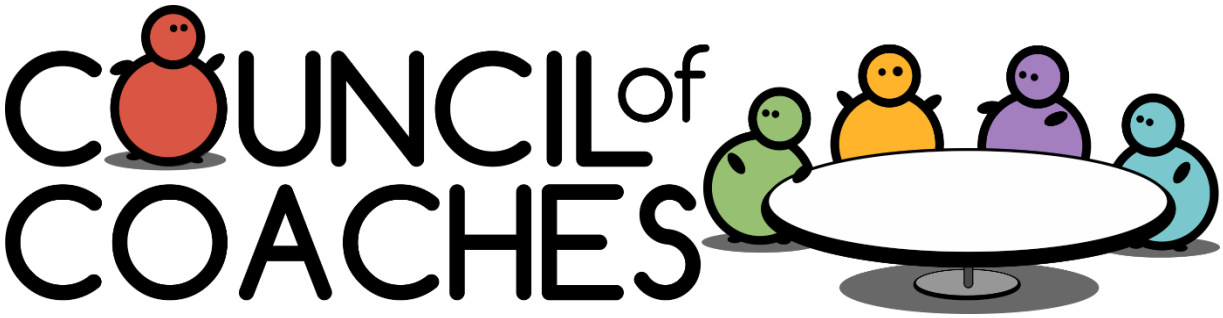


D6.5: Final prototype description and evaluations of the virtual coaches platform

Dissemination level: Public

Document type: Report

Version: 1.0.0

Date: January 22, 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Document Details

Project Number	769553
Project title	Council of Coaches
Title of deliverable	Final prototype description and evaluations of the virtual coaches platform
Due date of deliverable	November 30, 2019
Work package	WP6
Author(s)	Gerwin Huizing (HMI), Brice Donval (SU), Mukesh Barange (SU), Reshmashree Kantharaju (SU), Fajrian Yunus (SU)
Reviewer(s)	Catherine Pelachaud (SU), Harm op den Akker (RRD)
Approved by	Coordinator
Dissemination level	Public
Document type	Report
Total number of pages	44

Partners

- University of Twente – Centre for Monitoring and Coaching (CMC)
- Roessingh Research and Development (RRD)
- Danish Board of Technology Foundation (DBT)
- Sorbonne University (SU)
- University of Dundee (UDun)
- Universitat Politècnica de València, Grupo SABIEN (UPV)
- Innovation Sprint (iSPRINT)

Abstract

The goal of this Work Package (WP6) is to design, implement, and evaluate the Human Computer Interaction aspects of the Council of Coaches. In this deliverable D6.5, we present the work done so far on the development of the final technical prototype of the Council of Coaches system. We briefly summarize the architecture of the system and how different modules relate to enabling virtual agents to interact with each other and with the user. We present the preliminary study to analyse the effect of non-verbal cues and interruptions on group cohesion during the multiparty interaction. Moreover, we report on the two evaluation studies already performed in the context of the Council of Coaches system and their results.

Table of Contents

List of figures

List of tables

Symbols, abbreviations and acronyms

AMI	Augmented Multiparty Interaction
ASAP	Articulated Social Agents Platform
AU	Action Unit
BML	Behaviour Markup Language
CMC	Centre for Monitoring and Coaching
COUCH	Council of Coaches
D	Deliverable
DBT	Danish Board of Technology Foundation
EC	European Commission
ECA	Embodied Conversational Agent
ECC	Embodied Conversational Coach
GUI	Graphical User Interface
HCI	Human-Computer Interaction
ISPRINT	Innovation Sprint
M	Mean
M	Month
MS	Milestone
RRD	Roessingh Research and Development
SAIBA	Situation, Agent, Intention, Behaviour, Animation
SD	Standard Deviation
SU	Sorbonne University
UDun	University of Dundee
UPV	Universitat Politècnica de València
UT	University of Twente
WP	Work Package

1 Introduction

The Council of Coaches project aims to develop a tool to provide virtual coaching for ageing people to improve their physical, cognitive, mental and social health. The council consists of a number of Embodied Conversational Coaches (ECCs), each specialised in their own specific domain. They interact with each other and with the user to inform and motivate them, and discuss issues related to their health and well-being.

This deliverable D6.5 describes the final prototype developed for the Council of Coaches project (described in more detail in Section 3). It includes the architecture of the system and the brief description of functionalities of different components of the architecture, as they relate to enabling the interaction with our agents.

In the context of the Council of Coaches project, the prototype includes a multiparty interaction scenario where both the user and virtual coaches can interact with each other and among themselves. As explained in the previous deliverable D6.4, we have been focusing on understanding a higher group level phenomenon i.e., group cohesion in order to develop a computational model that simulates cohesive behaviour. We present the preliminary analysis of the effects of the non-verbal cues (such as gaze, head nod) and interruption phenomena on group cohesion during multiparty interaction (see Section 4).

In this project, the Greta and ASAP platforms are used for multimodal behaviour generation and for visualising Embodied Conversational Agents (ECA) into the Unity3D engine. In this deliverable, we describe the integration of the Greta agent(s) in the Unity platform, which includes the communication and synchronization between Greta and Unity platforms (see Section 5). Other improvements to the Council of Coaches systems are detailed in Section 6.

We present two evaluation studies in the context of the Council of Coaches system (see Section 7). The first study focuses on the evaluation for the final Council of Coaches *Technical Prototype* that focuses on the proof of concept. The second study aims to evaluate the effect of inter-coach discussion during a persuasive dialogue in a coaching session the project's *Functional Demonstrator*.

The software documentation to define new agents and behaviours using the Council of Coaches system is presented in Section 8, while Section 9 concludes the report.

2 Objectives

The main objective of this deliverable (D6.5) is to describe the final technical prototype of the virtual coach dialogue platform where virtual coaches are interacting with each other and the user and the evaluation of this model. This deliverable also includes a first multimodal analysis of group cohesion phenomenon and a software documentation that explains how we can customize the multi-agent platform to create different scenarios with different coaches.

3 Final Prototype

In this section, we describe the final prototype developed for the Council of Coaches project. We first briefly present the overall architecture of our system to give an overview of the components enabling our agents to interact. We then describe the improvements done to the different individual components in the part of the system that WP6 is responsible for, which are Flipper, Greta, and ASAP (part of the "THINK" and "ACT" sections of the architecture as depicted in Figure 1 below).

Below, in Figure 1 we present the current version of the architecture of the full Council of Coaches Technical Demonstrator.

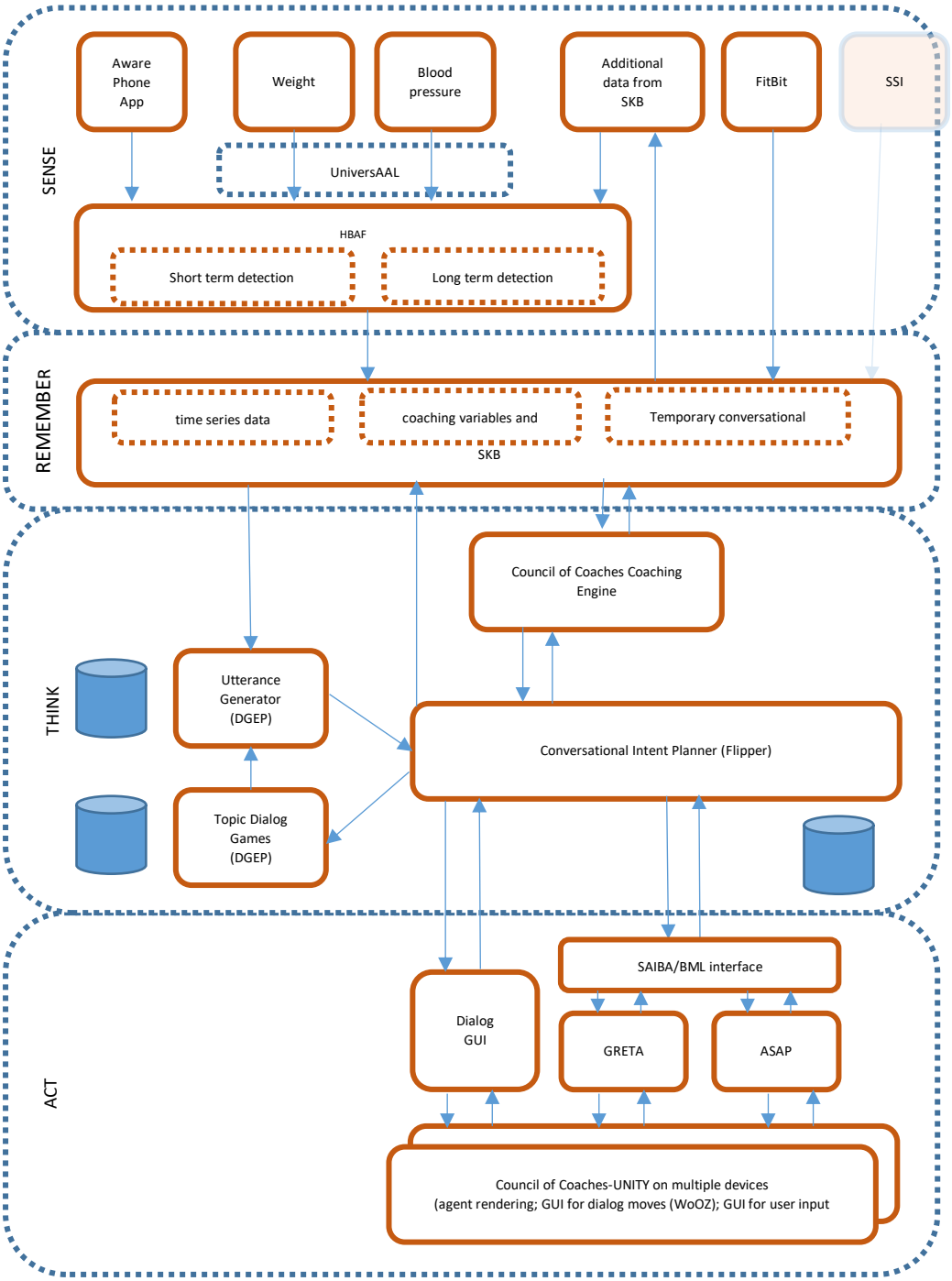


Figure 1: Current vision of the architecture.

In the following sections we will briefly outline each part of the system, the function it has, and the modules it contains.

3.1 SENSE

This part of the system is responsible for sensing things about the user through several devices, and captures and transforms this data so that it can be used by the rest of the system to provide meaningful coaching. It consists of the following modules:

- The aware phone app: An application that gathers data about the user through their smartphone, such as their movement.
- UniversAAL compliant devices, such as a weight scale and blood pressure measuring device.
- Additional data taken from the shared knowledge base (SKB), such as variables stored from parts of conversations had with the user, and the data from Fitbit that is in the SKB.
- Fitbit: A wristband that senses information about the user, such as heart rate and quality of sleep.
- SSI: The social signal integration system, which takes information from camera and sound input, and uses this to detect social signals from the user, such as whether they are interested, and whether they display positive or negative emotional valence.
- HBAF: The holistic behaviour analysis framework uses the various sources of sensor data in order to model long term and short term physical, social and affective behaviours, using time series analysis of sensor data.

3.2 REMEMBER

This part of the system is responsible for the storing and making available of all information that the HBAF provides, as well as information about the conversation the system is having, and other important variables coming from the Council of Coaches Coaching Engine. It consists of the following module:

- Shared knowledge base (SKB): contains the “memory” of the system. This is short term memory (at the level of the ongoing conversation), as well as longer term memory of the sensor data, personal characteristics, coaching goals, and interaction history for each user.

3.3 THINK

This part of the system is responsible for reasoning about the current state of the user (e.g. their goals, progress, emotional state) and the conversation (e.g. what topic we are discussing, what the goal of the current conversation is, whether the core message has been delivered in a previous utterance, whether we received new information from the user to be used later). It uses information from the SKB and the progress report regarding the current conversation from the ACT part of the system to do so. It consists of the following modules:

- Council of Coaches Coaching Engine: This engine decides on the topic to discuss and strategies (e.g. content, preferred style) to use by the coaches. The options it has are weighted based on things like the history of interactions, the current sensor data, and the knowledge about the user profile (all available from the SKB).
- Topic Dialog Games (DGEP): The modelling framework of DGEP for argumentation based on dialog structures has been used to form models of abstract structures for coaching dialogues (e.g. goal setting, informing user about health topic).
- Utterance Generator (DGEP): A new part of DGEP developed for the Council of Coaches project that fills abstract moves with concrete content and utterances for specific steps in the dialogue (Section 6.2.2: Innovation in Automated Personalized Multi-Party Dialogue; Deliverable D7.5: “Final Council of Coaches Technical Prototype”). It consists of utterance templates marked up with content and delivery style annotations. The Utterance Generator picks one of these templates based on the abstract “possible next move type” from the Topic Dialog Game, and the variables that the Council of Coaches Coaching Engine originally provided regarding the content and delivery style. It fills slots in the selected template with specific data from the SKB.

- Conversational Intent Planner (Flipper): This is an information state-based conversation engine (van Waterschoot, et al., 2018). It turns the provided abstract conversational moves and information provided by the other modules in this part of the system into a series of multimodal utterances for the virtual embodied agents. It also regulates turn taking and backchanneling, and keeps track of whether moves have been delivered or not (e.g. interruption took place). This module is constantly communicating with the Topic Dialog Games module to update which moves have been made and what the next moves are.

3.4 ACT

This part of the system performs the behaviours that the THINK part of the system has decided on, and gives feedback to the THINK part of the system regarding what behaviours were performed by all participants, including the user. It delivers its content to multiple devices using the Unity engine by employing two embodied virtual agent platforms; GRETA and ASAP. It consists of the following modules:

- SAIBA/BML interface: This interface translates the series of multimodal behaviours sent by the Conversational Intent Planner (Flipper) into a format that GRETA and ASAP can use. It also returns messages to the Conversational Intent Planner (Flipper) regarding the progress of these behaviours (e.g. were they performed, and which behaviour is currently being planned), which it receives from its connection with GRETA and ASAP. An overview of the SAIBA framework can be found in section 3: SAIBA Framework, of the deliverable D6.1: "Requirements and Concepts for Interaction Mobile and Web".
- GRETA: An embodied virtual agent platform developed to have advanced autonomous nonverbal behaviour accompanying the utterances by agents (Section 6.2.3: Innovation in Multi-Party Embodied Conversational Agent Systems; Deliverable D7.5: "Final Council of Coaches Technical Prototype").
- ASAP: A behaviour realiser for choreographing and realising multimodal behaviours of multiple agents (e.g. robots, chatbots, embodied virtual agents). It was developed specifically for adaptive timing and mutual coordination of behaviour between multiple agents and the user (Section 3.2.1: ASAP; Deliverable D6.3 "First Prototype description and evaluations of the virtual coach platform".).
- Dialog GUI: The graphical user interface (GUI) that allows the user to interact with the Council of Coaches system. It presents the user with several buttons with utterances on them. The user can pick the one best matching how they would like to respond. The system then sends this on to the Conversational Intent Planner (Flipper) to notify it of the dialog move made by the user.
- Council of Coaches-Unity on multiple devices: GRETA, ASAP and the Dialog GUI can all be displayed on multiple devices using the Unity engine. Some examples include the computer, tablet, and smartphone.

4 Group Model

In this section we describe the work we have conducted in order to develop our group interaction model for handling turn-taking between multiple virtual coaches and a human user and to generate appropriate behaviours. As explained in the previous deliverable, D6.4, we have been focusing on understanding a higher group level phenomenon i.e., group cohesion. Cohesion describes the tendency of group members' shared bond/attraction that drives the members to stay together and to want to work together (Casey-Campbell & Martens, 2009). The goal of our work is to develop a model that will be able to simulate cohesive behaviours for the coaching agents. The first step in order to develop a cohesive model is to recognize the behaviours associated to group cohesion. To achieve this, we first conducted a preliminary analysis on an existing database that provides annotations for cohesion, dialog acts and non-verbal behaviours. Following section provides a detailed description of the preliminary analysis.

4.1 Non-verbal cues

Non-verbal behavioural cues like gaze, facial expressions, gestures, and body postures etc., indicate the attitude of a given individual in any social situation (Richmond, McCroskey, & Payne, 1991) and convey information about affect, mental state, personality, and other traits (Vinciarelli, Pantic, & Bourlard, 2009). While there are several works in literature that provide a detailed analysis features e.g., prosody, visual energy that measure cohesion, they do not look at the social signal cues e.g., gaze, head movement. Therefore, for this preliminary study, we focus on gaze behaviour, head nods, facial action units (eyebrows and smile) and laughter. Since cohesion is associated with bonding, feedback and support, we hypothesize that behaviours corresponding to these i.e., mutual gaze, head nods and smile are frequent in highly cohesive segments. We also look at the presence of action unit AU4 i.e., brow lowerer which is often associated with negative emotions e.g., anger, disgust (Ekman, Davidson, & Friesen, 1990)

4.1.1 Dataset

We relied on an existing database, the Augmented Multiparty Interaction (AMI) corpus (Carletta, et al., 2005). It consists of 100 hours of multimodal recordings of four participants in realistic and scenario-driven meetings. We chose to work on this database for our preliminary analysis as it is annotated on different levels. This is not yet the case of the Council of Coaches database which has been annotated at the cohesion level, but not yet at the nonverbal behaviour level. We are currently working on annotating this level. Once the annotation is finalized, we will reproduce the analysis we have conducted with the AMI corpus. It will allow us to analyse how interaction context matters. Both databases, the AMI corpus and the Council of Coaches corpus, differ in settings, task, topic of interaction, role of the participants.

The AMI corpus has been annotated for speech transcription, dialogue acts, head and hand gestures, focus of attention along with several other properties. A portion of the AMI corpus was annotated for task and social cohesion values by Hung et. al., (Hung & Gatica-Perez, 2010). The meetings were divided into two-minute segments. A total of 100 segments were taken from the 10 meetings where the teams are asked to design a remote control and 20 segments from two groups involved in real discussions. The data was annotated manually by 21 annotators using a 27-item questionnaire on a 7-point Likert scale. Each segment was annotated by three different annotators and a kappa agreement was calculated. In total, 61 segments with a kappa score above 0.3 was retained. This consisted of 50 segments with high cohesion rating and 11 segments with low cohesion rating. Among the 61 segments annotated with cohesion, only 25 are annotated with dialogue act annotations. Specifically, these annotations are available for eight of the eleven low cohesion segments. Therefore, for our work, we consider a total of 16 segments i.e., eight high cohesion ($M = 2.995$, $SD = 0.3276$) and eight low cohesion ($M = 5.994$, $SD = 0.1929$) segments with $W = 0.94$, $p = 0.62$ and $W = 0.92$, $p = 0.45$ respectively.

4.1.2 Non-verbal cue annotation

We manually annotate the focus of attention i.e., gaze behaviour of each individual in the group. The annotation was carried out at the frame level using ELAN annotation tool. We defined four different gaze targets for a given participant: the other three participants and the "other" class e.g., looking at the table, slides. ***Mutual Gaze*** is calculated by computing the overlapping gaze between any two participants at a given point in time i.e., when two participants are mutually gazing at each other. ***Overall Gaze*** duration is calculated as the total amount of time spent by each participant in a group looking at the other participants. Finally, we calculate ***Gaze at Speaker*** which is the duration of time spent by a participant gazing at the speaker.

Further, we manually annotated ***Head nods*** i.e., vertical up-and-down movements of the head rhythmically raised and lowered. We made use of OpenFace (Baltrušaitis, Robinson, & Morency, 2016) to extract facial Action Units automatically. The tool offers two kinds of scores for the facial Action Unit AU: intensity and presence. The former provides the intensity on a continuous value scale from 1 (minimally present) to 5 (present at maximum intensity); a score of 0 indicates absence. The latter indicates the presence or absence. We segment the video data based on activation of a given action unit and calculate the duration and intensity for each segment. We extracted the laughter instances from the transcription files available with the corpus. Table 1 shows the number of instances annotated for all the segments. For each behavioural cue, we calculate the number of instances in each group, the total duration, the mean duration and additionally, mean intensity for Action Units.

Table 1: Total number of instances annotated for 16 low and high cohesion segments.

Annotation	Low Cohesion	High Cohesion
Mutual Gaze	202	258
Outer Brow Raiser (AU2)	28	26
Brow Lowerer (AU4)	77	59
Lip Corner Puller (AU12)	52	113
Head Nods	100	106
Laughter	31	108

4.1.3 Results

In order to verify our hypothesis for this preliminary study, we perform an independent t-test on the data. Initially, we verify the assumption of normality of the data distribution using Shapiro-Wilk test. For the non-normal data, we perform Mann-Whitney test. For each cue, as mentioned in the previous section, we have computed the total number of instances for each group, the total duration, mean duration for all and additionally, mean intensity for Action Units.

Gaze

We did not find any significant difference in the gaze behaviour at the segment level between the low and high cohesive segments with $p < 0.1$. Therefore, we observed the gaze behaviour at participant level. The duration of gaze at any given participant was significantly higher among participants, ($t(64) = -2.67$, $df = 60.75$, $p = .006$), in the high cohesion segments ($M = 76.64$, $SD = 27.83$) than the participants in the low cohesion segments ($M = 59.25$, $SD = 24.09$). Similarly, participant pairs mutually gazed at each other longer in the high cohesion segments than low cohesion segments and this difference was statistically significant, ($U = 857$, $p = .03$, $r = .31$) as shown in Figure 2.

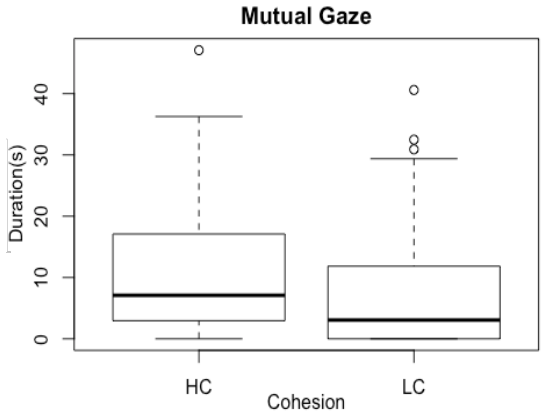


Figure 2: Box plots of mean duration of Mutual gaze for low and high cohesion segments (p = .030).

Facial Action Units

From our data annotations we observe that AU12 i.e., *Lip corner puller* (smile) was activated more frequently in highly cohesive groups. The duration of activation was significantly higher (t (16) = -2.57, df = 10.35, p = .026) in the high cohesive segments (M = 65.05, SD = 42.25) than in the low cohesive segments (M = 21.91, SD = 21.34). Further, as seen in Figure 3, the mean intensity of the activated AU12 was higher as well but the difference was not significant, (t (16) = -2.04, df = 13.77, p = .060). There was no significant difference in the duration or intensity of activation of AU2 i.e., *Outer brow raiser* (raise eyebrow) and AU4 i.e., *Brow lowerer* (frown).

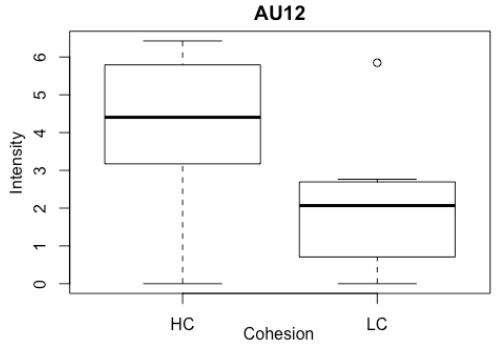


Figure 3: Box plots of mean intensity of Action Unit 12 for low and high cohesion segments (p = .026).

Head Nods

Even though there was not a huge difference in the occurrence of head nods for both the groups, there was a significant difference in the duration of the head nods, (t (16) = -4.33, df = 13.99, p = .0006), see Figure 4. In general, head nods in high cohesion segments lasted longer (M = 7.23, SD = 3.09) than in low cohesion segments (M = 3.38, SD = 3.23).

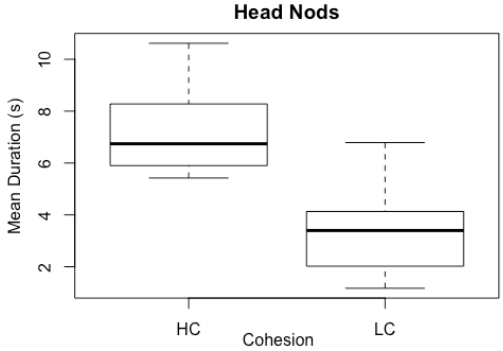


Figure 4: Box plots of mean duration of head nods for low and high cohesion segments (p = .0006).

Laughter

Laughter was observed more frequently in high cohesion segments. The duration of laughter was not significantly different but the average occurrence of laughter per segment was lower (t (16) = -2.59, df = 12.45, p = .022) in low cohesion segments (M = 0.96, SD = 2.22) than in high cohesion segments (M = 3.37, SD = 4.64) as seen in Figure 5.

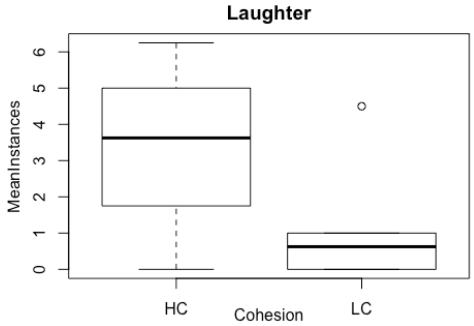


Figure 5: Box plots of mean instances of laughter for low and high cohesion segments (p = .022).

4.1.4 Discussion

As explained in Section 4.1, our aim was to recognize non-verbal social cues that are associated with low and high cohesion groups. In order to do this, we looked at gaze behaviour, facial action units, head nods and laughter. Our initial assumptions were that behavioural cues associated with positive affect, involvement and support e.g., gaze at locutor, laughter, head nods, will be higher in cohesive groups. The main findings of our result are that the instances of laughter are very high in cohesive groups. We observed that instances where more than one participant shared a laughter is higher. This is in line with several studies on laughter in groups which state that *"laughter establishes a form of bond in social groups and makes people feel more comfortable"* (Glenn, 2003). Laughter on an average lasted for 7 seconds in cohesive segments. Additionally, we observe that AU12, that is associated with happiness and smile (Ekman, Davidson, & Friesen, 1990), had a higher intensity value in these segments. Further, we observed that AU4, that is often associated with anger and contempt (Tian, Kanade, & Cohn, 2001), occurred more frequently in low cohesion segments, however, the differences were not significant. This could be attributed to the fact that we observe the interaction for short duration of time (2mn) and perhaps by considering more segments in the dataset this effect could be strengthened. The next assumption we looked at was head nods. The presence of head nods in conversation often creates a favourable environment for conversations (Hadar, Steiner, Grant, & Rose, 1984) and is commonly associated with attentive listening. In our data, there was almost no difference in the frequency of occurrence of head nods between the two groups. However, we did observe a significant difference in the average duration of the head nods. The final cue that we observe is the eye gaze of the participants. Overall, we assumed that in cohesive groups participants spend higher amount of time gazing at others and hold mutual gaze. Our results show that in high cohesive groups participants gazed at fellow members for longer duration than in low cohesive groups. This result supports the studies that state that eye-gazing regulates understanding in multi-party scenarios and is important for managing the flow of interaction (Kendon, 1967). Further, low cohesive groups spent a shorter amount of time holding the gaze with other participants, which is in line with Exline et.al., (Exline, 1963), where they state that the duration of eye-contact decreased in non-collaborative conditions.

4.1.5 Future Work

In the previous section we presented a preliminary analysis we conducted on the AMI dataset to verify our hypothesis about the non-verbal cues associated with group cohesion. Currently, we are in the process of finishing up annotating the database recorded for the Council of Coaches project by University of Dundee. Future work will include, replicating the results from the preliminary analysis using the Dundee dataset and incorporating these behaviours into the group model.

4.2 Interruptions

The effective multiparty dialogue interaction depends on the coordination of team members in conversation using turn taking (Bohus, 2011). In dialogue interaction, turn taking refers to the ability of participants to alternate speaking turns, where one of the participants intends to speak at any given point of time. However, during the multiparty interaction, overlapped utterances may occur where more than one participant may try to speak simultaneously (Heldner & Jens, 2010). These overlapped utterances can be a characteristic of cooperation (Tannen, 1994) as well as conflicts (West & Zimmerman, 2015) in the group. Furthermore, the violation of basic turn-taking rules may result in interruption when one speaker disrupts the turn of another with a new utterance. Based on the contents of the intervention, the interruption can be distinguished into cooperative and disruptive interruption (Li, 2001). Cooperative interruption includes support and agreement, finishing current speaker's phrase, asking for clarification etc. Disruptive interruptions include those showing rejection, topic change, disagreement.

Turn taking and interruption are important phenomena for the effective group interaction. The literature (Tannen, 1994) (Pontecorvo, Pirchio, & Sterponi, 2000) (Li, 2001) (Bangerter, Chevalley, & Derouwaux, 2010) illustrates the effects of turn taking and interruption in interactions and provides an insight into human behaviours during interactions. However, these phenomena have been little studied in the context of group cohesion in multiparty interactions. Thus, the objective of this study is to analyse the relation of turn taking and interruption with group cohesion in multiparty interactions. We hypothesize that occurrences of turns, overlaps and interruptions are higher in highly cohesive groups and lower in low cohesive groups.

4.2.1 Interruptions Annotation

In order to annotate the data with interruption, we define our annotation schema in two layers based on the interruption annotation schema described in (Cafaro, Ravenet, & Pelachaud, 2019)

Transition layer: defines the transition events from silence to speech and vice versa for the same speaker or between multiple speakers.

1. Pause within: a (long) silence within a speaking turn of speaker without speaker switch;
2. Pause between: a speaker switch from current speaker to other participant (or vice-versa) with a silence in between;
3. Perfect: a speaker changes without silence or an overlap in between;
4. Overlap within: an overlap without speaker switch;
5. Overlap between: an overlap with a speaker change.

This layer also makes the distinction between overlap and backchannel using the available dialogue act information along with the start and end time of the speech.

Interruption layer: defines the types of interruption depending upon the interruption time. It includes:

1. Overlapped interruption – interruption having an overlap with speaker change where the speaker does not manage to complete her sentence;
2. Paused interruption – interruption having a speaker switch from current speaker to other participant (or vice-versa) with a silence in between where the speaker does not manage to complete her sentence.

In order to annotate the data, we first perform semi-automatic annotation of communicative and of transition layers based on the start time and end time of each utterance and the dialogue act information. We then manually annotate the interruptions with the help of the multimodal information i.e., speech, verbal transcriptions, and the focus information where the speaker is looking at during the interaction.

4.2.2 Results

Our aim was to analyse the relation between turn taking, interruption and cohesion. We utilize Pearson correlation test to observe the relation between cohesive segments and the independent variables and perform a one-way ANOVA to measure the differences between the two groups.

Turn taking

The number of turns is positively correlated to cohesion score, Pearson's (r=0.624, p= 0.01). A one-way ANOVA shows that there is statistically significant effect of cohesion score on the number of turns during interaction (f (1, 14) = 6.465, p = .023). High cohesive groups alter turns more frequently (M =23.75, SD=7.741) than low cohesive group (M = 15.125, SD = 5.667).

Overlaps

The number of overlaps has a strong positive correlation with cohesion, Pearson's (r=0.519, p=0.039). The number of overlaps in high cohesive groups (M = 27.62, SD = 4.92) is significantly higher than in low cohesive groups (M = 16.5, SD = 11.46), with (F (1, 14) = 5.327, p = 0.037).

Overlapped Interruption

The result of Pearson correlation indicates a strong positive correlation between number of overlapped interruptions and cohesion (r = 0.613, p = 0.008). A one-way ANOVA shows statistically significant difference in number of overlapped interruptions in low and high cohesion (F (1, 14) = 9.847, p = 0.007). High cohesive groups appear to have more interruptions (M = 9.75, SD = 3.327) in comparison to low cohesive groups (M = 4.62, SD = 3.20).

Paused Interruptions

Although the number of paused interruptions is positively correlated to the group cohesion score, Pearson's (r = 0.258, p = 0.334), the relation is not statistically significant. A paired-sample t-test indicated that scores were significantly higher for overlapped interruptions (M = 7.187, SD = 4.118) than the paused interruptions (M = 2.937, SD = 2.205) in order to grab the turn even if the speaker has not completed her utterance (t(16) = 3.5, df = 15, p = 0.003).

4.2.3 Discussion

Our aim was to analyse the relationship of the turn taking and interruption with cohesion. Table 2 summarizes the correlation between cohesion and features related to turn taking and interruption.

Table 2: Pearson's Correlation coefficients and p-values between cohesion and features related to turn taking and interruption.

Feature	Correlation with Cohesion	P value
Number of turns	0.624	0.01
Number of overlaps	0.519	0.039
Number of overlapped interruptions	0.613	0.008
Number of paused interruptions	0.258	0.334

Our hypothesis that the number of turns is higher in high cohesive groups and lower in low cohesive groups during multiparty interaction is validated. This result supports the findings of Hung et al. (Hung & Gatica-Perez, 2010). Results show that participants exchange turns more frequently in high cohesive groups since all the members of the group are actively participating in the interaction, thus increasing the number of turns. It also results in reducing the duration between two successive speaking turns compared to the duration in low cohesive group. We further hypothesized that the number of overlaps is higher in high cohesive group and lower in low cohesive groups, which is also validated by the results. The occurrences of overlaps during interaction are also positively correlated to the group cohesion. The reason is that the subset of the AMI corpus that we have chosen, contains task-oriented meetings where participants collaborate and discuss with each other to achieve their common objective. Our next hypothesis that the number of interruptions is higher in high cohesive groups and lower in low cohesive groups during multiparty interaction is also validated. The result is also in-line with the finding of Tannen (Tannen, 1994), which describes that interruptions are good indicators for the cohesion in the group when people are able to complete each other's sentences.

4.2.4 Future Work

In the previous section, we presented a preliminary analysis of interruption in group cohesion during multiparty interaction. We used AMI dataset to verify our hypothesis turn taking and interruptions associated with group cohesion. Future work will include the analysis of other multimodal phenomena such as gaze and dialogues act along with the interruptions with respect to group cohesion during multiparty interaction, and incorporating these behaviours into the group model. We will replicate the analysis on the Council of Coaches database.

5 Greta in Unity

This section describes how the communication and synchronization between Greta and Unity is can be configured.

5.1 Communication between Greta and Unity player

To run a Unity scene that includes a Greta agent, one must first run Greta and load a configuration file that includes the following Thrift modules (see Figure 6):

- Thrift Command Receiver
- Thrift Audio Sender
- Thrift FAP Sender
- Thrift BAP Sender

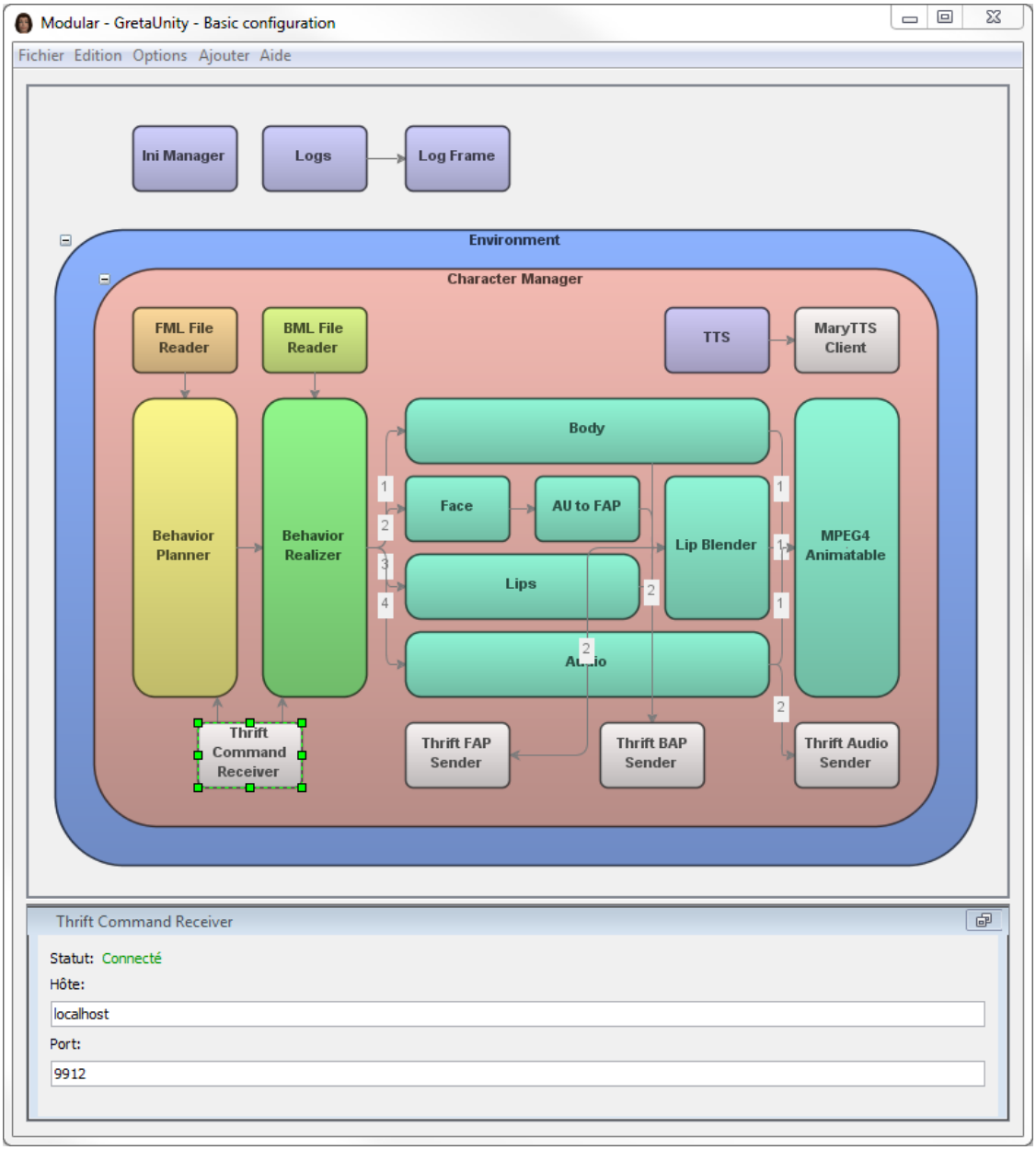


Figure 6: Greta configuration that includes the Thrift modules.

The "**Thrift Command Receiver**" is the main module which allows synchronizing agents, objects and events on them from Unity to Greta. The **Audio**, **FAP** (Facial Animation Parameters) and **BAP** (Body Animation Parameters) **Senders** are modules which send voice and animations from the Greta engine to the Unity player. Such Greta configuration can be found in the folder "{Greta}/bin/Configurations/GretaUnity/".

After that, it is possible to play a Unity scene that includes one or more GretaUnity agent(s). A **GretaUnity** agent is provided as a Prefab element for Unity, which includes the essential scripts to synchronize and animate it. Especially it includes the following scripts (Figure 7):

- **Greta Character Animator** (GretaCharacterAnimator.cs)
- **Greta Character Synchronizer** (GretaCharacterSynchronizer.cs)
- **Greta Object Tracker** (GretaObjectTracker.cs)
- **Animation Command Sender Tester** (AnimationCommandSenderTester.cs)

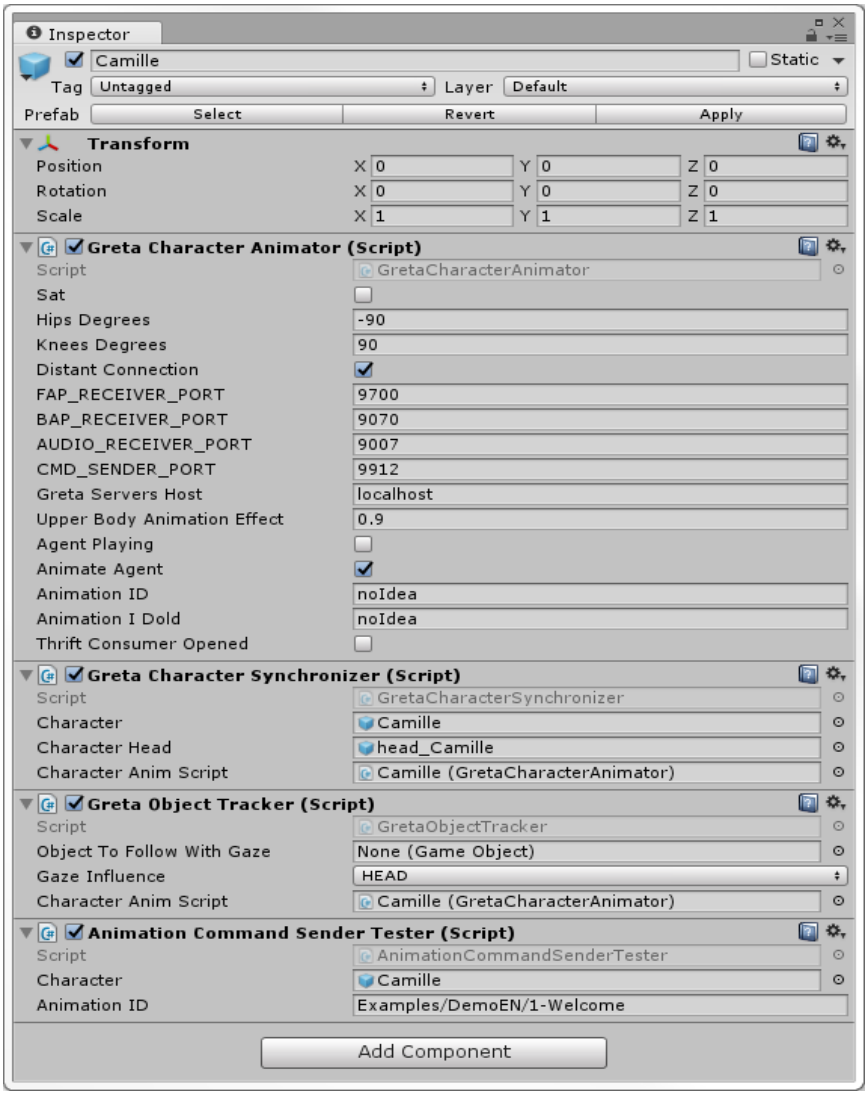


Figure 7: Prefab element for Unity, which includes the essential scripts to synchronize and animate a Greta character.

The "**Greta Character Animator**" is the main script which ensures responding to the Thrift modules of Greta and playing the corresponding animations. Indeed, it contains four elements:

- A Thrift command sender
- A Thrift audio receiver

- A Thrift FAP receiver
- A Thrift BAP receiver

The corresponding Thrift senders/receivers from Greta and Unity connect to each other with a specific host and port for each category of information. By default, in GretaUnity, these four ports are used (see Figure 8 below).

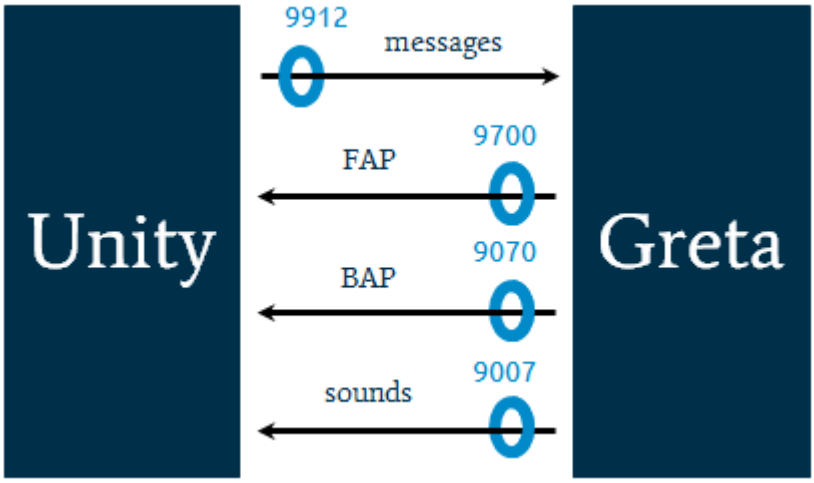


Figure 8: Different ports for communication between Unity3D and Greta agent.

The "**Greta Character Animator**" is also in charge to transform the character from the "T position" used by our characters to the "N position" used by Greta (Figure 9).

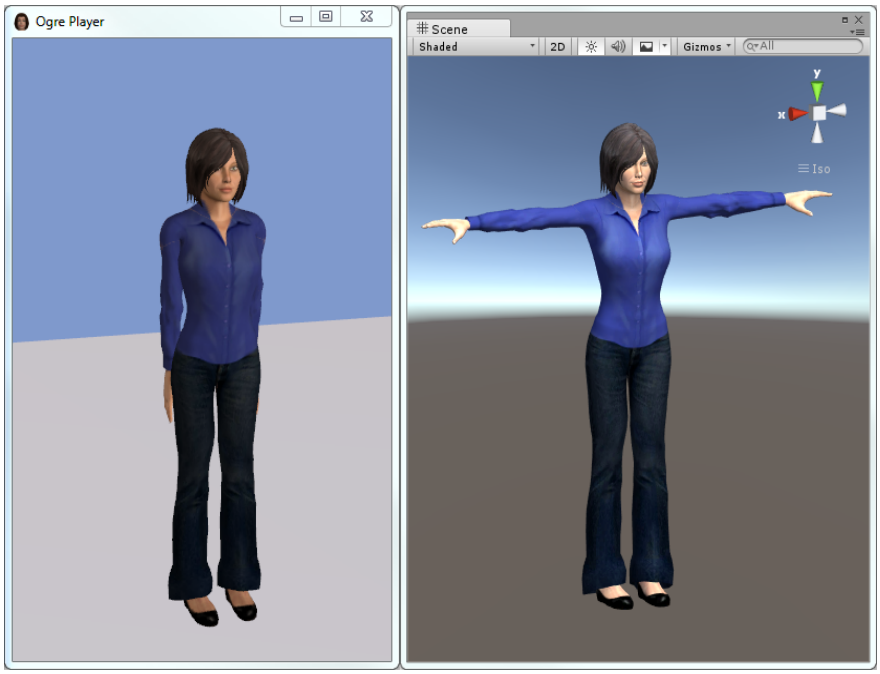


Figure 9: Visualisation of a Greta character in Ogre and Unity3D scene.

The "**Greta Character Synchronizer**" script is in charge of sending characters' position, rotation and size from Unity to Greta. It contains a parameter to know exactly where the character's head is in the Unity scene in order to calculate the correct angles for the eyes direction to simulate gaze behaviours.

The "**Greta Object Tracker**" script tracks an object in the scene. It takes an object to follow in parameter. It is possible to indicate how the gaze behaviour is performed. One needs to specify which body part is involved in the gaze behaviour. We consider five values: EYES, HEAD, SHOULDER, TORSO, WHOLE.

The "**Animation Command Sender Tester**" script sends an "Animation ID" (a FML file path without the .xml extension) in order to test GretaUnity capabilities.

On the Unity side, when the T key is pressed, the path to this FML file is sent to Greta which reads it and executes it. Internally, "**Animation Command Sender Tester**" asks "**Greta Character Animator**" to send this file path, which itself asks "**Thrift Command Sender**" to do so. In "**Thrift Command Sender**", a message with the type "*animID*" and with the file path is created. Then it is sent through the Thrift channel (port 9912 for messages).

On the Greta side, the "**Thrift Command Receiver**" module receives this message. It first looks at its type, then when it sees that it corresponds to "*animID*" it starts processing this file: it reads the FML file at the given location and propagates it content inside the Greta's internal system (from FML to Signals, from Signals to Keyframes and from Keyframes to FAP frames, BAP frames and Audio frames) to be executed by the virtual agent and be finally played by the Unity player after receiving corresponding animation parameters through the three other Thrift channels.

5.2 Synchronization between Greta and Unity player

In order to synchronize objects and external characters with Greta, there is one more script to add into the Unity scene: **Greta Environment Synchronizer** (GretaEnvironmentSynchronizer.cs) as shown in Figure 10.

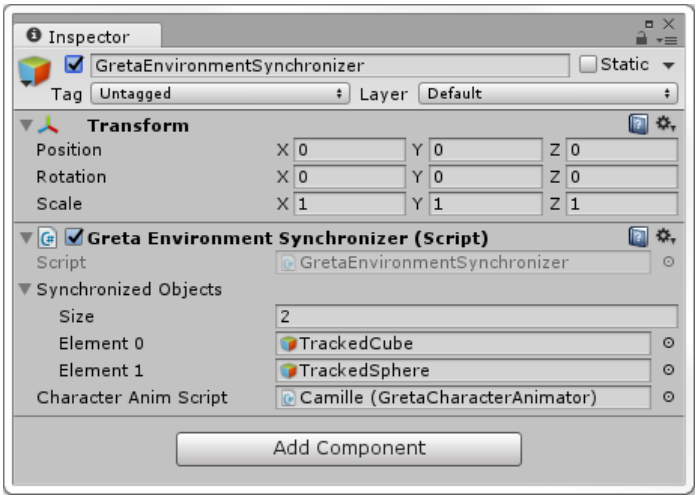


Figure 10: Greta Environment synchronization configuration in Unity.

This script indicates which objects will be synchronized from Unity to Greta. It is in charge to send objects position, rotation and size. It contains an array of objects and a parameter to indicate which Character Animation Script will be used to communicate this information. Indeed, as for "**Animation Command Sender Tester**", "**Greta Character Synchronizer**" and "**Greta Object Tracker**", this script uses "**Greta Character Animator**" to send and receive data to/from Greta.

At each frame, "**Greta Environment Synchronizer**" will check for each object in the list of synchronized objects if the object has changed its position, rotation or size. It is very easy to check this with the "*hasChanged*" boolean in the Transform of the object. Transform is a script attached to all Unity objects in a scene. It manages the position, rotation and size of the object. The "*hasChanged*" boolean is set to true as soon as a change occurs in the Transform. By overriding it, and checking when it changes to true, we can detect every change in the Transform. For each synchronized objects whose Transform has changed, the "**Greta Environment Synchronizer**" script directly asks the "**Thrift**

Command Sender" of "**Greta Character Animator**" to send a message to Greta by passing the related object to it.

In the "**Thrift Command Sender**", we retrieve the coordinates from the Transform of the object in order to communicate them to Greta. However, the coordinates cannot be sent as such, because the coordinate system is not the same in Greta and in Unity, and the objects do not have the same pivot (ref Figure 11, Figure 12). The Unity x-axis is inverted in Greta, so the x-coordinate must be inverted, but also the rotations in y and z must be inverted. The most delicate thing is to change the location of the pivot. The pivot of an object in Unity is in its center, while the pivot of a Greta object is in the bottom right of the object.

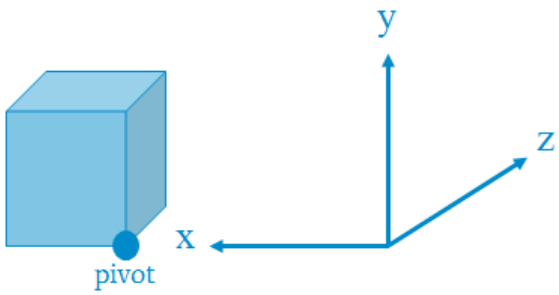


Figure 11: Coordinate system in Greta.

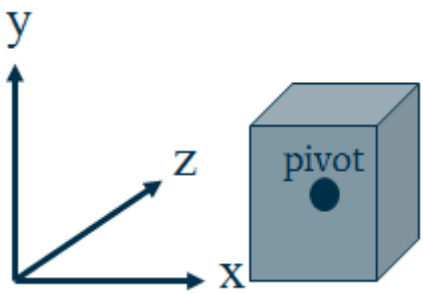


Figure 12: Coordinate system in Unity.

This means that when an object rotates or changes its size, its coordinates in Unity and Greta no longer match, even after reversing the x-axis. To fix this, we had to create a "shift" vector to add to the object's coordinates in order to adjust it according to the Greta pivot (Figure 13).

```
position = position vector of the object
quaternion = rotation quaternion of the object
scale = size vector of the object
shift = quaternion * (0.5 * scale.x, -0.5 * scale.y, -0.5 * scale.z)
positionGreta.x = -(position.x + shift.x)
positionGreta.y = position.y + shift.y
positionGreta.z = position.z + shift.z
```

Figure 13: Coordinate shift between Unity and Greta.

Once the coordinates have been updated for Greta, we create a message with the type "*object*". The content of this message is more complex than the animation file playback feature of the "**Animation Command Sender Tester**" script. We fill the property dictionary of the message with keys/values, both strings. We put the position in x, y and z, the rotation quaternion in x, y, z and w, the size of the object in x, y and z, as well as the unique identifier of the object in order to be able to recognize it later. Once the message is created, it is sent through the Thrift channel (port 9912 for messages).

On the Greta side, the "**Thrift Command Receiver**" module receives this message. Seeing that it has an "*object*" type, the *handleObjectMessage()* method is called. From the data in the received message, we

create the received object, or if it already exists in the environment of Greta, we update it (Figure 14).

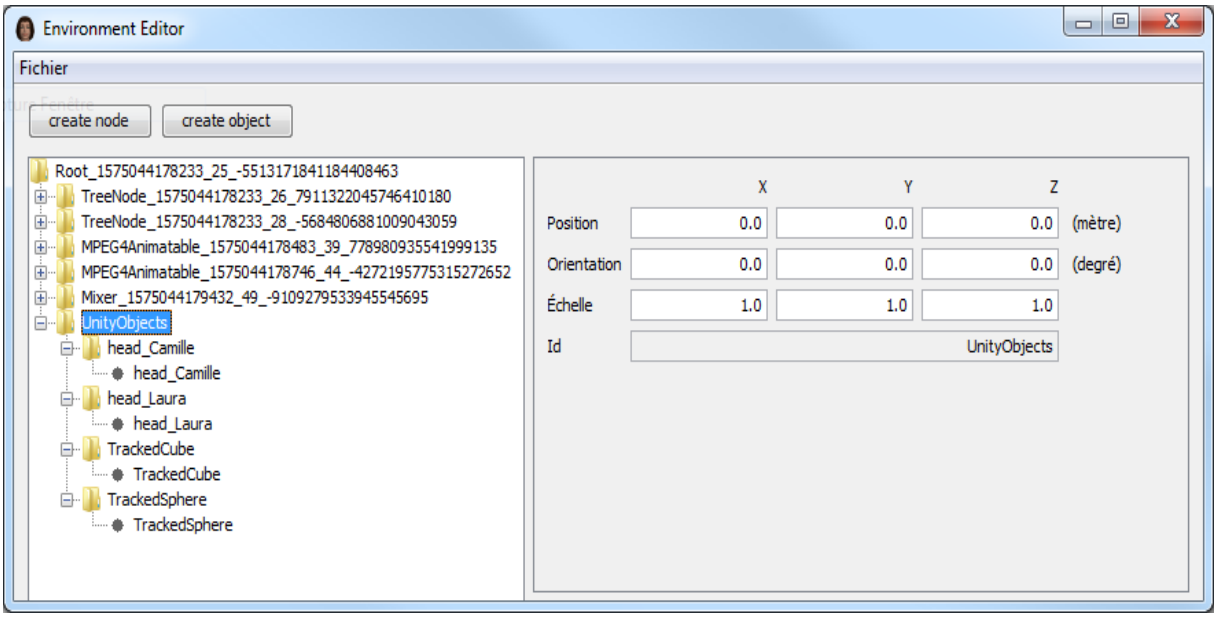


Figure 14: Visualisation of different elements in the environment through Environment Editor.

In order to better differentiate native Greta objects and the objects coming from GretaUnity, we add a "*UnityObjects*" node parent of all these GretaUnity objects. When a message is received, we first check if this node exists, otherwise we create it. Then, we check if the object is already in the Greta environment from its unique id sent by Unity. If it exists, we update its coordinates details based on those in the message properties. If it does not exist, we create it. This way the new objects are visible in the Greta environment and in the native Ogre player of Greta (Figure 15).

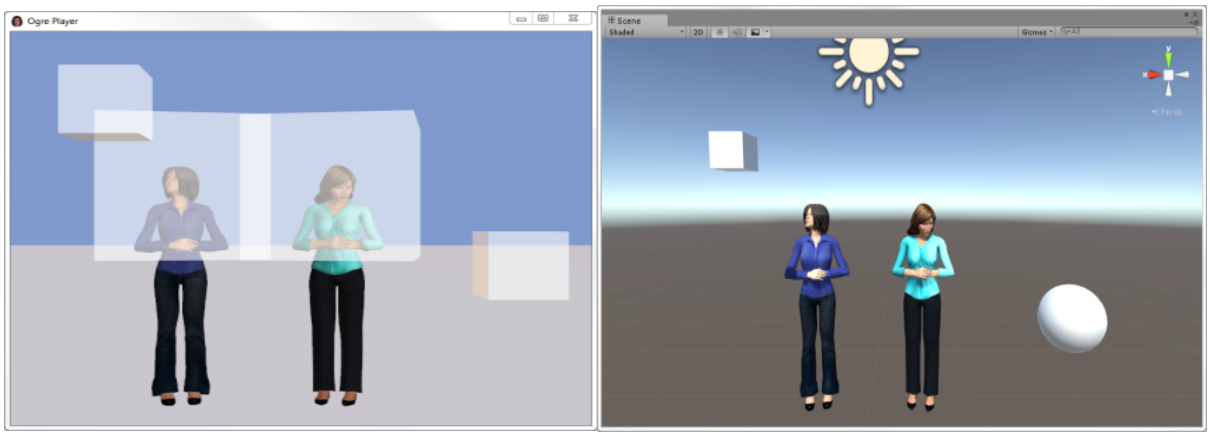


Figure 15: Visualisation of objects and virtual characters in Unity (left) and Ogre player of Greta (right).

The basic appearance of an object in the native Ogre player of Greta is a semi-transparent rectangular parallelepiped, which we considered sufficient to represent a Unity object. Indeed, it is more important to know the object coordinate in order to allow virtual agents to interact with it.

6 Further Council of Coaches system development

Besides the development of Greta, other improvements to the platform have also been made, and are still being made, that can enhance the dialogues with our coaches.

6.1 Multi-device capabilities

The Council of Coaches system now supports dialogue on multiple devices, such as tablets, and smartphones. These dialogues can be viewed as one cohesive group interaction with coaches represented on multiple devices at the same time. This allows for the use of different user interfaces, easier access to the coaches, and the option to have a coherent conversation with the coaches spread over multiple devices at the same time.

6.2 From sensors to dialogue

The connection between the sensors from the Holistic Behaviour Analysis Framework and the Knowledge Base has been made. The connection between the Knowledge Base and the dialogue components is being worked on as well. This will make the coaches able to use (processed) information gathered from the sensors in their dialogues. This can be used to support certain arguments, as well as decide on the best course of action in the dialogue. Furthermore, developments are being made to enable the coaches to store information they obtain from the dialogue with the user in the Knowledge Base. This way they can remember important information about the users to use at a later time. This can help build rapport with the users, as well as help gain valuable insight about the condition of the user.

6.3 Scenery and interface changes

The Unity scene, including the interface, has been changed. More photorealistic agents have been used to represent the ASAP agents. They are now more in line with the Greta agents. Furthermore, the scenery was changed by removing the large table which the coaches sat behind. They now sit in chairs with the user, mimicking a more open group talk with less of a divide between the user and the coaches. Finally, the user interface has been changed to show each message as it comes by, as well as allowing the user to control agents other than themselves. This can allow for the experience of helping oneself, as well as a way to pause the interaction if one needs a break by taking control of all involved parties.

7 Evaluation

In this section we describe two evaluation studies already performed related to the *Council of Coaches Technical Prototypes* and the *project's Functional Demonstrator*. In addition, we provide an overview of studies that will be performed in the period between January 2020 and August 2020 in the context of the Council of Coaches system.

7.1 Final Prototype Evaluation

In this section we describe the methodology used for the evaluation of the final technical prototype developed for the Council of Coaches.

7.1.1 System

The Council of Coaches system was installed on the laptop the researcher set up. An example of the setup in action from the perspective of the participant can be seen in Figure 16. It used the on-screen buttons for the participants to interact with the system, and had a scripted dialogue.

Figure 16: Screen shot of the setup from the perspective of the participant.

7.1.2 Roles and feedback

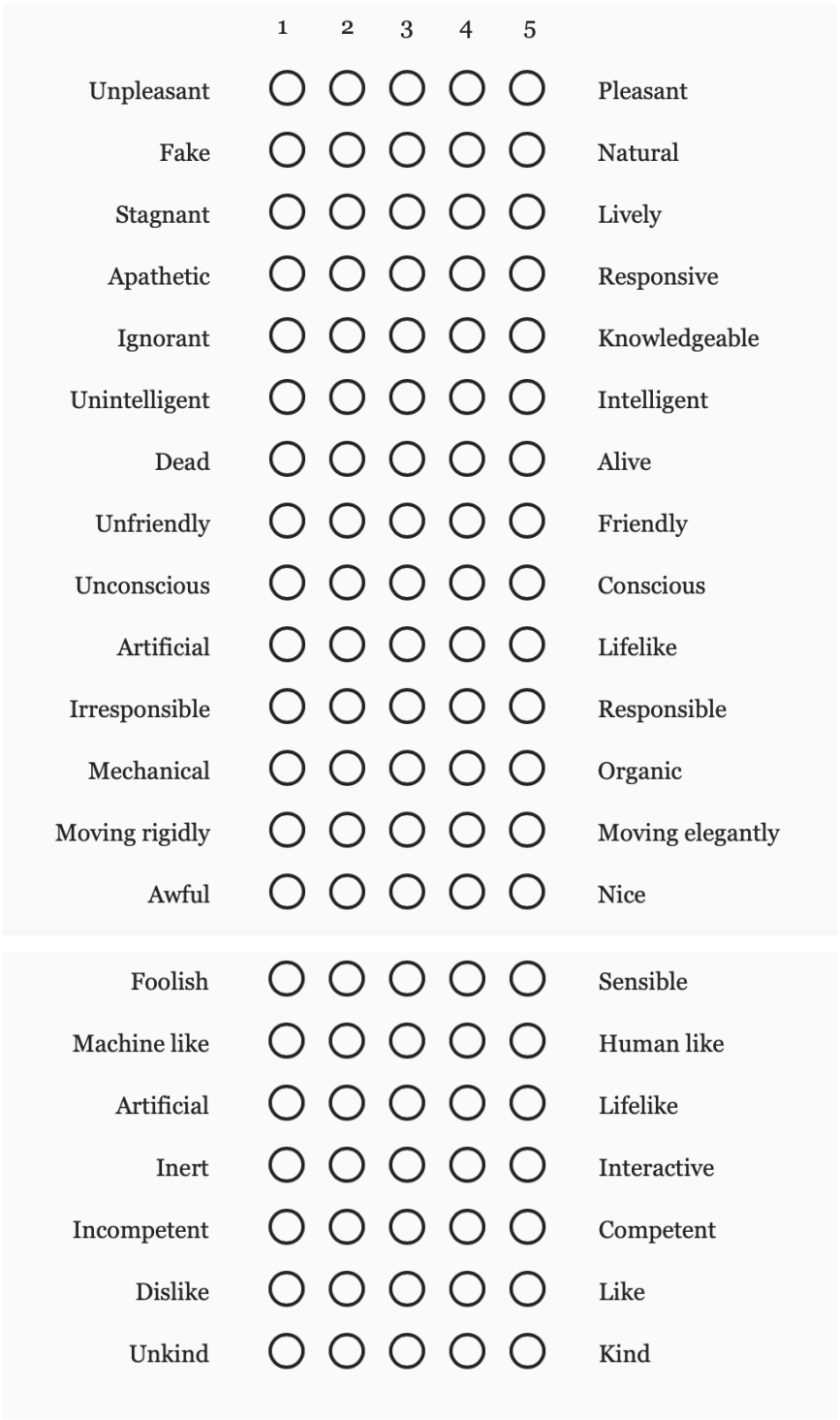
The council of coaches consisted of four different coaches and each had their own appearance, name, role, and expertise related to the topic of weight management.

1. Francois (Diet Coach): proposed a healthy recipe based on the user's dietary preferences that were collected during the interaction.
2. Olivia (Physical activity Coach): recommended the user go for a walk around the block once a day around meal time.
3. Emma (Social Coach): suggested the user to have a friend or a family member accompany them during the walk.
4. Carlos (Peer): provided supportive dialogue emphasizing the expertise of the coaches and the efficacy of their coaching.

7.1.3 Questionnaire

We made use of the Godspeed questionnaire to measure the *Animacy, Anthropomorphism, Likeability* and the *Perceived intelligence* of our coaching agents. We did not utilise the *Perceived safety*

questionnaire, as we did not expect the discussed subject to have a strong emotional impact with regards to anxiety, agitation, or surprise. The following is the questionnaire answered by the participants as shown in Figure 17.



	1	2	3	4	5	
Unpleasant	○	○	○	○	○	Pleasant
Fake	○	○	○	○	○	Natural
Stagnant	○	○	○	○	○	Lively
Apathetic	○	○	○	○	○	Responsive
Ignorant	○	○	○	○	○	Knowledgeable
Unintelligent	○	○	○	○	○	Intelligent
Dead	○	○	○	○	○	Alive
Unfriendly	○	○	○	○	○	Friendly
Unconscious	○	○	○	○	○	Conscious
Artificial	○	○	○	○	○	Lifelike
Irresponsible	○	○	○	○	○	Responsible
Mechanical	○	○	○	○	○	Organic
Moving rigidly	○	○	○	○	○	Moving elegantly
Awful	○	○	○	○	○	Nice
Foolish	○	○	○	○	○	Sensible
Machine like	○	○	○	○	○	Human like
Artificial	○	○	○	○	○	Lifelike
Inert	○	○	○	○	○	Interactive
Incompetent	○	○	○	○	○	Competent
Dislike	○	○	○	○	○	Like
Unkind	○	○	○	○	○	Kind

Figure 17: Questionnaire answered by the participants.

Further, we modified the System Usability Scale questionnaire to suit our study. We removed the questions related to use of the product in terms of a new technology since we used a simple computer interface. However, we still retained some questions related to the ease of use.

- I think that I would like to use this product frequently.

- I thought the product was easy to use.
- I think that I would need the support of a technical person to be able to use this product.
- I found the various functions in the product were well integrated.
- I thought there was too much inconsistency in this product.
- I felt very confident using the product.

Finally, we asked two general open-ended questions about the participant's opinion of the system and the agents to capture the overall impression, and to find out if they would recommend the system to others.

7.1.4 Design

The experiment for the final evaluation of the technical prototype involved one user interaction that lasted approximately five minutes. The target subject group for this study was adults in the age group of 40 - 65 years old. We chose this age group as it is expected to be (a) slightly younger and slightly more familiar with technology than the project's main target population of 55+, and therefore more likely to be able to successfully interact with the technology, and (b) easier to recruit for in the premises of the two technical partners conducting the experiment. The experiment was conducted in English. The dialogue for the coaches were derived from the content from the dialogues developed for the functional prototype. The scenario consisted of four coaches, where two are ASAP agents and two Greta agents.

7.1.5 Procedure

Every participant interacted with the Council of Coaches system individually. Each participant was let into the experiment room with the setup as seen in Figure 18 at Sorbonne University, and a similar setup at the University of Twente. The main difference was that the University of Twente used a tablet for the questionnaire instead of a second computer screen. Besides this, the setup was very similar.

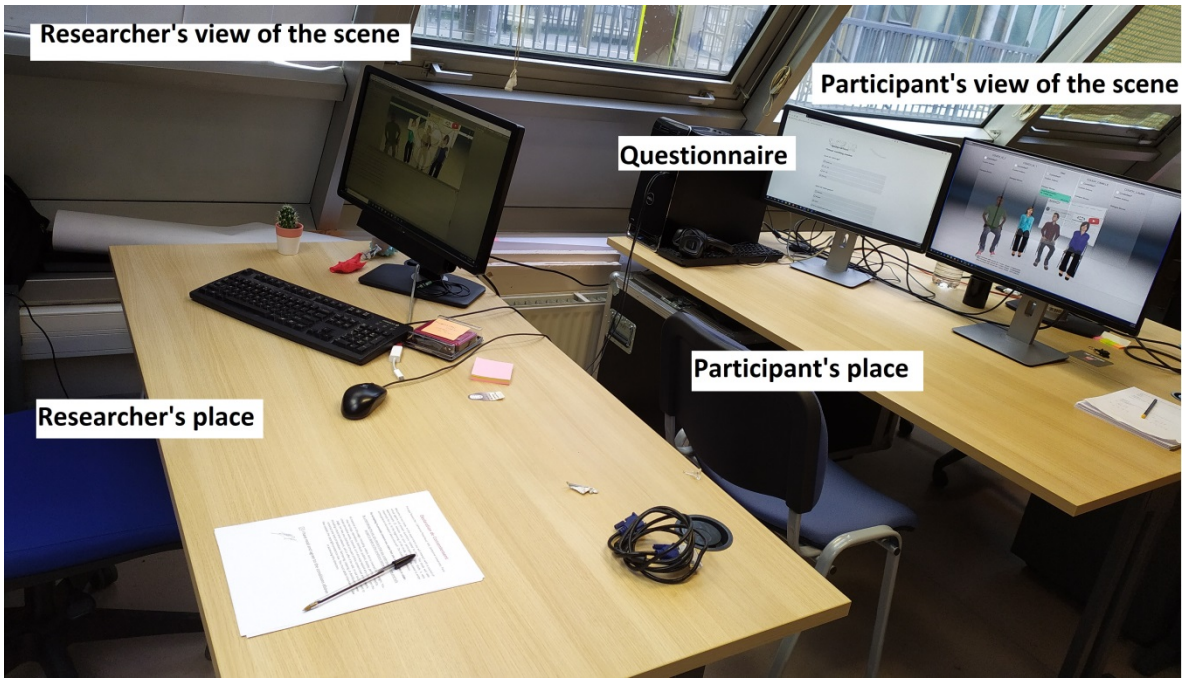


Figure 18: Initial experiment setup.

The participants were first asked to read the information letter and sign the informed consent form, as well as ask any questions they had. Once these questions were answered to the satisfaction of the participant, the researcher would also sign the informed consent form, and would offer a copy of them. Afterwards, the researcher provided a brief description of the experiment and explained tasks to be performed by the participant. This took around 10 minutes.

After the explanation was finished, the participant started by filling out their demographic information on a tablet. Then the researcher explained that the following interaction would be to get to know the coaching team, and to learn how to work with the interface. The participant was encouraged to ask the researcher any questions they had during the interaction. During the interaction the coaches introduced themselves and provided an interactive coaching session on healthy weight management. This interaction lasted about 5 minutes. After the interaction, the participants were asked to answer the selected part of the Godspeed questionnaire series, followed by SUS questionnaire on a tablet. Filling out the questionnaire took around 10 minutes.

7.1.6 Sample

The prototype was setup at Sorbonne University and the University of Twente, and we collected responses from 7 and 13 participants respectively. In total we had 20 participants, with 40% being female (n=8) and 60% being male (n=12). A total of 30% of the participants were below 55 years old while 70% were aged 55 years or older, thus falling within the primary target population for Council of Coaches. A total of 75% of the participants had never interacted with a virtual agent prior to this study.

7.1.7 Results

We made use of an adapted version of Godspeed questionnaire and System Usability Scale questionnaire for this study. The results indicate that participants rated the agents high on likability (m = 3.67) and perceived intelligence (m = 3.63), but did not have high score on anthropomorphism (m = 2.76) and animacy (m = 2.93). The agents were particularly perceived to be very friendly (m = 4.05) and kind (m = 3.85). 70% of the participants thought that the system was well integrated. 55% of the participants indicated that they would use the system frequently and 50% felt confident using the system. However, 35% of the participants reported that they would need technical assistance. Overall, 80% of the participants said that they would recommend the system to their friends and family.

We also summarized the main feedback from the open questions. Regarding the opinion of participants about the system and agents, the feedback was mixed. On the positive end, several participants mentioned they enjoyed using the system (6), remarking how easy it was to use, how pretty it looked, and that it felt comfortable to use. Several positive remarks about the agents were made as well (6), such as their advice being sensible, their behaviour making a calm impression, the speech of some of the coaches being quite clear, and some of the coaches behaving like a coach would. Some negative remarks about the system and agents were made as well. Several indicated disliking working with the system (3), saying the interaction was very much a one-way street and did not feel that useful, and that their choices felt quite limited. The remarks about the agents focused mostly on issues with their mechanical movement (10), often mentioning the agents not gazing at the user at the right times, the robotic voices of the agents (4), which were said to be hard to understand, and the coaches overacting (2), which made them seem overly friendly and a bit fake.

The feedbacks on whether or not participants would recommend the system to others were also mixed. Several remarks were positive (10), saying they would recommend it based on how nice it was to have multiple people with multiple opinions to talk to, and how this could help reflect (3), the advice being useful (2), the presentation format of information being good and easy to understand (5). There were also several negative remarks (11). They mentioned it not yet feeling human, personalized and interactive enough to interact with (7), the information being irrelevant and not useful (2), and the interface and animations feeling unfinished (2), for example the interface names being in the wrong order for the coaches and not containing their actual names.

7.1.8 Discussion

Since 70% of population were above 55 years, their expectations of agents were probably a bit technically unrealistic, and hence the agents scored low on the anthropomorphism and animacy. Furthermore, we need to consider the fact that 75% of them had never interacted with a virtual agent before and their perception was probably influenced by media (films). We got comments that the gaze behaviour of agents was not right as sometimes, the agents do not look at the user when the user speaks. This could also be related to the remarks about the interaction not being human enough yet. Thus, we aim to develop a group gaze model to improve the gaze behaviour of agents during the

interaction. Furthermore, we got feedback on voice quality of the agents. We will look into potential alternative voices. Moreover, the participants felt that the agents were friendly and perceived them to be knowledgeable in their domain.

Even though the interface was simple (clicking buttons), the overall system might have looked complicated for the older participants. If it was just launching an app or a browser, the user might have not felt that they required assistance. Thus, we will provide a simple tutorial/instruction at the start of the experiment to make the user feel comfortable using the system in our future study. Furthermore, we had feedback on the layout of the interface and the information it presented, and will make sure to update the interface for our future study using this feedback.

7.2 Multi-perspective persuasive discussion evaluation study

In this section, we describe a previously conducted evaluation study in which we used the Council of Coaches system. The subject of this study was multi-perspective persuasive discussion.

7.2.1 Objectives

To explore the effects of group discussion and multi-perspective persuasion using a virtual council of coaches trying to help achieve a health goal (weight management), we aimed to find answers to the following question: “What is the effect of inter-coach discussion during a persuasive dialogue in a coaching session on the perception of the council of coaches, the perception of the council's ability to coach, and the council's actual coaching ability?” To answer this question, we investigated the following research questions.

Does inter-coach discussion during a persuasive group dialogue lead to a change:

- in perception of a virtual council of coaches?
- in perception of a virtual council of coaches' ability to coach?
- in reflection on which approach to choose and why to choose it?
- in commitment to follow a chosen approach?
- in enjoyment of, and preference for an interaction?

7.2.2 Participants

We recruited 45 participants at the University of Twente. All of them had the ability to work with a computer, and could converse effectively in English. Our sample contained mostly students, as well as a few working adults. Due to technical issues disturbing our procedure during a few sessions, seven participants were excluded from analysis. The remaining group of 38 participants consisted of 22 male, and 16 female participants that were between 18 and 35 years old (M = 22.45, SD 3.438).

7.2.3 System

The virtual council system was installed on the laptop that the researcher set up. The laptop was connected to an external screen, external speakers, and a computer mouse. The user interface consisted of an environment with a browser in the background, a table at which a virtual council of three coaches sat down, and buttons that would appear for the participant to respond to the coaches. The buttons would only be on screen when the participant needed to respond to the coaches, and not while the coaches were talking. The council of coaches consisted of three coaches that each had their own appearance, name, role, and expertise related to the topic of weight loss. Figure 19 shows the coaches in the scene. From left to right, it shows Harm (discussion lead, and mental coach), Francois (diet coach), and Alexa (physical activity coach).

Figure 19: Coaching scene from the perspective of the participant.

Six tips were presented in two rounds of three tips, with a request for feedback on the tips by Harm between the rounds. Each coach would give one tip per round using their expertise. The tips were offered in the following order:

Round 1:

1. Francois (diet): Lower your sugar intake.
2. Alexa (physical activity): Start a daily exercise routine consisting of jogging.
3. Harm (mental): Identify troubling thoughts, and tell yourself out loud to stop. Then try to introduce healthier thoughts.

Round 2:

4. Alexa (physical activity): Do strength training two to three times per week.
5. Harm (mental): Make sure you get enough sleep every night.
6. Francois (diet): Drink more water, especially shortly before each meal.

7.2.4 Questionnaires and interview

We used a brief questionnaire asking for participants' age, gender, and experience interacting with virtual agents for demographic purposes. To answer our research questions, we used a part of the Godspeed questionnaire series (Bartneck, 2008), and an adjusted version of the Coaching Behaviour Scale for Sport (CBS-S) (Côte, Yardley, Hay, Sedgwick, & Baker, 1999). Within both of the questionnaires, the order of the questions was randomized for each condition for each participant. Furthermore, we used interview questions developed to get more in depth answers from the participants.

For the Godspeed questionnaire series, we selected the anthropomorphism, animacy, likeability, and perceived intelligence questionnaires. We left out the perceived safety questionnaire, as we did not expect our interactions to have a strong impact on participants with regards to anxiety, agitation, or surprise.

The original CBS-S we found contained several items that were not relevant to the interactions in our experiment. We did not use items on the scales of physical training and fitness, technical skills, competition strategies, personal rapport, and negative personal rapport (i.e. items 1 to 15, and items

27 to 47). On the scale of mental preparation, we did not use the item regarding performance under pressure (i.e. item 16). On the scale of goal settings, we did not use the items regarding monitoring of progress, identifying target dates for attaining goals, and setting long-term goals (i.e. items 22, 24, and 25). The items that we used, were rephrased from "the coach(es) most responsible for my" to "my coaching team" as we used a virtual council of coaches. Furthermore, several items relevant to the

interactions in our experiment were added under a new coaching quality" scale. These were the following items:

1. My coaching team helps me to be motivated and inspired by others.
2. My coaching team helps me to discover which things help me to attain and maintain my healthy weight better.
3. My coaching team had the right knowledge and abilities to give good coaching.
4. My coaching team gives advice of good quality.

The interview questions were about the experience of working with the system, behaviour of the coaching team and interactions with them, advice chosen by participants, commitment to the advice, and reasoning for the commitment, intention to use the system again, and recommendation of the system.

7.2.5 Experimental design

The experiment used a 1x2 within-subjects counterbalanced measure design. The independent variables were the following two interaction conditions:

- **Condition 1**: Each of the three coaches presented one tip. Then coach Harm asked how the participant liked the advice. Then the coaches presented another tip each. At the end, coach Harm asked the participant how they all liked the advice, and which of the six tips the participant preferred. Once the participant indicated the advice they liked best, the coaches would offer words of encouragement to the participant, wish the participant luck, and close out the conversation.
- **Condition 2**: The three coaches presented the same tips. Harm asked the same questions of the participant between the rounds, and after the second round, as he did in Condition 1. The coaches offered the same words of encouragement, and closed out the conversation in the same way. In contrast to Condition 1, when transitioning to the next advice, the coaches would briefly interact with each other regarding their advice, mimicking a real-life group discussion.

7.2.6 Procedure

Participants were individually tested. Each of them was let into the experiment room with the setup being ready. The experiment was conducted in English. The participants were asked to read the information letter and sign the informed consent form. Afterwards, the researcher would sign the informed consent form, and would offer a copy of the information letter and informed consent form. Then, the researcher would explain the procedure and tasks to the participant.

After the explanation, the participant would fill out their demographic information on a tablet. Then the researcher would explain that the introduction had the purpose of getting to know the coaching team, and learning how to work with the interface. They could ask the researcher questions. In the introduction the coaches gave their name, and briefly explained their expertise.

The participant then had two interactions with the coaches, specified in section 7.2.5 (conditions). After each interaction, they answered the Godspeed questionnaires, followed by the adjusted CBS-S on a tablet.

Once the participant was done with the two interactions and rounds of questionnaires, the researcher verbally asked for permission to record the interview. If consent was given, they proceeded to conduct an interview with the participant. The topics discussed are described in section 7.2.4.

7.2.7 Results and discussion

Research question 1: Perception of Council of Coaches

Our statistical analysis showed no significant effect on any of the used Godspeed questionnaires. We saw a trend towards significance for a more positive rating of Condition 1 on the animacy questionnaire ($t(37) = -1.928$, $p = .062$), with a medium effect size ($r = .30$). The higher animacy in Condition 1 could be related to participants mentioning during interviews that they noticed differences

between the conditions, such as the coaches moving and speaking more fluidly in Condition 1 as compared to Condition 2.

Research question 2: Perception of Council of Coaches' coaching ability

Our statistical analysis showed no significant effect on any of the adjusted CBS-S scales. We did have several participants stating during their interview that they picked advice in Condition 2, because the quality of the advice was better there. This leads us to believe that there could be a small effect on perceived coaching ability due to inter-coach discussion. It may have been masked here by a substantial number of participants not noticing the inter-coach discussion.

Research question 3: Reflection on choices that were made

During the interviews, participants often indicated they made a choice based on personal reasons, such as already being committed to the chosen advice, or the novelty of the information. As previously mentioned, some participants did mention picking advice in Condition 2, because the quality of the advice was better than in Condition 1. This did influence their choice, according to them. Considering the amount of times reasons were mentioned, some participants mentioning the better quality of advice in Condition 2 does suggest there was an impact of the inter-coach discussion on the reflection people had about what approach and advice to choose, and why to choose it, but only a small one.

Research question 4: Commitment to a chosen approach

In the interviews, participants indicated a stronger commitment to their chosen advice in Condition 2, as compared to Condition 1. This was the case for those that started with Condition 1 (Condition 1: M = 5.26, Condition 2: M = 5.53), and those that started with Condition 2 (Condition 1: M = 5.55, Condition 2: M = 5.71). Though the differences were not huge, and many participants gave similar ratings in both conditions, these differences do indicate that the inter-coach discussion increased reported commitment by the participants.

Research question 5: Interaction preference

During the interviews, participants were asked to indicate the differences they perceived between the conditions, and which condition they preferred (see Table 3).

Table 3: Identified differences Condition 1 and Condition 2, and preferences (N = 38).

Condition order	Identified difference	Condition 1	Condition 2	No preference
1-2	Correctly identified	0	9	2
1-2	Incorrectly identified	2	0	0
1-2	Not identified	0	1	5
2-1	Correctly identified	1	1	1
2-1	Incorrectly identified	6	0	1
2-1	Not identified	0	0	9

We saw that 15 of the participants could not identify any difference, and 9 participants identified a difference that was not present. The remaining 14 participants did identify the main manipulated difference. Those that could not identify the difference generally did not have a strong preference (14 of 15 participants), and those that incorrectly identified the difference generally had a preference for Condition 1 (8 of 9 participants). Finally, those that did identify the main manipulated difference generally preferred Condition 2 (10 of 14 participants). This indicates a potential change in preference for an interaction due to inter-coach discussion. Which direction this change goes seems to be linked to whether the participant perceived the inter-coach discussion (preference inter-coach discussion), thought they perceived another difference which was not there (preference no inter-coach discussion), or did not notice any difference (no preference).

7.3 Future Studies

In this section we provide a brief description of the evaluation studies we are planning to conduct in the coming months.

7.3.1 User interface usability evaluation

In this experiment, we plan to evaluate the user interface of the Council of Coaches system. It will involve one or more conditions, and will have a within-subject or between-subject comparison. The target age group for this study will be adults between 55 – 70 years of age. The experiment will be conducted in English and/or Dutch. The scenario will be the coaches discussing a general health topic. The interactions will be presented through live interaction with the Council of Coaches system, or as interactive video recordings of the Council of Coaches system including a user interface. Through this experiment we want to measure the usability of the interface, as well as user interface preferences of older adults. A questionnaire will be designed to ask about the usability of the system as a whole, as well as different facets of the interface, such as the size of text and buttons, as well as the need for subtitles. Furthermore, interview questions will be asked regarding these topics, to allow for further elaboration by participants.

7.3.2 Multi-device interaction evaluation

In this experiment, we plan to evaluate the multi-device interaction with the Council of Coaches system. It will involve one or more conditions, and will have a within-subject or between-subject comparison. The target age group for this study will be adults between 55 – 70 years of age. The experiment will be conducted in English and/or Dutch. The scenario will be the coaches discussing a general health topic, and will potentially also involve reminders and/or tasks to be done on the mobile phone. The interactions will be presented through live interaction with the Council of Coaches system, or as interactive video recordings of the Council of Coaches system including a user interface, as well as a mobile phone interface, potentially a tablet interface, and if possible, other devices. Through this experiment we want to measure the usability of multiple devices and interfaces at once during an interaction, as well as the experience of older adults with this multi-device interaction. A questionnaire will be designed to ask about the usability of the system as a whole, as well as the different devices and their interfaces, and the experience the participants had during the multi-device interaction. Furthermore, interview questions will be asked regarding these topics, to allow for further elaboration by participants.

7.3.3 Verbal conflict presentation style impact on group discussion evaluation

Part of the feedback from participants in the study described in section 7.2 was on the rather competitive and aggressive way the coaches spoke to each other when they discussed their opinions. Participants remarked this could be improved. This could make the coaches more realistic, likeable, and effective at coaching. In this experiment, we plan to compare several different forms of verbally presenting a conflict. The experiment will involve two or more conditions and will have a within-subject or between-subject comparison. The target age group for this study will be adults between 18 – 65 years of age. The experiment will be conducted in English and/or Dutch. The scenario will be the coaches discussing a difference of opinion they have on a general health topic. The different conditions showing the forms of verbal conflict presentation to display this conflict will be presented as (potentially interactive) video recordings of the agents in the system to the participants, or through live interaction with the Council of Coaches system. Through this experiment we want to measure the impact of different forms of verbal conflict presentation on the perception participants form of the coaches, as well as on the persuasion of participants. A questionnaire will be designed to measure perceptions of the group of agents, the persuasiveness of the interaction, potentially group cohesion, and, in case the (video) setup will be interactive, the experience participants had interacting with the agents.

7.3.4 Peer agent presence and behaviour impact on group discussion evaluation

In previous work (Dohsaka, Asai, Higashinaka, Minami, & Maeda, 2009), the impact of having a peer agent be present during an interaction has shown beneficial. One can imagine potential benefits in a coaching context. Having a peer that empathizes with participants could improve the experience participants have during the group interaction, and might make them more open to share information. The peer agent could also support proposals by coaches and emphasise how they used that advice to enhance their quality of life, potentially increasing the credibility of the coaches and the advice they give. Finally, they could mediate a discussion between coaches. In this experiment, we plan to compare a group coaching interaction including a peer agent to one without a peer agent. The experiment will involve two or more conditions and will have a within-subject or between-subject comparison. The target age group for this study will be adults between 18 – 65 years of age. The experiment will be conducted in English and/or Dutch. The scenario will be the coaches discussing a general health topic. The conditions consist of those with a peer coach, in which the behaviour of the peer coach varies between conditions, and the condition without a peer coach. The interaction will either be presented as (potentially interactive) video recordings of the agents in the Council of Coaches system to the participants, or through live interaction with the Council of Coaches system. Through this experiment we want to measure the impact of the presence of a peer agent and their behaviour on the perception participants form of the coaches, as well as on the persuasion of participants. A questionnaire will be designed to measure perceptions of the group of agents, the persuasiveness of the interaction, potentially group cohesion, and, in case the (video) setup will be interactive, the experience participants had interacting with the agents.

7.3.5 Gesture generator evaluation

The experiment for the gesture generator evaluation will involve two or more conditions and will have a within-subject or between-subject comparison. The target subject group for this study will be adults in the age group of 18 – 65 years old. The experiment will be conducted in English and/or French. The scenario will consist coaches performing various gestures corresponding to the status of the agents designed e.g., authoritative, peer. Literature shows that the gestures performed by the agent can have an effect on the perceived warmth and competence of the agent (Ravenet, Cafaro, Biancardi, Ochs, & Pelachaud, 2015) which in turn affects the persuasiveness of an agent (Kantharaju, De Franco, Pease, & Pelachaud, 2018). Through this experiment we want to measure the effectiveness of the automatic gesture generator. The main feature of the system that will be evaluated will consist of several types of gestures performed by the agents based on their characteristics. A questionnaire will be designed to measure the overall perception of the agents and the perceived level of warmth/competence.

7.3.6 Cohesive group evaluation

The cohesive group evaluation study will involve two or more conditions and will have a within-subject or between-subject comparison. The scenario will consist of three or more coaches (virtual agents) interacting with each other and the user. The main feature of the system that will be evaluated will consist of (a) the gaze behaviours corresponding to cohesive and non-cohesive group of agents, (b) different turn taking behaviours of the agents that might affect the overall perception of the group of agents by the user and also the level of user engagement, and (c) several non-verbal behaviours corresponding to the designed characteristics to simulate either a cohesive or a non-cohesive group of agents. The target subject group for this study will be adults in the age group of 18 – 65 years old. The experiment will be conducted in English and/or French. A questionnaire will be designed to measure the overall perception of agents, the perception of gaze behaviour, the overall perception of the group of agents, the level of engagement of the user with the agents, the perceived level of cohesion and the overall performance.

8 Software documentation

In this section, we describe how to define new agents and behaviours using the Council of Coaches system. This includes best practices, as well as example setups. This is an ongoing effort which will be continually updated as the project progresses to the final review.

8.1 Defining new agents

The agents can be generated using several software available for creating 3D animated agents. Here we show an example using AutoDesk Character generator to create our virtual coach (Figure 20).

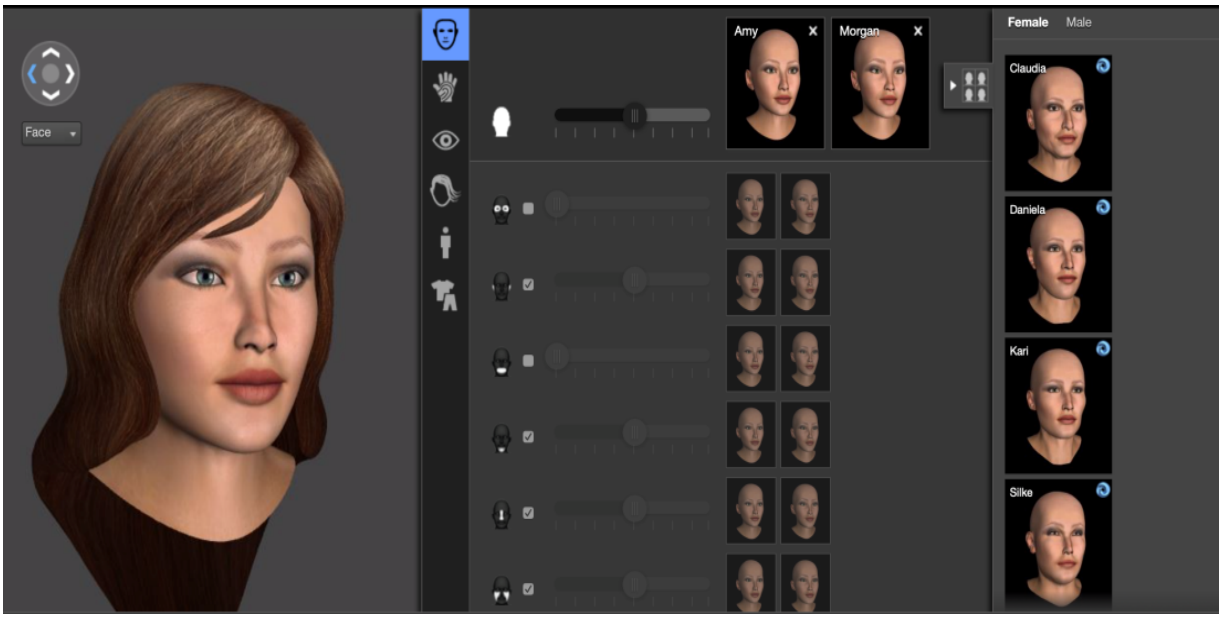


Figure 20: A screenshot of the several face and body parameters available to generate a virtual character.

Once the parameters are selected as per the preference, we need to download the character with the following parameters as shown in the figure below (Figure 21).

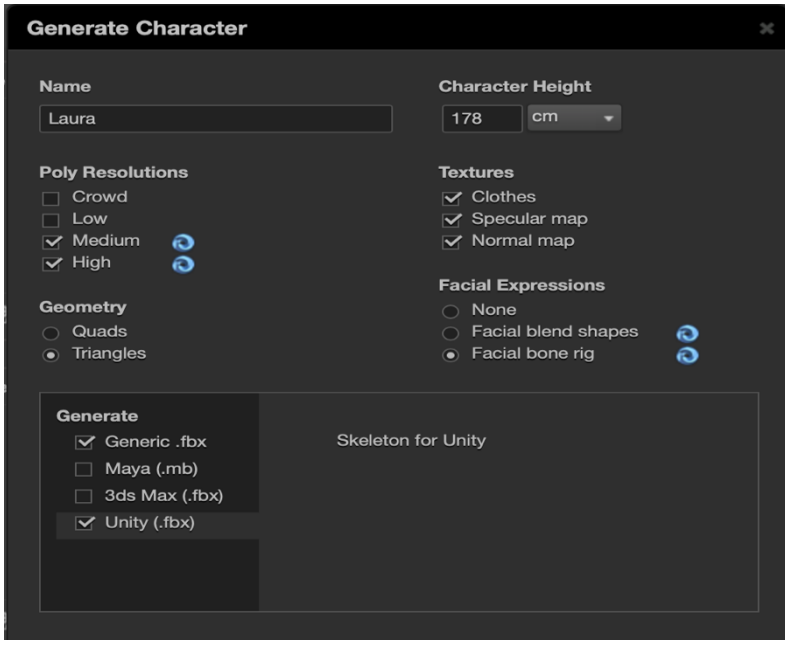


Figure 21: The parameters to be chosen for downloading a character.

Once the *.fbx* file is downloaded, we need to import the object to the Couch scene in the Unity environment.

The configuration of different modules in Greta has been described in the section 4 of Deliverable D6.3 “First Prototype description and evaluations of the virtual coach platform”. This configuration allows sending FML or BML files, processing them and scheduling verbal and non-verbal behaviours of the Greta agent. Moreover, the detailed description of Greta and its components are described on the wiki: https://github.com/isir/greta/wiki. Furthermore, the integration of gaze behaviour in the Greta agent has been detailed in the Section 4.2 of the deliverable D6.4 “First virtual coach design and model”.

The integration and synchronization of Greta with Unity3D have been detailed in the Section 5 of this deliverable D6.5. The connection between Greta and Unity3D is made through the Thrift modules. These modules allow to communicate between Greta and Unity3D, and to synchronize Greta and Unity environment with each other.

8.2 Authoring new dialogues

Due to the WOOL integration made with DGEP (see section 6: *The WOOL dialogue framework* of deliverable D3.4: *Final coaching actions and content*), dialogues can be authored rather easily. These dialogues would look like the brief bit of dialogue that can be seen below in Figure 22.

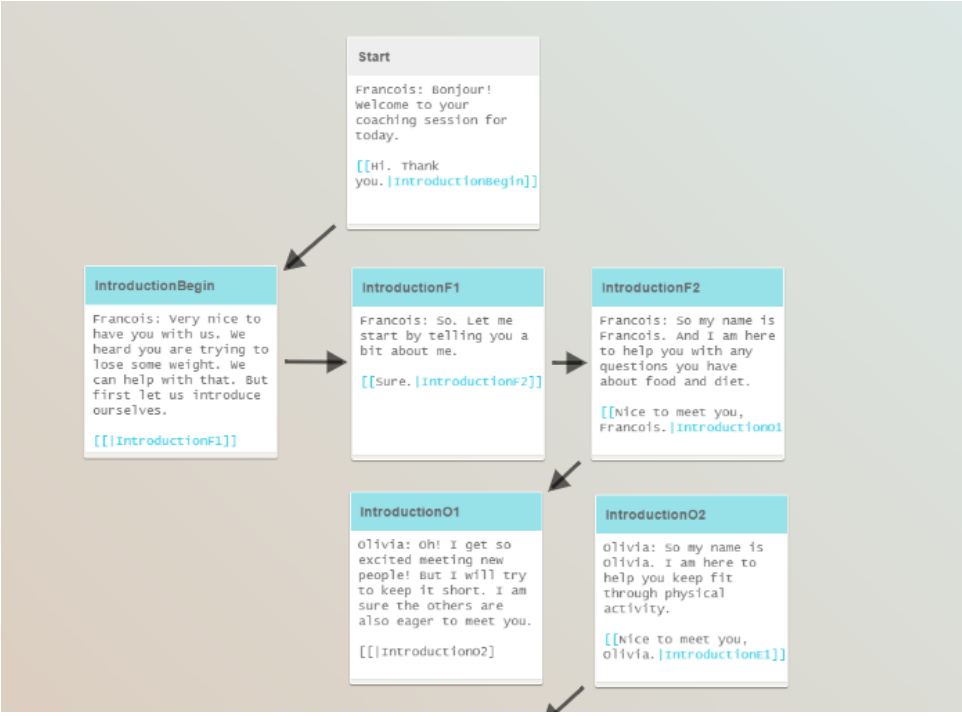


Figure 22: Example of a short WOOL dialogue.

Each dialogue box should consist of the following components:

- A title, of which the first one in the interaction chain should be named “Start”
- A body, containing
  - The (coaching) agent one wants to speak, if one does, in the form of a name followed by the : symbol, for example, “Francois:”.
  - The message they are conveying in plain text. Preferably using short sentences.
  - Add a white space in between.
  - Potential paths to other dialogue boxes, containing, in order
    - The [[symbols to open the message.
    - The text that should appear on the user interface button, if one wishes to give the user the option to reply. If one does not, this can be left empty.
    - The | symbol, to indicate the next part of text is the reference to the next dialogue step.

- The reference to the next dialogue step, which should be the title of the dialogue box.
- The]] symbols to close the message.

The last dialogue box of a dialogue cannot have a reference to a next box. However, it should still contain an empty reference of "[[|]]" if one wants the dialogue to automatically end, or "[[<whatever text one wants the user interface to show>|]]" if one wants to give the user the option to press a final button to end the dialogue. The file should be saved in JSON format.

To convert the JSON file to a DGDL file, which can be read by DGEP, some work is needed. The explanation here assumes one has a fully functional version of the Council of Coaches system installed and set up. First, one needs to place the WOOL file in the ...\UDUN_ArgTech\dgep\WOOL\WOOLfiles folder. Then, one needs to open a command window, and in there navigate to the ...\UDUN_ArgTech folder in the folder containing the Council of Coaches system. Once here, one needs to run the command docker-compose up. Next, one must navigate to the ActiveMQ web page they can now open up at the web address http://localhost:8161/. Here, one needs to navigate to the Topics header and find the topic DGEP\requests. Here, the following JSON code should be put in:

{"cmd":"new","params":{"protocol":"WOOL.<WOOLFILENAME>", "username":"<USERNAME>"}}

An example would be:

{"cmd":"new","params":{"protocol":"WOOL.new_dialogue","username":"Patient"}}

<WOOLFILENAME> is the file in the folder, without the extension (e.g. if the filename is new_dialogue.json, <WOOLFILENAME> = new_dialogue). <USERNAME> whatever you want the user to be called inside the protocol, for example Patient, or User.

Make sure to not copy the JSON code directly from this file, as Word uses symbols DGEP does not recognize. One needs to type out the code. Once this is done, send the message. A DGDL file should now be created in the ...\UDUN_ArgTech\dgep\src\protocols folder. This file can be used by the system. It is advised to manually check the file for issues if the dialogue does not run as expected, as DGEP is quite particular about the dialogue format in the DGDL file.

9 Conclusion

In this document, we report the final prototype developed for the Council of Coaches project. The overall architecture is composed of four layers which include “Sense”, “Remember”, “Think” and “Act” layers. We also elaborate the integration of Greta with Unity3D platform which is done through the Thrift modules. These modules allow to communicate between Greta and Unity3D, and to synchronize Greta and Unity environment with each other. Other improvements in the Council of Coaches system include the supports dialogue on multiple devices; connection between sensors from the Holistic Behaviour Analysis Framework and the Knowledge base, that allow coaches to use information gathered by sensors in their dialogues.

We have presented an analysis of cohesion in multi-party interactions which focuses on non-verbal social cues and interruption during conversation. The results show that certain non-verbal social cues and interruptions and have an impact on level of cohesion. The results from this work will contribute towards developing a computational model to simulate a cohesive group of virtual agents.

We described two evaluation studies in the context of the council of coaches system. The study focuses on the evaluation for the final *Council of Coaches Technical Prototype* that focuses on the proof of concept. The results indicate that participants rated the agents high on likability and perceived intelligence, however, the gaze model for group interaction of the agents need to be developed.

The second study aims to evaluate the effect of inter-coach discussion during a persuasive dialogue in a coaching session the *project’s Functional Demonstrator*. The results show that the inter-coach discussion during a persuasive group dialogue leads to a change in perception of a virtual council of coaches and their ability to coach, in commitment to follow a chosen approach, and in enjoyment of, and preference for an interaction. However, the results were not significantly different.

We aim to perform further evaluation studies include the user interface usability evaluation, multi-device interaction, impact of verbal conflict presentation styles, gesture generation, and cohesive group evaluation, in the context of council of coaches system.

10 Bibliography

scikit-learn. (2018). *scikit-learn: Machine Learning in Python*. Retrieved from www.scikit-learn.org

Pandas. (2018). *Pandas: powerful Python data analysis toolkit*. Retrieved from pandas.pydata.org: https://pandas.pydata.org/pandas-docs/stable/

Activity Recognition API. (2018). Retrieved from Google: https://developers.google.com/location-context/activity-recognition/

Stackoverflow. (2018). *Scikit-learn: How to obtain True Positive, True Negative, False Positive and False Negative*. Retrieved from Scikit-learn: https://stackoverflow.com/questions/31324218/ scikit-learn-how-to-obtain-true-positive-true-negative-false-positive-and-fal

Labrador, O. D. (2013). A Survey on Human Activity Recognition Using Wearable Sensors. pp. 1192-1209.

Y. P. Chen, J. Y. (2008). "Onlineclassifierconstructionalgorithmfor human activity detection using a tri-axial accelerometer," Applied Mathematics and Computation, vol. 205, no. 2, pp. 849-860.

Tapia, E. M. (2008). "Using machine learning for real-time activity recognition and estimation of energy expenditure," Massachusetts Institute of Technology. Dept. of Architecture. Program in Media Arts and Sciences., Massachusetts Institute of Technology.

M. Shoaib, S. B. (2015). "A Survey of Online Activity Recognition Using Mobile Phones," Sensors, vol. 15, no. 1, p. 2059.

F. Attal, S. M. (2015). "Physical Human Activity Recognition Using Wearable Sensors," Sensors, vol. 15, no. 12, p. 29858.

scipy.stats.pearsonr. (2018). Retrieved from https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html

M. Gjoreski, H. G. (2016). "How Accurately Can Your Wrist Device Recognize Daily Activities and Detect Falls?," Sensors (Basel), vol. 16, no. 6.

scikit-learn. (2018). *1.13. Feature selection*. Retrieved from http://scikit-learn.org/stable/modules/feature_selection.html

Zeng, Z., & Pantic, M. R. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. . *IEEE transactions on pattern analysis and machine intelligence*, 31(1), pp.39-58.

Ekman., P. (1997). What we have learned by measuring facial behavior.,. In *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)* (pp. 469–485).

Kroschel., M. G. (2005). Evaluation of natural emotions using self assessment manikins. *IEEE Workshop on In Automatic Speech Recognition and Understanding* (pp. 381–385). IEEE.

Glas, N. a. (2015). Definitions of engagement in human-agent interaction. *International Conference on Affective Computing and Intelligent Interaction*, (pp. 944–949).

Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication.*

Sidner, C. L. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

Peters, C. P. (2005). Engagement Capabilities for ECAs. *AAMAS'05 workshop Creating Bonds with ECAs.*

Yu, L. L. (2016). Building Chinese Affective Resources in Valence-Arousal Dimensions. . *HLT-NAACL.*

P. Ekman, W. V. (2002). *Facial action coding system.* Salt Lake City.

Baltrušaitis T., M. M. (2015). Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. *IEEE International Conference on Automatic Face and Facial Expression Recognition and Analysis Challenge, , IEEE International Conference on Automatic Face and Gesture Recognition.*

Camurri, A. C. (2004). Toward real-time multimodal processing: EyesWeb 4.0. *Proceedings of the artificial intelligence and the simulation of behaviour (AISB).*

Ringeval F., S. A. (2013). Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013).* IEEE.

Cafaro, A. W. (2017). The NoXi database: multimodal recordings of mediated novice-expert interactions. *19th ACM International Conference on Multimodal interaction.*

Corrigan, L. J. (2016). Engagement perception and generation for social robots and virtual agents. *Toward Robotic Socially Believable Behaving Systems*, 29-51.

Sidner, C. L. (2005). A first experiment in engagement for human-robot interaction in hosting activities. *Advances in natural multimodal dialogue systems*, 55-76.

Dermouche S., ,. P. (2018). From analysis to modeling of engagement as sequences of multimodal behaviors. *LREC.*

Bartneck, C. (2008, March 11). *http://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/*. Retrieved from http://www.bartneck.de: http://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/

Côte, J., Yardley, J., Hay, J., Sedgwick, W., & Baker, J. (1999). An exploratory examination of the coaching behavior scale for sport. *Avante Research Note, 5*(2), 82-92.

Richmond, V. P., McCroskey, J. C., & Payne, S. K. (1991). *Nonverbal behavior in interpersonal relations.* Prentice Hall Englewood Cliffs, NJ.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, 1743--1759.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of personality and social psychology*, 342.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., . . . others. (2005). The AMI meeting corpus: A pre-announcement. *International workshop on machine learning for multimodal interaction*, 28--39.

Hung, H., & Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 563--575.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, (pp. 1--10).

Glenn, P. (2003). *Laughter in interaction.* Cambridge University Press.

Tian, Y.-I., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 97--115.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). he timing of shifts of head postures during conservation. *Human Movement Science*, 237--245.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, 22--63.

Exline, R. V. (1963). Explorations in the process of person perception: Visual interaction in relation to competition, sex, and need for affiliation. *Journal of personality*.

Dohsaka, K., Asai, R., Higashinaka, R., Minami, Y., & Maeda, E. (2009). Effects of conversational agents on human communication in thought-evoking multi-party dialogues. *Proceedings of the {SIGDIAL} 2009 Conference* (pp. 217-224). London, UK: Association for Computational Linguistics.

Bohus, D. a. (2011). Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions. *Proceedings of the SIGDIAL 2011 Conference,*, (pp. 98--109).

Heldner, M., & Jens, E. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 555 - 568.

Tannen, D. (1994). *Gender and discourse*. Oxford University Press.

West, C., & Zimmerman, D. H. (2015). mall insults: A study of interruptions in cross-sex conversations between unacquainted persons. *American Sociological Association's Annual Meetings, Sep, 1978, San Francisco, CA, US*.

Pontecorvo, C., Pirchio, S., & Sterponi, L. (2000). Are there just two people in a dyad? Dyadic configurations in multiparty family conversations. *Schweizerische Zeitschrift fûr Bildungswissenschaften*, 535--558.

Li, H. Z. (2001). Cooperative and intrusive interruptions in inter-and intracultural dyadic discourse. *Journal of Language and Social Psychology*, 259--284.

Bangerter, A., Chevalley, E., & Derouwaux, S. (2010). Managing third-party interruptions in conversations: Effects of duration and conversational role. *Journal of Language and Social Psychology*, 235--244.

Cafaro, A., Ravenet, a. B., & Pelachaud, C. (2019). Exploiting evolutionary algorithms to model nonverbal reactions to conversational interruptions in user-agent interactions. *IEEE Transactions on Affective Computing*, 1-1.

Kantharaju, R., De Franco, D., Pease, A., & Pelachaud, C. (2018). Is Two Better than One?: Effects of Multiple Agents on User Persuasion. *Proc. of the 18th International Conference on Intelligent Virtual Agents* (pp. 255--262). ACM.

Ravenet, B., Cafaro, A., Biancardi, B., Ochs, M., & Pelachaud, C. (2015). Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. *Proc. International Conference on Intelligent Virtual Agents* (pp. 375--388). Springer.

Casey-Campbell, M., & Martens, M. (2009). Sticking it all together: A critical assessment of the group cohesion--performance literature. *International Journal of Management Reviews*, 223--246.

van Waterschoot, J., Bruijnes, M., Flokstra, J., Reidsma, D., Davison, D., Theune, M., & Heylen, D. (2018). Flipper 2.0: A Pragmatic Dialogue Engine for Embodied Conversational Agents. *In Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 43-50). ACM.

Acknowledgements

The Council of Coaches project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Headings and titles in this document, as well as the Council of Coaches logo use the Comfortaa font, designed by Johan Aakerlund and Cyreal and licensed under the Open Font License[1].

Additional text in this document uses the Roboto font, designed by Christian Robertson and licensed under the Apache License, Version 2.0[2].

The Council of Coaches logo and Blobmen graphics were *drawn freely* in Inkscape, licensed under the GNU General Public License[3].

[1] Open Font License: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=OFL_web
[2] Apache License, Version 2.0: http://www.apache.org/licenses/LICENSE-2.0
[3] Inkscape License Information: https://inkscape.org/about/license/