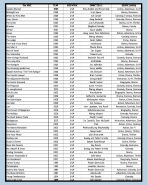
THE AMC-50 DATA QUICK REFERENCE

The **DATA REFERENCE** contains MDA DATA from previous studies - see *Opening Credits: Multi-Dimensional Analysis (MDA) DATA* - which uncover the linguistic and textual features of movie discourse and its similarity with face-to-face conversation and NEW 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 1 here). The AMC-50 DATA 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).



The AMC	YEAR	COUNTRY	DIRECTOR(S)	GENRE (IMDb)
Captain Marvel	2019	USA	Anna Boden and Ryan Fleck	Action, Adventure, Sci-Fi
Midnight Sun	2018	USA	Scott Speer	Drama, Romance
When we First Met	2018	USA	Ari Sandel	Comedy, Fantasy, Romance
Love, Simon	2018	USA	Greg Berlanti	Comedy, Drama, Romance
The Circle	2017	USA	James Ponsoldt	Drama, Sci-Fi, Thriller
Wonder	2017	USA	Stephen Chbosky	Drama, Family
Gifted	2017	USA	Marc Webb	Drama
Nerve	2016	USA	Henry Joost, Ariel Schulman	Action, Adventure, Crime
The Intern	2015	USA	Nancy Meyers	Comedy, Drama
The Judge	2014	USA	David Dobkin	Crime, Drama
The Fault in our Stars	2014	USA	Josh Boone	Drama, Romance
Iron Man 3	2013	USA	Shane Black	Action, Adventure, Sci-Fi
Man of Steel	2013	USA	Zack Snyder	Action, Adventure, Sci-Fi
The Internship	2013	USA	Shawn Levy	Comedy
Silver Linings Playbook	2012	USA	David O. Russell	Comedy, Drama, Romance
The Lucky One	2012	USA	Scott Hicks	Drama, Romance
The Avengers	2012	USA	Joss Whedon	Action, Adventure, Sci-Fi
The Amazing Spiderman	2012	USA	Marc Webb	Action, Adventure, Sci-Fi
Captain America: The First Avenger	2011	USA	Joe Johnston	Action, Adventure, Sci-Fi
The Lincoln Lawyer	2011	USA	Brad Furman	Crime, Drama, Thriller
The Adjustment Bureau	2011	USA	George Nolfi	Romance, Sci-Fi, Thriller
The Social Network	2010	USA	David Fincher	Biography, Drama
The Proposal	2009	USA	Anne Fletcher	Comedy, Drama, Romance
It's complicated	2009	USA	Nancy Meyers	Comedy, Drama, Romance
Julie & Julia	2009	USA	Nora Ephron	Biography, Drama, Romance
Twilight	2008	USA	Catherine Hardwicke	Drama, Fantasy, Romance
The Dark Knight	2008	USA	Christopher Nolan	Action, Crime, Drama
Iron Man	2008	USA	Jon Favreau	Action, Adventure, Sci-Fi
Cars	2006	USA	John Lasseter / Joe Ranft	Animation, Comedy, Family
The Pursuit of Happiness	2006	USA	Gabriele Muccino	Biography, Drama
The Holiday	2006	USA	Nancy Meyers	Comedy, Romance
The Devil Wears Prada	2006	USA	David Frankel	Comedy, Drama
Madagascar	2005	USA	Eric Darnell / Tom McGrath	Animation, Adventure, Comedy
Catwoman	2004	USA	Pitof	Action, Crime, Fantasy
The Matrix Reloaded	2003	USA	Andy and Larry Wachowsky	Action, Sci-Fi
Runaway Jury	2003	USA	Gary Fleder	Crime, Drama, Thriller
One Hour Photo	2002	USA	Mark Romanek	Drama, Thriller
Shallow Hal	2001	USA	Bobby and Peter Farrelly	Comedy, Drama, Fantasy
Ocean's Eleven	2001	USA	Steven Soderbergh	Crime, Thriller
Meet the Parents	2000	USA	Jay Roach	Comedy, Romance
Me, Myself & Irene	2000	USA	Bobby and Peter Farrelly	Comedy
Finding Forrester	2000	USA	Gus Van Sant	Drama
Mission: Impossible II	2000	USA	John Woo	Action, Adventure, Thriller
Erin Brockovich	2000	USA	Steven Soderbergh	Biography, Drama
Forrest Gump	1994	USA	Robert Zemeckis	Drama, Romance
Philadelphia	1993	USA	Jonathan Demme	Drama
Back to the Future	1985	USA	Robert Zemeckis	Adventure, Comedy, Sci-Fi
The Blues Brothers	1980	USA	John Landis	Adventure, Comedy, Crime
Young Frankenstein	1974	USA	Mel Brooks	Comedy
Some Like It Hot	1959	USA	Billy Wilder	Comedy, Music, Romance

Table 1. The AMC-50 movies

THE AMC-50 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.

SOFTWARE	TOKENS	TYPES
AntConc 4.0.5	≈ 560000	18216
#LancsBox 6.0	≈ 530000	20477
SketchEngine	≈ 580000	19950

CLAPPERBOARD

LEGAL DISCLAIMER:

The **A**merican **M**ovie **C**orpus (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. 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 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;

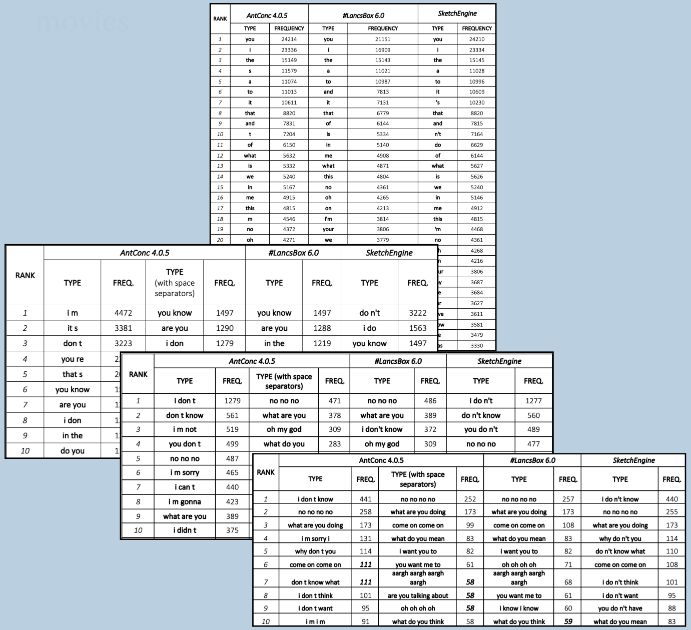
For LANGUAGE TEACHING: authoritative scholars have been emphasizing the crucial role played by spoken language in communication for almost a hundred years. In spite of this, attention given to the study of lexico-grammatical spoken features in educational settings has been scarce. The textual and linguistic 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.

WORD LISTS & MULTI-WORD SEQUENCES

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).

Table 1. The AMC-50
movies



RANK	*AntConc 4.0.5*		*#LancsBox 6.0*		*SketchEngine*	
	TYPE	FREQUENCY	TYPE	FREQUENCY	TYPE	FREQUENCY
1	you	24214	you	21151	you	24210
2	i	23336	i	16909	i	23334
3	the	15149	the	15143	the	15145
4	s	11579	a	11021	a	11028
5	a	11074	to	10987	to	10996
6	to	11013	and	7813	it	10609
7	it	10611	it	7131	's	10230
8	that	8820	that	6779	that	8820
9	and	7831	of	6144	and	7815
10	t	7204	is	5334	n't	7164
11	of	6150	in	5140	do	6629
12	what	5632	me	4908	of	6144
13	is	5332	what	4871	what	5627
14	we	5240	this	4804	is	5626
15	in	5167	no	4361	we	5240
16	me	4915	oh	4265	in	5146
17	this	4815	on	4213	me	4912
18	m	4546	i'm	3814	this	4815
19	no	4372	your	3806	'm	4468
20	oh	4271	we	3779	no	4361
					h	4268
					n	4216
					ur	3806
					y	3687
					e	3684
					r	3627
					ve	3611
					ow	3581
					e	3479
					as	3330

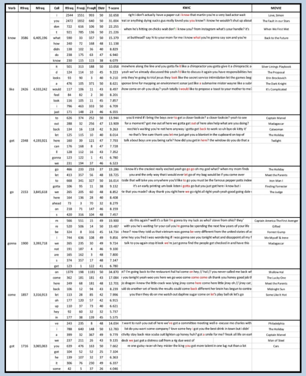
RANK	*AntConc 4.0.5*				*#LancsBox 6.0*		*SketchEngine*	
	TYPE	FREQ.	TYPE (with space separators)	FREQ.	TYPE	FREQ.	TYPE	FREQ.
1	i m	4472	you know	1497	you know	1497	do n't	3222
2	it s	3381	are you	1290	are you	1288	i do	1563
3	don t	3223	i don	1279	in the	1219	you know	1497
4	you re	2						
5	that s	2						
6	you know	1						
7	are you	1						
8	i don	1						
9	in the	1						
10	do you	1						

RANK	*AntConc 4.0.5*				*#LancsBox 6.0*		*SketchEngine*	
	TYPE	FREQ.	TYPE (with space separators)	FREQ.	TYPE	FREQ.	TYPE	FREQ.
1	i don t	1279	no no no	471	no no no	486	i do n't	1277
2	don t know	561	what are you	378	what are you	389	do n't know	560
3	i m not	519	oh my god	309	i don't know	372	you do n't	489
4	you don t	499	what do you	283	oh my god	309	no no no	477
5	no no no	487						
6	i m sorry	465						
7	i can t	440						
8	i m gonna	423						
9	what are you	389						
10	i didn t	375						

RANK	AntConc 4.0.5				*#LancsBox 6.0*		*SketchEngine*	
	TYPE	FREQ.	TYPE (with space separators)	FREQ.	TYPE	FREQ.	TYPE	FREQ.
1	i don t know	441	no no no no	252	no no no no	257	i do n't know	440
2	no no no no	258	what are you doing	173	what are you doing	173	no no no no	255
3	what are you doing	173	come on come on	99	come on come on	108	what are you doing	173
4	i m sorry i	131	what do you mean	83	what do you mean	83	why do n't you	114
5	why don t you	114	i want you to	82	i want you to	82	do n't know what	110
6	come on come on	***111***	you want me to	61	oh oh oh oh	71	come on come on	108
7	don t know what	***111***	aargh aargh aargh aargh	***58***	aargh aargh aargh aargh	68	i do n't think	101
8	i don t think	101	are you talking about	***58***	you want me to	61	i do n't want	95
9	i don t want	95	oh oh oh oh	***58***	i know i know	60	you do n't have	88
10	i m i m	91	what do you think	58	what do you think	***59***	what do you mean	83

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 gave, then, 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).



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 beg	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	how come on oh you okay? yeah totally i would like to propose a toast to your mother to me i	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	ie 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 a	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	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	gic 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		

MULTI-DIMENSIONAL ANALYSIS (MDA)

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 . More specifically, the following five Biberian dimensions, which are represented by Factors 1-5 respectively, are considered :

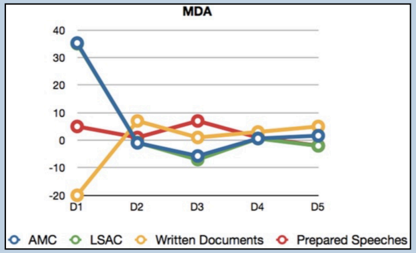
D1: the informational (negative) vs. involved (positive) production dimension, which identifies whether a text is marked by high informational density and exact informational content or, on the contrary, by affective, interactional, and generalized content (Biber 1988:107);

D2: the narrative (positive) vs. non-narrative (negative) concerns dimension, which distinguishes narrative discourse from other types of discourse (Biber 1988:109);

D3: 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);

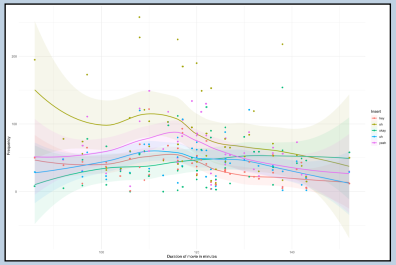
D4: the overt expression of persuasion (positive) dimension, which marks the degree to which persuasion is marked overtly employed (Biber 1988:111);

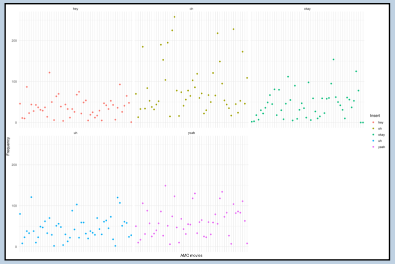
D5: the abstract (positive) vs. non-abstract (negative) information dimension, which "seems to mark informational discourse that is abstract, technical, and formal versus other types of discourse" (Biber 1988:113).



SCATTERPLOTS

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.

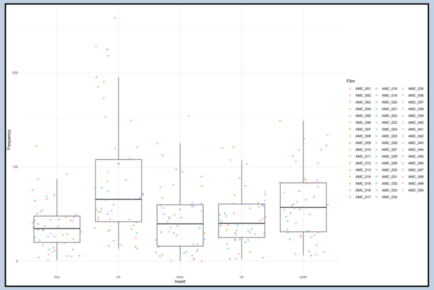


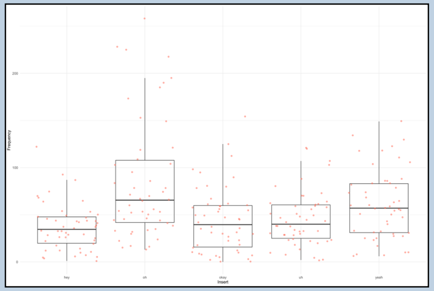


BOXPLOTS

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

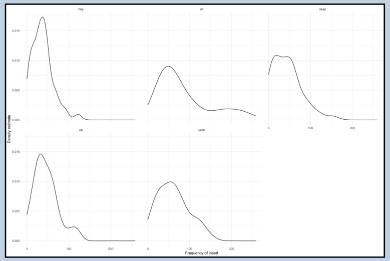
- the bold-typed horizontal lines represent the medians of the two vectors;
- the regular horizontal lines that make up the upper and lower boundary of the boxes represent the hinges (approximately the 75%- and the 25% quartiles);
- the whiskers – the dashed vertical lines extending from the box until 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.





DENSITY PLOTS

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).



HOW TO CITE US:

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