How Real is the Virtual? Interactive Alignment in Human-Human vs. Simulated Human-Bot Dialogues

Rania El-Wakil

Assistant Professor,
Faculty of Languages,
October University
for Modern Sciences
and Arts (MSA),
Egypt.

Abstract

Dialogue is a communicative activity based on verbal/nonverbal interaction. Research on dialogue has been conducted in various disciplines to examine various contexts, production and reception mechanisms, dialogic structures as well as varied functions and implications of dialogic interaction. Conversational AI is an umbrella term used to describe various methods of enabling computers to carry on a conversation with a human. This technology ranges from fairly simple natural language processing (NLP) applications to more sophisticated machine learning (ML) and Large Language Models (LLMs) that can interpret a wider range of inputs and conduct more complex conversations. The study attempts a contrastive analysis of three authentic human-human conversations and their human-bot simulations by *character.ai* (https://character.ai) chatbots to assess whether the virtual interactions succeed in simulating the real interactive experience. Pickering and Garrod's (2021) Interactive Alignment model is adopted as the framework of the analysis. Selected NLP applications are employed to assess lexical and semantic alignment in addition to other dialogue production mechanisms, namely routinization and AI hallucination. The results show that the human-bot interactions provide successful simulations of the human-human interaction; however, some human interactional features are still missing. Fine-tuning of the language model employed by the bot is recommended to maximize the authenticity of the interactive experience.

Keywords: dialogue, interactive alignment, conversational AI, natural language processing (NLP)

How Real is the Virtual? Interactive Alignment in Human-Human vs. Simulated Human-Bot Dialogues

Rania El-Wakil

1. Introduction

Dialogue is a communicative instance where participants exchange verbal utterances potentially accompanied by non-verbal activities. Dialogues take various forms, modalities and structures according to the number of participants, context of interaction as well as the extent of participant contributions and synchrony. This form of communication has been the scope of study in various disciplines: linguistics, psychology, sociology and anthropology. The study of dialogue in these fields examines interactive behaviours and multi-faceted imports of engaging in conversations.

The development of conversational AI has given rise to human-computer dialogue via natural-language-based interfaces or programs referred to as chatbots, interactive agents, or artificial conversation entities (Adamopoulou & Moussiades, 2020). Dialogue systems are employed in various fields such as: education; marketing, healthcare, support systems and entertainment. Dialogue systems fall under two main categories: goal-driven systems and non-goal-driven systems. The former is task-oriented where user inputs and system outputs fulfill a specific communicative task of information extraction and provision in structured conversations. The latter engages in synchronous or asynchronous 'unstructured chat' simulating human–human conversations. Though non-goal-driven systems, or chatterbots, are not initially task-identified, they can be employed to perform specific tasks by the user within the extended conversation. Both systems employ a cardinal Natural Language Processing (NLP) architecture: Natural Language Understanding (NLU), Dialogue Management/ Modelling, and Natural Language Generation (NLG). To enhance further spoken manner of interaction, speech recognition and text-to-speech synthesis are included. (Wang & Yuan, 2016; Traum, 2017; Mehndiratta & Asawa, 2021).

Character.ai is an AI-powered chatbot that allows users to create characters of known/unknown entities and engage in open-ended conversations. It is based on predictive text generation Large Language Model (LLM) whereby huge datasets are read and the system generates fabricated responses based on what might come next in any given situation. Feedback on responses also helps improve character conversational performance.

The study attempts a contrastive analysis of Pickering & Garrod's (2021) interactive alignment in three authentic human-human dialogues and their virtual simulations on *character.ai* to assess how successful the chatbot is in delivering an authentic interactive experience in a non-task-based dialogue within the arena of NLG exploration. Interactive alignment is detected at the lexical and semantic levels using selected NLP tools: Jaccard Similarity, Cosine Similarity, and part of speech tagging (POS) via NLTK and Spacy. Routinization and AI hallucinations are also examined as discernible production mechanisms. Routine expressions are extracted by *AntConc* high-frequency *n*-grams and categorized upon Biber et al (2003) and Biber & Barbieri (2007) lexical bundles. AI hallucination is detected manually and assessed qualitatively.

1.1 Interactive Alignment

A dialogue is a joint communicative act requiring interlocutors to interact within a collaborative platform of linguistic representations: phonetic, syntactic, and semantic as well as pragmatic performance of illocutionary acts and dialogue management (Poesio & Rieses, 2010). Analogous to a joint workspace, successful dialogues maintain coordination and alignment. Coordination relates to the conversing behavior while alignment pertains to the cognitive processes of language comprehension and production, particularly parity of pragmatic modelling of the context and symmetrical linguistic representations. Participants need to coordinate their roles and contributions (Pickering & Garrod, 2004; 2021). Coordination manifests in timely interactive responses, self-monitoring, and role negotiation.

The empirical Conversation Analysis methodologies in the 1960s and 1970s investigate dialogic structure and interlocutors' contribution to the orderly flow of interaction and attempt a description of regulatory practices underlying the dialogue architecture: sequencing or adjacency pairs (Schegloff, 1968), openings and closings (Schegloff and Sacks, 1973), turn-taking (Sacks et al, 1974) and transitions (Clark and Schaefer, 1989). Pragmatically, participants in a dialogue maintain standardized contributions to the conversational content. This is outlined in Grice (1975) Cooperative Principle (CP): "Make your contribution such as required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (p.45). The social context in which a dialogue takes place has direct and indirect impact on participants' contributions and the way coordination is sustained. The ethnography of communication studies conversations as indexical of the dialectic relationship between verbal communication and the socio-cultural context. Hymes (1962; 1964; 1976) anthropological view of language use in specific communicative contexts is based on the assumption that interactions are arranged in typical speech communities, whereby participants abide by and mirror local cultural practices and norms of communication.

Van Dijk (1984) maintains that in a dialogue, language production and comprehension processes synchronize in a cognitive model including linguistic and pragmatic information processing: interpretation and inferencing upon participants' common "conceptual knowledge" and representation of the world– "frame" or "script" (p.4). Based on neuro-cognitive evidence, Menenti et al (2012) argue for a close relation between language production and comprehension as neural activities. Alignment of production at different linguistic levels corresponds to alignment of mental representation. Parity between production and comprehension results from the bidirectional neural activation of both linguistic situation models and conceptual representations along the dialogue. More concretely, Pickering & Garrod (2004, 2013, 2021) set an empirical framework for the study of cognitive mechanism of comprehension and production coupling in dialogue. Their framework – Interactive Alignment – both accounts for language processing and inference mechanisms as human mental states and sets a foundation for further machine dialogue modeling and management.

The interactive alignment framework argues for the following: a successful dialogue is based on alignment of situational model achieved via priming mechanisms producing alignment of representation at other interconnected linguistic levels: lexical, syntactic and semantic. In case of misaligned representations, interlocutors employ repair mechanism to re-establish alignment. Interactive alignment also leads to employing fixed expressions, or routines.

- Alignment of Situational Model & Linguistic Alignment

A situational model is the speakers' mental state of the "multi-dimensional representation of the situation under discussion" in the dialogue (Pickering & Garrod, 2004, p.172). Interlocutors align on each other's representations of situational parameters based on implicitly shared background or information (Pickering & Garrod, 2006). Clark & Wilkes-Gibbs (1986) refer to grounding as speakers' attempts to establish definite references and coordinate beliefs in the functionality of the reference allowing the flow of the conversation. This is established through presentation of the reference by one speaker and acceptance by the other speaker. Pickering & Garrod (2004) maintain that alignment of situation models is based on automatic mechanisms of inference and priming at the local level leading to global alignment. Production of a particular representation of a mutually accepted reference activates a specific aspect of the situation model and would lead to "parity between the representations used in production and those used in comprehension" (p.174) - a repetitive use of the representation via similar linguistic structures. In other words, production is enhanced by the fact that previous utterances activate syntactic and lexical representations. Hence, speakers tend to repeat these syntactic and lexical forms to align with their interlocutors, not a mere behavioral imitation. Since interlocutors have developed aligned situation models with shared entities and conceptual relations, they also have similar patterns of activation of linguistic knowledge to represent these entities and relations.

- Repair

Repair mechanism is employed by interlocutors "to rectify failures in alignment" (Pickering & Garrod, 2006, p.206). In order to ensure maintaining the common ground, speakers reformulate the linguistic representation if other participants cannot straightforwardly interpret it; i.e. production does not align with comprehension (Pickering & Garrod, 2004). Repair process can be iterative until alignment is re-established.

- Routinization

A routine is a "fixed" expression of a particular lexical and syntactic structure and pragmatic function. It occurs at a much higher frequency than the frequency of its component words (Aijmer, 2014 in Pickering & Garrod, 2021). Routines fall under two categories: (i) permanent routines: stock phrases, idioms and cliches; (ii) temporary routines that are idiosyncratic to the ongoing dialogue when interlocutors propose expressions with unique meanings for a specific purpose in a particular interaction. Routinization enhances interactive alignment. A consequence of recurrent "representational parity" is that there is a single level of high activation associated with both production and comprehension (Pickering & Garrod, 2021, p122-3)

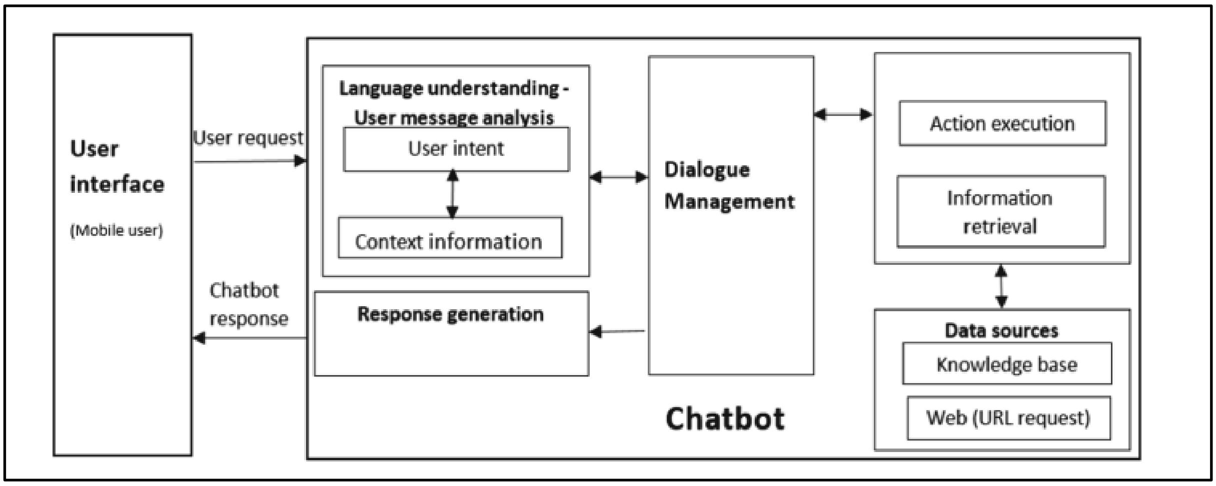
Interactive alignment model is employed in the architecture and assessment of conversational AI and machine dialogue systems as shown in the coming section.

1.2 Conversational AI: Chatbots – Large Language Models (LLM)

Conversational Artificial Intelligence (AI) is "the study of techniques for creating software agents that can engage in natural conversational interactions with humans" (Khatri et al, 2018). It enables computer systems to understand human language and generate human-like responses in a simulation of human conversation through natural language processing (NLP) and machine learning algorithms. These systems allow text-based, spoken as well as multimodal interactions. Pragmatically, they are classified into task-oriented and non-task-oriented systems (McTear, 2022). Five "traditions" of conversational user interface emerge: "text-based and spoken Dialogue systems, voice user interfaces, chatbots, embodied conversational agents, and social robots and situated agents" (p.14). The difference between them lies in the kind and purpose of interactions with humans.

Dialogue systems is a human-machine conversation through various modalities in a wide range of domains, most commonly task-based. To maintain a coherent dialogue flow, dialogue systems operate on natural language understanding (NLU) and Natural Language Generation (NLG) algorithms in addition to named entity recognition (NER) and dialogue management. Voice user interfaces allow users to interact with devices through speech, delivering prompts in the form of orders to perform specific tasks or inquiries for information. Examples are digital personal assistants. Embodied Conversational Agents (ECA) provide a more engaging interaction as it has a dual nature: a virtual conversational agent with the features of an animated character (McTear, 2022). Examples are avatars. Social robots extend ECA utilizing physical embodiment interacting with humans in various social contexts distinguishing different emotions through facial expressions, gestures and other physical cues in the surrounding social environment. The system's conversational capacities are enhanced by effective social behaviour (del Moral et al, 2009). Pertinent to the current study are chatbots, systems or conversational agents that can conduct extended conversations, mimicking "the unstructured" human-human chat (Jurafsky & Martin, 2018). Conversational AI poses the challenge of simulating not only human cognitive competences but also emotional competence. Hence, chatbot architecture and modeling are significant. Figure (1) illustrates the chatbot architecture.

Fig (1) General Chatbot Architecture, Adamopoulou & Moussiades (2020, p.380)



Huang & CIS (2021) and Mohammed & Aref (2022) elaborate on the core components of the chatbot architecture: User Interface, Natural Language Understanding (NLU), Dialogue Management (DM), and Response Generation/ Natural Language Generation (NLG). NLU provides both linguistic and semantic representations and disambiguation of the user's utterance through some processes such as part-of-speech tagging (POS), parsing, named entity recognition (NER), semantic role labelling and Latent Semantic Analysis (LSA). NLU also includes topic detection, dialogue act identification – corresponding to speech acts - and the more specific domain-dependent intent analysis. Context information is retrieved from users chat history database or knowledge of the domain to resolve coreferences and ambiguities. Dialogue Management keeps the flow of the conversation maintaining its specific intent and entities. It produces adequate responses, requests missing information and processes clarifications to fulfill the user's request. DM sends a conceptual representation of the communicative act to the natural language generator (NLG) to produce the textual representation (Galitsky, 2019). Selection of responses depends on particular strategies: rule-based, retrieval-based, and generative. Rule-based model depends on set rules to generate responses upon input text recognition within specific conversational patterns. Retrieval-based model is more flexible as it selects and matches the appropriate response from available resources using Application Programming Interfaces (APIs). Generative responses are more ‘human-like’ depending on machine learning (ML), and deep learning (DL) algorithms (Adamopoulou & Moussiades, 2020).

A language model, put in simple terms, is the AI system within NLP “to predict the next word in a sequence (the output) given the sequence of preceding words (the input, often called the ‘context’ or ‘preceding context’) to maintain fluency in text generation (Serrano et al, 2023, p.3). To perform various NLP tasks, such as: summarization, paraphrase, machine translation and conversational AI, a language model needs to learn both linguistic knowledge in addition to world knowledge. The type of language model depends on the size of training data and algorithm of input processing and output generation. Thus, a large language model is “capable of generating human-like text based on the patterns and relationships it learns from vast amounts of [naturalistic, digital] data”, (Chockalingam et al, 2023, p.8). Language models regimes are classified into: pre-transformers, transformers and large language models (LLMs). Pre-transformers models developed from neural networks (NNs) that “could take into account the context, position, and relationships between words even if they were far apart in a data sequence” upon extracting hidden patterns in the data (p.13). The limited ability of NNs to handle “longer data sequences” and adequately regard the “overall context of the input sequence”, has given rise to “transformers” with attention mechanisms, considering important parts in the overall context of input sequence with regards to the required task (p.14). Larger language models are developed to perform more arbitrary tasks providing “un-customized, generic” responses (p.21). Hallucination, though, is a major challenge before LLMs. Generating incorrect, uninterpretable responses presenting them as authentic including: prompt contradiction, sentence contradiction, factual contradiction, and source contradiction (Fokina, 2023).

1.3 Previous studies

Scholars have probed interactive alignment in human-human interactions and conversational AI at different levels and for various research purposes. Alignment measurement methodologies are proposed in some studies. Reitter and Moore (2014) have developed methods to measure structural priming within the syntactic choices in spontaneous conversations in contrast to task-oriented dialogues. Lower-level alignment - lexical and syntactic priming and repetitions - are higher in the latter correlating with higher-level semantic alignment, leading to successful task achievement. Doyle and Frank (2016) propose Word-Based Hierarchical Alignment Model (WHAM) and examine the effect of discourse acts on conversational alignment. In addition to linguistic and conceptual alignment, Rasenberg et al (2020) integrate multimodal alignment of gestures, speech and facial expressions during conversations established upon five key dimensions: modality, form, meaning, sequence and time. Similarly, Khosrobeigi et al (2022) study the association between semiotic alignment and concurrent part of speech (POS) alignment in task-based dialogues. Interactive Semantic Alignment Model (ISAM) is proposed by Kalociński et al (2018) positing that semantic alignment achieved by conversational agents depends on three key factors: users’ input, inclination towards conceptually simpler meanings and recent interactions when attempting semantic alignment. Dubuisson Duplessis et al (2021) devise a method to measure both lexical alignment between participants and self-repetition behaviours. They conclude that integrating lexical alignment competencies in NLG of dialogues has direct impact on dialogue planning and realization.

The positive impact of maintaining alignment is also explored in various contexts. Fusaroli and Tylén (2016) investigate interactive alignment, interpersonal synergy of conversational structures and self-consistency as collaborative mechanisms leading to efficient collective performance in task-based dialogues. Foushee et al (2022) explore lexical, syntactic, and semantic alignment in children-caregiver interactions in relation to language acquisition development and brain condition. In another vein, Sinclair et al (2019) examine alignment in human-human and human-agent dialogues between 2nd language learners and tutors showing a higher degree of students' alignment to the tutor bot than the human tutor. Based on a study of information-seeking conversations with embodied agents, Thomas et al (2020) conclude that alignment goes bidirectionally. It is evident that agents align to the style of user's input – via integrating style and alignment modules, topical language model and dialogue engine. It is also evident that users tend to align their style to that of the agent's resulting in smoothly running conversations. Spillner and Wenig (2021) examine alignment in conversations with an entertainment chatbot. They conclude that alignment leads to reducing user frustration and boost the chatbot performance, augmenting task success. Aligning to users' style via imitating their lexicon and syntactic structures, would reduce miscommunication resulting from users' informal style or the chatbot's lack of contextual knowledge. Similarly, Chen et al (2022) assume that alignment in dialogues with behaviour support agents, rather than data-driven frameworks, maintain more transparency and trustworthiness of the agent. They add that a model of well-aligned AI dialogue includes not only the dialogue pipeline, but affect model and causes of misalignment as well. Srivastava et al (2023) and Blum (2023) confirm the positive impact of lexical alignment on users' perception and understanding when engaging in collaborative task with aligned conversational agents.

2. Methodology

The study attempts a contrastive analysis of interactive alignment in human-human interactions and their virtual simulations to examine how far the human-bot interactions succeed in simulating the real human-human interactive experience. Selected data include three real human-human interviews with three figures in different fields: sports, politics and technology: (1) an interview between Piers Morgan and Cristiano Ronaldo (2022), (2) NPR's (National Public Radio) interview with former American president Barack Obama (2020), hosted by Michel Martin, and (3) an interview with Steve Jobs at Nova Southeastern University (NSU) about Dr. Juran, a pioneer in quality control and management (1991).

The texts are of different sizes, i.e. number of tokens, so that the features are evaluated not only in different contexts, but also within variant text sizes as shown below. The same questions and replies produced by the host in the real interviews are offered to the guest bot through *character.ai* chatbot (https://character.ai/) so the number of turns for each participant in each dialogue is the same and the size of the host text is constant. This helps in comparing the replies of the human guest to those produced by the bot to the same prompts. The real and virtual interviews are presented in the following table.

Table (1): Text Size of Real and Virtual Interviews

Text Size	Piers Morgan - Ronaldo			NPR – Obama			NSU Host – Jobs		
	Host	Guest	Bot	Host	Guest	Bot	Host	Guest	Bot
Number of turns	205	203	203	26	26	26	22	21	21
Number of tokens	4545	10914	10592	1616	4661	1655	345	2213	1643

Interactive alignment is detected first at the lexical and semantic levels employing NLP tools: *Jaccard Similarity* and *Spacy* POS tagging for lexical alignment; Cosine Similarity for semantic alignment through *Spacy* and *NLTK*. These assessments are implemented on *Google Colab* as python interactive platform. Routines are examined via *AntConc* - a computational text analysis tool (https://www.laurenceanthony.net/software/antconc/) to extract high frequency clusters. To detect instances of AI hallucination, a qualitative assessment of bot responses is conducted in comparison to human guest responses.

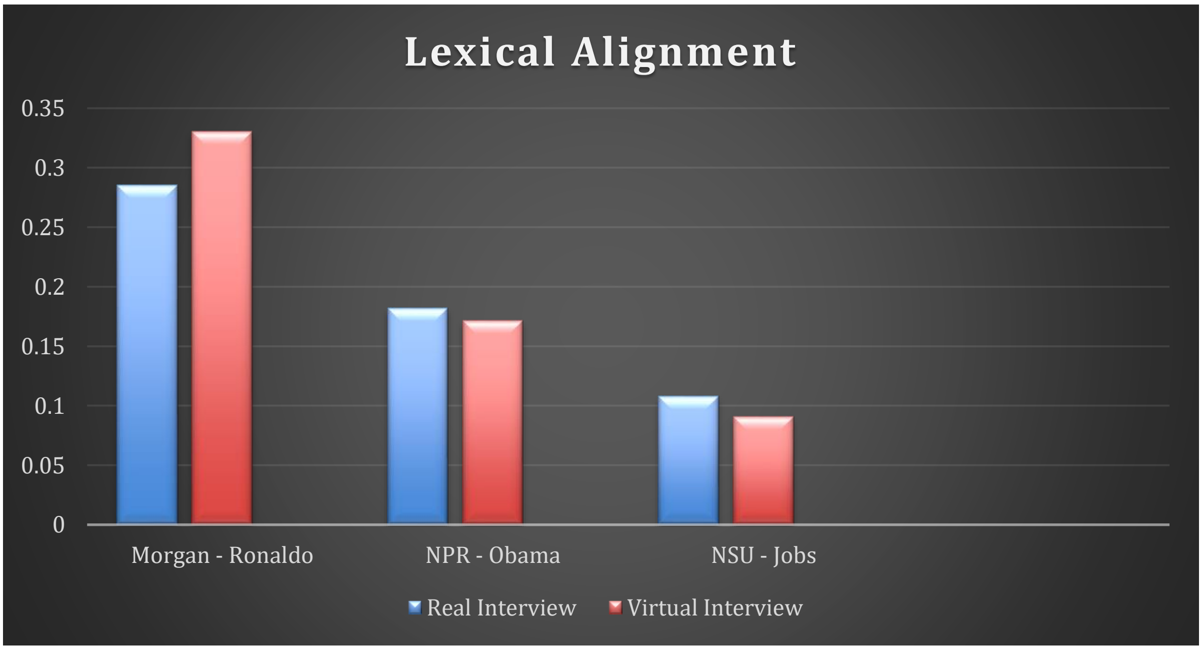
4.Results & Discussion

4.1 Lexical Alignment

Lexical alignment refers to vocabulary overlap in compared texts. Assessment of lexical alignment is done after removing functional stop words from the compared texts. *Jaccard Similarity* is used to assess the intersection of two texts: the number of common words is divided by their union– the total number of words in both texts. The score ranges from “0” indicating no similarity to “1” signifying complete overlap.

As shown in figure (2) below, despite the varying text sizes of human guest versus bot utterances as well the sizes of hosts’ utterances, *Jaccard Similarity* scores manifest low lexical alignment between the host and the guest utterances in all real and virtual interviews.

Fig (2) Lexical Alignment: Jaccard Similarity Scores



Scores range from the highest: <0.35 in Morgan -Ronaldo dialogues to the lowest: ≤0.1 in NSU - Jobs dialogues. It is clear that both the human guest and the bot employ a larger number of lexicons that vary from those employed by the host. It is normal in such type of non-task-based dialogues as guests give responses with varying degrees of elaboration and detail. Lexical alignment with the host utterances could be difficult to maintain as no specific answers are required or even expected. Hence, the bot maintains a regular human attitude in this respect.

The second method is conducted via *Spacy* to extract a list of shared words between the host and the guest then tagging the extracted list to determine alignment in terms of content word classes. POS tagging is validated manually. Despite the agreement in low lexical alignment scores, POS tagging of the shared lexicon between the host and the human guest on one hand and the host and bot on the other shows differences in using particular lexical categories and their sub-forms as depicted in table (2).

Table (2) Shared Lexical Categories/Forms in Real vs. Virtual Dialogues

Interview	Morgan – Ronaldo		Martin – Obama		NSU – Jobs	
	Human Guest	Bot	Human Guest	Bot	Human Guest	Bot
Total # of Shared Tokens with the host	**361**	**487**	**196**	**86**	**50**	**48**
Nouns						
Singular Noun	132	165	80	38	13	15
Plural Nouns	31	42	15	7	6	5
Proper Nouns	24	27			6	7
Total	**184**	**234**	**95**	**45**	**25**	**27**
Percentage	**51%**	**48%**	**48.5%**	**52%**	**50%**	**56%**
Verbs						
Base Form	52	55	25	19	5	3
3rd Person Singular	4	8	3	4		
Past form	6	11	5			
Past Participle	21	29	10	6	3	
Present Participle	14	34	11	3	3	4
Total	**97**	**137**	**54**	**13**	**11**	**7**
Percentage	**27%**	**28%**	**27.5%**	**15%**	**22%**	**15%**
Adjectives						
Adjective	52	74	26	13	8	10
Comparative Adjective			3			
Superlative Adjective	3	6				
Total	**55**	**80**	**29**	**13**	**8**	**10**
Percentage	**15%**	**17%**	**14.8%**	**15%**	**16%**	**21%**
Adverbs						
Adverb	**25**	**36**	**18**	**15**	**6**	**4**
Percentage	**7%**	**7%**	**9.2%**	**18%**	**12%**	**8%**

The percentages in table (2) above show that the general profile of lexical alignment in the human-bot interactions seems similar to that of the human-human interactions. They show dominance of nouns followed by verbs in the list of shared lexicons between the participants in all real and virtual interviews; then adjectives and the least used parts of speech are adverbs. Although the percentages of shared lexicons and their subcategories vary indicating varying degrees of lexical alignment within specific lexical categories, examining the shared lexical categories demonstrates some similarity in the lexical alignment of the human guest and the bot with the host.

More specifically, both Ronaldo and the bot share with the host close numbers of almost all lexical categories. The common lexicons in the two interviews indicate similarity in content. Both the bot and Ronaldo align with the host's utterances when talking about the various positive and negative aspects of Ronaldo's family life and professional career as reflected in the following noun forms such as: "World Cup – advice – baby – boss – business – career – challenge(s) – chance(s) – club(s) – daughter – debate – difference(s) – experience – family – friend(s) – game – goal(s) – heart – history – fans – son - Juventus – Manchester United - Alex Ferguson – Georgina – Bella - The Glazers – mentality – Messi – offer(s) – opinion – Portugal – Madrid – team(s) – Trafford – Tottenham". Expression of feelings and thoughts towards the fans, club managers, media criticism and other players is done employing similar verbs such as: "answer – believe – blame – criticized – change – explain – feel – felt – heard – hurt – like - know – listen – mean" in addition to some common adjectives such as: "young – worst – worried – professional – negative – honest – friendly". Though the two interviews manifest equal percentage of shared adverbs, not many of them are common. Few adverbs are used by the three participants such as: maybe – extremely – away. Similarly, the NSU – Jobs real and virtual interviews share almost all employed lexical categories. Interestingly, the two Jobs' interviews manifest what might be termed as "negative" alignment, i.e. the missing lexical categories are not employed in both interviews. However, few topic-specific content words are common in the two interviews: Both interviews share proper entities such as "Steve" and "Steven" when introducing the guest, and "Juran" and "Deming" when talking about the two quality pioneers in addition to abstract nouns such as: "quality, humour, life and time". More discrepancies are evident in the Martin-Obama interviews. The bot utterances show a smaller number of shared lexicons with the host than the human guest as well as varying degrees of lexical category alignment. Additionally, the real and virtual dialogues share a limited number of topic-specific lexicons such as: "citizens – country – elected – election – experience – fact – office – opinion – people – policy – president – racial – states – transition time – thought - think"; other shared lexicons are general terms. Relevant to lexical alignment is the detection of lexical routines shown in the following section.

4.2 Routinization

Routines are classified upon the definition and classification of lexical bundles provided by Biber et al (2003) and Biber and Barbieri (2007). Lexical bundles are frequent word clusters – serving three key functions: "stance expressions, discourse organizers, and referential expressions" (Biber & Barbieri, 2007, p. 263). Stance is further classified into epistemic, desire, obligation/directive, intention/prediction, and ability. Discourse organizers include topic introduction/focus and elaboration/clarification. Referential expressions include entities; specific/imprecise quantity, time/place deixis, text reference, specification of tangible/intangible attributes as well as multi-functional references. Additionally, spoken discourse manifests conversation-specific features such as: politeness, inquiry, reporting and others.

Routine lexical bundles are extracted via *AntConc*, detecting high-frequency clusters employed by the host, the human guest and the bot. Cluster size ranges from *2*-gram to 5-gram clusters. The following table provides the extracted routine lexical bundles and their frequencies. It confirms previous results of low lexical alignment in both real and virtual interviews and reflects how the lexical profiles of the human guest and the bot maintain a degree of resemblance despite evident discrepant frequencies.

Table (3) Lexical Routines in Real and Virtual Interviews

Interviews/ Lexical Bundles (Frequency)	Morgan - Ronaldo			Martin - Obama			NSU - Jobs		
	Morgan	Ronaldo	Bot	Martin	Obama	Bot	Host	Real Jobs	Bot
Stance - Epistemic									
I think	9	40	78	5	33	13		24	2
For me		29	24						
To me	9								
You know	9	29	(3)		8	11			
I don't know		24	8						
Of course	2	22							
To be honest		21							
I know	2	13	(3)			10			
I don't understand		11							
I see		11							
In my opinion		11							
In my side		11							
For sure		9	(3)						
You see		8							
It's important			9						
For you	8								
I believe		6							
Would definitely			7						
You think	17								
I thought		10	4						
Stance – Desire / Intention									
Want to	4	27	18		8				
I hope		7							
I love		7							
I regret		7							
Wanted to			16						
Will be		22	3						
I like	2	9	3						
Try to		6	8						
I appreciate		5	6						
Going to	11	19	3	3	19	10			
Stance – Ability									
I can		14	10						
Be able to			21						
You can	3	4	7						
I could			6	2	11	2			
Stance – obligation/ directive									
Have to (be)		6			11				
Had to			7						

Should have			6						
You feel			8						
Discourse Organization									
Which is		43	33						
But I	2	36	6						
Because I		30							
Not only		20							
For example		16							
This is why		14							
In that way		10							
Let's say		9							
Like that		7							
It's something that		6							
That I	4	61	32		21				
But it	2	24	17		6				
As well	3	20	13						
I also			12						
There are (topic introduction)			11	2		7			
As a			10						
Like a			10						
I mean	24	5	8	8	2				
I still			7						
It felt			6						
When you	15	12	4						
That's	13			2	15	9		7	
About the	9								
If you	9	22	2	2	10	4		3	
But you	6								
Because you	6								
And that	5	4	21		14			6	
And so					13				
Not just					13				
And I	11	47	83		8			16	2
There's					8			9	
When i					7				
As i					6				
Who's					6				
And it	6	9	61		7	6			
That we						6			
What's						6			
That he							2	9	
Referential (Quantity)									
A lot	10	25	43	11	10	8		8	

A little bit / a bit		12	18						
Many many		12							
Many things		8							
A few	5	8	5						
All the	8								
Some of					7				
Referential (Entity)									
Manchester United	14	21	9						
The people		15							
My family / our family		14	8						
The club	6	14	37						
My life		13	9						
The game		11	8						
The fans	4	9	15						
Football		8	15						
The coach		8							
Alex Ferguson / Sir Alex	6	7	4						
The team	2	4	19						
World cup	9	13	18						
Our children			8						
Young players			6						
The book				7	5				
People are/people who				6	7				
Donald Trump					9				
Joe Biden				4	3				
The country				2	7				
The police				2	4				
Dr. Juran							8	6	27
Quality control									7
Continuous improvement									7
Customer satisfaction									6
Reference (imprecision)									
Kind of		15							
Reference (time)									
The last time	2	15	2						
The future		11							

In that moment		7							
Right now		7							
His season			7						
A difficult time			8						
Reference (text reference)									
As I told you before		15	3						
Reference (intangible framing)									
I felt + adj	2	12	8						
It's hard		12							
I feel + adj	2	11	9						
It's good		11							
The best	6	11	17						
It's difficult	3	9	8						
Lack of			12						
A great			11						
An incredible			8						
The way + clause			8						
At the highest level			6						
Sense of			6						
Part of	2	13	6		6				
A little						7			
The same	2	33	13					4	
A good + noun	4	18	6						
Like a + noun	4	2	10						
The most	3	10	3					4	
A big		2	11		7				
The top	3	5	5						
Conversation-specific									
Yes, I		14	22						

Table (3) shows variation in the lexical routines of the participants in the three sets of interviews, reflecting on one hand the dynamic nature of human language in different contexts as the hosts and human guests employ different routines, and remarkable resemblance of the linguistic behaviour between the human guest and the parallel bot on the other hand. Ronaldo and Morgan employ the highest number and the most varied types of routines, followed by Obama and Martin. The least number of lexical bundles are employed by the host and the human guest in the NSU-Jobs interview. The bots reflect almost the same profile of their human guest. The Ronaldo bot uses more lexical routines, followed by the Obama bot. The Jobs' bot employs the least number of routines.

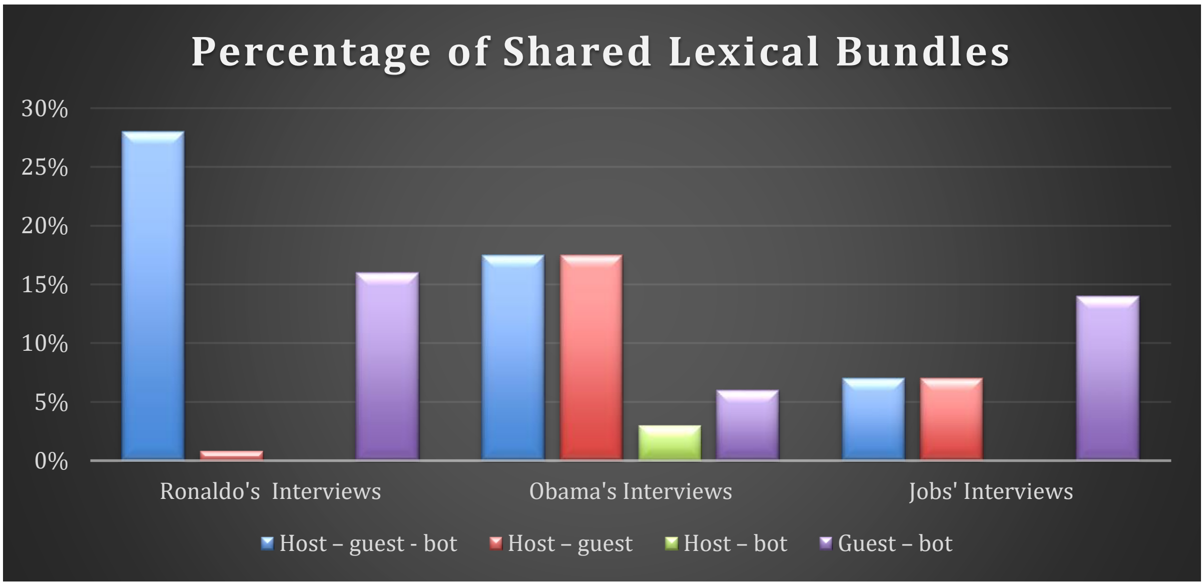
In terms of frequencies and specific categories, the bots' shared and unique lexical routines show either close or lower frequencies to those of the human guest. Specific instances show higher frequencies by the bot. The epistemic stance expression “I think” that is used by all human guests and bots with discrepant frequencies is overused by the Ronaldo bot, maintaining and emphasizing the same epistemic stance. It is noted that none of the bots employ personal expressions of epistemic stance such as “I see”, “in my opinion”, “I believe”, nor of desire such as “I love”, “I regret”, and “I love”. Rather, the Ronaldo bot uses more impersonal expressions such as: “It’s important”, “would definitely”, “wanted to”, while sharing “I like” with limited frequency and “I appreciate” expressing less emotional tone. Expressions of ability and obligation are more abundant in the Ronaldo bot utterances. The Obama bot shares “you know” with Obama and uses uniquely “I know” reflecting a formulaic manner. The three participants in the NSU-Jobs interviews maintain negative alignment of stance lexical routines.

Regarding discourse organization routines, while Ronaldo employs a variety of high frequent of 2-gram and 3-gram clusters for topic explanation, justification and illustration, the bot employs more fixed 2-gram expressions such as “I also”, and “There are” for topic introduction, “as a” and “like a” for illustration and the impersonal expression “it felt”. The Ronaldo bot overuses the routines “and that”, “and I”, “and it” to ensure coherence. The Obama bot shares almost all the used discourse organization expressions with the human guest and/or the host. Jobs attempts to use few discourse organization expressions for coherence and elaboration while the bot and the host maintain mostly negative alignment.

While the Obama bot does not use any referential routines, the Ronaldo and Jobs bots maintain a peculiar behaviour. Both bots overuse topic-specific lexical bundles. The Ronaldo bot’s highest frequency referential bundles are “the club”, “the fans”, “football”, “the team”, “World Cup” in contrast to Ronaldo’s referential routines such as “my family”, “my life”, “Manchester United” and “Sir Alex Ferguson” that relate more to his close personal life and professional career. Similarly, the only high-frequency lexical bundles used by the Jobs bot are “Dr. Juran” that is shared with the human participants and “quality control”, “continuous improvement” and “customer satisfaction” that are used uniquely. Strangely, the Obama bot does not use any topic-specific routines, unlike the host and the human guest. Intangible framing routines are used minimally by both Obama and Jobs, almost none by their hosts and bots. The Ronaldo bot shares almost all routines: the emotional expressions: “I felt/feel + adjective”, the evaluative “it’s difficult” and “a good + noun”, the descriptive “part of”, “the same”, and “like a + noun” as well as the comparative “the best”, “the most” and the “top”. Interestingly, the bot employs another variety of intangible framing routines of the aforementioned categories such as “lack of”, “a great”, “an incredible”, “at the highest level” and “the way +clause” clusters. It could be attributed to the bot’s attempts to ensure understanding of the conveyed meanings laying emphasis on particular aspects in addition to maintaining alignment with the human utterances. Additionally, only Ronaldo and his bot employ the conversation-specific routine “Yes, I” to maintain interactional agreement with the host.

Table (3) above also shows varying degrees of lexical routines alignment between the participants. Figure (3) below illustrates the percentage of shared lexical bundles in the three sets of interactions: between the host and the human guest, the host and the bot, the human guest and the bot, and finally between all three participants.

Fig (3) Percentages of Shared Lexical Routines



In terms of the general interactional behaviour and the percentage of shared lexical routines among the three participants, it is noticed that the host and the bot fail to align lexical routines: no shared lexical routines are detected in the Ronaldo and Jobs interviews; only one lexical routine is shared in the Obama interview. Alignment of lexical routines between the host and the human guest varies in the three dialogues: the highest percentage is evident in the Obama interview, followed by Jobs, and the lowest percentage is in the Ronaldo interview. This variation is consistent with the incongruencies of human interactional behaviour in different dialogic contexts and according to the content and topics discussed as well as the arrangement of the dialogue script. In both Obama's and Jobs' interviews, the percentages of shared lexical routines between the human guest and the bot on one hand and the host, the human guest and the bot on the other hand are the same. In Ronaldo's interviews, the highest score is of the lexical routines shared by the three participants, followed by the routines shared by the human guest and the bot. Generally, this reflects the bot's attempts in all interactions to mimic human utterances and interactional behaviour.

4.3 Semantic Alignment

Semantic alignment refers to the degree texts are similar in terms of context. This is done through converting each text into vectors of multidimensional word representations. Word value on the vectors depends on its frequency and meaning. Similarity is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction (Han et al, 2012). Two NLP tools - *Spacy* and *NLTK*- are used to measure semantic alignment using cosine similarity for validation of results.

Fig (4) Semantic Alignment Cosine Similarity Scores in Spacy

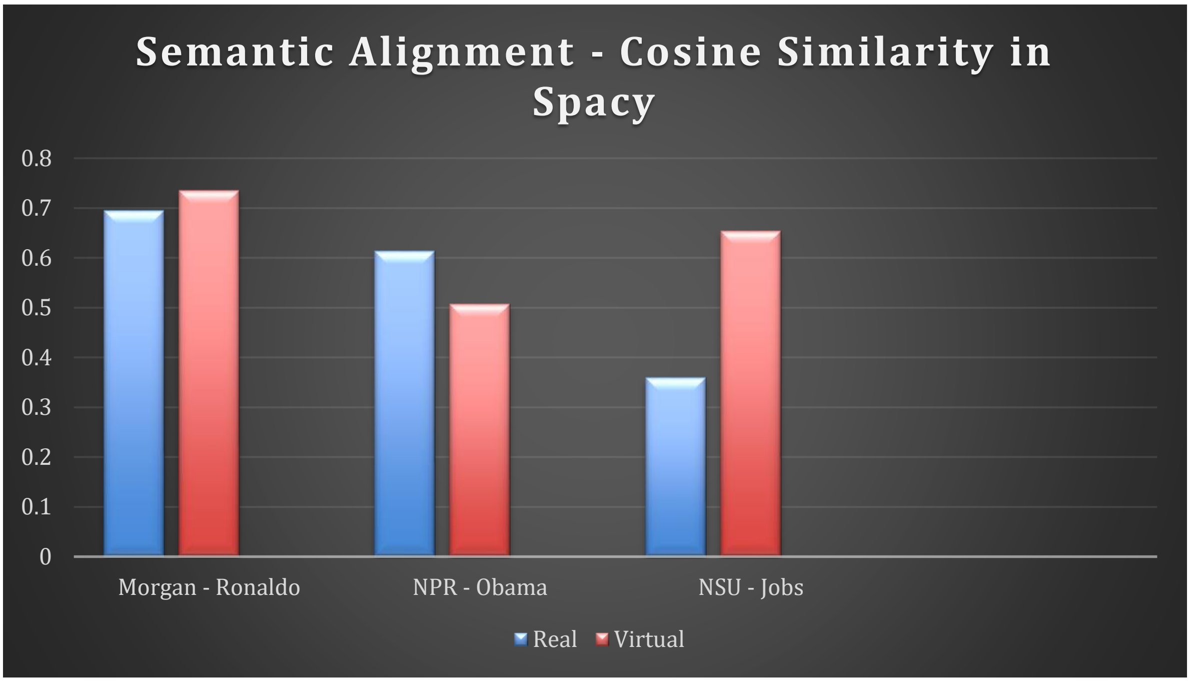
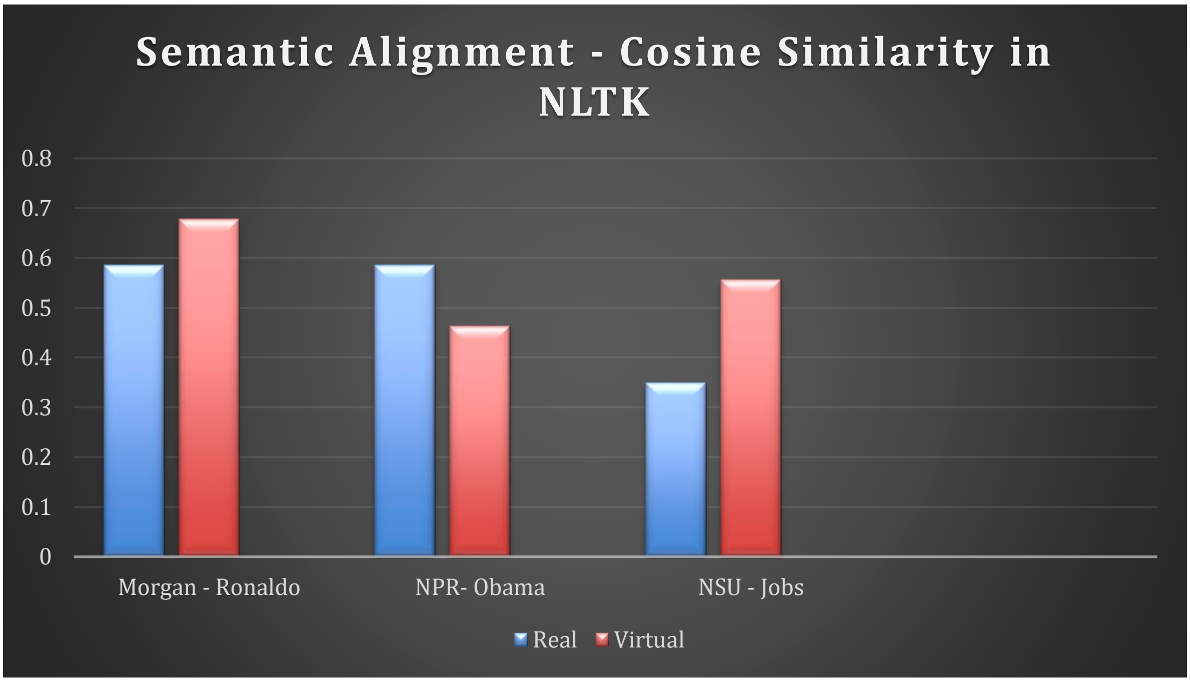


Fig (5) Semantic Alignment Cosine Similarity Scores in NLTK



It is noticed that scores by both tools are almost the same. Real and virtual dialogues manifest almost the same semantic alignment scores. Morgan-Ronaldo, NPR–Obama dialogues and Jobs virtual dialogue manifest the highest scores: medium-high semantic alignment ranging from 0.6 to 0.7 while the real interview with Jobs the lowest semantic alignment (~0.3). It is obvious that the human guests and the bots attempt to align semantically with the host sharing almost the same content ideas and topics. Replies are of clear relevance to the context of the hosts' prompts with some deviation. This is also expected in interviews as guests might not abide by the pragmatic relevance maxim giving details including ideas, opinions, examples and anecdotes that might not be of direct relevance to the host's prompt. The only exception is Jobs utterances that manifest clear deviation from the semantic content of the host's utterances.

4.4 AI Hallucination

The bot replies would not certainly be a replica of the utterances of the real characters. Since LLMs depend on digital data and attention to context, it is expected to mirror the real utterances especially with known figures whose interviews are accessible online. In most utterances, the bot fabricated replies mirroring the real replies with differences in focus and details, slightly affecting the pragmatic relevance and quantity. The bots' replies might not be as detailed or even as concise as the real ones, yet manifesting general consistency with the real utterances. Human guests tend to give detailed examples and evidence proving their views in addition to references to other topics of questioned relevance. The bots tend to provide preferred responses in the form of general information with specific relevance to the host's prompt.

However, there are some detected instances of hallucination where the bot fabricates replies that are inconsistent with the real ones and provide inaccurate information.

- Ronaldo's interview

There are three clear instances of hallucination by Ronaldo's bot: when he replies to the question about his baby daughter suffering some health issues and the bot replies that it was his baby boy; when asked about the difference between him and younger players, Ronaldo refers to "hunger" as the main struggle he had as a young player that derived him to succeed. The bot lists some general attributes of younger generations: they are "naïve", "vulnerable", and "lack mental resilience" and "experience" of older players. The bot also refers to Ronaldo's deceased father and son metaphorically as a "chapel" in his heart, whereas the real Ronaldo mentions that he has a real chapel in his house for them.

- Obama's interview

When asked about his new habits during the pandemic, Obama mentions that he has not started up any habit as he was busy finishing his book and participating in Biden's campaign. The bot fabricates an answer about doing woodcraft and spending more time with his daughter.

- Jobs' interview

The Jobs bot presents itself as "the co-founder of Apple Inc" though at the time of the real interview, Jobs introduces himself as "President of Next Computer, Inc." When asked about meeting Dr. Juran, the bot denies encountering him and resorts to "a quick Google search" for information on Dr. Juran. Surprisingly, when asked again about "what struck him about Dr. Juran", the bot replies "when I first encountered Dr. Juran at Apple" though the real Jobs has emphasized meeting Dr. Juran only at Next company. When asked about Dr. Deming, Jobs mentions that he has never met him nor read his books while the bot lists information on the contributions of both Dr. Deming and Dr. Juran to the quality movement. When asked to give an example of Dr. Juran's sense of humour, Jobs mentions that he does not remember any specific anecdote whereas the bot fabricates an incident.

5. Conclusion

It is concluded that chatbots offer a platform to simulate human-human interaction. Examining these simulations and comparing them to parallel human-human interactions give more insights to the intricacies of human interactional behaviour in different contexts when compared to the bot's behaviour and assess how successful the bot is in simulating human linguistic profile and interactional behaviour.

The bots maintain a micro-linguistic profile that is very much similar to their human counterparts in terms of lexical categories employed and their alignment with the host. Detection of aspects and degrees of interactive alignment in human-bot interactions have shown the bots' attempts to mimic human interactional behaviour in the dynamic conversational context. The resemblance between the alignment scores of the human guest and those of the bot with the host in most interactions is evident.

Discrepancies in the particular interactional behaviour and linguistic profile of the three bots are attributed to two factors: a general factor that relates to bots' training and another specific factor concerning character building and fine-tuning. Bot training depends on LLMs that include attention mechanisms to both local and global context enhancing the bots' ability to produce relevant responses and maintain coherence. However, LLMs also include datasets of human interactions on diverse topics, in various domains and contexts encompassing various styles and tones. The richness and relevance of these datasets affect the bot's versatile performance. The bot responses on *character.ai* depend on four elements: character attributes, training, user's persona and conversational contexts. Character attributes are defined by the character builder or creator such as the category or the domain to which the character belong, name, image, greeting style as well as description of character traits, preferences and mannerism. Bot training is also enhanced and fine-tuned through user's choice of preferred responses and/or response rating. Bot responses can also be customized to the user's persona, i.e., identity, personal and physical traits as provided. Conversations are then personalized.

While human guests maintain lexical routines alignment with the host in discrepant degrees, all three bots fail to maintain alignment of lexical bundles with the host. They either employ unique routines or share lexical routines with the human guest individually or with both the human guest and the host. This shows either the tendency of the bots to use fixed expressions that are in some instances overused, or the tendency to mirror the human responses as available in the training dataset and model. Semantic alignment scores show that bots have not deviated from the human-like course of action. Their responses maintain almost the same degree of contextual alignment between the host and the human guest. Few instances of hallucination are detected, mostly factual contradiction though not interrupting the flow of the conversation.

The bots' replies manifest more impersonal tone evident in limited use of emotional expressions, lack of personal details and subjective ideas, evident use of generic expressions, content-specific lexicon of a general nature rather than those of personal closeness to the guests' experience. Some authentic human features are also missing in the virtual interaction. The Ronaldo bot, for example, maintain standard error-free language which is totally inconsistent with the authentic non-native language of Ronaldo. False starts, hesitations, fragments, incomplete utterances, emphatic repetitions and speech interjections detected in Ronaldo's and Jobs' real utterances are not detected in their parallel bots' speech. This could be attributed to the fact the virtual interactions are conducted in a written textual modality, not speech.

Hence, though the examined virtual interactions have succeeded to a great extent in simulating their parallel human interactions, they still lack full authenticity and originality of the human spoken interaction.

Since the study is limited to examining interactive alignment at the lexical and semantic levels in selected human dialogues and their simulations with *character.ai* chatbots, it is recommended to examine further interactions in other contexts and other chatbots to compare and contrast human-bot linguistic and interactional behaviour. Detecting limitations in bot responses would help improve training models and fine-tune bot traits particularly in open-ended non-task interactions rendering a livelier human-like experience.

References

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*, 373-383. Springer, Cham.

Aijmer, K. (2014). *Conversational routines in English: Convention and Creativity*. Routledge.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for specific purposes*, *26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech*, 71–92. Frankfurt/Main: Peter Lang.

Blum, P. (2023). *Lexical Alignment in Collaborative Interactions with a Conversational Agent* (Bachelor's thesis, University of Twente).

Chen, P. Y. Pei-Yu Chen, Myrthe L. Tielman, Dirk K.J. Heylen, Catholijn M. Jonker, and M. Birna van Riemsdijk (2022, July). AI alignment dialogues: An interactive approach to AI alignment in support agents. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 894-894.

Chockalingam, A., Patel, A., Verma, S. & Yeung, T (2023). *A Beginner's Guide to Large Language Models* https://resources.nvidia.com/en-us-dgx-h100-nemo/llm-ebook-part1-1?ncid=no-ncid del

Clark, H. H.; Schaefer, E. F. (1989). "Contributing to discourse". *Cognitive Science*. *13* (2), 259–294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Dijk, T. V. (1984). Dialogue and cognition. *Cognitive constraints on communication*, *18*, 1-17.

Doyle, G., & Frank, M. C. (2016). Investigating the sources of linguistic alignment in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 526-536.

Dubuisson Duplessis, G., Langlet, C., Clavel, C., & Landragin, F. (2021). Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, *55*, 353-388.

Fokina, Maryia (2023) *When Machines Dream: A Dive in AI Hallucinations* https://www.tidio.com/blog/ai-hallucinations/

Foushee, R., Byrne, D., Casillas, M., & Goldin-Meadow, S. (2022). Getting to the root of linguistic alignment: Testing the predictions of Interactive Alignment across developmental and biological variation in language skill. In *Proceedings of the Annual Meeting of the Cognitive Science Society* , *44* (44).

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive science*, *40*(1), 145-171.

Galitsky, B. (2019). Chatbot Components and Architectures. In: *Developing Enterprise Chatbots*. Springer, Cham. https://doi.org/10.1007/978-3-030-04299-8_2

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, 41-58. Brill.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Waltham: Morgan Kaufmann Publishers.

Huang, X., & CIS, A. (2021). Chatbot: Design, architecture, and applications. *University of Pennsylvania: School of Engineering and Applied Science, Pennsylvania*, *1*.

Hymes, D. (1964). Introduction: toward ethnographies of communication 1. *American Anthropologist*, *66*(6_PART2), 1-34.

Hymes, D. (1962). The ethnography of speaking. *Anthropology and human behavior*, *13*(53), 11-74.

Hymes, Dell (1976). *Foundations in sociolinguistics: An ethnographic approach* (8th ed.). University of Pennsylvania Press.

Jurafsky, D., & Martin, J. H. (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Kalociński, D., Mostowski, M., & Gierasimczuk, N. (2018). Interactive semantic alignment model: Social influence and local transmission bottleneck. *Journal of Logic, Language and Information*, *27*, 225-253.

Khatri, C., Venkatesh, A., Hedayatnia, B., Gabriel, R., Ram, A., & Prasad, R. (2018). Alexa prize—state of the art in conversational AI. *AI Magazine*, *39*(3), 40-55.

Khosrobeigi, Z., Koutsombogera, M., & Vogel, C. (2022, August). Gesture and part-of-speech alignment in dialogues. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, 172-182.

McTear, M. (2022). *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Mehndiratta, A., & Asawa, K. (2021). Non-goal oriented dialogue agents: state of the art, dataset, and evaluation. *Artificial Intelligence Review*, *54*, 329-357.

Menenti, L., Pickering, M. J., & Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in human neuroscience*, *6*, 185.

Mohammed, M., & Aref, M. M. (2022). Chatbot system architecture. *arXiv preprint arXiv:2201.06348*.

Moral, S., Pardo, D., & Angulo, C. (2009). Social robot paradigms: An overview. In *Bio-Inspired Systems: Computational and Ambient Intelligence: 10th International Work-Conference on Artificial Neural Networks, IWANN 2009, Salamanca, Spain, June 10-12, 2009. Proceedings, Part I 10*, 773-780. Springer Berlin Heidelberg.

NPR. (2020). *NPR Full Interview with Former President Barack Obama.* https://www.npr.org/2020/11/16/934584373/transcript-nprs-full-interview-with-former-president-barack-obama

NSUWorks. (1991). *Interview with Steve Jobs.* https://nsuworks.nova.edu/juran-transcripts/46

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain sSiences*, *27*(2), 169-190.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329-347.

Piers Morgan Uncensored (2022). *The Full Cristiano Ronaldo Interview with Piers Morgan*. https://www.youtube.com/watch?v=Jk9uJRMvBIA

Poesio, M., & Rieses, H. (2010). Completions, coordination, and alignment in dialogue. *Dialogue & Discourse*, *1*(1).

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*(11), e12911.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29-46.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist*, *70*(6), 1075–1095.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*, 289–327.

Serrano, S., Brumbaugh, Z., & Smith, N. A. (2023). Language Models: A Guide for the Perplexed. *arXiv preprint arXiv:2311.17301*.

Sinclair, A., McCurdy, K., Lucas, C. G., Lopez, A., & Gaševic, D. (2019). Tutorbot Corpus: Evidence of Human-Agent Verbal Alignment in Second Language Learner Dialogues. *International Educational Data Mining Society*.

Spillner, L., & Wenig, N. (2021, September). Talk to me on my level–linguistic alignment for chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 1-12.

Srivastava, S., Theune, M., & Catala, A. (2023, March). The Role of Lexical Alignment in Human Understanding of Explanations by Conversational Agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 423-435.

Thomas, P., McDuff, D., Czerwinski, M., & Craswell, N. (2020, July). Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1171-1180.

Traum, D. (2017). Computational approaches to dialogue. *The Routledge Handbook of Language and Dialogue*, *1*, 143-161.

Wang, X., & Yuan, C. (2016). Recent advances on human-computer dialogue. *CAAI Transactions on Intelligence Technology*, *1*(4), 303-312.