

מדעי הרוח הדיגיטליים:

קריאה רחוקה

אוניברסיטת תל-אביב

הספרייה המרכזית ע"ש סוראסקי

Adapted from:

Heather Froehlich, "[Corpus Analysis with Antconc](#)," Programming Historian 4 (2015)

Itay Marienberg-Milikowsky, "[Hebrew Guide for CATMA 5](#)," catma.de (2017)

עידו אהרון

idoah@tauex.tau.ac.il

עודד זרחיה

odedzarchia@tauex.tau.ac.il



מהי קריאה רחוקה?

”קריאה שבה המרחק, אם יורשה לי לחזור על עצמי, הוא תנאי לידע, שכן הוא מאפשר לנו להתמקד ביחידות שהן קטנות בהרבה או גדולות בהרבה מן הטקסט עצמו: תחבולות, תמות, טרופים (tropes) – או ז'אנרים ומערכות שלמות”

(פרנקו מורטי, [השערות על ספרות העולם](#), 2000)

- השוואה בין מושאים טקסטואליים בקנה מידה גדול
- איתור דפוסים לשוניים חוזרים, תבניות וצירופי מילים
- ויזואליזציה של טקסטים

מטרות :

1. יצירה/הורדה של קורפוס טקסטואלי
2. ניתוח ממוחשב של טקסטים בתוכנת AntConc (KWIC, collocates, wildcards)
3. ויזואליזציה של טקסטים בתוכנת CATMA (distribution, doubletree, wordcloud)
4. תיוג חופשי של טקסטים בתוכנת CATMA

סוגים של קבצי טקסט



embargo [em-bahr-goh]

See synonyms for: embargo / embargoed / embargoing

noun, plural em·bar·goes.

1. an order of a government prohibiting the movement of merchant ships into or out of its ports.
2. an injunction from a government commerce agency to refuse freight for shipment, as in case of congestion or insufficient facilities.
3. any restriction imposed upon commerce by edict.
4. a restraint or hindrance; prohibition.

Plain text

(***.txt**)



embargo [em-bahr-goh]

See synonyms for: [embargo](#) / [embargoed](#) / [embargoing](#)

noun, plural **em·bar·goes**.

- 1 an order of a government prohibiting the movement of merchant ships into or out of its ports.
- 2 an injunction from a government commerce agency to refuse freight for shipment, as in case of congestion or insufficient facilities.
- 3 any restriction imposed upon commerce by edict.
- 4 a restraint or hindrance; prohibition.

Rich text

(***.docx**, ***.pages**)

עבודה עם קבצי תמליל פשוט (Plain Text)

- מובנית לתוך מערכת ההפעלה:
- חלונות – **Notepad** (C:\Windows\System32\notepad.exe)
- מקינטוש – **TextEdit** (Applications)
- לינוקס – **Vim** (/etc/vim/vimrc)
- חלופות:



עבודה עם קבצי תמליל פשוט (המשך)

Plain Text

קולנוע | אוסקר 2022
אוסקר 2022: כריס רוק לעג לאשתו של ויל סמית – שחבט בו בתגובה
רוק התבדח בזמן הטקס על ראשה המגולח של ג'יידה פינקט סמית. בתגובה, זוכה האוסקר

שתפו בפייסבוק
שתפו וחבריכם יקראו את הכתבה ללא תשלום
שתפו כתבה במיילשתפו כתבה במייל
מעבר לטוקבקים
80

שמרו
שמרו את הכתבה לרשימת הקריאה

קריאת זן
הדפיסו כתבה
עברו לתצוגת גלריה
ויל סמית מחטיף אגרוף
ויל סמית מכה את כריס רוק, לפנות בוקר צילום: BRIAN SNYDER/רויטרס
נירית אנדרמן
נירית אנדרמן
קבל התראות בתיבת הדואר האלקטרוני שלך עבור כתבות מ נירית אנדרמן
התראות במייל
08:03, 28 במרץ 2022

ויל סמית זכה לפנות בוקר בפרס האוסקר לשחקן הטוב ביותר על תפקידו בסרט "משפחה מ
רוק, שהגיש את פרס הסרט התינודי הטוב ביותר ל"הקיץ של סול", הרשה לעצמו להתלוצץ
על ראשה המגולח של פינקט סמית (שחשפה בעבר כי הסיבה לכך היא רפואית), ואמר שהוא מצפה כבר לראות את סרט ההמשך ל"ניי איי ניו". סמית צחק

HTML

קולנוע | אוסקר 2022

אוסקר 2022: כריס רוק לעג לאשתו של ויל סמית - שחבט בו בתגובה

רוק התבדח בזמן הטקס על ראשה המגולח של ג'יידה פינקט סמית. בתגובה, זוכה האוסקר הטרי עלה לבמה והכה אותו לתדהמת הצופים. "בעסק הזה אתה צריך לדעת להתמודד עם אנשים שלא מכבדים אותך", אמר לאחר מכן

80 שמרו קריאת זן



ויל סמית מכה את כריס רוק, לפנות בוקר צילום: BRIAN SNYDER/רויטרס

ויל סמית זכה לפנות בוקר בפרס האוסקר לשחקן הטוב ביותר על תפקידו בסרט "משפחה מנצחת", אולם הדרמה האמיתית בסיכנו התרחשה מוקדם יותר. סמית, שהיה את מיליוני הצופים בטקס כאשר ניגש לבמה והחטיף סטירה עזה לקומיקאי כריס רוק שהתבדח על חשבונה של אשתו של סמית, ג'יידה פינקט סמית.

רוק, שהגיש את פרס הסרט התינודי הטוב ביותר ל"הקיץ של סול", הרשה לעצמו להתלוצץ על ראשה המגולח של פינקט סמית (שחשפה בעבר כי הסיבה לכך היא רפואית), ואמר שהוא מצפה כבר לראות את סרט ההמשך ל"ניי איי ניו". סמית צחק



נירית אנדרמן

התראות במייל

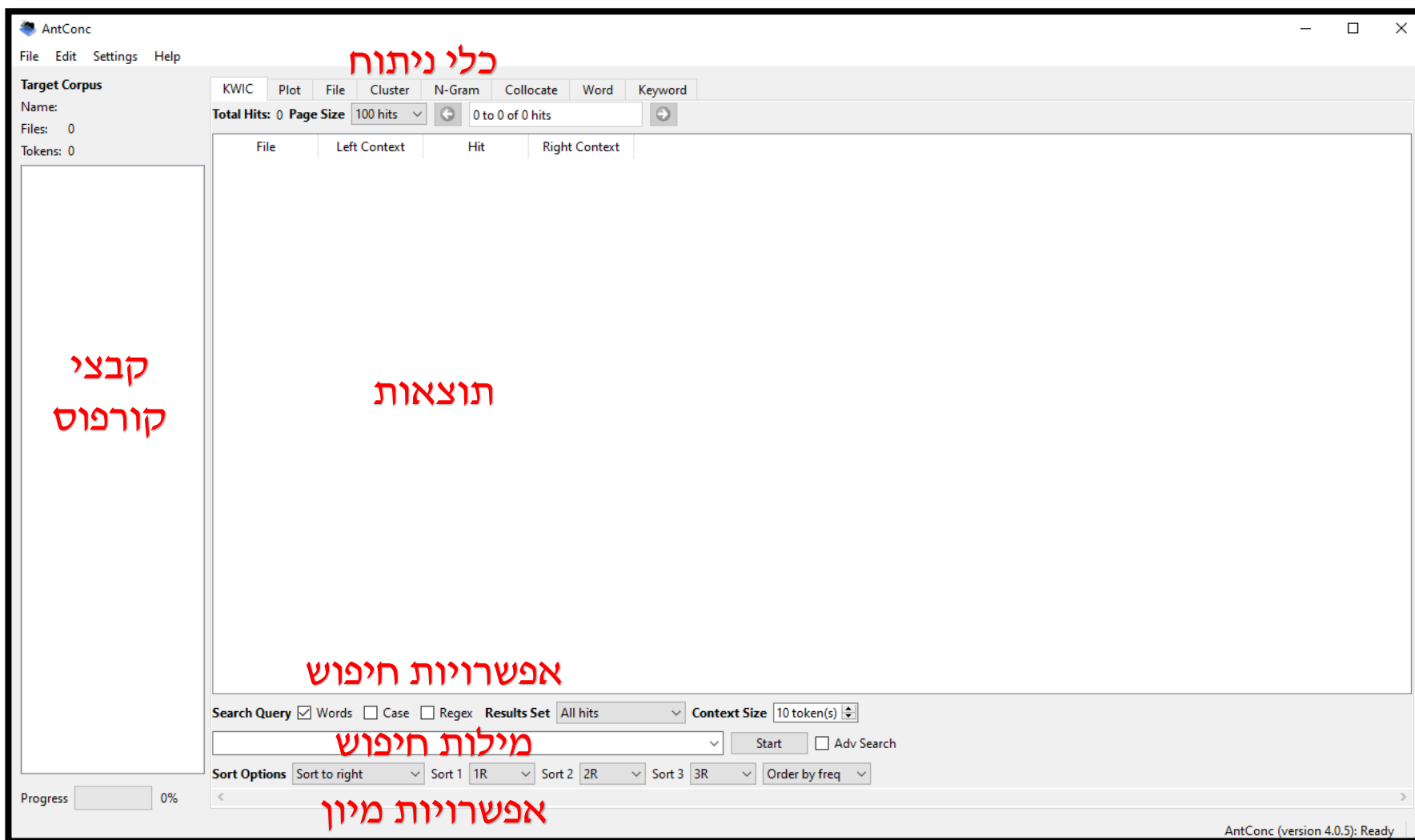
28, 08:03 בנרץ 2022



AntConc

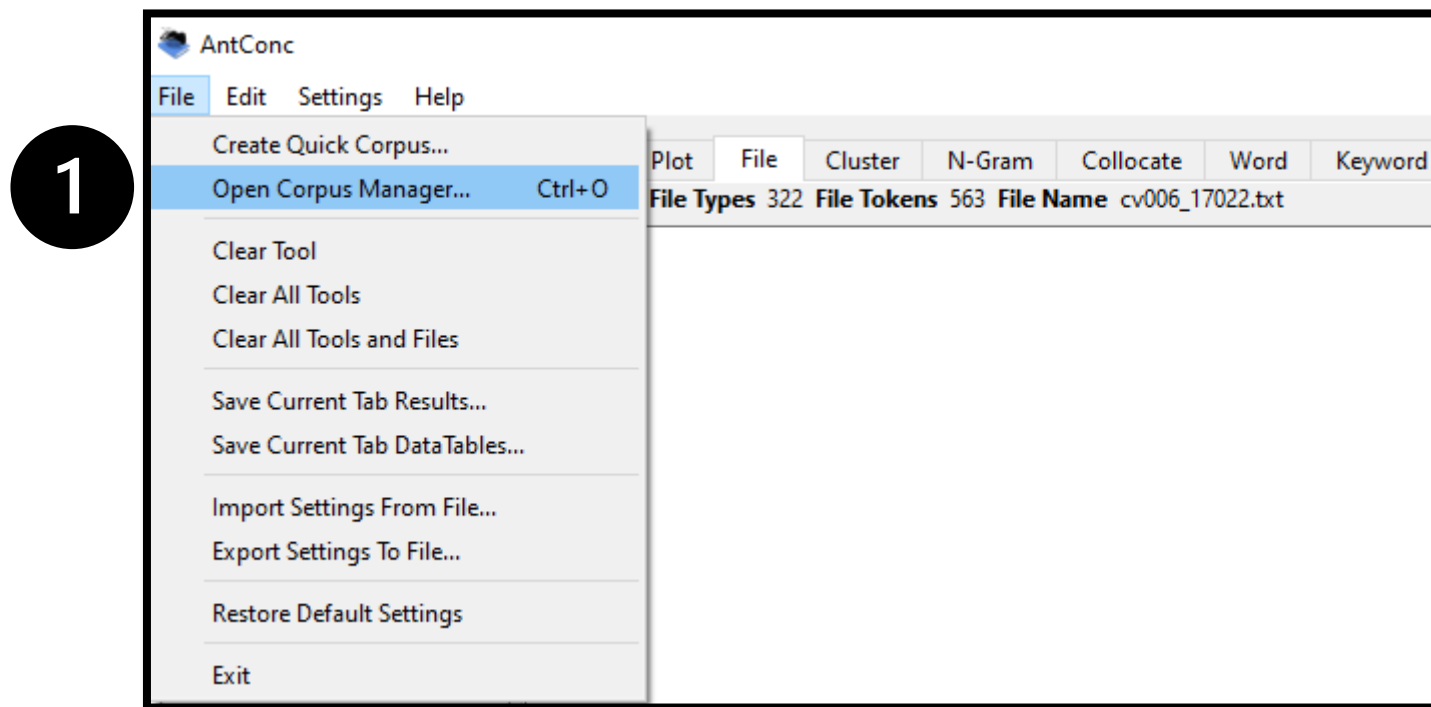
- תוכנה חופשית (Freeware) לניתוח טקסטואלי ממוחשב שפותחה ע"י פרופ' לורנס אנטוני (אוניברסיטת ואסדה, יפן)
- מערכות הפעלה נתמכות : Windows, MacOS, Linux
- גרסה אחרונה : 4.0.5 (2022)
- כתובת להורדה :
- <https://www.laurenceanthony.net/software/antconc/>

AntConc - מסך ראשי



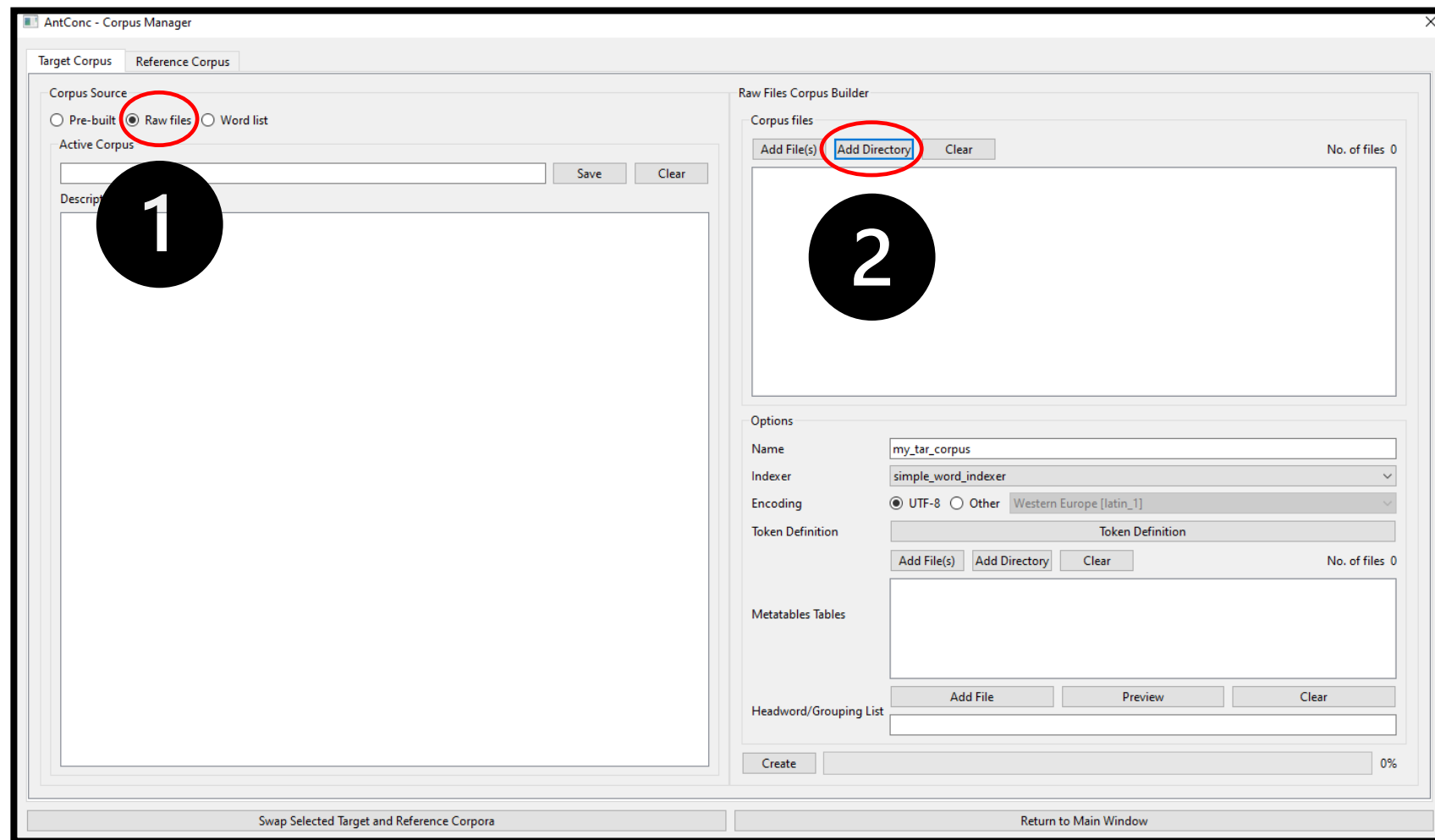
AntConc - העלאת קבצי טקסט

1. Go to 'File' and click on 'Open Corpus Manager'



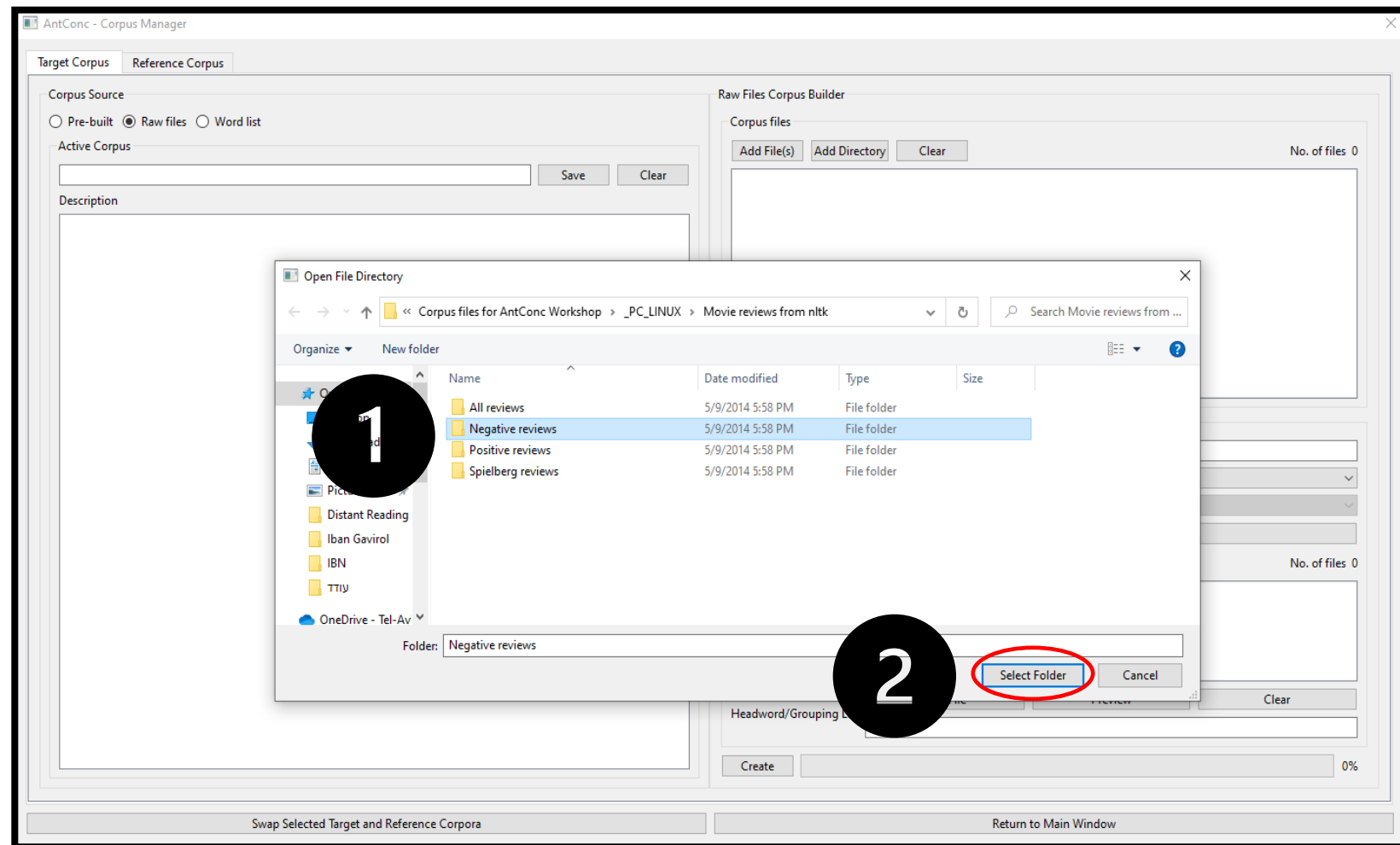
AntConc - העלאת קבצי טקסט #1

1. Under 'Corpus Source' select 'Raw files'
2. Under 'Raw Files Corpus Builder' hit 'Add Directory'



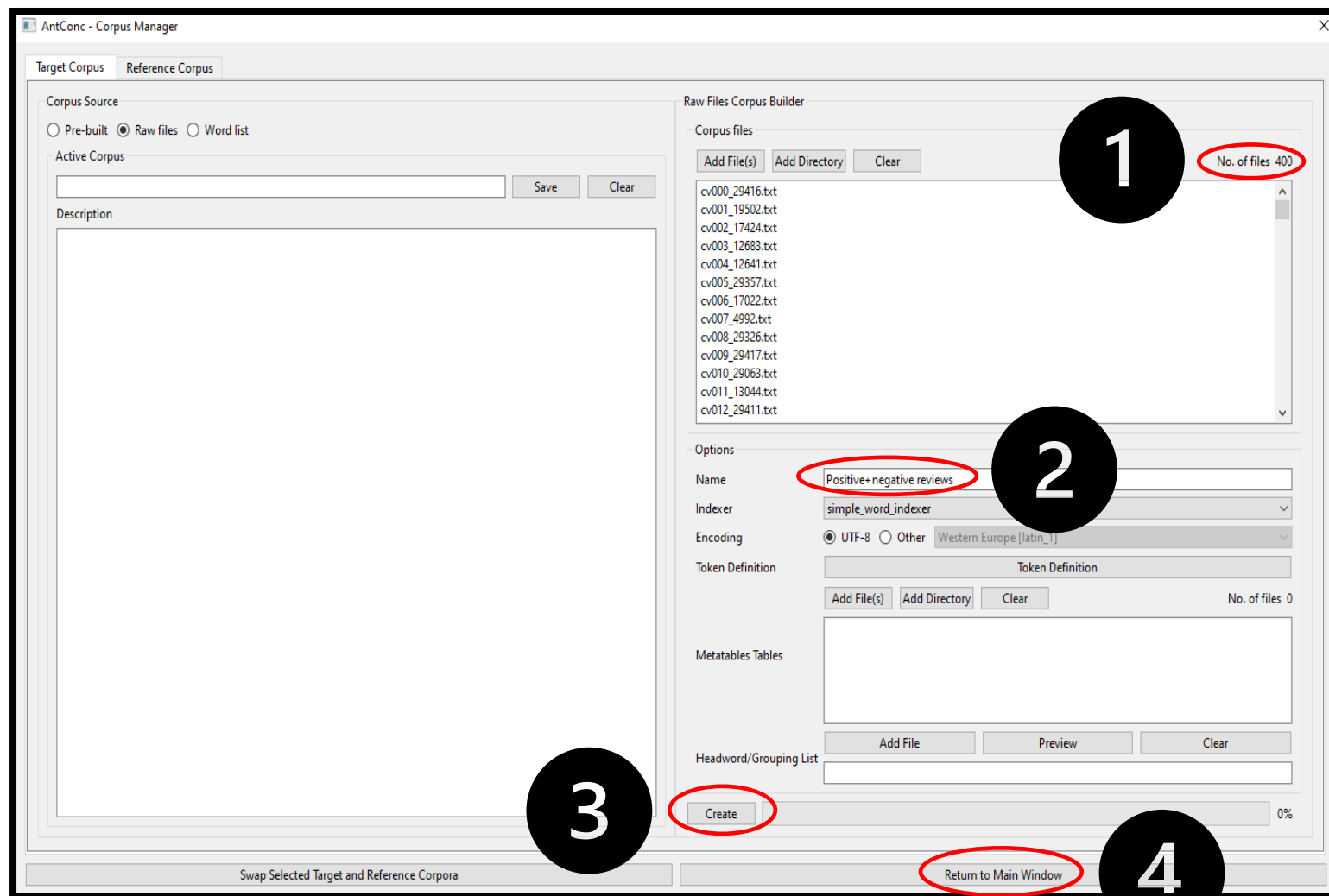
AntConc - העלאת קבצי טקסט #2

1. Navigate to our movie reviews folder and select 'Negative reviews'
2. Hit 'Select Folder'
3. Repeat actions for the 'Positive reviews' folder



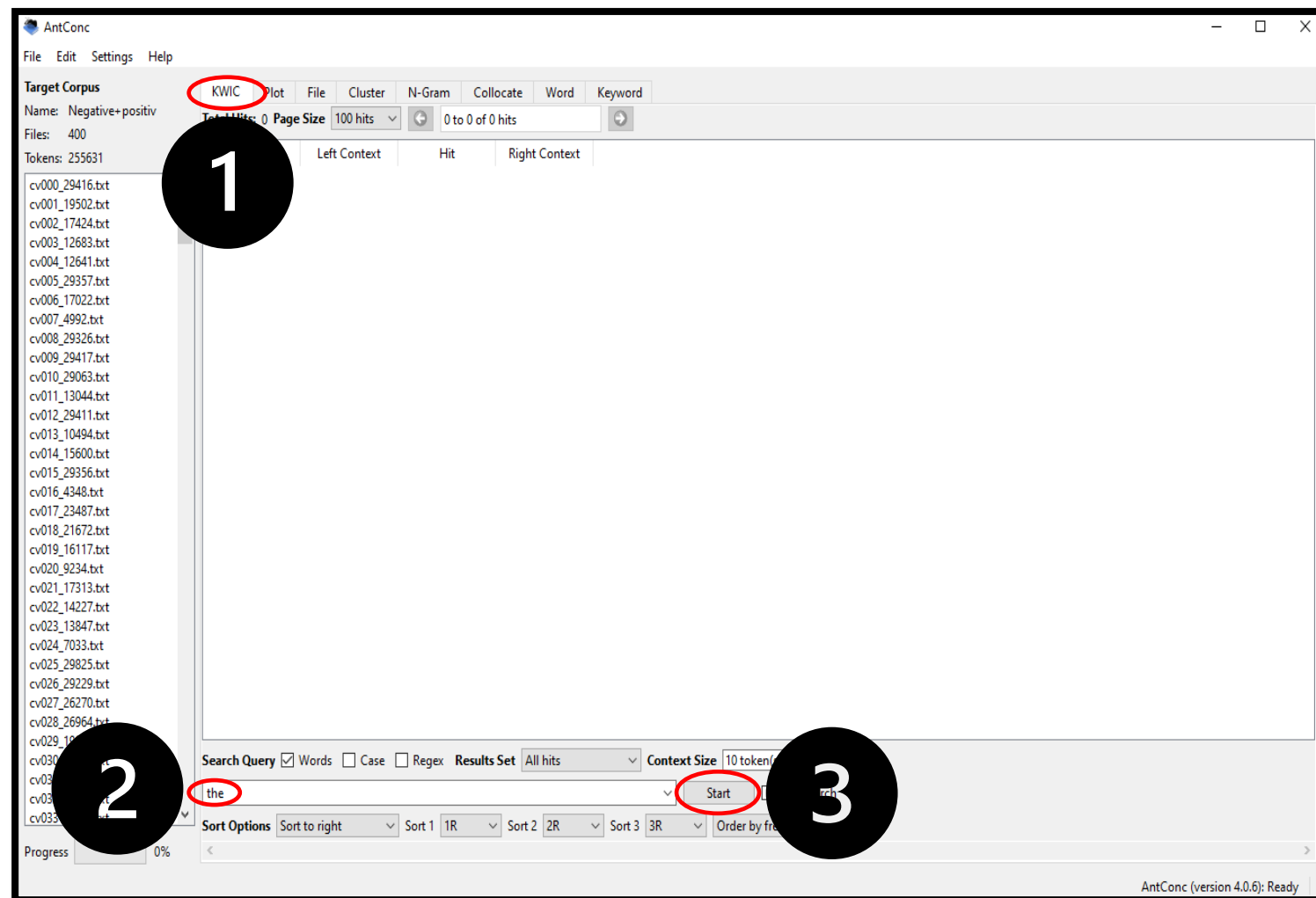
AntConc - העלאת קבצי טקסט #3

1. The 'Corpus Manager' should now display 400 .txt files
2. Under 'Options' name your corpus as 'Negative+positive reviews'
3. Hit 'Create'
4. Once loading is finished, click on 'Return to Main Window'



KWIC - חיפוש בסיסי

1. Navigate to the 'KWIC' tab
2. Type in the search box a stop word such as 'the'
3. Hit 'Start'



KWIC - חיפוש בסיסי (תוצאות)

1. In the 'Hit' column
AntConc displays every
time your stop word
appears in the corpus,
and some context for it
(‘Left Context’/Right
Context’)

The screenshot shows the AntConc application window. The 'KWIC' tab is selected. The search query is 'the'. The results are displayed in a table with columns: File, Left Context, Hit, and Right Context. A red circle with the number '1' highlights the 'Hit' column. The table shows 17 hits, each with a file path, a left context sentence, the word 'the', and a right context sentence. The search options at the bottom are: Search Query: the, Words checked, Case unchecked, Regex unchecked, Results Set: All hits, Context Size: 10 token(s), Sort Options: Sort to right, Sort 1: 1R, Sort 2: 2R, Sort 3: 3R, Order by freq.

File	Left Context	Hit	Right Context
1 cv031_19540.txt	with little gift for ad-lib . about three-quarters of	the	film is a somewhat unnecessary origin story for both
2 cv073_23039.txt	otherwise given up on this pathetic attempt at romantic comedy .	the	film is a multi-million dollar testament to everything
3 cv098_17021.txt	levity . in some respects , it works , but the majority of	the	film is a convoluted and confusing mess . characters keep
4 cv170_29808.txt	earthworms , but there are some decidedly strange david lynch touches .	the	film is a little long for the subject matter .
5 cv191_29539.txt	whole team listens to it in the morning . " everyone in	the	film is a device , and it's painful to
6 cv019_14482.txt	drug addiction as maintenance . one of the earliest observations in	the	film is a casual reference to " naked lunch " author
7 cv035_3954.txt	of isaac mizrahi , an artist whose palette is fabric . ostensibly ,	the	film is a documentary , but use of that term
8 cv084_13566.txt	the issues ; sex , love , murder , and betrayal . the setting of	the	film is a town named blue bay . it consists
9 cv180_16113.txt	it rather than going to the authorities . the rest of	the	film is a question of who is going to
10 cv083_25491.txt	audience to become attached to . but the main problem with	the	film is its inability to create any type of
11 cv150_14279.txt	want the role of straight man . no , the problem with	the	film is its direction . the first five minutes of
12 cv102_7846.txt	him solve the mystery . one of the greatest charms to	the	film is its sense of humor . there is a
13 cv122_7392.txt	within the confines of its plot . a key ingredient to	the	film is its ability to create an ongoing polemic
14 cv026_29325.txt	the public sitting in those seats . the only drawback to	the	film is the ending , which unfortunately i must reveal .
15 cv160_10362.txt	brothers studio back to the forefront of film production . secondly ,	the	film is the re-discovery of yuen biao . the
16 cv162_10424.txt	overblown . in terms of its characters , the strong part of	the	film is the interesting relationship between gale and dewey (
17 cv182_7281.txt	of paper , from shue . one of the enjoyable features in	the	film is the disguises that kilmer dawns to complete

KWIC - חיפוש בסיסי (מיון)

1. Search for the word 'shot' (noun & verb)
2. AntConc sorts information based on the 'Sort Options' parameters:
 - a. 'Sort to right' = words after hit
 - b. 'L' corresponds with 'left' and 'R' corresponds with 'right (e.g., '1R' = first word after hit)

AntConc

File Edit Settings Help

Target Corpus
Name: Negative+positiv
Files: 400
Tokens: 255631

KWIC Plot File Cluster N-Gram Collocate Word Keyword

Total Hits: 65 Page Size 100 hits 1 to 65 of 65 hits

	File	Left Context	Hit	Right Context
1	cv084_15183.txt	some fantastic shots by de palma, including a 15 minute steadicam	shot	at the start of the film, halfway through the
2	cv151_15771.txt	was borderline-abysmal, and can't really be considered a	shot	at the title, mulan had the makings of winner.
3	cv147_22625.txt	fight each other, the winner of which is "promised" a	shot	at a future title fight. the film is divided
4	cv023_12672.txt	i like that the film doesn't take a clear	shot	at any of the political parties or affiliations, but
5	cv121_18621.txt	ve guessed that this is director alan cohn's first	shot	at directing a feature film? when respecting a director,
6	cv131_11568.txt	the jackal, "a character wandered around as people were being	shot	at in a big suspense sequence, and one of
7	cv183_18612.txt	the matrix and existenz, but it certainly gives a great	shot	at it. in the matrix, we were told that
8	cv056_14663.txt	my mind every time i saw a medium or long	shot.	at times, i was hoping to see it. one
9	cv131_10713.txt	young man with big dreams. in its bleak, foreboding opening,	shot	in grainy black and white, he gets a lesson
10	cv138_12721.txt	portrayed in ghandi. many scenes from this film were actually	shot	in montreal, canada (joblo's haven of restfulness). donald
11	cv004_11636.txt	hideous title sequence. this also explains why the movie was	shot	in odd places like morocco and spain. (chan's
12	cv179_9228.txt	it wouldn't feel so contrived if it weren't	shot	in such a tasteful manner. there are two characters
13	cv172_11131.txt	upside down, a shot rings out, a cop is down.	shot	in the back and paralyzed for life. the young
14	cv022_14227.txt	no exaggeration. we are even treated to a five second	shot	from the point of view of van damme's
15	cv141_17179.txt	it). imagine the first attack of the aliens in "aliens,"	shot	from the pov of the soldiers and then the
16	cv003_11664.txt	sharks (all nicknamed "bruce" by the crew) built to be	shot	from different angles. many have forgotten, but "jaws" was
17	cv197_29328.txt	excellent camera work by antoine heberle and jeanne lapoirie, who	shot	the film in a dreamy haze, and the lead

Search Query ☒ Words ☐ Case ☐ Regex Results Set All hits Context Size 10 token(s)

☐ Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Progress 100%

Time taken (creating KWIC display): 0.0895 ms

KWIC – חיפוש מתקדם (wildcards)

טכנולוגיית המידע (תשס"ז, 2007)

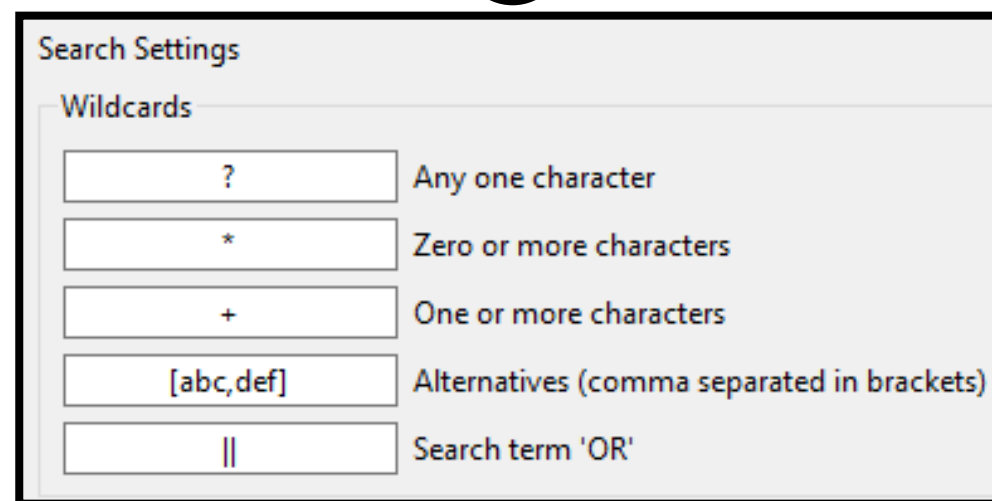
תו-כול [תו המשמש לחיפוש של מחרוזת שחלק מאותיותיה נשאר בלתי מוגדר, כגון כוכבית]

wildcard

מילון ארגון נתונים [טכנולוגיית המידע] (תשס"ג, 2002)

1

1. Go to 'Settings' -> 'Global Settings' and select 'Searches' to see what wildcards are available in AntConc



KWIC – חיפוש מתקדם (#1 operators)

1. (*) = zero or more characters (e.g., [dog*] = dog or dogs; [qualit*] = quality or qualities)
2. (?) = any one character (e.g., [m?n] = men or man; [wom*n] = woman or women)

AntConc

File Edit Settings Help

Target Corpus
Name: Negative+positiv
Files: 400
Tokens: 255631

KWIC Plot File Cluster N-Gram Collocate Word Keyword

Total Hits: 37 Page Size 100 hits 1 to 37 of 37 hits

File	Left Context	Hit	Right Context
1 cv173_4295.txt	a bad print with only half-faces and poor sound	quality ,	but i couldn't help myself . . . i got it .
2 cv019_16117.txt	the words " family entertainment " have never been a stamp of	quality ,	but might joe young is just a silly special
3 cv043_16808.txt	s not that i think exploitation is necessarily a good	quality ,	but when you are talking about a dull turkey
4 cv153_10779.txt	a funny way of playing characters with very few redeeming	qualities ,	but whom you end up liking anyway . in as
5 cv146_19587.txt	though they are , have an unmistakable been there-done that	quality ;	in fact , whole sequences , not to mention the cloaked ,
6 cv024_6778.txt	have been a great actor , but he had that unique	quality	in his eyes . in gridlock'd , an oddly effective
7 cv069_11613.txt	of the world , plot , etc . there's not one redeeming	quality	in the picture that is not technical . directed by
8 cv025_29825.txt	focus for a film that purports to articulate the specific	qualities	of couplehood . it just goes to show , an intelligent
9 cv032_23718.txt	the best and worst of american filmmaking with the same	qualities	of hong kong films . in a nutshell , the current
10 cv188_20687.txt	in any league) . you will be able to determine the	quality	of the film by examining the opening scene . sylvester
11 cv089_12222.txt	a filmed version of performance art -- lifeless and tedious . the	qualities	that make it captivating when seen in person are
12 cv115_25396.txt	remarkable " chicken run " really is . this movie possesses that same	quality	that makes pixar studios' animation great : precise and unwavering
13 cv081_16582.txt	or she's all that , it certainly holds an undeniable	quality	that older crowds will lean towards . it's a
14 cv157_29302.txt	in order to give him credibility as a publisher of	quality ,	and he quickly wins favor with her by publishing
15 cv028_26746.txt	of course in these situations , quantity is more important than	quality	and profit is more important than the artistical values .
16 cv172_12037.txt	haven't learned a thing about ergonomics . one vaguely redeeming	quality	is that the end credits succeed where the rest
17 cv195_14528.txt	well , and it's another good bet this month . since	quality	is usually sparse during the early fall months , it'

Search Query ☒ Words ☐ Case ☐ Regex Results Set All hits Context Size 10 token(s)

1 qualit* Start ☐ Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Progress 100%

Time taken (creating KWIC display): 0.0857 ms

KWIC – חיפוש מתקדם (#2 operators)

1. (||) = or (e.g., [hell~~l~~she]
= **he** or **she**)
2. (+) = one or more
characters (e.g.,
[he+] = **head** or **here**
but not **he**; [she+] =
sheriff or **shelves** but
not **she**)

The screenshot shows the AntConc software interface. On the left, a list of text files is displayed. The main window shows a KWIC search results table. The search query is 'he||she', and the results show 17 hits. A red circle with the number '1' highlights the search query field.

File	Left Context	Hit	Right Context
cv109_22599.txt	in a monty python sketch , and the headless horseman . . . well ,	he'	s a guy without a head . . . on a horse (
cv113_24354.txt	years , he's unwilling to marry her . his reasons ? well	he'	s a guy you see , and well , guys are
cv140_7963.txt	a time to kill ") but a lot of them suck .	he'	s a really eccentric person and i loved one
cv082_11080.txt	this was another of my fears for the film , but	he'	s a really good dancer , and plays a mean
cv010_29063.txt	more complex than the usual child molesters seen in movies --	he'	s a beast , but ashamed of it .) I . i .
cv065_16909.txt	but in the end we're supposed to believe that	she'	s a better person than that . we're supposed
cv168_7435.txt	the movie's main characters . " he's a prick " or "	she'	s a bitch " is no defense . the movie is
cv184_2673.txt	use some highlights . they both do fine with their roles :	he'	s a bored columnist with the chicago chronicle and
cv071_12969.txt	though the world's ending in a matter of months ,	she'	s a cable news personality . so there she is ,
cv067_19774.txt	performances , and their dynamic is manifest . willis convinces us that	he'	s a doctor , something he was unable to do
cv028_26964.txt	almost single-handedly foisted on a wisecracking macaw who thinks	he'	s a dog , voiced by eric idle) , its rabid
cv066_11668.txt	ever . jackie chan plays , who else , jackie . this time out ,	he'	s a famous chef who hosts an australian cooking
cv140_7963.txt	been with bruce all his life . he's a god .	he'	s a father figure . he's also dying . this
cv140_7963.txt	t know , who has been with bruce all his life .	he'	s a god . he's a father figure . he'
cv073_21785.txt	christina ricci , who plays the love interest in the film .	she'	s a good witch too , how cute . one of
cv023_13847.txt	just because this is a brian depalma film , and since	he'	s a great director and one who's films
cv146_19587.txt	de bont is a plagiarist , i mean to suggest that	he'	s a hack , having found no new ways to

Search Query ☒ Words ☐ Case ☐ Regex Results Set All hits Context Size 10 token(s)

1 he||she Start Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Progress 100%

Time taken (creating KWIC display): 0.2893 ms

צירופי שכנות (collocates #1)

COL·LO·CATE:

IN LINGUISTICS

(of a word) be habitually juxtaposed with another with a frequency greater than chance.

"“maiden” collocates with “voyage.”"

Have	Take	Come	Make
Have a bath	Take on	Come along	Make trouble
Have a drink	Take up	Come on	Make noise
Have a goodtime	Take a chance	Come up	Make a wish
Have a problem	Take a walk	Come by	Make a point
Have lunch	Take a break	Come over	Make sense

צירופי שכינות (collocates #2)

1. To generate collocates first go to the 'Collocate' tab
2. Type in 'm?n' in the search query and hit 'start'. Try generating another collocate for 'wom?n'

AntConc

File Edit Settings Help

Target Corpus

Name: Negative+positiv

Files: 400

Tokens: 255631

Collocate Types 15 Collocate Tokens 100 hits

Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1 who	1	88	17	71	69	93.335	1.781
2 a	2	296	196	100	136	75.830	0.781
3 young	3	25	24	1	21	48.128	2.584
4 old	4	23	22	1	14	44.179	2.581
5 black	5	14	9	5	10	30.716	2.813
6 ho	6	8	5	3	1	27.935	3.847
7 grumpy	7	5	5	0	2	24.830	4.947
8 suri	8	3	1	2	1	23.825	7.017
9 importance	9	4	1	3	2	22.555	5.432
10 mano	9	4	0	4	1	22.555	5.432
11 whose	11	11	0	11	10	21.440	2.606
12 group	12	10	7	3	7	20.704	2.710
13 jarmusch	13	3	2	1	1	19.401	6.017
14 grumpier	13	3	2	1	2	19.401	6.017
15 grumpiest	13	3	2	1	1	19.401	6.017

Search Query ☒ Words ☐ Case ☐ Regex Window Span From 8L To 8R Min. Freq 1 Min. Range 1

☒ Start ☐ Adv Search

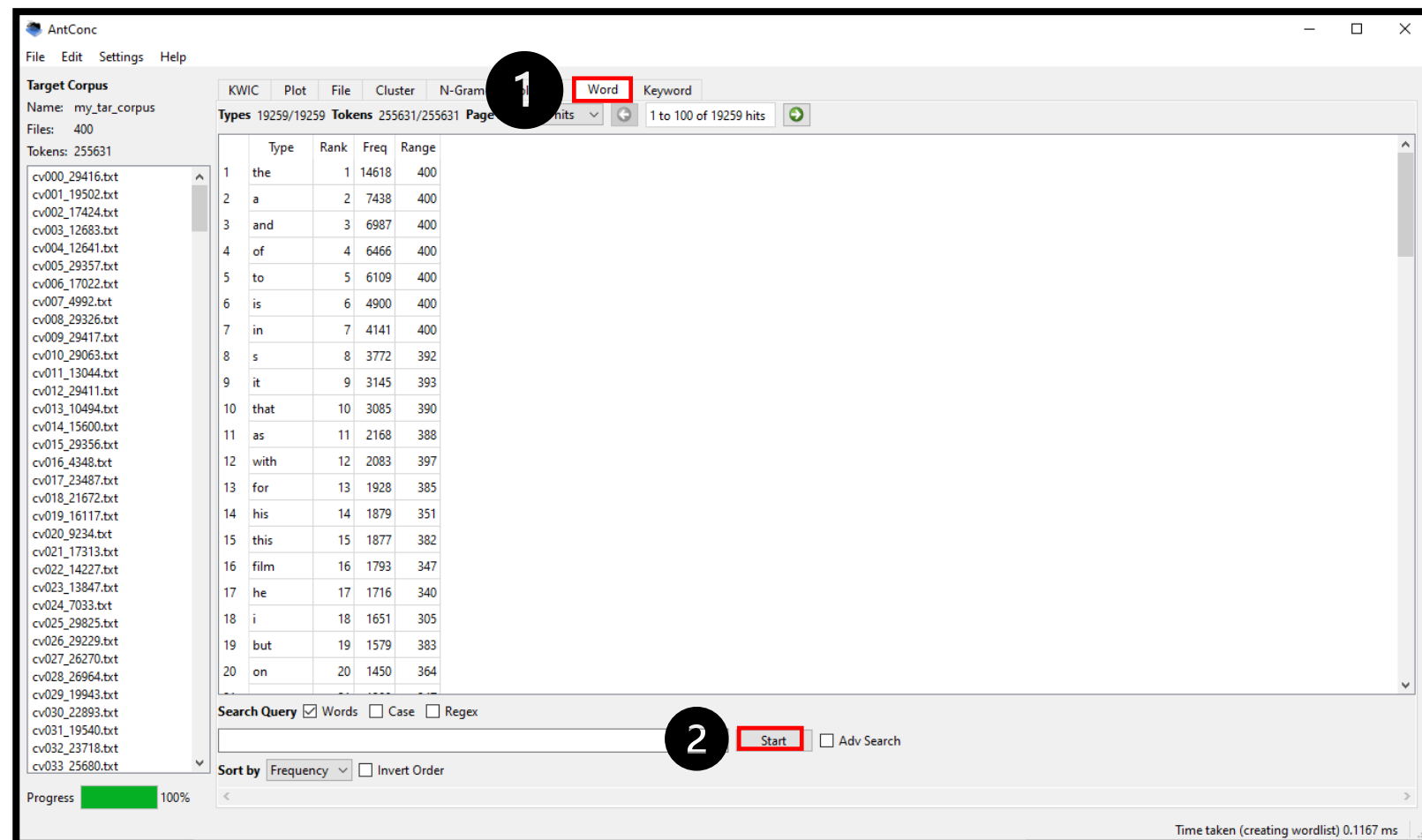
Sort by Likelihood ☐ Invert Order

Progress 100%

Time taken (creating KWIC display): 0.0298 ms

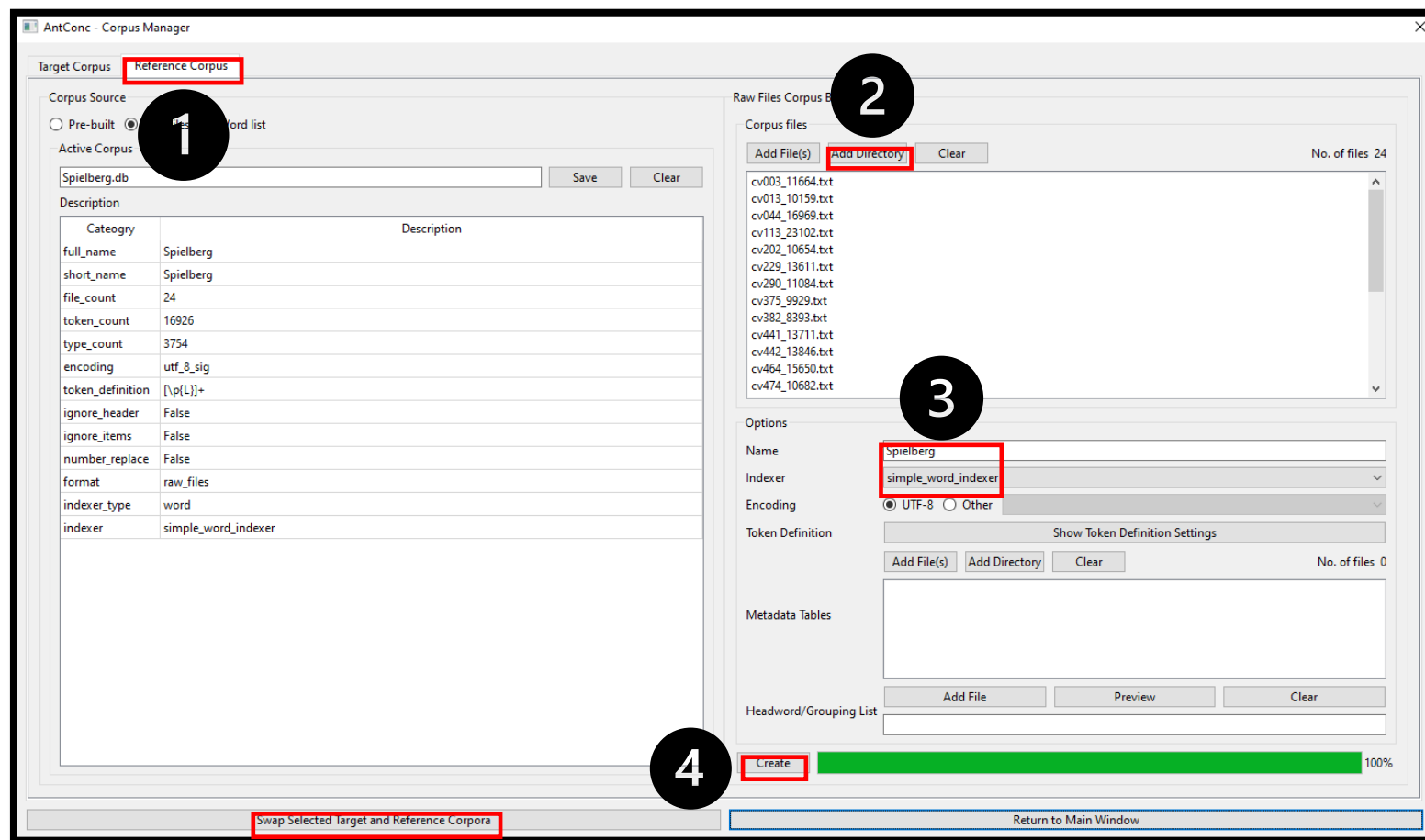
רשימת מילים (word)

1. To generate a wordlist go to the 'Word' tab
2. Hit 'start'



השוואה בין קורפוסים

1. Open the 'Corpus Manager' and hit the 'Reference Corpus' tab
2. Select 'Raw files' (clear files if necessary) and add the reviews from our 'Spielberg' directory
3. Name the corpus accordingly and select the 'Indexer'
4. Hit 'Create' and then 'Swap Selected Target and Reference Corpora'



השוואה בין קורפוסים #2

1. Go to the 'Keyword' tab
2. Keep the search box empty and hit 'Start'
3. Look at the 'Keyness (Likelihood)' column. What do you think it reflects?

Keyness: the statistical significance of a keyword's frequency in a given corpus, relative to a reference corpus

AntConc

File Edit Settings Help

Target Corpus
Name: Spielberg
Files: 24
Tokens: 16926

Reference Corpus
Name: Negative+positiv
Files: 400
Tokens: 255631

Keyword Types: 58/3754 Keyword Tokens: 2575/16926

1 Keyword

1 to 58 of 58 hits

Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1 spielberg	1	95	22	23	11	418.210	0.011
2 ryan	2	58	34	8	19	205.691	0.007
3 amistad	3	32	0	6	0	177.913	0.004
4 private	4	38	29	9	23	123.325	0.004
5 war	5	48	77	11	34	110.262	0.006
6 africans	6	19	0	4	0	105.622	0.002
7 saving	7	26	15	8	15	92.616	0.003
8 cinque	8	19	5	5	1	81.699	0.002
9 schindler	8	19	5	6	5	81.699	0.002
10 slavery	10	17	5	3	3	71.560	0.002
11 hanks	11	21	15	3	4	69.762	0.002
12 soldiers	12	17	7	3	3	66.424	0.002
13 battle	13	24	33	8	26	60.058	0.003
14 miller	14	15	10	5	7	51.013	0.002
15 slave	15	15	13	5	9	46.375	0.002
16 hounsou	16	10	2	5	1	45.028	0.001
17 list	17	17	21	7	19	44.935	0.002
18 buren	18	8	0	4	0	44.467	0.001
19 legal	18	8	0	4	0	44.467	0.001
20 men	20	32	103	11	65	43.235	0.004

3

Search Query ☒ Words ☐ Case ☐ Regex

2 Start

Sort by Likelihood ☐ Invert Order

Progress 100%

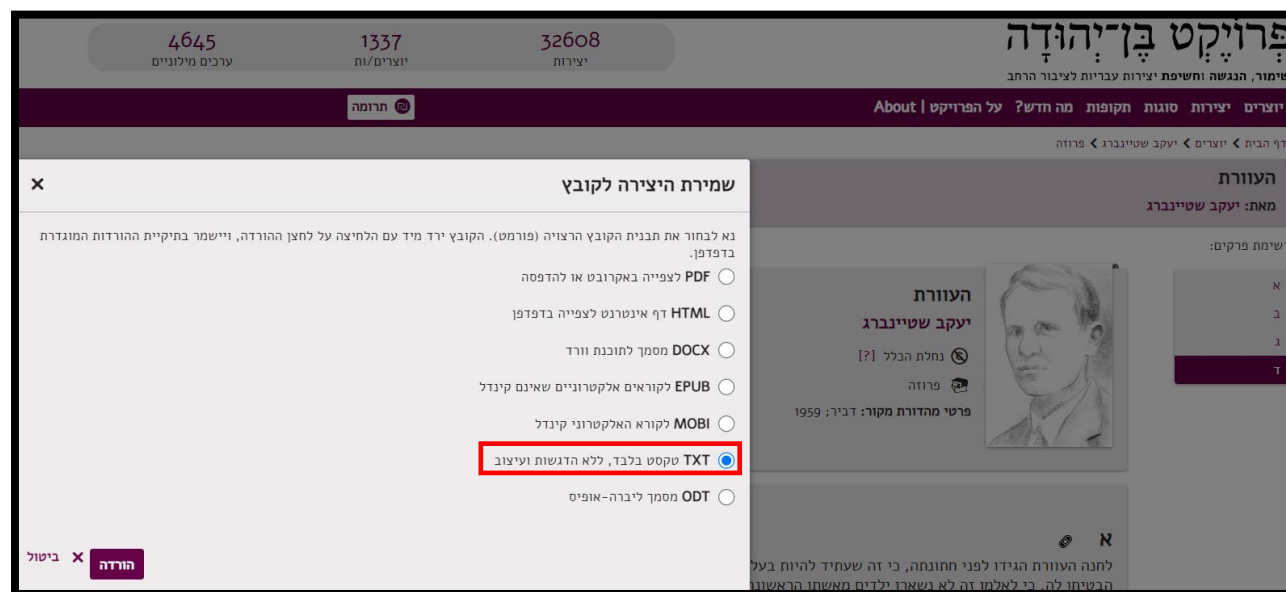
Time taken (creating keyword list) 0.1856 ms



- CATMA (Computer Aided Textual Markup and Analysis) היא תוכנה חופשית (GPL 3.0) לניתוח אוטומטי ממוחשב ותיוג חופשי של טקסטים מאוניברסיטת המבורג, גרמניה. תומכת (כמעט) בכל השפות.
- מבוססת ווב ותומכת ברוב סוגי הדפדפנים (Edge, Firefox, Chrome, Safari)
- גרסה אחרונה : 6.5.0 (2019)
- תומכת במגוון פורמטים של קבצי טקסט (doc, pdf, txt)
- כתובת לגישה (נדרש רישום מוקדם):
- <https://catma.de/>

CATMA - הכנת קורפוס

1. הכנסו אל פרויקט בן יהודה בכתובת <https://benyehuda.org/>
2. חפשו את היצירה "העוורת" מאת יעקב שטיינברג
3. שימרו את היצירה במחשבכם האישי כקובץ TXT. נקו את הקורפוס ושמרו בשנית



CATMA - מסך ראשי

CATMA 6.5.1

All Projects

CREATE NEW PROJECT

JOIN PROJECT

Project

Tags

Annotate

Analyze

ניהול פרויקטים

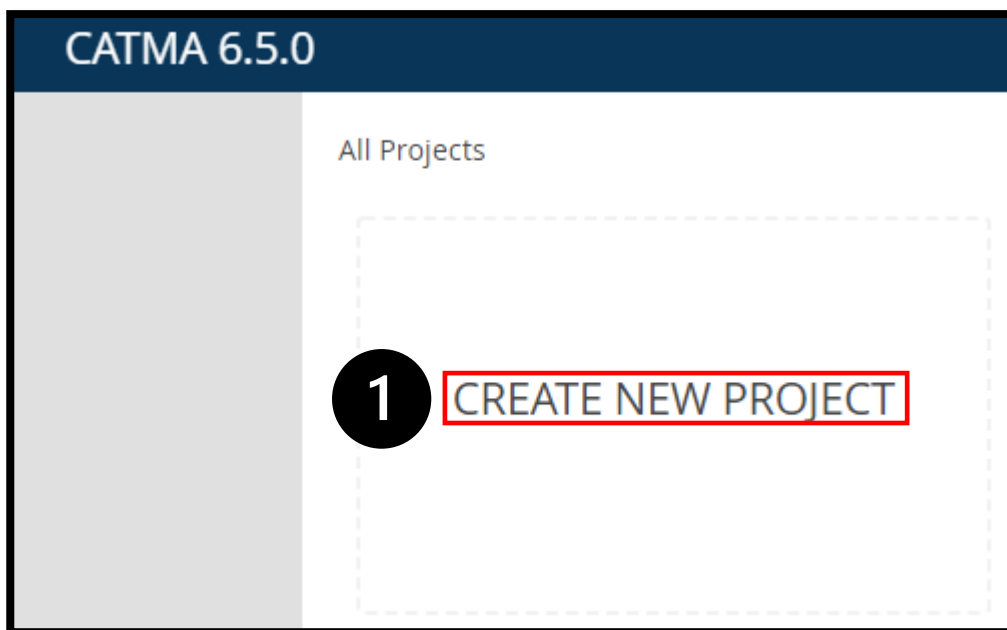
תגים

אנוטציה

ניתוח אוטומטי/ויזואליזציה

CATMA - יצירת פרויקט חדש

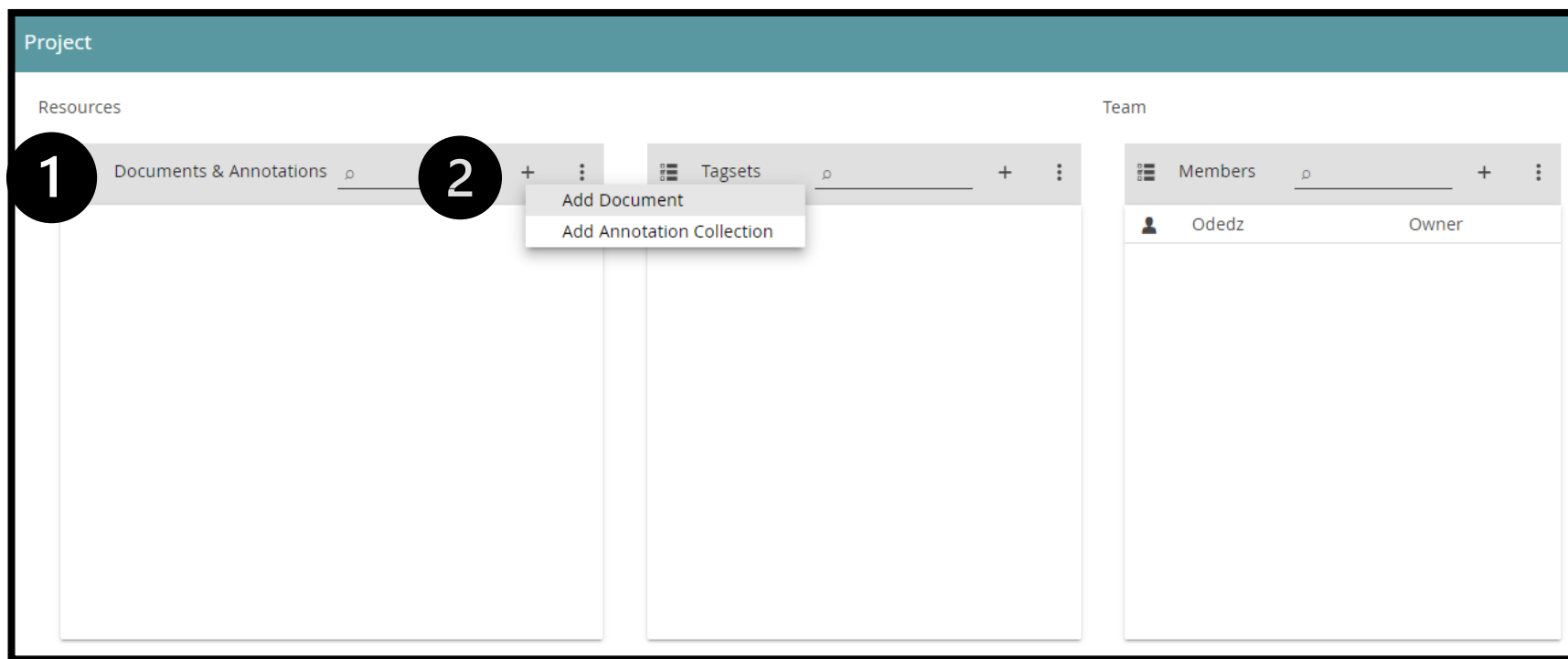
1. Log in to <https://catma.de/> and hit 'Create New Project'
2. Name your project and click 'OK'



The screenshot shows a form titled 'Create a new Project'. It has two main input fields: 'Name' and 'Description'. The 'Name' field contains the text 'העוורת' (He'evort), which is highlighted with a red rectangular border. To the right of this field, there is a black circle containing the white number '2'. The 'Description' field is empty. At the bottom right of the form, there are two buttons: 'OK' and 'CANCEL'.

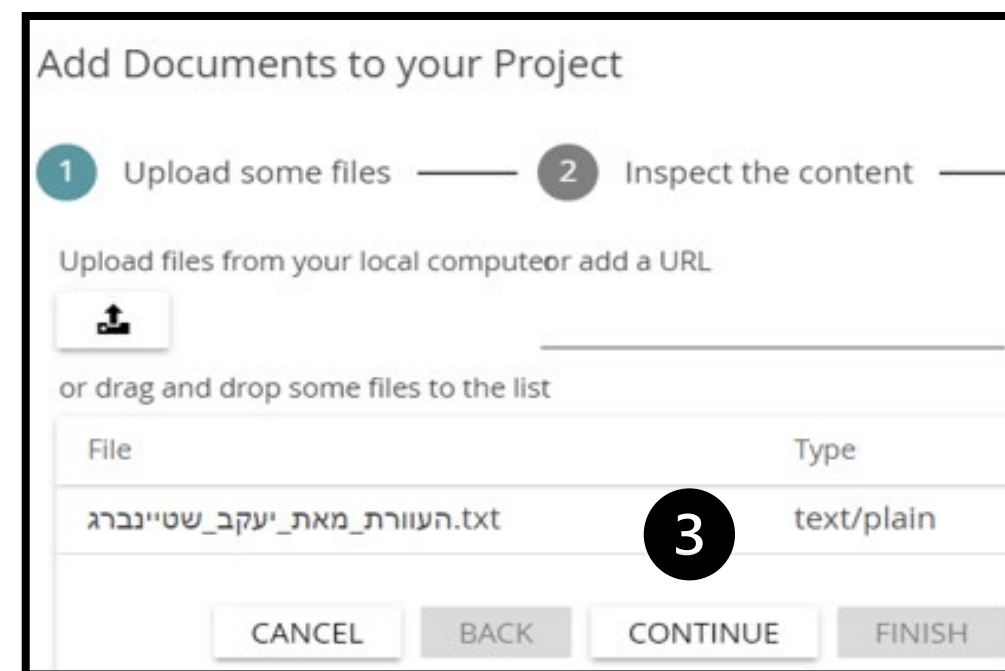
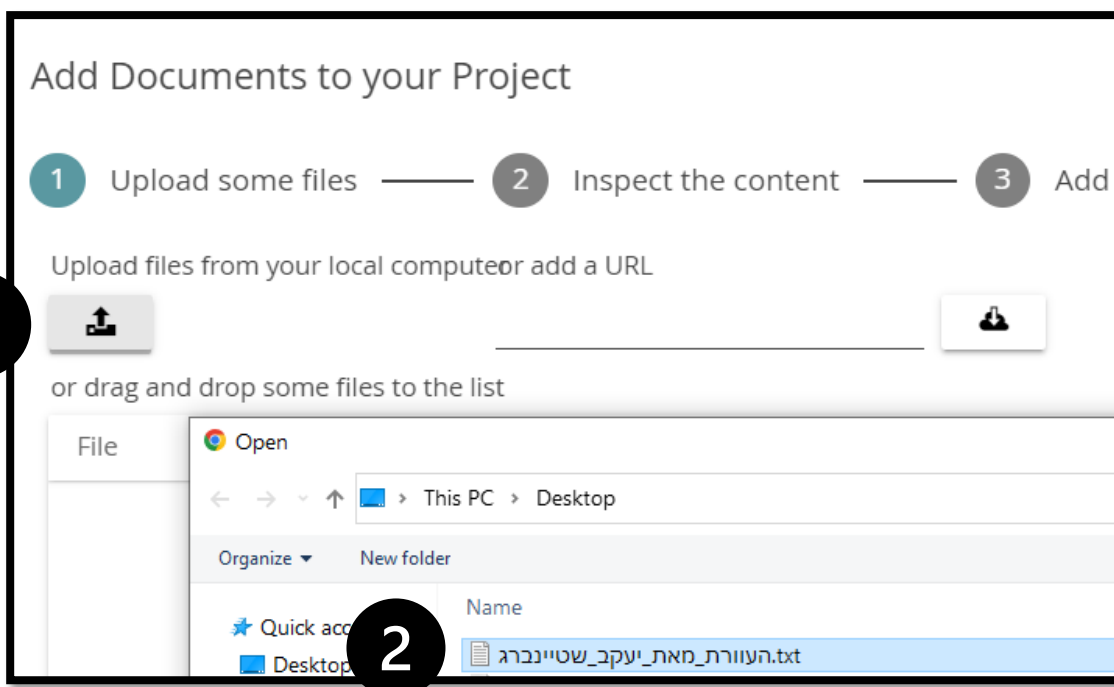
CATMA - העלאת קורפוס #1

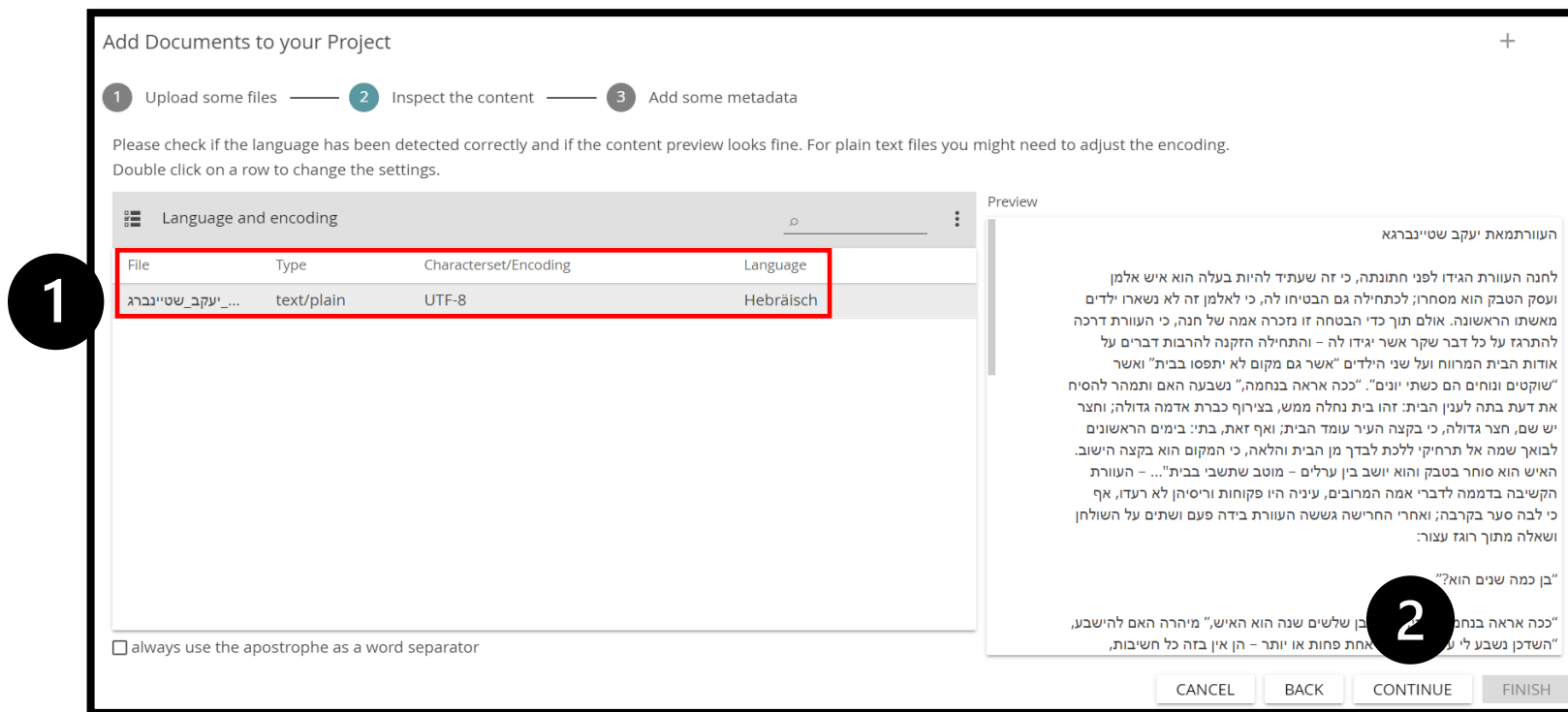
1. Open your project and go to 'Documents & Annotations'
2. Click on the '+' symbol and hit 'Add Document'



CATMA - העלאת קורפוס #2

1. Click on the icon titled 'Upload files from your local computer'
2. Select your corpus files and hit 'Open'
3. Click 'Continue'





CATMA – העלאת קורפוס #4

1. Add the relevant metadata for your files (author, publisher, description)
2. Hit 'Continue' then 'Finish'

Add Documents to your Project

3 Add some metadata 4 Create Annotation Collections

Add some metadata by doubleclicking on a row

File	Type	Title	Author	Publisher	Description
יעקב_שטיינברג_...	text/plain	העוורת מאת יעקב שטיינברג	יעקב שטיינברג	פרויקט בן יהודה	סיפור קצר

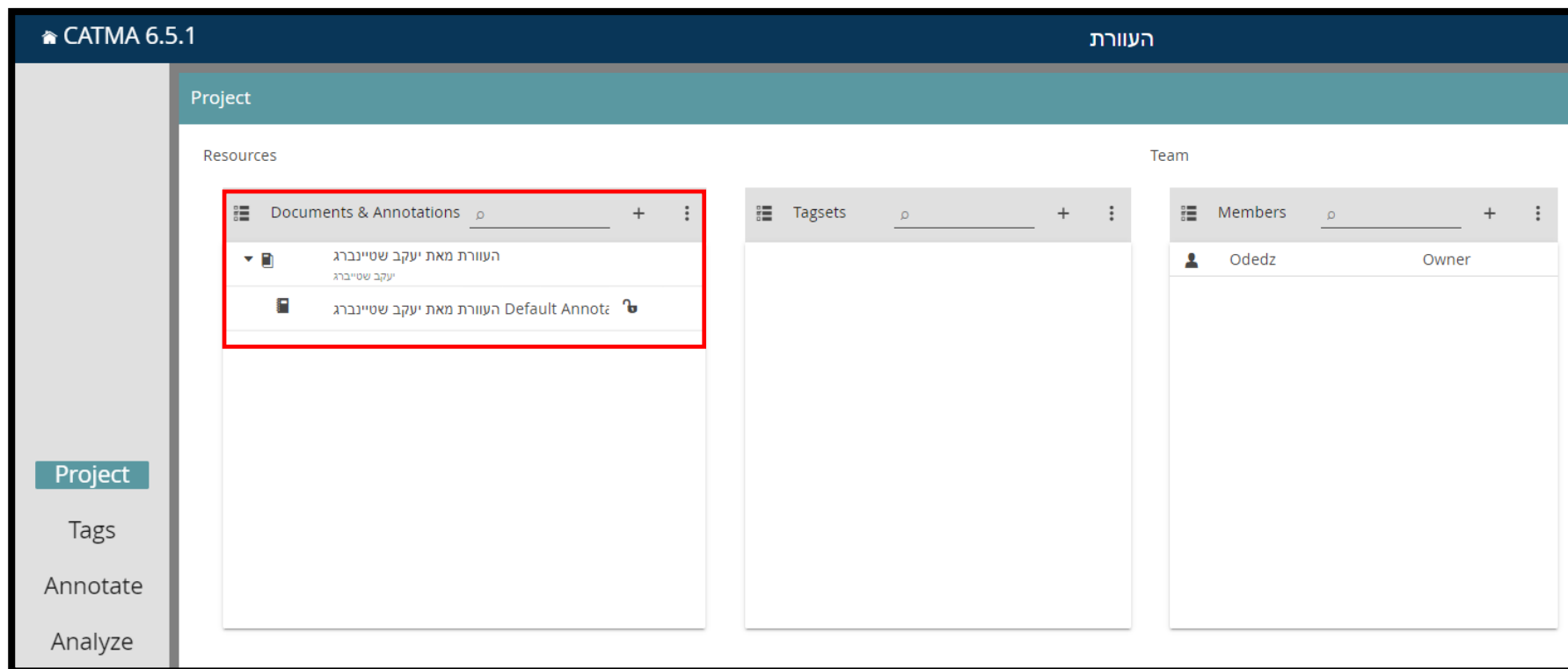
CANCEL BACK CONTINUE FINISH

1

2

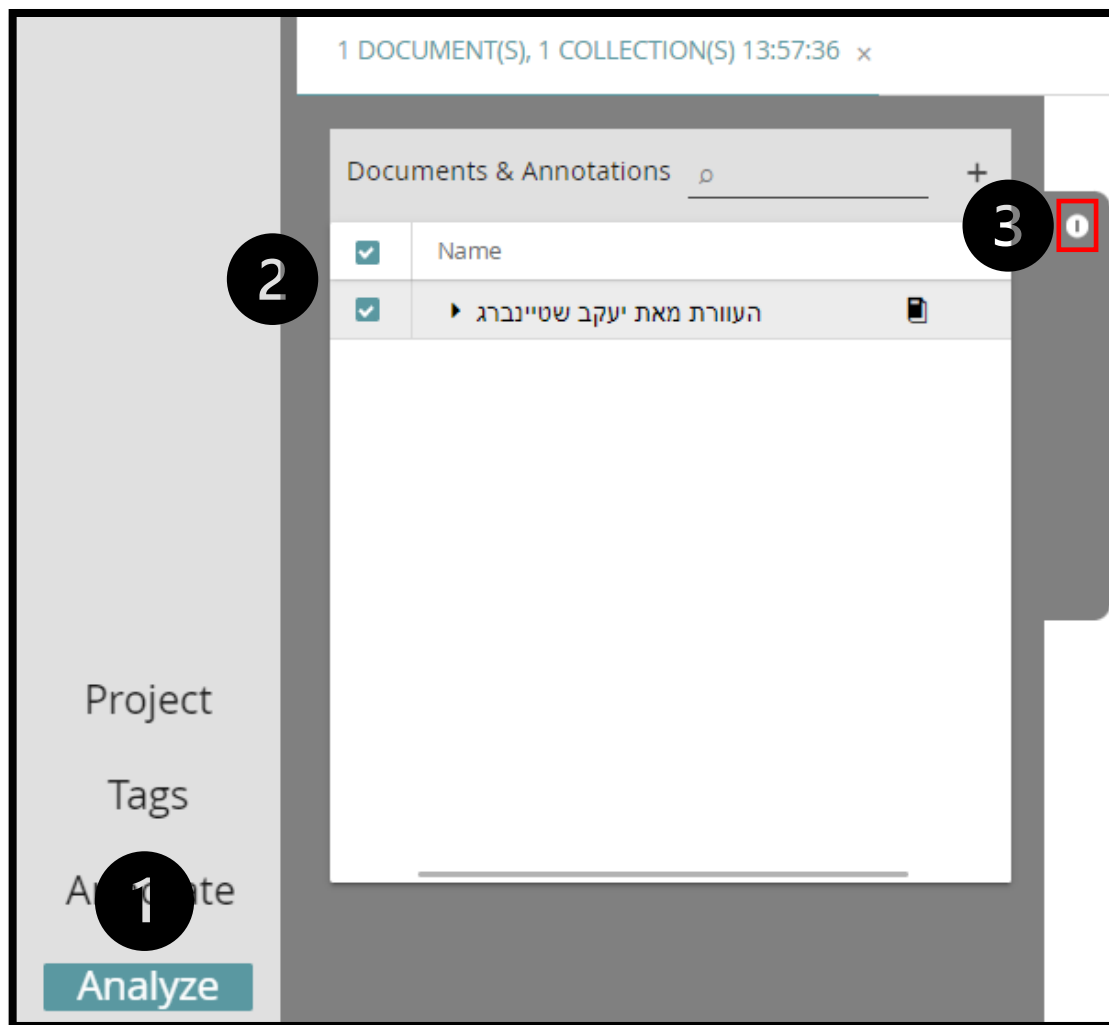
CATMA – העלאת קורפוס #5

1. Once processed, CATMA should display your files under 'Documents & Annotations'



CATMA – ניתוח אוטומטי

1. To access CATMA's automatic analysis tools first hit 'Analyze'
2. Check the files you want to analyze
3. Hit the 'I' symbol



CATMA – ניתוח אוטומטי (wordlist)

1. To generate a vocabulary, choose 'Wordlist (freq>0)' under 'Queries'
2. CATMA should now display the wordlist of your corpus with frequency
3. Clicking on the arrow located before each word will display its context

The screenshot shows the CATMA interface with a query 'Wordlist (freq>0)' selected. The results table displays the following data:

Phrase	Frequency
את	88
כי	63
העוורת	62
העוורת מאת יעקב שטיינברג	62
ה את העולל. ארעוורתהזקנה חוששת שמא תחנק האם	1
בשנתה את העולל. ארעוורתהזקנה חוששת שמא תחנק האם	1
ה, כי היולדת אינה העוורתהזקנה, אשר התבוננה למעשה	1
את ראשה לתוך חלל הלילה העוורתהבית. רגע אחד מטה	1
ממעטת בדברים, ואולם מיום העוורת. גם עד כה היתה	1

Types: 1477 Tokens: 3243

CATMA – ניתוח אוטומטי (distribution)

1. To generate a distribution graph, choose 'DISTRIBUTION' under 'Visualizations'

The screenshot displays the CATMA web interface. On the left, the 'Queries' panel shows a search for 'Wordlist (freq>0)' with results for 'freq>0 11:13:50 (2417)'. A table lists phrases and their frequencies. On the right, the 'Visualizations' panel contains four icons: KWIC, DISTRIBUTION (highlighted with a red box and a large black circle with the number 1), WORDCLOUD, and DOUBLETREE.

Phrase	Frequency
של	34
את	33
אלי	27
לא	26
על	24
הילדים	21
הילד	18
היה	17
היו	15

Types: 1524
Tokens: 2417

CATMA – ניתוח אוטומטי (distribution #2)

1. Select a phrase from the wordlist by clicking the black arrow under the 'Select' column



CATMA – ניתוח אוטומטי (wordcloud #1)

1. To generate a wordcloud graph, choose 'WORDCLOUD' under 'Visualizations'

The screenshot displays the CATMA interface. On the left, the 'Queries' panel shows a search for 'Wordlist (freq>0)' with results for 'freq>0 11:13:50 (2417)'. A table lists phrases and their frequencies:

Phrase	Frequency
של	34
את	33
אלי	27
לא	26
על	24
הילדים	21
הילד	18
היה	17
היו	15

Summary statistics: Types: 1524, Tokens: 2417.

On the right, the 'Visualizations' panel shows four options: KWIC, DISTRIBUTION, WORDCLOUD, and DOUBLETREE. The 'WORDCLOUD' option is highlighted with a red box and a large black circle with the number '1' below it, indicating the first step in the process.

CATMA – ניתוח אוטומטי (wordcloud #2)

1. Select all phrases from the wordlist by hitting the three dots of the query result

Wordcloud 14:52:17

העוורת מאת יעקב שטיינברג

Phrase	Frequency	Select
את	88	⋮
כי	63	⋮
העוורת	62	⋮
על	50	⋮
היא	38	⋮
בעלה	36	⋮
לא	35	⋮
לה	33	⋮

Types: 1477 Tokens: 3243

1

Group by Phrase
Group by Tag Path
Display Annotations as flat table
Display Properties as columns
Export
✓ Filter punctuation
Select all

העוורת
היא
אשר
כל
היתה
כבר
לכה
שומעת
אין

No. of Types 158
Tile size 9.2
Fontsize min 7.3
Fontsize max 33
Word padding 2.7

Change number of words, size of wordcloud, distance between words

CATMA – ניתוח אוטומטי (doubletree)

1. To generate a doubletree graph, choose 'DOUBLETREE' under 'Visualizations'

The screenshot displays the CATMA web interface. On the left, the 'Queries' panel shows a search for 'Wordlist (freq>0)' with results for 'freq>0 11:13:50 (2417)'. A table lists phrases and their frequencies:

Phrase	Frequency
של	34
את	33
אלי	27
לא	26
על	24
הילדים	21
הילד	18
היה	17
היו	15

Summary statistics: Types: 1524, Tokens: 2417.

On the right, the 'Visualizations' panel shows four options: KWIC, DISTRIBUTION, WORDCLOUD, and DOUBLETREE. The DOUBLETREE option is highlighted with a red box and a large black circle containing the number 1, indicating the correct choice for generating a doubletree graph.

CATMA – ניתוח אוטומטי (doubletree #2)

1. Select a phrase from the wordlist by clicking the black arrow under the 'Select' column

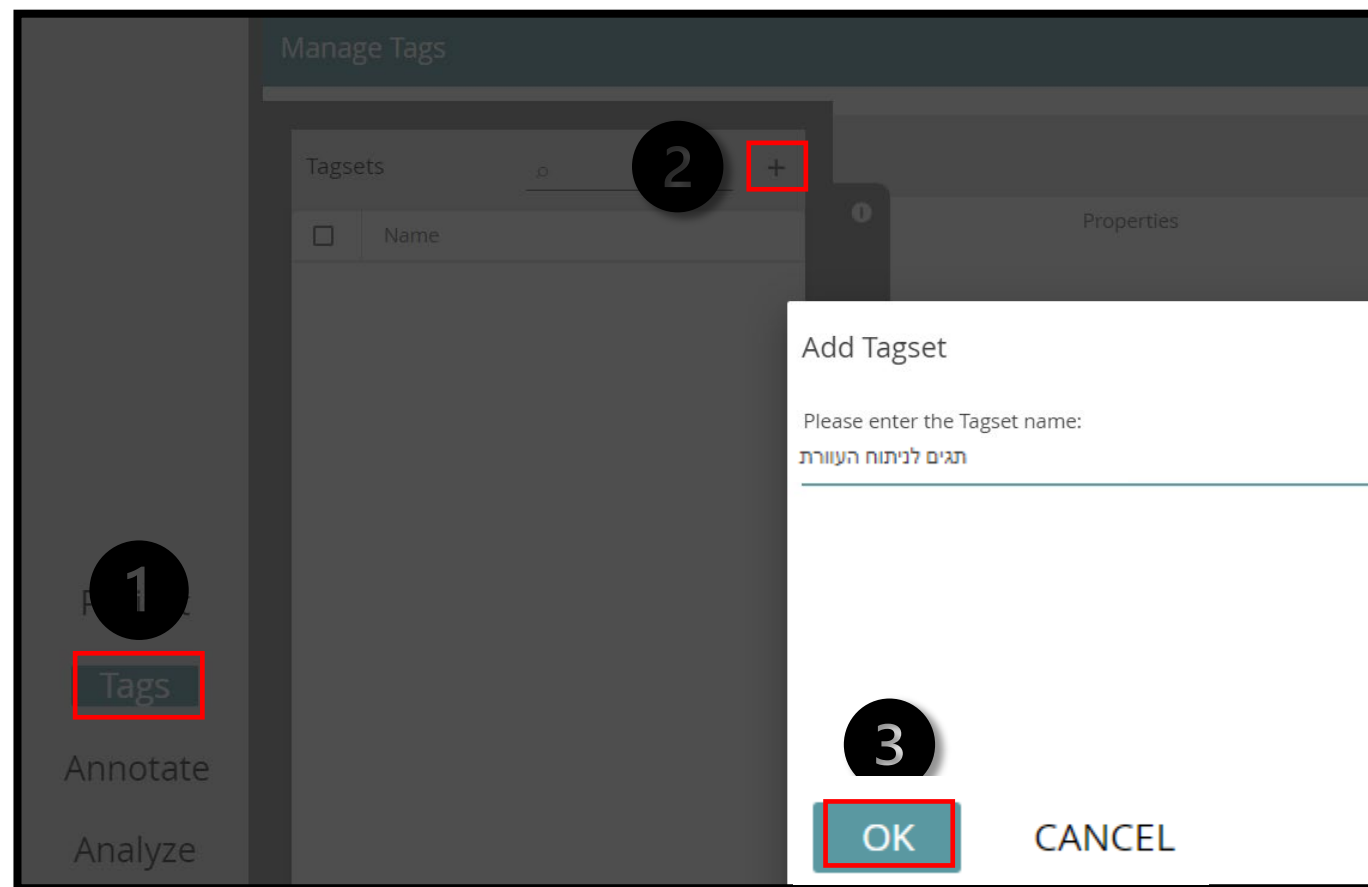
The screenshot displays the DoubleTreeJS interface. On the left, a table lists phrases and their frequencies. A red box highlights the 'Select' column for the phrase 'בעלה' (frequency 36), with a black circle containing the number '1' next to it. On the right, a dependency graph shows the phrase 'בעלה' at the center, with various words connected to it by lines. A red arrow points to the word 'כ' in the graph. Below the graph, a text box explains that clicking on any word in the graph will extend the context for the rest of the sentence and highlight possible combinations.

Phrase	Frequency	Select
על	50	⬇
היא	38	⬇
בעלה	36	⬇
לא	35	⬇
לה	33	⬇
הוא	32	⬇
אשר	32	⬇
של	29	⬇
Types: 1477	Tokens: 3243	

Clicking on any word in the graph will extend the context for the rest of the sentence and highlight possible combinations

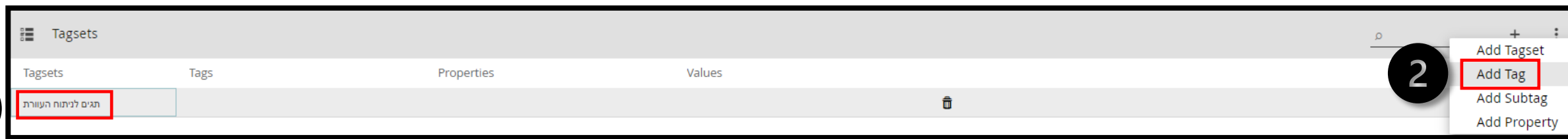
CATMA – תיוג (יצירת ספריית תגים)

1. Hit 'Tags' in the main menu
2. Click on the '+' sign in the 'Tagsets' window
3. Name your tagset 'תגים' 'לניתוח העוורת' and hit 'OK'



CATMA – תיוג (יצירת תגים #1)

1. Select you tagset by clicking on it
2. Hit the '+' sign and select 'Add Tag'





CATMA – תיוג (יצירת תגים #2)

1. Name your tag 'פעולות רצוניות'
2. Hit 'ADD TAG'

Add Tag

Tagset

תגים לניתוח העוורת

Enter name, color and Properties (optional) of the Tag you want to add

פעולות רצוניות

1

#35072A

Add Properties

Enter a Property name

ADD PROPERTY

2

ADD TAG

CANCEL

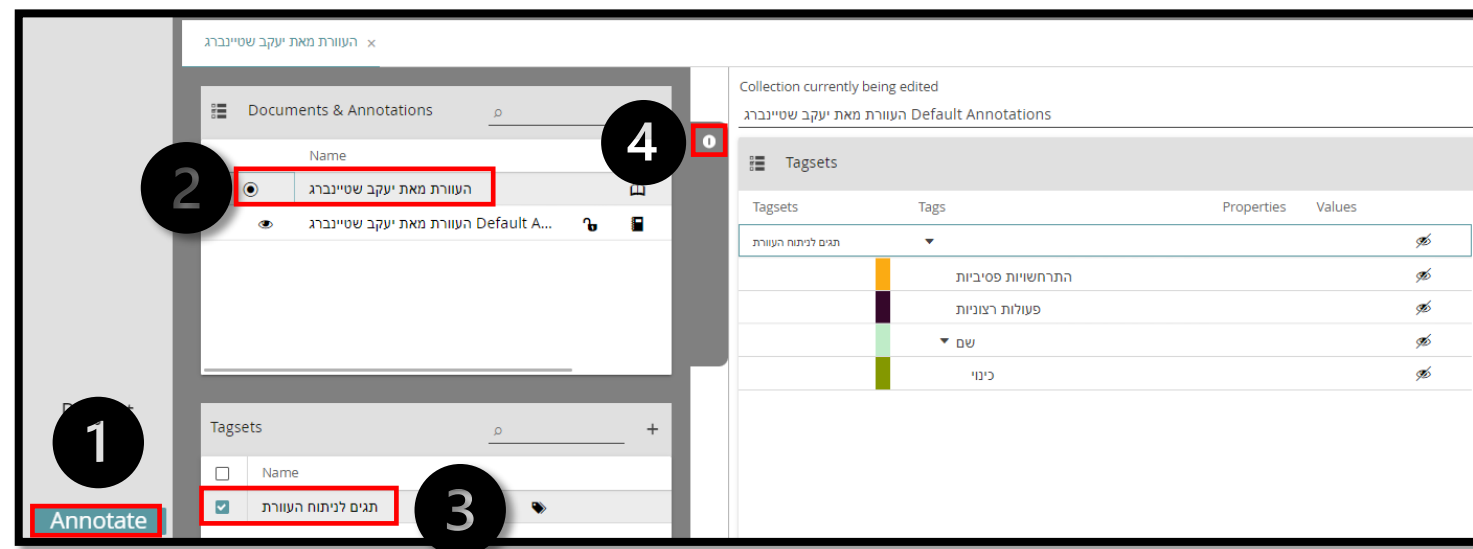
CATMA – תיוג (יצירת תגים #3)

1. Add the following tags to your tagsets:

Tagsets				
Tagsets	Tags	Properties	Values	
תגים לניתוח העוורת	▼			🗑️
	פעולות רצוניות			🗑️
	התרחשויות פסיביות			🗑️
	שם ▼			🗑️
	כינוי			🗑️

CATMA – אנוטציה #1

1. To start annotating hit ‘Annotate’ in the left menu
2. Select your text in ‘Documents & Annotations’
3. Load your tagsets by checking them under ‘Tagsets’
4. Once finished, hit the ‘|’ button to view your text



CATMA – אנוטציה #2

1. To annotate, mark a word and select the tag you wish to apply

העיוור ממת יעקב שטיינברג

Collection currently being edited
Default Annotations העיוור ממת יעקב שטיינברג

Tagsets

Tagsets	Tags	Properties	Values
תגים לניתוח העיוור	▼		✕
	התרשויות פסיביות		✕
	פעולות רצויות		✕
	שם ▼		✕
	כינוי		✕

לחנה העיוור הגידו לפני חתונתה, כי זה שעתידי להיות בעלה הוא איש אלמן
ועסק הטבק הוא מסחרו; לכתחילה גם הבטיחו לה, כי לאלמן זה לא נשארו
ילדים
מאשתו הראשונה. אולם תוך כדי הבטחה זו מזכרה אמה של חנה, כי העיוור
דרכה
להתרגז על כל דבר שקר אשר יאידו לה – והתחילה הזקנה להרבות דברים על
אודות הבית המרווח ועל שני הילדים "אשר גם מקום לא יתפסו בבית" ואשר
"שוקטים ומנוחים הם כשתי יונים". "ככה אראה בנחמה", נשבעה האם ותמרה
להסית
את דעות בתה לענין הבית: זהו בית נחלה ממש, בעירוף כבוד אדמה גדולה;
וחצר
יש שם, חצר גדולה, כי בקצה העיר עומד הבית, ואף זאת, בתי בימים
הראשונים
לבואך שמה אל תרחיקי ללכת לבדך מן הבית והלאה, כי המקום הוא בקצה
הישוב
האיש הוא סוחר בטבק והוא יושב בין ערלים – מוטב שתשבי בבית" ...
העיוור
הקשיבה בדממה לדברי אמה המרובים, עיניה היו פקוחות וריסיהן לא רעדו,
אף
כי לבה סער בקרבה; ואחרי החרישה גששה העיוור בידה פעם ושתים על
השולחן
ושאלה מתוך רוגז עצור:

CATMA – אנוטציה #3

1. To generate a taglist go to 'Analyze'
2. Verify that your annotation collection is marked
3. Under 'Queries' select 'Taglist (tag="")'

The screenshot shows the CATMA web interface. On the left sidebar, the 'Analyze' button is highlighted. The main panel displays the 'Queries' section with a 'BUILT QUERY' button and a search bar containing 'Taglist (tag="")'. Below this, a table shows the results of the query for the tag 'tag=""'.

Phrase	Frequency
נזכרה	1
העוורת מאת יעקב שטיינברג	1
ושאלה	1
הבסיוח	1
לבה סער בקרבה	1
הגידו	1
ככה אראה בנחמה	1
הקשיבה	1

CATMA – אנוטציה #4

1. You can also group the list by tag names by hitting ‘Group by Tag Path’
2. Try to generate a distribution graph for your tags

The screenshot shows the CATMA interface with a query titled "Taglist (tag='%')". The search bar contains "tag='%'" and the results show 10 annotations. A table displays the tag paths and their frequencies:

Tag Path	Frequency
▶ התרחשויות פסיביות /	8
▶ פעולות רצופיות /	2

A context menu is open on the right, with the option "Group by Tag Path" highlighted in red. Other options include "Group by Phrase", "Display Annotations as flat table", "Display Properties as columns", "Export", and "Filter punctuation" (which is checked).

At the bottom of the interface, it shows "Tags: 2" and "Annotations: 10".

CATMA – אנוטציה #5

1. You can also annotate all occurrences of a selected word in the text by going to the KWIC tool, selecting a word and then 'Annotate selected rows'

1 DOCUMENT(S), 1 COLLECTION(S) 07:47:05 x

KWIC - KeyWord In Context 07:52:15

freq>0 07:52:07

Phrase	Frequency	Select
את	87	Ⓢ
כי	63	Ⓢ
העוורת	62	Ⓢ
על	50	Ⓢ
היא	38	Ⓢ
בעלה	36	Ⓢ
לא	35	Ⓢ
לה	33	Ⓢ
אשר	32	Ⓢ
Types: 1460 Tokens: 3223		

Keyword in context

	Document	Left Context	Keyword	Right Context		
☒	העוורת	הקשיבה בדממה לדברי א...	העוורת	שתשבי בבית" ... -		
☒	העוורת	מורידה עוד יותר את ראשה	העוורת	האיש הכבדה פסקה פתא...	15612	15618
☒	העוורת	ידעה זה כבר, כי	העוורת	העריסה והקשיבה לנשימי...	15252	15258
☒	העוורת	מדברת את דבריה נפל על...	העוורת	נשימתה כבדה. " ובעוד	14069	14075
☒	העוורת	, כי בעלה מונה כסף	העוורת	על פי צלצול המטבעות מכ...	9515	9521
☒	העוורת	עמדה בלי נוע, ורק	העוורת	" האם חדלה מדבר.	1067	1073
☒	העוורת	בצעקות, אך ליד העריסה	העוורת	. כל הלילה עובר על	15888	15894
☒	העוורת	יותר ויותר ברויסי.	העוורת	של בעלה נעשו הרהוריה ...	1655	1661
☒	העוורת	, שנעשתה דוממה ומקשי...	העוורת	" בנהימה נמלט האיש מידי	14909	14915
☒	העוורת	לעמוד ליד הנער ולא ניסתה	העוורת	. רגע אחרי רגע הוסיפה	6657	6663
☒	העוורת	לעצמה את דמות בעלה:	העוורת	אף. לפי המגפיים תיארה	2083	2089
☒	העוורת	, בשעה שקשה לה לתאר	העוורת	ביתנו? " - וכדרך	4940	4946
☒	העוורת	ולתנועותיה, כי היולדת אינה	העוורת	הזקנה, אשר התבוננה למ...	10963	10969
☒	העוורת	יושבת בראש מורד ושפתי...	העוורת	יבוא נא..." ובעוד	15413	15419

Annotate selected rows
Remove selected Annotations
Export

יעץ והדרכה בספרייה

cenlib@tauex.tau.ac.il 

03-6408423 

050-5075982 

תודה!

עודד זרחיה

odedzarchia@tauex.tau.ac.il