

Chatbot abuse: the case of ELIZA

Darja Fišer and Tomaž Erjavec

Jožef Stefan Institute and University of Ljubljana
Slovenia

Overview of the talk

1. Introduction
2. The ELIZA chatbot
3. The ELIZA corpus
4. Lexical analysis of the corpus
5. Conclusions

Introduction: chatbots

- Conversational agents (Cassell 2000)
 - software agents which perform tasks or services for their users
 - chatbots on the web, virtual assistants on smartphones, automated call handling systems
 - efficiency, special needs, entertainment, companionship
 - introduced into everyday tasks and technologies
- Human-computer interaction
 - increasingly important research topic
 - computer science: how to create a machine that can hold a conversation
 - communication studies and social sciences: the social role of conversational agents

Abusive behaviour towards conversational agents

- Nothing new:
observed in many novel mechanical inventions, e.g. cars - fear (Billerter 1997)
- Increasingly human-like agents result in anthropomorphism (Bicmore and Picard 2005)
- Still heavily underresearched (De Angeli et al. 2005)
- Estimated to be much higher in both volume and intensity than in human-human interaction (20-30% of all interactions abusive)

Related work

- Veletsianos, Scharber, and Doering (2008):
 - analysis of conversations between teenagers and a pedagogical agent in a school environment
- De Angeli and Brahnham (2008):
 - analysis of a corpus of conversation with a chatterbot designed mainly for entertainment and companionship
- Rehm (2008):
 - lab. evaluation of a game where two humans and an embodied conversational agent participate in a game requiring deception

Related work: results

- The social acceptability of disinhibition tends to increase in these studies
- Verbal abuse & sexual conversations are common in anonymous interaction with conversational agents
- Intended focus of the dialogue (learning vs. fun), interaction setting (school vs. home), and interface (text only chatterbot vs. embodied agent) has little effect on the occurrence of verbal abuse (10% of communication)
- Independent of the context, people enjoy provoking, teasing, and humiliating the conversational agent
- Conversational agents are conducive towards disinhibited behavior and not bound by moral constraints

The ELIZA chatbot

ELIZA

- ELIZA is one of the first, and still best known computer chat(ter)bots
- Named after Eliza Doolittle (Shaw's Pygmalion)
- Created in 1966 by Joseph Weizenbaum (1923–2008) at MIT AI Laboratory
- ELIZA mimics a Rogerian psychologist: open-ended questions to encourage patients to communicate more effectively with therapists
- Weizenbaum is considered one of the fathers of AI
- He was ambivalent towards computer technology: computers should not be allowed to make important decisions because they lack human qualities such as compassion and wisdom
- ELIZA used as case in point: his secretary asked him to leave the room after starting to converse with the chatbot

Joseph Weizenbaum

Computer Power and Human Reason (1976)

young woman: Men are all alike.

eliza: In what way?

young woman: They're always bugging us about something specific or other.

eliza: Can you think of a specific example?

young woman: Well, my boyfriend made me come here.

eliza: Your boyfriend made you come here?

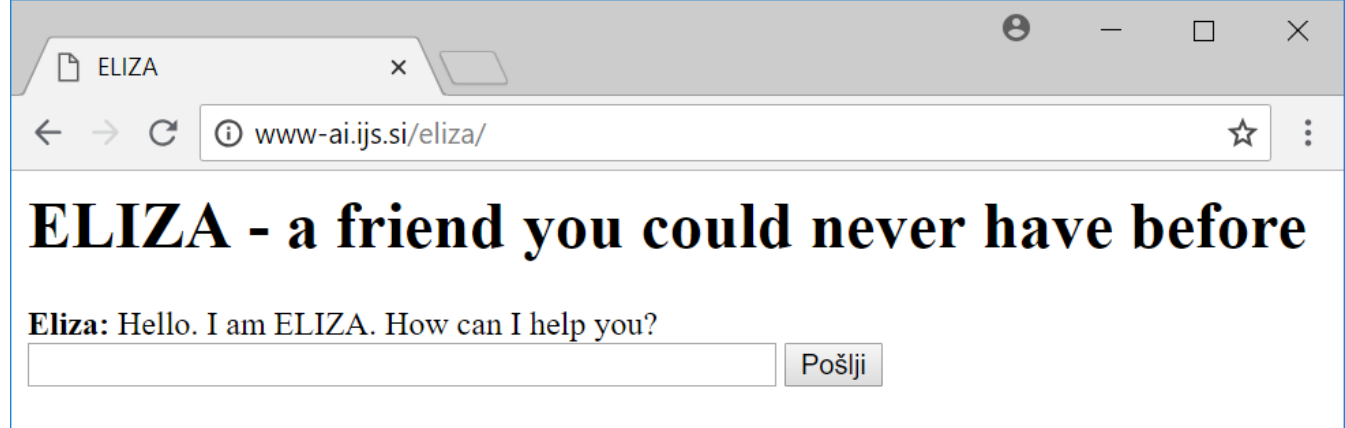
young woman: He says I'm depressed much of the time.

eliza: I am sorry to hear you are depressed.

young woman: It's true. I'm unhappy.

eliza: Do you think coming here will help you not to be unhappy?

ELIZA in action

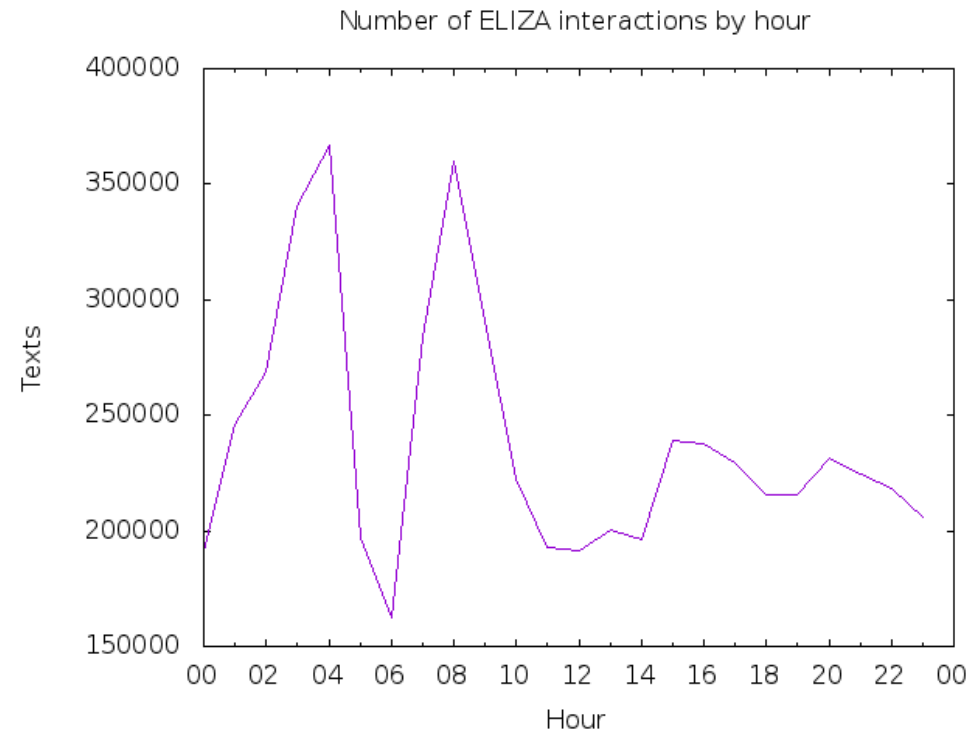
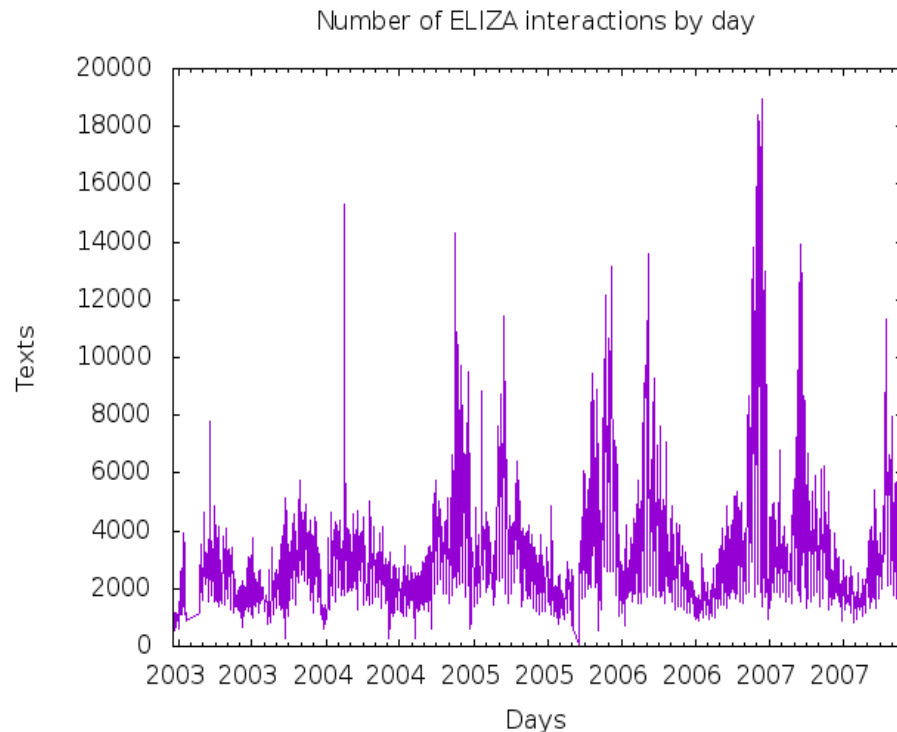


- ELIZA uses pattern matching and reformulates users' sentences into questions
- Re-implemented countless times and in different computer languages like LISP (part of Emacs) and Prolog
- The Prolog version was made available at www-ai.ijs.si/eliza/ by Marko Grobelnik soon after the beginning of the Internet
- URL linked from many sites, e.g. from Wikipedia, therefore very popular
- Web version worse than original:
users are anonymous, so can't use previous answers for new questions
- Keeps a log of interactions

The ELIZA corpus

The ELIZA Corpus

- Source: conversation logs of www-ai.ijs.si/eliza
- 5 years of conversations: 2002-12-19 to 2007-11-26
- Cleaned, de-duplicated
- 5.7 million conversations, 25 million words



Thu Dec 19 20:54:14 CET 2002

i' am in love

How long have you been in love?

Thu Dec 19 20:54:39 CET 2002

for foor months

Please go on.

Thu Dec 19 20:56:54 CET 2002

Mi name is David

I have told you before, I do not care
about names.

Thu Dec 19 20:57:25 CET 2002

Ok. I need a piece of advice

Why do you want a piece of advice?

Thu Dec 19 20:57:58 CET 2002

Because i 'm not feeling well

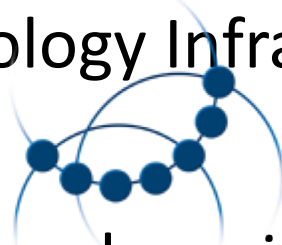
Do any other reasons not come to mind?

Structure of the corpus

- One interaction: time + input + question
- No user identification (overlapping conversations)
- Tokenised, part-of-speech tagged and lemmatised

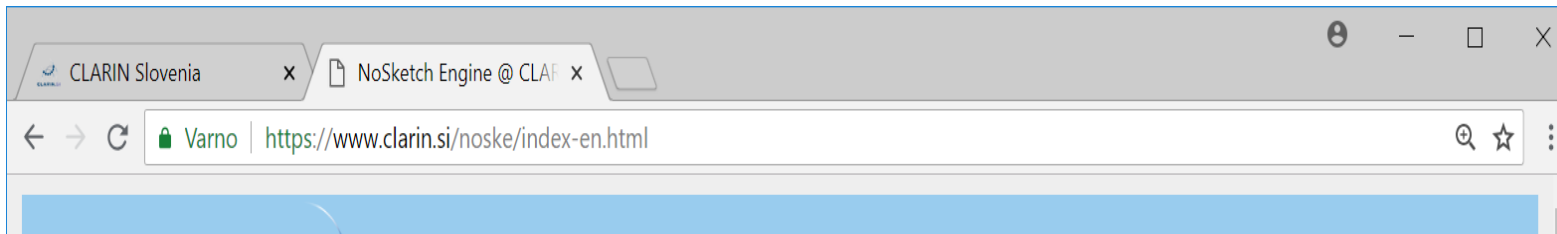
```
<text date="2002-12-19" time="20:54:39"  
  prev_q="How long have you been in love ?"  
  next_q="Please go on .">  
  <w lemma="for" ctag="Sp">for</w><c> </c>  
  <w lemma="foor" ctag="Afp">foor</w><c> </c>  
  <w lemma="month" ctag="Ncnp">months</w>  
</text>
```

Availability of the ELIZA corpus

- European research infrastructure CLARIN
„Common Language Resources and Technology Infrastructure“
- 20 members, also Slovenia: **CLARIN.SI** 
- Support for events and language resource and service creation for Slovene (and other languages)
- Certified repository of language resources (downloadable datasets)
- Two on-line concordancers: noSketch Engine and KonText
- ELIZA corpus publicly available through both

noSketchEngine

- search with powerful CQL language: regular expressions and annotations
- concordances
- frequency lexica
- keywords
- collocations



A screenshot of the 'Corpus info' page for the ELIZA chatbot corpus. The browser's address bar shows the URL 'https://www.clarin.si/noske/run.cgi/corp_info?corpname=eliza_en&struct_attr_stats=1&subcorpora=1'. The page is titled 'ELIZA (chatbot)' and displays 'Logs of conversations with ELIZA@ijs.si'. It features four main tables: 'Counts', 'General info', 'Lexicon sizes', and 'Tags legend'. The 'Counts' table shows statistics for Tokens, Words, Paragraphs, and Documents. The 'General info' table provides details about the Corpus description, Language, Encoding, and Tagset. The 'Lexicon sizes' table lists word, lemmas, and tag counts. The 'Tags legend' table defines various grammatical tags. At the bottom, the 'Structures and attributes' section shows a dropdown menu for 'text' with 5,728,568 items and a list of attributes like date, next_q, prev_q, and time.

Lexical analysis of the ELIZA corpus

Keyword analysis

- Profanity wordlist (von Ahn @ CMU)
 - list of 1400 English words that could be found offensive
 - <http://www.cs.cmu.edu/~biglou/resources/bad-words.txt>
- Comparison with corpus of the English web (ukWaC)
- Categories of swearing and abuse vocabulary (McEnery et al. 1998)
 - traditional swearwords (*fuck, piss, shit*)
 - animal terms of abuse (*pig, cow, bitch*)
 - sexist terms of abuse (*bitch, whore, slut*)
 - intellect-based terms of abuse (*idiot, prat, imbecile*)
 - racist terms of abuse (*paki, nigger, chink*)
 - homophobic terms of abuse (*queer, puff, lezza*)
- Categorization of top 100 keywords

Top 30 ELIZA

lc	ELIZA (chatbot)		ukWaC (British Web)		Score
	frequency	frequency/mill ?	frequency	frequency/mill	
fuck	104,320	4174.4	8,312	3.9	853.5
suck	29,215	1169.0	5,068	2.4	346.9
bitch	25,254	1010.5	4,869	2.3	308.4
horny	5,337	213.6	1,038	0.5	144.4
stupid	42,940	1718.2	27,961	13.1	122.0
penis	7,518	300.8	3,488	1.6	114.6
asshole	3,029	121.2	504	0.2	98.9
pussy	4,109	164.4	1,529	0.7	96.4
shit	16,665	666.8	12,901	6.0	94.9
dumb	9,582	383.4	7,352	3.4	86.5
fucking	10,170	407.0	8,363	3.9	83.0
fucker	2,359	94.4	391	0.2	80.6
ass	10,980	439.4	9,775	4.6	79.0
slut	2,395	95.8	543	0.3	77.2
idiot	7,926	317.2	7,446	3.5	70.9
whore	2,961	118.5	1,541	0.7	69.4
cunt	2,383	95.4	944	0.4	66.8
cock	5,936	237.5	5,902	2.8	63.4
moron	2,255	90.2	1,107	0.5	60.1
fucked	2,221	88.9	1,641	0.8	50.8
loser	3,152	126.1	3,764	1.8	46.0
poop	1,617	64.7	1,082	0.5	43.6
dick	10,568	422.9	18,770	8.8	43.3
boobs	1,446	57.9	785	0.4	43.0
retarded	1,869	74.8	1,786	0.8	41.3
motherfucker	1,090	43.6	241	0.1	40.1
blowjob	999	40.0	96	0.0	39.2
sex	38,075	1523.6	85,747	40.2	37.0
vagina	1,717	68.7	1,913	0.9	36.8
dumbass	923	36.9	138	0.1	35.6

Top 30 ukWaC

lc	ukWaC (British Web)		ELIZA (chatbot)		Score
	frequency	frequency/mill ?	frequency	frequency/mill	
uk	2,262,553	1059.4	178	7.1	130.6
european	438,379	205.3	38	1.5	81.8
remains	161,090	75.4	38	1.5	30.3
ethnic	74,641	34.9	6	0.2	29.0
poverty	96,100	45.0	23	0.9	24.0
church	474,259	222.1	215	8.6	23.2
minority	62,632	29.3	8	0.3	23.0
welfare	96,294	45.1	30	1.2	20.9
joint	167,799	78.6	79	3.2	19.1
radical	49,872	23.4	10	0.4	17.4
liberal	64,535	30.2	20	0.8	17.3
conservative	59,021	27.6	17	0.7	17.0
crime	176,595	82.7	105	4.2	16.1
diseases	65,865	30.8	31	1.2	14.2
israel	92,077	43.1	57	2.3	13.4
period	454,150	212.7	389	15.6	12.9
execution	32,950	15.4	8	0.3	12.4
coloured	34,633	16.2	12	0.5	11.6
cemetery	24,331	11.4	2	0.1	11.5
soviet	36,621	17.1	15	0.6	11.3
racial	29,607	13.9	8	0.3	11.3
israeli	34,667	16.2	14	0.6	11.0
crimes	31,434	14.7	11	0.4	10.9
african	83,817	39.2	68	2.7	10.8
refugee	26,305	12.3	6	0.2	10.7
palestinian	26,388	12.4	7	0.3	10.4
deposit	49,882	23.4	34	1.4	10.3
criminal	103,280	48.4	95	3.8	10.3
australian	53,042	24.8	38	1.5	10.3
executed	26,513	12.4	9	0.4	9.9

Categories

Swear and abuse (McEnery et al. 1998)	
sexist terms of abuse	13
traditional swearwords	13
intellect-based terms of abuse	10
homophobic terms of abuse	5
racist terms of abuse	2
general terms of abuse	3
total	46

Profanity	
sex	32
excretion	9
violence	3
emotion	3
body parts	2
clothing	2
spelling error	1
language	1
religion	1
total	54

- 10 keywords with more than 1 spelling variant
- nearly half of the keywords swear & abusive words
- animal terms of abuse not among 100 top-ranking keywords
- McEnery et al. (1998) missing general terms of abuse (*sucker, wanker*)

Collocation analysis

- traditional swearwords: FUCK

	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>MI3</u>
you	69,094	1,638,053	35.454
off	12,405	23,260	34.159
fuck	15,149	104,320	32.858
i	29,949	1,385,109	32.077
yes	23,117	759,436	31.824
?	23,958	1,024,521	31.546
no	15,037	447,869	30.724
me	12,752	349,427	30.369
!	10,784	279,731	29.964
want	8,847	180,104	29.743
what	10,928	422,597	29.427
to	11,902	655,020	29.164
.	11,963	668,537	29.157
bitch	3,843	25,254	28.968
are	10,985	590,151	28.967

- sexist terms of abuse: BITCH

	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>MI3</u>
you	12,252	1,638,053	30.038
bitch	2,754	25,254	29.597
a	6,932	435,130	29.486
fuck	3,620	104,320	28.734
are	5,384	590,151	27.952
yes	5,115	759,436	27.367
fat	1,052	6,854	27.314
i	5,491	1,385,109	26.807
son	626	2,487	26.530
fucking	972	10,170	26.402
no	3,287	447,869	26.215
!	2,756	279,731	26.131
,	3,260	481,830	26.073
stupid	1,310	42,940	25.616
up	1,149	46,209	24.942
shut	894	25,369	24.721
?	3,064	1,024,521	24.717

- intellect-based terms of abuse: IDIOT

	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>MI3</u>
an	3,438	28,506	32.043
you	5,070	1,638,053	27.879
are	3,047	590,151	27.148
're	612	32,838	24.369
yes	1,732	759,436	24.340
i	1,784	1,385,109	23.601
,	1,119	481,830	23.105
.	1,219	668,537	23.003
no	1,004	447,869	22.741
!	842	279,731	22.659
fucking	277	10,170	22.629
a	919	435,130	22.400
idiot	234	7,926	22.258
?	1,086	1,024,521	21.887
stupid	367	42,940	21.768
your	508	182,977	21.084
is	616	347,571	20.993
u	336	58,457	20.941

- sexist terms of abuse: SLUT

	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>MI3</u>
a	1,042	435,130	24.668
you	1,335	1,638,053	23.828
whore	136	2,961	23.054
dirty	112	1,957	22.811
bitch	241	25,254	22.438
are	653	590,151	22.206
slut	101	2,395	22.072
fuck	240	104,320	20.374
fucking	100	10,170	19.943
yes	420	759,436	19.932
i	505	1,385,109	19.862
u	149	58,457	19.146
no	266	447,869	18.717
your	190	182,977	18.552
,	262	481,830	18.546
're	97	32,838	18.120
?	300	1,024,521	18.043
!	189	279,731	17.917

- racist terms of abuse: NIGGER

	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>MI3</u>
nigger	36	421	22.192
niggers	14	157	19.527
fuck	101	104,320	18.703
bastards	8	67	18.333
you	230	1,638,053	18.292
a	138	435,130	17.994
pappy	3	5	17.832
lowlife	4	15	17.493
sand	7	92	17.298
creole	3	9	16.984
black	17	1,649	16.974
fucker	16	2,359	16.196
fucking	26	10,170	16.189
garbage	7	232	15.964
i	122	1,385,109	15.790
?	110	1,024,521	15.777
are	91	590,151	15.752
hate	32	36,874	15.229

- homophobic terms of abuse: FAG

	<u>Cooccurrence count</u>	<u>Candidate count</u>	<u>MI3</u>
a	415	435,130	22.289
fag	34	753	20.635
you	340	1,638,053	19.513
are	181	590,151	18.258
gay	50	20,057	17.568
is	123	347,571	17.349
u	66	58,457	17.227
your	91	182,977	16.971
i	177	1,385,109	16.930
no	120	447,869	16.877
ur	29	7,300	16.669
yes	125	759,436	16.292
russel	5	50	16.251
faf	3	11	16.224
?	126	1,024,521	15.894
such	16	3,056	15.351
fags	4	48	15.344
fuck	44	104,320	14.636

Abusive collocates

- Observing the top ranking keywords from each of the categories by McEnery et al. Inspected top 20 collocates in window -3 .. +3 and using the MI3 score
- How many are abusive?

Abusive collocates	<i>n</i>
fuck	3
bitch	6
idiot	3
slut	6
nigger	11
fag	6

Conclusions

- Nice inversion: Weizenbaum's secretary vs. real users
- Due to stupidity of program, quick descent into aggression: much more abuse than is reported for other chatbots
- So, a large publicly accessible corpus of abusive texts
- Possible to download ELIZA corpus on request (GDPR)
- Presented a preliminary lexical analysis of the corpus
- Further work: better PoS annotation, further analyses

References

- Cassell, J., 2000. Embodied conversational agents. MIT press.
- Billerter, J., 1997. Look at what happens to telltales and buffaloes. In: AAI Fall Symposium on Socially Intelligent Agents, Cambridge, MA, pp. 7–9.
- Bickmore, T.W., Picard, R.W., 2005. Establishing and maintaining long- term human–computer relationships. ACM Transactions on Computer–Human Interaction 12, 293–327.
- De Angeli, A., Brahnam, S., Wallis, P., 2005. ABUSE: the dark side of human–computer interaction. In: Buono, P., Costabile, M.F., Paterno, F., Santoro, C. (Eds.), Interact 2005 Adjunct Proceedings, Rome, pp. 91–92.
- Veletsianos, G., Scharber, C. and Doering, A., 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. Interacting with Computers, 20(3), pp.292-301.
- De Angeli, A. and Brahnam, S., 2008. I hate you! Disinhibition with virtual partners. Interacting with computers, 20(3), pp.302-310.
- Rehm, M., 2008. “She is just stupid”—Analyzing user–agent interactions in emotional game situations. Interacting with Computers, 20(3), pp.311-325.
- Weizenbaum, J., 1976. Computer power and human reason: From judgment to calculation.
- McEnery, A., Baker, J.P. and Hardie, A., 2000. Assessing claims about language use with corpus data: Swearing and abuse. Language and Computers, 30, pp.45-56.