AMC-Xplorer

Forchini P. & Palmero Aprosio A.

HOW TO QUOTE:
Forchini, P. & Palmero Aprosio A. 2025. AMC-Xplorer (web-based interface).
https://amc-xplorer.americanmoviecorpus.net

The present guide provides a simplified overview of **AMC-Xplorer** and its **tools**. It also outlines various **search functions** and **regular expressions** with examples, serving as a practical reference for performing searches.

What is AMC-Xplorer?

AMC-Xplorer is a web-based interface developed by **Alessio Palmero Aprosio** (Università di Trento, Italy – contact: a.palmeroaprosio@unitn.it) and **Pierfranca Forchini** (Università Cattolica del Sacro Cuore, Milan, Italy – contact: pierfranca.forchini@unicatt.it). It enables the exploration of the **American Movie Corpus (AMC)**, which consists of authentic dialogs from American movies.

The interface is freely available for **research**, **scholarship, teaching**, **criticism**, and **comment**. If you use AMC-Xplorer, please cite it in your work. Additionally, we welcome bibliographies that reference AMC data, as they may be added to the official AMC website.

Getting Started: Movie or Character Selection

In order to conduct a search, **movies or characters must be selected first**. They can then be deselected for new searches.

Movies are listed alphabetically and can be selected either individually or all at once using the section shown in Figure 1. They can be analyzed using the KWIC, Plot, N-Gram, Collocate, and Word Frequency tools. **Characters** are listed in chronological order based on their first appearance in the movie. They can be selected individually or all at once using the section shown in Figure 2. Character data can be analyzed using the KWIC, N-Gram, Collocate, and Word Frequency tools, but not the Plot tool.

Figure 1:

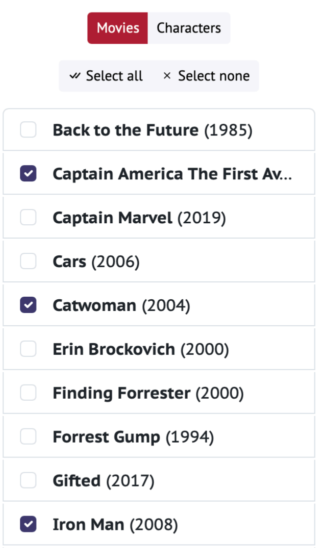
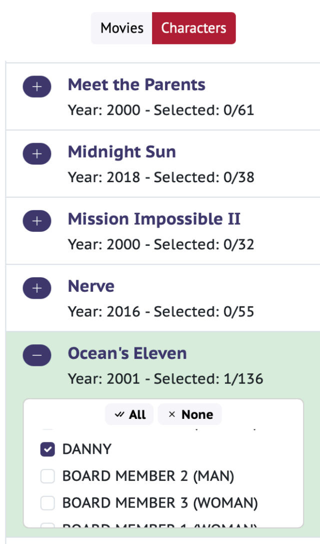


Figure 2:

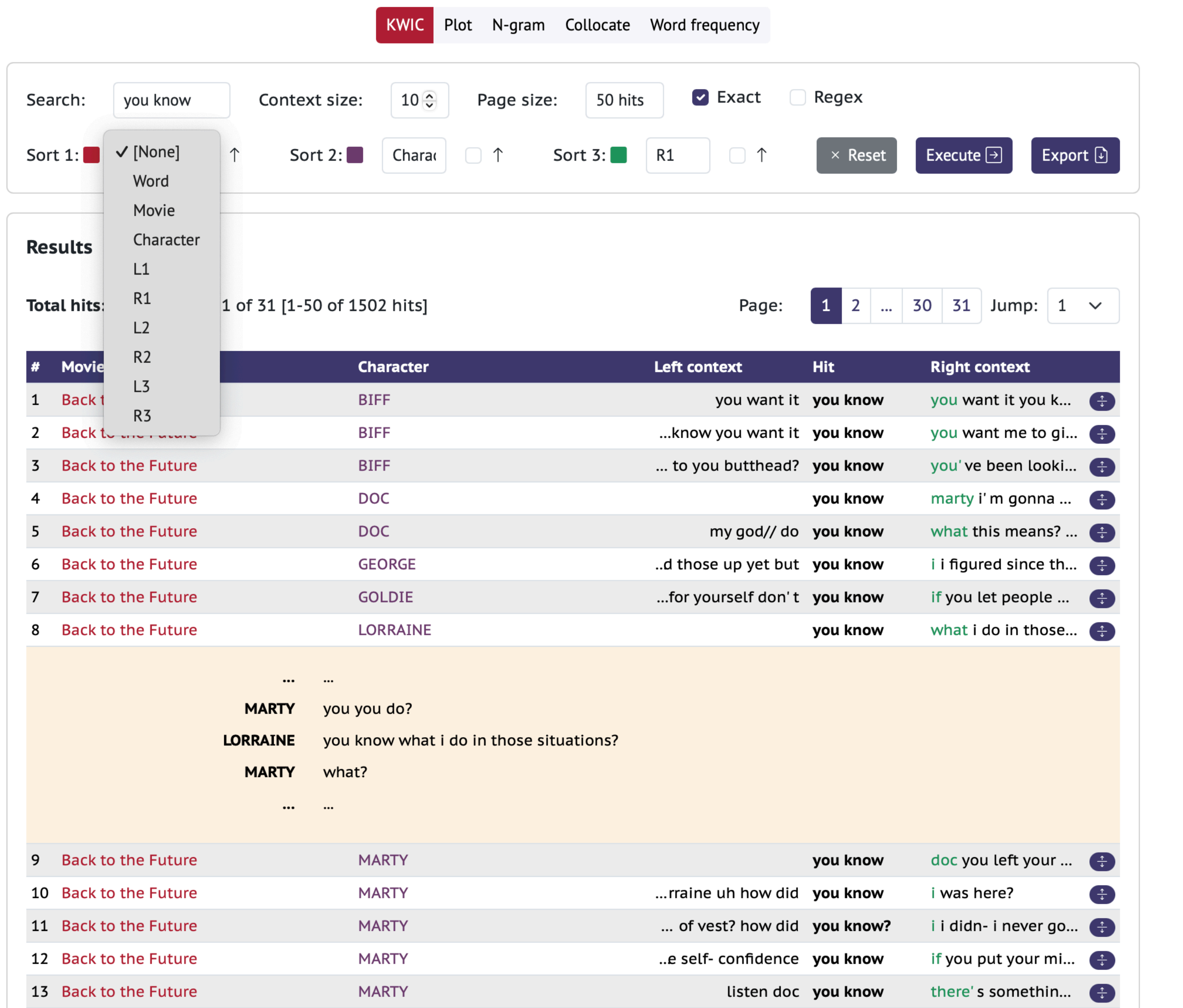


KWIC (Key-Word-In-Context) Tool

The KWIC tool displays search results in a concordance format, allowing users to examine how linguistic items are used in the AMC. The **Sort** options (**Sort 1, Sort 2, and Sort 3**) enable the rearrangement of concordance lines based on criteria such as movie, character, or the item's position (up to three words to the left or right).

The surrounding context of linguistic items can be expanded using the **Context Size** option or by clicking the **Expand** button, as shown in Figure 3.

Figure 3:



Plot Tool

The Plot Tool visualizes the positions where search results appear within individual movies in the AMC. By using the **Search** options (cf. **Search 1**, **Search 2**, and **Search 3** in Figure 4), users can compare up to three linguistic items to analyze their distribution and relationships across the AMC.

Figure 4:

KWIC | Plot | N-gram | Collocate | Word frequency

Search (1): i — Exact — Regex

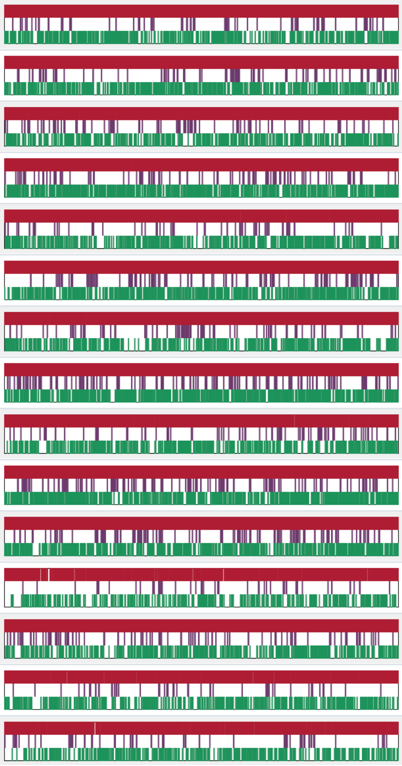
Search (2): ^(she|he)$ — Exact — Regex ☑

Search (3): you — Exact — Regex

× Reset | Execute | Export

Results

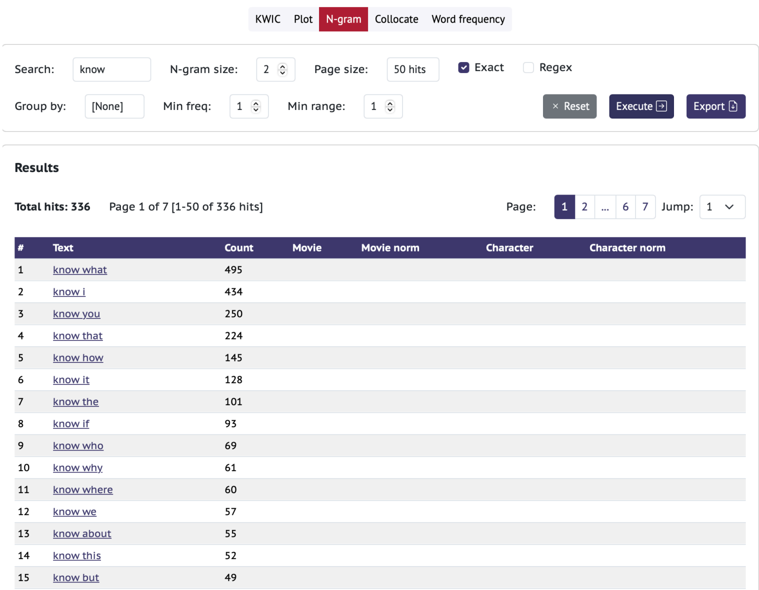
#	Movie	Tokens	Freq	Norm freq	Dispersion	Plot
1	Erin Brockovich	14459	4183	289.30	0.994	
2	Some like it Hot	13995	4034	288.25	0.991	
3	Me Myself & Irene	11716	3371	287.73	0.989	
4	Shallow Hal	12434	3851	309.72	0.989	
5	Back to the Future	9901	3114	314.51	0.988	
6	The Intern	15489	4649	300.15	0.987	
7	Ocean's Eleven	10336	3164	306.11	0.986	
8	Silver Linings Playbook	18346	5892	321.16	0.986	
9	Twilight	9198	2817	306.26	0.986	
10	The Judge	13516	4267	315.70	0.985	
11	The Social Network	18645	5788	310.43	0.985	
12	Catwoman	5894	1905	323.21	0.984	
13	The Devil Wears Prada	11599	3667	316.15	0.984	
14	Captain Marvel	7986	2353	294.64	0.983	
15	Midnight Sun	7736	2399	310.11	0.983	



N-Gram Tool

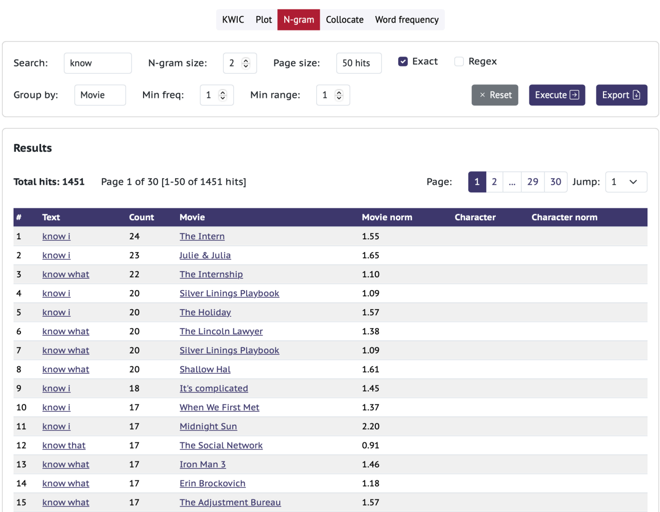
The N-Gram Tool identifies common multi-word expressions in the AMC. Results can be grouped by overall corpus frequency (cf. Figure 5) or filtered by movie (cf. Figure 6) or character (cf. Figure 7). The n-gram size can be expanded up to 20 words. Clicking on a result redirects to the KWIC tool, displaying the specific contexts in which the expression occurs.

Figure 5:



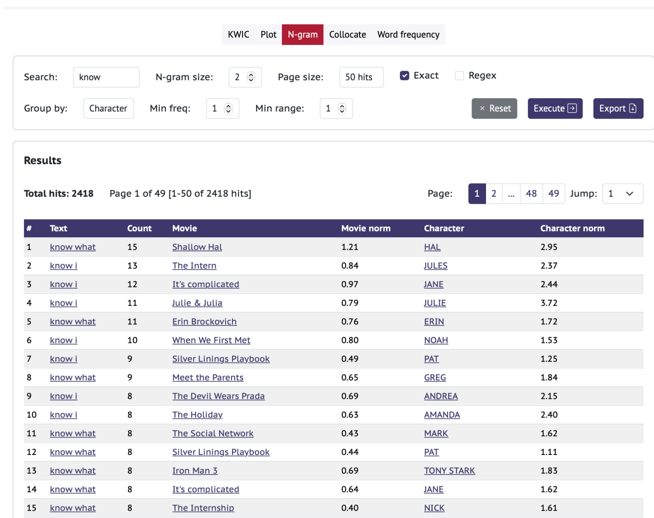
#	Text	Count	Movie	Movie norm	Character	Character norm
1	know what	495				
2	know i	434				
3	know you	250				
4	know that	224				
5	know how	145				
6	know it	128				
7	know the	101				
8	know if	93				
9	know who	69				
10	know why	61				
11	know where	60				
12	know we	57				
13	know about	55				
14	know this	52				
15	know but	49				

Figure 6:



#	Text	Count	Movie	Movie norm	Character	Character norm
1	know i	24	The Intern	1.55		
2	know i	23	Julie & Julia	1.65		
3	know what	22	The Internship	1.10		
4	know i	20	Silver Linings Playbook	1.09		
5	know i	20	The Holiday	1.57		
6	know what	20	The Lincoln Lawyer	1.38		
7	know what	20	Silver Linings Playbook	1.09		
8	know what	20	Shallow Hal	1.61		
9	know i	18	It's complicated	1.45		
10	know i	17	When We First Met	1.37		
11	know i	17	Midnight Sun	2.20		
12	know that	17	The Social Network	0.91		
13	know what	17	Iron Man 3	1.46		
14	know what	17	Erin Brockovich	1.18		
15	know what	17	The Adjustment Bureau	1.57		

Figure 7:



#	Text	Count	Movie	Movie norm	Character	Character norm
1	know what	15	Shallow Hal	1.21	HAL	2.95
2	know i	13	The Intern	0.84	JULES	2.37
3	know i	12	It's complicated	0.97	JANE	2.44
4	know i	11	Julie & Julia	0.79	JULIE	3.72
5	know what	11	Erin Brockovich	0.76	ERIN	1.72
6	know i	10	When We First Met	0.80	NOAH	1.53
7	know i	9	Silver Linings Playbook	0.49	PAT	1.25
8	know what	9	Meet the Parents	0.65	GREG	1.84
9	know i	8	The Devil Wears Prada	0.69	ANDREA	2.15
10	know i	8	The Holiday	0.63	AMANDA	2.40
11	know what	8	The Social Network	0.43	MARK	1.62
12	know what	8	Silver Linings Playbook	0.44	PAT	1.11
13	know what	8	Iron Man 3	0.69	TONY STARK	1.83
14	know what	8	It's complicated	0.64	JANE	1.62
15	know what	8	The Internship	0.40	NICK	1.61

Collocate Tool

The Collocate Tool identifies words that co-occur with a given search term within a specified distance in the AMC. The search range can be adjusted using the **L (Left)** and **R (Right)** options. In addition to raw frequency, five statistical measures are provided (cf. Figure 8):

1. **Log-likelihood (LL)** – Compares the observed co-occurrence of a word pair to its expected frequency (if the words were independent). It highlights frequent collocations but does not measure the strength of association. Unlike MI, it is not biased toward low-frequency words. Higher values indicate stronger evidence that co-occurrence is not random.

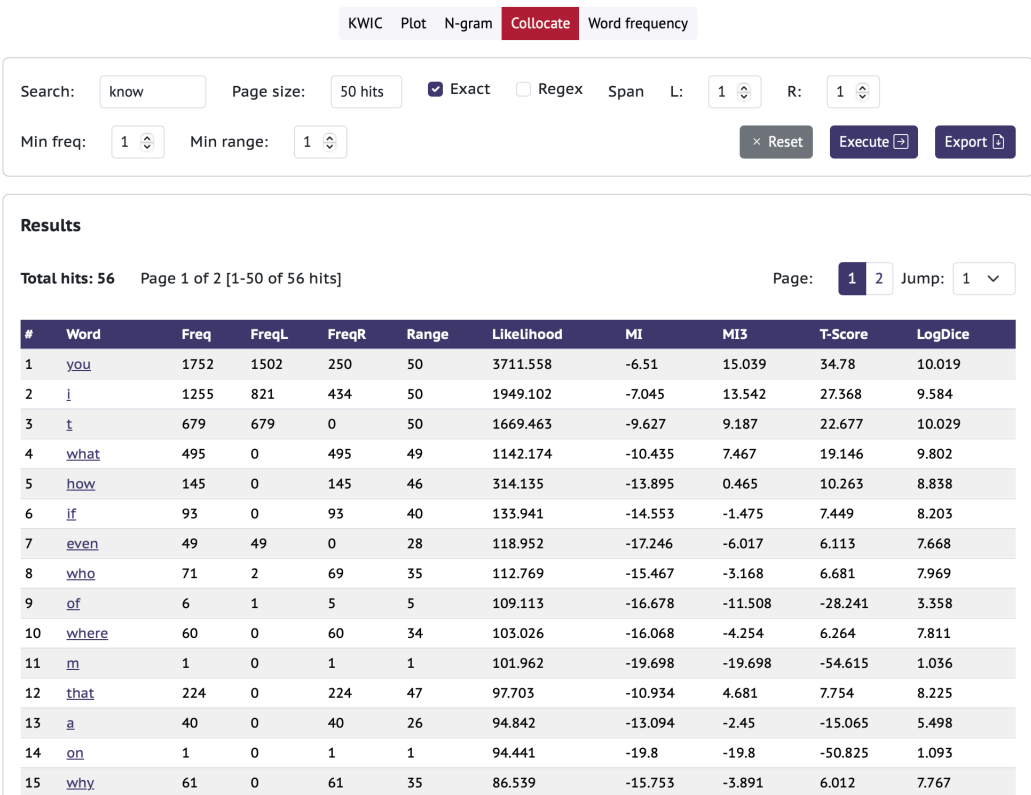
2. **Mutual Information (MI)** – Quantifies how much information is gained by knowing that two words occur together. It favors rare but exclusive word combinations, making it useful for detecting unexpected collocations. However, it can overemphasize low-frequency pairs, sometimes highlighting unusual or idiosyncratic word associations.

3. **MI3 (Mutual Information Cubed)** – Adjusts MI by giving more weight to frequency while maintaining selectivity. It strikes a balance between informativeness and reliability, making it particularly effective for identifying mid-frequency but meaningful collocations.

4. **T-score** – Measures the standardized difference between observed and expected co-occurrence. It prioritizes high-frequency word pairs, even if their association is weak. This measure is useful for identifying common and stable collocations but may overlook rare but meaningful patterns.

5. **LogDice** – A frequency-sensitive and symmetrical measure, bounded between 0 and 14. Designed for cross-linguistic comparisons (e.g., used in Sketch Engine), it is effective for analyzing collocations across different corpora or languages. While less intuitive statistically, it provides stable and reliable results.

Figure 8:

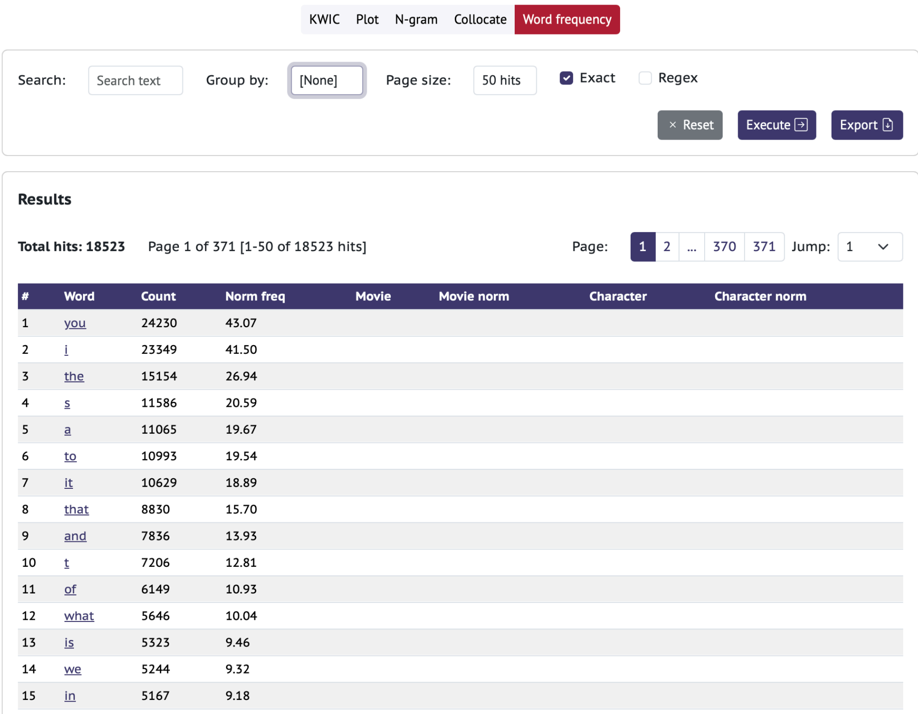


#	Word	Freq	FreqL	FreqR	Range	Likelihood	MI	MI3	T-Score	LogDice
1	you	1752	1502	250	50	3711.558	-6.51	15.039	34.78	10.019
2	i	1255	821	434	50	1949.102	-7.045	13.542	27.368	9.584
3	t	679	679	0	50	1669.463	-9.627	9.187	22.677	10.029
4	what	495	0	495	49	1142.174	-10.435	7.467	19.146	9.802
5	how	145	0	145	46	314.135	-13.895	0.465	10.263	8.838
6	if	93	0	93	40	133.941	-14.553	-1.475	7.449	8.203
7	even	49	49	0	28	118.952	-17.246	-6.017	6.113	7.668
8	who	71	2	69	35	112.769	-15.467	-3.168	6.681	7.969
9	of	6	1	5	5	109.113	-16.678	-11.508	-28.241	3.358
10	where	60	0	60	34	103.026	-16.068	-4.254	6.264	7.811
11	m	1	0	1	1	101.962	-19.698	-19.698	-54.615	1.036
12	that	224	0	224	47	97.703	-10.934	4.681	7.754	8.225
13	a	40	0	40	26	94.842	-13.094	-2.45	-15.065	5.498
14	on	1	0	1	1	94.441	-19.8	-19.8	-50.825	1.093
15	why	61	0	61	35	86.539	-15.753	-3.891	6.012	7.767

Frequency Tool

The Frequency Tool counts all words in the AMC and presents them as a frequency list, enabling users to identify the most common words in the corpus. Frequency results can be grouped by overall frequency (see Figure 9), by movie (see Figure 10), or by character (see Figure 11). Clicking on a result redirects to the KWIC tool, where the word's occurrences can be examined in context.

Figure 9·



KWIC Plot N-gram Collocate Word frequency

Search: Search text Group by: [None] Page size: 50 hits Exact Regex

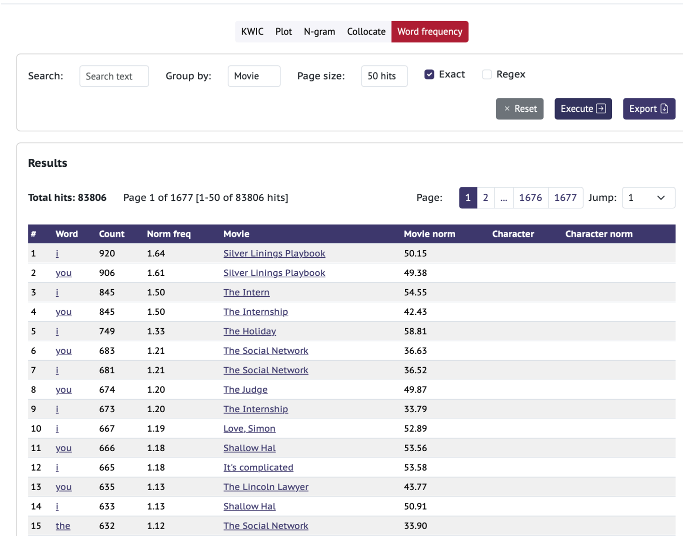
× Reset Execute Export

Results

Total hits: 18523 Page 1 of 371 [1-50 of 18523 hits] Page: 1 2 ... 370 371 Jump: 1

#	Word	Count	Norm freq	Movie	Movie norm	Character	Character norm
1	you	24230	43.07				
2	i	23349	41.50				
3	the	15154	26.94				
4	s	11586	20.59				
5	a	11065	19.67				
6	to	10993	19.54				
7	it	10629	18.89				
8	that	8830	15.70				
9	and	7836	13.93				
10	t	7206	12.81				
11	of	6149	10.93				
12	what	5646	10.04				
13	is	5323	9.46				
14	we	5244	9.32				
15	in	5167	9.18				

Figure 10:



KWIC Plot N-gram Collocate Word frequency

Search: Search text Group by: Movie Page size: 50 hits Exact Regex

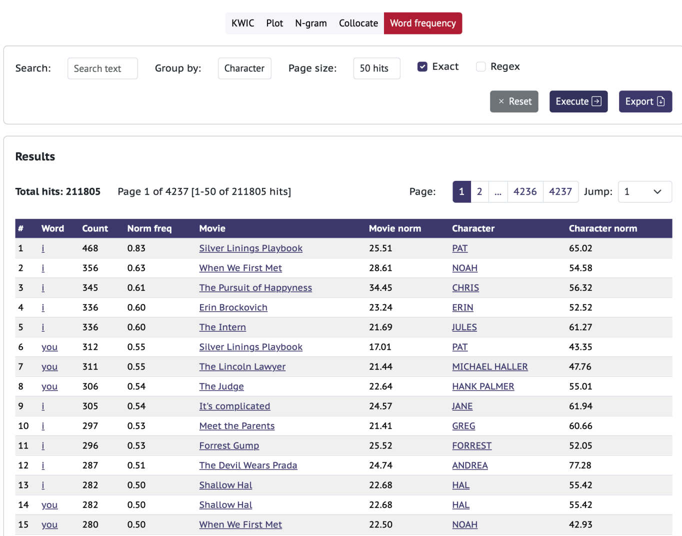
× Reset Execute Export

Results

Total hits: 83806 Page 1 of 1677 [1-50 of 83806 hits] Page: 1 2 ... 1676 1677 Jump: 1

#	Word	Count	Norm freq	Movie	Movie norm	Character	Character norm
1	i	920	1.64	Silver Linings Playbook	50.15		
2	you	906	1.61	Silver Linings Playbook	49.38		
3	i	845	1.50	The Intern	54.55		
4	you	845	1.50	The Internship	42.43		
5	i	749	1.33	The Holiday	58.81		
6	you	683	1.21	The Social Network	36.63		
7	i	681	1.21	The Social Network	36.52		
8	you	674	1.20	The Judge	49.87		
9	i	673	1.20	The Internship	33.79		
10	i	667	1.19	Love, Simon	52.89		
11	you	666	1.18	Shallow Hal	53.56		
12	i	665	1.18	It's complicated	53.58		
13	you	635	1.13	The Lincoln Lawyer	43.77		
14	i	633	1.13	Shallow Hal	50.91		
15	the	632	1.12	The Social Network	33.90		

Figure 11:



KWIC Plot N-gram Collocate Word frequency

Search: Search text Group by: Character Page size: 50 hits Exact Regex

× Reset Execute Export

Results

Total hits: 211805 Page 1 of 4237 [1-50 of 211805 hits] Page: 1 2 ... 4236 4237 Jump: 1

#	Word	Count	Norm freq	Movie	Movie norm	Character	Character norm
1	i	468	0.83	Silver Linings Playbook	25.51	PAT	65.02
2	i	356	0.63	When We First Met	28.61	NOAH	54.58
3	i	345	0.61	The Pursuit of Happyness	34.45	CHRIS	56.32
4	i	336	0.60	Erin Brockovich	23.24	ERIN	52.52
5	i	336	0.60	The Intern	21.69	JULES	61.27
6	you	312	0.55	Silver Linings Playbook	17.01	PAT	43.35
7	you	311	0.55	The Lincoln Lawyer	21.44	MICHAEL HALLER	47.76
8	you	306	0.54	The Judge	22.64	HANK PALMER	55.01
9	i	305	0.54	It's complicated	24.57	JANE	61.94
10	i	297	0.53	Meet the Parents	21.41	GREG	60.66
11	i	296	0.53	Forrest Gump	25.52	FORREST	52.05
12	i	287	0.51	The Devil Wears Prada	24.74	ANDREA	77.28
13	i	282	0.50	Shallow Hal	22.68	HAL	55.42
14	you	282	0.50	Shallow Hal	22.68	HAL	55.42
15	you	280	0.50	When We First Met	22.50	NOAH	42.93

Search Cheat Sheet

Search Function		Example	Result
Exact	when selected, it finds the exact word	you	**you** / **you?**
	If not selected, it finds any word containing the item	you	**you** / **you? / bayou / yourself / ...**
Regex	when selected, it finds word(s) through regular expressions		
^	starts with	^you	any item starting with **you**
$	ends with	ed$	any item ending with **ed**
*	zero or more characters	yo*	any item containing **y**, **yo**, the**y**, etc.
		^you*	any item starting with **yo**
.	any single character	c.t	any item containing **c.t** such as a**c**u**t**e, advo**c**a**t**e, etc.
		^c.t	any item such as **c**a**t**, **c**u**t**, **c**o**t**, etc.
.*	zero or more occurrences of any character	sa.*	accu**sa**tion
		^sa.*	any item starting with **sa**, such as **sa**lary, **sa**id, etc.
+	one or more characters	yo+	any item containing **yo**, **you**, etc.
?	zero or more characters	y?	zero or one occurrence of **y**
		yo?	any item containing **y**, **yo**, **you**, etc.
		^you+	any item starting with **you**
[]	any character inside the brackets	[0123456789]	any digit
		[0-9]	any digit
		[abc]	**a**, **b**, or **c**
(a|b|c)	any one character in the set	sa(y|id)	**essay**, **say** or **said**
		^sa(y|id)	**say** or **said**
a{n}	matches exactly n occurrences of a	a{2}	**aa**rgh, is**aa**c
a{n,}	matches n or more occurrences of a	a{3,}	**aaa**, **aaaa**, etc.
a{n,m}	matches between n and m occurrences of a	a{2,4}	**aa**, **aaa**, or **aaaa**

Feedback, Bugs, and Movie Transcription Contributions

If you have any suggestions for improving the interface or notice any bugs, please contact us at: a.palmeroaprosio@unitn.it (Alessio Palmero Aprosio) and pierfranca.forchini@unicatt.it (Pierfranca Forchini). Contributions for adding new movies are also welcome, provided transcription guidelines are strictly followed. This ensures that both authenticity and representativeness are preserved. The authenticity of the American movie dialogs is what makes this corpus stand out, offering a unique resource for linguistic analysis and spoken language teaching.

ACKNOWLEDGEMENTS

We would like to thank Lawrence Anthony for inspiring the idea behind this tool, following many years of development and free access to AntConc within corpus linguistics.