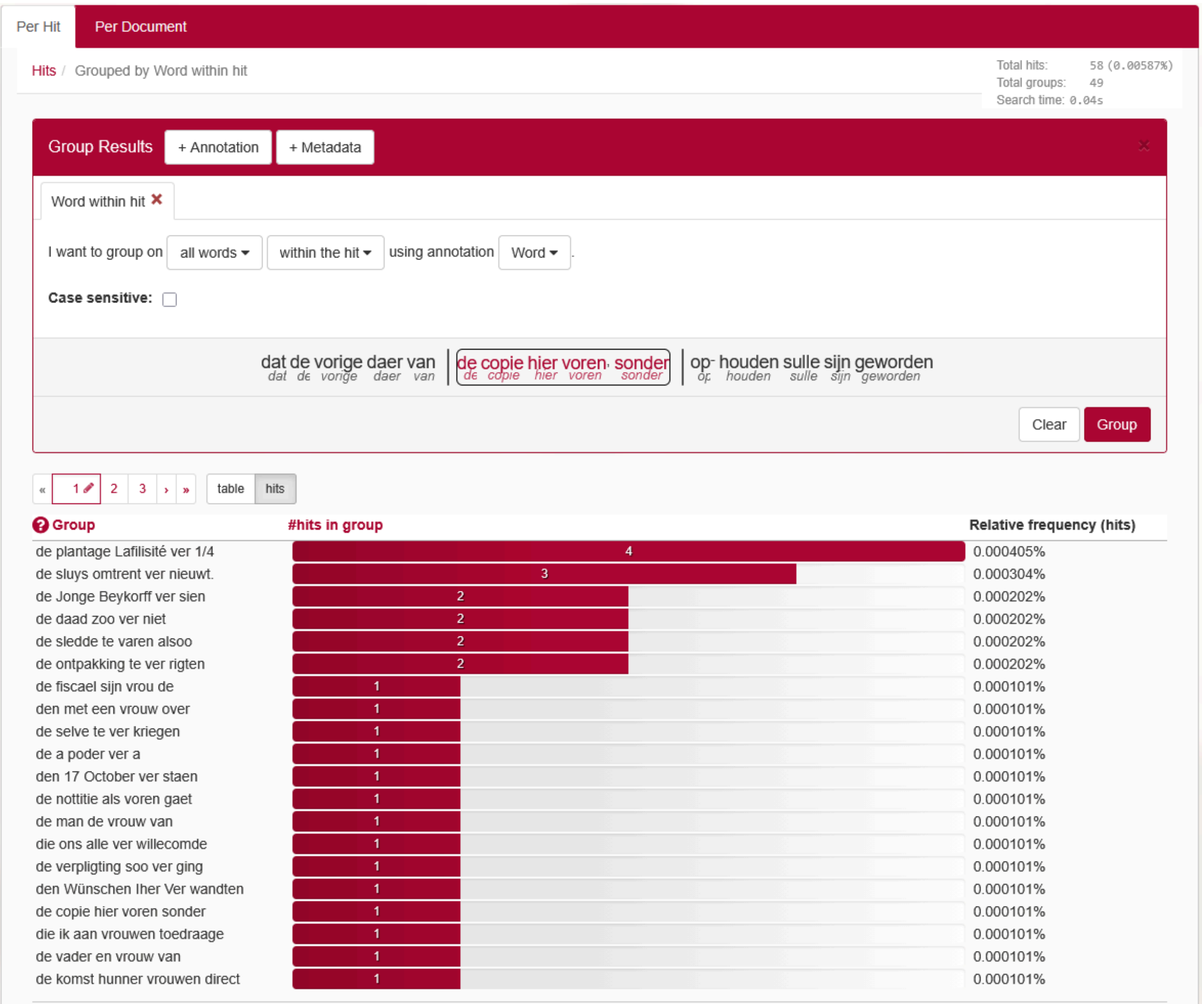
☐ de (PD ART)
☑ die (PD)
☐ dijen (VRB)
☐ du (PD)
☐ dij (NOU-C)

☑ vrouw
☑ frouw
☑ varen
☑ ver
☑ vere
☑ veren
☑ voren
☑ vrou
☑ vrouen
☑ vrous
☑ vrouwe
☑ vrouwen
☑ vrouwens

Limit to Part of Speech

☐ vrouwenleven (NOU-C)
☑ vrouw (NOU-C)

Within all the documents of the *Gekaapte Brieven*, you will find 58 occurrences of this so-called 5-gram (choose the option ‘Select all’ for both words and limit your search to Part of Speech *die* (PD) respectively *vrouw* (NOU-C)).



Total hits: 58 (0.00587%)
Total groups: 49
Search time: 0.04s

I want to group on all words within the hit using annotation Word.

Case sensitive: □

dat de vorige daer van | de copie hier voren. sonder | op- houden sulle sijn geworden

| Group | #hits in group | Relative frequency (hits) |
|---|---|---|
| de plantage Lafilisité ver 1/4 | 4 | 0.000405% |
| de sluys omtrent ver nieuwt. | 3 | 0.000304% |
| de Jonge Beykorff ver sien | 2 | 0.000202% |
| de daad zoo ver niet | 2 | 0.000202% |
| de sledde te varen alsoo | 2 | 0.000202% |
| de ontpakking te ver rigten | 2 | 0.000202% |
| de fiscael sijn vrou de | 1 | 0.000101% |
| den met een vrouw over | 1 | 0.000101% |
| de selve te ver kriegen | 1 | 0.000101% |
| de a poder ver a | 1 | 0.000101% |
| den 17 October ver staen | 1 | 0.000101% |
| de nottitie als voren gaet | 1 | 0.000101% |
| de man de vrouw van | 1 | 0.000101% |
| die ons alle ver willecomde | 1 | 0.000101% |
| de verpligting soo ver ging | 1 | 0.000101% |
| den Wünschen Iher Ver wandten | 1 | 0.000101% |
| de copie hier voren sonder | 1 | 0.000101% |
| die ik aan vrouwen toedraage | 1 | 0.000101% |
| de vader en vrouw van | 1 | 0.000101% |
| de komst hunner vrouwen direct | 1 | 0.000101% |

## Statistics (frequency lists)

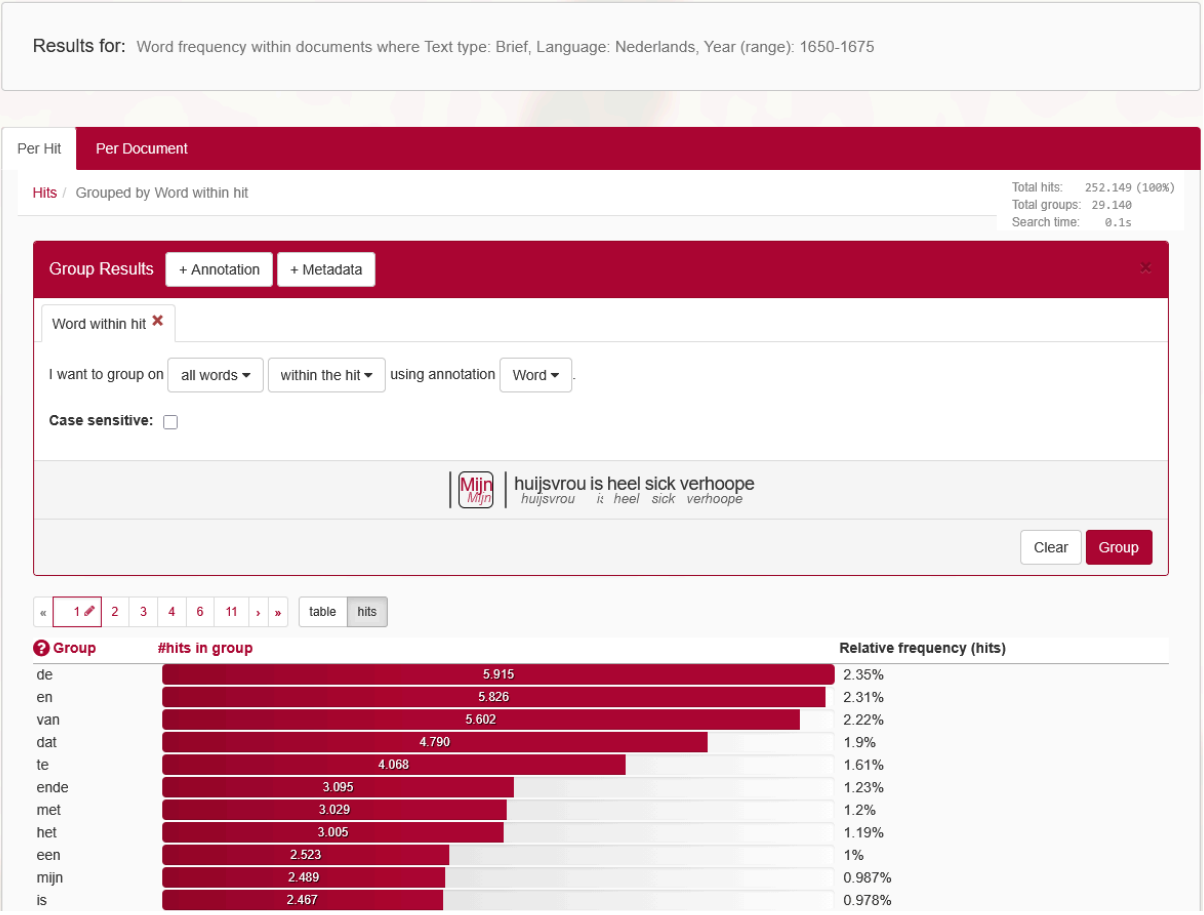
Here, you can produce frequency lists for the corpus. It is rather similar to the previous option, but restricted to 1-grams.

### 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 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: *Nederlands*, Text type: *Brief* and Year (range): *1650-1675*. This results in:



Results for: Word frequency within documents where Text type: Brief, Language: Nederlands, Year (range): 1650-1675

Total hits: 252.149 (100%)
Total groups: 29.140
Search time: 0.1s

I want to group on all words within the hit using annotation Word.

Case sensitive: □

Mijn huijsvrou is heel sick verhoope

| Group | #hits in group | Relative frequency (hits) |
|---|---|---|
| de | 5.915 | 2.35% |
| en | 5.826 | 2.31% |
| van | 5.602 | 2.22% |
| dat | 4.790 | 1.9% |
| te | 4.068 | 1.61% |
| ende | 3.095 | 1.23% |
| met | 3.029 | 1.2% |
| het | 3.005 | 1.19% |
| een | 2.523 | 1% |
| mijn | 2.489 | 0.987% |
| is | 2.467 | 0.978% |

# 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!"

## 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.

It is possible to search for punctuation marks. E.g. to find occurrences of the word "want" preceded by a comma use the following query:

```
[punctBefore="," & word="want"]
```

To find occurrences of the lemma "krant" that are followed by an exclamation mark, use:

```
[lemma="krant" & punctAfter="!"]
```

Some punctuation marks have a special function in regular expressions and therefore must be preceded by a backslash (\) when used in queries. For instance, to search for a period (`.`) after the word **"geweest"**, use:

```
[word="sentence" & punctAfter="\."]
```

## 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.