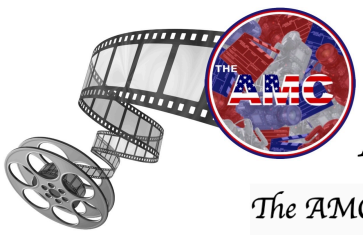
The AMC-50 DATA Reference

Pierfranca Forchini
Francesca Poli



April 11, 2022
Approved by
The AMC Internal Board

www.americanmoviecorpus.net

Contents

The AMC-50 DATA Reference

How to cite us:

Forchini, Pierfranca and Francesca Poli. 2022. *The AMC-50 DATA Reference*. www.americanmoviecorpus.net

1.0 Clapperboard

LEGAL DISCLAIMER: The American Movie Corpus (AMC) is conceived as a repository of data transformed into a new monomodal (textual) utility. As such, the AMC is a collection of movie dialogs transcribed by the AMC team which does not include any audiovisual (multimodal) material or script from the web. The copyright of the movies resides with the original copyright holder(s). The resulting texts are used for noncommercial/nonprofit purposes, such as linguistic research, scholarship, teaching, criticism and comment. The AMC is not publicly available, but data extracted from the corpus can be shared free of charge. Any scholars, language learners and/or teachers (henceforth USERS) who are interested in word lists, lexical bundles and collocations can use the data presented here by citing us (see above). USERS can also access *The AMC LAB* for further updates and contact *The AMC Team* to obtain concordances and snippets of dialog (cf. www.americanmoviecorpus.net). USERS declare and accept that the data obtained from the corpus will be used by them for the exclusive purposes of research, scholarship, teaching, criticism and comment. USERS assume complete responsibility: neither the AMC team nor the AMC Board are a party to or are in any way responsible for any copyright infringement.

BACKSTORY: The AMC LAB has been created to share data with scholars, teachers and learners who aim to investigate, teach and/or learn the lexico-grammatical features characterizing spoken language. Although movies are artifacts by nature, recent investigations of the AMC (see Opening Credits here and Publications on the AMC site) have, in fact, revealed that their dialogs share the same textuality and linguistic features of natural face-to-face conversation. These revolutionary findings have opened up new avenues:

- For SCHOLARLY RESEARCH: for many years movie language has been considered as artificially written-to-be-spoken and deemed unlikely to comprise the features that characterize conversation. Data from the AMC corpus offers new ways of approaching the study of movie language;

- For LANGUAGE LEARNING: language learners can improve their spoken competence through practice on movie conversation;

- For LANGUAGE TEACHING: authoritative scholars have been emphasizing the crucial role played by spoken language in communication for almost a hundred years. In spite of this, attention given to the study of lexico-grammatical spoken features in educational settings has been scarce. The textual and linguistic similarity of movie dialogs with face-to-face conversation and the rich resource of spoken language features which the AMC represents mean that teachers now have the chance to give spoken language its rightful place.

ASIDE: The data shared here are intended as a mere QUANTITATIVE REFERENCE for learners, teachers and scholars interested in movie discourse and are just an example of the role that movie language and corpora can have in the mastering of the most recurrent linguistic patterns found in conversation.

2.0 Opening Credits: Multi-Dimensional Analysis (MDA) DATA

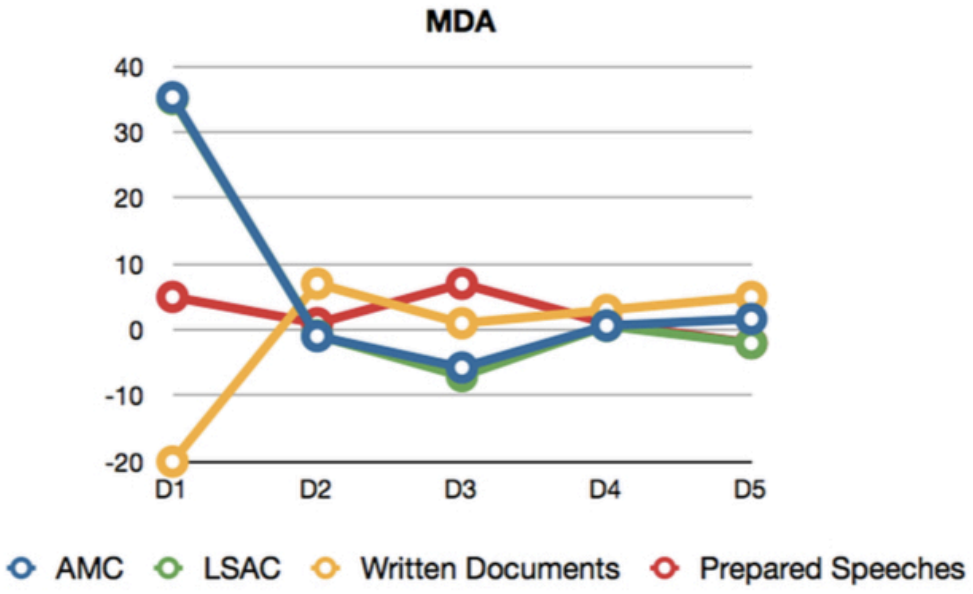
The present section illustrates findings (see Table 1, Graphs 1, 2, 3 and 4) from previous MDA studies which uncover the linguistic and textual features of movie discourse and its similarity with face-to-face conversation. All the corpora mentioned in this section were kindly tagged and processed for MDA by Douglas Biber with the *Biber grammatical tagger* he developed and with the *SAS software package* for statistical analyses he adapted for linguistic studies. With the aid of the *SAS software package*, the identified grammatical features were turned into the underlying *Dimensions* characterizing the various corpora. Technically, via factor analysis a large number of linguistic features characterizing a text are reduced to a small set of derived variables called *Factors*. Then, through a calculation of the communicative functions most widely shared by the linguistic features in question, each *Factor* is interpreted functionally as a *Dimension* of variation which underlines each set of co-occurring linguistic features[1]. More specifically, the following five Biberian *dimensions*, which are represented by *Factors 1-5* respectively, are considered here[2]:

1) *Dimension 1*: the *informational (negative) vs. involved (positive) production* dimension, which identifies whether a text is marked by high informational density and exact informational content or, on the contrary, by affective, interactional, and generalized content (Biber 1988:107);
2) *Dimension 2*: the *narrative (positive) vs. non-narrative (negative) concerns* dimension, which distinguishes narrative discourse from other types of discourse (Biber 1988:109);
3) *Dimension 3*: the *explicit (positive) vs. situation-dependent (negative) reference* dimension, which distinguishes between highly explicit, context-independent reference and non-specific, situation-dependent reference (Biber 1988:110);
4) *Dimension 4*: the *overt expression of persuasion (positive)* dimension, which marks the degree to which persuasion is marked overtly employed (Biber 1988:111);
5) *Dimension 5*: the *abstract (positive) vs. non-abstract (negative) information* dimension, which "seems to mark informational discourse that is abstract, technical, and formal versus other types of discourse" (Biber 1988:113).[3]

[1] This interpretation is based on the assumption that frequently co-occurring linguistic features in texts share at least one communicative function, and that it is possible to identify a unified *Dimension* underlying each set of co-occurring linguistic features (cf. Biber 1988).
[2] See Biber (1988 and 1995) and Biber and Conrad (2001) for details about *Dimensions* and *Factors 6* and *7*, which were also initially investigated, but then not included in later studies. Quoting Biber (1988, 114-115): *Dimension 6*, which is labelled *On-line Informational Elaboration*, "seems to distinguish discourse that is informational but produced under real-time conditions from other types of discourse", whereas *Dimension 7* "seems to mark academic hedging or qualification but is not sufficiently represented for a full interpretation".
[3] Source: Forchini (2021:14-15).

MULTI DIMENSIONAL ANALYSIS					
American Movie Corpus					
Variable	N	Mean	Std Dev	Minimum	Maximum
dim1	3	**35.3166667**	1.7013622	33.4100000	36.6800000
dim2	3	**-0.9700000**	0.2778489	-1.1500000	-0.6500000
dim3	3	**-5.7233333**	0.1887679	-5.9300000	-5.5600000
dim4	3	**0.6466667**	0.8195324	0.1100000	1.5900000
dim5	3	**1.6633333**	0.7011657	0.8700000	2.2000000
American Face-to-Face Conversation Corpus					
Variable	N	Mean	Std Dev	Minimum	Maximum
dim1	327	**35.0451070**	7.0665176	9.9800000	53.5800000
dim2	327	**-0.8459327**	1.3330098	-5.1000000	3.7900000
dim3	327	**-7.0434557**	2.0445282	-14.8100000	-1.1300000
dim4	327	**0.6002141**	2.0707167	-6.6100000	8.3400000
dim5	327	**-2.0426911**	0.7711589	-3.6300000	3.5000000

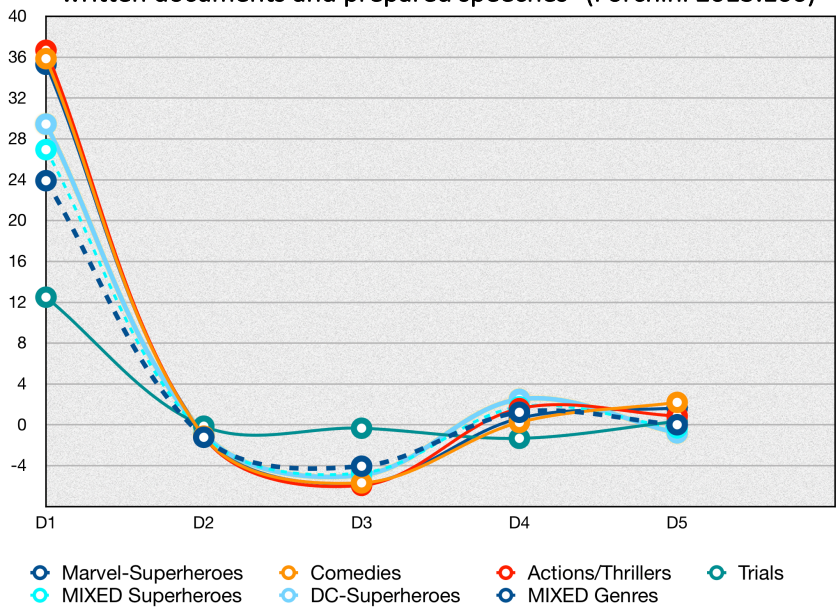
Table 1. MDA of American movie and face-to-face conversation[4] (Forchini 2012:64)



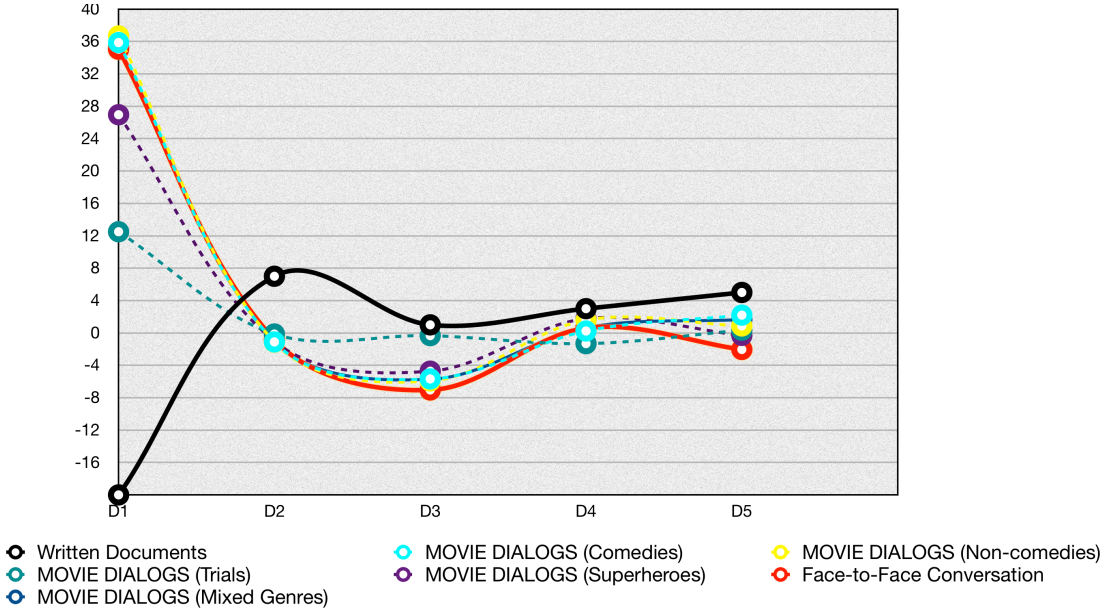
Graph 1. Movie conversation vs. face-to-face conversation,

[4] The label Variable stands for the 5 Dimensions (or Factors, i.e. dim1-5 in the table) taken into account; N for the number of texts (or sub-corpora) in the two corpora considered; *Mean* for the mean (average) frequency of items (the higher it is, the more frequent the items are); *Std Dev* for standard deviation, namely, a measure of the spread of the distribution; and a *Minimum* and *Maximum* for the minimum and maximum frequencies of items, respectively. Biber, Conrad and Reppen (1998: 280) explain that in all Multi-Dimensional studies "frequencies are standardized to a mean of 0.0 and a standard deviation of 1.0 before factor scores are computed. This process translates the scores for all features to scales representing standard deviation units, thus, regardless of whether a feature is extremely rare or extremely common in absolute terms, a standard score of +1 represents one standard deviation unit above the mean score for the feature in question. That is, standardized scores measure whether a feature is common or rare in a text relative to the overall average occurrence of that feature. The raw frequencies are transformed to standard scores so that all features on a factor will have equivalent weights in the computation of Dimension scores. If this process was not followed, extremely common features would have much greater influence than rare features on the Dimension scores." (Forchini 2012:64).

written documents and prepared speeches[5] (Forchini 2013:100)

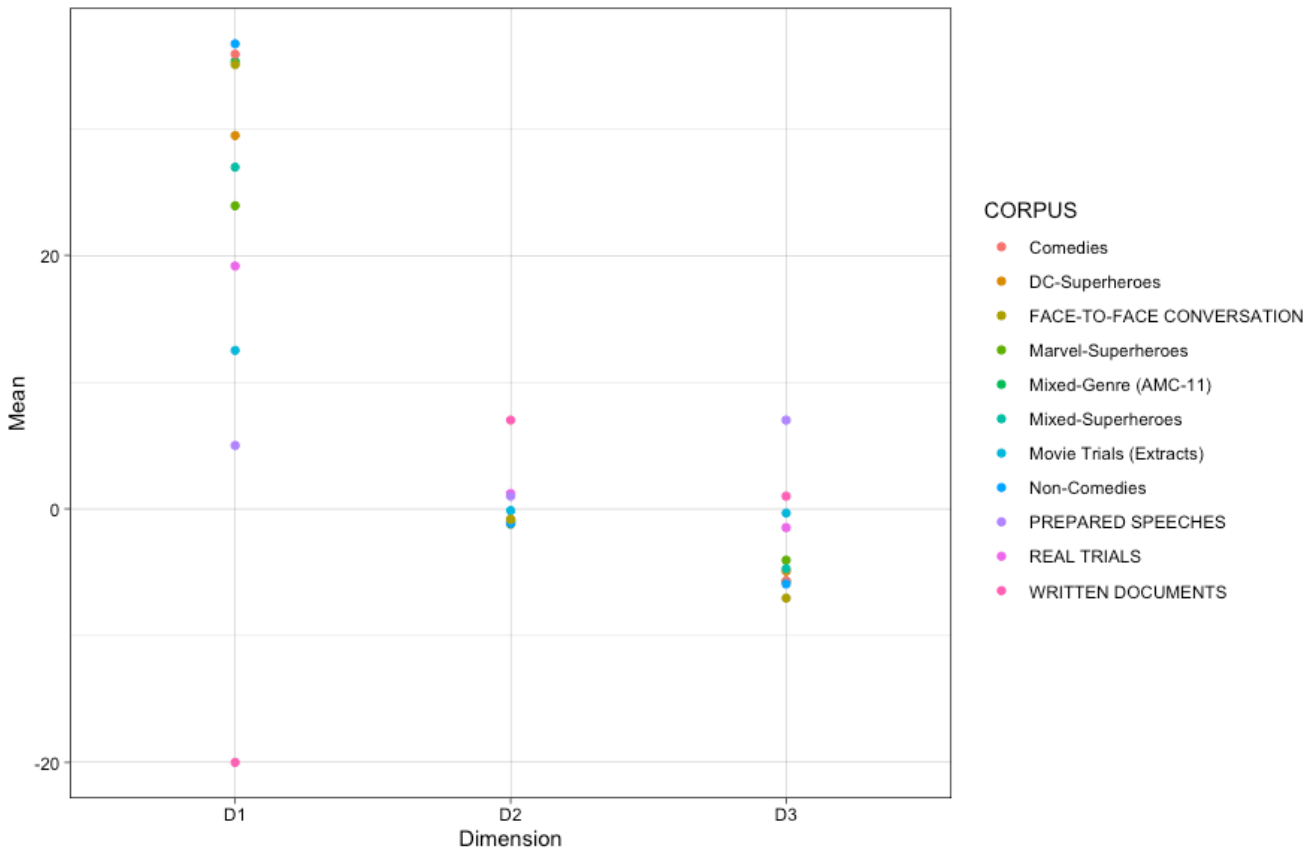


Graph 2. Movie genres (Forchini 2021:16)



Graph 3. Movie genres vs. face-to-face conversation and written documents (Forchini 2021:20)

[5]AMC stands for *American Movie Conversation* and LSAC stands for the *Longman Spoken American Corpus* which is taken from the *Longman Spoken and Written English Corpus* and belongs, together with the *Longman Written American Corpus*, to the *Longman Corpus Network*. In particular, the *Longman Spoken American Corpus* is owned by Pearson Education and was gathered by Professor Jack Du Bois and his team at the University of California at Santa Barbara (UCSB). I was kindly given access to the LSAC by Douglas Biber and Randi Reppen (Forchini 2013:100).



Graph 4. MDA Scatterplot[6]: movie genres vs. face-to-face conversation, prepared speeches, real trials and written documents

[6] MDA Scatterplot (Dimensions 1, 2 and 3) of all the movie genres considered so far compared to face-to-face conversation, prepared speeches and written documents.

3.0 The AMC-50 DATA

From this section on, **data are new** (i.e. they have never been presented/published before). More specifically, the present REFERENCE contains DATA from the dialogs of 50 movies (henceforth AMC-50 DATA) produced in the USA from 1959 to 2019 (see Forchini 2021:33-35 and Table 2 here) which have been processed and extracted via *AntConc 4.0.5* (www.laurenceanthony.net/software/antconc), *#LancsBox 6.0* (http://corpora.lancs.ac.uk/lancsbox), *SketchEngine* (www.sketchengine.eu), *WordSmith Tools* 8.0 (https://lexically.net) and *RStudio* (2020).

#	MOVIES	YEAR	DIRECTOR(S)	GENRE (IMDb[7])
01	*Captain Marvel*	2019	Anna Boden and Ryan Fleck	Action, Adventure, Sci-Fi
02	*Midnight Sun*	2018	Scott Speer	Drama, Romance
03	*When We First Met*	2018	Ari Sandel	Comedy, Fantasy, Romance
04	*Love, Simon*	2018	Greg Berlanti	Comedy, Drama, Romance
05	*The Circle*	2017	James Ponsoldt	Drama, Sci-Fi, Thriller
06	*Wonder*	2017	Stephen Chbosky	Drama, Family
07	*Gifted*	2017	Marc Webb	Drama
08	*Nerve*	2016	Henry Joost, Ariel Schulman	Action, Adventure, Crime
09	*The Intern*	2015	Nancy Meyers	Comedy, Drama
10	*The Fault in our Stars*	2014	Josh Boone	Drama, Romance
11	*The Judge*	2014	David Dobkin	Crime, Drama
12	*Iron Man 3*	2013	Shane Black	Action, Adventure, Sci-Fi
13	*Man of Steel*	2013	Zan Snyder	Action, Adventure, Sci-Fi
14	*The Internship*	2013	Shawn Levy	Comedy
15	*The Lucky One*	2012	Scott Hicks	Drama, Romance
16	*The Avengers*	2012	Joss Whedon	Action, Adventure, Sci-Fi
17	*The Amazing Spiderman*	2012	Marc Webb	Action, Adventure, Sci-Fi
18	*Silver Linings Playbook*	2012	David O. Russell	Comedy, Drama, Romance
19	*The Adjustment Bureau*	2011	George Nolfi	Romance, Sci-Fi, Thriller
20	*Captain America: The First Avenger*	2011	Joe Johnston	Action, Adventure, Sci-Fi
21	*The Lincoln Lawyer*	2011	Brad Furman	Crime, Drama, Thriller
22	*The Social Network*	2010	David Fincher	Biography, Drama

[7] www.imdb.com

23	*The Proposal*	2009	Anne Fletcher	Comedy, Drama, Romance
24	*It's Complicated*	2009	Nancy Meyers	Comedy, Drama, Romance
25	*Julie & Julia*	2009	Nora Ephron	Biography, Drama, Romance
26	*Twilight*	2008	Catherine Hardwicke	Drama, Fantasy, Romance
27	*The Dark Knight*	2008	Christopher Nolan	Action, Crime, Drama
28	*Iron Man*	2008	Jon Favreau	Action, Adventure, Sci-Fi
29	*The Holiday*	2006	Nancy Meyers	Comedy, Romance
30	*Cars*	2006	John Lasseter / Joe Ranft	Animation, Comedy, Family
31	*The Pursuit of Happiness*	2006	Gabriele Muccino	Biography, Drama
32	*The Devil Wears Prada*	2006	David Frankel	Comedy, Drama
33	*Madagascar*	2005	Eric Darnell / Tom McGrath	Animation, Adventure, Comedy
34	*Catwoman*	2004	Pitof	Action, Crime, Fantasy
35	*The Matrix Reloaded*	2003	Andy and Larry Wachowsky	Action, Sci-Fi
36	*Runaway Jury*	2003	Gary Fleder	Crime, Drama, Thriller
37	*One Hour Photo*	2002	Mark Romanek	Drama, Thriller
38	*Ocean's Eleven*	2001	Steven Soderbergh	Crime, Thriller
39	*Shallow Hal*	2001	Bobby and Peter Farrelly	Comedy, Drama, Fantasy
40	*Meet the Parents*	2000	Jay Roach	Comedy, Romance
41	*Erin Brockovich*	2000	Steven Soderbergh	Biography, Drama
42	*Mission: Impossible II*	2000	John Woo	Action, Adventure, Thriller
43	*Me, Myself & Irene*	2000	Bobby and Peter Farrelly	Comedy
44	*Finding Forrester*	2000	Gus Van Sant	Drama
45	*Forrest Gump*	1994	Robert Zemeckis	Drama, Romance
46	*Philadelphia*	1993	Jonathan Demme	Drama
47	*Back to the Future*	1985	Robert Zemeckis	Adventure, Comedy, Sci-Fi
48	*The Blues Brothers*	1980	John Landis	Adventure, Comedy, Crime
49	*Young Frankenstein*	1974	Mel Brooks	Comedy
50	*Some Like it Hot*	1959	Billy Wilder	Comedy, Music, Romance

Table 2. The AMC-50 movies

3.1 SIZE

The size of the AMC varies depending on what is being counted and what counts as a word: a TOKEN is the smallest unit that a corpus consists of, consequently, the number of tokens in a corpus represents the total number of individual words it contains. A TYPE, instead, is a unique word form in a corpus, consequently, the number of types in a corpus represents the number of unique word forms it contains. As for the four tools used here, it goes beyond our scope to cover how they count and process words and we trust the interested USER to have all the useful information required or, if necessary, to check the developers' manuals[8]. Table 3 illustrates the number of tokens and types present in the AMC calculated via *AntConc 4.0.5*, *#LancsBox 6.*0, *SketchEngine* and *WordSmith Tools 8.0*.

SOFTWARE	TOKENS	TYPES
AntConc 4.0.5	≈ 560000	18216
#LancsBox 6.0	≈ 530000	20477
SketchEngine	≈ 580000	19950
WordSmith Tools 8.0	≈ 533000	18703

Table 3. The AMC-50 tokens and types

[8] Cf. http://www.laurenceanthony.net/software/antconc/resources/help_AntConc321_english.pdf for *AntConc 4.0.5*, http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_6.0_manual.pdf for *#LancsBox 6.*0, www.sketchengine.eu/guide/glossary for *SketchEngine* and https://lexically.net for *WordSmith Tools 8.0*.

3.2 FREQUENCIES

3.2.1 Word lists

Table 4 shows the 100 most frequent words in the AMC-50 processed and extracted via *AntConc 4.0.5*, *#LancsBox 6.0*, *SketchEngine* and *WordSmith Tools 8.0*.

RANK	*AntConc 4.0.5*		*#LancsBox 6.0*		*SketchEngine*		*WordSmith Tools 8.0*	
	TYPE	**FREQUENCY**	**TYPE**	**FREQUENCY**	**TYPE**	**FREQUENCY**	**TYPE**	**FREQUENCY**
01	you	24214	you	21151	you	24210	you	21620
02	i	23336	i	16909	i	23334	i	17902
03	the	15149	the	15143	the	15145	the	15149
04	s	11579	a	11021	a	11028	a	11069
05	a	11074	to	10987	to	10996	to	11013
06	to	11013	and	7813	it	10609	and	7831
07	it	10611	it	7131	's	10230	it	7616
08	that	8820	that	6779	that	8820	that	7113
09	and	7831	of	6144	and	7815	of	6150
10	t	7204	is	5334	n't	7164	is	16401
11	of	6150	in	5140	do	6629	in	5167
12	what	5632	me	4908	of	6144	what	5008
13	is	5332	what	4871	what	5627	me	4914
14	we	5240	this	4804	is	5626	this	4810
15	in	5167	no	4361	we	5240	no	4368
16	me	4915	oh	4265	in	5146	oh	4271
17	this	4815	on	4213	me	4912	on	4227
18	m	4546	i'm	3814	this	4815	we	3998
19	no	4372	your	3806	'm	4468	i'm	3821
20	oh	4271	we	3779	no	4361	your	3808
21	on	4227	my	3686	oh	4268	my	3688
22	your	3808	for	3626	on	4216	for	3629
23	re	3714	know	3580	your	3806	know	3585
24	my	3688	have	3441	my	3687	have	3442
25	for	3629	do	3411	're	3684	do	3406
26	know	3586	just	3280	for	3627	just	3280
27	he	3479	not	3201	have	3611	don't	3223
28	have	3442	are	3181	know	3581	not	3204

29	do	3407	was	3069	he	3479	are	3183
30	just	3280	yeah	3048	was	3330	was	3070
31	don	3227	it's	2912	are	3297	yeah	3054
32	not	3204	all	2906	just	3280	all	2929
33	are	3183	be	2883	not	3258	it's	2915
34	was	3070	so	2799	yeah	3054	be	2887
35	yeah	3054	right	2667	all	2915	so	2808
36	all	2937	with	2664	be	2884	right	2676
37	can	2920	don't	2661	so	2799	with	2664
38	be	2887	but	2568	right	2673	but	2569
39	so	2808	like	2424	with	2664	he	2546
40	right	2676	he	2391	but	2569	like	2426
41	with	2664	get	2346	here	2441	get	2348
42	but	2569	here	2320	like	2425	here	2345
43	here	2441	okay	2204	get	2347	uh	2270
44	like	2426	uh	2172	did	2291	okay	2207
45	get	2348	go	2150	there	2207	go	2152
46	uh	2271	about	2144	okay	2206	about	2145
47	there	2208	out	2054	uh	2172	out	2065
48	okay	2207	up	1990	they	2159	up	2038
49	they	2159	can	1985	go	2152	can	1985
50	go	2153	you're	1943	about	2145	you're	1947
51	about	2145	gonna	1900	out	2057	gonna	1900
52	out	2065	hey	1857	can	2042	hey	1858
53	up	2038	come	1855	up	1994	s	1857
54	gonna	1900	well	1837	gonna	1900	come	1856
55	she	1888	at	1825	she	1888	well	1849
56	hey	1858	one	1740	hey	1858	at	1828
57	come	1857	if	1725	come	1855	one	1787
58	well	1849	got	1715	well	1839	if	1725
59	at	1828	now	1700	at	1826	got	1716
60	one	1813	that's	1673	'll	1779	they	1713
61	ll	1781	good	1645	one	1770	now	1702
62	how	1755	how	1610	how	1754	there	169
63	if	1725	they	1589	if	1725	that's	168
64	got	1716	there	1569	got	1715	good	1658
65	now	1703	him	1488	now	1703	how	1638

66	good	1659	did	1408	good	1649	him	1489
67	him	1489	mm	1400	him	1489	she	143
68	let	1407	think	1393	mm	1401	did	1402
69	did	1402	her	1351	think	1393	mm	1401
70	mm	1401	she	1342	her	1351	think	14
71	think	1400	yes	1270	've	1312	her	1351
72	her	1351	see	1266	let	1308	yes	1272
73	ve	1318	look	1207	yes	1272	see	1267
74	yes	1272	why	1135	see	1267	look	1207
75	see	1267	an	1123	look	1207	why	1136
76	look	1207	from	1113	who	1200	an	1124
77	who	1203	time	1075	would	1177	from	1117
78	why	1149	really	1067	why	1149	time	1102
79	an	1124	back	1060	an	1123	really	107
80	from	1117	when	1036	from	1117	back	1062
81	time	1105	as	1012	time	1077	when	1037
82	really	1070	take	1007	really	1070	who	1017
83	back	1064	were	979	back	1063	as	1012
84	when	1046	who	977	when	1046	take	101
85	as	1012	want	968	were	1045	were	978
86	take	1010	or	945	as	1012	want	970
87	were	978	would	944	take	1007	would	949
88	want	970	will	937	want	970	or	948
89	would	965	thank	931	could	961	will	939
90	or	948	sorry	930	or	945	can't	935
91	where	939	them	918	ca	942	thank	931
91	will	939	say	917	will	939	sorry	930
93	sorry	931	going	912	where	939	them	918
93	thank	931	little	898	sorry	931	say	917
95	man	923	his	894	thank	931	going	915
96	them	918	man	890	them	918	man	911
97	say	917	something	876	say	917	little	898
98	going	915	us	864	going	912	his	897
99	little	898	need	860	man	900	didn't	890
100	his	897	down	859	little	898	something	882

Table 4. The AMC-50 most frequent words

3.2.2 Scatterplots, boxplots and density plots[9]

A SCATTERPLOT is "a two-dimensional coordinate system in which the values of the vector are interpreted as coordinates of the *y*-axis, and the order in which they appear in the vector are the coordinates of the *x*-axis" (Gries 2009:98). A *regression line* is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x. Put simply, a scatterplot is a graph which displays a vector according to two dimensions on its x and y axes. A vector is an object with a magnitude and a direction, so that the y axis represents the magnitude, that is the size, while the x axis represents the order in which it appears (i.e. the direction). It is also possible to superimpose a line on the scatterplot which summarizes the relationship between the y and x axes' variables. This line is called "regression line" and it is helpful when we wish to visualize the trend of the vector which will then need to be verified with appropriate tests.

A BOXPLOT contains various types of valuable information (Gries 2009:119):

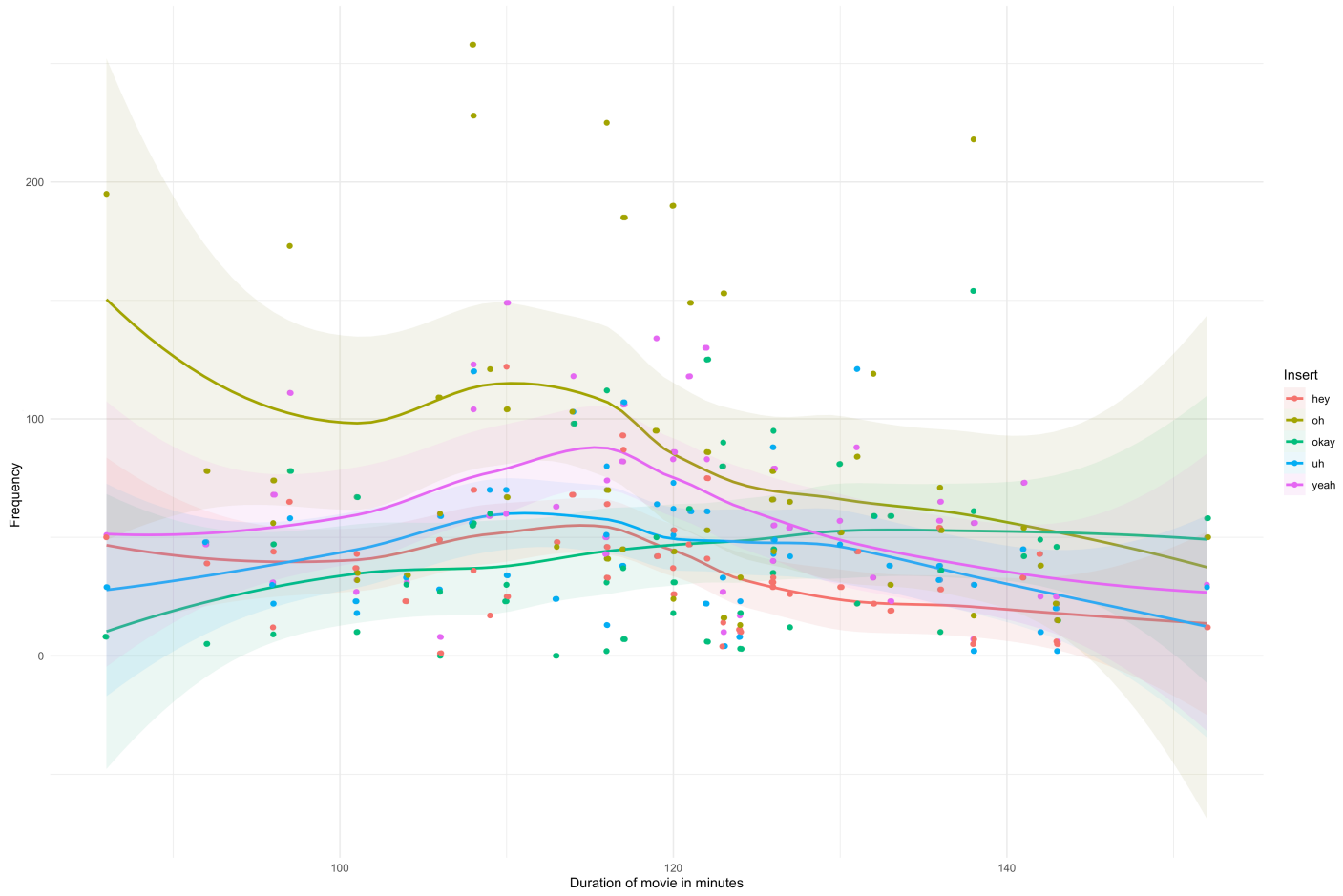
- the bold-typed horizontal lines represent the medians of the two vectors;
- the regular horizontal lines that make up the upper and lower boundary of the boxes represent the hinges (approximately the 75%- and the 25% quartiles);
- the whiskers – the dashed vertical lines extending from the box until the upper and lower limit – represent the largest and smallest values that are not more than 1.5 interquartile ranges away from the box;
- each outlier that would be outside of the range of the whiskers would be represented with an individual dot;
- the notches on the left and right sides of the boxes extend across the range ±1.58*IQR/sqrt(n): if the notches of two boxplots overlap, then these will most likely not be significantly different.

A DENSITY PLOT shows the ordered numerical values of a variable *x* on the horizontal axis, and the probability density of *x* on the vertical axis (Levshina 2015:51). Put simply, a density plot is a useful graph when one wishes to display the distribution of the data. If the distribution appears to be normal (and this is verified with a statistical test called *Shapiro-Wilk test*), then parametric tests can be employed to explore the dataset. However, if the density plot shows that the data is skewed, that is, not normally distributed (and this is verified via means of a *Shapiro-Wilk test*), then the non-parametric version of the tests should be preferred. This graph plots all the numerical values of the data on the x axis, while the y axis corresponds to the probability density. Thus, any peaks in the curve are to be interpreted as follows: the majority of numerical values are distributed along this peak, while the remaining ones are found along its tail(s).

[9] Retrieved with *AntConc 4.0.5* and processed with *RStudio*.

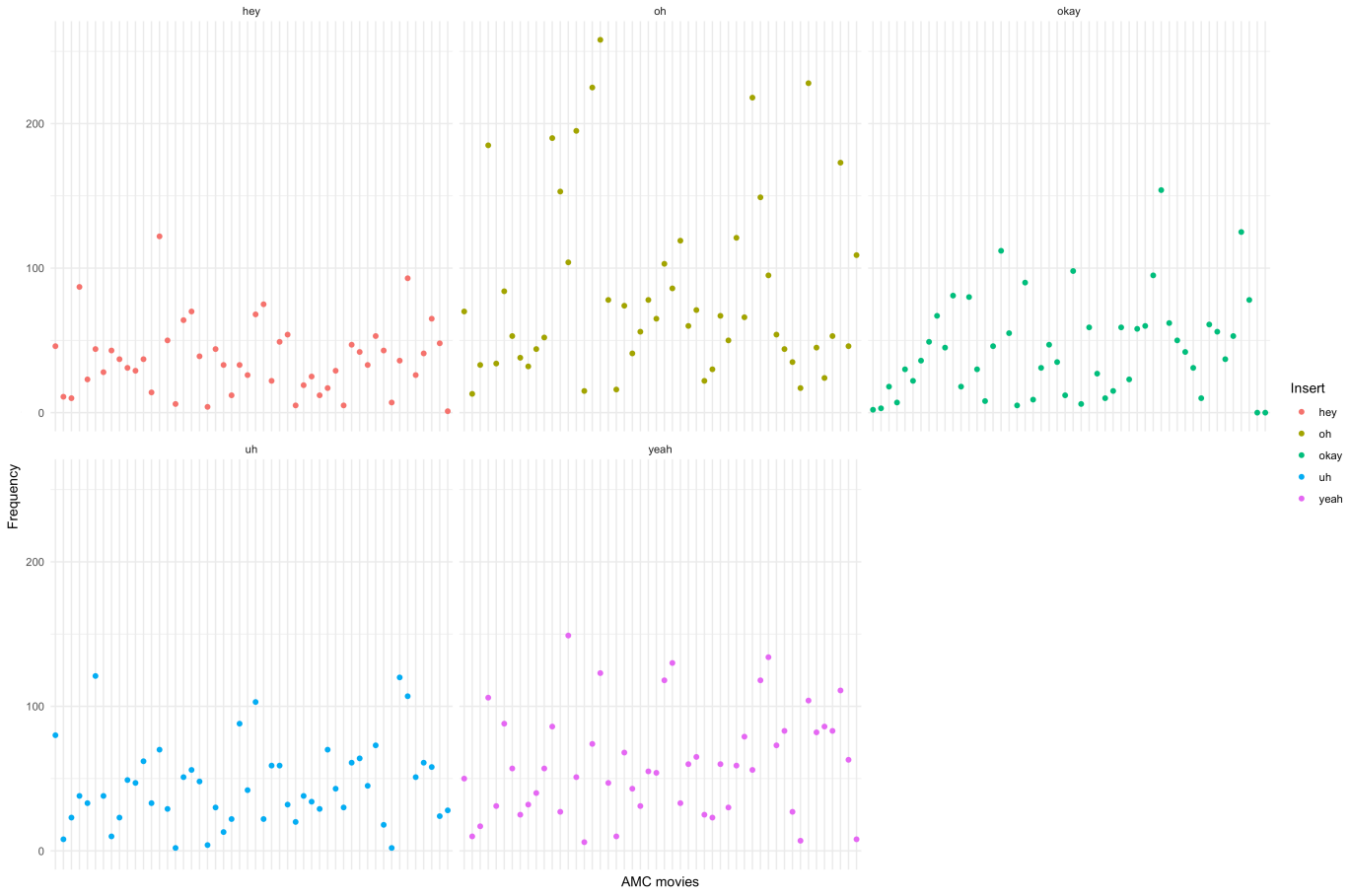
3.2.2.1 Inserts

Graph 5 is a scatterplot with regression line of the top 5 most frequent inserts extracted from the AMC-50. Each dot represents the frequency of a given insert in one movie of the corpus. The x axis is the length of the movies in minutes, while the y axis is the frequency of the insert. The dots at the top of the graph are outliers, that is, observations (i.e. inserts) whose frequency exceeds the average. These may or may not be considered during the analysis. The lines superimposed over the top summarize the relationship between variables: they are called regression lines. The shaded areas around the lines are the 95% confidence intervals. Confidence intervals refer to the fact that any random value selected from the dataset has a 95% probability of falling between the two given values.



Graph 5. Scatterplot of the top 5 most frequent inserts present in the AMC-50

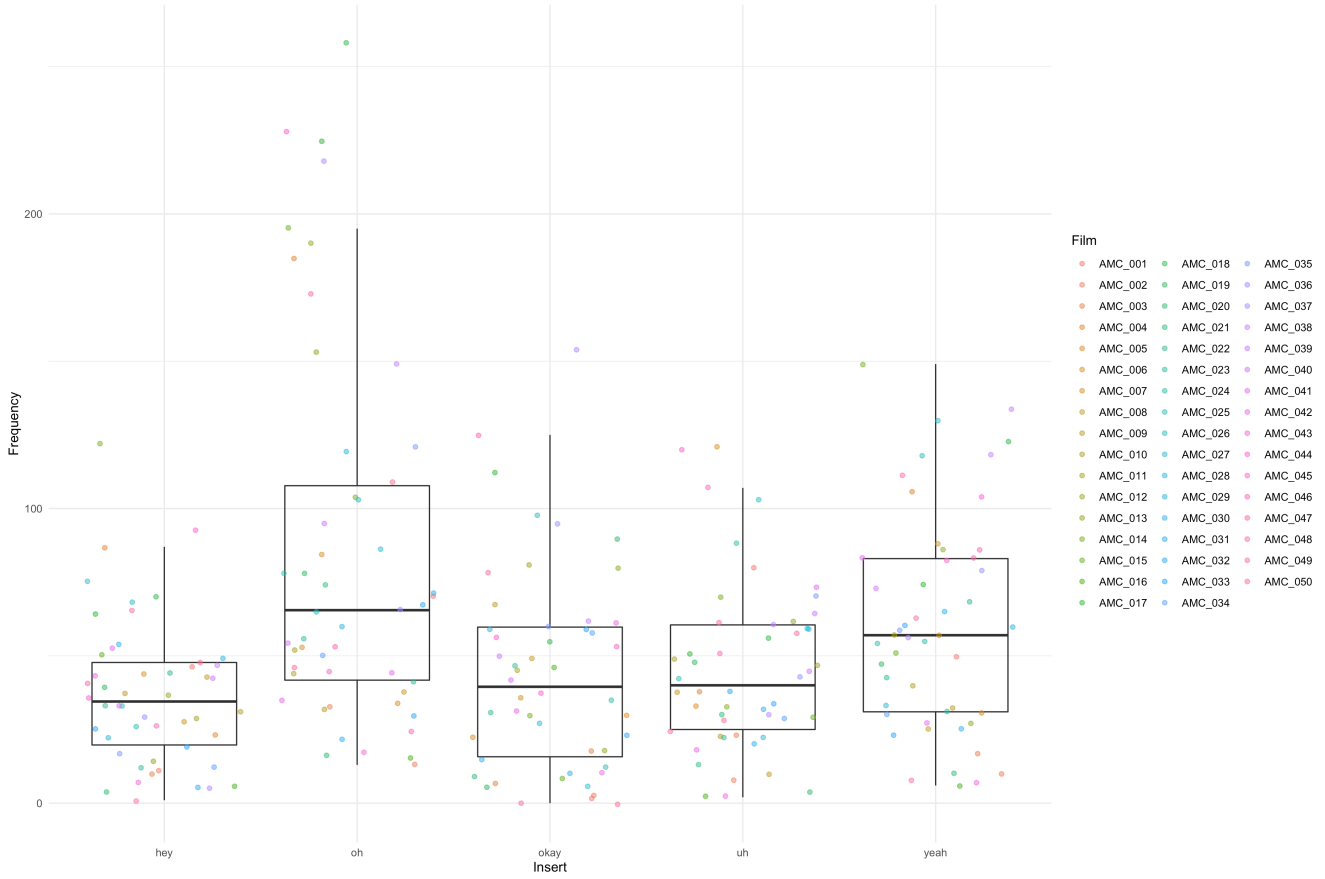
Graph 6 displays grouped scatterplots of the top 5 most frequent inserts extracted from the AMC-50 via *AntConc 4.0.5* and processed with *RStudio*. Each line on the x axes corresponds to a movie, while the y axes indicate the frequency of the insert. The dots represent the inserts. The scatterplots have been grouped by inserts so as to offer a bird's-eye view of the most frequent inserts and their frequency side by side.



Graph 6. Grouped scatterplots of the top 5 most frequent inserts present in the AMC-50

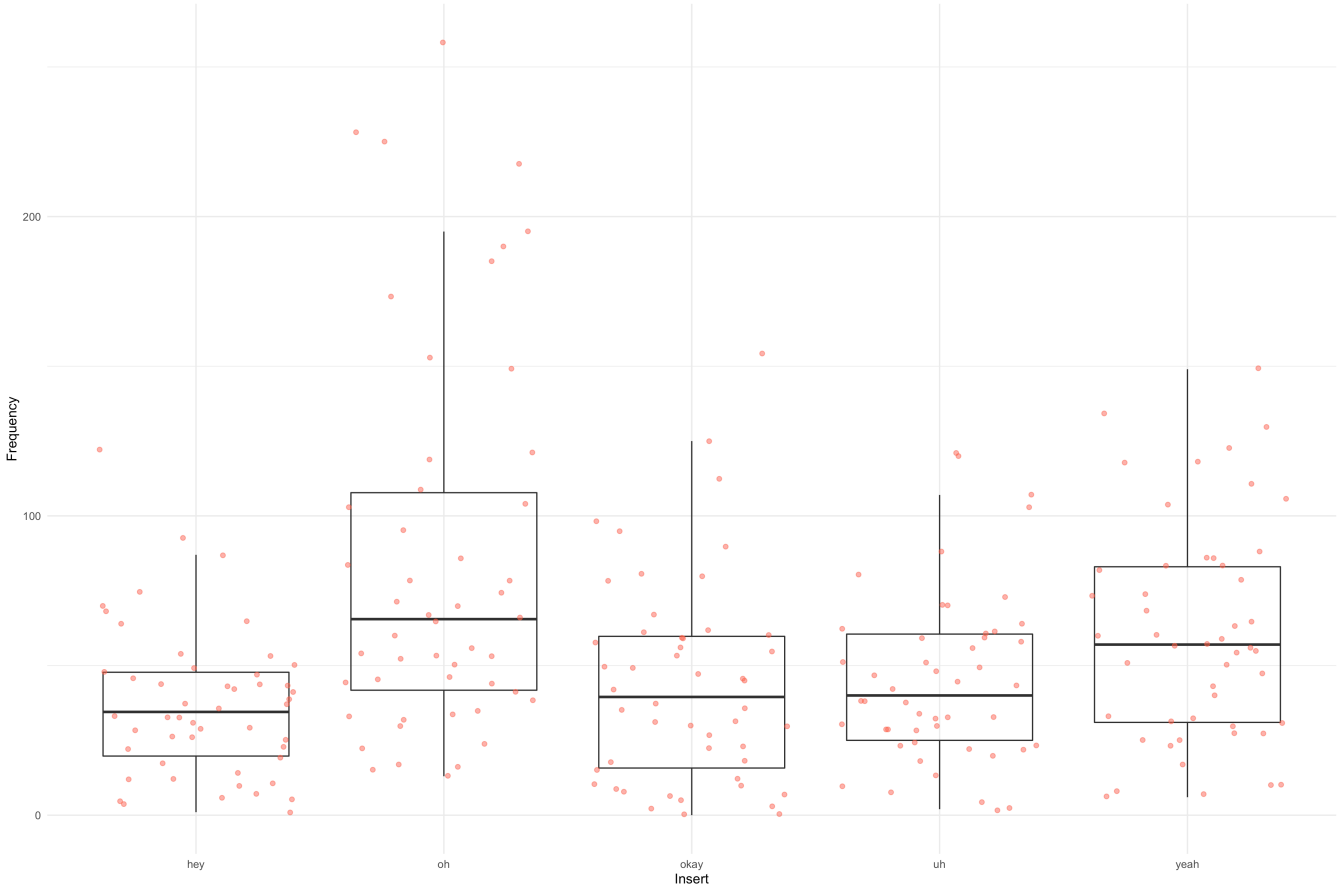
Graph 7 displays boxplots for each of the five most frequent inserts extracted from the AMC-50 via *AntConc 4.0.5* and processed with *RStudio*. On the x axis are the inserts, while on the y axis the raw frequency of each insert in each movie of the AMC-50 is displayed. Furthermore, each boxplot indicates several key values for each insert: the lowest point of the bottom whisker (or a dot below it) indicates the lowest frequency and the highest point of the whisker (or dot above it) indicates the highest frequency. For example, the boxplot of *mm* shows that the highest frequency of the insert is greater than that of *hey*. The lowest edge of the white box is the first quartile (i.e. each of the four parts into which the data has been divided), so that the distance between the bottom of the vertical line and the lowest

edge of the white box is the range between which the lowest 25% of frequencies fall. This range is larger in *hey* compared to *mm*, which means that there is more variability in the frequency of *hey* compared to that of *mm*. The box shows the interquartile range, in other words, 50% of the frequencies are bigger than the lower part of the box area, but smaller than the top part. The boxes of *hey* and *uh* are of similar size, as well as those of *okay* and *yeah*. The top edge of the box shows the value of the upper quartile, thus the distance between the top edge of the box and the top of the vertical line shows the range between which the top 25% of frequencies fall. In the middle of the box is a line that represents the median (i.e. the number found in the middle of the data). The median for *hey* is higher than for *mm*. Lastly, if the whiskers are the same length, then the distribution is symmetrical (cf. the description of the density plot); however, if the top or bottom whisker is much longer, then the distribution is asymmetrical. The dots above the boxplots represent the outliers (i.e. a value that differs significantly from the rest of the data). The legend displays the different colors used to group the frequencies by movie.



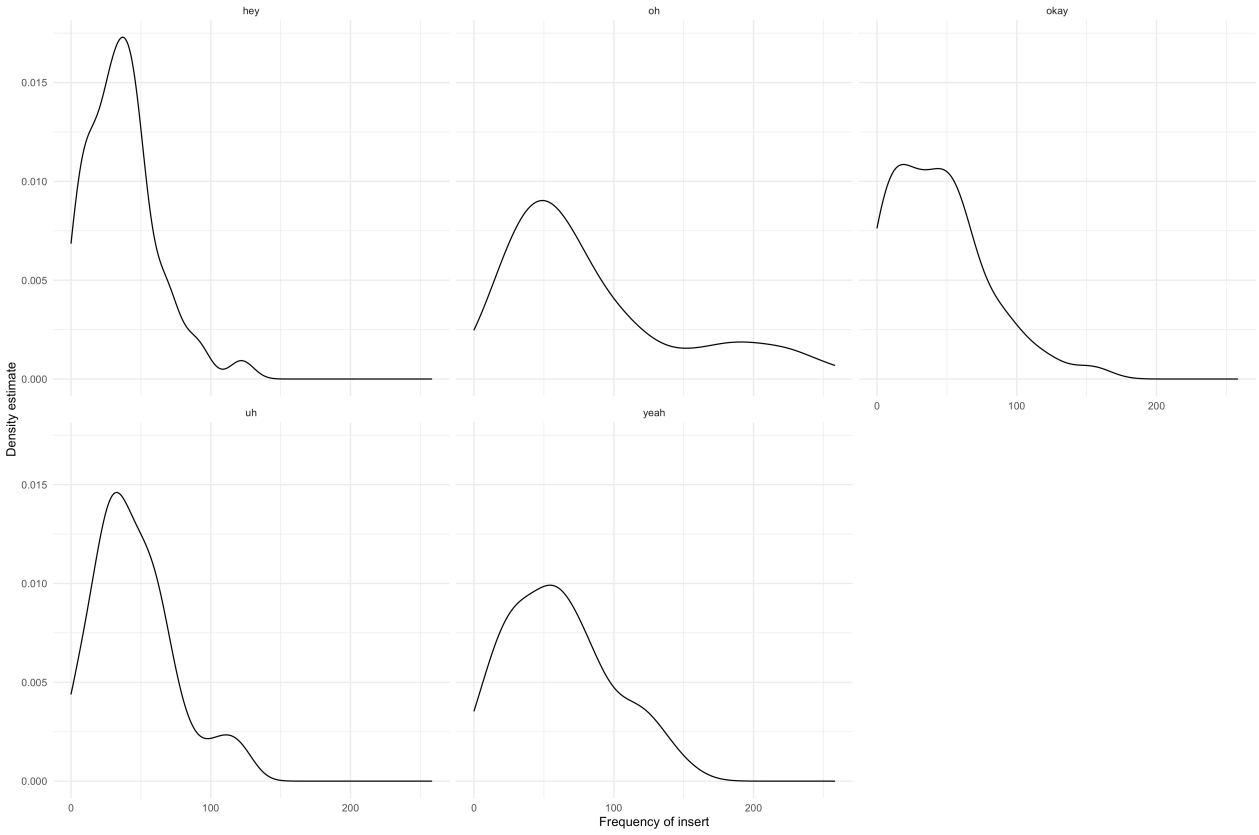
Graph 8. Grouped boxplots of the top 5 most frequent inserts present in the AMC-50 according to the movie in which they occur

As opposed to Graph 8, Graph 9 does not differentiate the dots/inserts on the basis of the movie in which they occur.



Graph 9. Grouped boxplots of the top 5 most frequent inserts present in the AMC-50

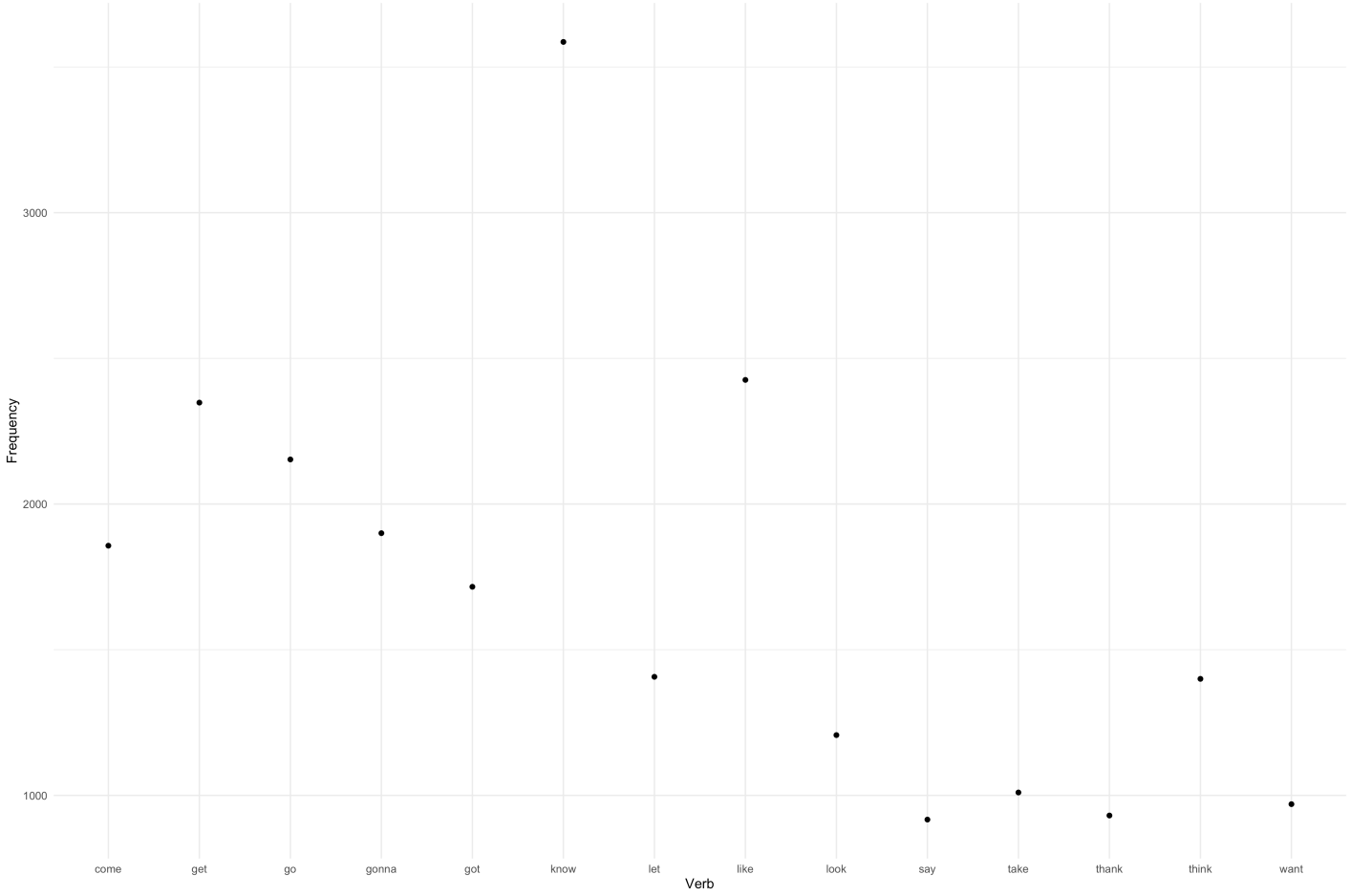
Graph 10 displays the density plots of the top 5 most frequent inserts extracted from the AMC-50. The curve represents the skewness (that is, the asymmetry or non-normality in the distribution) of the data, so that if the peak of the curve is found on the left, the data are right skewed, whereas if the data peaks on the right, it is left skewed. If the density curve is left skewed, the mean is smaller than the median; if the density curve is right skewed, the mean is greater than the median. Lastly, if the curve has no skewness, the mean and the median are equal. The graph shows that all the data regarding the frequency of the top 5 inserts are right skewed and unimodal, except for *uh* which shows two minor peaks, indicating that the data might be bimodal.



Graph 10. Density plots of the top 5 most frequent inserts present in the AMC-50

3.2.2.2 Verbs

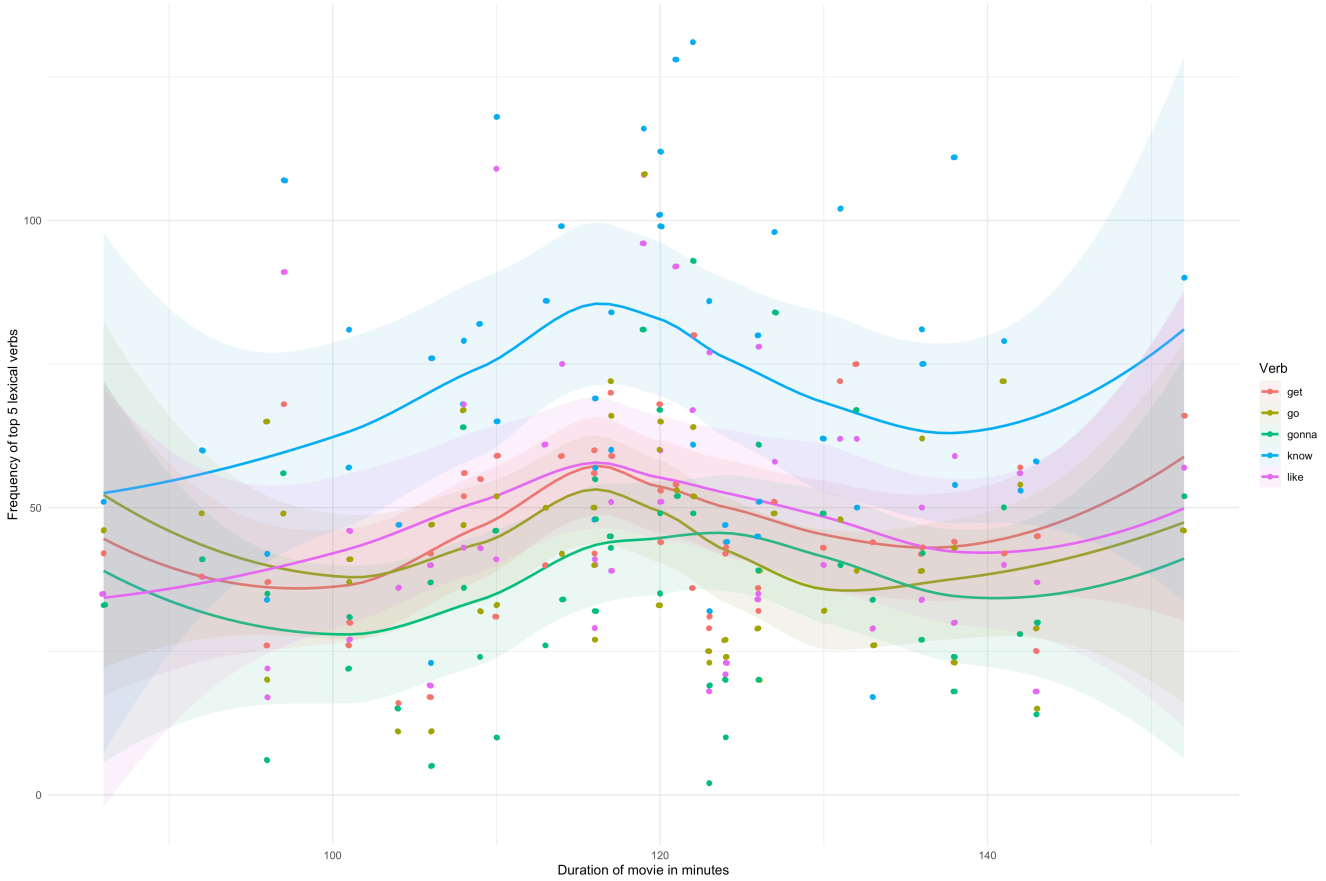
Graph 11 is a simple scatterplot of the most frequent verbs extracted from the AMC-50. Each dot represents the frequency of a given verb. The frequency is the overall frequency of the verb in all 50 movies of the corpus. The x axis displays the verbs, while the y axis indicates the overall raw frequency.



Graph 11. Scatterplot of the most frequent verbs present in the AMC-50

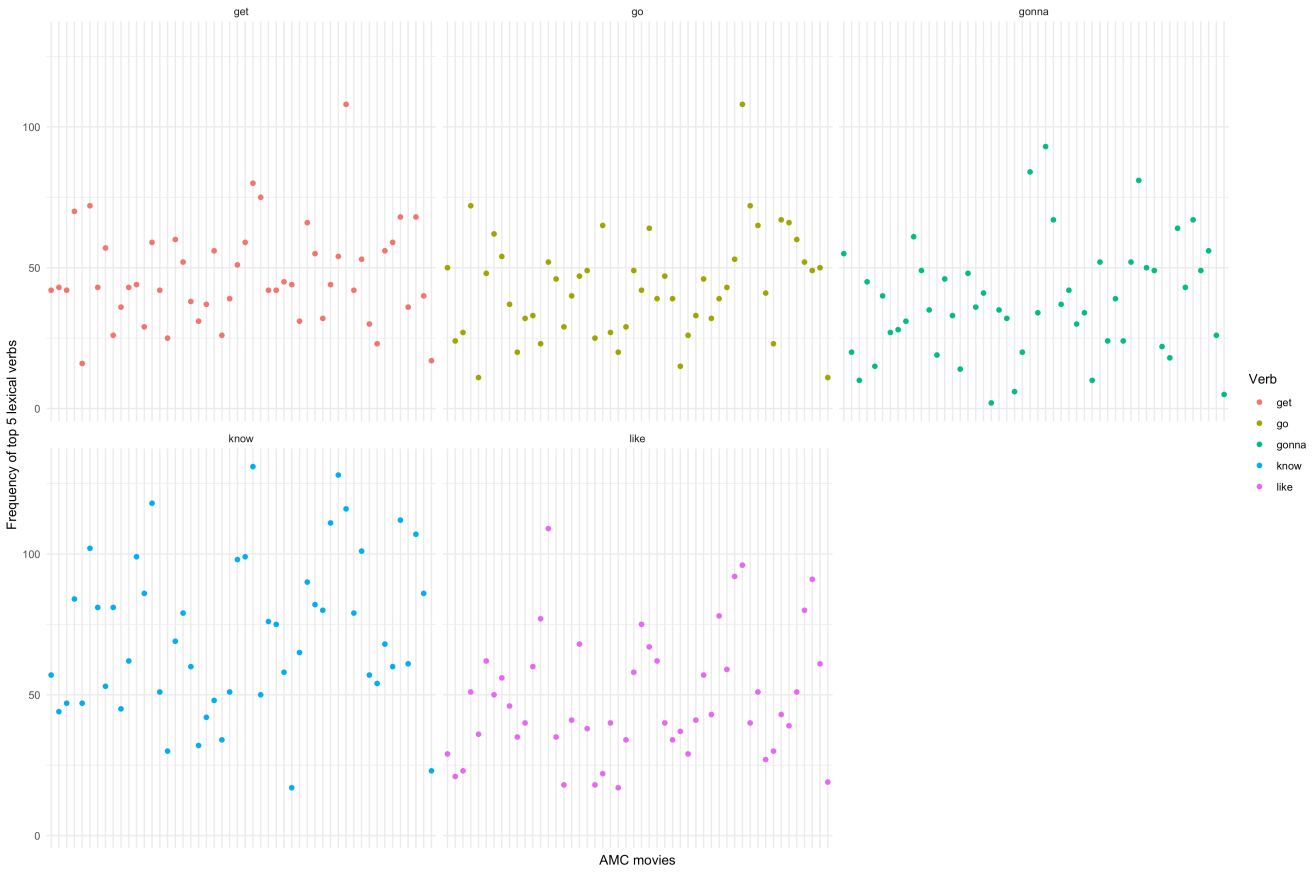
Graph 12 is a scatterplot with regression line of the top 5 most frequent lexical verbs extracted from the AMC-50. Each dot represents the frequency of a given verb in one movie of the corpus. The x axis is the length of the movies in minutes, while the y axis is the frequency of the verb. The dots at the top of the graph are outliers, that is, cases in which the frequency of the verb exceeds the average. These may or may not be considered during the analysis. The lines superimposed over the top summarize the

relationship between variables: they are called regression lines. The shaded areas around the lines are the 95% confidence intervals.



Graph 12. Scatterplot with regression line of the most frequent verbs present in the AMC-50

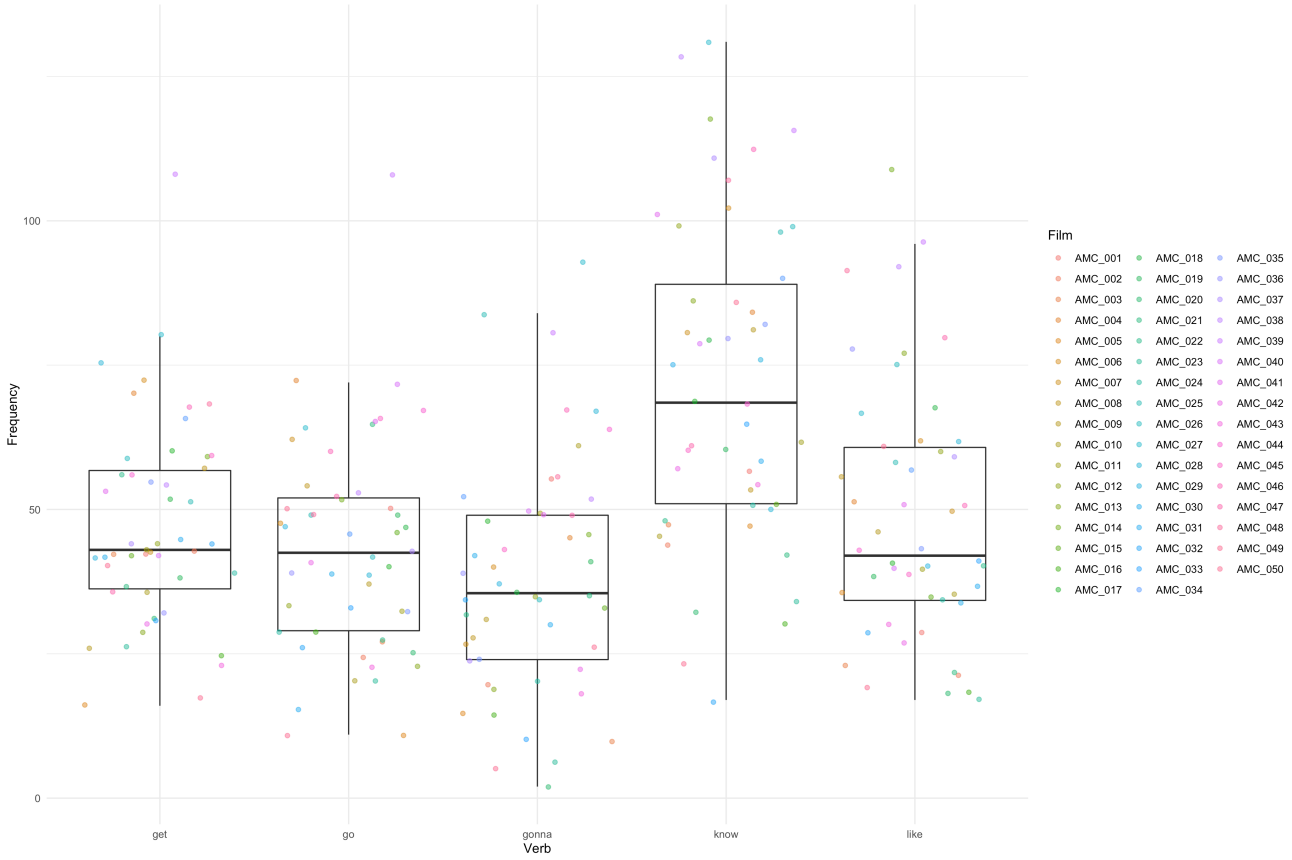
Graph 13 displays grouped scatterplots of the top 5 most frequent lexical verbs extracted from the AMC-50. Each line on the x axes corresponds to a movie, while the y axes indicate the frequency of the verb. The dots represent the verbs. The scatterplots have been grouped by verbs so as to offer a bird's-eye view of the frequent verbs and their frequency side by side.



Graph 13. Grouped scatterplots of the most frequent verbs present in the AMC-50

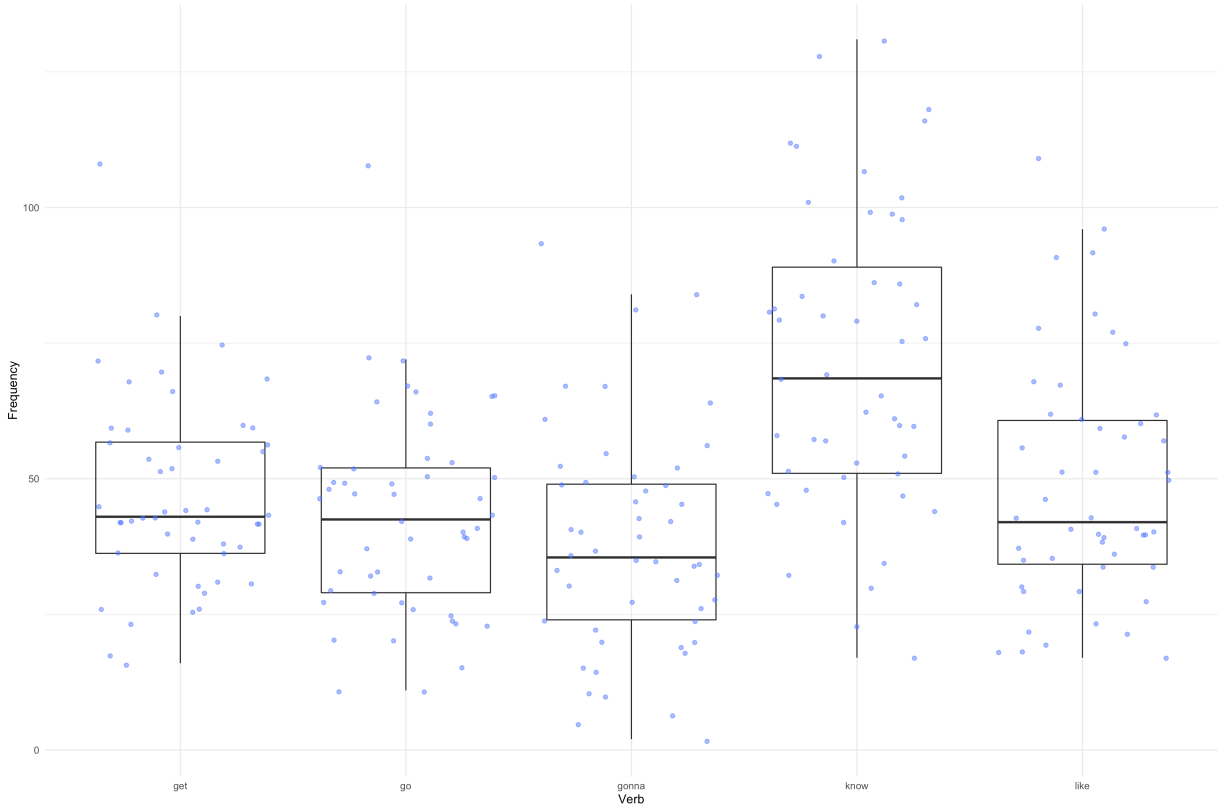
Graph 14 displays grouped boxplots for each of the 5 most frequent lexical verbs extracted from the AMC-50. On the x axis are the verbs, while on the y axis the raw frequency of each verb in each movie of the AMC-50 is displayed. Furthermore, each boxplot indicates several key values for each verb: the lowest point of the bottom whisker (or a dot below it) indicate the lowest frequency and the highest point of the whisker (or dot above it) indicates the highest frequency. For example, the boxplot of *know* shows that the highest frequency of the verb is greater than that of *gonna*. The lowest edge of the white box is the first quartile, so that the distance between the bottom of the vertical line and the lowest edge of the white box is the range between which the lowest 25% of frequencies fall. This range is larger in *know* compared to *gonna*, which means that there is more variability in the frequency of *know* compared to that of *gonna*. The box shows the interquartile range, in other words, 50% of the frequencies are bigger than the lower part of the box area, but smaller than the top part. The boxes of *gonna* and *go* are of similar size, whereas the remaining ones are different. The top edge of the box shows the value of the upper quartile, thus the distance between the top edge of the box and the top of the vertical line shows the range between which the top 25% of frequencies fall. In the middle of the box is a line that represents the median. The median for *know* is higher than the one for *gonna*. Lastly, if the whiskers are the same length, then the distribution is symmetrical; however, if the top or bottom whisker is much longer, then

the distribution is asymmetrical. The dots above the boxplots represent the outliers. The legend displays the different colors used to group the frequencies by movie.



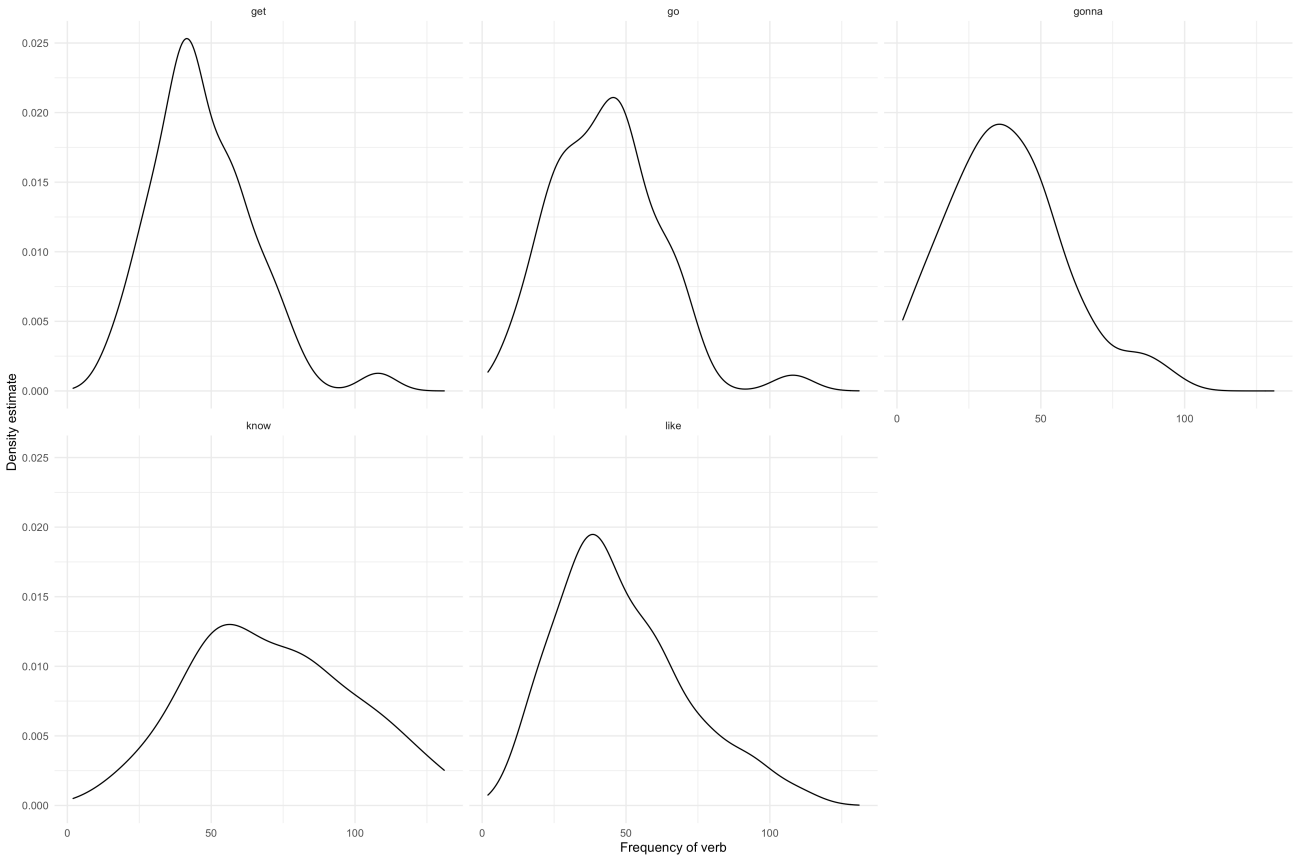
Graph 14. Grouped boxplots of the 5 most frequent verbs present in the AMC-50 according to the movies in which they occur

As opposed to the previous graph, Graph 15 does not differentiate the dots/verbs on the basis of the movie in which they occur.



Graph 15. Grouped boxplots of the 5 most frequent verbs present in the AMC-50

Graph 16 displays grouped density plots of the top 5 most frequent lexical verbs extracted from the AMC-50. The curve represents the skewness (that is, the asymmetry in the distribution) of the data, so that if the peak of the curve is found on the left, the data are right skewed; whereas, if the data peaks on the right, it is left skewed. If the density curve is left skewed, the mean is smaller than the median; if the density curve is right skewed, the mean is greater than the median. Lastly, if the curve has no skewness, the mean and the median are equal. The graph shows that all the data regarding the frequency of the top 5 verbs are right skewed and unimodal, except for *known* which appears to have a slightly more normal, and thus not skewed, distribution.



Graph 16. Grouped density plots of the 5 most frequent verbs present in the AMC-50

3.2.3 Multi-word sequences (N-Grams)

The concept of *multi-word sequences*, also called *lexical items* (Sinclair 1998, 2004)*, clusters* (Scott 1998, Scott and Tribble 2006), *lexical bundles* (Biber *et al.* 1999), *n-grams* (www.laurenceanthony.net), *sequences of words* (Hunston 2006), or *phrasal units* (Stubbs 2006), is directly linked to the Firthian notion of collocation in that it expands the category in terms of number of words involved: words do not only come in sets of two (as collocates do), but also in groups of three, four (or more) items which together create a meaning which is different from the meaning of the single items taken in isolation (Sinclair 2004). More recent studies have distinguished *lexical bundles* from *n-grams*, defining the former specifically as uninterrupted strings of three or more words which, in order to be considered as such, have to be extremely frequent in a given register (Cortes 2015)[10].

3.2.3.1 Two-Grams

Table 5 shows the 100 most frequent 2-grams in the AMC-50 processed and extracted via *AntConc 4.0.5*, *#LancsBox 6.0*, *SketchEngine* and *WordSmith Tools 8.0*.

	AntConc 4.0.5				*#LancsBox 6.0*		*SketchEngine*		*WordSmith Tools 8.0*	
RANK	TYPE	FREQ.	TYPE (with space separators)	FREQ.	TYPE	FREQ.	TYPE	FREQ.	TYPE	FREQ.
01	i m	4472	you know	1497	you know	1497	do n't	3222	are you	1288
02	it s	3381	are you	1290	are you	1288	i do	1563	in the	1222
03	don t	3223	i don	1279	in the	1219	you know	1497	do you	1129
04	you re	2294	in the	1218	do you	1127	are you	1291	come on	1055
05	that s	2010	do you	1115	come on	1053	in the	1218	this is	1054
06	you know	1500	come on	1054	this is	1053	do you	1123	no no	989
07	are you	1293	this is	1050	i don't	1051	this is	1112	all right	974
08	i don	1280	and i	1011	no no	986	come on	1054	thank you	884
09	in the	1223	all right	972	all right	971	and i	1012	of the	823
10	do you	1136	no no	944	thank you	882	no no	974	on the	795
11	come on	1055	thank you	884	i know	821	all right	974	and i	790
12	this is	1054	s a	882	of the	821	ca n't	935	have to	683
13	and i	1015	i can	869	on the	795	did n't	889	it was	582
14	no no	990	i know	821	and i	752	thank you	884	out of	567
15	all right	974	of the	821	i was	747	of the	821	have a	566
16	we re	964	on the	792	have to	683	i know	821	don't know	561
17	he s	959	i was	747	i have	667	i was	811	if you	557

[10] Source: Forchini (2021:25).

18	i ll	947	have to	683	to the	649	on the	795	did you	533
19	can t	935	t know	681	to be	638	you do	769	know what	484
20	didn t	890	i have	668	it was	581	i have	746	oh oh	469
21	s a	884	you can	663	i think	567	have to	683	what are	468
22	thank you	884	s not	650	have a	566	n't know	680	it s	466
23	i can	869	to the	648	a little	565	it was	666	but i	448
24	of the	823	to be	638	out of	565	to the	649	for the	443
25	i know	821	if you	631	if you	548	to be	638	i'm not	439
26	on the	795	it was	581	you have	548	if you	631	in a	430
27	i was	747	i think	568	did you	533	you have	580	at the	411
28	let s	743	have a	566	i just	529	i think	568	that i	408
29	what s	715	a little	565	i mean	481	have a	566	i'm sorry	394
30	i ve	684	out of	564	know what	477	a little	565	what i	390
31	have to	683	you have	546	you are	474	out of	564	oh my	389
32	t know	681	s the	533	don't know	469	i did	535	for you	382
33	you can	674	did you	528	to do	468	i i	534	and you	375
34	i have	668	i just	527	what are	468	did you	533	that you	373
35	s not	650	i i	525	i i	458	but i	531	i'm gonna	373
36	to the	649	but i	522	you can	458	i just	528	for a	368
37	to be	640	m not	520	oh oh	453	know what	500	know i	366
38	if you	632	know what	491	for the	443	i mean	481	no i	349
39	there s	629	you don	487	i'm not	438	you are	478	kind of	348
40	it was	582	i mean	481	in a	430	and you	477	is that	345
41	i think	568	and you	473	i can	429	that i	475	going to	344
42	out of	567	to do	467	i got	428	to do	468	it is	337
43	have a	566	m sorry	466	but i	427	you can	467	what do	336
44	a little	565	you are	466	i am	419	no i	456	let me	335
45	you have	549	what are	455	you to	419	what are	455	is a	334
46	that i	546	that i	444	you were	413	oh oh	453	that s	330
47	i i	545	for the	441	at the	411	i can	449	my god	324
48	but i	536	no i	438	you don't	397	n't you	449	with the	322
49	did you	536	t you	434	i'm sorry	394	i ca	445	look at	318
50	s the	536	in a	429	i can't	392	for the	441	and the	317
51	i just	529	i got	428	oh my	389	you were	441	is it	315
52	m not	520	m gonna	424	that i	385	that you	436	got a	312
53	know what	505	i am	418	for you	379	know i	433	is the	311
54	you don	500	you to	418	i'm gonna	372	in a	430	if i	311
55	what i	491	re gonna	416	to me	372	i got	428	yeah i	309

56	i mean	481	that you	415	for a	368	you to	420	it's a	307
57	and you	478	know i	412	you think	364	i am	418	gonna be	306
58	you are	477	at the	411	and you	360	at the	411	what you	304
59	she s	476	you were	411	what i	355	yeah i	398	of a	300
60	to do	468	oh my	381	that you	354	oh my	389	we have	299
61	what are	468	for you	379	kind of	348	for you	380	all the	297
62	oh oh	467	i didn	377	know i	348	what i	377	see you	295
63	m sorry	466	oh oh	376	to you	348	n't have	373	want to	289
64	you i	466	what i	376	a lot	344	to me	372	well i	282
65	know i	462	t have	372	going to	344	does n't	371	it i	270
66	no i	462	to me	371	is that	343	for a	368	that was	267
67	that you	462	for a	368	it is	337	you think	366	it's not	267
68	t you	451	you think	366	what do	336	it is	358	like a	266
69	for the	443	kind of	348	let me	334	well i	351	yeah yeah	265
70	in a	430	re not	348	is a	333	to you	349	me to	265
71	i got	428	to you	348	to get	329	kind of	348	do it	265
72	they re	428	yeah i	348	my god	323	is that	347	with you	264
73	m gonna	424	well i	346	with the	322	a lot	344	he was	264
74	you to	422	is that	345	look at	318	you i	343	aargh aargh	263
75	i am	419	a lot	344	and the	317	if i	341	tell you	260
76	re gonna	416	going to	341	no i	317	going to	341	was a	258
77	you were	412	if i	341	you got	316	what you	336	so i	257
78	at the	411	let me	335	is it	314	let me	335	don't you	256
79	yeah i	409	what you	334	got a	312	what do	334	and then	256
80	what you	405	is a	332	is the	310	is a	333	hey hey	254
81	oh my	388	it is	332	to go	310	is it	329	what is	250
82	for you	383	what do	332	if i	308	to get	329	tell me	250
83	i didn	377	to get	329	it's a	307	oh i	327	from the	248
84	t have	373	is it	327	i didn't	306	my god	324	of course	246
85	to me	372	ll be	324	gonna be	305	with the	322	right now	243
86	doesn t	371	my god	324	of a	300	look at	318	can you	243
87	for a	369	with the	321	we have	299	so i	316	be a	242
88	you think	366	look at	318	all the	296	you got	316	need to	240
89	well i	354	and the	316	see you	294	and the	316	me i	240
90	to you	351	you got	316	you do	292	got a	312	can i	239
91	it i	350	got a	312	you i	291	you you	311	one of	237
92	is that	348	to go	309	what you	290	we have	311	oh i	235
93	kind of	348	is the	308	you doing	290	to go	310	go to	234

94	re not	348	gonna be	306	you want	289	is the	308	we are	232
95	a lot	345	so i	306	want to	288	gonna be	306	of my	229
96	going to	344	of a	300	yeah i	288	of a	300	let's go	224
97	you you	344	we have	299	i do	284	i could	299	about the	221
98	if i	341	all the	297	i need	280	see you	298	is this	220
99	you ve	340	see you	295	i want	277	all the	297	with a	217
100	it is	337	ve been	291	you and	270	is n't	294	when i	217

Table 5. The AMC-50 most frequent 2-grams

3.2.3.2 Three-Grams

Table 6 shows the 100 most frequent 3-grams in the AMC-50 processed and extracted via *AntConc 4.0.5*, *#LancsBox 6.0*, *SketchEngine* and *WordSmith Tools 8.0*.

	AntConc 4.0.5				***#LancsBox 6.0***		***SketchEngine***		***WordSmith Tools 8.0***	
RANK	**TYPE**	**FREQ.**	**TYPE** (with space separators)	**FREQ.**	**TYPE**	**FREQ.**	**TYPE**	**FREQ.**	**TYPE**	**FREQ.**
01	i don t	1279	no no no	471	no no no	486	i do n't	1277	no no no	487
02	don t know	561	what are you	378	what are you	389	do n't know	560	what are you	389
03	i m not	519	oh my god	309	i don't know	372	you do n't	489	oh my god	310
04	you don t	499	what do you	283	oh my god	309	no no no	477	what do you	285
05	no no no	487	you know what	242	what do you	285	i ca n't	440	are you doing	212
06	i m sorry	465	you know i	220	you know what	240	what are you	378	oh oh oh	162
07	i can t	440	are you doing	212	are you doing	212	i did n't	375	do you think	162
08	i m gonna	423	a lot of	193	a lot of	193	oh my god	310	do you know	120
09	what are you	389	do you think	160	you know i	182	what do you	283	aargh aargh aargh	120
10	i didn t	375	i have to	148	do you think	162	do n't you	260	want you to	117
11	it s a	360	you have to	142	oh oh oh	156	you know what	244	this is the	117
12	it s not	326	i know i	140	i have to	148	you know i	222	come on come	117

13	oh my god	310	oh oh oh	124	you have to	142	do n't have	216	out of here	**115**
14	what do you	285	m sorry i	121	i have a	120	you ca n't	215	on come on	**115**
15	don t you	260	i have a	120	aargh aargh aargh	120	are you doing	212	why don't you	**114**
16	let s go	253	i think i	119	do you know	120	a lot of	193	this is a	**110**
17	you know what	248	t know what	119	i know i	120	do you think	162	get out of	**108**
18	you know i	239	do you know	117	this is the	117	oh oh oh	157	what the hell	**105**
19	you re not	229	this is the	117	want you to	115	why do n't	154	how are you	**105**
20	don t have	216	what the hell	116	out of here	115	i know i	153	don't know what	**102**
21	you can t	215	i love you	115	i love you	114	do n't want	148	out of the	**100**
22	are you doing	212	out of here	115	come on come	113	i have to	148	do you want	**92**
23	you re gonna	210	want you to	115	on come on	111	do n't think	146	do you mean	**91**
24	a lot of	193	this is a	110	this is a	110	you have to	142	mm mm mm	**90**
25	i m a	187	aargh aargh aargh	108	get out of	108	we do n't	141	this is my	**88**
26	i m just	178	get out of	108	i told you	106	you did n't	138	how do you	**87**
27	i ve been	169	i know you	108	how are you	105	n't know what	122	what is it	**85**
28	that s what	163	i told you	106	what the hell	103	i think i	121	do you have	**83**
29	do you think	162	how are you	105	i want you	100	i have a	120	know what i	**81**
30	it s just	162	come on come	104	out of the	99	aargh aargh aargh	120	don't have to	**81**
31	oh oh oh	161	t have to	104	why don't you	98	i know you	118	want me to	**79**
32	that s a	161	on come on	101	you have a	92	this is the	117	it was a	**79**
33	i know i	160	i want you	100	do you want	92	do you know	117	hey hey hey	**78**
34	we re gonna	160	out of the	99	do you mean	91	out of here	115	what did you	**75**
35	and i m	156	s gonna be	98	mm mm mm	90	i love you	115	all the time	**75**
36	you re a	156	s going on	96	to see you	89	what the hell	115	where are you	**74**

37	why don t	154	t want to	94	this is my	88	want you to	115	don't want to	**74**
38	don t want	148	do you want	92	how do you	87	come on come	113	by the way	**73**
39	i have to	148	i think it	92	i don't think	87	on come on	110	what's going on	**72**
40	don t think	146	to see you	92	to meet you	87	do n't worry	110	give me a	**70**
41	there s a	143	do you mean	91	what is it	85	this is a	110	know i know	**68**
42	you have to	142	you have a	91	do you have	83	get out of	108	i'm sorry i	**68**
43	we don t	141	this is my	88	a little bit	83	i told you	106	oh come on	**67**
44	you didn t	140	you know that	88	i know you	83	how are you	105	i'm going to	**67**
45	that s not	139	how do you	87	you know that	81	do n't wanna	105	here we go	**67**
46	that s the	136	to meet you	85	i got a	81	n't have to	104	talk to you	**66**
47	m sorry i	132	what is it	85	don't know what	80	i want you	100	listen to me	**66**
48	no i m	126	a little bit	83	it was a	79	out of the	99	have to do	**66**
49	i ll be	124	do you have	83	the rest of	79	n't want to	95	why are you	**64**
50	i think i	124	know what i	83	i don't want	78	do n't do	93	no no i	**64**
51	t know what	123	i got a	81	hey hey hey	78	do you want	92	we have a	**63**
52	i m going	122	no no i	80	i think i	78	you have a	92	one of the	**63**
53	i m so	122	m going to	79	want me to	78	did n't know	92	in front of	**63**
54	you i m	122	the rest of	79	what did you	74	i think it	92	i'm not gonna	**63**
55	do you know	120	want me to	79	all the time	74	to see you	92	are you talking	**63**
56	i have a	120	it was a	78	where are you	74	do you mean	91	wait a minute	**62**
57	it s okay	120	ll see you	78	by the way	73	mm mm mm	90	go to the	**62**
58	aargh aargh aargh	119	all the time	75	you thank you	73	this is my	88	be able to	**62**
59	i know you	118	hey hey hey	75	know what i	73	you know that	88	are you gonna	**61**
60	it s the	118	mm mm mm	75	you don't have	72	to meet you	87	we have to	**59**

61	this is the	117	you know you	75	what's going on	72	how do you	87	wait wait wait	59
62	want you to	117	where are you	74	you all right	72	no no i	86	thank you for	59
63	come on come	116	by the way	73	give me a	70	what is it	85	it s not	59
64	no it s	116	m not gonna	72	to be a	69	it was n't	84	all right i	59
65	what the hell	116	d like to	71	i need to	68	know what i	83	supposed to be	58
66	i love you	115	i mean i	71	i don’t know	68	i would n't	83	need you to	58
67	out of here	115	give me a	70	here we go	67	a little bit	83	look at me	58
68	on come on	114	what did you	70	oh come on	67	do you have	83	i'll see you	58
69	that i m	114	i think you	69	i'm going to	67	i got a	81	how you doing	58
70	what i m	114	to be a	69	listen to me	66	want me to	79	gonna have to	58
71	it s all	111	i need to	68	have to do	66	the rest of	79	thank you very	56
72	don t worry	110	here we go	67	know i know	66	i have n't	78	how did you	56
73	this is a	110	have to do	66	talk to you	66	hey hey hey	78	have to be	56
74	get out of	108	i wanted to	66	i wanted to	66	it was a	78	yeah yeah yeah	55
75	i m i	106	listen to me	66	don't have to	66	it does n't	77	thank you thank	54
76	i told you	106	talk to you	66	a couple of	65	you know you	76	my name is	54
77	know what i	106	a couple of	65	i'm sorry i	65	all the time	75	i'm so sorry	54
78	don t wanna	105	i thought you	65	you want to	65	where are you	74	if you don't	54
79	how are you	105	s what i	64	don't want to	64	by the way	73	have to go	54
80	t have to	104	you want to	64	you know you	64	i mean i	73	did you get	54
81	i ve got	103	are you talking	63	why are you	64	you thank you	72	no i don't	53
82	there s no	101	in front of	63	i'm not gonna	63	what did you	70	when i was	52
83	what s the	101	one of the	63	we have a	63	give me a	70	nice to meet	52
84	i want you	100	we have a	63	in front of	63	you all right	70	son of a	51

85	out of the	100	why are you	63	one of the	63	i think you	69	oh no no	**51**
86	that s it	99	you think you	63	are you talking	63	all right i	69	it s a	**51**
87	s gonna be	98	you want me	63	i thought you	63	to be a	69	in love with	**51**
88	all right i	97	be able to	62	you want me	63	if you do	68	are you going	**51**
89	he s a	97	go to the	62	go to the	62	i need to	68	and i don't	**51**
90	it s like	97	i got it	62	i got it	62	oh come on	67	whoa whoa whoa	**50**
91	s going on	96	s not a	62	be able to	62	here we go	67	of a bitch	**50**
92	t want to	95	wait a minute	62	wait a minute	62	do n't get	67	in the world	**50**
93	and you re	94	are you gonna	61	i mean i	61	listen to me	66	take care of	**48**
94	it it s	94	i i i	61	are you gonna	61	do n't understa nd	66	ladies and gentlemen	**48**
95	m i m	93	know i know	61	i need you	60	i wanted to	66	have no idea	**48**
96	you re the	93	re going to	61	you need to	59	talk to you	66	come on let's	**48**
97	didn t know	92	t do that	61	we have to	59	he did n't	66	there you go	**47**
98	do you want	92	i need you	60	you talking about	59	have to do	66	of my life	**47**
99	i m sure	92	know what you	60	thank you for	59	i thought you	65	do you wanna	**47**
100	i think it	92	m so sorry	60	gonna have to	58	a couple of	65	come on you	**47**

Table 6. The AMC-50 most frequent 3-grams

3.2.3.3 Four-Grams

Table 7 shows the 100 most frequent 4-grams in the AMC-50 processed and extracted via *AntConc 4.0.5*, *#LancsBox 6.0*, *SketchEngine* and *WordSmith Tools 8.0*.

	AntConc 4.0.5				*#LancsBox 6.0*		*SketchEngine*		*WordSmith Tools 80*	
RANK	**TYPE**	**FREQ.**	**TYPE (with space separators)**	**FREQ.**	**TYPE**	**FREQ.**	**TYPE**	**FREQ.**	**TYPE**	**FREQ.**
01	i don t know	441	no no no no	252	no no no no	257	i do n't know	440	no no no no	258
02	no no no no	258	what are you doing	173	what are you doing	173	no no no no	255	what are you doing	173
03	what are you doing	173	come on come on	99	come on come on	108	what are you doing	173	come on come on	112
04	i m sorry i	131	what do you mean	83	what do you mean	83	why do n't you	114	what do you mean	83
05	why don t you	114	i want you to	82	i want you to	82	do n't know what	110	oh oh oh oh	75
06	come on come on	111	you want me to	61	oh oh oh oh	71	come on come on	108	aargh aargh aargh aargh	68
07	don t know what	111	aargh aargh aargh aargh	58	aargh aargh aargh aargh	68	i do n't think	101	what do you think	59
08	i don t think	101	are you talking about	58	you want me to	61	i do n't want	95	are you talking about	58
09	i don t want	95	oh oh oh oh	58	i know i know	60	you do n't have	88	get out of here	56
10	i m i m	91	what do you think	58	what do you think	59	what do you mean	83	thank you thank you	54
11	what s going on	88	i know i know	57	i don't know what	59	i want you to	82	what are you talking	53
12	you don t have	88	you know what i	57	are you talking about	58	do n't have to	81	thank you very much	53
13	i want you to	83	get out of here	56	you know what i	57	do n't want to	74	nice to meet you	52
14	what do you mean	83	thank you very much	53	get out of here	56	oh oh oh oh	72	son of a bitch	49
15	don t have to	81	what are you talking	53	thank you very much	53	aargh aargh aargh aargh	68	mm mm mm mm	48
16	you know what i	80	nice to meet you	52	what are you talking	53	i do n't wanna	68	in the middle of	40
17	i m going to	79	son of a bitch	49	nice to meet you	52	i do n't have	65	all right all right	40
18	oh oh oh oh	75	thank you thank you	43	thank you thank you	50	you do n't know	65	what do you want	38

19	don t want to	74	i need you to	42	son of a bitch	49	you want me to	61	good to see you	36
20	i m not gonna	72	mm mm mm mm	41	mm mm mm mm	48	i know i know	60	are you doing here	35
21	i ll see you	69	in the middle of	40	you don't have to	42	i i do n't	59	on come on come	32
22	aargh aargh aargh aargh	68	what do you want	38	i don't want to	42	i did n't know	59	hey hey hey hey	31
23	i don t wanna	68	good to see you	37	i need you to	42	are you talking about	58	thank you so much	30
24	it s it s	66	i just wanted to	36	in the middle of	40	what do you think	58	oh no no no	29
25	i don t have	65	are you doing here	35	all right all right	39	you know what i	57	where are you going	28
26	you don t know	65	do you think you	32	what do you want	38	get out of here	56	what the hell is	28
27	i know i know	62	on come on come	31	i just wanted to	36	if you do n't	54	come on let's go	28
28	you want me to	61	hey hey hey hey	30	good to see you	35	thank you very much	53	ow ow ow ow	27
29	i i don t	60	no no no i	30	are you doing here	35	what are you talking	53	oh my god oh	27
30	i m so sorry	60	thank you so much	30	on come on come	32	nice to meet you	52	wait wait wait wait	26
31	i didn t know	59	the hell are you	30	hey hey hey hey	31	no i do n't	52	how do you know	25
32	what do you think	59	oh no no no	28	the hell are you	30	thank you thank you	52	god oh my god	25
33	are you talking about	58	what the hell is	28	thank you so much	30	and i do n't	51	cat who's the cat	25
34	get out of here	56	where are you going	28	oh no no no	29	son of a bitch	49	would you like to	24
35	i think it s	56	all right all right	27	where are you going	28	mm mm mm mm	48	let me tell you	24
36	if you don t	54	ow ow ow ow	27	come on let's go	28	i do n't care	47	do you want to	24
37	thank you thank you	54	the rest of the	27	what the hell is	28	i ca n't i	47	do you want me	24
38	that s what i	54	i have no idea	26	ow ow ow ow	27	i ca n't believe	42	do you know what	24
39	come on let s	53	i have to go	26	the rest of the	27	i need you to	42	whoa whoa whoa whoa	23
40	no i don t	53	re gonna have to	26	i don't know how	27	in the middle of	40	oh my god i	23
41	thank you very much	53	how do you know	25	i don't know i	27	do n't have a	40	no no no i	23

42	what are you talking	53	the end of the	25	i have no idea	26	do n't know how	40	do you think you	23
43	nice to meet you	52	wait wait wait wait	25	i have to go	26	we do n't have	40	are you gonna do	23
44	and i don t	51	a lot of people	24	wait wait wait wait	25	what do you want	38	are you all right	23
45	it s gonna be	51	do you want me	24	oh my god oh	25	do n't know i	37	what do you say	22
46	son of a bitch	49	i thought you were	24	how do you know	25	do n't worry about	37	what are you gonna	22
47	i m gonna go	48	let me tell you	24	cat who's the cat	25	do n't think i	37	what the hell are	21
48	mm mm mm mm	48	would you like to	24	god oh my god	25	good to see you	37	my god oh my	21
49	i can t i	47	are you gonna do	23	i can't i can't	25	but i do n't	36	am i supposed to	21
50	i don t care	47	do you want to	23	the end of the	25	i just wanted to	36	where are we going	20
51	i d like to	44	to talk to you	23	the cat who's the	25	are you doing here	35	i'm a girl i'm	20
52	i can t believe	42	whoa whoa whoa whoa	23	do you want to	24	do n't know if	35	go go go go	20
53	i need you to	42	are you all right	22	do you want me	24	i do n't understand	34	girl i'm a girl	20
54	m sorry i m	42	do you know what	22	i thought you were	24	do n't you think	34	don't worry about it	20
55	no i m not	42	oh my god i	22	a lot of people	24	i just do n't	33	can i help you	20
56	all right all right	40	what are you gonna	22	let me tell you	24	on come on come	32	who's the cat who's	19
57	don t have a	40	what do you say	22	do you know what	24	i do n't like	32	what did you do	19
58	don t know how	40	am i supposed to	21	would you like to	24	oh my god i	32	talk to you about	19
59	don t know i	40	god oh my god	21	are you all right	23	do you think you	32	in the first place	19
60	i m not a	40	oh my god oh	21	are you gonna do	23	i do n't really	31	got a lot of	19
61	in the middle of	40	t worry about it	21	to talk to you	23	hey hey hey hey	31	for the rest of	19
62	we don t have	40	the middle of the	21	i don't know if	23	thank you so much	30	do you have any	19
63	i m trying to	38	what the hell are	21	whoa whoa whoa whoa	23	no no no i	30	stop it stop it	18
64	what do you want	38	can i help you	20	what do you say	22	i do n't believe	30	oh oh my god	18
65	but i don t	37	do you think i	20	do you think you	22	know i do n't	30	let's go let's go	18

66	don t think i	37	go go go go	20	what are you gonna	22	i do n't i	30	in love with you	18
67	don t worry about	37	m a girl i	20	a girl i'm a	21	the hell are you	30	all right come on	18
68	good to see you	37	where are we going	20	the middle of the	21	do n't do that	29	yeah yeah yeah yeah	17
69	it s all right	37	you know i don	20	what the hell are	21	do n't even know	29	why don't you just	17
70	i just wanted to	36	do you have any	19	oh my god i	21	n't i ca n't	29	run run run run	17
71	all right let s	35	for the rest of	19	my god oh my	21	oh no no no	29	oh my god what	17
72	are you doing here	35	got a lot of	19	am i supposed to	21	well i do n't	28	nothing to do with	17
73	do you think you	35	in the first place	19	no no no i	21	where are you going	28	know i know i	17
74	don t know if	35	my god oh my	19	go go go go	20	what the hell is	28	how did you get	17
75	i know it s	35	t know what you	19	where are we going	20	do n't do n't	28	get out of the	17
76	that s that s	35	talk to you about	19	can i help you	20	ow ow ow ow	27	do you think i	17
77	don t you think	34	what did you do	19	i'm a girl i'm	20	all right all right	27	bomb bomb bomb bomb	17
78	i don t understand	34	you mind if i	19	girl i'm a girl	20	i did n't mean	27	at the end of	17
79	it s just a	34	i know what you	18	do you have any	19	the rest of the	27	are you going to	17
80	no it s not	34	i think we should	18	in the first place	19	ca n't i ca	27	and i don't know	17
81	oh my god i	34	t have to be	18	who's the cat who's	19	i have to go	26	what s going on	16
82	we re going to	34	t know how to	18	for the rest of	19	why do n't we	26	this is gonna be	16
83	i just don t	33	the hell out of	18	talk to you about	19	you do n't want	26	let's go come on	16
84	i know i m	33	at the end of	17	what did you do	19	n't i do n't	26	know what it is	16
85	i m talking about	33	bomb bomb bomb bomb	17	got a lot of	19	i thought you were	26	know what i mean	16
86	what s the matter	33	get out of the	17	don't worry about it	18	i do n't even	26	how do you do	16
87	i don t like	32	how did you get	17	you know i don't	18	oh my god oh	26	hell are you doing	16
88	i m just gonna	32	in love with you	17	the hell out of	18	i have no idea	26	do you have a	16
89	on come on come	32	know i know i	17	stop it stop it	18	i ca n't do	25	do me a favor	16

90	on let s go	32	m gonna have to	17	let's go let's go	18	do n't i do	25	alex alex alex alex	16
91	hey hey hey hey	31	nothing to do with	17	you mind if i	18	how do you know	25	why don't you go	15
92	i don t really	31	oh my god you	17	i think we should	18	wait wait wait wait	25	what what are you	15
93	i m not sure	31	run run run run	17	i don't think so	17	the end of the	25	what do you do	15
94	no no no i	31	stop it stop it	17	at the end of	17	a lot of people	24	what can i do	15
95	you know i m	31	t want you to	17	are you going to	17	let me tell you	24	we don't have to	15
96	you re not gonna	31	you have to do	17	nothing to do with	17	do n't think so	24	want you to know	15
97	i don t believe	30	alex alex alex alex	16	run run run run	17	do you want me	24	thank you for your	15
98	i don t i	30	are you going to	16	get out of the	17	god oh my god	24	so what are you	15
99	i m gonna be	30	do me a favor	16	all right come on	17	would you like to	24	let me ask you	15
100	know i don t	30	do you have a	16	in love with you	17	i i ca n't	23	know what to do	15

Table 7. The AMC-50 most frequent 3-grams

3.3 COLLOCATIONS AND CONCORDANCES

Chronologically, the notion of ***collocation*** was first introduced by Palmer (1933) who defined collocation as a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts and, some years later, by Firth (1957b:14) who defined it as "actual words in habitual company". Firth (1951, 1957a) particularly emphasized the habituality which distinguishes collocation and the limited possibility of co-occurrence of words, or, in Sinclairian modern terms, the *phraseological tendency* of language: "One of the meanings of *ass* is its habitual collocation with an immediately preceding *you silly*, and with other phrases of address or of personal reference. […] There are only limited possibilities of collocation with preceding adjectives, among which the commonest are *silly*, *obstinate*, *stupid*, *awful*, occasionally *egregious*" (Firth 1957a:195). Other scholars then gave a slightly different definition of *collocation*: Leech (1974), for example, pointed out the psychological association "a word acquires on account of the meanings of words which tend to occur in its environment" (Leech 1974:20). Sinclair (1991:170), instead, emphasized the textual trait of collocation, i.e. "the occurrence of two or more words within a short space of each other in a text". Both Hoey (1991) and Stubbs (2001) highlighted its statistical aspect, namely the chance of relationship that "a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991:6-7), or, simply, "frequent co-occurrence" (Stubbs 2001:29). However, despite these different slants (i.e. contextual, psychological, textual, and statistical), what remains at the basis of the notion of collocation is the Firthian intuition that the meaning created by the co-occurrence of two items in a given context is a product of those two co-occurring words in that particular context, or in Hallidayan terms, "of the relationship between the system and its environment" (Halliday 2003:196, cf. also Halliday 1985). An example of this creation of new meaning, provided by Sinclair (1998, 2004:135), is the use of the adjective *white* which, when followed by the noun *wine*, implies a different color range from when it is used in isolation.[11] Key Word In Context (henceforth KWIC) is the most common format for ***concordance lines***.

[11] Source: Forchini (2021:25).

3.3.1 Verbs[12]

Verb	Rfreq	Nfreq	Coll	Rfreq	FreqL	FreqR	Distr	T-score	KWIC	MOVIE
know	3586	6,405,196	i	2544	1551	993	50	32.658	right i don't actually have a paper cut i know that martin you're a very bad actor wait	Love, Simon
			you	2472	1832	640	50	31.004	eat or anything dying sucks gus really loved you you know? i know he wouldn't shut up about	The Fault in our Stars
			don	722	616	106	50	22.255	when he's hitting on chicks wait don't i know you? from instagram what's your handle? it's	When We First Met
			t	921	785	136	50	21.226		
			what	590	33	557	50	15.379	at butthead? say hi to your mom for me i know what you're gonna say son and you're	Back to the Future
			how	240	72	168	48	11.138		
			didn	138	102	36	40	8.829		
			do	238	175	63	47	6.940		
			know	230	115	115	38	6.079		
like	2426	4,333,242	it	501	313	188	50	10.058	mewhere along the line and you gotta fix it like a chiropractor you gotta give it a chiropractic a	Silver Linings Playbook
			d	124	114	10	45	9.223	yeah we've already discussed this yeah i'd like to discuss it again you have responsibilities her	The Proposal
			looks	93	90	3	40	9.210	ents they're going to trial jesus they look like the secret service intimidation let the games begi	Erin Brockovich
			a	476	105	371	50	8.621	sponse time for mapping an environment sonar just like a submarine mister wayne like a subn	The Dark Knight
			would	117	106	11	43	8.497	show come on oh you okay? yeah totally i would like to propose a toast to your mother to me?	It's Complicated
			feel	84	82	2	30	8.201		
			look	116	105	11	45	7.857		
			i	796	463	333	50	6.709		
			just	171	148	23	46	6.555		
get	2348	4,193,921	to	626	374	252	50	13.944	you'd mind if i bring the boys over to get a closer looksie? a closer looksie? yeah to see	Captain Marvel
			out	288	32	256	47	13.909	for a moment? get me out of here we gotta get out of here alex help what are you doing?	Madagascar
			back	134	16	118	42	9.263	recchio's waiting you're not here anyway i gotta get back to work so uh bye ok kitty it'	Catwoman
			let	125	115	10	40	8.014	no that's fine sure thank you let me just get you a blanket in the cupboard on top of	The Holiday
			here	160	39	121	47	7.793	talk about boys are you being safe? how did you get in here? the window do you do that a	Twilight
			can	176	168	8	47	7.728		
			ll	128	112	16	43	7.352		
			gonna	123	122	1	41	6.780		
			we	231	194	37	46	6.523		
go	2153	3,845,618	go	466	233	233	37	19.286	i know it's the craziest really excited yeah go go go oh my god what? when my mom finds	The Holiday
			let	413	357	56	48	18.725	you and the only way that i would ever let go of my bag would be if you came over	Meet the Parents
			to	668	341	327	50	16.014	tside that will take you anywhere you'd like to go you must be the famous pepper potts indee	Iron Man I
			gotta	106	95	11	38	9.132	it's an early printing um look listen i gotta go but you just got here i know but i	Finding Forrester
			we	265	205	60	48	8.852	rie that you made? okay thank you right here we go right all right yeah yeah good going dale	The Judge
			here	164	136	28	40	8.408		
			ahead	73	3	70	32	8.279		
			on	218	71	147	46	8.159		
			s	420	316	104	48	7.457		
gonna	1900	3,393,718	m	566	551	15	49	19.900	do this again? well it's a fair i'm gonna try my luck as who? steve from ohio? they'	Captain America The First Avenger
			re	520	506	14	50	19.487	with you he's waiting for your call you're gonna be spending the next few years of your life	Gifted
			be	324	8	316	45	14.734	hear? now they told us that vietnam was gonna be very different from the united states of ar	Forrest Gump
			i	744	636	108	49	9.856	ome hey you fred i was wondering if i was gonna see you tonight what and disappoint all my f	Me Myself & Irene
			we	265	235	30	49	9.724	talk to you again stop it look we're just gonna find the people get checked in and have this	Madagascar
			not	191	187	4	46	9.100		
			are	165	162	3	48	7.800		
			s	374	357	17	48	7.147		
			get	123	1	122	41	6.780		
come	1857	3,316,913	on	1379	198	1181	50	34.870	at? i'm going back to the restaurant hal hal come on hey // hal // you never called me back wh	Shallow Hal
			come	362	181	181	43	17.084	e you tonight yeah woo yes here we go woo come come come oh thank you honey good job th	The Lucky One
			here	249	68	181	48	12.701	ic dragon i knew the little crack was lying jinxy come here come here little jinxy oh // jinxy cat,	Meet the Parents
			back	106	12	94	43	8.239	uld do another set of tests the results could come back different her brain has begun to contra	Midnight Sun
			let	113	28	85	42	7.996	you than they do on me watch out daphne sugar come on let's play ball ok let's go	Some Like It Hot
			oh	177	120	57	42	6.915		
			up	110	37	73	40	6.621		
			hey	92	60	32	32	5.737		
			in	177	38	139	45	5.575		
got	1716	3,065,063	ve	243	235	8	48	14.034	t want to rush you out of here we've got a committee meeting well e- excuse me charles with	Philadelphia
			i	788	640	148	50	12.783	list do you want some company? love some hey i got you the best drink in town but i didn'	The Holiday
			a	399	32	367	49	9.779	ctivity stay back nice scuba suit lighten up honey huh? got a smile for me? freak all life on eart	Captain Marvel
			we	237	211	26	43	9.135	deck we just got a distress call from a rig due west of	Man of Steel
			you	639	476	163	50	7.662	re one gutsy racer oh hey mister the king you got more talent in one lug nut than a lot	Cars
			got	104	52	52	25	7.104		
			he	139	107	32	37	6.363		
			it	306	76	230	49	6.337		
			some	42	5	37	26	4.046		

[12] Collocations and KWIC concordances of the most frequent verbs: the first column contains the type of verb – we have included the most frequent lexical verbs as extracted from the top 100 frequent items in the wordlist. The verb is followed by its raw (Rfreq) and normalized frequency (Nfreq). The raw frequency is the number of times the verb occurs in the corpus, while the normalized frequency is the number of times the verb would occur in a corpus of 1 million words. The normalization is needed to compare data between corpora of different sizes. Column D contains the top ten most frequent collocates of the verb; these were selected by adjusting the search parameters. The collocational span (number of words to the left and right of the node, i.e. the verb) was set to ± 3 (that is three words to the left and three words to the right), the distribution (i.e. the minimum number of texts the verb has to occur in) and the minimum frequency to 5. In addition, since collocations can be identified according to different statistical tests, the choice was made to use the *t*-score as indicator of collocational strength. A t-score equal or greater than 3 was considered the cut-off points for collocates (Hunston 2002). Lastly, the collocate column is followed by the raw frequency of the collocation and its left and right frequency; the last two columns before the text strings are the distribution and the t-score result. We have decided to include some context to showcase the verb with its collocate in a sentence in the hope of helping the interpretation of the data.

let	1407	2,513,137	s	851	98	753	50	23.187	next missus joy of cooking maybe mm all right well let's try let's just give it a woo	Julie & Julia
			go	413	56	357	48	18.725	i don't yeah let's do it come on let's go it'll be fun it'll be	Love, Simon
			me	387	44	343	50	15.905	ow we're saving lives what? uh what? what? what? let me just recap this for you real quick w	The Internship
			let	128	64	64	26	9.438	ou saying? it's the truth what are you saying? let let him speak well what's the problem here?	Runaway Jury
			get	125	10	115	40	8.014	but you can't stay here so till next time let's get the hell out of here shit i'	The Blues Brothers
			come	113	85	28	42	7.996		
			see	91	10	81	30	7.537		
			him	91	22	69	29	7.186		
			just	111	59	52	37	5.841		
think	1400	2,500,634	i	1109	786	323	50	22.788	le jailbird joey? he's your brother mom yeah i think it's a major embarrassment having an uncl	Back to the Future
			you	776	511	265	50	14.815	ather you can predict the price of heating oil i think you asked me that because you think the fi	The Social Network
			do	217	190	27	47	11.261	you think we do all day? that just sit around eating	The Amazing Spiderman
			don	196	186	10	48	10.542	at miracles happen every day some people don't think so but they do hey dummy are you ret	Forrest Gump
			t	253	234	19	49	9.111		
			should	63	7	56	37	6.841		
			about	92	8	84	39	6.236		
			it	249	33	216	49	5.691		
			what	142	117	25	49	4.825		
look	1207	2,155,904	at	359	29	330	49	17.699	yes you died what? i didn't die i mean look at me i'm right here you died but	Catwoman
			look	158	79	79	27	11.328	on no come on come on here i come look look look look it's gloria it's gloria oh	Madagascar
			like	116	11	105	45	7.857	of the cloud bank you need to see what i look like in the sunlight this is why we don'	Twilight
			me	133	56	77	40	6.020	wish that one time she would use it to look at me yeah good job good job <unintelligible> mirar	Wonder
			oh	113	73	40	35	5.433	us a present oh isn't that nice? oh oh look at this it's a flower pot with the	Meet the Parents
			this	108	27	81	40	4.399		
			good	52	19	33	25	4.235		
			you	397	229	168	49	4.205		
			hey	55	46	9	23	4.175		
take	1010	1,804,029	care	68	1	67	33	7.976	yeah ok oh oh let's go oh can you take care of that? right lou give me a milk	Back to the Future
			to	224	157	67	49	7.002	ke painkillers 'cause they made me too groggy to take care of my kids matthew's eight and ka	Erin Brockovich
			it	217	70	147	46	6.934	't move i'm not moving you want something? take it although the guns are all fake because th	Iron Man I
			ll	81	78	3	41	6.858	tell him to call back right? no no i'll take it hello? // yes // you want to bring who to	Julie & Julia
			take	64	32	32	13	6.633	minute all right? all right all right all right take take it out of my pay all right ? you don'	The Lincoln Lawyer
			gonna	79	79	0	34	6.574		
			off	53	7	46	29	6.440		
			a	213	19	194	50	6.381		
			easy	34	6	28	16	5.564		
want	970	1,732,582	to	597	16	581	50	19.749	fit you in may i remind you i don't want to see you you want to see me? yeah	It's Complicated
			you	619	392	227	50	14.764	ways to get to the courthouse using the doors i want you to learn how to stay in the substrate	The Adjustment Bureau
			i	538	460	78	50	12.738	come on you gotta let me know the brand i want to be sneaker brothers oh i see more phones	Love, Simon
			don	163	156	7	46	10.140	fit you in may i remind you i don't want to see you you want to see me? yeah	Julie & Julia
			do	163	122	41	43	9.994	best there is you're in cooperstown what do you want? nothing i got a duplex now i got wall-	Ocean's Eleven
			t	214	207	7	48	9.510		
			me	119	16	103	44	6.226		
			if	54	49	5	33	4.909		
			what	107	88	19	39	4.685		
thank	931	1,662,922	you	1224	201	1023	50	28.081	m sorry great thank you is there anything else? no thank you thank you bye bye i think it's	The Circle
			thank	152	76	76	28	11.576		
			much	93	6	87	30	9.100	go i have so much to do get out hey thank you so much hey man how you doing? are	Julie & Julia
			very	76	16	60	24	8.011	ove to you during the course of the trial thank you very much your honor mister cable? thank	Runaway Jury
			oh	126	86	40	39	7.429	than i pictured it thank you shall we cruise? oh thank you dear i'd love to no no no	Cars
			for	99	29	70	38	6.311		
			okay	71	50	21	28	5.813		
			god	39	7	32	22	5.170		
			well	48	34	14	26	4.266		
say	917	1,637,915	to	282	202	80	49	10.348	say to me? i said do you have something to say to me? i'll take it from here hey	Love, Simon
			i	387	223	164	49	8.015	and i had my foot out like this and i say i will devote my entire reign as miss wichita	Erin Brockovich
			did	80	71	9	31	7.404	shit creek aargh i beg your pardon what did you say? i i offered to help you mm you refused	The Blues Brothers
			what	139	106	33	45	7.095	" honey how about your platelets? what did she say? oh even my platelets look good great n	Philadelphia
			you	350	218	132	49	5.989	t tell him i said that what did you just say? it's moving faster than any of us ever	The Social Network
			t	135	109	26	42	5.526		
			anything	38	9	29	22	5.437		
			d	42	39	3	25	5.239		
			say	42	21	21	16	5.090		

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas, and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Cortes, Viviana. 2015. "Situating lexical bundles in the formulaic language spectrum. Origins and functional analysis developments." In *Corpus-based research in applied linguistics. Studies in honor of Doug Biber*, edited by Viviana Cortes, and Eniko Csomay, 197-216. Amsterdam/Philadelphia: John Benjamins.

Firth, John Rupert. 1951. "General linguistics and descriptive grammar." *Transactions of the Philological Society* 216-228.

Firth, John Rupert. 1957a. *Papers in linguistics 1934-1951*. London: Oxford University Press.

Firth, John Rupert. 1957b. "A synopsis of linguistic theory, 1930-1955." In *Studies in Linguistic Analysis. Special volume of the Philological Society*, edited by John Rupert Firth et al., 1-32. Oxford: Blackwell.

Forchini, Pierfranca. 2013. "Using movie corpora to explore spoken American English. Evidence from Multi-Dimensional Analysis." In *Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistics*, edited by Julia Bamford, Silvia Cavalieri, and Giuliana Diani, 123-136. Amsterdam / Philadelphia: John Benjamins.

Forchini, Pierfranca (Ed.). 2021. *The American Movie Corpus: A Tool for the Developent of Spoken Lexico-Grammatical Competence*. Milano: EDUCatt.

Gries, Stefan Th. 2009. Statistics for Linguistics with R. A practical introduction. Berlin, New Your: De Gruyter.

Halliday, Michael Alexander Kirkwood. 2003 (first printed in 1985). Systemic background. In *On language and linguistics*, ed. Jonathan J. Webster, 185-198. London / New York: Continuum.

Halliday, Michael A. K. 1985c. Systemic background. In *Systemic perspectives on discourse, Vol. 1. Selected theoretical papers from the 9th International Systemic Workshop*, edited by James D. Benson, and William S. Greaves, (1): 1-15. Norwood / New Jersey: Ablex Publishing Corporation.

Hoey, Michael. 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.

Hunston, Susan. 2006. "Phraseology and system: A contribution to the debate." In *System and Corpus: Exploring Connections,* edited by Susan Hunston, and Geoff Thompson, 55-80. London: Equinox Publishing.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Leech, Geoffrey. 1974. *Semantics*. Harmondsworth: Penguin.

Levshina, Natalia. 2015. How to do Linguistics with R. Data exploration and statistical analysis. Amsterdam / Philadelphia: John Benjamins,

Palmer, Harold E. 1933. *Second interim report on English collocations: Submitted to the Tenth Annual Conference of English teachers*. Tokyo: Institute for Research in English Teaching.

RStudio Team 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL *http://www.rstudio.com/*.

Sinclair, John McHardy. 1991. *Corpus concordance collocation*. Oxford: Oxford University Press.

Sinclair, John McHardy. 1998. "The lexical item." In *Contrastive lexical semantics*, edited by Edda Weigand, 1-24. Amsterdam / Philadelphia: John Benjamins.

Sinclair, John McHardy. 2004b. *Trust the text: Language, corpus and discourse*. London / New York: Routledge.

Scott, Mike. 1998. *WordSmith Tools*. Oxford: Oxford University.

Scott, Mike, and Chris Tribble. 2006. *Textual patterns*. Amsterdam / Philadelphia: John Benjamins.

Stubbs, Michael. 2001. *Words and phrases: Corpus studies in lexical semantics*. Oxford / Massachusetts: Blackwell.

Stubbs, Michael. 2006. *Quantitative data on multi-word sequences in English: The case of prepositional phrases*. Paper presented at the Berlin-Brandenburgische Akademie der Wissenschaften, 3rd November 2006, Berlin, Germany.