GRK 2839 – Winter School 2023

Corpus linguistics: Representation, indexing & corpus queries

Prof. Dr. Stephanie Evert
Chair of Computational Corpus Linguistics
www.linguistik.uni-erlangen.de

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Corpus annotation: lemmatization

<s> It$_{PP}^{it}$ seemed$_{VBD}^{seem}$ a$_{DT}^{a}$ day$_{NN}^{day}$ much$_{RB}^{much}$ as$_{IN}^{as}$ any$_{DT}^{any}$ other$_{JJ}^{other}$ until$_{IN}^{until}$ I$_{PP}^{I}$ happened$_{VBD}^{happen}$ to$_{TO}^{to}$ look$_{VB}^{look}$ out$_{RP}^{out}$ of$_{IN}^{of}$ the$_{DT}^{the}$ back$_{JJ}^{back}$ window$_{NN}^{window}$.$_{SENT}^{.}$ </s>

<s> There$_{EX}^{there}$ was$_{VBD}^{be}$ a$_{DT}^{a}$ little$_{JJ}^{little}$ garden$_{NN}^{garden}$ behind$_{IN}^{behind}$ the$_{DT}^{the}$ house$_{NN}^{house}$;$_{:}^{;}$ a$_{DT}^{a}$ well-mown$_{VBN}^{???}$ lawn$_{NN}^{lawn}$ surrounded$_{VBN}^{surround}$ by$_{IN}^{by}$ a$_{DT}^{a}$ neatly$_{RB}^{neatly}$ cut$_{VBN}^{cut}$ hedge$_{NN}^{hedge}$,$_{,}^{,}$ a$_{DT}^{a}$ few$_{JJ}^{few}$ bushes$_{NNS}^{bush}$ and$_{CC}^{and}$ colourful$_{JJ}^{colorful}$ flowers$_{NNS}^{flower}$.$_{SENT}^{.}$ </s>

need better representation format

XML markup of annotation

Standard for data interchange & archiving

root element

element annotated with attributes

start tag of XML element

```
<corpus>
  <story num="6" title="The Garden">
    <p>
      <s>
        <token pos="PP"  lemma="it">It</token>
        <token pos="VBD" lemma="seem">seemed</token>
        <token pos="DT"  lemma="a">a</token>
        <token pos="NN"  lemma="day">day</token>
        <token pos="RB"  lemma="much">much</token>
        <token pos="IN"  lemma="as">as</token>
        <token pos="DT"  lemma="any">any</token>
        <token pos="JJ"  lemma="other">other</token>
        <token pos="IN"  lemma="until">until</token>
        <token pos="PP"  lemma="I">I</token>
      </s>
    </p>
  </story>
</corpus>
```

corresponding end tag

XML markup of annotation
Standard for data interchange & archiving

XML declaration

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <story num="6" title="The Garden">
    <p>
      <s>
        <token pos="PP"  lemma="it">It</token>
        <token pos="VBD" lemma="seem">seemed</token>
        <token pos="DT"  lemma="a">a</token>
        <token pos="NN"  lemma="day">day</token>
        <token pos="RB"  lemma="much">much</token>
        <token pos="IN"  lemma="as">as</token>
        <token pos="DT"  lemma="any">any</token>
        <token pos="JJ"  lemma="other">other</token>
        <token pos="IN"  lemma="until">until</token>
        <token pos="PP"  lemma="I">I</token>
        ...
      </s>
    </p>
  </story>
</corpus>
```

XML markup of annotation

Standard for data interchange & archiving

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
 <metadata>
    <author>
      <name>Stefan Evert</name>
      <sex>male</sex>
    </author>
    <publication>
      <title>Very Short Stories</title>
      <type>collection</type>
      <genre>fiction</genre>
    </publication>
  </metadata>
  <story num="6" title="The Garden">
    <p>
      <s>
        <token pos="PP"  lemma="it">It</token>
        <token pos="VBD" lemma="seem">seemed</token>
        <token pos="DT"  lemma="a">a</token>
        <token pos="NN"  lemma="day">day</token>
        ...
```

metadata header

XML standards

- XML (Extensible Markup Language) is a widely-used standard for structured annotation

- A well-formed XML document guarantees consistent markup and hierarchical structure, but neither names nor semantics of elements & attributes

- DTD (document type declaration) or XML Schema
  specify valid element & attribute names and their nesting
  - still doesn't explain semantics without documentation!

- Exchange formats for text corpora:
  TEI (Text Encoding Initiative), XCES (Corpus Encoding Standard),
  ISO 24612: LAF (Linguistic Annotation Framework)
  - but simpler and more efficient representation is preferable for corpus search & quantitative analysis

TEI standard (BNC)

H9C.xml*

```
<bncDoc xml:id="H9C">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title> The prince of darkness. Sample containing about 44223 words from a book
                    (domain: imaginative) </title>
                <respStmt>
                    <resp> Data capture and transcription </resp>
                    <name> Oxford University Press </name>
                </respStmt>
            </titleStmt>
            <editionStmt>
                <edition>BNC XML Edition, December 2006</edition>
            </editionStmt>
            <extent> 44223 tokens; 44797 w-units; 3933 s-units </extent>
            <publicationStmt>
                <distributor>Distributed under licence by Oxford University Computing Services on
                    behalf of the BNC Consortium.</distributor>
                <availability> This material is protected by international copyright laws and may
                    not be copied or redistributed in any way. Consult the BNC Web Site at
                    http://www.natcorp.ox.ac.uk for full licencing and distribution
                    conditions.</availability>
                <idno type="bnc">H9C</idno>
                <idno type="old"> PDarkn </idno>
            </publicationStmt>
            <sourceDesc>
                <bibl>
                    <title>The prince of darkness. </title>
                    <author domicile="Epping" n="DoherP1">Doherty, P C</author>
                    <imprint n="HEADLI1">
                        <publisher>Headline Book Publishing plc</publisher>
                        <pubPlace>London</pubPlace>
                        <date value="1992">1992</date>
                    </imprint>
                </bibl>
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            <tagsDecl>
                <namespace name="">
```

TEI header = metadata

text from British National Corpus

information about this text

TEI standard (BNC)

```
<wtext type="FICTION">
    <pb n="69"/>
    <div level="1">
        <head>
            <s n="2">
                <w c5="NN1" hw="chapter" pos="SUBST">Chapter </w>
                <w c5="CRD" hw="5" pos="ADJ">5</w>
            </s>
        </head>
        <p>
            <s n="3">
                <w c5="VVB-NN1" hw="ranulf" pos="VERB">Ranulf </w>
                <w c5="CJC" hw="and" pos="CONJ">and </w>
                <w c5="NP0" hw="dame" pos="SUBST">Dame </w>
                <w c5="NP0" hw="agatha" pos="SUBST">Agatha </w>
                <w c5="VBD" hw="be" pos="VERB">were </w>
                <w c5="VVG" hw="wait" pos="VERB">waiting </w>
                <w c5="PRP" hw="for" pos="PREP">for </w>
                <w c5="PNP" hw="he" pos="PRON">him </w>
                <w c5="PRP" hw="near" pos="PREP">near </w>
                <w c5="AT0" hw="the" pos="ART">the </w>
                <w c5="NN1-NP0" hw="galilee" pos="SUBST">Galilee </w>
                <w c5="NN1" hw="gate" pos="SUBST">Gate</w>
                <c c5="PUN">, </c>
                <w c5="AT0" hw="the" pos="ART">the </w>
                <w c5="AJ0" hw="young" pos="ADJ">young </w>
                <w c5="NN1" hw="nun" pos="SUBST">nun </w>
                <w c5="AV0" hw="apparently" pos="ADV">apparently </w>
                <w c5="VVG" hw="enjoy" pos="VERB">enjoying </w>
                <w c5="AT0" hw="an" pos="ART">an </w>
                <w c5="NN1" hw="account" pos="SUBST">account </w>
                <w c5="PRF" hw="of" pos="PREP">of </w>
                <w c5="CRD" hw="one" pos="ADJ">one </w>
                <w c5="PRF" hw="of" pos="PREP">of </w>
                <w c5="DPS" hw="he" pos="PRON">his </w>
                <w c5="NN1" hw="manservant" pos="SUBST">manservant</w>
                <w c5="POS" hw="'s" pos="UNC">'s </w>
                <w c5="DT0" hw="many" pos="ADJ">many </w>
                <w c5="NN2" hw="escapade" pos="SUBST">escapades </w>
                <w c5="PRP" hw="in" pos="PREP">in </w>
                <w c5="NP0" hw="london" pos="SUBST">London</w>
                <c c5="PUN">.</c>
```

TEI body = object data + annotation

structure & typographic markup

tokens + token-level annotations

principle:
raw text (= object data)
can be reconstructed by
deleting all XML tags

Vertical text format (.vrt)

Simpler, more efficient format → used by CWB & NLP tools

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

```
<corpus>
<story title="The Garden">
<p>
<s>
It        PP    it
seemed    VBD   seem
a         DT    a
day       NN    day
much      RB    much
as        IN    as
any       DT    any
other     JJ    other
until     IN    until
I         PP    I
...
</s>
</p>
</story>
</corpus>
```

TAB characters (\t, \x09)

metadata

title:	The Garden
author:	Stefan Evert
author sex:	male
date:	05.08.1991

Vertical text format (.vrt)

Text metadata encoded in XML start tags (not in header!)

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

```
<corpus>
<text title="The Garden" author="Stefan Evert" author_sex="male"
      date="1991-08-05">
<p num="1">
<s>
It        PP    it
seemed    VBD   seem
a         DT    a
day       NN    day
much      RB    much
as        IN    as
any       DT    any
other     JJ    other
until     IN    until
I         PP    I
...
</s>
</p>
</text>
</corpus>
```

CQPweb requires **<text>**,
SketchEngine prefers **<doc>**

sub-text level metadata

CoNLL format

Vertical text without the metadata

http://universaldependencies.org/docs/format.html

```
# story: "The Garden"
# paragraph #1
1   It       PP   it
2   seemed   VBD  seem
3   a        DT   a
4   fine     JJ   fine
5   day      NN   day
6   .        SENT .

1   There    EX   there
2   was      VBD  be
3   an       DT   a
4   elephant NN   elephant
5   .        SENT .

# this is the end of the file
```

these are just comments

blank lines = sentence boundaries

token numbers (within sentence)

CWB data model (~ column store)

Efficient binary format used by indexing & query tools

#	word	pos	lemma
0	A	DET	a
1	fine	ADJ	fine
2	example	NN	example
3	.	PUN	.
4	Very	ADV	very
5	fine	ADJ	fine
6	examples	NN	example
7	!	PUN	!

corpus position (“cpos”)

CWB data model (~ column store)

Efficient binary format used by indexing & query tools

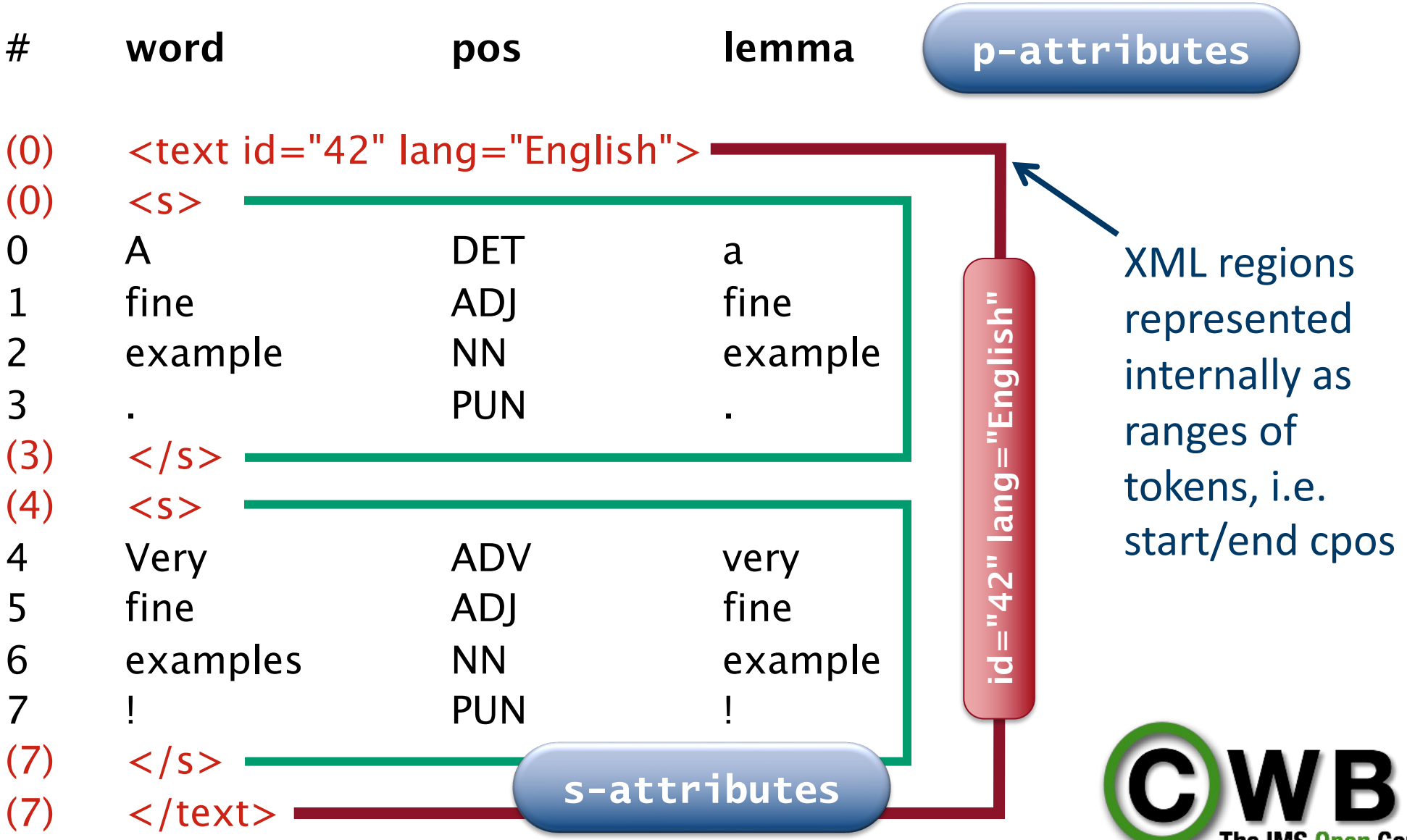
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

#	word	pos	lemma
(0)	<text id="42" lang="English">		
(0)	<s>		
0	A	DET	a
1	fine	ADJ	fine
2	example	NN	example
3	.	PUN	.
(3)	</s>		
(4)	<s>		
4	Very	ADV	very
5	fine	ADJ	fine
6	examples	NN	example
7	!	PUN	!
(7)	</s>		
(7)	</text>		

XML tags inserted as "invisible" tokens

CWB data model (~ column store)

Efficient binary format used by indexing & query tools



CWB data model (~ column store)

Efficient binary format used by indexing & query tools

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

#	word		pos		lemma	
(0)	<text id="42" lang="English">					
(0)	<s>					
0	A	0	DET	0	a	0
1	fine	1	ADJ	1	fine	1
2	example	2	NN	2	example	2
3	.	3	PUN	3	.	3
(3)	</s>					
(4)	<s>					
4	Very	4	ADV	4	very	4
5	fine	1	ADJ	1	fine	1
6	examples	5	NN	2	example	2
7	!	6	PUN	3	!	5
(7)	</s>					
(7)	</text>					

lexicon IDs for annotation strings (per column)

Corpus query & analysis with CQPweb

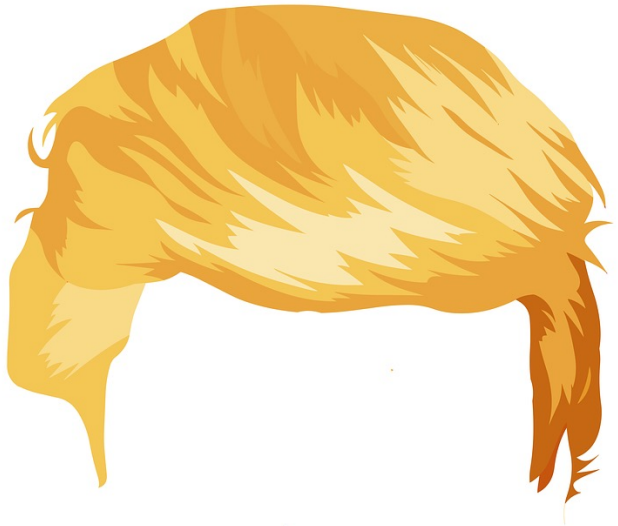
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Techniques of corpus linguistics: A hands-on example

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

- Goal: Understand this [→] guy's rhetoric, terminology, phraseology and topics just in case he comes back …
- How would you approach this task?



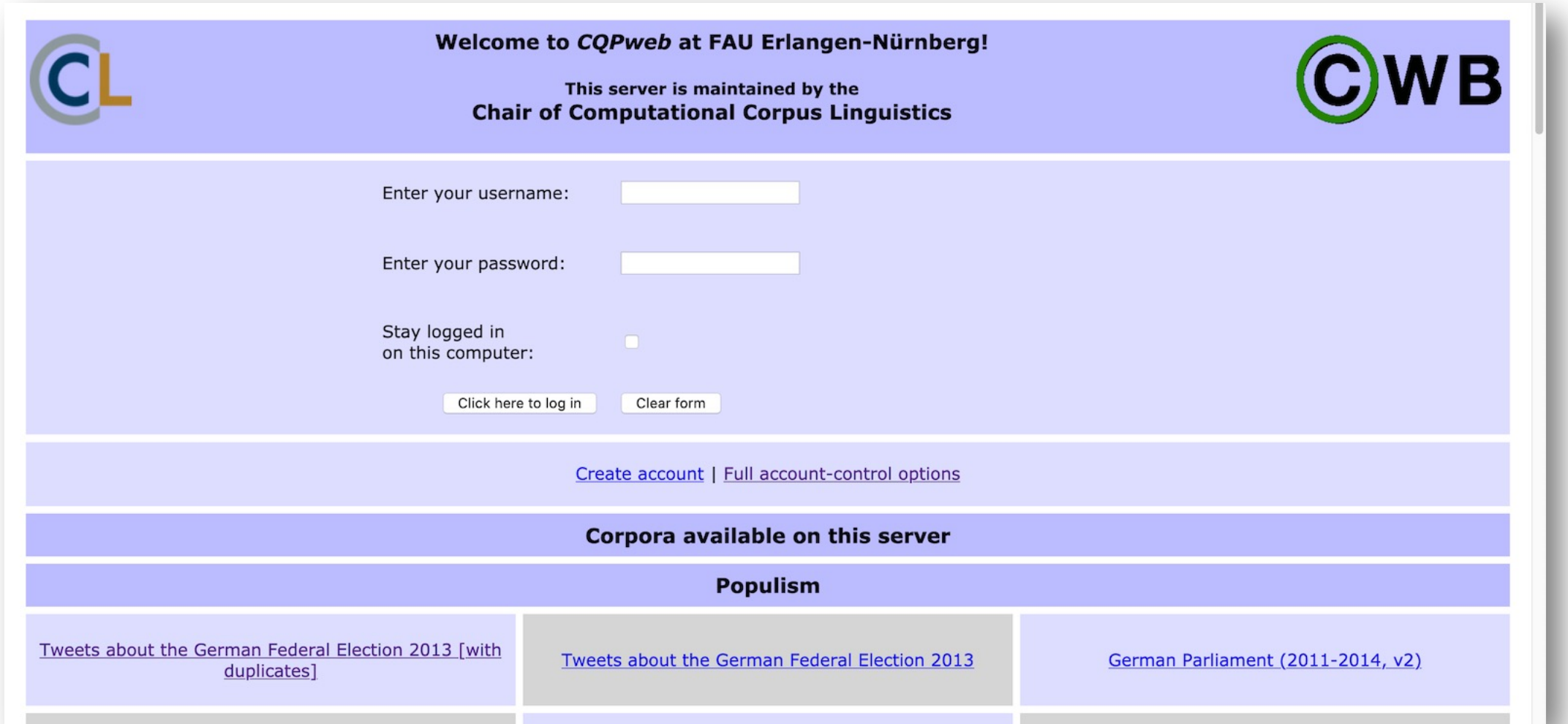
- We need a large collection of texts written by Trump, e.g. the Trump Twitter Archive [www.thetrumparchive.com]

- Step 1: Compile a linguistically annotated corpus (TTA)
- Step 2: Upload to indexing & search engine
- Step 3: Analyse with corpus-linguistic techniques

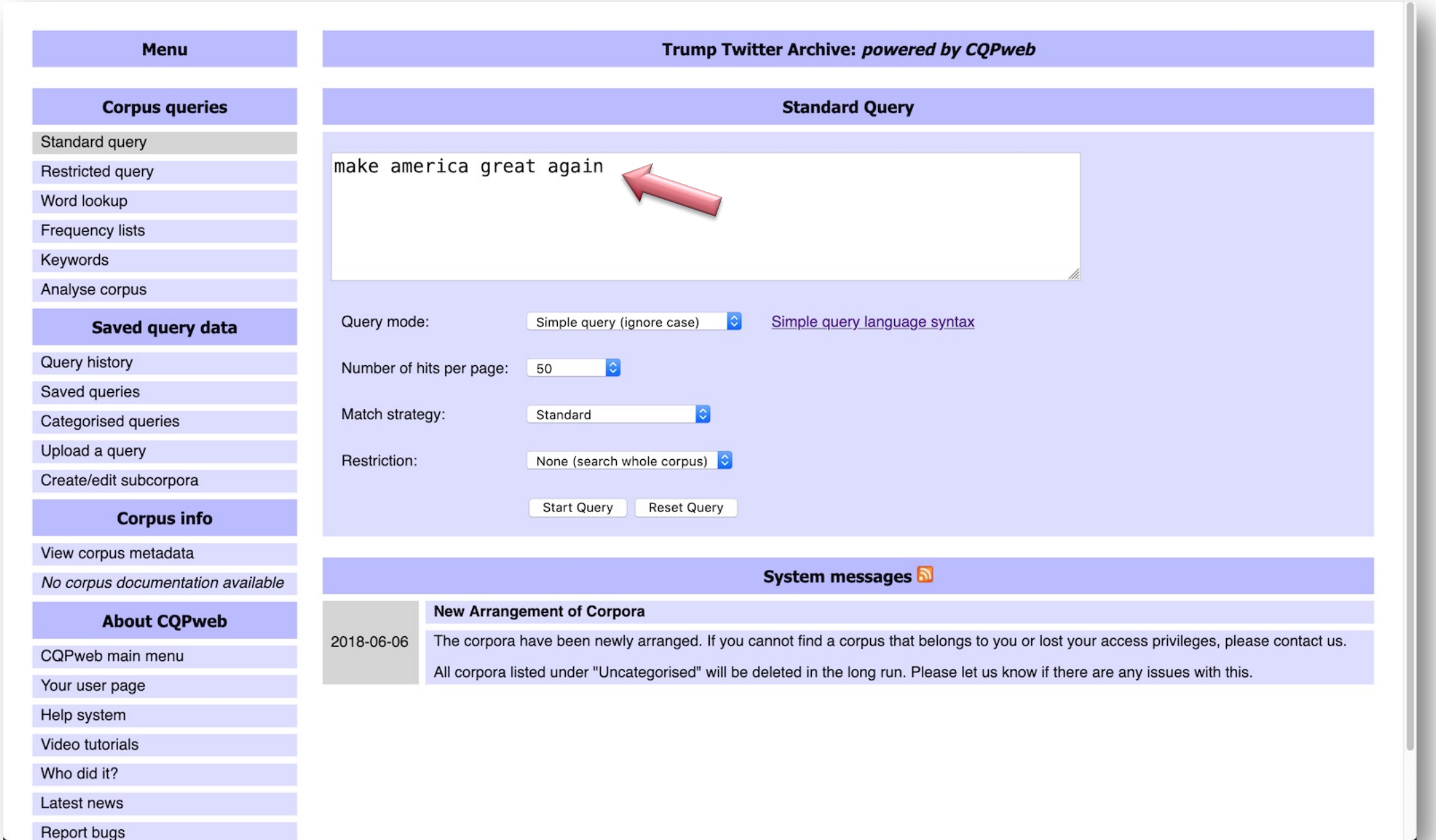
Techniques of corpus linguistics: A hands-on example

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

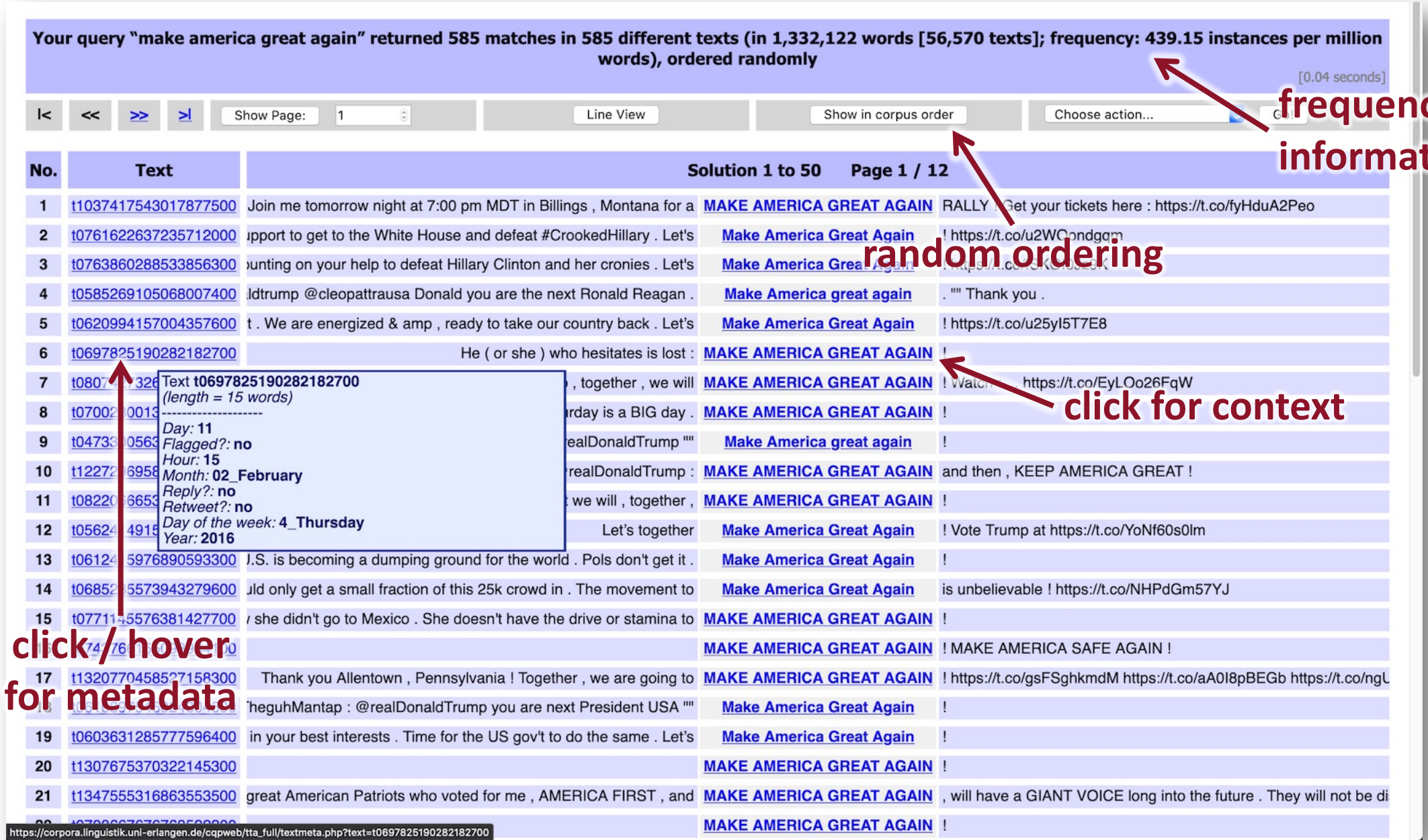
- Hands-on experience with CQPweb and the Trump Twitter Archive
  - go to https://corpora.linguistik.uni-erlangen.de/cqpweb/ and login as `student1` … `student15`
  - select corpus *Trump Twitter Archive* from "Political Tweets" section



Simple search



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]

|< << >> >| Show Page: 1 Line View Show in corpus order Choose action... Go

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	upport to get to the White House and defeat #CrookedHillary . Let's	Make America Great Again	! https://t.co/u2WCondgqm
3	t0763860288533856300	unting on your help to defeat Hillary Clinton and her cronies . Let's	Make America Great Again	
4	t0585269105068007400	ldtrump @cleopattrausa Donald you are the next Ronald Reagan .	Make America great again	. "" Thank you .
5	t0620994157004357600	t . We are energized & amp , ready to take our country back . Let's	Make America Great Again	! https://t.co/u25yI5T7E8
6	t0697825190282182700	He (or she) who hesitates is lost :	MAKE AMERICA GREAT AGAIN	!
7	t080 326	, together , we will	MAKE AMERICA GREAT AGAIN	! Watch https://t.co/EyLOo26FqW
8	t0700 0013	rday is a BIG day .	MAKE AMERICA GREAT AGAIN	!
9	t0473 0563	ealDonaldTrump ""	Make America great again	!
10	t1227 6958	realDonaldTrump :	MAKE AMERICA GREAT AGAIN	and then , KEEP AMERICA GREAT !
11	t0822 6653	we will , together ,	MAKE AMERICA GREAT AGAIN	!
12	t0562 4915	Let's together	Make America Great Again	! Vote Trump at https://t.co/YoNf60s0lm
13	t0612 5976890593300	J.S. is becoming a dumping ground for the world . Pols don't get it .	Make America Great Again	!
14	t0685 5573943279600	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	!
			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/aA0I8pBEGb https://t.co/ngU
		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
			MAKE AMERICA GREAT AGAIN	!

Text **t0697825190282182700**
(length = 15 words)

Day: **11**
Flagged?: **no**
Hour: **15**
Month: **02_February**
Reply?: **no**
Retweet?: **no**
Day of the week: **4_Thursday**
Year: **2016**

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

frequency information

random ordering

click for context

click/hover for metadata

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

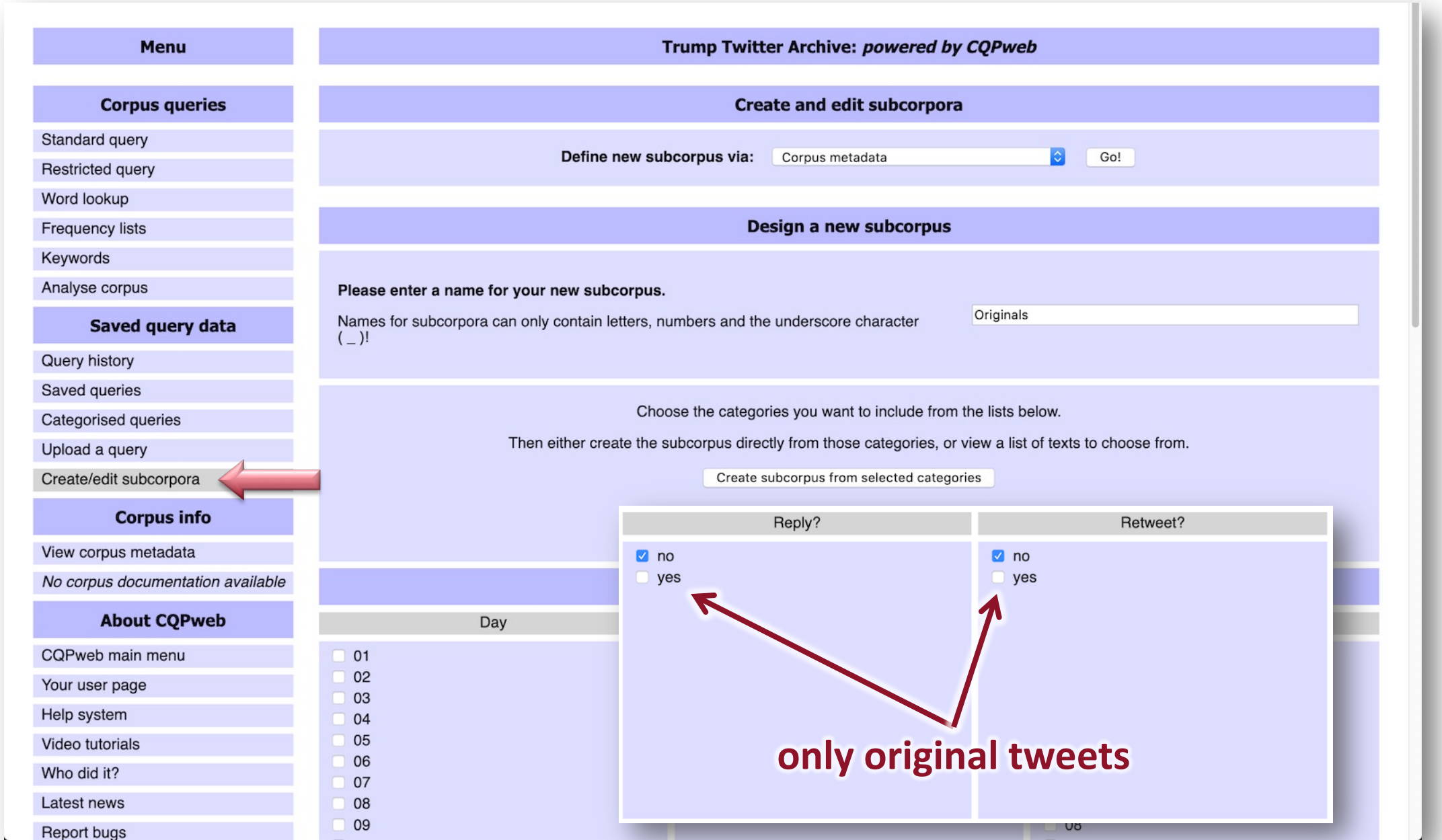
2016-02-12 *02:28:28* [⏩ 3822, ❤️ 9677] THANK YOU- Baton Rouge , Louisiana ! WE will #MakeAmericaGreatAgain ! #Trump2016 https://t.co/XV7Ele7A2l

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 & amp , others ? https://t.co/Tqej1euLVL

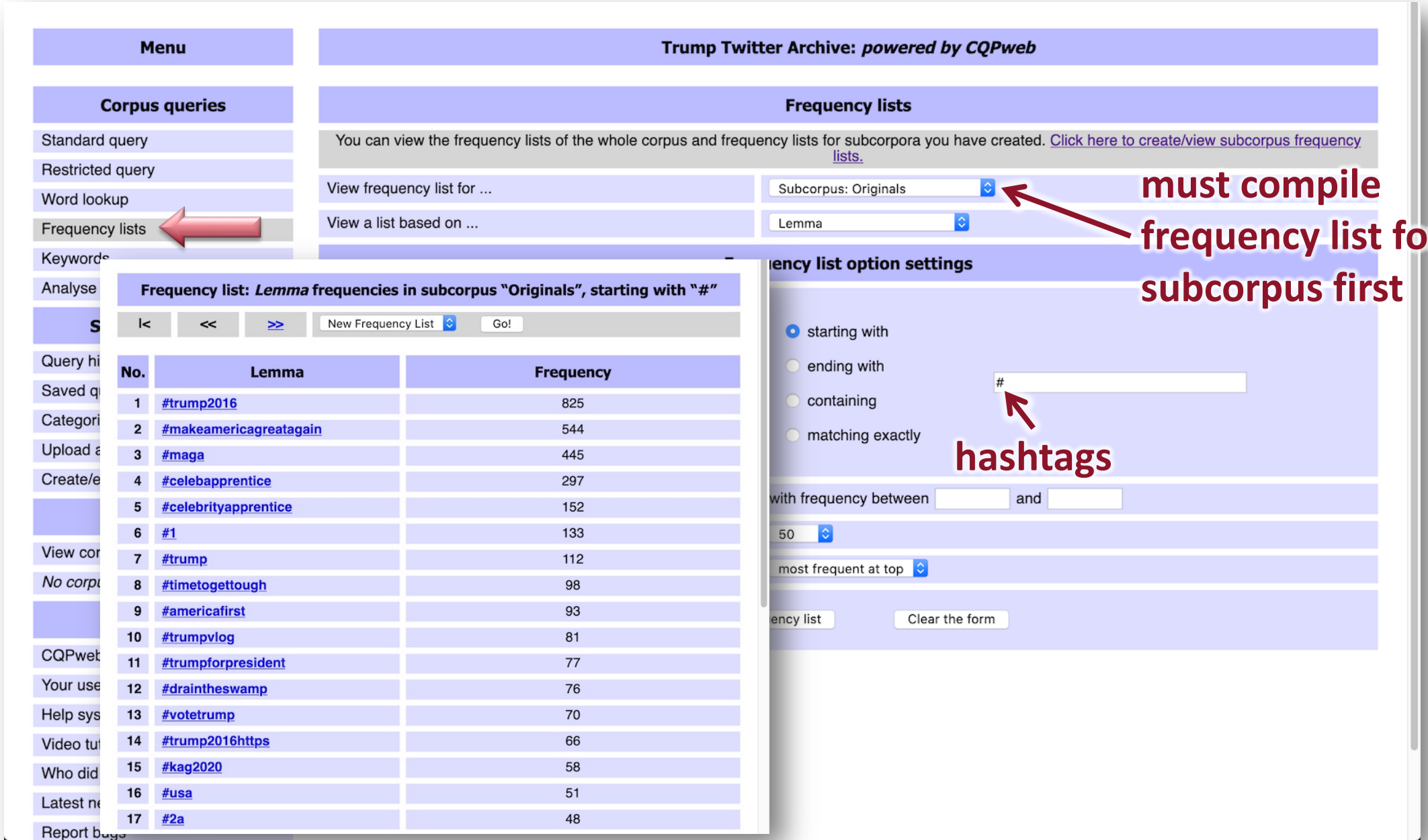
2016-02-12 *03:20:33* [⏩ 3357, ❤️ 9104] Lying Cruz put out a statement , " Trump &

CQPweb v3.2.41 © 2008-2020 Help! on CQPweb You are logged in as user [student13]

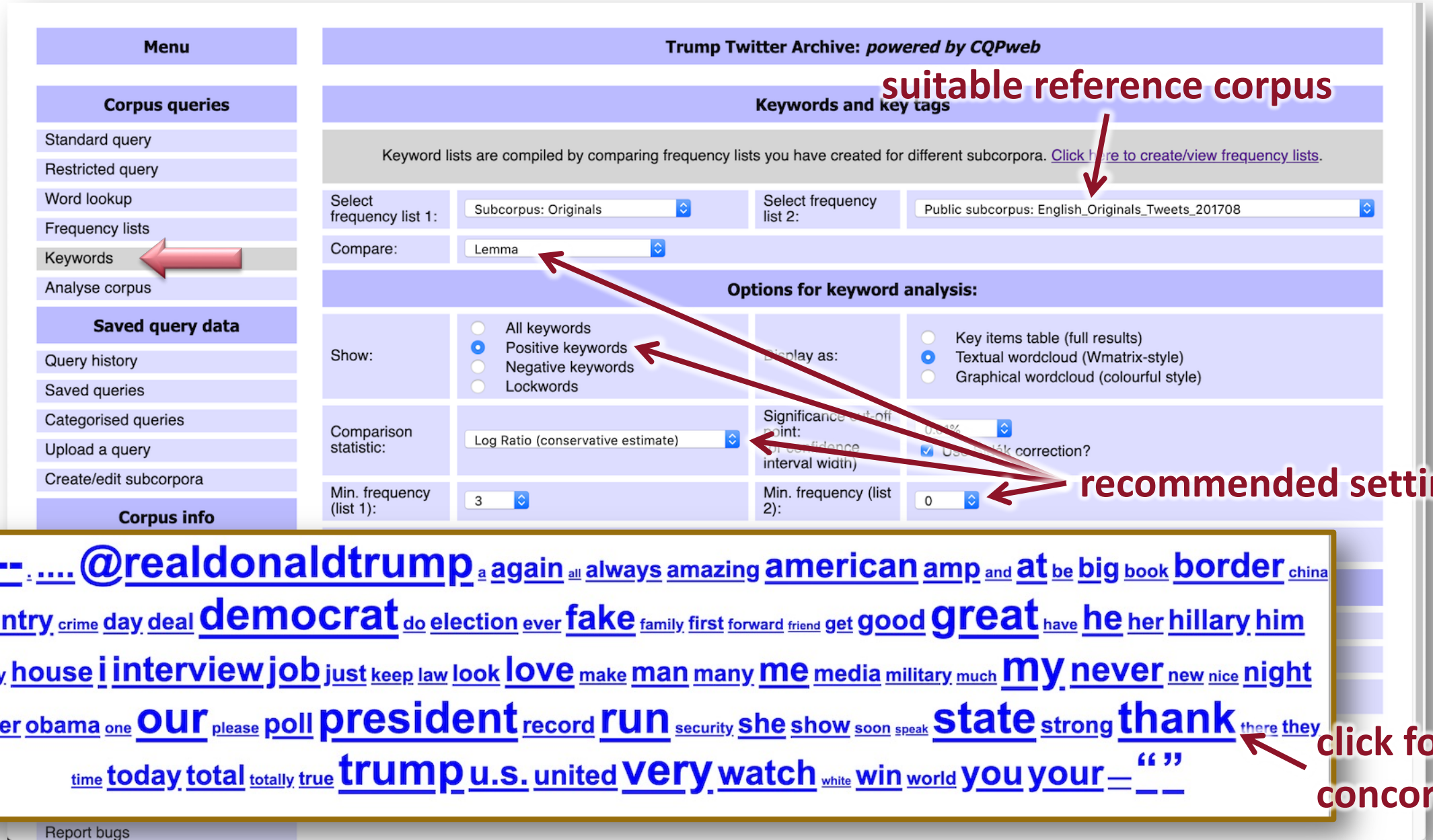
Create subcorpus for targeted search



Frequency list for hashtags



Keywords vs. reference corpus



Sorted concordance: *great*

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

action menu

Your query "[lemma="great"%cd]", 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]

|< << >> >| Show Page: 1 | Line View | Show in random order | ✓Choose action... Go!

Sort control: | Position: 1 Right | Tag restriction: None exclude | Starting with: exclude

- ✓Choose action...
- New query
- Thin...
- Frequency breakdown
- Distribution
- Dispersion
- Sort
- Collocations...
- Download...
- Categorise...
- Save current query result...

No.	Text	Solution 2101 to 2150		Page 43 / 147
2101	t1122270956192272400	Sincerest THANK YOU to our	great	**Border** Patrol Agent who st gue in Poway , Ca
2102	t1104017664131907600	ending 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/BHPu6IIdAN
2111	t0564215204154966000	ldTrump @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/p
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*

full process history

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

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]

I< << >> >I 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*

Distribution across metadata categories

Collocation analysis

key parameters

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

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:	and/or tag: V V	Submit changed parameters 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	[illegible]
5	doesn't	249	2.071	18	18	47.059
6	mention	129	1.073	13	12	42.165
7	say	1,987	16.529	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	4.725	14	14	12.027

click for examples

Choose settings for proximity-based collocations:

Include annotation:	Lemma	Include / Exclude
	POS-disambiguated lemma	Include / Exclude
	Part of speech	Include / Exclude
Maximum window span:	+/- 5	

Create collocation database

Thank you for listening!

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE