VivA Corpus Portal user guide

Version 1.0 - May 2017

Liesbeth Augustinus

09/05/2017

VIRTUELE INSTITUUT VIR AFRIKAANS

Table of contents

1 What is corpus linguistics? ... 1

2 What is Corpus Portal? ... 1

3 Corpus Information ... 3

3.1 Corpus ... 3

3.2 Statistics ... 3

3.3 N-grams ... 4

4 Search ... 5

4.1 Simple ... 5

4.1.1 Search command ... 5

4.1.2 Results ... 5

4.2 Expanded ... 9

4.2.1 Search command ... 9

4.2.2 Results ... 10

4.2.3 Batch queries ... 11

4.3 Advanced ... 13

4.3.1 Search command ... 13

4.3.2 Results ... 14

4.3.3 Batch queries ... 15

4.4 Expert (‘Kundig’) ... 15

4.4.1 Search command ... 15

4.4.2 Results ... 16

5 Corpus Query Language (CQL) ... 17

5.1 Simple search commands ... 18

5.1.1 A few basic principles ... 18

5.1.2 Search for a word, lemma or part of speech. ... 18

5.1.3 Search for adjacent words, lemmas and/or parts of speech. ... 18

5.1.4 Search adjecent constructions ... 19

5.2 Boolean operators.. 20

5.2.1 AND 20

5.2.2 OR.. 20

5.2.3 NOT 20

5.3 Fixed expressions 21

5.3.1 Wild card for a random character 21

5.3.2 Grouping of characters and patterns 21

5.3.3 Quantification 22

5.3.4 Escape sign 23

5.3.5 Case sensitive search commands... 23

6 Addendum .... **Error! Bookmark not defined.**

6.1 Corpora in Corpus Portal 24

6.2 Linguistic annotation 25

6.2.1 Labels 25

6.2.2 General parts of speech 25

6.2.3 Detailled parts of speech 25

6.2.4 Boolean operators 29

6.2.5 Regex operators 29

6.2.6 Quantification of search environment... 30

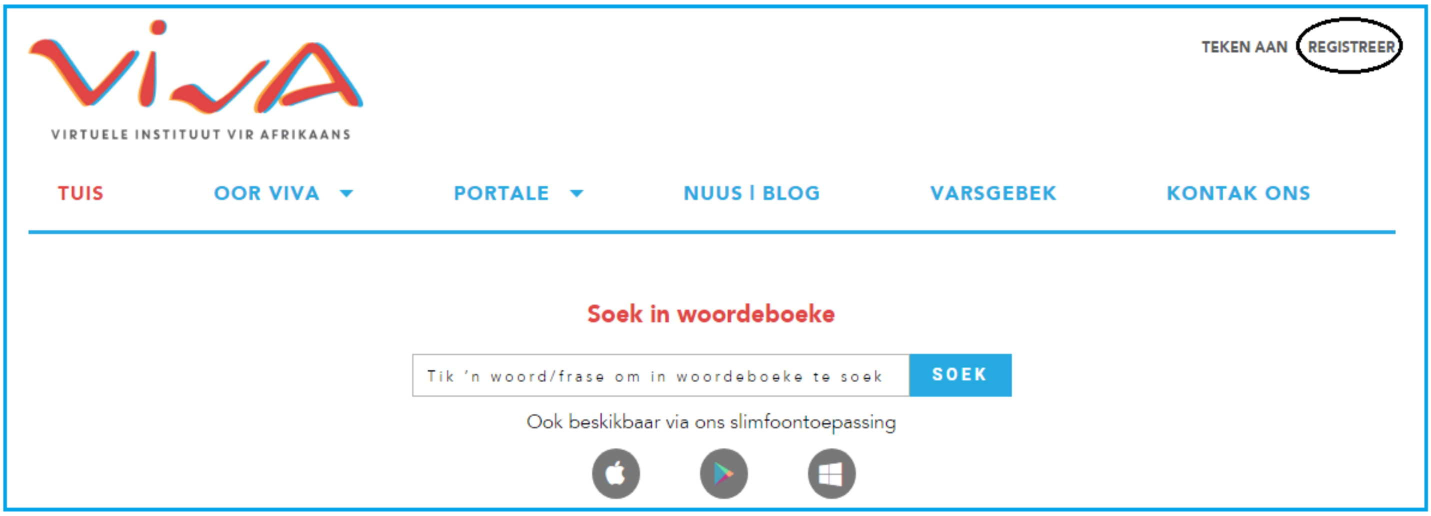
1 What is corpus linguistics?

Within the domain of linguistics, various research methods exist. For decades now, introspective approaches have paved the way for more empirical research techniques, which are based on actual language usage rather than on the grammatical judgement of the linguist. Corpus linguistics forms part of quantitative, empirical approaches, by means of which linguistic phenomena are researched with the aid of large databases, i.e. corpora, (singular: corpus) containing written or spoken language.

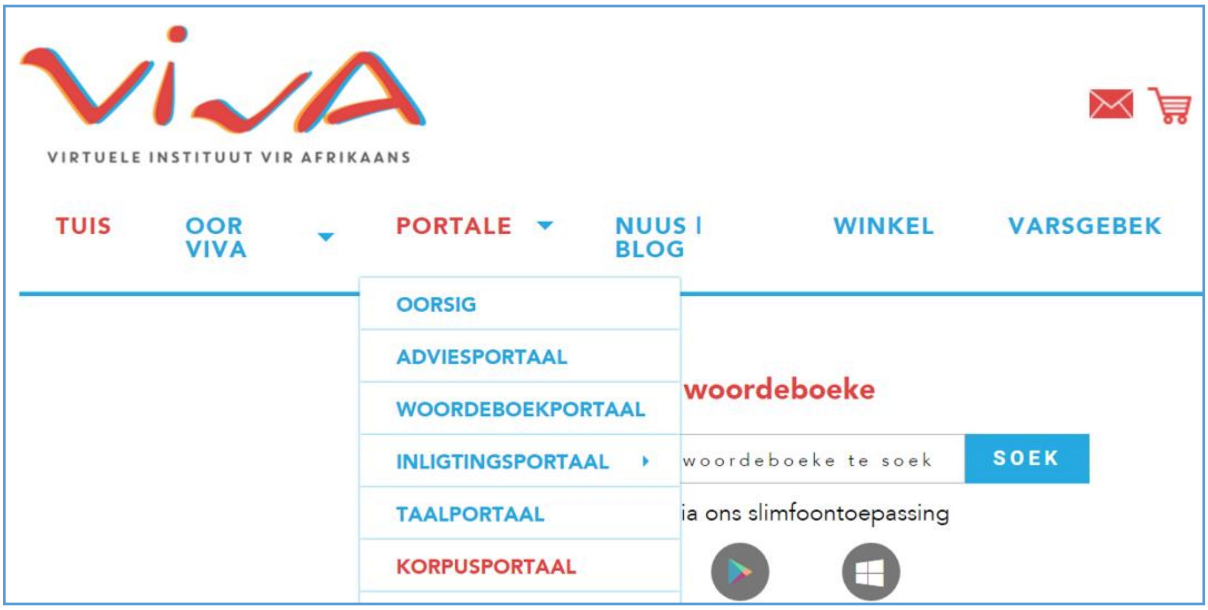
2 What is Corpus Portal?

Within the scope of the VivA project (http://viva-afrikaans.org) Corpus Portal has been made available in October 2015, whereby more than **85 million words** have been made accessible in written Afrikaans. Apart from the words, Corpus Portal provides an **extended search interface** to search the data.

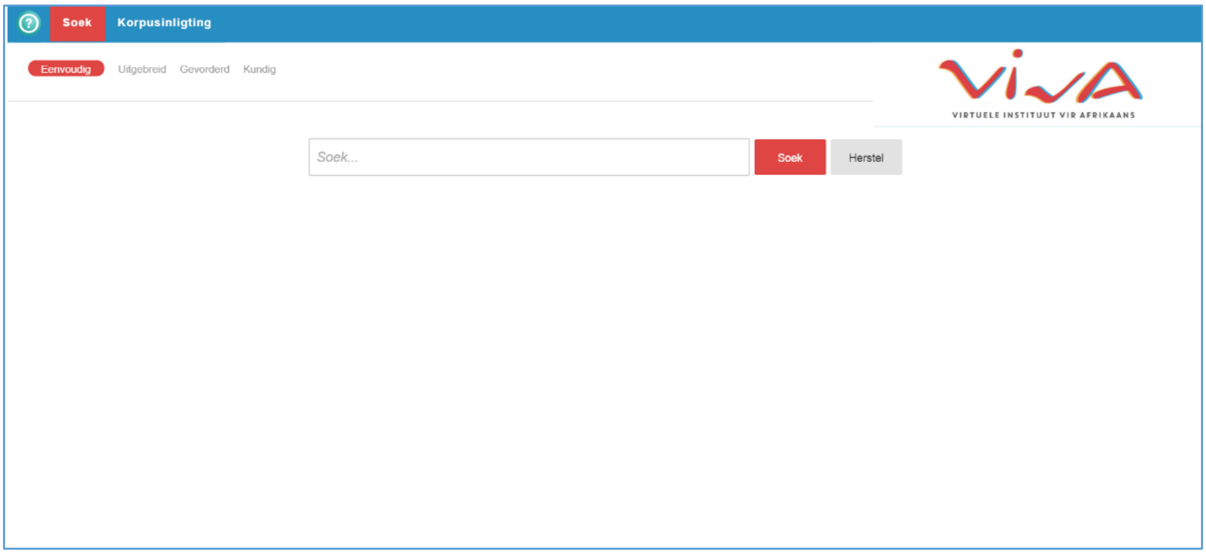
In order to gain access to Corpus Portal, users have to be registered VivA users. Registration can be done by navigating to the home page and clicking on 'Registreer'.


VIVA
VIRTUELE INSTITUUT VIR AFRIKAANS
TEKEN AAN
REGISTREER
TUIS
OOR VIVA
PORTALE
NUUS | BLOG
VARSGEBEK
KONTAK ONS
Soek in woordeboeke
Tik 'n woord/frase om in woordeboeke te soek
SOEK
Ook beskikbaar via ons slimfoontoepassing

As soon as registration has been completed, the user may log on by clicking on the link. Hover over 'Portale' in the menu and click on 'Korpusportaal', which will take you to the correct page.



This is what the Corpus Portal page looks like:



At the top of the page there are 2 options: ‘Soek’ [Search] and ‘Korpusinligting’ [Corpus Information]. Section 3 deals with the other information that may be accessed via Corpus Information. The search interface is discussed in section 4.

By using the ? button, you can undertake a systematic tour of the Corpus Portal, which will introduce you to every functionality.

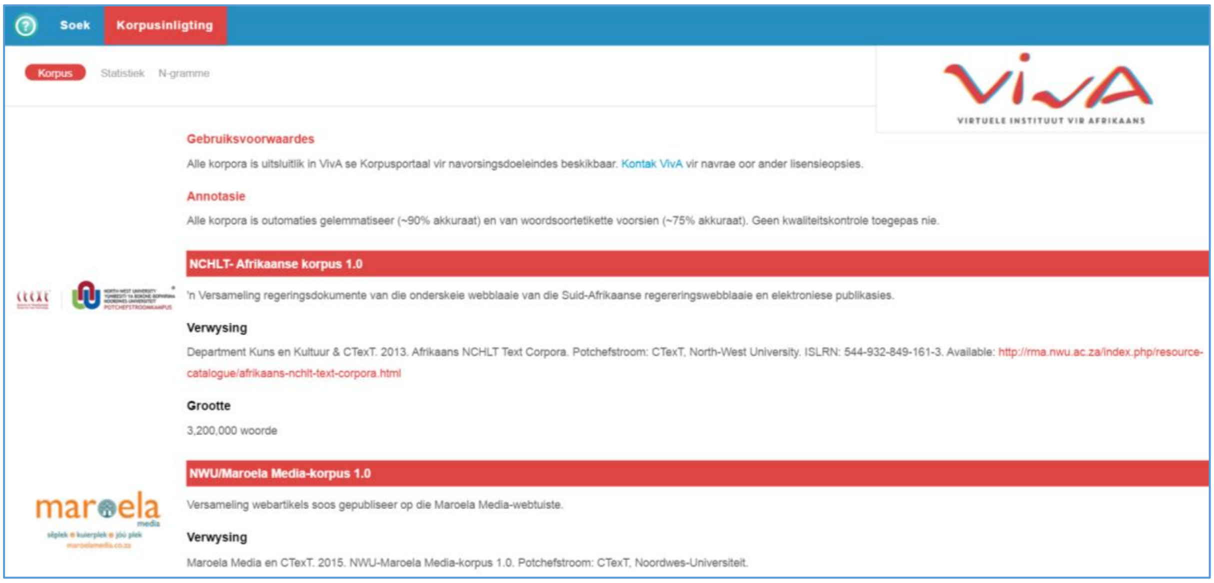
3 Corpus Information

On the Corpus Information page in Corpus Portal, three options are given pertaining to information on the compilation of the corpora: ‘Korpus’ [Corpus], ‘Statistiek’ [Statistics] and ‘N-gramme’ [N-grams].

3.1 Corpus

On the Corpus page, you will find an overview of the various corpora and their specifications, including genre and scope (number of words) of the texts (refer to Addendum A). In total, Corpus Portal contains more than 85 million words in written Afrikaans. Most source texts are very recent, and dates from 1995-2015.

All corpora have been automatically lemmatized (approximately 90% accurate) and provided with parts of speech (approximately 75% accurate). The annotations have not been corrected; users should take into account possible annotation errors.



Gebruiksvoorwaardes

Alle korpora is uitsluitlik in VivA se Korpusportaal vir navorsingsdoeleindes beskikbaar. Kontak VivA vir navrae oor ander lisensieopsies.

Annotasie

Alle korpora is outomaties gelemmatiseer (~90% akkuraat) en van woordsoortetikette voorsien (~75% akkuraat). Geen kwaliteitskontrole toegepas nie.

NCHLT- Afrikaanse korpus 1.0

'n Versameling regeringsdokumente van die onderskeie webblaaie van die Suid-Afrikaanse regeringswebblaaie en elektroniese publikasies.

Verwysing

Department Kuns en Kultuur & CTexT. 2013. Afrikaans NCHLT Text Corpora. Potchefstroom: CTexT, North-West University. ISLRN: 544-932-849-161-3. Available: http://rma.nwu.ac.za/index.php/resource-catalogue/afrikaans-nchlt-text-corpora.html

Grootte

3,200,000 woorde

NWU/Maroela Media-korpus 1.0

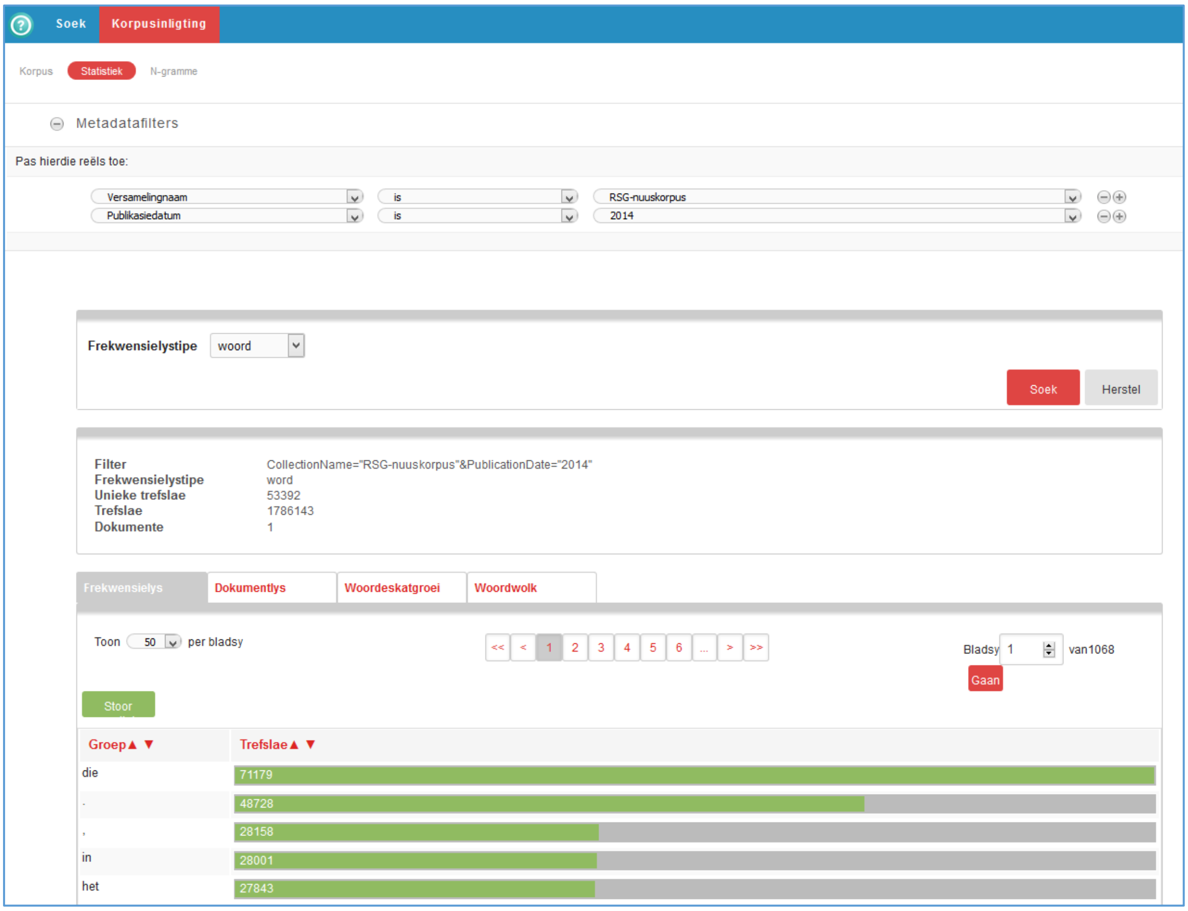
Versameling webartikels soos gepubliseer op die Maroela Media-webtuiste.

Verwysing

Maroela Media en CTexT. 2015. NWU-Maroela Media-korpus 1.0. Potchefstroom: CTexT, Noordwes-Universiteit.

3.2 Statistics

Under Statistics you can request **frequency lists** for words, lemmas and parts of speech for every author, publication date and corpus component (i.e. metadata filters). By clicking on the plus sign to the right you can create more metadata filters; by clicking on the minus, they will be removed. The figure below illustrates the search command to find the most frequent word in the RSG corpus for 2014. The ‘Stoor’ [save] button enables you to export the result. It is also possible to request a document list, a graph of the frequency of the entry (‘Woordeskatgroei’) and to generate a word cloud.



Pas hierdie reëls toe:

Versamelingnaam	is	RSG-nuuskorpus
Publikasiedatum	is	2014

Frekwensielystipe: woord

Filter	CollectionName="RSG-nuuskorpus"&PublicationDate="2014"
Frekwensielystipe	word
Unieke trefslae	53392
Trefslae	1786143
Dokumente	1

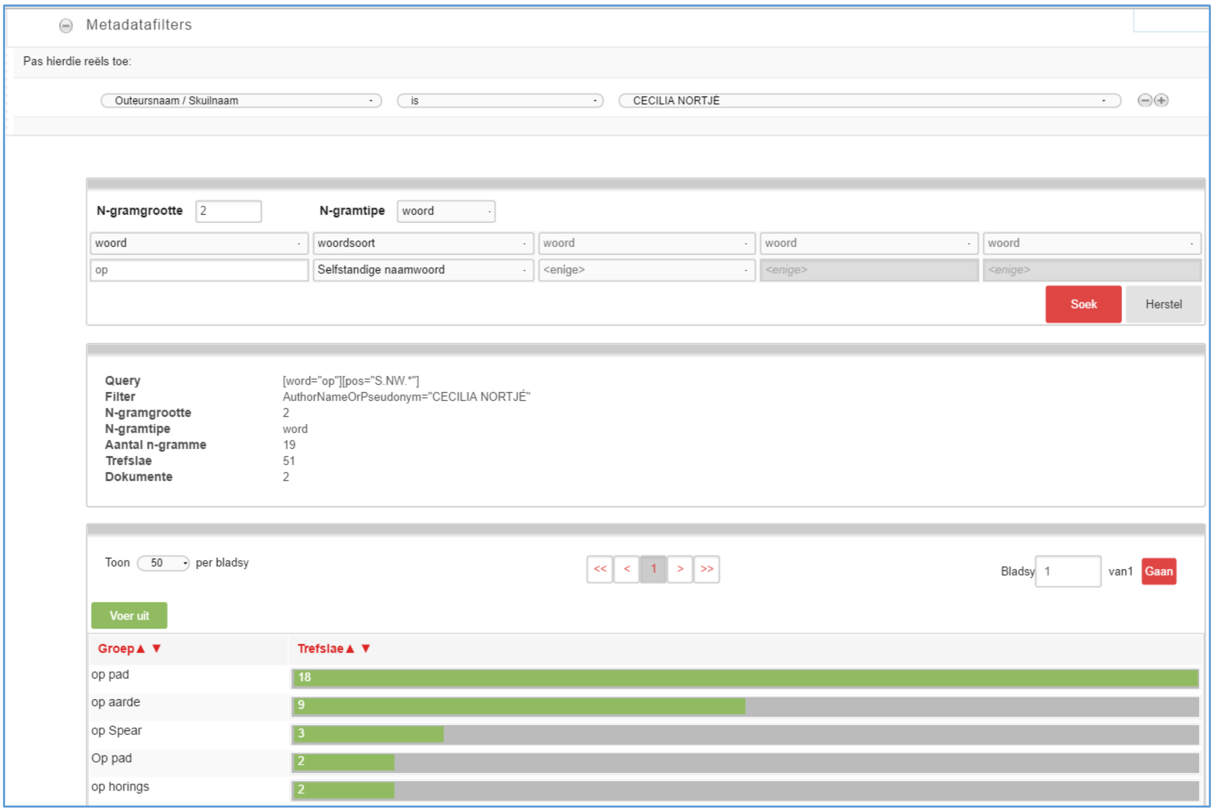
Toon 50 per bladsy

Bladsy 1 van1068

Groep▲ ▼	Trefslae▲ ▼
die	71179
.	48728
,	28158
in	28001
het	27843

3.3 N-grams

Here you can generate a frequency list of N-grams. Within this context an N-gram refers to a sequence of N words, lemmas or parts of speech (or a combination thereof), to a maximum length of 5. The search command below generates an answer to the question which substantives are most frequent after the preposition or in the work of Cecilia Nortjé.



Metadatafilters

Pas hierdie reëls toe:

Outeursnaam / Skuilnaam | is | CECILIA NORTJÉ

N-gramgrootte 2 N-gramtipe woord

woord	woordsoort	woord	woord	woord
op	Selfstandige naamwoord	<enige>	<enige>	<enige>

Query	[word="op"][pos="S.NW.*"]
Filter	AuthorNameOrPseudonym="CECILIA NORTJÉ"
N-gramgrootte	2
N-gramtipe	word
Aantal n-gramme	19
Trefslae	51
Dokumente	2

Toon 50 per bladsy

Groep▲ ▼	Trefslae▲ ▼
op pad	18
op aarde	9
op Spear	3
Op pad	2
op horings	2

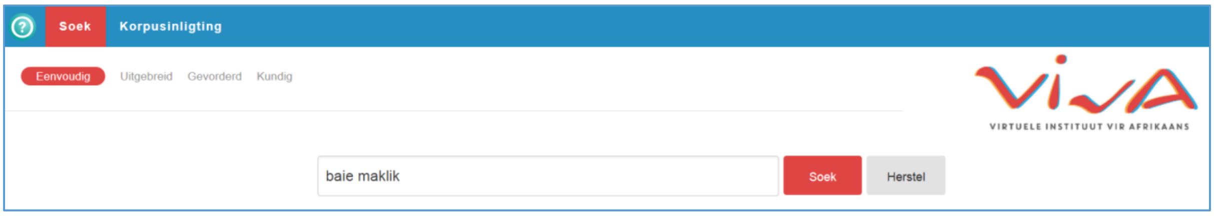
4 Search

The search interface ('Soek') offers four ways to search the corpora: 'eenvoudig' [simple], 'uitgebreid' [extended], 'gevorderd' [advanced] and 'kundig' [expert].

4.1 Simple

4.1.1 Search command

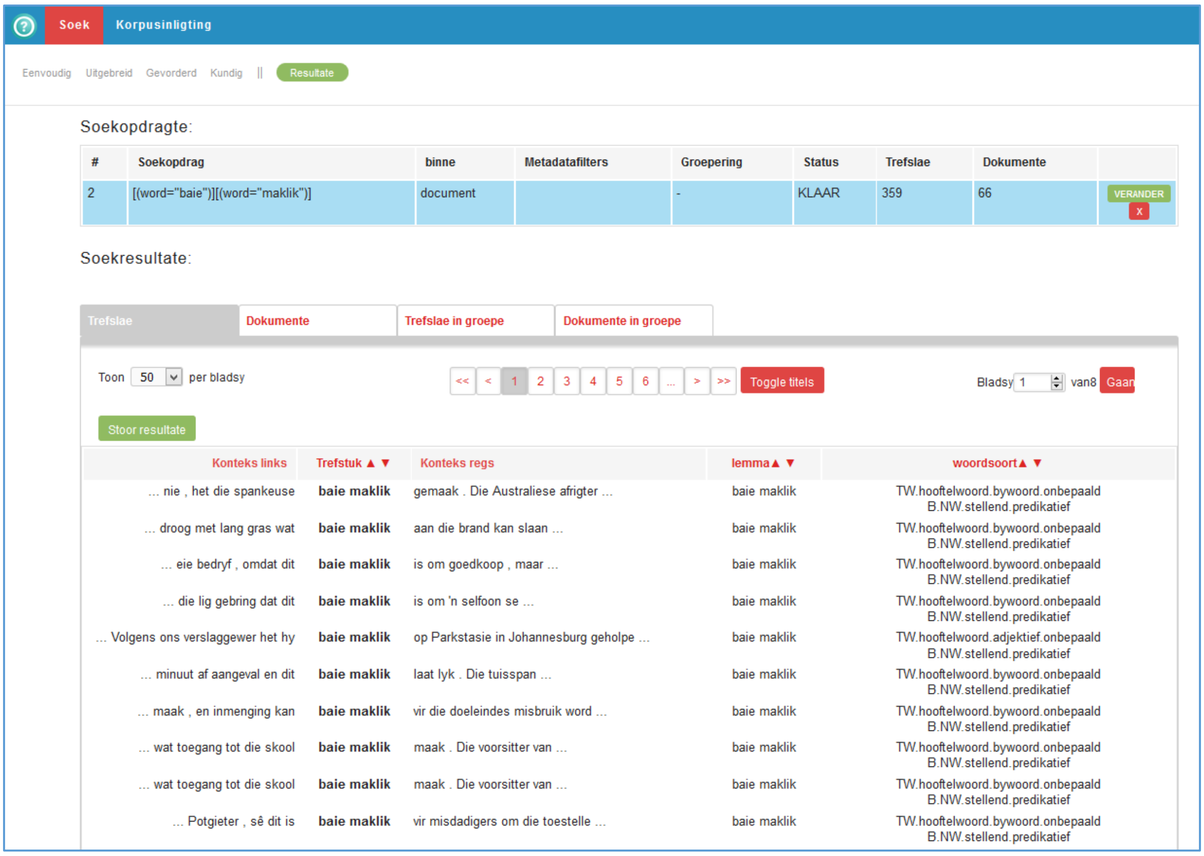
With the simple option you can search for a word or a sequence of words in Corpus Portal via a search button. There is no option to filter metadata. The search term is also not case sensitive, which means that capitals and accents are excluded from the search.



baie maklik

4.1.2 Results

After clicking on 'Soek' [search], a search is launched for the search term in all corpora of the Corpus Portal. Below are a few hits for the search query 'baie maklik'.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
2	[(word="baie")][(word="maklik")]	document		-	KLAAR	359	66	VERANDER X

Soekresultate:

Konteks links	Trefstuk ▲ ▼	Konteks regs	lemma▲ ▼	woordsoort▲ ▼
... nie , het die spankeuse	**baie maklik**	gemaak . Die Australiese afrigter ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... droog met lang gras wat	**baie maklik**	aan die brand kan slaan ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... eie bedryf , omdat dit	**baie maklik**	is om goedkoop , maar ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... die lig gebring dat dit	**baie maklik**	is om 'n selfoon se ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... Volgens ons verslaggewer het hy	**baie maklik**	op Parkstasie in Johannesburg geholpe ...	baie maklik	TW.hooftelwoord.adjektief.onbepaald B.NW.stellend.predikatief
... minuut af aangeval en dit	**baie maklik**	laat lyk . Die tuisspan ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... maak , en inmenging kan	**baie maklik**	vir die doeleindes misbruik word ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... wat toegang tot die skool	**baie maklik**	maak . Die voorsitter van ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... wat toegang tot die skool	**baie maklik**	maak . Die voorsitter van ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
... Potgieter , sê dit is	**baie maklik**	vir misdadigers om die toestelle ...	baie maklik	TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief

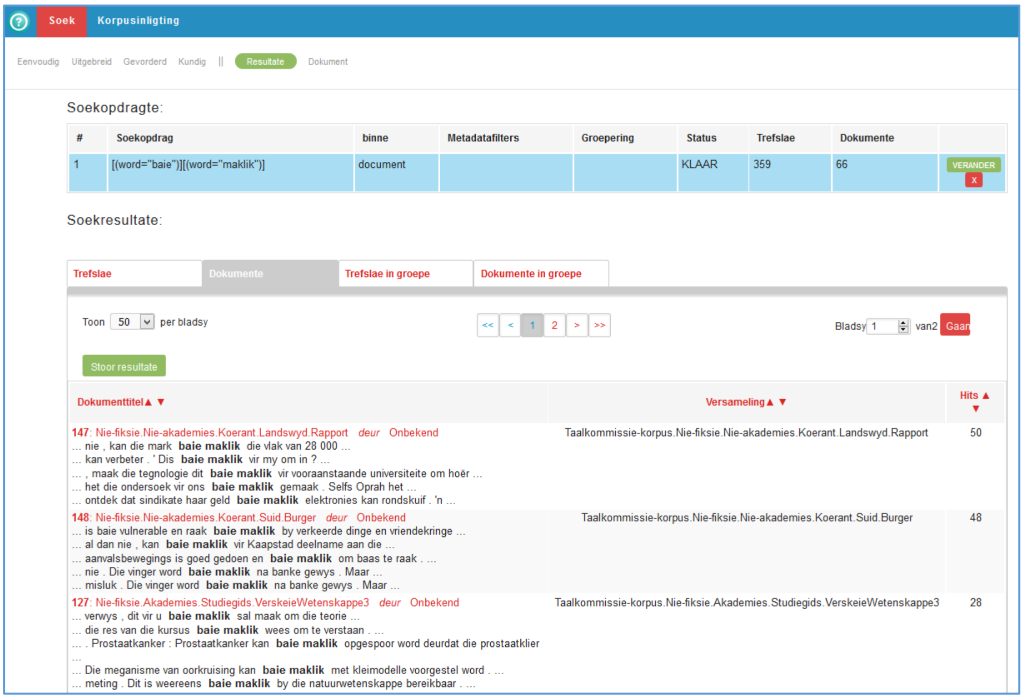
At the top of the search results page ('Soekresultate') the following information is given: The search query in Corpus Query Language (see section 5), the status (whether the search is completed or not), the number of hits and the number of documents in which the hits occurred. You can customize the search by clicking 'verander'.

Warning: If a search yields more than 5 million hits, only 5000000 will be displayed. The figure should therefore be interpreted as '5 million or more'.

Under the search command is the search results window. Under the heading **'Trefslae'** the results are given as keywords in context (KWIC).

The words that match the search command (i.e. the hit or the 'match') are bold and are visually separated from the left and right. In addition, the lemma and parts of speech are displayed. The results can be sorted by clicking on one of the column names. You can choose whether you want to see 50, 100 or 200 results per page. By clicking 'Toggle titles' you can see information about the source of the sentence, i.e. the author and the publication date of the text. By clicking on the search result, you can see more context. You can save the results by clicking 'stoor resultate'.

Under the heading **‘Dokumente’**, the hits are sorted by the text (‘dokumenttitel’) and corpus (‘versameling’). This enables you to determine which texts produce the most hits.



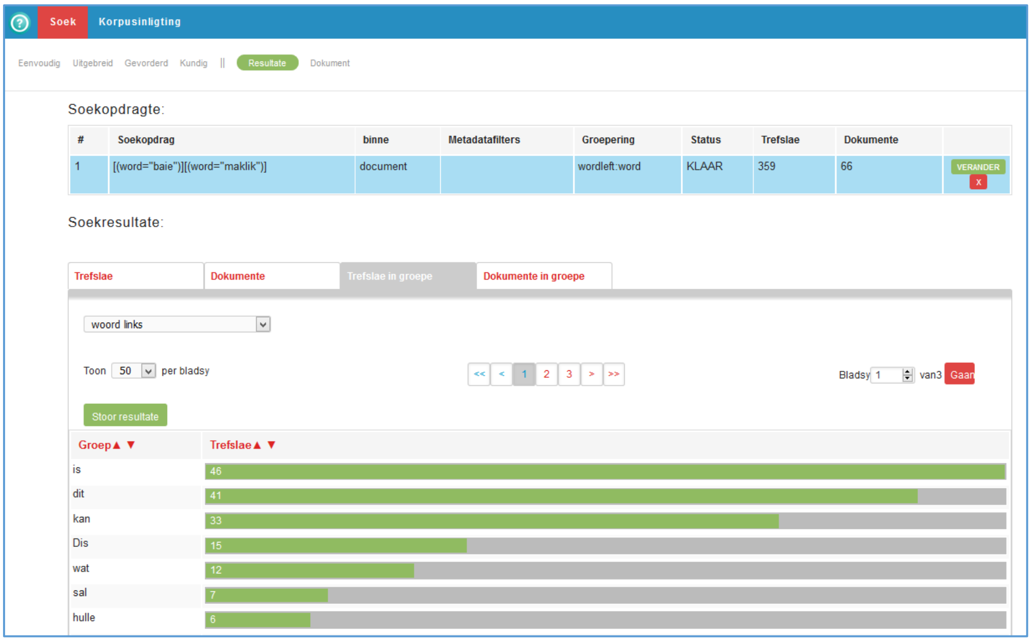
Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[(word="baie")][(word="maklik")]	document			KLAAR	359	66	VERANDER x

Soekresultate:

Dokumenttitel▲ ▼	Versameling▲ ▼	Hits ▲ ▼
147: Nie-fiksie.Nie-akademies.Koerant.Landswyd.Rapport *deur* Onbekend ... nie , kan die mark **baie maklik** die vlak van 28 000 ... ... kan verbeter . ' Dis **baie maklik** vir my om in ? ... ... , maak die tegnologie dit **baie maklik** vir vooraanstaande universiteite om hoër ... ... het die ondersoek vir ons **baie maklik** gemaak . Selfs Oprah het ... ... ontdek dat sindikate haar geld **baie maklik** elektronies kan rondskuif . 'n ...	Taalkommissie-korpus.Nie-fiksie.Nie-akademies.Koerant.Landswyd.Rapport	50
148: Nie-fiksie.Nie-akademies.Koerant.Suid.Burger *deur* Onbekend ... is baie vulnerable en raak **baie maklik** by verkeerde dinge en vriendekringe ... ... al dan nie , kan **baie maklik** vir Kaapstad deelname aan die ... ... aanvalsbewegings is goed gedoen en **baie maklik** om baas te raak . ... nie . Die vinger word **baie maklik** na banke gewys . Maar ... ... misluk . Die vinger word **baie maklik** na banke gewys . Maar ...	Taalkommissie-korpus.Nie-fiksie.Nie-akademies.Koerant.Suid.Burger	48
127: Nie-fiksie.Akademies.Studiegids.VerskeieWetenskappe3 *deur* Onbekend ... verwys , dit vir u **baie maklik** sal maak om die teorie ... ... die res van die kursus **baie maklik** wees om te verstaan . ... Prostaatkanker : Prostaatkanker kan **baie maklik** opgespoor word deurdat die prostaatklier ... ... Die meganisme van oorkruising kan **baie maklik** met kleimodelle voorgestel word . ... meting . Dit is weereens **baie maklik** by die natuurwetenskappe bereikbaar	Taalkommissie-korpus.Nie-fiksie.Akademies.Studiegids.VerskeieWetenskappe3	28

Under the heading **‘Trefslae in groepe’** you can group results by a specific category, for example by author name, word to the left, part of speech of the hit, etc. Below are the results sorted by ‘woord links’ [word to the left]. It appears that ‘is’ appears most frequently before the string ‘baie maklik’ in the corpora, for example in the sentence ‘Dit is baie maklik om te sê wat die meervoud van hond is’.



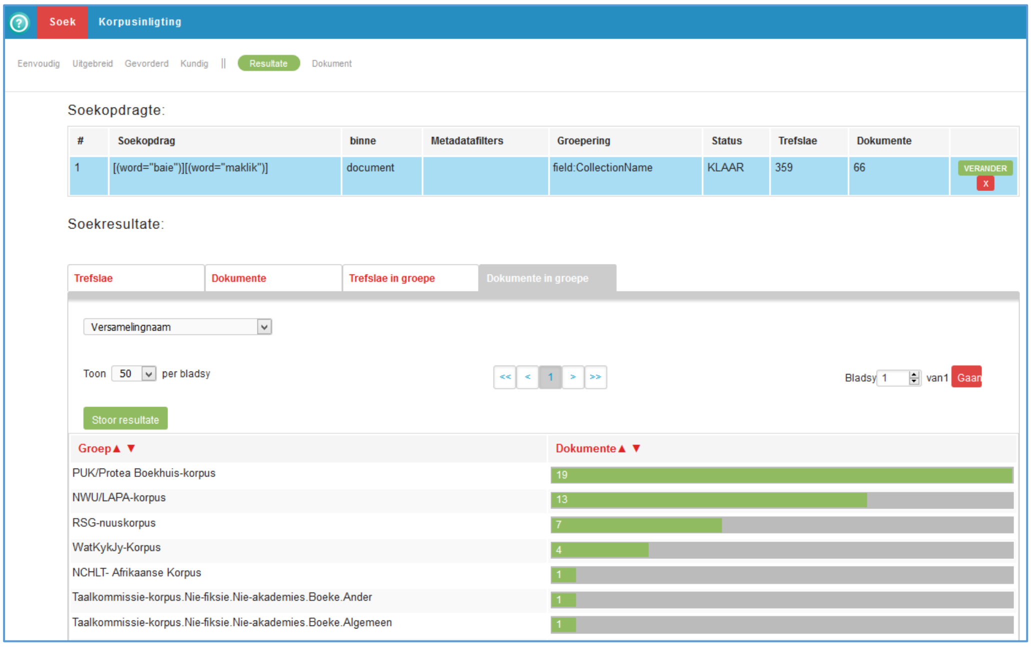
Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[(word="baie")][(word="maklik")]	document		wordleft:word	KLAAR	359	66	VERANDER x

Soekresultate:

Groep▲ ▼	Trefslae▲ ▼
is	46
dit	41
kan	33
Dis	15
wat	12
sal	7
hulle	6

Under the heading **'Dokumente in groepe',** you can sort the source texts according to three major categories: author, publication date and corpus ('versamelingnaam'). Below are the results sorted by corpus. It appears that most hits are found in 19 documents of the PUK/Protea Boekhuis corpus.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente
1	[(word="baie")][(word="maklik")]	document		field:CollectionName	KLAAR	359	66

Soekresultate:

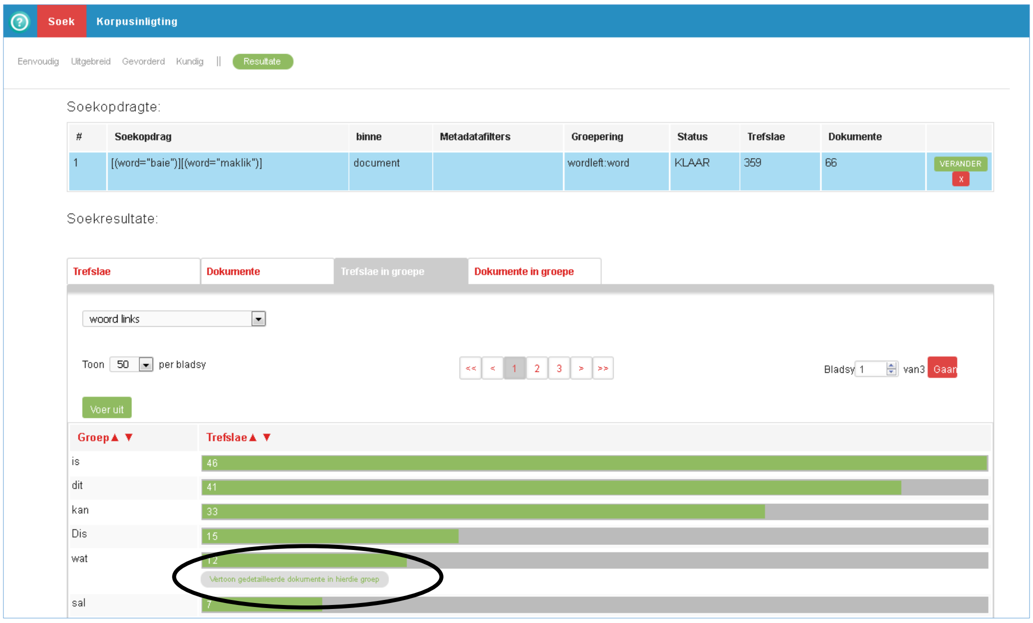
Versamelingnaam

Toon 50 per bladsy

Bladsy 1 van1

Groep▲ ▼	Dokumente▲ ▼
PUK/Protea Boekhuis-korpus	19
NWU/LAPA-korpus	13
RSG-nuuskorpus	7
WatKykJy-Korpus	4
NCHLT- Afrikaanse Korpus	1
Taalkommissie-korpus.Nie-fiksie.Nie-akademies.Boeke.Ander	1
Taalkommissie-korpus.Nie-fiksie.Nie-akademies.Boeke.Algemeen	1

Under the headings 'Trefslae in groepe' and 'Dokumente in groepe' you can click on the green bar of a search result. Then another button appears: 'Vertoon gedetailleerde dokumente in hierdie groep' [Show detailed documents in this group].



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente
1	[(word="baie")][(word="maklik")]	document		wordleft:word	KLAAR	359	66

Soekresultate:

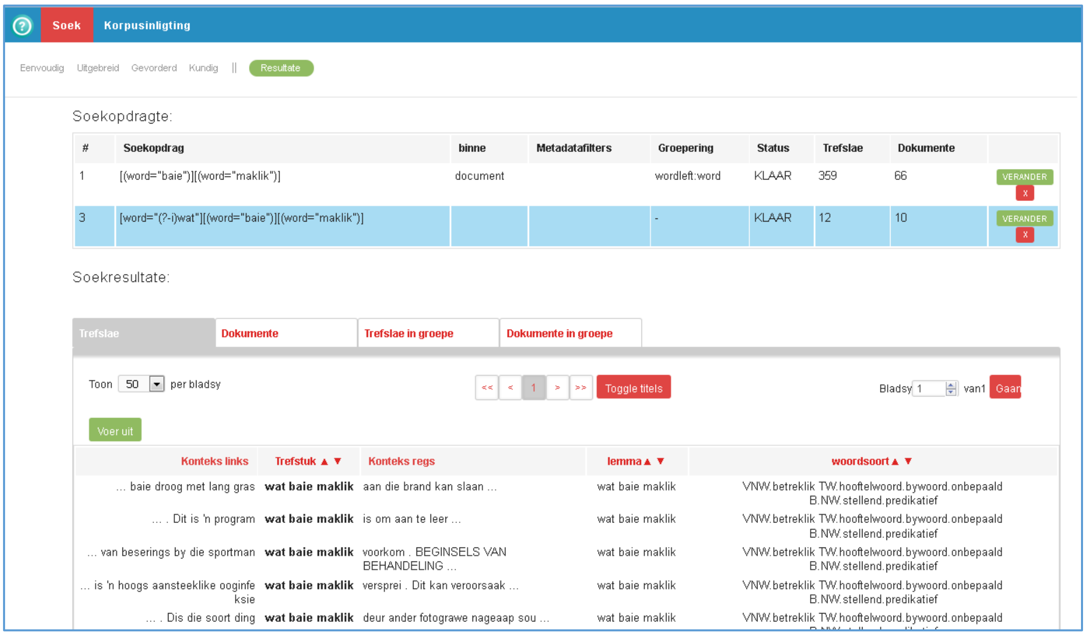
woord links

Toon 50 per bladsy

Bladsy 1 van3

Groep▲ ▼	Trefslae▲ ▼
is	46
dit	41
kan	33
Dis	15
wat	12
sal	7

Clicking this button will perform a new search, rendering only the search results of the item on which you have clicked. Below you will see the search command and the first results for a search on words that appear to the left of ‘baie maklik’, as in ‘lang gras wat baie maklik aan die brand slaan’.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[(word="baie")][(word="maklik")]	document		wordleft:word	KLAAR	359	66	VERANDER X
3	[word="(?-i)wat"][(word="baie")][(word="maklik")]			-	KLAAR	12	10	VERANDER X

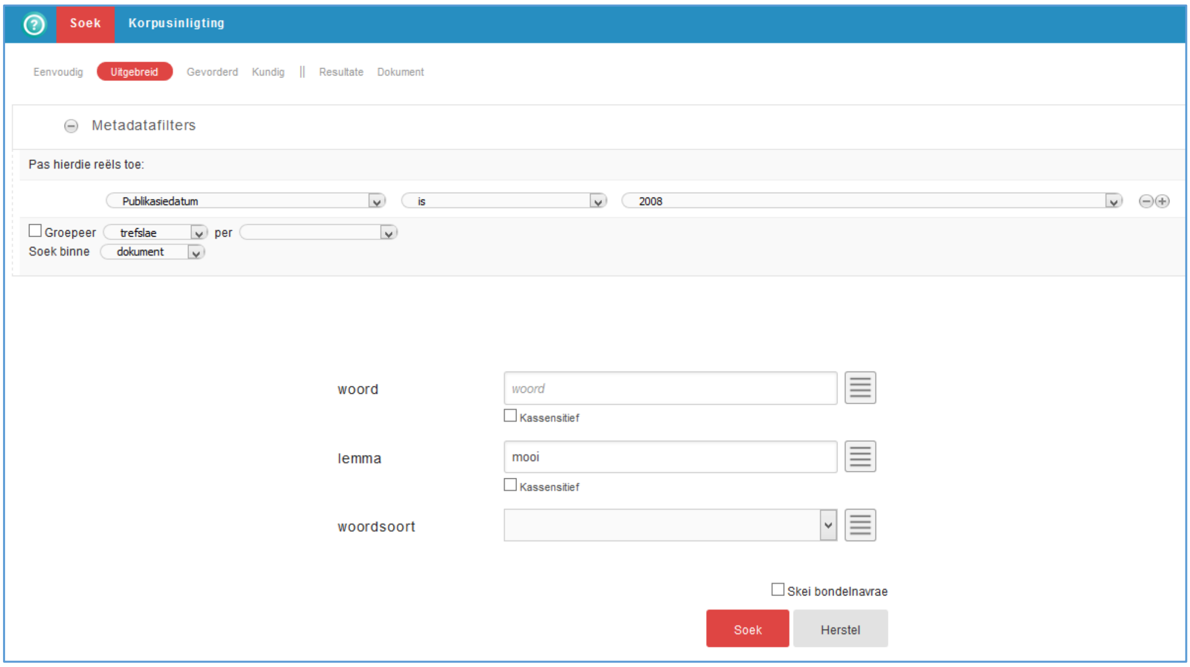
Soekresultate:

Konteks links	Trefstuk ▲ ▼	Konteks regs	lemma ▲ ▼	woordsoort ▲ ▼
… baie droog met lang gras	**wat baie maklik**	aan die brand kan slaan …	wat baie maklik	VNW.betreklik TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
… . Dit is 'n program	**wat baie maklik**	is om aan te leer …	wat baie maklik	VNW.betreklik TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
… van beserings by die sportman	**wat baie maklik**	voorkom . BEGINSELS VAN BEHANDELING …	wat baie maklik	VNW.betreklik TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
… is 'n hoogs aansteeklike ooginfeksie	**wat baie maklik**	versprei . Dit kan veroorsaak …	wat baie maklik	VNW.betreklik TW.hooftelwoord.bywoord.onbepaald B.NW.stellend.predikatief
… . Dis die soort ding	**wat baie maklik**	deur ander fotograwe nageaap sou …	wat baie maklik	VNW.betreklik TW.hooftelwoord.bywoord.onbepaald

4.2 Extended

4.2.1 Search command

The extended search method allows you to filter metadata and search for words, lemmas and/or parts of speech. To filter the metadata, you can set a number of ‘rules’. For example, you can search within a particular corpus, or you can choose a publication date or author name. You can set multiple filters by right-clicking on the + character. The search command below is to search for ‘mooi’ in texts published in 2008.



Pas hierdie reëls toe:

Publikasiedatum is 2008

☐ Groepeer trefslae per

Soek binne dokument

woord: woord
☐ Kassensitief

lemma: mooi
☐ Kassensitief

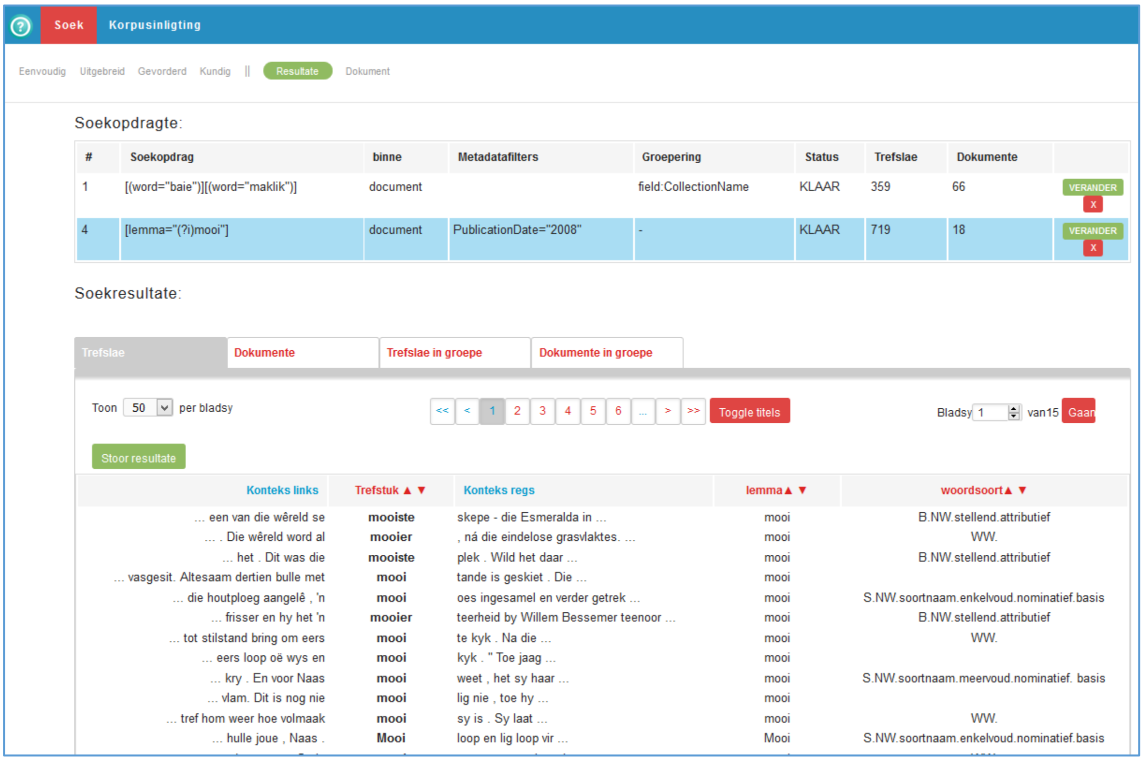
woordsoort:

☐ Skei bondelnavrae

Usually search queries are not case sensitive. If you want to create a case sensitive search, you must tick the box ‘Kassensitief’.

4.2.2 Results

The results are displayed in the ‘Soekresultate’ window. As indicated in section 4.1.2, you can sort and download the results.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[(word="baie")][(word="maklik")]	document		field:CollectionName	KLAAR	359	66	VERANDER x
4	[lemma="(?i)mooi"]	document	PublicationDate="2008"	-	KLAAR	719	18	VERANDER x

Soekresultate:

Toon 50 per bladsy

Bladsy 1 van15

Konteks links	Trefstuk ▲ ▼	Konteks regs	lemma ▲ ▼	woordsoort ▲ ▼
... een van die wêreld se	**mooiste**	skepe - die Esmeralda in ...	mooi	B.NW.stellend.attributief
... . Die wêreld word al	**mooier**	, ná die eindelose grasvlaktes. ...	mooi	WW.
... het . Dit was die	**mooiste**	plek . Wild het daar ...	mooi	B.NW.stellend.attributief
... vasgesit. Altesaam dertien bulle met	**mooi**	tande is geskiet . Die ...	mooi	
... die houtploeg aangelê , 'n	**mooi**	oes ingesamel en verder getrek ...	mooi	S.NW.soortnaam.enkelvoud.nominatief.basis
... frisser en hy het 'n	**mooier**	teerheid by Willem Bessemer teenoor ...	mooi	B.NW.stellend.attributief
... tot stilstand bring om eers	**mooi**	te kyk . Na die ...	mooi	WW.
... eers loop oë wys en	**mooi**	kyk . " Toe jaag ...	mooi	
... kry . En voor Naas	**mooi**	weet , het sy haar ...	mooi	S.NW.soortnaam.meervoud.nominatief. basis
... vlam. Dit is nog nie	**mooi**	lig nie , toe hy ...	mooi	
... tref hom weer hoe volmaak	**mooi**	sy is . Sy laat ...	mooi	WW.
... hulle joue , Naas .	**Mooi**	loop en lig loop vir ...	Mooi	S.NW.soortnaam.enkelvoud.nominatief.basis

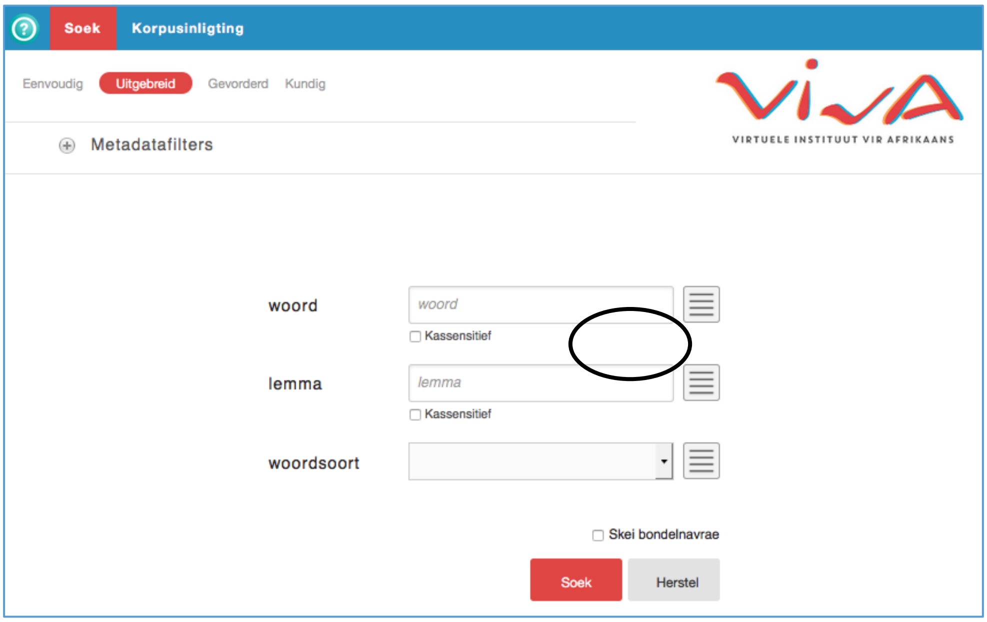
In the command, the metadata filter was created: Only documents created in 2008 are searched. Note that the lemma ‘mooi’ was searched. Therefore, the formal variants such as ‘mooie’, ‘mooier’ and ‘mooiste’ are also rendered.

Above the new search command, the previous search command is still visible. Clicking on it allows you to re-examine the search results.

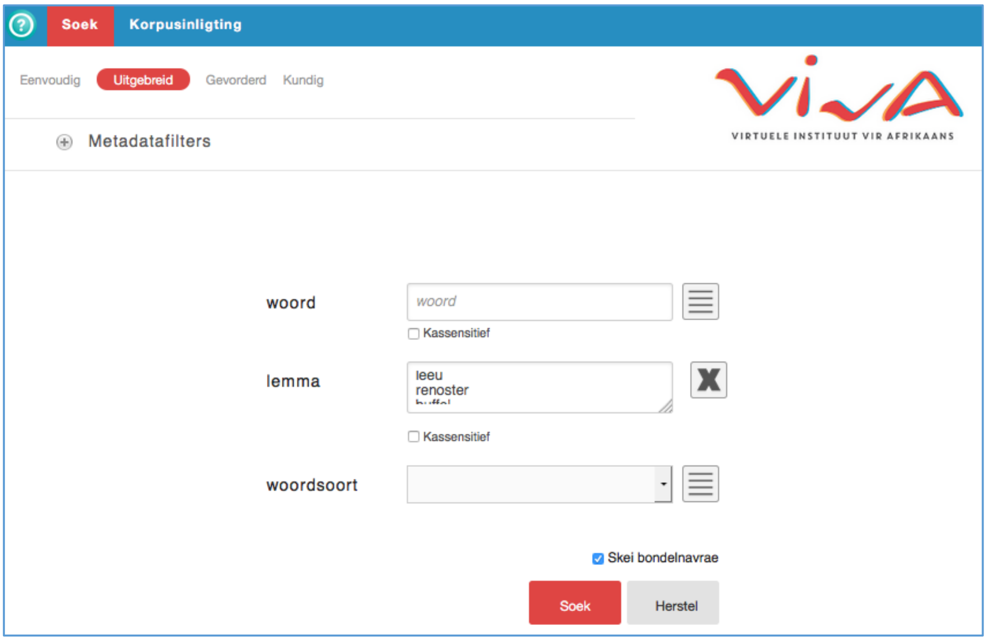
4.2.3 Batch queries

In expanded search mode, it is possible to run batch queries (‘bondelnavrae’). That means that it is possible to execute different search commands simultaneously. If you have a long list of words, lemmas or parts of speech that you want to reference in Corpus Portal, you do not have to enter each command separately.

By clicking on the ☰ button to the right of word, lemma or part of speech, you can upload an excerpt of the text (flat text, NOT a Word document) with a list of search commands. The items in the list must be the same type (words OR lemmas OR parts of speech).



In the example below, lemmas are loaded.



woord: woord

□ Kassensitief

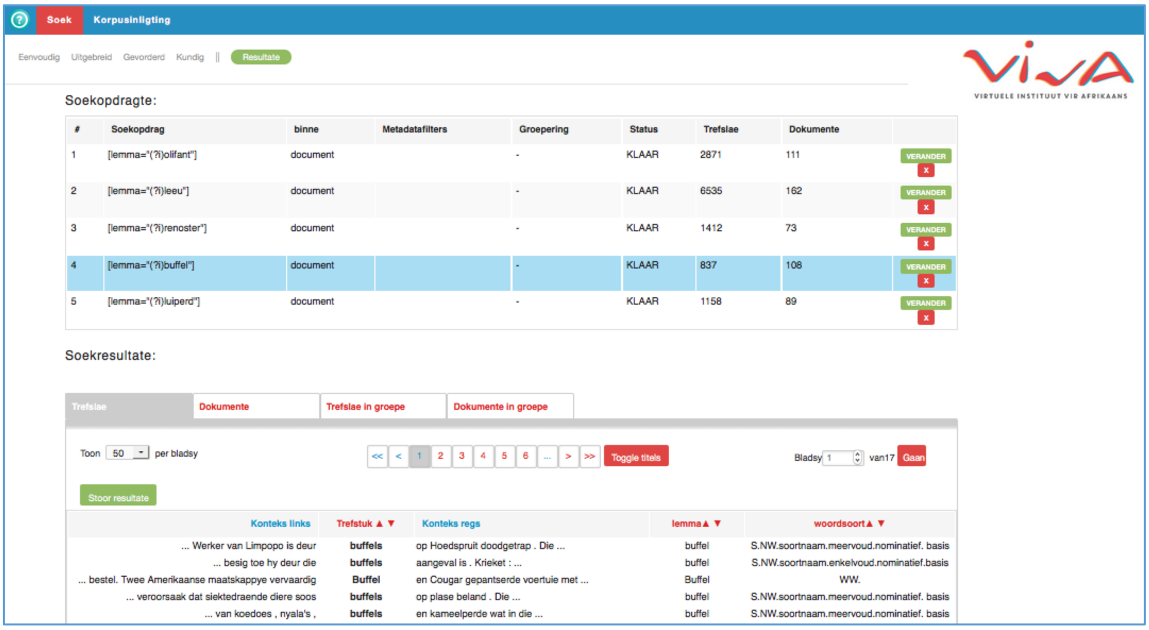
lemma: leeu
renoster
buffel

□ Kassensitief

woordsoort:

☑ Skei bondelnavrae

The ‘skei bondelnavrae’ box is always marked. That means every line in the search window is treated as a separate query. The results page looks like this:



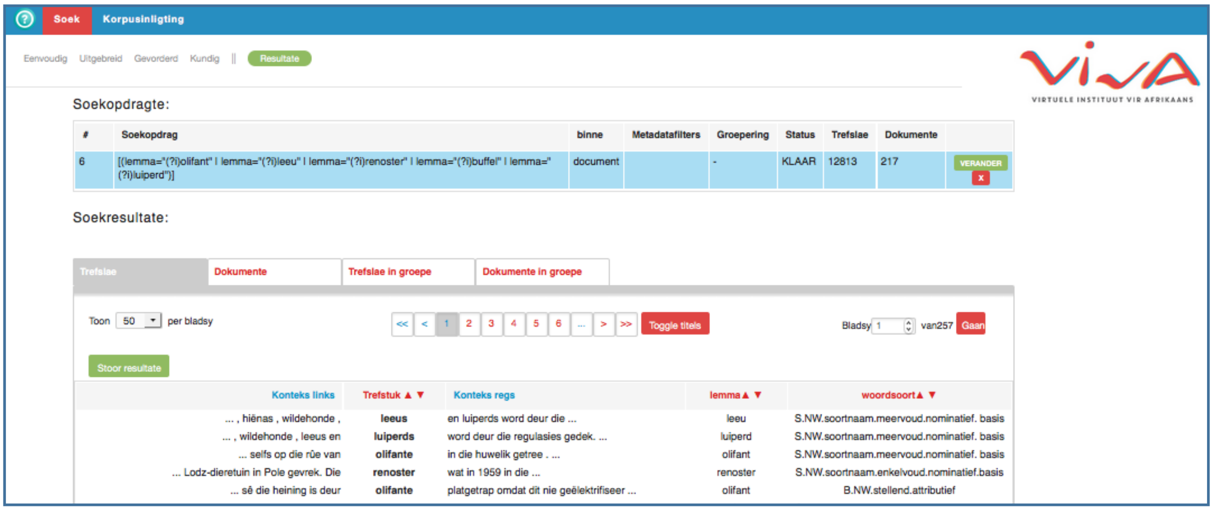
Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[lemma="(?i)olifant"]	document		-	KLAAR	2871	111	VERANDER x
2	[lemma="(?i)leeu"]	document		-	KLAAR	6535	162	VERANDER x
3	[lemma="(?i)renoster"]	document		-	KLAAR	1412	73	VERANDER x
4	[lemma="(?i)buffel"]	document		-	KLAAR	837	108	VERANDER x
5	[lemma="(?i)luiperd"]	document		-	KLAAR	1158	89	VERANDER x

Soekresultate:

Konteks links	Trefstuk ▲ ▼	Konteks regs	lemma ▲ ▼	woordsoort ▲ ▼
... Werker van Limpopo is deur	buffels	op Hoedspruit doodgetrap . Die ...	buffel	S.NW.soortnaam.meervoud.nominatief. basis
... besig toe hy deur die	buffels	aangeval is . Krieket : ...	buffel	S.NW.soortnaam.enkelvoud.nominatief.basis
... bestel. Twee Amerikaanse maatskappye vervaardig	Buffel	en Cougar gepantserde voertuie met ...	Buffel	WW.
... veroorsaak dat siektedraende diere soos	buffels	op plase beland . Die ...	buffel	S.NW.soortnaam.meervoud.nominatief. basis
... van koedoes , nyala's ,	buffels	en kameelperde wat in die ...	buffel	S.NW.soortnaam.meervoud.nominatief. basis

If you delete the ‘skei bondelnavrae’ box, all the items in the query will be converted to a single query.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
6	[(lemma="(?i)olifant" I lemma="(?i)leeu" I lemma="(?i)renoster" I lemma="(?i)buffel" I lemma="(?i)luiperd")]	document		-	KLAAR	12813	217	VERANDER x

Soekresultate:

Konteks links	Trefstuk ▲ ▼	Konteks regs	lemma ▲ ▼	woordsoort ▲ ▼
... , hiënas , wildehonde ,	**leeus**	en luiperds word deur die ...	leeu	S.NW.soortnaam.meervoud.nominatief. basis
... , wildehonde , leeus en	**luiperds**	word deur die regulasies gedek. ...	luiperd	S.NW.soortnaam.meervoud.nominatief. basis
... selfs op die rûe van	**olifante**	in die huwelik getree	olifant	S.NW.soortnaam.meervoud.nominatief. basis
... Lodz-dieretuin in Pole gevrek. Die	**renoster**	wat in 1959 in die ...	renoster	S.NW.soortnaam.enkelvoud.nominatief.basis
... sê die heining is deur	**olifante**	platgetrap omdat dit nie geëlektrifiseer ...	olifant	B.NW.stellend.attributief

The results can be consulted in the same way as described in section 4.2.1.

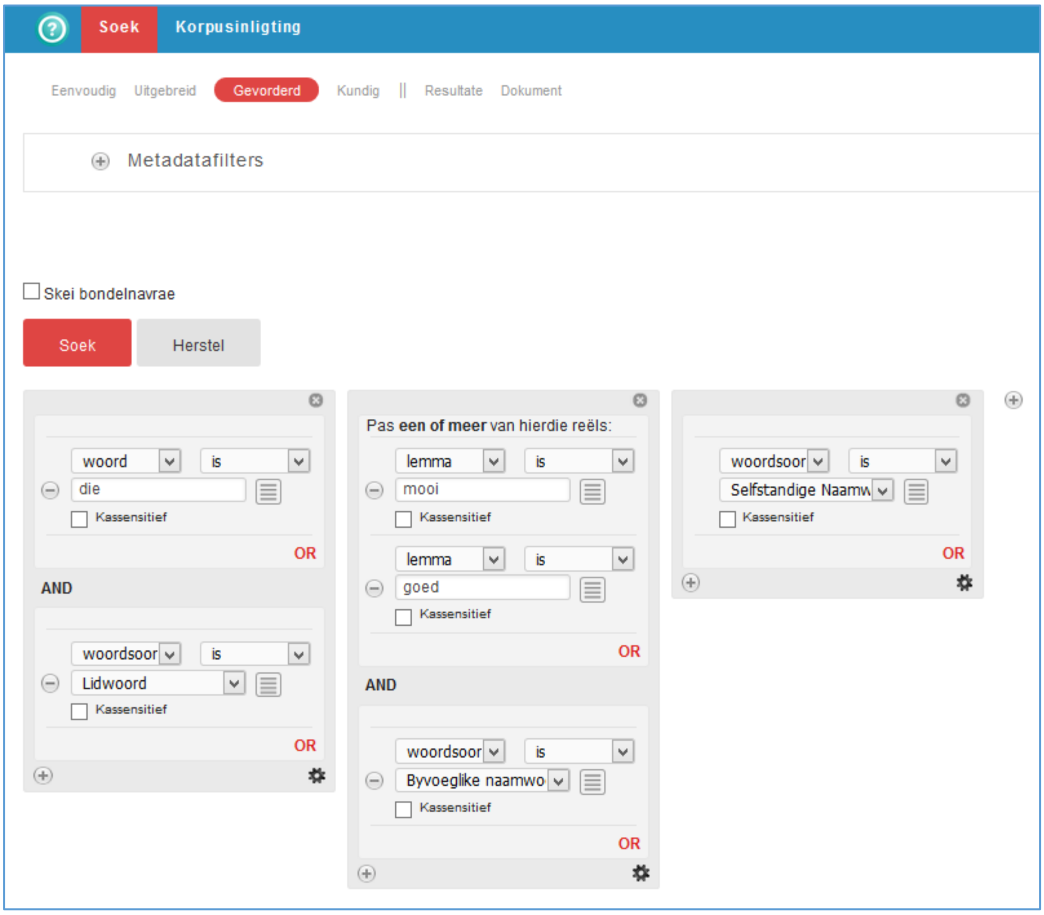
4.3 Advanced

4.3.1 Search command

The advanced search method allows you to filter metadata, as set out in section 4.2.1. This input method is a combination of the previous two. As in simple search mode, it is possible to define one or more search terms. As in the expanded search mode, it is possible to indicate whether you are interested in the exact word form, the lemma or the part of speech for each item. Combinations are also possible in this search mode. Each block indicates an element of the query. In that block, you can provide information about the word form, the lemma and/or the part of speech. By clicking on the + symbol to the right, you can link multiple elements to the search command. The search command below searches for combinations of ‘die’ as article, followed by ‘mooi’ of ‘goed’ as adjective, followed by a noun.

By clicking the + symbol at the bottom left of the box, you can apply an AND restriction. In this manner, you can indicate that you are looking for the word ‘die’ AND that it must be an article. By clicking on OR you can apply an OR restriction. For example, you can indicate that you are looking for the lemma ‘mooi’ OR the lemma ‘goed’.

The ⚙ button allows you to set additional conditions, for example, to display only hits that occur at the beginning of a sentence.



□ Skei bondelnavrae

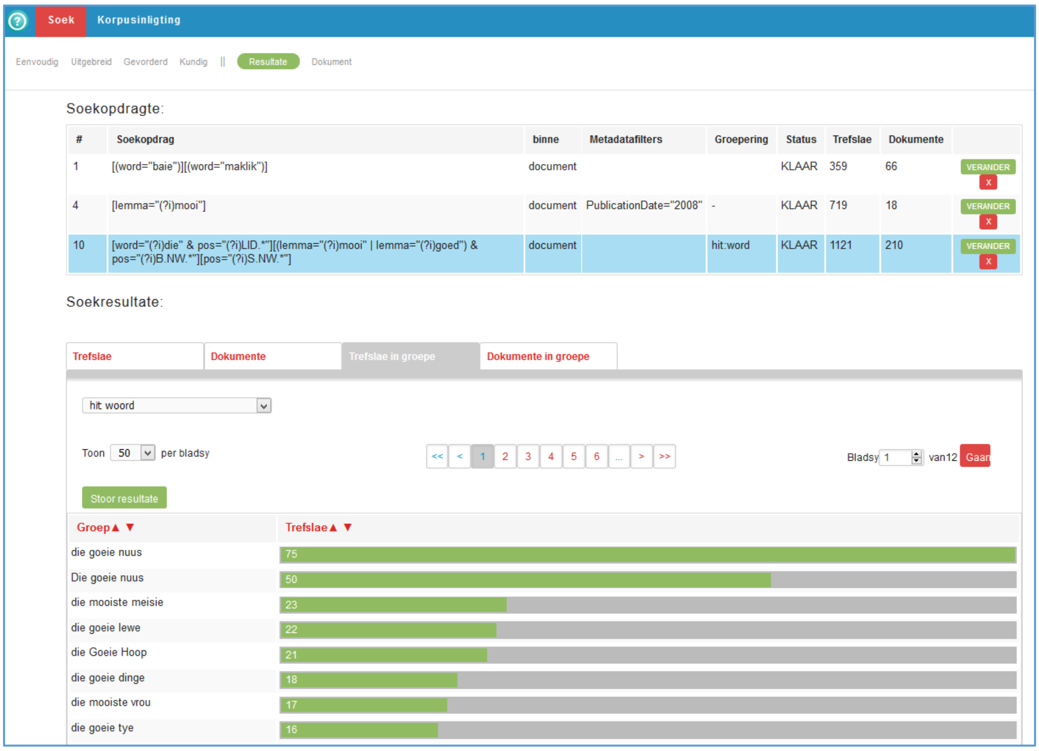
woord is die □ Kassensitief OR
AND
woordsoort is Lidwoord □ Kassensitief OR

Pas **een of meer** van hierdie reëls:
lemma is mooi □ Kassensitief
lemma is goed □ Kassensitief OR
AND
woordsoort is Byvoeglike naamwo □ Kassensitief OR

woordsoort is Selfstandige Naamw □ Kassensitief OR

4.3.2 Results

The results are displayed in the ‘Soekresultate’ window. As indicated in section 4.1.2, you can sort and download the results. The figure below shows the grouped hits, indicating that the string ‘die goeie nuus’ has the highest number of hits in terms of the given search command.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[(word="baie")][(word="maklik")]	document			KLAAR	359	66	VERANDER X
4	[lemma="(?i)mooi"]	document	PublicationDate="2008"	-	KLAAR	719	18	VERANDER X
10	[word="(?i)die" & pos="(?i)LID.*"][(lemma="(?i)mooi" | lemma="(?i)goed") & pos="(?i)B.NW.*"][pos="(?i)S.NW.*"]	document		hit:word	KLAAR	1121	210	VERANDER X

Soekresultate:

Groep▲ ▼	Trefslae▲ ▼
die goeie nuus	75
Die goeie nuus	50
die mooiste meisie	23
die goeie lewe	22
die Goeie Hoop	21
die goeie dinge	18
die mooiste vrou	17
die goeie tye	16

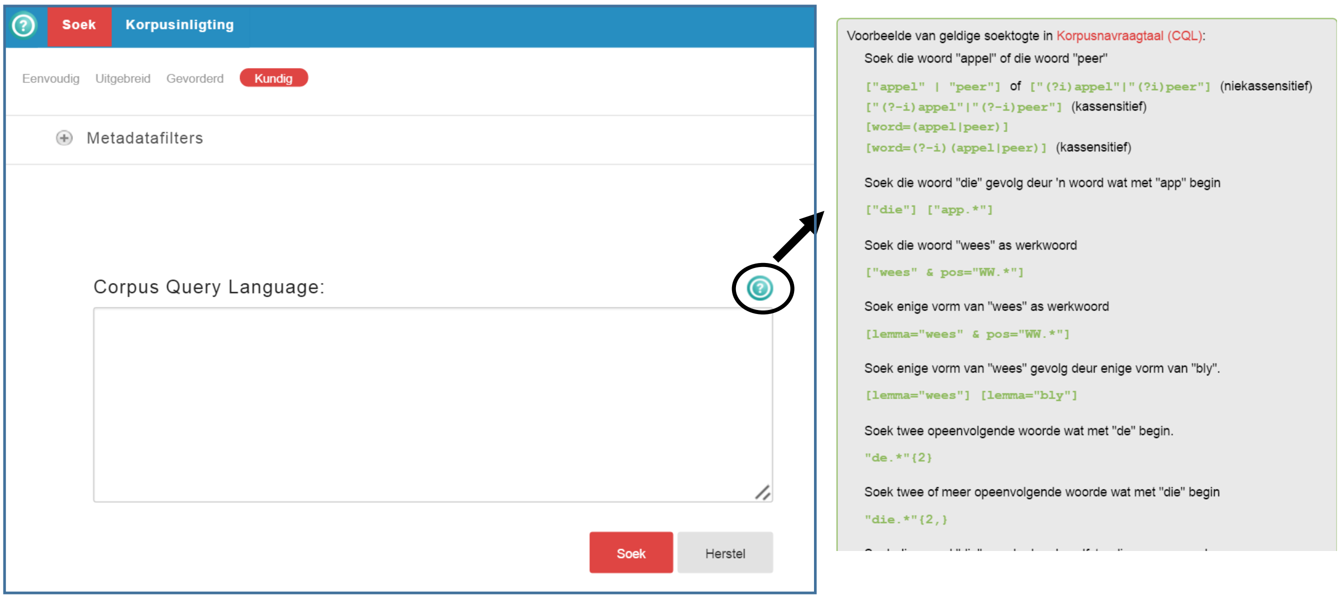
4.3.3 Batch queries

It is also possible in this search mode to perform bundle queries by loading a list of queries via the ☰ button. This function works in a similar way as described in section 4.2.3.

4.4 Expert [‘Kundig’]

4.4.1 Search command

‘Kundig’ is the most advanced search mode. The metadata can be filtered as set out in section 4.2.1. In addition, the interface consists of a text field in which you can give a search command in **Corpus Query Language (CQL)**, which makes it possible to define complex search assignments. For this, you need to know the CQL format and the labels used for the annotation of parts of speech. The advantage is that you have more control over the patterns you are looking for. If you click on the question mark icon, you will get a few examples of search commands in CQL format.



Voorbeelde van geldige soektogte in Korpusnavraagtaal (CQL):

Soek die woord "appel" of die woord "peer"

`["appel" | "peer"]` of `["(?i)appel"|"(?i)peer"]` (niekassensitief)
`["(?-i)appel"|"(?-i)peer"]` (kassensitief)
`[word=(appel|peer)]`
`[word=(?-i)(appel|peer)]` (kassensitief)

Soek die woord "die" gevolg deur 'n woord wat met "app" begin

`["die"] ["app.*"]`

Soek die woord "wees" as werkwoord

`["wees" & pos="WW.*"]`

Soek enige vorm van "wees" as werkwoord

`[lemma="wees" & pos="WW.*"]`

Soek enige vorm van "wees" gevolg deur enige vorm van "bly".

`[lemma="wees"] [lemma="bly"]`

Soek twee opeenvolgende woorde wat met "de" begin.

`"de.*"{2}`

Soek twee of meer opeenvolgende woorde wat met "die" begin

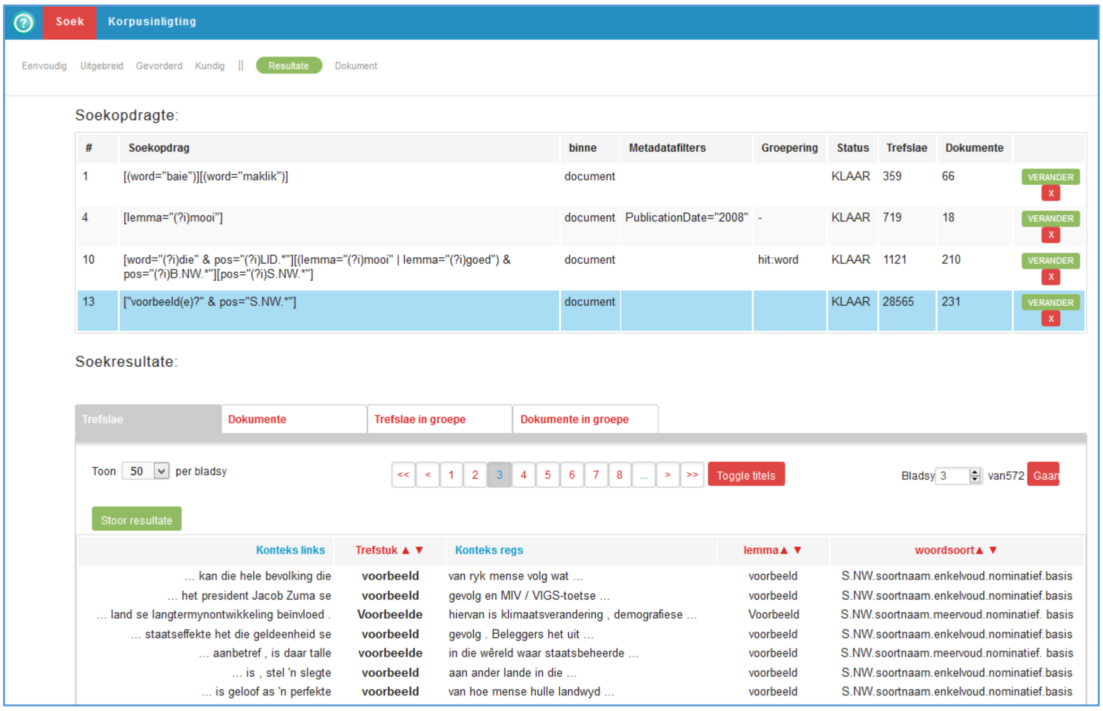
`"die.*"{2,}`

The following CQL query searches for ‘voorbeeld’ or ‘voorbeelde’ as noun. **[‘voorbeeld(e)?’ & pos=‘S. NW.*’].**

Section 5 contains an introduction to CQL, with reference to Corpus Portal. An overview of all abbreviations and operators may be found in the Addendum in the back of this user guide.

4.4.2 Results

The results are also displayed in the ‘Soekresultate’ window. As indicated in section 4.1.2, you can sort and download the results. The figure below indicates a number of hits where ‘voorbeeld’, ‘voorbeelde’ as well as ‘Voorbeeld’ and ‘Voorbeelde’ were annotated as nouns, i.e. where they have been labelled ‘S.NW.’.



Soekopdragte:

#	Soekopdrag	binne	Metadatafilters	Groepering	Status	Trefslae	Dokumente	
1	[(word="baie")][(word="maklik")]	document			KLAAR	359	66	VERANDER X
4	[lemma="(?i)mooi"]	document	PublicationDate="2008"	-	KLAAR	719	18	VERANDER X
10	[word="(?i)die" & pos="(?i)LID.*"][(lemma="(?i)mooi" | lemma="(?i)goed") & pos="(?i)B.NW.*"][pos="(?i)S.NW.*"]	document		hit:word	KLAAR	1121	210	VERANDER X
13	["voorbeeld(e)?" & pos="S.NW.*"]	document			KLAAR	28565	231	VERANDER X

Soekresultate:

Konteks links	Trefstuk ▲ ▼	Konteks regs	lemma▲ ▼	woordsoort▲ ▼
... kan die hele bevolking die	voorbeeld	van ryk mense volg wat ...	voorbeeld	S.NW.soortnaam.enkelvoud.nominatief.basis
... het president Jacob Zuma se	voorbeeld	gevolg en MIV / VIGS-toetse ...	voorbeeld	S.NW.soortnaam.enkelvoud.nominatief.basis
... land se langtermynontwikkeling beïnvloed .	Voorbeelde	hiervan is klimaatsverandering , demografiese ...	Voorbeeld	S.NW.soortnaam.meervoud.nominatief. basis
... staatseffekte het die geldeenheid se	voorbeeld	gevolg . Beleggers het uit ...	voorbeeld	S.NW.soortnaam.enkelvoud.nominatief.basis
... aanbetref , is daar talle	voorbeelde	in die wêreld waar staatsbeheerde ...	voorbeeld	S.NW.soortnaam.meervoud.nominatief. basis
... is , stel 'n slegte	voorbeeld	aan ander lande in die ...	voorbeeld	S.NW.soortnaam.enkelvoud.nominatief.basis
... is geloof as 'n perfekte	voorbeeld	van hoe mense hulle landwyd ...	voorbeeld	S.NW.soortnaam.enkelvoud.nominatief.basis

5 Corpus Query Language (CQL)

Corpus Query Language (CQL) is a formalized search language used to search annotated corpora. It is used by other search engines besides Corpus Portal. As indicated in section 4, CQL makes it possible to formulate more complex search commands. In order to accomplish this, you need to know a little about CQL and the labels used for the annotation of parts of speech. The advantage is that you have more control over the patterns you are looking for. This section provides an introduction the CQL format, as applicable to Corpus Portal. An overview of all abbreviations and operators may be found in the Addendum.

You may use CQL as input method in the search mode **'Kundig'** of Corpus Portal (see section 4). If you use a different search mode, the question will be converted automatically into a CQL query. You will be able to see this in the results window (under 'Soekopdragte'). In this manner, you may become familiar with the formulation without actively using it. You may also copy automatically generated CQL queries into the search mode 'kundig' and refine them. This may be simpler than to formulate the queries from scratch. And yet it may be useful to grasp some of the basics of CQL before you embark on this search. It is discussed in the next section.

Warning: It is important to use the **correct quotation marks** in CQL, or you will receive an error message. Keep this in mind when copying search commands from word processing programs such as MS Word or OpenOffice. Search commands copied from flat texts should not be a problem.

5.1 Simple search commands

In this section, a few simple CQL search commands are demonstrated. Similar search commands may form the building blocks for more complex search instructions.

5.1.1 A few basic principles

- Every element (word, lemma and/or part of speech) must be defined in **square brackets**: **[ ]**
- You can define the element you are looking for by typing **word=, lemma=,** or **pos=** before it.
- When you define a word, lemma or label for a part of speech, you should place it between **straight (single ' ' or double " ") quotation marks**.

5.1.2 Search for a word, lemma or part of speech.

- Search for all the instances of the word ‘lekker’. **[word='lekker']** or **['lekker']** or **'lekker'**
  This search command searches for the exact word ‘lekker’ but is not case sensitive. Sentences with ‘Lekker’ are therefore also displayed in the results, but sentences with ‘lekkerder’ or ‘lekkerste’ are not displayed.
  If you search for the word form only, you do not have to specify the label ‘word=‘ and the brackets.
- Search for all the instances of the lemma ‘lekker’. **[lemma='lekker']**
  This search command searches for sentences in which ‘lekker’ and all its derivatives appear, i.e. ‘lekker’, ‘lekkerder’, ‘lekkerste’, etc.
- Search for constructions with a definite article. **[pos='LID.bepaald']**
  This search command will search for items labelled ‘LID.bepaald’, i.e. ‘die’ and ‘Die’.
- With the & symbol, you are able to define more characteristics of an element, such as the lemma and the part of speech. This enables you to search for ‘groot’ as attributive adjective. This will exclude all instances of ‘groot’ as predicative adjective.
  [lemma='groot' & pos='B.NW.stellend.attributief']
  The search results will render hits such as ‘groot deurbraak’, ‘groot kinders’, ‘die groot familie’.

5.1.3 Search for adjacent words, lemmas and/or parts of speech.

You are able to search various elements, such as two adjacent words or combination of words, lemmas or parts of speech by defining the various elements in square brackets.

- Search for instances of ‘baie lekker’. **[word='baie'] [word='lekker']** or **['baie'] ['lekker']** or **'baie' 'lekker'**
  The results contain constructions with ‘baie lekker’, ‘Baie lekker’ and ‘BAIE LEKKER’. In this case the brackets and the label **'word='** are optional as well.
- You are also able to combine elements. For instance, you can search for constructions that contain the word ‘die’, and an attributive adjective and the lemma ‘man’.
  [word='die'] [pos='B.NW.stellend.attributief'][lemma='man']
  Examples of hits are: ‘die 40-jarige man’, ‘die gesogte man’, ‘Die gewapende mans’.
- You can define multiple characteristics per element. In this manner, you can search for instances of ‘geveg’ as noun, where it is followed by ‘het’.
  ['geveg' & pos='S.NW.soortnaam.enkelvoud.nominatief.basis'] ['het']
  Hits will include constructions such as ‘die geveg het begin’. Constructions containing ‘geveg’ as verb will be excluded.

5.1.4 Search non-adjacent constructions

You can search for constructions where the elements are not adjacent, such as the combination of a verb and a particle. By using square brackets **[ ]** you are able to indicate the number of items between the two elements.

- Search for constructions with ‘die man’ containing an arbitrary item between ‘die’ and ‘man’.
  ‘die’ [] ‘man’
  The result contains constructions such as ‘die ander man’, ‘die 50-jarige man’, ‘die jong man’.

You can indicate more elements between the two items by using more bracket pairs: **'die' [] [] 'man'** or **'die' [][][] 'man'.** For searches where the two elements are placed far apart, this is not an effective search method. In such cases, accolades **{ }** should be used to indicate the distance between the elements. The search command **'die' [][][] 'man'** can therefore also be formulated as **'die' []{3} 'man'.**

- Search for constructions with ‘aflaai’ which contain two items between the verb and the particle.
  'laai' []{2} [lemma='af' & pos='U.partikel.ww. ']
  The result contains constructions like ‘toe gaan laai ek haar af’, ‘laai dit nou af’, ‘Joe laai sy dogter af’.

You can also define an interval, for example, that between two and five elements may occur between the verb and the particle. **'laai' []{2,5} [lemma='af' & pos='U.partikel.ww.']** That generates hits such as ‘die ander man’, ‘laai my asseblief by die kantoor af’, ‘laai jou weer veilig af’.

5.2 Boolean operators

Boolean operators are **logical operators** used regularly in informatics and mathematics. CQL also uses them. The three operators in CQL are AND, OR and NOT. Operators enable you to express both conjunction, disjunction and negation. This allows you to formulate more advanced search commands.

5.2.1 AND

The use of the AND operator is discussed in section 5.1.2. With the **&** symbol you can define an AND ratio. For example, you can search for combinations of lemmas and part of speech annotations. This way you can find instances of ‘groot’ as predicative adjective in Corpus Portal using the search query **[lemma='groot' & pos='B.NW.stellend.predikatief']**. Examples of hits are ‘die impak sal groot wees’, ‘die risiko is te groot’.

5.2.2 OR

With the OR operator, you can search for more than one element at a time in one search. The OR ratio is expressed with the **|** symbol. The CQL command **['klein'|'groot']** searches for instances of the word ‘klein’ or the word ‘groot’. Examples of hits are ‘'n groot gedeelte’, ‘klein entrepreneurs’. Another formulation that will yield the same result is **[word='klein|groot'], [word='klein'|word='groot'].**

The OR function not only works with words; you can also search for different lemmas or parts of speech at the same time. The search command **[pos='VNW.aanwysend|VNW.vraend']** renders results that contain a demonstrative (e.g, ‘hierdie’, ‘dié’) or interrogative pronoun (e.g, ‘wie’, ‘hoe’, ‘wat’).

This search command can also be formulated as follows:
[pos='VNW.aanwysend'|pos='VNW.vraend'].
The notation **[pos='VNW.aanwysend'|'VNW.vraend']** is not an alternative, because the search will be for the demonstrative pronoun as part of speech, but for the interrogative pronoun as word.

5.2.3 NOT

The NOT function allows you to exclude certain elements from a search. You can do this in CQL by placing an **exclamation point before the equal sign (!=)**. The search **['suid'] [word!='Afrika']** is going to search for hits in which the word ‘suid’ is followed by another word, where the second word may not be ‘Afrika’. For example, the results will include ‘suid van die Sahara’.

It is possible to indicate that one element must comply with condition A, but not with condition B. The search command
['suid'] [word!= 'Afrika' & pos='EIE.eienaam.enkelvoud.nominatief.basis']
will render hits where the second element is a proper noun, but not ‘Afrika’. The results contain, among other things, the erroneous ‘Suid Korea’ and ‘Suid Soedan’, but no combinations such as ‘suid van ...’

5.3 Fixed expressions

The command for **fixed expressions** (regex) enables you to define a search pattern. As with Boolean operators, fixed expressions have not been created specifically for CQL, but different applications do make use of it. A simple application is the ‘find and replace’ function in text processors. This allows you to search for specific patterns (or strings) in a text and you can use your wildcard to search for different patterns simultaneously. In Corpus Portal, you can use fixed expressions to search for words, but you can also search for lemmas or parts of speech.

The usual letters and numbers in the fixed expression correspond with the same characters in the text: the search **[word='3-duisend']** will find the word ‘3-duisend’ in the corpus.

This section lists the most important **special characters** that you can use in conjunction with CQL.

5.3.1 Wild card for a random character

A **fullstop** (.) indicates a random character. In this way, the search command **['b.l']** renders *inter alia* ‘bal’, ‘bol’, ‘bil’ and ‘BOL’.

Fixed expressions are by default **not case sensitive.**

5.3.2 Grouping of characters and patterns

By using **square brackets ([])** you can define a character class, i.e. a list of possible characters. The fixed expression **['b[aie]l']** searches for ‘bal’, ‘bil’ and ‘bel’ (but not ‘bol’ or ‘bul’).

In section 5.2.2, the OR operator | is explained. You can use it to indicate **optionality** between full words or parts of speech. For example, if you are looking for all instances of plural possessive pronouns, you can use the following search command:

[pos='VNW.eerste.meervoud.besitlik'|pos='VNW.tweede.meervoud.besitlik'|pos='VNW. derde.meervoud.besitlik'].
An alternative would be **[pos='VNW.(eerste|tweede|derde).meervoud.besitlik'],** where you indicate differences in round brackets.

An additional example is **['ou(m|p)a'],** which will render ‘ouma’ as well as ‘oupa’ and thus correspond with [‘ouma|oupa’].

5.3.3 Quantification

The **quantifiers +, ? and *** indicate how often a particular character would appear in the search pattern.

- The **question mark (?)** indicates that the previous character can occur 0 or 1 time. **['voorbeelde? ']** searches for instances of ‘voorbeeld’ and ‘voorbeelde’. The quantifiers can be used in the same way as the OR operator in conjunction with groupings in round brackets. **['seun(tjie)?']** renders ‘seun’ and ‘seuntjie’.
- The **plus sign (+)** indicates that the previous character (or group of characters) may occur 1 or more times. **['do+r']** searches for ‘dor’, ‘door’, ‘dooor’, etc.
- The **asterisk (*)** indicates that the previous character (or group of characters) may occur 0, 1 or more than 1 times. **['do*r']** searches for ‘dr’, ‘dor’, ‘door’, ‘dooor’, etc.

A regular fixed expression is **['. *']** that detects all text. The search is also useful to do general part of speech annotation searches. For example, if you search for all pronouns, you can use the search request **[pos='VNW.*']**.

You can use the quantifier within an element, or to define 0, 1 or multiple occurrences of the same elements. In this way you can search for constructions with one or more verbs using the query **[pos='WW.*']+.**

You can also use **accolades** to indicate how often a particular element may occur. In this way the fixed expression **['do{2}r']** searches for ‘door’, **['do{1,2}r']** searches for ‘dor’ and ‘door’. **['hallo{2,}']** searches for constructions with two or more o characters, such as ‘halloo’ and ‘hallooo’.

This method of quantification also allows you to use CQL for the complete element.

[pos='WW.*']{2} searches for constructions with two consecutive verbs, such as ‘opgelos word’. **[pos='WW.*']{2,3}** searches for constructions with two or three consecutive verbs, such as ‘opgelos word’ and ‘probeer insmokkel het’. **[pos='WW.*']{2}** searches for constructions with two or more consecutive verbs.

5.3.4 Escape sign

Sometimes you want to search for the **actual version of the special characters** in the corpus. You can do this by placing a backslash (\) before that character. **['\.']** searches for a full stop, **['\?']** for a question mark, **['\(']** for a bracket. The backslash itself can be found with **['\\'].**

5.3.5 Case sensitive search commands

In the previous sections it was indicated that CQL searches are not case sensitive. If you wish to formulate a case sensitive command, you can add the operator (?-i) to the case sensitive section of the search.

['(?-i) Lekker'] searches for instances of ‘Lekker’. Sentences with ‘lekker’ or ‘LEKKER’ are not shown in these results.

In search commands with multiple elements, or if the OR operator is used, the (?-i) add-on does not relate to the full command. **['(?-i)Lekker'|'dag']** renders constructions with ‘Lekker’, ‘dag’, ‘Dag’ and ‘DAG’. If you also want to make the second element of the disjunction case sensitive, you should do the following: **['(?-i)Lekker'|'(?-i)dag']** or **['(?-i)(Lekker|dag)'].** The search engine only looks for ‘Lekker’ and ‘dag’.

6 Addendum

6.1 Corpora in Corpus Portal

There are three levels of access to VivA's Corpus Portal:

- **OPEN**: The corpora in this group are freely available to all VivA users for research purposes.
- **EXTENSIVE**: The corpora in this group also include the corpora in the OPEN level. Access is restricted to *bona fide* researchers. Apply for access to these corpora by completing the online application form.
- **EXCLUSIVE**: The corpora in this group are only available to bona fide researchers who can demonstrate that the corpora are required for a specific project. Apply for access to these corpora by completing the online application form .

Level	Name	Number of words	Number of text units
OPEN	NCHLT- Afrikaanse korpus 1.0	2 229 214	2 489 716
EXTENSIVE	Afrikaanse Leipzig-korpus 1.0	28 776 800	32 269 153
EXTENSIVE	NWU/Maroela Media-korpus 1.2	9 173 430	10 378 957
EXTENSIVE	NWU/Lapa-korpus 1.1	9 804 270	11 639 129
EXTENSIVE	PUK/Protea Boekhuis-korpus 2.1	8 022 403	9 255 228
EXTENSIVE	RSG-nuuskorpus 2.1	14 829 223	16 160 631
EXTENSIVE	Taalkommissie-korpus 1.1	47 321 344	53 622 677
EXTENSIVE	Wikipedia- Afrikaanse korpus 1.0	11 523 680	13 119 966
EXCLUSIVE	Watkykjy.co.za-korpus 1.2	1 236 214	1 406 709
	TOTAL	132 916 578	150 342 166

All corpora in Corpus Portal are automatically lemmatized (approximately 90% accurate) and provided with word types (approximately 75% accurate). The annotations are not corrected; users should take into account possible annotation errors.

Use the following reference to refer to this corpus collection:

Virtuele Instituut vir Afrikaans (VivA). 2016. Korpusportaal. Available at: http://viva-afrikaans.org.

6.2 Linguistic annotation

6.2.1 Labels

Annotation	Label	CQL example	Examples
Word	word	‘rooi’ [word='rooi']	*rooi, Rooi*
Lemma	lemma	[lemma='rooi']	*rooi, Rooi, rooie, Rooie*
Part of speech	pos	[pos='B.NW.*']	*toegesneeude, flikkerende, goeie*

6.2.2 General parts of speech

Part of speech	Abbreviation	CQL search term
Adjective	B.NW.	[pos='B.NW.*']
Adverb	BW.	[pos='BW.*']
Article	LID	[pos='LID.*']
Noun - Substantive - Proper noun	 S.NW. EIE.	[pos='S.*|EIE.*'] [pos='S.NW.*'] [pos='EIE.*']
Punctuation	U.	[pos='U.(sinseinde|sinmiddel|links.*|regs.*)']
Numeral	TW.	[pos='TW.*']
Interjection	TSW.	[pos='TSW.*']
Conjunction	VG.	[pos='VG.*']
Pronoun	VNW.	[pos='VNW.*']
Preposition	VS.	[pos='VS.*']
Verb	WW.	[pos='WW.*']

6.2.3 Parts of speech

Labels in Corpus Portal. Found by part of speech via ‘Kundig’, e.g. [pos='S.NW.*'], Hits in groups, sorted on 'hit woordsoort'.

Part of speech	Label
Adjective	
	B.NW.stellend.attributief
	B.NW.stellend.predikatief

	B.NW.oortreffend.attributief
	B.NW.vergrotend.attributief
	B.NW.vergrotend.predikatief
	B.NW.oortreffend.predikatief
Adverb	
	BW.oortreffend
	BW.stellend
	BW.vergrotend
Article	
	LID.bepaald
	LID.onbepaald
Noun	
	S.NW.abstrak
	S.NW.maatnaam.enkelvoud.basis
	S.NW.massanaam
	S.NW.soortnaam.enkelvoud.basis
	S.NW.soortnaam.enkelvoud.diminutief
	S.NW.soortnaam.meervoud.basis
	S.NW.soortnaam.meervoud.diminutief
Proper noun	
	EIE.eienaam.enkelvoud.basis
Punctuation	
	U.sinseinde
	U.sinmiddel
	U.links-parentese
	U.regs-parentese
Particle	
	U.partikel.infinitief
	U.partikel.ontkenning
	U.partikel.ww.
	U.partikel.genitief
	U.partikel.vergelyking
	U.partikel.deel
	U.partikel.graad
	U.partikel.betreklik

Unique	
	U.eks-daar
	U.woorddeel
Residue	
	R.afkorting
	R.akroniem.letterklankwoord
	R.akroniem.letternaamwoord
	R.ongeklassifiseerd
	R.simbool
	R.vreemdetaalwoord
Numeral	
	TW.hooftelwoord.adjektief.bepaald
	TW.hooftelwoord.adjektief.onbepaald
	TW.hooftelwoord.bywoord.bepaald
	TW.hooftelwoord.bywoord.onbepaald
	TW.hooftelwoord.voornaamwoord.bepaald
	TW.hooftelwoord.voornaamwoord.onbepaald
	TW.rangtelwoord.adjektief.bepaald
	TW.rangtelwoord.adjektief.onbepaald
	TW.rangtelwoord.bywoord.bepaald
	TW.rangtelwoord.bywoord.onbepaald
Conjunction	
	VG.neweskikkend
	VG.onderskikkend
Pronoun	
	VNW.aanwysend
	VNW.betreklik
	VNW.derde.manlik.enkelvoud.besitlik
	VNW.derde.manlik.enkelvoud.gemarkeerd.persoonlik
	VNW.derde.manlik.enkelvoud.ongemarkeerd.persoonlik
	VNW.derde.manlik.enkelvoud.wederkerend
	VNW.derde.meervoud.besitlik
	VNW.derde.meervoud.persoonlik
	VNW.derde.meervoud.wederkerend
	VNW.derde.onsydig.enkelvoud.ongemarkeerd.persoonlik

	VNW.derde.onsydig.enkelvoud.wederkerend
	VNW.derde.vroulik.enkelvoud.besitlik
	VNW.derde.vroulik.enkelvoud.gemarkeerd.persoonlik
	VNW.derde.vroulik.enkelvoud.ongemarkeerd.persoonlik
	VNW.eerste.enkelvoud.besitlik
	VNW.eerste.enkelvoud.gemarkeerd.persoonlik
	VNW.eerste.enkelvoud.ongemarkeerd.persoonlik
	VNW.eerste.meervoud.besitlik
	VNW.eerste.meervoud.persoonlik
	VNW.eerste.meervoud.wederkerend
	VNW.onbepaald
	VNW.tweede.enkelvoud.besitlik
	VNW.tweede.enkelvoud.gemarkeerd.persoonlik
	VNW.tweede.enkelvoud.ongemarkeerd.persoonlik
	VNW.tweede.enkelvoud.wederkerend
	VNW.tweede.meervoud.persoonlik
	VNW.vraend
	VNW.wederkerig
Preposition	
	VS.voorsetsel
Verb	
	WW.gemarkeerd.hoof.onskeibaar.koppel
	WW.gemarkeerd.hoof.onskeibaar.onoorganklik
	WW.gemarkeerd.hoof.onskeibaar.oorganklik
	WW.hulp.onskeibaar.tyd
	WW.ongemarkeerd.hoof.onskeibaar.koppel
	WW.ongemarkeerd.hoof.onskeibaar.onoorganklik
	WW.ongemarkeerd.hoof.onskeibaar.oorganklik
	WW.ongemarkeerd.hoof.onskeibaar.voorsetsel
	WW.ongemarkeerd.hoof.skeibaar.onoorganklik
	WW.ongemarkeerd.hoof.skeibaar.oorganklik
	WW.teenwoordig.hulp.onskeibaar.aspek
	WW.teenwoordig.hulp.onskeibaar.modaliteit
	WW.teenwoordig.hulp.onskeibaar.modus
	WW.verlede.hulp.onskeibaar.modaliteit

	WW.verlede.hulp.onskeibaar.modus
	WW.gemarkeerd.hoof.onskeibaar.koppel
	WW.gemarkeerd.hoof.onskeibaar.onoorganklik
	WW.gemarkeerd.hoof.onskeibaar.oorganklik
	WW.hulp.onskeibaar.tyd
	WW.ongemarkeerd.hoof.onskeibaar.koppel
	WW.ongemarkeerd.hoof.onskeibaar.onoorganklik
	WW.ongemarkeerd.hoof.onskeibaar.oorganklik
	WW.ongemarkeerd.hoof.onskeibaar.voorsetsel
	WW.ongemarkeerd.hoof.skeibaar.onoorganklik
	WW.ongemarkeerd.hoof.skeibaar.oorganklik
	WW.teenwoordig.hulp.onskeibaar.aspek
	WW.teenwoordig.hulp.onskeibaar.modaliteit
	WW.teenwoordig.hulp.onskeibaar.modus

6.2.4 Boolean operators

Operator	**Symbol**	**Example**	**Some results**
OR	|	'die|hierdie|daardie' 'hond'	*die hond, daardie hond, Die hond, Hierdie hond*
AND	&	[lemma='sy' & pos='VNW.*']	*langs sy bevrore lyk, Tom draai weer na sy papier*
NOT	!	[lemma='assos.*' & pos!='S.*']	*assosieer, geassosieer, Outomobiel Assosiasie*

6.2.5 Regex operators

Search command	**Input**	**Example**	**Some results**
Random sign	.	'ma.'	*man, mal, mat, mag*

Character class	[ ]	['b[aie]l']	*bal, bil, bel*
Disjunction	|	'ouma|oupa'	*Ouma, oupa*
Grouping	( )	['ou(m|p)a']	*Ouma, oupa*
Quantification: 0 or 1	?	['voorbeelde?']	*voorbeeld, voorbeelde*
Quantification: 1 or more	+	['do+r']	*dor, door dooor*
Quantification: 0, 1 or more	*	['do*r']	*dr, dor, door, door*
Undetermined number of random characters (zero included)	.*	'assos.*'	*assosieer, assosiasie, assosiaatdirekteur*
Determined number	{*number*}	'do{2}r’	*door*
Interval	{*minimum number*, *maximum number*}	'do{1,2}r'	*dor ,door*
Minimum number	{*minimum number*,	'hallo{2,}'	*halloo, hallooo*
Wild card	\		
Case sensitive	(?-i)	[lemma='(?-i)Man']	*Man, Mans, Manne, Mannetjies*

6.2.6 Quantification of search environment

Operator	**Symbol**	**Example**	**Some results**
Random word	[]	'die' []'vrou'	*die eerste vrou, dié 24-jarige vrou,...*
Determined number	{*number*}	'baie' []{3} 'huise'	*baie van die groot huise, baie bly dat hulle huise*

Interval	{*minimum*, *maximum*}	[pos='WW.*']{2,3}	*kan help, geplant het, ontdek kan word, gaan sien het*
Minimum	{*minimum number*,	'die' [pos='B.NW.*']{2,} 'man'	*Dié merkwaardige jong man, die ware testosteroon-belaaide Afrikaanse man*