

Guide to the LANCHART search engine (dgcsearch.ku.dk)

Philip Diderichsen, september 2020

1. Short introduction to the LANCHART corpus

The LANCHART Centre (LANCHART is short for *language change in real time*) is a research centre at the University of Copenhagen focusing on Danish spoken language change. The core resource of the centre is the LANCHART corpus, a world-class corpus of sociolinguistic interviews. The corpus has been built up through several rounds of recordings of the same informants across several decades. New recordings, transcriptions, and linguistic markup from various research projects are being added on a regular basis. The corpus is searchable through a web interface, which is described in this document.

2. Access to the search engine

The search engine can be accessed online if you are a registered associate of the University of Copenhagen and have an official ID (KU-ID, of the form abc123). It is further required that you sign a non-disclosure agreement, and that you are registered as a user of the search engine through identity.ku.dk.

See the separate guide "Access to dgcsearch.ku.dk".

3. Search

Searches are performed by specifying a subset of the corpus and a set of search criteria. The results can be displayed in various formats. Searches are performed from the search engine front page, see Figure 1.

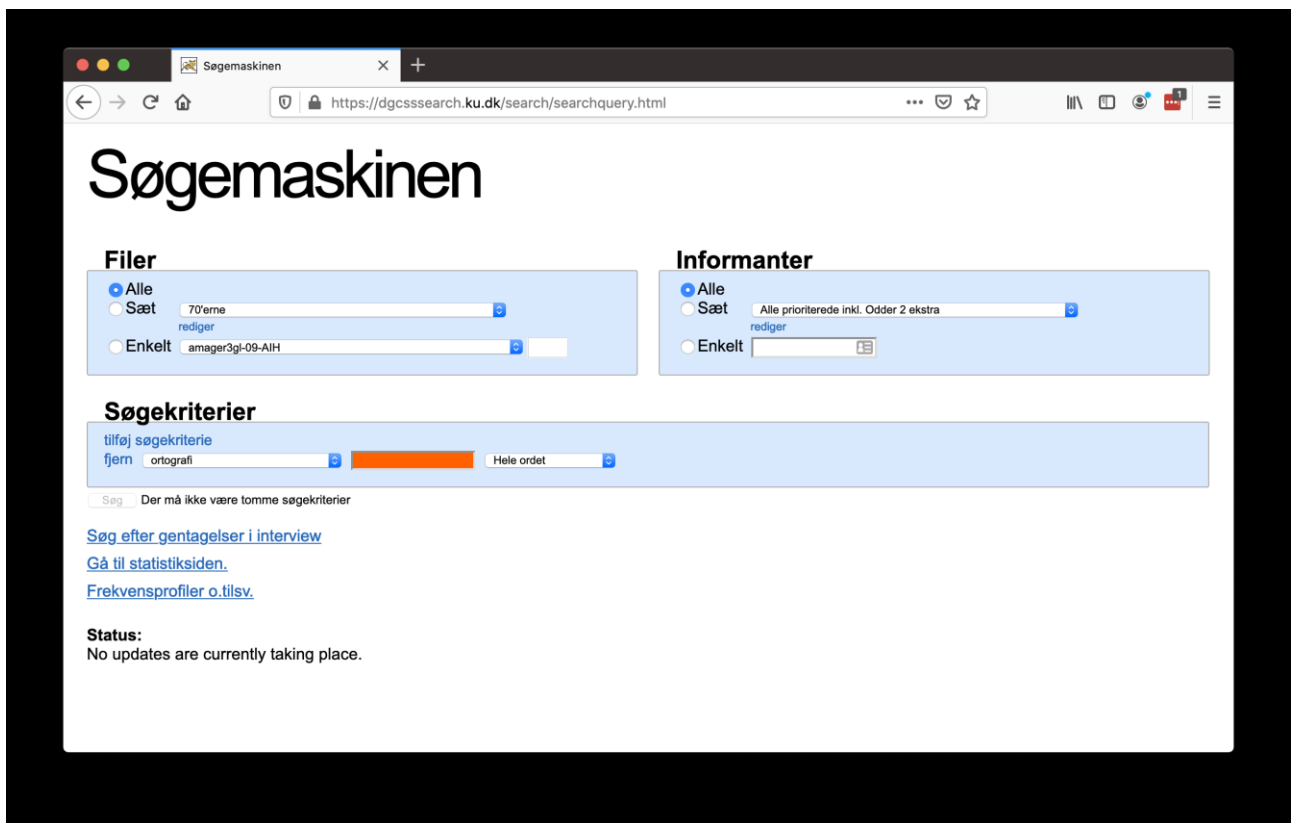


Figure 1. Front page of the search engine. The search term is entered in the orange field.

3.1. Files

Under "Filer" ('files'), one may specify which subset of the LANCHART corpus one wishes to search in. The corpus consists of a sizeable collection of transcriptions of the interviews and conversations belonging to the various projects of the LANCHART Centre. Each conversation corresponds to a single file in the corpus. Read more about the individual projects here: <https://dgcshum.ku.dk/forskning/undersogelsesomraader>.

<p>Filer</p> <p> <input checked="" type="radio"/> Alle <input type="radio"/> Sæt 70'erne rediger <input type="radio"/> Enkelt amager3gl-09-AIH </p>	<p>Select the option "Alle" ('all') if you wish to search all transcriptions.</p>
<p>Filer</p> <p> <input type="radio"/> Alle <input checked="" type="radio"/> Sæt <input type="radio"/> Enkelt </p> <div> 70'erne 80'erne 90'erne alle filer fra København alle filer fra København alle filer fra Næstved </div>	<p>To search in a specific part of the corpus, choose the option "Sæt" ('set'). A specific, named subset of files can then be chosen from the drop-down menu.</p>
<p>Filer</p> <p> <input type="radio"/> Alle <input type="radio"/> Sæt 70'erne rediger <input checked="" type="radio"/> Enkelt </p> <div> 70'erne 80'erne 90'erne alle filer fra København alle filer fra København alle filer fra Næstved </div> <div> amager3gl-09-AIH amager3gl-09-FWA amager3gl-09-HCM </div>	<p>It is also possible to search in a single file. Choose the option "Enkelt" ('single'), and select the desired file.</p>

It is even possible to create a new subset of files if the existing sets are not sufficient. Click "Rediger" ('edit') just below the "Sæt" ('set') field. This will send you to the page "Rediger interviewsæt" ('edit interview set').

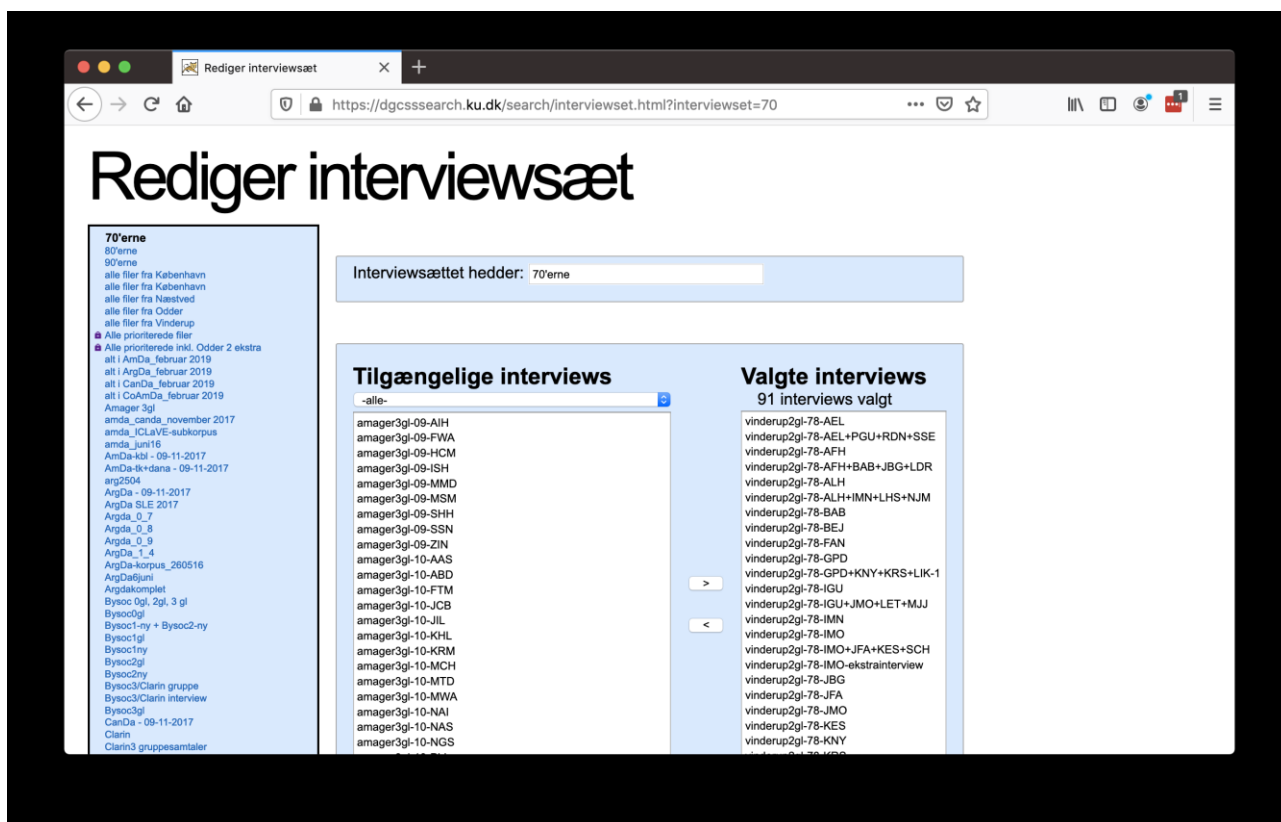


Figure 2. The page "Rediger interviewsæt" ('edit interview set'). On this page one can select a subset of transcriptions to search in.

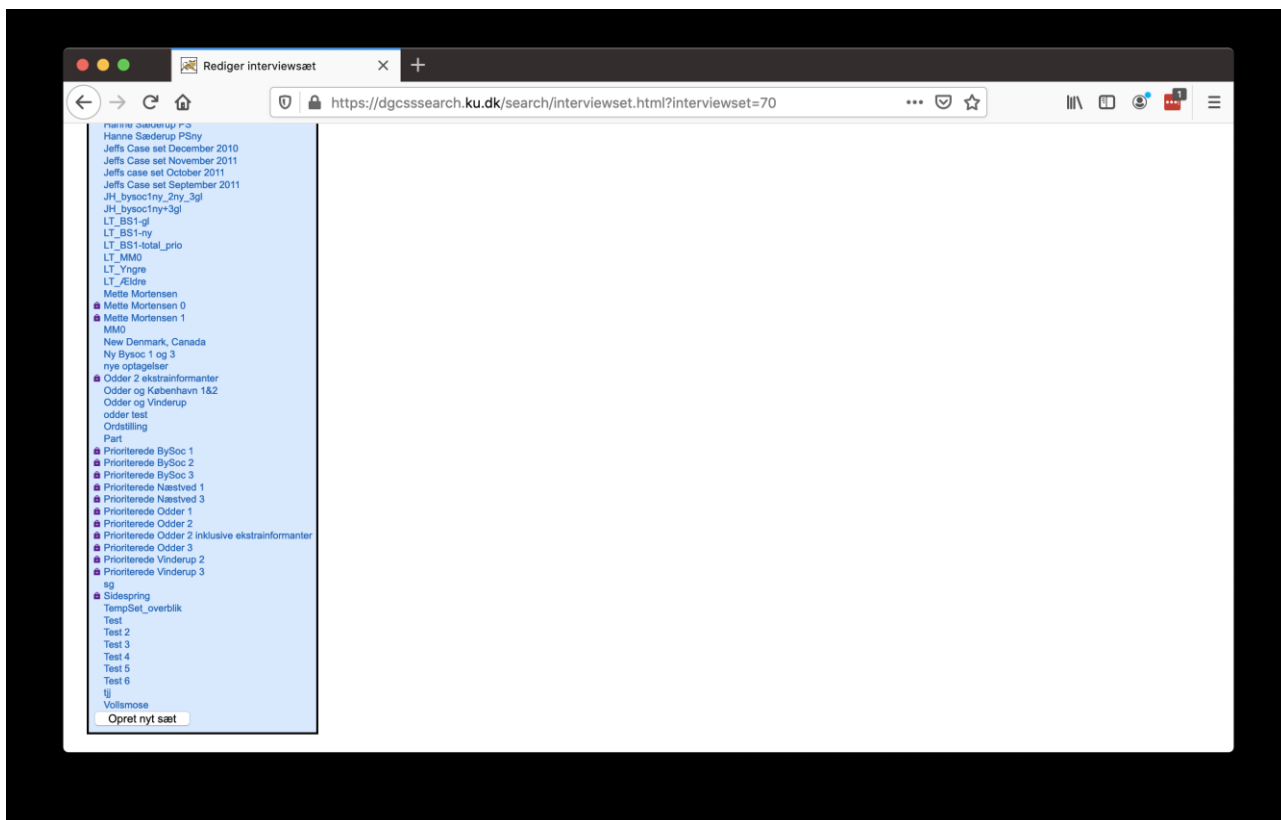
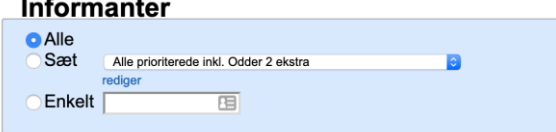
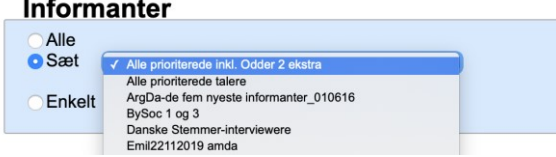
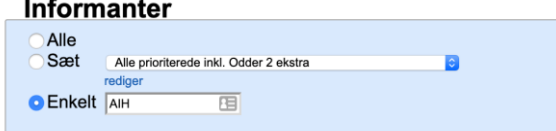


Figure 3. The bottom of the page "Rediger interviewsæt", where new interview sets are created by clicking "Opret nyt sæt" ('create new set').

On the page "Rediger interviewsæt" ('edit interview set'), scroll to the bottom, and click the button "Opret nyt sæt" ('create new set'). Choose a (new, nonexistent) name for the interview set, and enter it in the field "Interviewsættet hedder" ('name of the interview set'). The desired files can then be selected in the left column and moved to the right column using the right arrow button between the columns. Save the changes to create the new interview set.

3.2. Informants

Under "Informanter" ('informants') on the front page, you can specify which informants in the corpus you are interested in. Only results from the selected informants will then appear in the results.

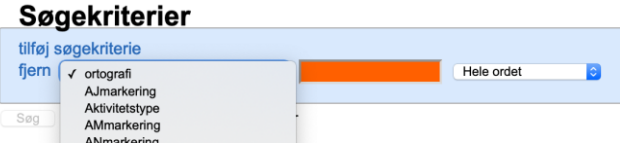
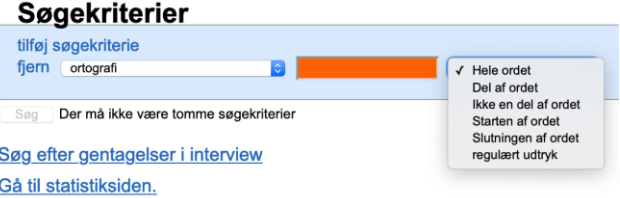
	<p>If you want to search in data from all informants, select the option "Alle" ('all'). This will include everyone, even incidental informants like people only appearing with a single 'hi' in passing, babies babbling, and the like.</p>
	<p>It is also possible to select named subsets of informants. For instance, all interviewers, all informants in a certain age group, or all informants of a certain gender. The option "Sæt" ('set') is selected, and the desired subset is selected in the drop-down menu. "Prioriterede" ('prioritized') informants are informants systematically singled out according to various sociolinguistic variables - see https://dgcsc.hum.ku.dk/forskning/undersogelsesomraader.</p>
	<p>It is also possible to choose single informants. Select the option "Enkelt" ('single'), and enter the informant code for the relevant informant.</p>

It is even possible to create a custom subset of informants. The procedure is very similar to creating a custom interview set - see above.

3.3. Search criteria

Under "Søgekriterier" ('search criteria'), the query can be specified. The corpus is based on a collection of files (Praat TextGrids), each associated with several annotation tiers (with phonetic annotations, grammatical annotations, etc.). All tiers are available for search, although not all files contain annotations in all tiers. Each file contains at least the principal annotation tier, i.e. an orthographic transcription of the speech of the informant in question.

A basic search is performed as follows.

	<p>Start by specifying which tier the search should query. (The drop-down menu shows a complete list of all tiers, whether the individual tiers contain annotations for the current subcorpus or not).</p> <p>Next, enter the search term in the orange search field. The field is orange as long as it is empty; empty search criteria are not allowed.</p>
	<p>When the search term has been entered, pick the appropriate category of the search term:</p> <ul style="list-style-type: none">• Hele ordet ('the whole word'): Matches the whole word.• Del af ordet ('part of the word'): Matches that contain the search term anywhere.• Ikke en del af ordet ('not a part of the word'): The inverse of the above.• Starten/Slutningen af ordet ('the beginning/end of the word'): Matches that start/end with the search term.• Regulært udtryk: Regular expression match.

The tiers contain one time interval per content element. For instance, each orthographic tier contains a single word for each interval. It is thus not possible to search for strings of several elements (several words, for instance) from one and the same search field. In order to do this, an additional search field has to be added for each additional search term. This context search is accomplished as follows.

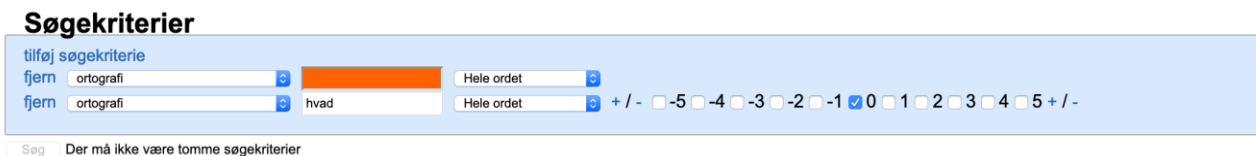


Figure 4. Context search. Click "tilføj søgekriterie" ('add search criterion') to display an additional search field. Don't forget to check the correct position relative to the topmost search term. Otherwise you might end up trying to search for two different words in the same position, which is a logical impossibility.

Click "tilføj søgekriterie" ('add search criterion'). This displays an additional search field. The criterion line is filled in as described above. In addition, the position relative to the primary search term must be specified. Select 0 to search in the same position as the primary search term (logically, this has to be in a different tier). Select -1 to search in the position to the left of the primary search term, +2 to search in the position two words to the right, etc. As long as the position is not 0 (or more generally, not equal to the position of any other search field), the search can be performed in the same tier as the primary search term.

By using regular expressions, it is possible to perform searches for empty intervals or a set of different strings using a single search term. Don't forget to declare the search term a regular expression using the drop-down menu. Examples of regular expressions are listed in Table 1.

Symbol	Search term	Explanation
.*	.*	Zero or more characters. Finds all intervals, including empty intervals.
.+	.+	One or more characters. Finds all non-empty intervals.
	man du	Logical OR. Finds intervals containing <i>man</i> or <i>du</i> .
()	l(æ)gge	Logical OR within string. Finds intervals containing <i>ligge</i> or <i>lægge</i> .
^	^G	Beginning of string. Finds intervals beginning with <i>G</i> .
\$	ik\$	End of string. Finds intervals ending on <i>ik</i> .

Symbol	Search term	Explanation
token.*	sur.*	Zero or more optional characters within string. Finds intervals containing <i>sur</i> followed by zero or more characters, e.g. <i>sure</i> , <i>glasuren</i> , <i>armbåndsuret</i> etc.

Table 1. Regular expressions. Don't forget to choose "regulært udtryk" in the menu to the right of the search field in order to search using regular expressions.

The various regular expression symbols can be used together in a regular expression. Note that the symbols `^` and `$` must be used to specify matches at the beginning and/or end of words. For instance, if the exact words *ligge* and *lægge* are intended, and not *indlægge*, *lægger* etc., the following regular expression can be used: `^(l|i|æ)gge$`. The symbols `^` and `$` can even be used multiple times in the same regular expression. To match the exact forms *ligget* and *lagt*, for instance, the following regular expression can be used: `^ligget$|^lagt$`.

4. Results

When the search is completed, a result overview is shown.

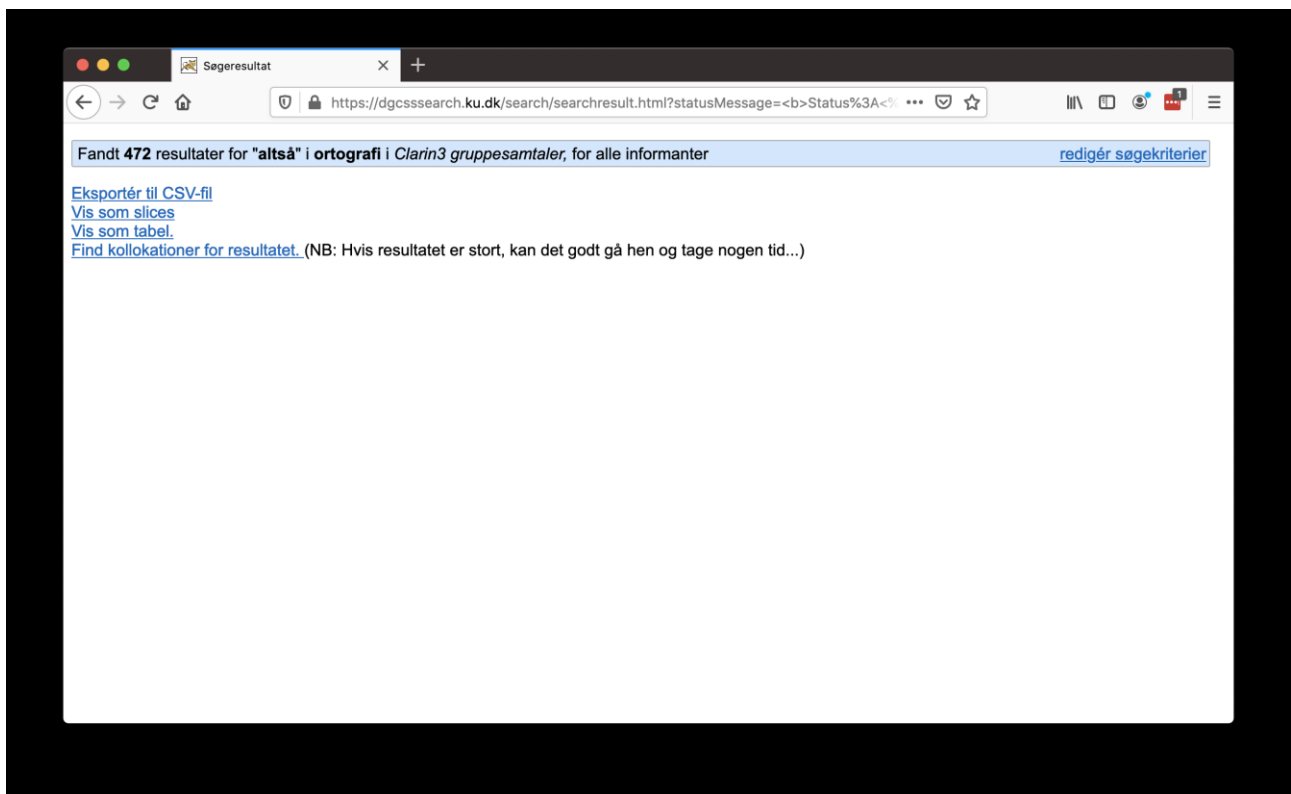


Figure 5. Result overview presented when a search has been performed.

From here, several views of the search results are available: "Eksporter til CSV" ('Export to CSV'), "vis som slices" ('view as slices'), "vis som tabel" ('view as table' (also supports CSV export)), "kollokationer" ('collocation').

4.1. Export to CSV

The result can be exported directly to Excel. Click "Eksporter til CSV-fil" ('export to CSV file'). This will open a prompt to open or save the file.

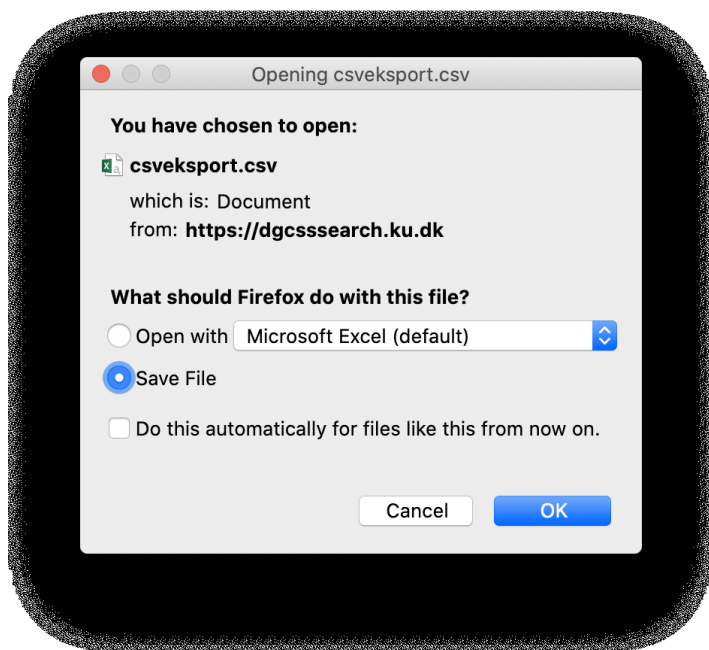


Figure 6. The prompt that appears when "Eksporter til CSV-fil" ('export to CSV file') is clicked.

The CSV file shows each matching interval in the corpus. Each match is shown in a separate row along with file and speaker information as well as the content of the other tiers in the given time interval.

This format is well suited for further quantitative processing and statistic analysis.

4.2. View as slices (concordance view)

The link "Vis som slices" ('view as slices') reveals a concordance view of the result.

Søgeresultaterne 1 - 15 ud af 472 for "altså" i ortografi i Clarin3 gruppesamtaler, for alle informanter [redigér søgekriterier](#)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

clarin3gl-10-AEB+AOE+BFG+LML
AEB (273.472) 272.4844 273.472 Afspil lyd ... 274.466

ortografi

← forrige [følgende](#) ⇒

clarin3gl-10-AEB+AOE+BFG+LML
AEB (338.342) 337.694 338.342 Afspil lyd ... 339.32

ortografi

← forrige [følgende](#) ⇒

clarin3gl-10-AEB+AOE+BFG+LML
AEB (342.505) 341.259 342.505 Afspil lyd ... 343.3825

ortografi

Figure 7. Concordance view. Shown when "Vis som slices" ('view as slices') was chosen.

The view contains the following:

- At the top of the view, an overview in the form of the number of search results is still shown.
- On the left of each concordance line it is possible to choose which tiers are shown along with the transcription.
- Above each concordance line, time information for the match is shown (in seconds).
- The match is marked in blue. The context is white.
- Below each concordance line, a click on "forrige" ('previous') will expand the context to the left, and "følgende" ('next') to the right.
- The recording can be played at the current position by clicking "Afspil lys ..." above each concordance line. For information security reasons, this is however only supported for a limited amount of the data - currently the interview set "Clarin3 gruppesamtaler" ('Clarin3 group conversations') and the files from the project "Danish Voices in the Americas", i.e. the files beginning with "amda" or "argda".

4.3. View (and export) table

An overview table of the matches in context is available from the link "Vis som tabel" ('view as table'). The page contains an interface to choose how much context (how many words) to the left and right of the match are shown, and which tiers are shown.

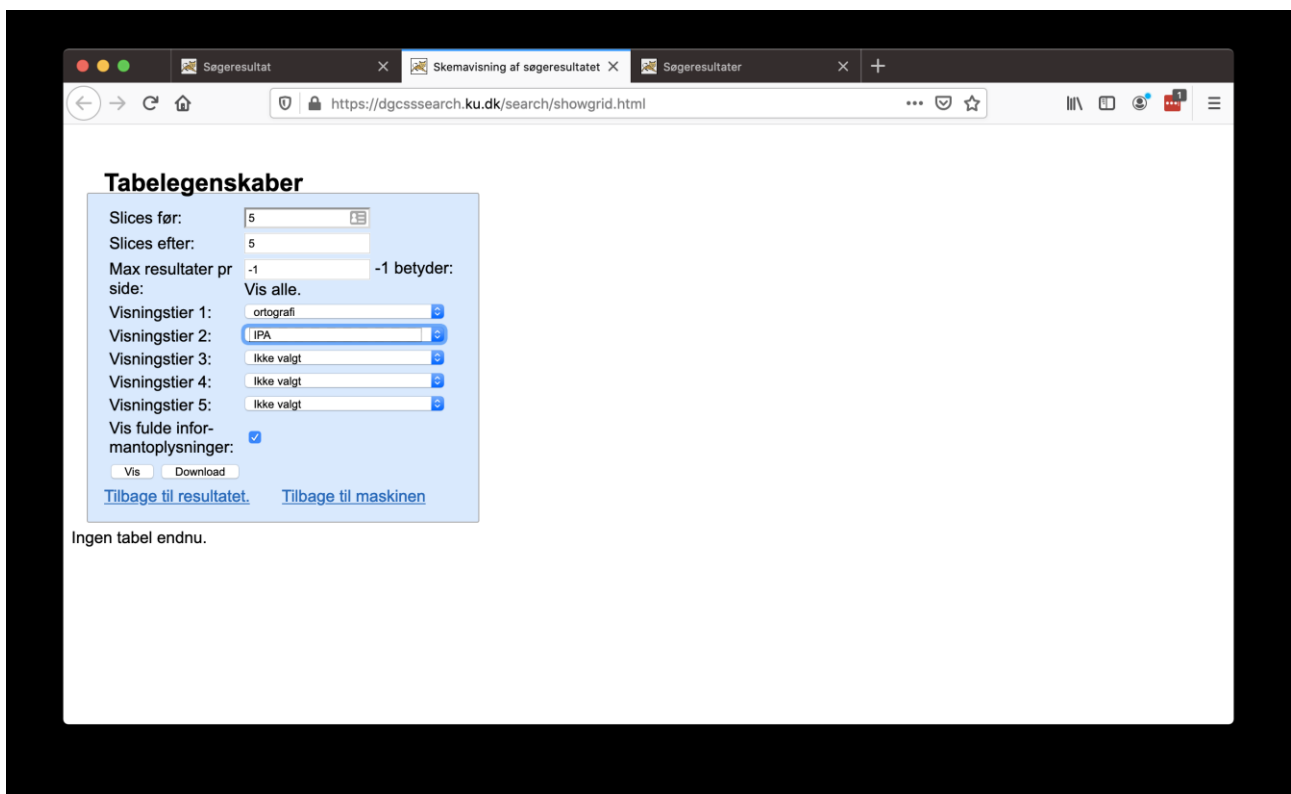


Figure 8. Page shown after clicking "Vis som tabel" ('view as table'). Here, you can specify the parameters of the result table.

When these parameters have been set, the table can either be viewed on the page by clicking "Vis" ('view') or be downloaded as a CSV file by clicking "Download".

In Figure 9, the results are viewed on the page.

Skematisering af søgeresultatet

https://dgcsearch.ku.dk/search/showgrid.html

50%

Tabelegenskaber

Slices for: 5

Slices efter: 5

Vis alle

Vis fulde informationer

Download

Tilbage til maskinen

Fund	Filnavn	Opt.	Opt.år	Tier	Inf	XMin	Kan	S.M.	F.år	Projekt	-5	-4	-3	-2	-1	0	1	2	3	4	5
1	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	273.47	K	NIA	1991	CLARIN	ej			jeg	tror	altså	eh	si	gengæld	at	
2	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	338.94	K	NIA	1991	CLARIN	varer	så	meget	i	mederne	altså	han	har	varer	så	sådan
3	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	342.5	K	NIA	1991	CLARIN	luge	det	så	læst		altså	han	har	jo	haft	
4	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	353.62	K	NIA	1991	CLARIN	man	lige	pludselig	ikke		altså	eh	at	at	han	
5	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	422.82	K	NIA	1991	CLARIN	jeg	var	eh	jeg	synes	altså		det	var	ret	eh
6	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	423.96	K	NIA	1991	CLARIN		det	var	ret	eh	altså				at	overgået
7	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	625.68	K	NIA	1991	CLARIN		det	er	jo	klart	altså			men		men
8	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	637.94	K	NIA	1991	CLARIN	giver	jo	ingen	mening		altså				men	
9	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	656.48	K	NIA	1991	CLARIN	men	det	er	jo		altså	det	er	jo	klart	fordi
10	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	669.27	K	NIA	1991	CLARIN	er	så	mange			altså		ad	dem	der	ah
11	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	732.47	K	NIA	1991	CLARIN	er	jo	fra	Jehovas	Vidne	altså	der	er	jo	ikke	noget
12	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	787.96	K	NIA	1991	CLARIN	så	en	dokumentar	det	var	altså	for	så	noget		True
13	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	811.78	K	NIA	1991	CLARIN	var		hel	hel	ille	altså				elve	
14	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	900.2	K	NIA	1991	CLARIN	ti	al	godt		ikke	altså				nej	nej
15	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	903.26	K	NIA	1991	CLARIN		fuldstændig		nej	men	altså			det	er	jo
16	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	934.81	K	NIA	1991	CLARIN	stue	de		f-	læst	altså	fyfset	alle	v-	eh	mabler
17	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	975.96	K	NIA	1991	CLARIN						altså		de	er		
18	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1029.14	K	NIA	1991	CLARIN		ja		han	er	altså	også	bare	grineren	ikke	
19	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1044.88	K	NIA	1991	CLARIN			og	det		altså	det	var	så	noget	med
20	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1077.91	K	NIA	1991	CLARIN	er	det	grineren		men	altså	pointen	med	det	er	jo
21	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1281.42	K	NIA	1991	CLARIN	er	bare	vildt	med	fordbold	altså	lad	nu	barnet	ikke	
22	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1291.84	K	NIA	1991	CLARIN	nu	bare	Kære	tit	nes	altså	det	skal	jeg	da	ikke
23	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1312.66	K	NIA	1991	CLARIN	skulle	have	varer	sådan		altså	ligesom	sådt	men	der	
24	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1330.84	K	NIA	1991	CLARIN	ikke		det	synes	jeg	altså	er	for	vildt	at	
25	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1382.9	K	NIA	1991	CLARIN	det	er	slet	ikke		altså		på	den	ene	side
26	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1387.77	K	NIA	1991	CLARIN	også	påent	frygteligt	ikke		altså		at	eh	at	de
27	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1391.99	K	NIA	1991	CLARIN	den	der	måde			altså		det	det	er	
28	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1396.66	K	NIA	1991	CLARIN	fred-	ah	fredelig	gjort		altså		der	ku-	det	kunne
29	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1490.98	K	NIA	1991	CLARIN	jeg	har	se			altså	jeg	kommer	bare	lign	med
30	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1638.9	K	NIA	1991	CLARIN		ah	vi	lever		altså		vi	vi	lever	
31	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1640.87	K	NIA	1991	CLARIN	vi	lever	jo	vi	i-	altså	vi	lever	jo	rigtig	rigtig
32	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1649.89	K	NIA	1991	CLARIN	at	være	bange			altså				ja	
33	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1731.92	K	NIA	1991	CLARIN	slet	ikke	komme	nogen	ind	altså	på	lufthavnen	som	ikke	skulle
34	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1736.98	K	NIA	1991	CLARIN	derind	alene	niende	ki-	i	altså		niendeklassen		ung	
35	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1749.9	K	NIA	1991	CLARIN	pas	og	eventuelle	sådan		altså	medicin	eller	så	noget	
36	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1755.94	K	NIA	1991	CLARIN		overførsel				altså	med	ah	op	i	flyet
37	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1775.43	K	NIA	1991	CLARIN	sådan	at	gemme	det	ah	altså	at	dysse	det	ned	ikke
38	clarin3gl-10-AEB+AOE+BFQ+LML	gi	2010	ortografi	AEB	1835.99	K	NIA	1991	CLARIN	det	var	for		vildt	altså	jeg	synes	fandeme	det	var

Figure 9. Table view. The table appears on the page when "Vis" ('view') is clicked. The other option is "Download", which will prompt a CSV download of the table.

In each row, speaker information is available. Position 0 contains the query match. Note that the table can contain multiple rows per search result - one extra row per additional tier selected.

The table can be downloaded as a CSV file by clicking "Download".

The table shown on this page only contains the most important corpus metadata. To obtain the full set of metadata, use "Eksportér til CSV-fil" ('export to CSV file') instead.

4.4. Collocations

If the query only consists of a single string, a collocation analysis can be performed. Collocations are calculated statistically as word pairs cooccurring more frequently than would be expected from each word's individual frequency in the corpus.

To perform a collocation analysis, click "Find kollokationer for resultatet" ('find collocations for the result').

At the top of the results page, the number of tokens in the selected subcorpus is listed along with the number of unique tokens. Below that, the right and left context of the query string are shown.

The tables show the context word, the number of times the context word occurs with the query word, and the Mutual Information score - a measure of the unexpectedness of the context word and the query word occurring together. The collocations are ordered by the MI score.