

HS Corpus Linguistics / Korpuslinguistik

1. Introduction & Techniques of corpus linguistics

Prof. Dr. Stephanie Evert

Chair of Computational Corpus Linguistics
www.linguistik.uni-erlangen.de



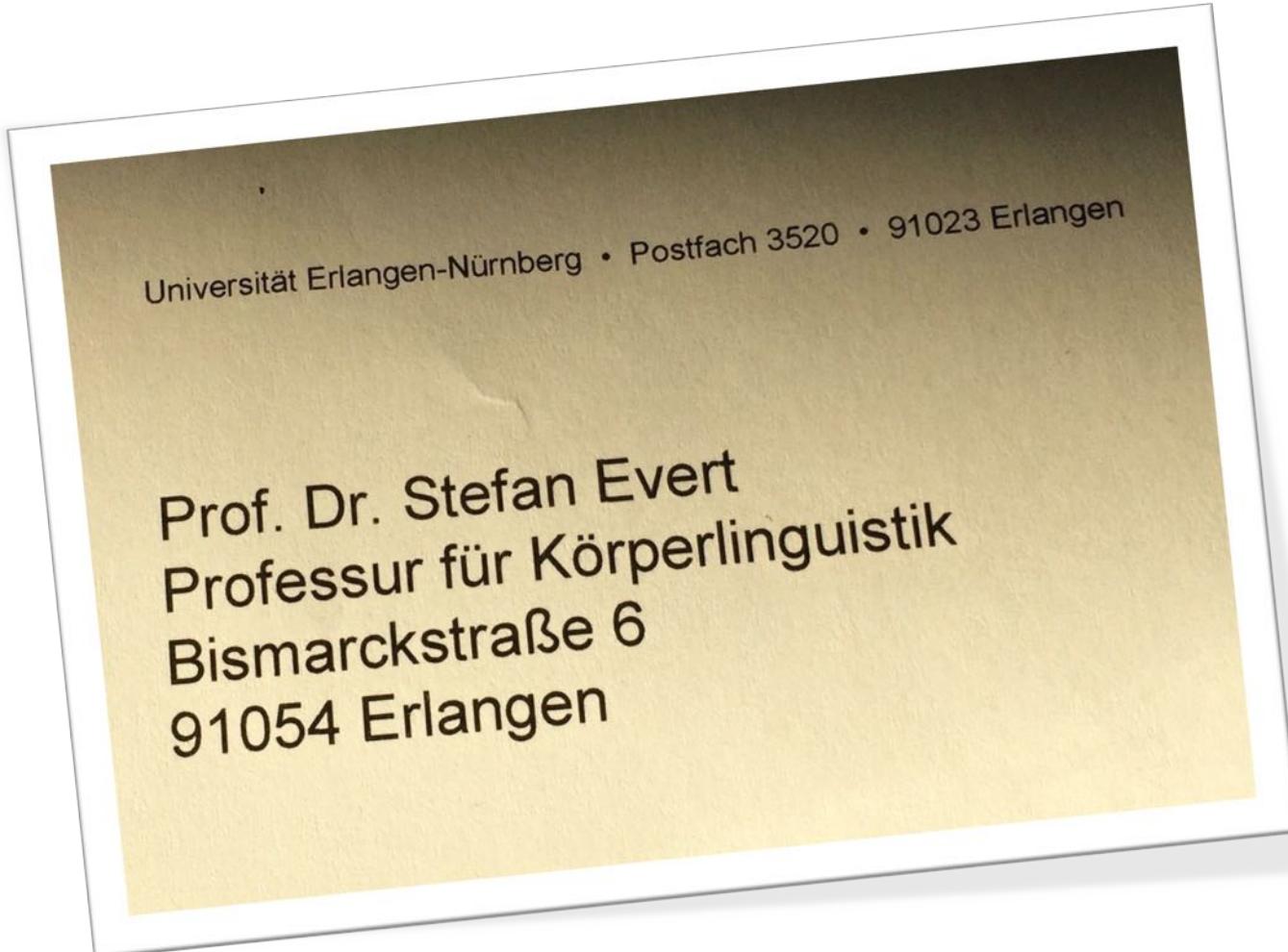
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE



contributions allowed
(except MA English Studies)

What is a corpus?



What is a corpus?

corpus | 'kɔ:pəs |

noun (pl. **corpora** | 'kɔ:pərə | or **corpuses**)

1 a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: *the Darwinian corpus*.

- a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research.

2 Anatomy the main body or mass of a structure.

- the central part of the stomach, between the fundus and the antrum.

ORIGIN late Middle English (denoting a human or animal body): from Latin, literally '**body**'. **SENSE 1** dates from the early 18th cent.

- Corpus as **electronic text collection** (archive / database)
 - wide sense: common in computational linguistics & digital humanities
- Corpus as **representative sample** of language
 - narrow sense: dominant in corpus linguistics (in narrow sense)

What is corpus linguistics?

- Corpus linguistics as a **methodology**
 - (one form of) empirical evidence for linguistic research
 - corpus design & practical aspects
 - development of software tools
 - quantitative analysis techniques, often in terms of frequency comparison
- Corpus linguistics as **discipline**
 - concordance, collocations, keywords
 - language in use, not language as system
 - language as a medium → applied CL
 - typical research topics
- Corpus-based research
 - verification of hypotheses and theoretical predictions
 - compilation of specialised corpora
 - exploration of specific constructions
 - statistical significance of differences, data analysis as discovery procedure
- Corpus-driven research
 - “trust the text” (Sinclair 2004)
 - purely empirical, inductive, contextualised
 - socio-political discourses & rhetoric
 - sociolinguistic & functional aspects
 - language variation
 - computational lexicography
 - language pedagogy (→ CALL)

...

Corpus linguistics vs. Theoretical linguistics

Corpus linguistics

- long tradition in esp. British linguistics since 1950s (Quirk, Firth, Sinclair, Leech, ...)
- based on corpus data,
i.e. authentic language use
- empiricist / inductive
- patterns & preferences
- frequency
- “ordinary” language in use
- empirical validation of theory,
induction of new hypotheses
- goals: language use, language as medium

Generative linguistics

- mainstream of theoretical linguistics since early 1960s (Chomsky 1957)
- based on introspection and grammaticality judgements
- rationalist / deductive
- rules
- yes / no (grammaticality)
- language system, border cases
- theory developed on basis of individual (invented) examples
- goal: cognitive account of language faculty

Requirements

- Module **Corpus Linguistics** (MA English Studies, MA Linguistik)
 - HS *Corpus Linguistics* (7 credits) | **Fri 10–12**, CIP pool (0.320, Bismarckstr. 12)
 - Ü *The corpus linguist's essential toolbox* (3 credits) | **Fri 12–14**, CIP pool | lecturer: Dr. Besim Kabashi
- Module **Korpuslinguistik** (BA Linguistische Informatik)
 - HS *Korpuslinguistik* (5 credits) | **Fri 10–12**, CIP pool
 - Ü *Statistik für Korpuslinguisten* (5 credits) | **Tue 14–16**, CIP pool | lecturer: Philipp Heinrich, M.Sc.
- Presentation
 - discussion of research paper(s), ideally combined with small corpus study / replication of findings
 - with presentation slides, duration: 15–20 minutes + 10 min discussion
- Term paper (only MA)
 - ca. 15 pages, usually based on the presentation (and feedback from discussion)
 - grade: term paper (HS, 70%) + exercises (Ü, 30%)

Semester overview

- 29.04. Introduction + Techniques of CL
 - get-to-know, a first look at a corpus
- 06.05. Fundamentals
 - list of presentation topics
- 13.05. Statistical analysis: frequency data
- 20.05. Corpus compilation & annotation
 - discussion on corpus design
 - assignment of presentation topics
- 27.05. Hands-on: **CADS & MMDA** 1
- 03.06. Hands-on: **CADS & MMDA** 2
- 10.06. Hands-on: **CADS & MMDA** 3
- 17.06. Corpus queries & indexing
- 24.06. Statistical analysis: non-randomness
- 01.07. Student presentations 1–3
- 08.07. Student presentations 4–6
- 15.07. Student presentations 7–9
- 22.07. Student presentations 10–12
- 29.07. Student presentations 13–15
 - summary & final discussion
 - instructions and Q&A for term papers

Special feature

- Hands-on block of three sessions with practical corpus study
- Ties into cutting-edge research on **digital hermeneutics**
 - theoretical **framework** and hermeneutic methodology for corpus-assisted discourse analysis (CADS)
 - innovative software support for qualitative-quantitative studies: **MMDA toolkit**
- Preliminary schedule
 1. Introduction to CADS
 2. Group work: corpus study on hermeneutic methodology
 3. Group work: corpus study using MMDA toolkit vs. traditional corpus software
- Co-teachers for this block (PhD researchers @ CCL)
 - **Nathan Dykes**, M.A. (framework & methodology for CADS)
 - **Philipp Heinrich**, M.Sc. (MMDA development)

Introductions

- Who are you? And how do I pronounce your name?
- Which degree programme are you enrolled in?
- What background do you have in linguistics, corpus research, computer science, NLP?
- What do you know about corpus linguistics?
- Have you ever worked with corpora?
- How could your own research benefit from corpora?
- What do you expect from this seminar?
- Any other questions?

All materials in the StudOn course
HS Corpus Linguistics /
Korpuslinguistik – SS 2022



Techniques of corpus linguistics: A hands-on example

- Goal: Understand this [→] guy's rhetoric, terminology, phraseology and topics just in case he comes back ...
- How would you approach this task?
- We need a large collection of texts written by Trump, e.g. the **Trump Twitter Archive** [www.thetrumparchive.com]
- Step 1: Compile a linguistically annotated corpus (**TTA**)
- Step 2: Upload to indexing & search engine
- Step 3: Analyse with corpus-linguistic techniques



Techniques of corpus linguistics: A hands-on example

- Hands-on experience with **CQPweb** and the **Trump Twitter Archive**
 - go to <https://corpora.linguistik.uni-erlangen.de/cqpweb/> and login as **student1 ... student15**
 - select corpus *Trump Twitter Archive* from “Political Tweets” section

The screenshot shows the login interface for CQPweb. At the top, there is a purple header bar with the text "Welcome to CQPweb at FAU Erlangen-Nürnberg!" and "This server is maintained by the Chair of Computational Corpus Linguistics". On the left is the CL logo, and on the right is the CWB logo. Below the header is a form for entering a username and password. There is also a checkbox for staying logged in. At the bottom of the form are two buttons: "Click here to log in" and "Clear form". Below the form, there are links for "Create account" and "Full account-control options". A section titled "Corpora available on this server" lists "Populism", "Tweets about the German Federal Election 2013 [with duplicates]", and "German Parliament (2011-2014, v2)".

Welcome to CQPweb at FAU Erlangen-Nürnberg!

This server is maintained by the
Chair of Computational Corpus Linguistics

Enter your username:

Enter your password:

Stay logged in
on this computer:

[Click here to log in](#) [Clear form](#)

[Create account](#) | [Full account-control options](#)

Corpora available on this server

Populism

[Tweets about the German Federal Election 2013 \[with duplicates\]](#)

[Tweets about the German Federal Election 2013](#)

[German Parliament \(2011-2014, v2\)](#)

Simple search

Menu

Corpus queries

- Standard query
- Restricted query
- Word lookup
- Frequency lists
- Keywords
- Analyse corpus

Saved query data

- Query history
- Saved queries
- Categorised queries
- Upload a query
- Create/edit subcorpora

Corpus info

- View corpus metadata
- No corpus documentation available*

About CQPweb

- CQPweb main menu
- Your user page
- Help system
- Video tutorials
- Who did it?
- Latest news
- Report bugs

Trump Twitter Archive: powered by CQPweb

Standard Query

make america great again



Query mode: Simple query (ignore case) [Simple query language syntax](#)

Number of hits per page: 50

Match strategy: Standard

Restriction: None (search whole corpus)

[Start Query](#) [Reset Query](#)

System messages 

2018-06-06 **New Arrangement of Corpora**
The corpora have been newly arranged. If you cannot find a corpus that belongs to you or lost your access privileges, please contact us.
All corpora listed under "Uncategorised" will be deleted in the long run. Please let us know if there are any issues with this.

Concordance (kwic = key word in context)

Your query "make america great again" returned 585 matches in 585 different texts (in 1,332,122 words [56,570 texts]; frequency: 439.15 instances per million words), ordered randomly [0.04 seconds]

I < << >> > Show Page: 1 Line View Show in corpus order Choose action... G

No.	Text	Solution 1 to 50	Page 1 / 12
1	t1037417543017877500	Join me tomorrow night at 7:00 pm MDT in Billings , Montana for a	MAKE AMERICA GREAT AGAIN ! RALLY ! Get your tickets here : https://t.co/fyHduA2Peo
2	t0761622637235712000	pport to get to the White House and defeat #CrookedHillary . Let's	Make America Great Again ! https://t.co/u2WCoondgqm
3	t0763860288533856300	ounting on your help to defeat Hillary Clinton and her cronies . Let's	Make America Great Again ! https://t.co/2KJQDkKJQDk
4	t0585269105068007400	ldtrump @cleopatrausa Donald you are the next Ronald Reagan .	Make America great again . !!! Thank you .
5	t0620994157004357600	t . We are energized & , ready to take our country back . Let's	Make America Great Again ! https://t.co/u25yl5T7E8
6	t0697825190282182700	He (or she) who hesitates is lost : MAKE AMERICA GREAT AGAIN !	
7	t0807326	Text t0697825190282182700 (length = 15 words)	, together , we will MAKE AMERICA GREAT AGAIN ! Watch https://t.co/EyLOo26FqW
8	t070020013	-----	MAKE AMERICA GREAT AGAIN !
9	t047330563	Day: 11 Flagged?: no Hour: 15	realDonaldTrump !!! Make America great again !
10	t122726958	Month: 02_February Reply?: no Retweet?: no	realDonaldTrump : MAKE AMERICA GREAT AGAIN and then , KEEP AMERICA GREAT !
11	t082206653	Day of the week: 4_Thursday Year: 2016	we will , together , MAKE AMERICA GREAT AGAIN !
12	t056244915		Let's together Make America Great Again ! Vote Trump at https://t.co/YoNf60s0lm
13	t061245976890593300	J.S. is becoming a dumping ground for the world . Pols don't get it .	Make America Great Again !
14	t068525573943279600	uld only get a small fraction of this 25k crowd in . The movement to	Make America Great Again is unbelievable ! https://t.co/NHPdGm57YJ
15	t0771145576381427700	/ she didn't go to Mexico . She doesn't have the drive or stamina to	MAKE AMERICA GREAT AGAIN !
16	t07476700		MAKE AMERICA GREAT AGAIN ! MAKE AMERICA SAFE AGAIN !
17	t1320770458527158300	Thank you Allentown , Pennsylvania ! Together , we are going to	MAKE AMERICA GREAT AGAIN ! https://t.co/gsFSghkmdM https://t.co/aA0l8pBEGb https://t.co/ngL
18	t0602007070700500000	'heguhMantap : @realDonaldTrump you are next President USA !!!	Make America Great Again !
19	t0603631285777596400	in your best interests . Time for the US gov't to do the same . Let's	Make America Great Again !
20	t1307675370322145300		MAKE AMERICA GREAT AGAIN !
21	t1347555316863553500	great American Patriots who voted for me , AMERICA FIRST , and	MAKE AMERICA GREAT AGAIN , will have a GIANT VOICE long into the future . They will not be di
22	t0700007070700500000		MAKE AMERICA GREAT AGAIN !

click/hover
for metadata

frequency
information

random ordering

click for context

https://corpora.linguistik.uni-erlangen.de/cqpweb/tta_full/textmeta.php?text=t0697825190282182700

Extended context display with formatting

Displaying extended context for query match in text **t0697825190282182700**

Select action... Go! Show tags

2016-02-11 00:34:25 [1279, ❤ 4331] I was referring to the fact that Jeb Bush wants to keep common core .

2016-02-11 00:37:51 [1184, ❤ 4258] "" @HARyder : Which is it @realDonaldTrump ? Are you planning on getting rid of Common Core or keeping it ? Get rid of it fast .

2016-02-11 01:12:19 [2927, ❤ 8568] THANK YOU- Clemson , South Carolina ! #MakeAmericaGreatAgain #SCPrimary <https://t.co/FgACmaFxxc>

2016-02-11 01:51:40 [5097, ❤ 13386] I have been consistent in my opposition to Common Core . Get rid of Common Core --- keep education local !

2016-02-11 13:13:09 [1640, ❤ 5432] There are no buyers for the worthless @NYDailyNews but little Mort Zuckerman is frantically looking . It is bleeding red ink - a total loser !

2016-02-11 14:11:56 [3469, ❤ 9767] Jeb Bush spent more than \$ 40,000,000 in New Hampshire to come in 4 or 5 , I spent \$ 3,000,000 to come in 1st . Big difference in capability !

2016-02-11 14:40:04 [1436, ❤ 4032] . @MarkHalperin showed a focus group on @Morning_Joe me using a very bad word . I never said the word , left an open blank . Please apologize !

2016-02-11 15:14:10 [2589, ❤ 6342] Remember , it was the Republican Party , with the help of Conservatives , that made so many promises to their base , BUT DIDN'T KEEP THEM ! Hi DT

2016-02-11 15:51:04 [2596, ❤ 7190] He (or she) who hesitates is lost : **MAKE AMERICA GREAT AGAIN !** ← this is the match

2016-02-11 18:45:59 [2277, ❤ 7161] Jeb failed as Jeb ! He gave up and enlisted Mommy and his brother (who got us into the quicksand of Iraq) . Spent \$120 million . Weak-no chance !

2016-02-11 19:39:33 [3659, ❤ 9115] We are getting reports from many voters that the Cruz people are back to doing very sleazy and dishonest "" pushpolls "" on me . We are watching !

2016-02-11 22:03:54 [1604, ❤ 5195] Heading to Baton Rouge , Louisiana for a speech . Expecting a very large crowd ! See you soon . #Trump2016 #MakeAmericaGreatAgain

2016-02-11 22:19:33 [1646, ❤ 4445] "" @truthinvest : @CNN @tedcruz @realDonaldTrump Ted Cruz is the definition of sleaze . Just ask @RealBenCarson ""

2016-02-11 23:44:06 [2478, ❤ 5355] Cruz caught cold in lie after denial of push polls like lies w/ @RealBenCarson . How can he preach Christian values ? <https://t.co/p3yGL02ABA>

2016-02-12 00:06:30 [2179, ❤ 7547] Just landed in Baton Rouge , Louisiana . Reports are out that lines are three quarters of a mile to get in . Wow ! #MakeAmericaGreatAgain

2016-02-12 02:28:28 [3822, ❤ 9677] THANK YOU- Baton Rouge , Louisiana ! WE will #MakeAmericaGreatAgain ! #Trump2016 <https://t.co/XV7Ele7A2l>

2016-02-12 03:10:52 [2110, ❤ 5156] Weak JEB getting thrown out by management during speech . Do you think he will be this tough on Putin & others ? <https://t.co/Tqej1euLVL>

2016-02-12 03:20:33 [3357, ❤ 9104] Lying Cruz put out a statement , " Trump &

Create subcorpus for targeted search

Menu

Corpus queries

- Standard query
- Restricted query
- Word lookup
- Frequency lists
- Keywords
- Analyse corpus

Saved query data

- Query history
- Saved queries
- Categorised queries
- Upload a query
- Create/edit subcorpora** 

Corpus info

- View corpus metadata
- No corpus documentation available

About CQPweb

- CQPweb main menu
- Your user page
- Help system
- Video tutorials
- Who did it?
- Latest news
- Report bugs

Trump Twitter Archive: powered by CQPweb

Create and edit subcorpora

Define new subcorpus via: Go!

Design a new subcorpus

Please enter a name for your new subcorpus.

Names for subcorpora can only contain letters, numbers and the underscore character (_)! 

Choose the categories you want to include from the lists below.

Then either create the subcorpus directly from those categories, or view a list of texts to choose from.

Reply?	Retweet?
<input checked="" type="checkbox"/> no <input type="checkbox"/> yes	<input checked="" type="checkbox"/> no <input type="checkbox"/> yes

Day

01
02
03
04
05
06
07
08
09

only original tweets

Frequency list for hashtags

Menu

Corpus queries

- Standard query
- Restricted query
- Word lookup
- Frequency lists** 
- Keywords

Analyse

S

Query history

Saved queries

Categories

Upload a file

Create/export

View corpus

No corpora

CQPweb

Your usage

Help system

Video tutorial

Who did what

Latest news

Report bugs

Trump Twitter Archive: powered by CQPweb

Frequency lists

You can view the frequency lists of the whole corpus and frequency lists for subcorpora you have created. [Click here to create/view subcorpus frequency lists.](#)

View frequency list for ... Subcorpus: Originals 

View a list based on ... Lemma 

Frequency list option settings

starting with 

ending with

containing

matching exactly



with frequency between _____ and _____

50 

most frequent at top 

Frequency list Clear the form

must compile frequency list for subcorpus first

hashtags

No.	Lemma	Frequency
1	#trump2016	825
2	#makeamericagreatagain	544
3	#maga	445
4	#celebapprentice	297
5	#celebrityapprentice	152
6	#1	133
7	#trump	112
8	#timetogettough	98
9	#americafirst	93
10	#trumpvlog	81
11	#trumpforpresident	77
12	#draintheswamp	76
13	#votetrump	70
14	#trump2016https	66
15	#kag2020	58
16	#usa	51
17	#2a	48

Keywords vs. reference corpus

Menu

Corpus queries

- Standard query
- Restricted query
- Word lookup
- Frequency lists
- Keywords** ←
- Analyse corpus

Saved query data

- Query history
- Saved queries
- Categorised queries
- Upload a query
- Create/edit subcorpora

Corpus info

Trump Twitter Archive: powered by CQPweb

suitable reference corpus

Keywords and key tags

Keyword lists are compiled by comparing frequency lists you have created for different subcorpora. [Click here to create/view frequency lists.](#)

Select frequency list 1: Subcorpus: Originals Select frequency list 2: Public subcorpus: English_Originals_Tweets_201708

Compare: Lemma ←

Options for keyword analysis:

Show:	<input checked="" type="radio"/> Positive keywords ←	Display as:	<input checked="" type="radio"/> Textual wordcloud (Wmatrix-style) ←
Comparison statistic:	Log Ratio (conservative estimate) ←	Significance cut-off point: 0.04% ←	<input checked="" type="checkbox"/> Use 15k correction?
Min. frequency (list 1):	3 ←	Min. frequency (list 2):	0 ←

recommended settings

!'"--....@realdonaldtrump a again all always amazing american amp and at be big book border china country crime day deal democrat do election ever fake family first forward friend get good great have he her hillary him history house i interview job just keep law look love make man many me media military much my never new nice night number obama one our please poll president record run security she show soon speak state strong thank there they time today total totally true trump u.s. united very watch white win world you your --"

click for concordance

Sorted concordance: great

action menu

Your query "[lemma='great'%ocd]", in subcorpus "Originals", returned 7,332 matches in 6,771 different texts (in 1,084,300 words [43,878 texts]; frequency: 6,761.97 instances per million words), sorted on position +1 (7,332 hits)

[0.126 seconds - retrieved from cache]

No.	Text	Solution 2101 to 2150	Page 43 / 147
2101	t1122270956192272400	Sincerest THANK YOU to our	great Border Patrol Agent who sta
2102	t1104017664131907600	hending record numbers of illegal immigrants - but we need the Wall to help our	great Border Patrol Agents !
2103	t1075954128222871600	ans voted and won , 217-185 . Nancy does not have to apologize . All I want is	GREAT BORDER SECURITY !
2104	t0876899250071904300 getting	great border security and healthcare . #VoteRalphNorman tomorrow !
2105	t1024248479386923000	One of the reasons we need	Great Border Security is that Mexico's murder rate in 2017 increased by 27% to 31,1
2106	t0977855968364171300	ild our Military , many jobs are created and our Military is again rich . Building a	great Border Wall , with drugs (poison) and enemy combatants pouring into our Co
2107	t0968704442110545900	Administration and rejected the attempt to stop the government from building a	great Border Wall on the Southern Border . Now this important project can go forwar
2108	t0830405706255913000	I am reading that the	great border WALL will cost more than the government originally thought , but I have
2109	t0741286391200505900	The	great boxing promoter , Don King , just endorsed me . Nice !
2110	t1198679802913415200	Agnes , your	great boy Ronald is looking down , very proud of you ! https://t.co/BHPu6IldAN
2111	t0564215204154960000	IdTrump @Newsmax_Media @dpatten32 Time for Trump to take charge of the	greatest brand in the world , the USA ! ""
2112	t0263720538923470850	's acquisition of Lucas Film is a smart deal for both sides . Disney just bought a	great brand which will keep producing revenue .
2113	t0862283792559616000	years , as a pol in Connecticut , Blumenthal would talk of his	great bravery and conquests in Vietnam - except he was never there . When
2114	t0334834807483818000	I totally respect that Angelina Jolie has shown such	great bravery in the face of danger - she has really come a long and positive way !
2115	t1060515116700045300	Great bravery shown by police . California Highway Patrol was on scene within 3 mir
2116	t0979082457340407800		Great briefing this afternoon on the start of our Southern Border WALL ! https://t.co/c
2117	t0746458701565988900	Many people are equating BREXIT , and what is going on in	Great Britain , with what is happening in the U.S. People want their country back !
2118	t0800887087780294700	Many people would like to see @Nigel_Farage represent	Great Britain as their Ambassador to the United States . He would do a great job !
2119	t0589633925837881300	You rocked both events . The @washingtonpost knows it too !! Thank you for a	great brunch and a great talk !!
2120	t0585243017705062400	"" @StaceyHunter10 : DonaldTrump	great brunch yesterday at Mar a Lago thank you for stopping to take a pic with our k

Frequency breakdown: fake

Query "{fake}", in subcorpus "Originals", returned 1,194 matches in 1,117 different texts (in 1,084,300 words [43,878 texts]; frequency: 1,101.17 instances per million words)

Showing frequency breakdown of words in this query, at position 1 to the Right; there are 104 different types and 1,194 tokens at this concordance position.

[0.006 seconds - retrieved from cache]

<

<<

>>

>

Breakdown position: 1 Right



Frequency breakdown of words only

Go!

No.	Query result	No. of occurrences	Percent
1	NEWS	902	75.54%
2	and	24	2.01%
3	Dossier	21	1.76%
4	media	19	1.59%
5	&	17	1.42%
6	,	15	1.26%
7	!	9	0.75%
8)	9	0.75%
9	Polls	7	0.59%
10	reporting	7	0.59%
11	Story	7	0.59%
12	Whistleblower	7	0.59%
13	Suppression	6	0.5%
14	..	5	0.42%
15	ballots	5	0.42%
16	(4	0.34%
17	""	3	0.25%
18	@NBCNews	3	0.25%
19	ad	3	0.25%
20	as	3	0.25%
21	Book	3	0.25%

Time for a closer look: *fake news*

Menu

Corpus queries

- Standard query
- Restricted query
- Word lookup
- Frequency lists
- Keywords
- Analyse corpus

Saved query data

- Query history
- Saved queries
- Categorised queries
- Upload a query
- Create/edit subcorpora

Corpus info

- View corpus metadata
- No corpus documentation available

About CQPweb

- CQPweb main menu
- Your user page
- Help system
- Video tutorials
- Who did it?
- Latest news
- Report bugs

Trump Twitter Archive: powered by CQPweb

Standard Query

fake news

Query mode:

Simple query (ignore case) 

[Simple query language syntax](#)

Number of hits per page:

50 

Match strategy:

Standard 

Restriction:

Subcorpus: Originals (1,084,300 words in 43,878 texts) 

Start Query
Resubmit query



System messages 

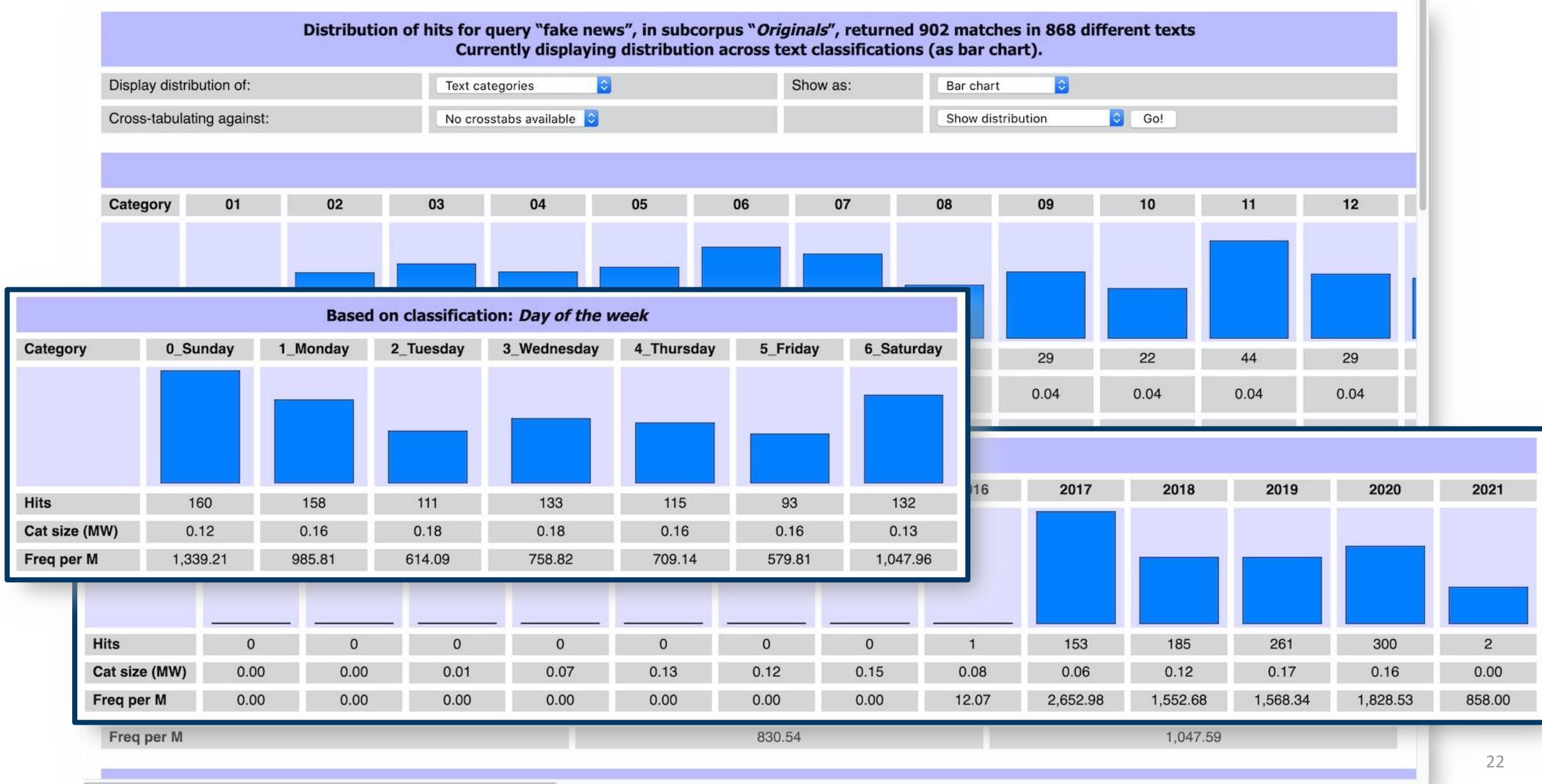
2018-06-06

New Arrangement of Corpora

The corpora have been newly arranged. If you cannot find a corpus that belongs to you or lost your access privileges, please contact us.

All corpora listed under "Uncategorised" will be deleted in the long run. Please let us know if there are any issues with this.

Distribution across metadata categories



Collocation analysis

Collocation controls

Collocation based on:	Lemma	Statistic:	Log-likelihood
Collocation window from:	5 to the Left	Collocation window to:	5 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate: <input type="text"/>	and/or tag: <input type="text"/> V <input type="button" value="V"/>	Submit changed parameters <input type="button"/> Go!

Extra information:

Log-likelihood (LL) scores collocations by significance: the higher the score, the more evidence you have that the association is not due to chance. More frequent words tend to get higher log-likelihood scores, because there is more evidence for such words.

There are 1,497 different lemmas in the collocation database for this query (Query "fake news", in subcorpus "Originals", returned 902 matches in 868 different texts)

[0.175 seconds - retrieved from cache]

No.	Lemma	Total no. in this subcorpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	report	615	5.116	39	39	92.712
2	cover	156	1.298	18	18	63.171
3	refuse	176	1.464	17	17	53.748
4	be	35,694	296.929	423	61	41.166
5	doesn't	249	2.071	18	18	47.059
6	mention	129	1.073	13	12	42.165
7	say	1,987	16.520	48	47	40.017
8	talk	677				
9	refer	52				
10	hate	242				
11	try	548				
12	go	2,877				
13	fail	314				
14	put	568				

click for examples

Choose settings for proximity-based collocations:

Include annotation:	Lemma <input checked="" type="radio"/> Include <input type="radio"/> Exclude
	POS-disambiguated lemma <input checked="" type="radio"/> Include <input type="radio"/> Exclude
	Part of speech <input checked="" type="radio"/> Include <input type="radio"/> Exclude
Maximum window span:	+ / - 5 <input type="button"/>
<input type="button" value="Create collocation database"/>	