# Nonverbal Indicators of Comprehension Among L2 Users of English Interacting with Smart Verbal Software Agents

**Abdulmalik Yusuf Ofemile**

**Abstract** Pervasive computing has engendered increasing interaction between speakers of English as a second language (EL2) and intelligent software agents using English as a first language. This extends discourse to contexts such as satellite navigation systems giving drivers directions, self-service systems in banks, and computer aided language learning (CALL) devices in Nigeria. Additionally, most research around listenership in Human-Agent Interaction (HAI) has focused on assessing listener feedback using verbal feedback or posed nonverbal behaviours with little attention paid to listener spontaneous nonverbal behaviours. This chapter reports a scoping study aimed at developing a better understanding of the nature of marked spontaneous nonverbal listenership behaviours displayed and their impact on listener-comprehension during interaction. Ten student-teachers of English were tasked with assembling two Lego models using vague verbal instructions from a computer interface and one human instructor within two 15-min interactions. The study used a continuum of four voices comprising two synthesised voices, one by a voice actor and another by a human instructor. A 5-h long multimodal corpus was built and analysed from these interactions. The results suggest that it is possible for humans to show their level and process of comprehending agent instructions through facial actions, nonverbal private talk and repairs during interaction. Furthermore, there is a potential for formulating a theoretical basis for researching interaction in similar contexts. Findings suggest that enhancing agents' emotive functionality may enhance HAI in English language learning contexts, but this requires further research.

**Keywords** Human-Agent Interaction (HAI) · Facial actions · Nonverbal listenership · Software agents · Private talk · Comprehension

A. Y. Ofemile (✉)
FCT College of Education, Zuba Abuja, Nigeria

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

R. Al-Mahrooqi, C. J. Denman (eds.), *Individual and Contextual Factors in the English Language Classroom*, English Language Education 24,
https://doi.org/10.1007/978-3-030-91881-1_18

# 1 Introduction

This chapter reports a scoping study carried out in Nigeria in 2015. Nigeria has about 519 living languages (Simons & Fennig, 2018), where Hausa, Igbo, and Yoruba are the most widely spoken, and English is a national language. While Nigerian languages share some roles with English in national life, English is predominantly the language of education, governance, commerce, and general interaction among Nigerians.

The Nigerian interaction context is changing rapidly and shifting from being solely for human-human interaction (HHI) to Human-Agent Interaction (HAI). Agents are described as highly inter-connected computational components capable of acting autonomously and intelligently (Jennings et al., 2014). These include intelligent personal assistants (IPAs), like Amazon's Alexa and Apple's Siri, and Embodied Conversational Agents (ECAs) used as instructors/advice-givers (e.g. sat-navs and map applications, automated checkouts in supermarkets, ATM cash dispensers etc.). This chapter discusses the use of simulated agents, i.e. rather than being commercially available software agents. The simulated agent is a simpler bespoke interface that takes in keyboard-based commands and returns appropriate speech-based output.

Results from the author's own study in the UK indicated that participants who were speakers of English as a first language nonverbally projected their comprehension and incomprehension of the agent's vague instructions and language use. Hence, the conclusion was drawn that agents should be adaptive to user linguistic capabilities and context since no one size fits all. This conclusion motivated a scoping study in Nigeria that aimed to understand how users in other contexts of English language usage comprehend and display their comprehension or incomprehension of L1 agent instructions during interaction. The Nigerian study enlarged the UK study's scope in terms of eliciting circumstances, research population, expanding the cline of voices from three to four, and replicating the study in a lower technology context where English is used as a second language.

# 2 Literature Review

## 2.1 *Emerging Hybrid Space in Nigeria*

Hybrid spaces of interaction are created from the meeting of interlocutors from different linguistic/cultural groups as outlined in Bhabha's (1994) theory of hybridity. Hybridity is also created from differences between entities with distinct interactional competences as seen between humans and smart devices within hybrid spaces (Ofemile, 2018). Furthermore, with pervasive ubiquitous computing, agents possessing the communicative abilities of users of English as a L1 and multilingual users of English as a L2 intersect, and a hybrid interaction context is created in

Nigeria. This, according to Simpson (2017), enables intersemiotic and interdiscursive practices to evolve in a participatory manner. The first describes switches between spoken and written, visual and verbal language and non-linguistic signs, while the second describes language use that occurs when unfamiliar discourse is experienced in intercultural interaction.

Humans, unlike agents, are capable of sensory functions like the detection of stimuli, perception, flexible and ingenuous innovation, inductive and deductive reasoning, and judgement. Agents, on the other hand, have speed, multitasking, computational and deductive reasoning, flexible autonomy, agile teaming and crowd sourced information gathering attributes, which constitute a culture that gives them a shared way of doing things in a way that is distinct from humans (Ofemile, 2018). Furthermore, human body gestures, unlike agent gestures, evolve over time and agents do not have natural gestures, faces or limbs (Dautenhahn, 2013).

These differences between humans and agents are so significant that each has what can be described as natural behaviour to them. Thus, whenever humans interact with agents, each of them brings distinct interaction patterns derived from their cultural backgrounds to play during interaction. For example, when withdrawing money from ATMs, one is made to conform to specific patterns of behaviour to get desired results. In the course of this interaction, agents give verbal and/or written instructions while humans react in specific ways to get money out. Humans bring to play experience, reading, writing, speaking, and listening skills, as well as knowledge about the language of communication within that space. Agents bring to play speed, deductive reasoning, multitasking and computational skills by recognising passwords, disbursing correct amounts of money, and taking the withdrawer's picture.

The scenario is replicated severally when we access our emails, make phone calls, upload documents, learn or teach online, swipe our identity cards at entrances of offices or restricted areas to gain access, or even when making payments online. These behaviours are normalised in us and so are taken for granted in our daily lives, but when we remember our very first attempts at these actions, memories of false starts, negotiations and trials come back.

Similarly, agent characteristics are normalised internally in the agent's personal identity and linked to its verbal capabilities which may cause users to categorise agents as having a particular linguistic or vocal property. This is linked to the notion that agents conforming or not conforming to listener expectations may inform researchers on how users socially position agents during interaction (Clark et al., 2015).

## 2.2 *Listeners Comprehension Process*

Listening is a vital process for effective communication that provides input from interlocutors receiving aural stimuli and giving meaning to it (Nunan, 2002; Oxford, 1993). Researchers in Teaching English as a Second Language (TESOL) and applied linguistics generally accept that top-down processing and bottom-up processing interpretation theories can be used to explain how listeners decode speaker input (Harmer, 2007). Bottom-up processing holds that listening is a process of decoding speaker input incrementally beginning with the smallest meaningful unit of language reshaped into larger complex texts (Field, 2004). Thus, listeners progress upwards, decoding and linking smaller units to larger ones in order to make meaning of speaker input.

The top-down processing view argues that, "Larger units exercise an influence over the way in which smaller ones are perceived" (Field, 2004, p. 364). This suggests that listeners reconstruct speaker-meaning using aural stimuli as a guide; for example, listener interpretation of phonemes depends on their knowledge of that particular word. Therefore, listening is not sequential; rather, it is a framework of the two strategic actions of 'decoding' and 'meaning building' in which one runs into the other (Field, 2008). Although these two views seem like opposites, they are actually complimentary as outlined below.

Krashen's (1982) Comprehension Hypothesis maintains that we acquire language and develop literacy when we understand messages as 'comprehensible inputs' which are comprised of the things we read and hear. Thus, the more comprehensible speaker utterances are, the better a listener's comprehension during interaction. Listeners understand speaker utterances using their own knowledge, systemic information derived from context, inferential schemata and systemic processes (see Fig. 1).

In the figure, the downwards solid arrows indicate the predictive nature of listening where listeners continue to guess or be in a state of anticipation of what speakers will say or mean with each utterance, also called 'listening out' by Lacey (2013), based on information available. The upwards broken arrows indicate how listeners incrementally process speaker input as information while giving feedback. These arrows do not signify separate routes for distinct processes; rather, they suggest a continuous loop that meets at different points within seamless boundaries indicated by horizontal broken lines. Interaction between information sources and meaning making during comprehension may depend on listener role, listening purpose, text listened to and speaker verbal characteristics.

Harmer (2007) suggests that comprehension is activated by listener's schemata. Here schemata refer to inferential schemata, described as "the ways that successive turns in talk can be interpreted" (Coulthard et al., 2016, p. 10). Thus, when listeners relate speaker input to specific interaction contexts, such as instruction-giving, they interpret successive instructions using schemata and this may support and lead to the development of procedural knowledge.

Schematic knowledge comprises 'background knowledge' and procedural knowledge featured in Fig. 1. Background knowledge, or propositional knowledge, includes facts that listeners bring into interactions, such as knowledge about topics of discussion and implicit knowledge. Vandergrift (2011) suggests that strategic listeners unintentionally develop implicit listening comprehension knowledge performatively by using it unconsciously as a social asset without being aware of such knowledge, as seen in spontaneous turn-taking, facial actions and gestures displayed

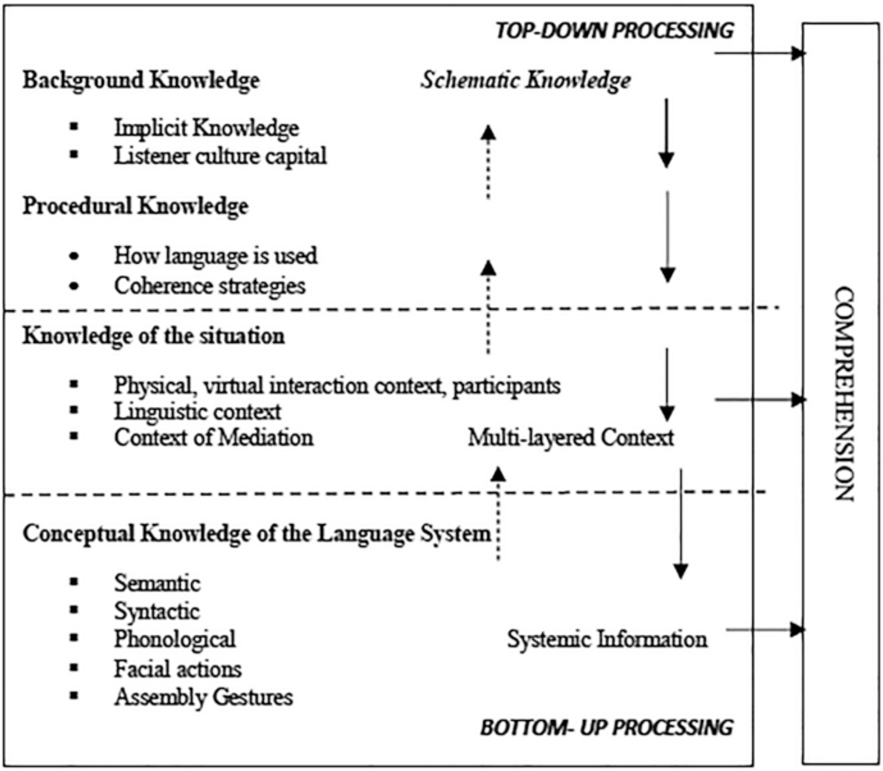


**Fig. 1** Listener's comprehension process. (Ofemile, 2018, p. 14)

during interaction. Conversely, explicit knowledge is deliberately and consciously developed by listeners.

Listeners also use contextual information to compensate for inadequacies in communication media and, as Field (2008) observes, listeners seem to use more top-down information to compensate for gaps in their understanding during interaction. Listener-compensation strategies include schemata activation by making inferences from explicit meanings of speaker-utterances to intended conclusions (Mazzone, 2015). In task contexts, a frame like “select X” activates in listeners the schema of ‘instruction taking’ and the need to respond with suitable language frame or action, such as picking X or asking for clarification. The activated schema represents listener-culture capital brought into a new context.

Procedural knowledge is the sum of what listeners know of steps and actions done to achieve a goal (Rittle-Johnson et al., 2015), with a focus on how language is used in given contexts. Procedural knowledge compares favourably with Canale and Swain's (1980) strategic communicative competence as procedural knowledge allows listeners to manage their communication and to negotiate meanings, codes, and identities in order to achieve interaction goals.

Another source of knowledge is the multilayered interaction context. The notion of multilayered context is used here to project the understanding that, when people interact with technology, the context has layers, such as interaction context, linguistic context and mediation context. Interaction context refers to socially contextual situations where speech takes place as speech events or speech activities (Levinson, 2016). The concept is applied here to focus on listener comprehension processes in unidirectional instruction-giving contexts.

Forms of discourse used in interaction constitute another dimension for characterising context as linguistic, where linguistic contexts refer to descriptions of occurrences of semantic and syntactic forms of language used in interactions, including parts of speech (Ofemile, 2018). However, linguistic contexts require socio-affective strategies in order to build interaction among interlocutors, such as bootstrapping, which is the ability to predict relationships between linguistic forms and meanings (Huang & Arnold, 2016).

In addition, listeners require contextual cues to successfully make meaning of speaker utterance. Adolphs (2008) explains that contextualisation cues are used to analyse relationships between surface structures and context. These could be lexical, linguistic, paralinguistic or prosodic, indicating contextual suppositions at different discourse levels, and are used to encode speaker-expectations about upcoming discourse. Listeners use these features to enrich meaning, make it relevant, aid decoding processes and influence listener-orientation towards interactions.

Using a simulated agent-instructor with three voices in this study creates the context of mediation, which has been described as, "The physical medium of utterances and how it interacts with other interlocutors and layers of context" (Chun et al., 2016, p. 68). This implies that listeners' experiences in HHI and HAI contexts are influenced by media used in interactions, hence the need to compare behaviours emerging from both contexts.

The systemic source of information relates to listeners' conceptual knowledge that provides building blocks for bottom-up processing. Conceptual knowledge is the knowledge of abstract and general principles of language systems (Rittle-Johnson et al., 2015). This comprises verbal (semantic, syntactic, phonological and prosodic) and nonverbal (facial actions and gestures) components of languages that form aspects of co-text used by listeners during interaction. It also comprises knowledge of the language code (grammar and vocabulary), the conventions of its nonverbal and spoken forms, as well as written representations. These are critical to listeners' competent processing of speaker input, i.e. comprehension, because competence influences listener feedback and attitudes towards interaction.

Thus, Field (2008), citing Osada (2004), holds that listeners who are less skilled tend to either spend more time decoding unfamiliar words or have greater reliance on context to decode as they either pay too much attention to details or lack the linguistic competence necessary to properly decode utterances. Further, Hendrikse et al. (2016) posit that attitudinal reactions are concerned with a listener's willingness and ability to react and/or respond to the speaker's utterances and to reject or accept the message verbally and nonverbally.

### 2.3 Projecting Nonverbal Listenership

In order to assess how L2 users of English language comprehend L1 agent instructions and project their comprehension or incomprehension it is useful to understand the role of nonverbal listenership in interaction. Listenership is “the active, responsive role that listeners have in conversation” (O’Keeffe et al., 2007, p. 142). This implies that, in collaborative language use, listeners participate actively in any discourse even when they are not talking using backchannels as expected of them by speakers.

Backchannels, as proposed by Yngve (1970), are listener responses during one-way communication that can be verbal or nonverbal expressions, such as gestures and facial actions. White (1989) explains that backchannels imply that there are two channels of communication used by speakers and listeners. Speakers use the main channel, while listeners use the backchannel to interject speakers without claiming the floor. Backchannels are used to maintain the flow of conversation, indicate listener agreement with speakers, show listeners are paying attention to the speaker, and indicate that the information uttered is of interest to listeners and may be evaluative (Zimmerman, 1998). For the purpose of this chapter, backchannels include marked spontaneous facial actions and gestures emerging from interaction.

Spontaneous facial actions are “unmodulated emotional facial expressions that are congruent with an underlying emotional state” (Hess & Kleck, 1997, p. 271), while Givens (2015), citing Soukhanov (1992, p. 762), describes a gesture as “a motion of the limbs or body to express or help express thought or to emphasize speech”. These definitions imply that spontaneous facial expressions will often agree or align with associative expressions, including in terms of voice, gesture or posture indicating fluency in communication. Facial actions are fundamentally related to emotions that are universal to people because specific facial muscles express specific emotions (Ekman, 2007). However, some facial actions express emotional attitude.

Gestures evolved alongside speaking and listening and with the advent of literacy, reading and writing. There are two broad classes of gestures. The first co-occur with speech and are variously called ‘co-speech’ (Mol et al., 2012), coverbal gestures (Xu et al., 2009), or illustrators (Ekman, 2007), that depict some content of the message. In a related development, Kita et al. (2017) suggest that people gesture when they think silently using co-thought gestures.

There are representational gestures, or emblems (McNeil, 1992), that convey semantic meaning through hand shape, position, or motion, and that do not occur with speech or content. They are used to produce and deliver spontaneous verbal messages encoded in body movements in an enculturation process. Examples include, “The peace sign (forefinger and middle finger up, palm facing outward) or ‘good’ (thumb up, hand in fist)” (Matsumoto & Hwang, 2013, p. 2). Emblems are useful for communicating from long distances and in noisy places, such as crowded halls. As communicative and discourse-oriented gestures, they offer a channel for observing psychological activities that take place during interaction (McNeil, 1985).

In intrapersonal communication, they perform expressive functions, such as externalising listener comprehension of speaker-utterances through repairs as dialogic and monologic hesitation during interaction.

Representational gestures are also used to represent forms of task objects and the nature of actions to be used with those objects, scaffold conceptual development, provide, clarify and coordinate instructions during assisted assembly tasks (Kirk et al., 2005). Listenership repair systems are actual corrections of factual errors or faults in content (Frenečik, 2005; Knight, 2009). The types of repairs that occur in conversation include self-initiated self-repair (SISR), other-initiated self-repair (OISR), self-initiated other-repair (SIOR), and other-initiated other-repair (OIOR) (Clark, 2012).

This chapter reports on a study that attempts to understand how people use nonverbal behaviour to project their comprehension or incomprehension of agent instructions in a unidirectional instruction-giving context; thus, the type of repair relevant to this study is SISR because the "other" in this case is a simulated agent that can only give instructions. Clark (2012), citing Levinson (1983, pp. 340–341), defines SISR as repairs that speakers of utterances that need repair make without prompting from another participant. The concept is extended here to describe repairs of assembling errors that listeners taking instructions carry out without prompting from instructors during tasks.

The Computers are Social Actors (CASA) paradigm, later elaborated as Media Equation (M-E) theory by Reeves and his colleagues, is used to explain why and how humans behave when interacting with agents. The theories present user responses to physical and social features of computers and software agents in various settings of Human-Computer Interaction (HCI).

The CASA paradigm holds that, during interaction, people treat computers and computerised spaces as real people and spaces (Nass et al., 1994) based on a number of premises. The first is that people applied social norms and notions of self and other to computers or agents, responded socially to the computer itself, and did not see the computer as a medium of social interaction with the programmer. Secondly, CASA suggests that basic human communication devices are powerful because of their control and influence over people and events while social responses, such as facial actions and gestures, are automatic because they are naturally ingrained in us.

As a progression from CASA, M-E holds that social rules guiding interactions with people can apply equally to HCI, thus an "individual's interactions with computers, television, and new media are fundamentally social and natural, just like interaction in real life" (Reeves & Nass, 1996, p. 5). 'Social' refers to the disposition to treat media as if you were interacting with another person, while 'natural' refers to the disposition to treat media as if you were dealing with a natural physical environment. The medium becomes invisible and the human is oriented to its socialness or what is being seen (Reeves & Nass, 2014). Thus, people can be flattered by computers, similar to how they would be with other people, and perceive computers as having personalities similar to humans, while even small changes in creating these perceived personalities could elicit social behaviours from their users (Nass et al., 1999).

Reeves and Nass (2014) suggest that people respond socially and naturally to media because human brains have not evolved to adapt to twentieth century technology as the human brain evolved to accommodate a world where emphasis is placed on human-displayed social skills and perception of objects as physically real. However, Barrett (2012) posits that the human brain's plasticity enables it to adapt to emerging technologies due to pervasive computing in the everyday interaction context.

This chapter argues that nonverbal behaviours can project listener comprehension which is constantly changing, depending on the text interacted with. These changes occur naturally due to text contents and structure as mediated by our cognition (i.e. how we interpret information), disposition (how we feel at the time), and environmental factors (things taking place outside the person's body). Thus, it becomes important to understand how listener comprehension is affected by verbal and linguistic characteristics of co-interlocutors, how listeners project the impact using gestures and facial actions, and how these can be used to improve the agent's emotive functionality and associated user-experience in English language learning contexts.

## 3 Methodology

As stated above, this scoping study examined the listenership behaviours of participants from Nigeria where English is spoken as a second language. It aimed to understand user nonverbal projection of their comprehension of assembly instructions from L1 speaking verbal agents. Specifically, it sought to answer these research questions:

RQ1: Do participants nonverbally display comprehension signals (facial actions and gestures)?

RQ2: Are there differences in user nonverbal projection of their comprehension or incomprehension of instructions across the voice cline, e.g. human versus simulated agent?

### *3.1 Agent Design and Human Instructor Choice*

A simulated agent was created on a computer interface for the study instead of a real agent because it provides users with experiences similar to those which actual agents provide (Clark et al., 2014). The interface allows participants to repeat instructions, but they cannot return to previous instructions – a condition imposed to ascertain self-propelled behaviours in participants.

There are three ranges of voice progression in the continuum, namely synthesised, human-like, and the target voice (see Fig. 2). The synthesised voices include

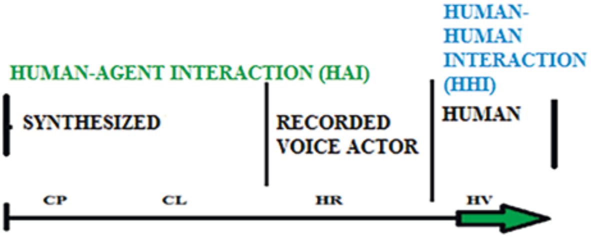


**Fig. 2** The voice continuum

Cepstral Lawrence (CL)[1] and Giles from CereProc (CP)[2]. The recorded human voice was provided by a professional voice actor (HR) hired from http://voicebunny.com and the human instructor (HV) used his natural voice.

The three agent voices were male, had a southern RP English accent, and were aged between 45 and 60 years old. In creating the synthesised speech instructions, text files were inputted into a text-to-speech program (Text2SpeechPro) and exported as .wav files. For the human recording, a voice actor recorded the same text which was edited into individual .wav files using the software program Audacity[3].

HV is a Nigerian, 45-year-old male L2 speaker of English with an accent that is generally understood by most educated Nigerians. HV was used for this study because he is an experienced teacher of English and teacher trainer and has an excellent understanding of Educated Nigeria English (ENE) or Popular Nigerian English.

The instructors provide interaction as envisaged in real life HAI and HHI contexts, which made interactions natural and familiar as participants are used to taking instructions from agents and people in various contexts.

## *3.2 Participants and Task*

Purposive sampling was used to select EL2 speakers as the target population in order to understand how they will respond to vague instructions during interaction. Ten participants were self-selected from among student-teachers of English at the FCT College of Education Zuba-Abuja, Nigeria. There were five males (50%) and five females (50%) aged 18–24 years old. These students have studied English at the further education level for at least two years, and can be classified as independent speakers of the language using Common European Framework of Reference for Languages (CEFR) descriptors. Participants were given consent forms to indicate agreement and willingness to participate in the experiment followed by

[1] See: https://www.cepstral.com

[2] See: https://www.cereproc.com

[3] For more information: http://www.audacityteam.org

demographics forms for personal details. The task assignment was randomised to make sampling counterbalanced when allocating slots, voices, tasks and timings to participants, because it is straight forward, simple, and eliminates clustered selection (Dörnyei, 2007) and makes population and corpora balanced without researcher bias.

Participants were provided two Lego models in two separate tasks and were briefed that they will construct two different Lego models using verbal instructions from a simulated agent on a computer interface (Fig. 2) and/or HV within a 15-min time limit per model. In each task, the interaction pattern involves participants asking for instructions (by pressing the start button in HAI or giving the thumbs up sign in HHI). Participants can ask for a repeat of the same instruction by pressing the repeat button in HAI or raising the forefinger in HHI.

After executing the instruction, participants ask for the next instruction by pressing the next button in HAI and giving the thumbs up sign in HHI. The first model was given to them and, after 15 min or after the model was completed, the second model was assigned.

### *3.3 Data Collection*

A clear record was obtained of interactions observed for measurements using two digital video cameras (Knight, 2009) that recorded the interactions from two angles – face and side (see Fig. 3).

The two cameras enabled the researcher to record both the individual sequences of body movements of different positions of the listeners during interaction and allowed for the analysis of synchronised videos in order to enable the examination of coordinated movement (across each view) following Knight (2009). However, power cuts forced the second camera to malfunction and the bright light from the window in the poorly lit laboratory made shots from the side camera blurry; thus, one digital camera was used to record face views of other interactions. While this made recordings less dynamic, it ensured that all recordings were acceptably clearer.

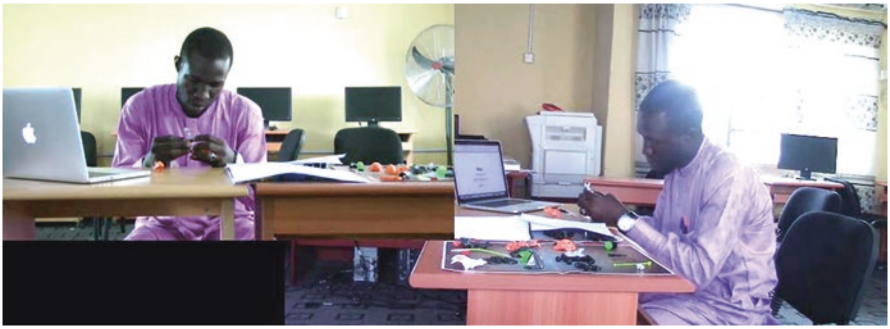


**Fig. 3** Side and front camera views of interaction

A 5-h-long multimodal corpus was built and analysed from these interactions. ELAN-EUDICO Linguistic Annotator[4] (Wittenburg et al., 2006) and CLAN[5] (Computerized Language Analysis) software were used to manage and annotate the corpora for facial actions and gestures using adapted schemas, coding systems, and hierarchies (Ekman et al., 2013; Feng & O'Halloran 2013).

### 3.4 Data Analysis

Only samples that are representative of families of nonverbal behaviours displayed are presented because they have all the basic features of each family. Data analysis aims at developing a 'thick description' (Dörnyei, 2007) of the emerging nonverbal behaviour which relies on an annotation scheme that identifies annotation tiers, values and their descriptions. Thick description occurs in three stages: (1) linguistic annotation and segmentation; (2) classification; and (3) establishing interactive, communicative and task functions of nonverbal listenership behaviour (Ofemile, 2015).

Linguistic annotation of facial actions is done by first describing the neurobiological processes generating the facial action, such as the facial muscles responsible for an expression. These are described as Action Units (AU) outlined in Ekman's Facial Action Coding Scheme (FACS) (Ekman & Friesen, 1978a, b). The neurological processes also relate to the five senses that elicit emotions through stimulation (Ekman & Friesen, 1978a, b; Mixdorff et al., 2017). For example, when people enjoy touching, seeing, hearing, smelling, tasting, or experiencing something they smile.

Gestures are described at the segmental level in order to capture every movement (Kendon, 1980; Kirk et al., 2005; Kita et al., 1998) along the G-phrase. Kendon (1980) designed a detailed kinesis structure – the Gesture unit (G-Unit) – which was later expanded and called the Gesture phrase (G-phrase) (Ofemile, 2018). The G-phrase comes with hierarchies for analysing gestures and defined terms required for implementation. The G-phrase begins with a rest pose then several gestures consecutively occur in succession, and it ends with another rest pose. The five procedural phases of gestures are preparation, stroke, retraction, holds, and recoils (Kendon, 1980; Kita et al., 1998; Ofemile, 2018). Following Zwitserlood et al. (2008), this study uses descriptive tiers grouping gesture under major headings, such as hand shape, position, and orientation.

[4] See: https://tla.mpi.nl/tools/tla-tools/elan/download/

[5] See: http://childes.talkbank.org/clan/

### 3.5 Colour Coding

Research indicates that colour coding reduces confusion on the part of the reader and aids object detection (Dalal & Triggs, 2005; Papageorgiou et al., 1998). Thus, to make gesture analysis clearer, this study uses a simple colour and letter coding as this draws attention to specific aspects of the annotation. Table 1 features the colour and letter codes used. When the participants' clothes are the same or have similar colours with those specified, a contrasting colour is used.

The next stage is to determine the classification of the nonverbal listenership behaviour. Such categorisation enhances systematic analysis as a form of typology; thus, gestures and facial actions are categorised according to families (Ekman, 2007; Kendon, 2004). While the kinesic structure enables easy and systematic annotation, transcription and classification of gestures within the expanding boundaries of the G-phrase, research suggests that evolutionary and innate factors are responsible for nonverbal action characteristics in facial actions (Ekman, 2007). Within each family, there are varieties of nonverbal behaviours that are identifiable as distinct because of their manner of execution and communicative functions.

Linguistic segmentation and categorisation also focus on descriptive arrangement and discussion of the most frequent linguistic and multimodal bundles considered as indicators of agreement and variation in listenership behaviours, such as those observed in the listener comprehension process (Field, 2008; Oxford, 1993). As outlined earlier, listener nonverbal behaviour and their emerging functions may be shaped by listener cognition disposition, culture, relationship between interlocutors, interlocutor's state of mind, environmental factors and interaction context with the aim of identifying emerging comprehension patterns useful for rule setting (Ekman, 2016; Kita, 2013; Sekine & Kita, 2015). Following this, the analytical focus shifts to understanding the communicative practices of participants as listeners and motivation for communicative behaviour during interaction.

A second rater analysed and annotated videos of randomly selected interactions using the annotation scheme designed to provide another perception of listener nonverbal listenership behaviour and ascertain inter-rater reliability (IRR) as separately

**Table 1** Colour codes

| S. No | Colour code | Meaning |
|---|---|---|
| 1 | RH<br>→ | Right Hand code and movement |
| 2 | LH<br>→ | Left Hand code and movement |
| 3 | ↔ | Bidirectional Right Hand |
| 4 | ↔ | Bidirectional Left Hand |
| 5 | - - - - ▶ | Intended movements |

done in Clark et al. (2016) and Kita et al. (1998). Multiple methods were used for measuring IRR. These include percentage agreement or rule of the thumb analysis of first and second annotators' perceptions to establish the coefficient correlation, which is traced to their Kappa values (Gwet, 2012; Lombard et al., 2010). The other method uses inter-rater agreement calculator software (Geertzen, 2012) to assesses IRR and measure the corresponding Fleiss' Kappa (*K*) and Krippendorff's alpha (α) to establish an acceptable benchmark of 75% as suggested by Gwet (2012) and Wongpakaran et al. (2013). The percentage agreement between the two annotators for nonverbal behaviours displayed is 96% and 88% respectively. The resulting Kappa indicates almost perfect agreement and falls within the Landis and Koch, as well as the Altman, benchmarks of .81–1.00 (Gwet, 2012).

## 4 Results

The study examined the two research questions detailed in Sect. 3. The research questions are premised on research indicating that facial actions as emotive cognitive activities and bodily responses are controlled by the brain and may enable us to understand listener nonverbal feedback (Fortin et al., 2010).

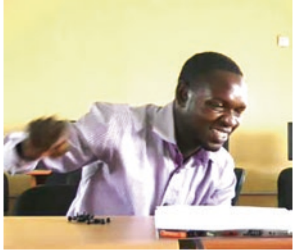
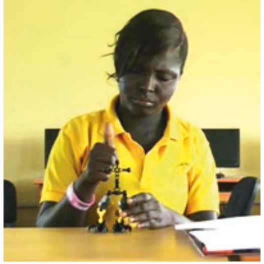
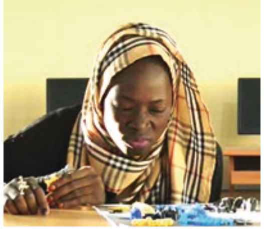
### *4.1 Listener Facial Actions*

Facial actions may externalise a person's attitude towards co-interactants, situation or task as positive, neutral or negative (Mehrabian & Ferris, 1967). Consequently, facial actions displayed may also signal listener comprehension during interaction. Positive facial actions suggest likeability; thus, participants' smiles can be elicited by positive stimulation, including amusement, delight, contentment, satisfaction, beatific experiences, relief from pain, pressure or tension and success (Ekman et al., 2013; Ofemile, 2018), such as the felt smile.

The felt or Duchenne smile (A in Table 2) is made possible by the following facial movements: AU6 (Cheek raiser – orbicularis oculi; pars orbitalis) raises the cheek, gathers the skin around the eyes inwards, narrows the eyes apertures and produces crow's feet wrinkles; AU12 (Lip corner puller – zygomaticus major) pulls the lips sideways exposing the teeth; and AU7 (Lid tightener – orbicularis oculi; pars palpebralis) tightens the eye lids, raises the lower eye lid creating wrinkles below the lower eye lid. Participants also displayed other smiles, such as tight-lipped with closed lips, partial-half open lips, and nervous smiles indicated by sadness in the eyes during interaction.

The neutral face (AU0) indicating indifference (Ekman, 2007), as displayed by B in Table 2, does not show any emotion as facial muscles are at rest. This occurs when participants are listening to instructions (neutral concentration on instructions), engaged in task execution (neutral concentration on task), face down (when

**Table 2** Comprehension levels inferred from spontaneous listener facial actions

| Representative listener facial actions | | | |
|---|---|---|---|
| Ref | A-Positive | B-Neutral | C-Negative |
| |  |  |  |

cognitively processing information), neutral hard (when experiencing difficulties in cognitive processing or task execution), and when about to initiate turn-taking by asking for fresh and repeated instructions.

Participants also displayed the following negative microexpressions indicating anger, disgust and sadness. Microexpressions, also known as hot spots, are described as very fast facial actions that last between 100 microseconds and 500 microseconds. They provide the greatest source of information leakage from the human face even when people try to conceal emotion during interaction (Ekman, 2007; Ofemile, 2018). The participant (C in Table 2) displays full scale disgust produced by AU44, a separate strand of AU4 brow lowerer that narrows the eyes, AU9 nose wrinkler, AU5 upper lid lowerer, AU7 lid tightener, AU15 lip corner depressor, and AU16 lower lip depressor. Even though the participants are trying to remain calm, disgust leaks out probably due to perplexity (Ekman, 2007) during the task. This finding affirms Bartlett et al.'s (2009) research indicating that, when a person is experiencing emotions, the physiology takes over so that, even when people try to mask their true feelings, they still leak out.

## 4.2 Nonverbal Private Talk

Research indicates that people use nonverbal private talk for self-regulation that helps them to plan, monitor and guide a set of activity in demanding situations (Montazeri et al., 2015), such as during assembly tasks in unidirectional interaction contexts. Nonverbal private talk includes head nods as featured in Fig. 4. Sideways head nods occur in three Left-Right-Left movements using the following muscles: (1) AU51 – head turn left; (2) AU52 – head turn right; and (3) AU51 – head turn left again. Sideways head nods as a composite communicative action are used by Nigerians to indicate negation, disagreement with a co-interlocutor's view, and self-recognition of one's errors or inability to execute an action during interaction. The participant in Fig. 4 uses sideways head nods with a frown to indicate self-recognition of his incomprehension of the instructions and consequent errors during the task.

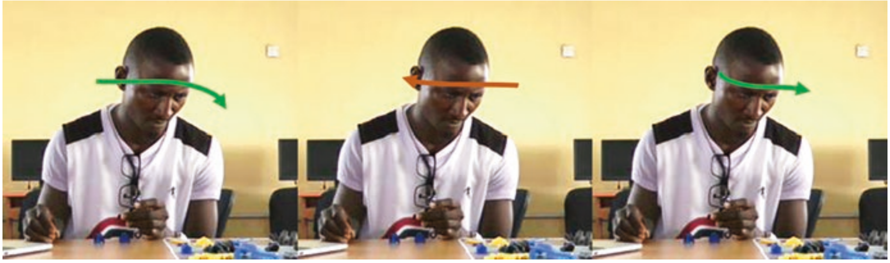
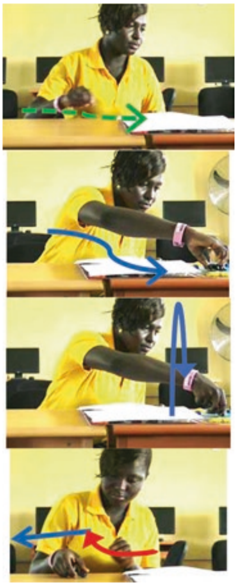


**Fig. 4** L-R-L sideways head nods as self-recognition of self-errors



**Preparation:** P2 initiates the gesture with the **LH** in a hold, palm down with cupped digits. The **RH** is palm lateral and in a hold, too.
**Stroke 1:** Here, N9's **RH** zooms across the task space leftwards, palm down then grasps the first piece with coupler-shaped digits and pulls it up. N9's **LH** goes beneath the table.
**Stroke 2:** N9's **RH** spins ulnarly to grasp and pull up the second piece with coupler-shaped digits pulling it up.
**Retraction:** N9's **RH** goes backwards to drop the selected pieces on the table just as N9's **LH** comes up from beneath the table to a hold.

**Fig. 5** The knowing hand indicating listener comprehension

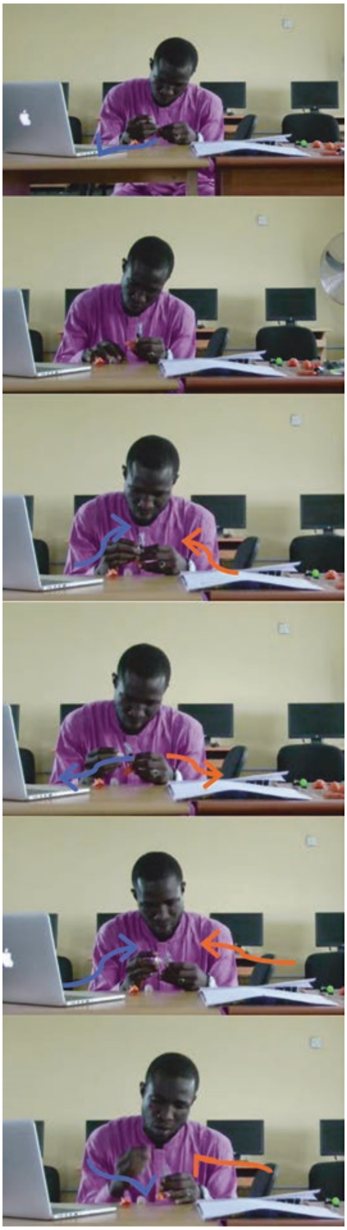
## *4.3 Assembly Gestures*

This section outlines gestures displayed during interaction. However, the knowing hand and the failed joining hand gestures presented are representative of all families indicating listener comprehension and incomprehension of instructions.

***The Knowing Hand Gesture*** This is a beat gesture that is the opposite of the wavering hands gestures (Kirk et al., 2005). The knowing hand indicates that the listener correctly decodes instructions and selects required piece(s) without hesitation. This is an assistive assembly gesture that aids the execution of operative gestures. The form and function of the knowing hand gesture are outlined in the vignette in Fig. 5.

***Joining Hands*** The joining hand is an operative gesture that occurs when participants correctly connect assembly pieces in the right position with the right orientation. In contrast, the *failed joining hands* gesture suggests that the listener did not correctly decode the instruction in either one or all the stages. The gesture presented below is used to assemble parts of the feet by attaching the ball joints of the yellow pieces to the black sockets in order to build the shins of the Lego kit. The form and function of the gesture is shown in the vignette featured in Fig. 6.

Listeners could select wrong assembly kits; this failure is foundational and makes subsequent assembly stages incorrect. In addition, listeners could select the correct assembly kits but wrongly decode the assembly instruction and would thus fix them in the wrong positions with the wrong orientation (see Fig. 6). Participants



**Preparation:** N2 initiates this gesture with both hands holding the kit.

**Preparation** continues as N2’s **RH** goes down palm down with digits held claw-like to pull up the ‘white piece’ from the task space.

**Failed Stroke 1:** N2’s **RH** moves up, leftwards, palm lateral to connect the piece with the kit in the **LH** (receiving hand). Unable to successfully join the assembly bits, N2 aligns them to measure fit – Incomprehension of instruction leads to self-assessment as nonverbal private talk.

**Retraction 1**: As the piece did not fit, N2 retracts the **RH** rightwards while visually assessing the assembly bits (self-initiated self-correction suggesting a possible change in strategy from nonverbal private talk).

**Failed Stroke 2:** N2’s **RH** moves leftwards, palm lateral to connect the piece with the kit in the **LH** (receiving hand). However, N2 only aligns them to measure fit again. **(Testing self-comprehension of assembly instruction again while holding concurrent private talk.)**

**Retraction 2:** As the piece did not fit, N2 retracts the **RH** rightwards again while visually assessing the assembly bits. The **LH** is in a hold.

**Failed Stroke 3:** N2’s **RH** moves leftwards, palm lateral to connect the piece with the kit in the **LH** (receiving hand). The **LH** turns the receiving kit clockwise. However, N2 only aligns them to measure fit again. **(Testing self-comprehension of assembly instruction again.)**

**Retraction 3**: As the piece did not fit, N2 retracts the **RH** upwards in preparation for another picking hand in the next assembly gesture. While the **LH** rests on the table in a post stroke hold as presentation to visually assess the assembly bits.

**Fig. 6** Failed joining hands indicating listener incomprehension

used the picking gesture to select the specified assembly bits in an extended preparation (P-P) and then tried to attach the pieces.

In the process, participants use the aligning hand, which is similar to Kirk et al.'s (2005, p. 11) 'mimicking hands', in an assisted assembly task because they enable the listener order and discover the fit of the assembly pieces before joining them together. This action is similar to taking aim before shooting. Concatenated gestures have inbuilt monologic gestures that externalise listeners' internal cognition processing procedures, such as self-repairs emanating from nonverbal private talk.

## *4.4 Repeat Sequence*

Repetition patterns are potentially present in language and language users employ the various forms of repetition to project their comprehension of speaker input, way of seeing things and coping during interaction (Carter, 2004). The results suggest that repeats occur at different times and for different purposes during the assembly process. Repeats are consistent, reliable and rational communication strategies that promote active listening, set a communication standard and involve a continuous chain of events (Clark et al., 2015; Oxford, 1993). The results shown in Fig. 7 indicate that repeats occur at three different times during the task: before the assembly action, during the assembly action and after the initial instruction or the assembly action has taken place.

Participants ask for repeats before the assembly action or the next instruction occurs in order to get clarification or confirmation. Repeats that co-occur with the assembly action are strategic and are used to demarcate the task self-correct and to confirm assembly process in one instruction. Others occur after the assembly action has occurred or when the current instruction has been provided, and these could be for confirmation, task demarcation, or error correction. However, repeats operate discreetly and are often combined by listeners due to timing and task objectives. Repeats are not fixed as the illustration may suggest because, in real life, they tend to overlap in different combinations.

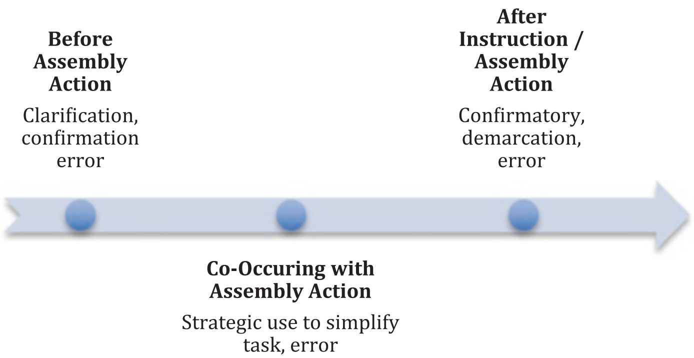

Before Assembly Action
Clarification, confirmation error
After Instruction / Assembly Action
Confirmatory, demarcation, error
Co-Occuring with Assembly Action
Strategic use to simplify task, error

**Fig. 7** The sequence of asking for instruction repeats

# 5 Discussion

## *5.1 Listener Attitudes and Comprehension*

The results suggest that it is possible for humans to show their level and process of comprehending agent instructions through their facial actions, nonverbal private talk and repairs during interaction. Thus, the agent in the TESOL classroom should be adaptive enough to meet the needs of the language learning context.

The results reaffirm the view that attitudes might be reliably detected and measured through facial expressions, just as Meadors and Murray (2014) measured and classified bias through body language. Listener attitudes towards the interaction may also be distinguishable as positive, neutral or negative (Mehrabian & Ferris, 1967), and these may also reflect listener comprehension of speaker input. Positive facial actions include felt smile, slight smile, and controlled laughter, and these represent 18% of the distribution. Neutral includes neutral, neutral concentration, workman effort face, and static searching head, representing 57% of distribution. Negative facial actions represent 25% of the distribution and include puzzled face, compressed or swallowed lips, disgust, slight disgust, slightly compressed lips, micro-frown and nervous smile.

Positive, neutral and negative facial actions constitute the three degrees of attitudes (Mehrabian & Ferris, 1967, p. 249) used in this study to assess listener perceptions about instructions. In addition, Mehrabian and Ferris' (1967) study, citing Gates and Levitt's findings, suggests that focusing on facial actions and gestures may provide feedback comparable to that obtained from interpreting connected strands of multimodal resources or channels of communication as done in Bezemer and Jewitt's (2010) study.

Although the results indicated that there were more neutral facial actions than any other, the listeners were generally more consistent in the display of positive attitudes than neutral ones. This is evident to some extent in Table 3 as positive attitudes have lower standard deviations (1.73, 1.50, 1.26) for CP, CL, and HV instructors respectively compared to neutral (7.05, 3.83, 6.71), in addition to HV when compared to negative (1.36).

When viewed from the prism of instructors along the voice cline (Fig. 2), attitudinal and communicative results (see Tables 3, 4a and 4b) suggest that listeners are more consistent in having a positive attitude towards the human instructor, while experiencing more negative interactions with L1 speaking agents due to their easy comprehension of HV's instructions. Within group results for the positive attitudes suggest that listeners had a better comprehension of HR's instruction than the instructions from the synthesised L1 voices. These L2 speaker attitudes and comprehension levels suggest that, the closer an agent's voice quality is to a human voice, the more the agent is perceived as likeable and easier to understand.

These confirm earlier studies indicating that, the further away a speaker's voice quality and language use are from a hearer's norm, the more a speaker is perceived as less attractive and more incomprehensible (Babel et al., 2014). By implication, a

**Table 3** Inferences from listener facial actions: Degree of listener comprehension of verbal instructions

| | Inferred comprehension scores corresponding to instructor | | | | | |
|---|---|---|---|---|---|---|
| Inferred attitude | Positive | | Neutral | | Negative | |
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Instructor | | | | | | |
| CP | 2.00 | 1.73 | 5.80 | 7.05 | 1.00 | 0.38 |
| CL | 2.75 | 1.50 | 3.80 | 3.83 | 1.70 | 0.86 |
| HR | 2.33 | 1.41 | 7.50 | 0.89 | 1.40 | 0.89 |
| HV | 0.63 | 1.26 | 2.00 | 6.71 | 0.88 | 1.36 |

software agent with verbal and linguistic characteristics that are adaptable to L2 contexts has greater potential for being more successful at teaching English language in L2 contexts such as Nigeria than one that is not. Adaptability is required in the areas of verbal qualities and language usage, making it closer to, for example, Educated Nigerian English.

## 5.2 *Pragmatics of Interaction in Hybrid ELT Classrooms*

As outlined above, language use is a form of joint action between instructors and instructees where each has responsibilities and communicative expectations. The results show that listeners use a combination of communicative gestures (Table 4c), such as self-initiated self-repairs, to focus on meaning-making, reposition task execution, identify errors and complete tasks using experiences gained from route knowledge. Route knowledge is used here to describe the spontaneous know-how a participant develops from carrying out repetitive assembly tasks.

Route knowledge parallels Tannen and Wallat's (1987) knowledge schemas to the extent that they both refer to the interlocutor's use of prior knowledge in current interaction contexts. Specifically, listeners mediate meaning using facts they bring as well as implicit knowledge developed unintentionally to interaction (Vandergrift, 2011). However, unlike repairs in HHI, repairs in HAI are more elaborate and reinforce the notion that there is a potential for all talk to be embedded in a power relationship (Hutchby, 2001).

Furthermore, the interaction is tied to L2 listeners' meaning making; thus, repetition is used to set interaction and learning expectations at different stages of task execution while developing reflective learning practices through self-regulation. These affirm earlier findings that self-initiated self-repairs provide a self-correcting mechanism for the organisation of language during interaction (Clark, 2012). In addition, when people interact with agents, they not only rely on cognitive processes, but also on other strategies in making sense of interaction (Murdoch et al., 2013.)

**Table 4a** Summary of spontaneous listener facial actions implying emotions and their communicative functions

| Facial action family | Illustration | | | |
|---|---|---|---|---|
| | Child | Sibling buckets | Action units (AU) | Communicative function |
| Basic facial actions | Neutral | Neutral | AU0 | May indicate indifference, relaxed composure, cognitive processing of on-going instruction while multitasking and initiation of turn-taking. |
| | | Neutral face down | AU0+AU54 | Cognitive processing of speaker input. |
| | | Neutral concentration on task | AU0+AU8 | Concentration on task execution, i.e. translating instructions into action. |
| | | Neutral hard | AU0+AU44 | Experiencing difficulty in cognitive or processing. |
| | Smile | Felt smile | AU6+AU12+AU7 | May indicate positive interaction with instructor, task and self-resulting from correct comprehension of speaker input. |
| | | Tight-lipped | AU6+AU6B+AU12B | Positive mask for real feelings of incomprehension. |
| | Disgust | Disgust | AU9+AU15B+AU16B+AU44B (strand of AU4) | Indicates feeling of aversion towards self-efforts or interaction experience and incomprehension of speaker input. |
| Eye action | Eye closure | Eye closure | AU7+AU9 +AU12+AU43B +AU56B (slight tilt left) | May indicate listener's degree of self-belief or certainty and ongoing cognitive processes regarding their comprehension of instructions or appropriacy of task execution. |
| Micro-expressions | Basic hot spots | Anger | AUB4B +AU5+AU7B | Emotional leakages indicating true emotions depending on the context. |
| | | Disgust | AU9B+AU15B+AU16 | See C2 above with lesser intensity. |

**Table 4b** Summary of spontaneous listener facial actions implying emotions and their communicative functions

| Facial action family | Illustration | | | |
|---|---|---|---|---|
| | Child | Sibling buckets | Action units (AU) | Communicative function |
| Non-basic blends | Whuck! | Whuck! Moment | AU13+AU24+AU34+ AU55 (head tilt left) | Indicates a mix of surprise, astonishment, incredulity, shock (eye-action), confusion and difficulty in cognitive processing; disruption of cognitive processes as listener concentrates on instructor input (puffy pout) negative experiences. |
| Emotional attitudes/ Moods | Frown | Frown | AU44+AU46+AU24+ AU21+AU41+AU17 +AU15 | May indicate concentration while processing instructions or assessing action taken but may also suggest difficulty with comprehension. |
| | Workman face | Work face | AU9+AU15+AU16 | Indicates listener's exertion of force when under pressure during tasks. Has no impact on comprehension. |
| | Tense mouth and lip action | Compressed lips | AU8 | May indicate concentration, challenging cognitive processing, anxiety, nervousness, mood shift. |
| | Emotional build-up | Frustration process | | May indicate listener burden as their ability to deal with specific challenges diminishes due to increasing incomprehension during interaction. |
| | Nonverbal private talk | L-R-L Sideways Head nods | AU51+AU52+AU51 | A sign of negation in all Nigerian cultures used here to suggest listener self-recognition of their incomprehension of instructions during interaction. |
| | | Pouty face | AU17+AU25, AU21+AU22+AU23 | May be used for concentration or self-comfort. |

Following these findings, agents destined for L2 contexts should possess adaptable relational capabilities that would enable them to build and maintain long-term socio-emotional relationships with L2 users, as well as remember past communications and manage future expectations in their interactions. In addition, when integrating agents into TESOL classrooms in L2 contexts, consideration should be

**Table 4c** Summary of spontaneous listener assembly gestures and their communicative functions

| Gesture family | Sub-group | Description | Function |
|---|---|---|---|
| Aligning hand | Basic aligning hand | Assembly bits are brought close, but they do not touch. | Enables listener to visually and mentally assess the fit of one or more assembly parts into others. |
| Picking hand | Knowing hand | A beat gesture executed without hesitation. | Enables listeners to select appropriate assembly piece and indicates correct comprehension of instruction. |
| | Searching hand | Executed with hands wavering. It is the opposite of the knowing hand. | Indicates that listeners are unable to select assembly kit due to incomprehension of instruction. |
| Joining hand | Concatenated joining hand | Executed with both hands placing, pushing, or sliding one piece into another with alignment built in between. Has inbuilt monologic communicative gestures. | Enables listener to fix assembly pieces together within a 3D location with the appropriate orientation. In built sub-gestures enable listeners to externalise their comprehension of speaker input. |
| Monologic communicative gestures | Self-initiated self-repairs | Proceeds in three stages: trouble source identification contains two full gesture phrases before repair initiation and actual repairs take place. | Used to correct picking and joining errors as well as testing of self-comprehension of assembly instructions. |
| | Presentation | Enacted as the participant places assembly pieces before their eyes for examination. | Used to assess correct interpretation of instructions by examining assembly piece selected or attachment done. |

given to the ability of teachers and agents to collaboratively recognise subtle learner initiatives and expressions of comprehension or incomprehension of speaker input as this will enable them to devise appropriate L2 language learning support measures.

Although, this is a unidirectional instruction-giving context, the Gricean notion that speaker-information must be clear and adequately informative (Grice, 2006) remains relevant because, when this notion is flouted, listener communication expectations may not be met. To meet L2 communication expectations, enhance cognitive activities and use emotions positively, listeners should be encouraged to use strategic repeat sequences to manage interaction and reduce task difficulty in hybrid TESOL classrooms.

## 6 Conclusion and Further Research

This chapter presented a scoping study in which L2 English speakers took assembly instructions from L1 English speaking agents. The findings suggest that adaptable affect-aware relational agents (sentiment-aware, emotion-aware, courteousness-aware) with knowledgeability and multimodal understanding to develop an engaging response generation system will be most suitable for hybrid TESOL contexts, rather than those available, as no one size fits all. Such affect-aware relational agents are desirable to understand L2 listener sentiments and emotions while generating responses. This will make agents more user-friendly. L2 listeners showed that they understood human and human-like voices better than synthesised ones. However, they were able to devise self-regulatory strategies that enhanced their meaning-making in challenging comprehension contexts.

People interact and perceive their environment multimodally in real life and not as separate layers. For example, a listener in HHI sees and hears the speaker's utterances, gestures, posture, distance, and facial actions at the same time and uses these aspects of the interaction context to make meaning of utterances and interaction. This study analysed nonverbal listenership in two layers of facial actions and gestures following established procedures of applied linguistic research that focus on separate but combinable semiotic resources, such as gestures and facial actions in nonverbal listenership as espoused by Ekman (2016), McNeil (2005), and Kendon (2004).

Applying such research to TESOL aims to meaningfully reconcile these fragments into a coherent discourse at the level of analysis in order to devise multimodal corpus linguistics coding matrixes useful for annotating various co-occurring nonverbal listenership behaviours. These linguistic code matrixes may potentially be used to derive laws that drive agents' high emotive functionality for enhanced HAI in English language learning contexts. However, this process requires more understanding in order to enable researchers to knowledgeably perceive how smart agents integrate co-interlocutor's facial actions, gestures, voice, utterances and posture to arrive at a multimodal interpretation of information exchanged during interaction.

## References

Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. John Benjamins.

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS ONE, 9*(2), e88616. https://doi.org/10.1371/journal.pone.0088616

Barrett, J. (2012). John Barrett: The internet of things. *TEDx Talks*. Retrieved from www.youtube.com/watch?v=QaTIt1C5R-M

Bartlett, M., Littlewort, G., Vural, E., Whitehill, J., Wu, T., Lee, K., & Movellan, J. (2009). Insights on spontaneous facial expressions from automatic expression measurement. In M. C. Giese & H. Bulthoff (Eds.), *Dynamic faces: Insights from experiments and computation* (pp. 211–238). MIT Press.

Bezemer, J., & Jewitt, C. (2010). Multimodal analysis: Key issues. In L. Litosseliti (Ed.), *Research methods in linguistics* (pp. 180–197). Continuum.

Bhabha, H. K. (1994). *The location of culture*. Routledge.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Carter, R. (2004). *Language and creativity: The art of common talk*. Routledge.

Chun, D., Ker, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *The Modern Language Journal, 100*(S1), 64–80.

Clark, L. (2012). *On the lingua franca core and L1 influenced pronunciation variations: A repair-based study on the speech of ELF users*. Unpublished bachelor thesis, York St. John University, York, UK.

Clark, L., Bachour, K., Ofemile, A. Y., Adolphs, S., & Rodden, T. (2014). Potential of imprecision: Exploring vague language in agent instructors. In *Proceedings of the second international conference on human-agent Interaction (HAI'14)* (pp. 339–344). ACM. https://doi.org/10.1145/2658861.2658895

Clark, L., Ofemile, A. Y., Adolphs, S., & Rodden, T. (2015). Language and identity in human computer interaction: A multimodal corpus approach. Paper presented at *British Association of Applied Linguistics (BAAL) Conference*, Aston University, Birmingham, UK.

Clark, L., Ofemile, A. Y., Adolphs, S., & Rodden, T. (2016). A multimodal approach to assessing user experiences with agent helpers. *ACM Transactions on Interactive Intelligence Systems, 6*(4), 1–31. https://doi.org/10.1145/2983926

Coulthard, M., Johnson, A., & Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). Routledge.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Paper presented at *IEEE computer society conference on computer vision and pattern recognition (CVPR)*, San Diego, USA.

Dautenhahn, K. (2013). Human-robot interaction. In M. Soegaard & R. F. Dam (Eds.), *The encyclopaedia of human-computer interaction* (2nd ed., Chapter 38). The Interaction Design Foundation. Available at: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Ekman, P. (2007). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan.

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science, 11*(1), 31–34.

Ekman, P., & Friesen, W. V. (1978a). *Facial action coding system (FACS): A technique for the measurement of facial action*. Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1978b). *Facial action coding system (FACS): Part two*. Consulting Psychologist Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier Science.

Feng, D., & O'Halloran, K. L. (2013). The multimodal representation of emotion in film: Integrating cognitive and semiotic approaches. *Semiotica, 197*, 79–100. https://doi.org/10.1515/sem-2013-0082

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down. *System, 32*(3), 363–377.

Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.

Fortin, S., Long, W., & Lord, M. (2010). Three voices: Researching how somatic education informs contemporary dance technique classes. *Research in Dance Education, 3*(2), 155–179. https://doi.org/10.1080/1464789022000034712

Frenečik, M. (2005). Organization of repair in talk-in-interaction and politeness. In *Theory and practice in English studies 3. Proceedings from the eighth conference of British, American and Canadian studies* (pp. 1–10). Masarykova Univerzita.

Geertzen, J. (2012). *Inter-rater agreement with multiple raters and variables*. Retrieved from https://nlp-ml.io/jg/software/ira/

Givens, D. B. (2015). Measuring gestures. In D. Chadee & A. Kostić (Eds.), *Social psychology in nonverbal communication* (pp. 66–91). Palgrave-Macmillan.

Grice, H. P. (2006). Logic and conversation. In A. Jaworski & N. Coupland (Eds.), *The discourse reader* (2nd ed., pp. 66–77). Routledge.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measure the extent of agreement among raters* (3rd ed.). Advanced Analytics.

Harmer, J. (2007). *The practice of English language teaching* (4th ed.). Pearson Longman.

Hendrikse, A. P., Nomdebevana, N., & Allwood, J. (2016). An exploration of the nature, functions and subcategories of the discourse functional category, interactive, in spoken Xhosa. *South African Journal of African Languages, 36*(1), 93–101.

Hess, U., & Kleck, R. E. (1997). Differentiating emotion elicited and deliberate emotional facial expression. In P. Ekman & E. Rosenberg (Eds.), *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (pp. 271–286). Oxford University Press.

Huang, Y. T., & Arnold, A. R. (2016). Word learning in linguistic context: Processing and memory effects. *Cognition, 156*, 71–87.

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology, 35*(2), 441–456.

Jennings, N. R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., & Rogers, A. (2014). Human-agent collectives. *Communications of the ACM, 57*(12), 80–88.

Kendon, A. (1980). Gesture and speech: Two aspects of the process utterance. In M. R. Key (Ed.), *Nonverbal communication and language* (pp. 207–227). Mouton.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kirk, D., Crabtree, A., & Rodden, T. (2005). Ways of the hands. In H. Gellersen, K. Schmidt, M. Beaudouin-Lafon, & W. Mackay (Eds.), *ECSC 2005: The ninth european conference on computer-supported cooperative work* (pp. 1–21). Springer.

Kita, S. (2013). *Architectural issue in the model of speech-gesture production: Gesture, action language*. Retrieved from https://tiger.uvt.nl/pdf/presentations/tiger2013_architecture_v1.pdf

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures and their transcriptions by human coders. In I. Wachsmuth & M. Erhlich (Eds.), *Gesture and sign language in human-computer interaction* (Lecture notes in computer science) (Vol. 1371, pp. 23–35). Springer.

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review, 124*(3), 245–266.

Knight, D. (2009). *A multi-modal corpus approach to the analysis of backchanneling behaviour*. Unpublished doctoral dissertation, University of Nottingham, Nottingham, UK.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Lacey, K. (2013). *Listening publics: The politics and experience of listening in the media age*. Wiley.

Levinson, S. C. (2016). Turn-taking in human communication – Origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2010). *Practical resources for assessing and reporting intercoder reliability in content analysis research projects*. Retrieved from http://matthewlombard.com/reliability/#what_is_intercoder_reliability

Matsumoto, D., & Hwang, H. C. (2013). Cultural similarities and differences in emblematic gestures. *Journal of Nonverbal Behaviour, 37*(1), 1–27.

Mazzone, M. (2015). Constructing the context through goals and schemata: Top-down processes in comprehension and beyond. *Frontiers in Psychology, 6*, 1–13.

McNeil, D. (1985). So, you think gestures are nonverbal? *Psychological Review, 92*(3), 350–371.

McNeil, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeil, D. (2005). *Gesture and thought*. University of Chicago Press.

Meadors, J. D., & Murray, C. B. (2014). Measuring nonverbal bias through body language responses to stereotypes. *Journal of Nonverbal Behaviour, 38*(2), 209–229.

Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31*(3), 248–252.

Mixdorff, H., Hönemann, A., Rilliard, A., Lee, T., & Ma, M. K. (2017). Audio-visual expressions of attitude: How many different attitudes can perceivers decode? *Speech Communication, 95*, 114–126.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language, 66*(1), 249–264.

Montazeri, M., Hamidi, H., & Hamidi, B. (2015). A closer look at different aspects of private speech in SLA. *Theory and Practice in Language Studies, 5*(3), 478–484.

Murdoch, J., Salter, C., Cross, J., & Poland, F. (2013). Misunderstandings, communicative expectations and resources in illness narratives: Insights from beyond interview transcripts. *Communication & Medicine, 10*(2), 153–163.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of CHI conference* (pp. 72–78). ACM.

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology, 29*(5), 1093–1110.

Nunan, D. (2002). Listening in language learning. In J. C. Richards & W. A. Renandya (Eds.), *Methodology in language teaching* (pp. 238–241). Cambridge University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Ofemile, A. Y. (2015). A framework for analysing discourse between humans and software agents. *MAUTECH Journal of Language and Communication (JLC), 1*(1). Retrieved from https://www.researchgate.net/publication/291945249_A_FRAMEWORK_FOR_ANALYSING_DISCOURSE_BETWEEN_HUMANS_AND_SOFTWARE_AGENTS

Ofemile, A. Y. (2018). *Listenership in human-agent collectives: A study of unidirectional instruction-giving*. Unpublished doctoral dissertation, University of Nottingham, Nottingham, UK.

Oxford, R. (1993). Research update on teaching L2 listening. *System, 21*(2), 243–250.

Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. In *Proceedings of IEEE sixth international conference on computer vision* (pp. 555–562). IEEE: Bombay. https://doi.org/10.1109/ICCV.1998.710772

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications/Cambridge University Press.

Reeves, B., & Nass, C. (2014, January 29). *The media equation*. (E. Griffin, Interviewer). Available at: https://www.youtube.com/watch?v=26BclMJQUwo

Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review, 27*(4), 587–597.

Sekine, K., & Kita, S. (2015). The parallel development of the form and meaning of two-handed gestures and linguistic information packaging within a clause in narrative. *Open Linguistics, 1*(1), 490–502.

Simons, G. F., & Fennig, C. D. (2018). *Ethnologue: Languages of the world*. SIL International. Online version: http://www. ethnologue.com

Simpson, J. (2017). Translanguaging in the contact zone: Language use in superdiverse urban areas. In *Multilingualisms and development: Selected proceedings of the 11th language and development conference* (pp. 207–223). British Council India.

Tannen, D., & Wallat, C. (1987). Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview. *Social Psychology Quarterly Special Issue: Language and Social Interaction, 50*(2), 205–216.

Vandergrift, L. (2011). L2 listening: Presage, process, product and pedagogy. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning (Vol. II)* (pp. 455–471). Routledge.

White, Z. S. (1989). Backchannels across cultures: A study of Americans and Japanese. *Language in Society, 18*(1), 59–76.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, fifth international conference on language resources and evaluation* (pp. 1556–1559) Retrieved from http://www.lrecconf.org/proceedings/lrec2006/pdf/153_pdf.pdf

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *MBC Medical Research Methodology, 13*(1), 61–67.

Xu, J., Gannon, P. J., Emmorey, K., Smith, J. F., & Braun, A. R. (2009). Symbolic gestures and spoken language processed by a common neural system. *Proceedings of the National Academy of Sciences, 106*(49), 20664–20669. https://doi.org/10.1073/pnas.0909197106

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting, Chicago linguistic society* (pp. 567–578). Chicago Linguistic Society.

Zimmerman, D. H. (1998). Identity, context and interaction. In C. Antaki & S. Widdicombe (Eds.), *Identities in talk* (pp. 87–106). Sage.

Zwitserlood, I., Özyürek, A., & Perniss, P. (2008). Annotation of sign and gesture cross-linguistically. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and exploitation of sign language corpora. 3rd workshop on the representation and processing of sign languages* (pp. 185–190). ELDA.