



D5.5: Methods for evaluating the Dialogue and Argumentation Framework

Dissemination level: Public

Document type: Report

Version: 1.0.0

Date: March 2, 2020



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	D5.5: Methods for evaluating the Dialogue and Argumentation Framework
Due date of deliverable	February 29, 2020
Work package	WP5
Author(s)	Mark Snaith (UDun), Tessa Beinema (RRD), Harm op den Akker (RRD), Alison Pease (UDun)
Reviewer(s)	Daniel Davison (CMC), Harm op den Akker (RRD)
Approved by	Coordinator
Dissemination level	Public
Document type	Report
Total number of pages	17

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, Grupa SABIEN (UPV)
- Innovation Sprint (iSPRINT)

Abstract

This document describes two different methods in which the Dialogue and Argumentation Framework will be evaluated, one focusing on recreating the scripted dialogues used in the Council of Coaches evaluation, and one focused on an expert-based evaluation.

Table of Contents

1	Introduction.....	5
2	Objectives	6
3	Background: Scripted dialogues as input	7
4	From scripted dialogues to dialogue games	10
4.1	WOOL dialogue creation	10
4.2	Dialogue structure extraction	10
4.3	Dialogue structure analysis.....	11
4.4	Implementation of dialogue game and content	11
5	Evaluation Method 1: Recreating concrete dialogues	12
5.1	Description	12
5.2	Process	12
5.2.1	Dialogue game execution	12
5.2.2	Evaluation	12
6	Evaluation Method 2: Expert-based evaluation	13
6.1	Description.....	13
6.2	Process	13
6.2.1	Dialogue execution	13
6.2.2	Expert evaluation.....	13
7	Bibliography	15

List of figures

Figure 1: Expert evaluation to record different evaluation metrics during different stages of a conversation.....	14
---	----

Symbols, abbreviations and acronyms

CMC	Centre for Monitoring and Coaching
COUCH	Council of Coaches
D	Deliverable
DAF	Dialogue and Argumentation Framework
DBT	Danish Board of Technology Foundation
EC	European Commission
ISPRINT	Innovation Sprint
M	Month
MS	Milestone
RRD	Roessingh Research and Development
SU	Sorbonne University
UDun	University of Dundee
UPV	Universitat Politècnica de València
UT	University of Twente
WP	Work Package

1 Introduction

The Dialogue and Argumentation Framework provides the core logic component to the Council of Coaches technical demonstrator and the Open Agent Platform, allowing for structured dialogue between the coaches and the user and the coaches themselves. It is important that these dialogues are realistic, both in terms of user interaction and the outcomes that they lead to.

2 Objectives

The objectives of this deliverable are to:

1. Provide a scenario and associated shared knowledge base that will form input for the evaluations to be reported in D5.6.
2. Provide methods for the evaluation of the Dialogue and Argumentation Framework.

3 Background: Scripted dialogues as input

The dialogues that will result from the Dialogue and Argumentation Framework will be the main form of interaction between the user and the Council of Coaches in the project's Technical Demonstrator and the resulting Open Agent Platform. The setting in which these interactions take place are the coaching sessions that the user has with their coaches.

As described in more detail in Deliverable 3.4 (Beinema, op den Akker, Kosterink, ter Stal, & van den Boer, 2019), the Council of Coaches features six coaches (as well as a peer agent and an assistant agent). Each of these coaches in the council has their own domain of expertise (physical activity, nutrition, cognition, social, diabetes type 2 and chronic pain).

In the coaching sessions that take place between the user and their coaches, multiple topics of conversation can be discussed with each coach (again, see D3.4 for elaboration). These topics can, for example, be an introduction between the coach and the user (when they meet for the first time) or the coach and the user who discuss what the user's long-term physical activity goal will be.

Consider the following example, where the user meets their coach François for the first time (in the example, the dialogue replies chosen by the user are **highlighted in teal**):

Example 1: User Bob meets nutrition coach François.

François: Bonjour! Allow me to introduce myself... I am... François Dubois! Your personal Chef and diet coach!

Bob: **Hello Francois.**

François: Hello to you Bob! Very nice to have you with us. Let me tell a little bit about moi.

François: So, like I said, my name is François and I'm here to help you out with any questions you have about food and diet!

François: Oh, François loves to talk about food. Eating food and cooking food! What about you, do you like to cook?

Bob: **Yes, I like cooking.** | I don't really like cooking.

François: I see you and me will get along just fine!

François: Or just for a chat of course, I always love to chat about... food, cheese, recipes... hmmm, anything really!

Bob: **Okay, thank you François!**

François: Adieu! Talk to you later.

This introductory conversation between François and the user is a dialogue in which we can highlight several aspects. First, the conversation in this case is between François and the user, but if we observe introductions between the other coaches and the user, there tends to be a common pattern to the conversation structure (these are the types of things you say and these are the responses that make sense). Second, there is a clear use of knowledge about the user. That is, François calls the user 'Bob'. This is information that was provided by the user to the assistant agent during the intake and stored, and which was later conveyed to the other coaches by the assistant saying 'Coaches, this is Bob' (which is a move meant to let the user know that the other coaches can also access this information). Third, in this dialogue we have an example of the coach asking the user a question and receiving an answer that can be stored as knowledge about the user (namely, they like cooking).

For another, slightly longer example on the topic of setting a long-term physical activity goal, consider the following (again the dialogue replies chosen by the user are **highlighted in teal**):

Example 2: Physical activity coach Olivia wants to set a long-term goal with the user (Bob).

Olivia: Setting a goal for yourself is the first step towards becoming more active. Actually, did you know that people who commit to a goal are more likely to become more active?

Olivia: You can set goals for the upcoming days or weeks. That's what we call "short-term" goals, or for the upcoming months. That's what we call "long-term" goals.

Olivia: So, let's talk about your long-term goal...

Bob: Do you have a suggestion? | *Why do I need a long-term goal?* | Can we talk about something else?

Olivia: Why? Well, there are two good reasons for setting a long-term goal for yourself.

Olivia: First, it can be a great motivation for yourself. It gives you something to work towards.

Olivia: So, your long-term goal should be ambitious! Something that will make you feel really proud once you achieve it.

Olivia: And the second reason is that it gives us something to cut into little pieces, something to tackle day-by-day.

Olivia: So having a long-term goal makes it possible to set relevant and personal daily goals. And those daily goals give us a concrete objective for every day!

Bob: *Okay, so can you suggest a goal?* | Can we talk about something else?

Olivia: First you should decide what kind of goal you want to set. We can look at the number of steps taken or the minutes of higher activity per week or per day.

Olivia: I can tell you about each of the options, or let me know if you're ready to decide.

Bob: Steps taken? | *Active minutes?* | I am ready to decide.

Olivia: Although every little bit of activity helps, it is also important to do things that require a little more effort. Basically, this is when you feel your heart beat going faster, maybe you become sweaty, and you feel tired at the end of the activity.

Olivia: Your activity tracker can see how much time you spent doing a little extra, which we call "active minutes".

Olivia: So, would you like to set a goal for active minutes?

Bob: *I'm ready to decide.* | What about steps?

Olivia: That's good to hear. What will it be?

Bob: Steps. | *Active minutes.* | Actually, could you explain the options again?

Olivia: Good! I will help you work towards a long-term goal for active minutes then!

Olivia: The next step is to decide how many minutes would be a challenging but achievable goal for you in the long-term.

Olivia: The World Health Organization recommends that people aged above 65 years old participate in at least 2.5 hours of moderate to high intensity physical activity per week. Think of 3 sessions a bit shorter than 1 hour per week in which your heart rate is elevated. What do you think about this goal?

Bob: *It sounds good, let's do that!* | That sounds too easy for me. Could you suggest a higher goal? | *It might be a bit high for me. Could you please suggest a lower goal?*

Olivia: Ok, then I would like to suggest a long-term goal of 120 minutes of moderate to high physical activity intensity. Remember that we are discussing the long-term goal, not your goal for next week. Is that ok?

Bob: I still think that is too high. | Oh, no, I think that is too low. | **Yes, let's do that.**

Olivia: That is great! So just to make sure, we are working towards 120 minutes of moderate to high intensity physical activity per day.

Emma: Group activities are a great way to motivate you to reach your physical activity goal! And you might meet some interesting people as well!

Olivia: Yes Emma, that's true, and if you need any other tips on how to become more active you know where to find me.

Olivia: So, is there anything else I can help you with?

Bob: I want to know more about goals. | I want to discuss something else. | **Goodbye.**

Again, as with the previous example, this type of conversation has some typical statements and responses to it and there are a few examples of knowledge use (e.g. the initially suggested goal in this case is adjusted to the fact that the user is 65 years or older) and a number of examples of knowledge storage. In this conversation however, the example above is just one of the possible resulting dialogues. That is, there is a whole set of statements and replies that discuss setting a long-term step goal (instead of the active minutes goal that is shown), but these are never visited. The structure for the part of the dialogue discussing steps is almost the same as the structure for active minutes, but when scripting a dialogue in the traditional manner every different statement and its replies needs to be defined separately (and linked to other statements). Furthermore, the dialogue shows some examples of knowledge in the domain of physical activity that is implicitly embedded in the statements (e.g. 150 active minutes is a healthy goal for our user who is 65 years or older).

As stated previously, each of the possible topics in a coaching conversation has certain dialogues associated with it. In the Functional Demonstrator these dialogues are fully scripted. These scripts are written to cover a range of coaching scenarios (as elaborated on in D3.4). They also serve as examples that can be used when determining the underlying structure of coaching dialogues on various topics. Furthermore, they implicitly represent knowledge – each statement is a representation of facts that a domain expert put in – and explicitly represent knowledge, by asking the user questions and storing their responses.

In the Technical Demonstrator/Open Agent Platform, the scripted dialogues representing different scenarios and domain knowledge serve as an input for the development of dynamic conversations between a user and their coaches by the Dialogue and Argumentation Framework. The process of turning these dialogue scripts into dialogue games and extracting content to instantiate the moves in those dialogue games is described in the next section.

4 From scripted dialogues to dialogue games

An important feature of the dialogue and argumentation framework is its ability to permit dialogues that accurately reflect the structure, content and outcomes that those dialogues might contain in real life. We therefore, in this section, present a pipeline that takes as input dialogues created using the WOOL framework, and produces as outputs 1) an abstract dialogue game, expressed in Dialogue Game Description Language (DGDL); and 2) content that can be used to create concrete dialogue moves in those dialogue games. These two outputs will then form the input to the two evaluation methods proposed in Section 5 and Section 6.

4.1 WOOL dialogue creation

The first step in the dialogue generation pipe-line is to create scripted dialogues that represent (part of) a natural conversation on a certain topic (e.g. long-term goal-setting for physical activity). The WOOL dialogue framework¹ (Roessingh Research and Development, 2020) allows domain experts to write and test WOOL dialogue scripts without needing to learn a complicated programming/scripting language.

The dialogue creation process is broadly executed in five phases:

1. First, a technical expert and domain expert discuss which content to include in the dialogue, based on the selected domain and topic.
2. Second, the domain expert will use the WOOL editor to write a first version of that dialogue. They will define dialogue steps that represent what the agent says, how the user can respond, and how these steps are connected. While doing so, they use the built-in test-functionality to test the dialogue and to get a feel for the result. Alternating editing and testing, they will iteratively add to their dialogue until they feel it is done.
3. Third, when the domain expert has completed a dialogue, this dialogue will be checked by a dialogue expert. This expert will ensure the resulting file is structurally sound, i.e. that the variables that are used by the author are present in the Council of Coaches system and that the dialogue flow has no errors in conditionals and node order (e.g. authors might create conditionals that are syntactically correct, but when executed show incorrect parts of the dialogue). The dialogue expert will also add in variables that are relevant for using knowledge or storing knowledge in the system (adding to variables that the domain expert might have already included). The information stored through these variables can involve any type of information, ranging from 'this dialogue was completed' to 'the user has answered that they accept a certain step goal'.
4. Fourth, the completed dialogue undergoes a final check to ensure that the flow for the completed dialogue is realistic and natural. That is, the order and content of dialogue steps that are shown should make sense in terms of available and expected information for the reply that the user selects. E.g. if the user selects 'Okay, so can you suggest a long-term goal?' they probably expect the coach to respond to that statement with a 'yes'/'no' and an actual suggestion. The final check is therefore performed to ensure that there are no unexpected or too far-fetched jumps in content.
5. Finally, the completed WOOL dialogue can be fully translated to other natural languages (e.g., Danish, Dutch, or English).

4.2 Dialogue structure extraction

Following the creation of WOOL dialogues, the next stage is to extract from the WOOL scripts a structure that can be analysed in order to determine the abstract structure. This extraction is necessary to clearly see every possible path through the dialogue in terms of alternative utterances – the WOOL editor simplifies this structure by capturing several possible moves, and several possible alternative next moves, in one node.

¹ For more background on WOOL also see Part B of Deliverable 3.4 (Beinema, op den Akker, Kosterink, ter Stal, & van den Boer, 2019).

Extraction of this structure will involve expanding a WOOL representation of a dialogue into a directed graph with each utterance represented by a unique node, and transitions (i.e. a move following another move) represented by the edges.

4.3 Dialogue structure analysis

This step will manually annotate the extracted dialogue structure to determine sequences of abstract move types. An abstract move type might consist of a single utterance (e.g. “Shall we set a goal?” is a *question*) type or multiple utterances (e.g. “There are several reasons you should set a goal...” is an *explain* type).

It is not envisaged that this step will be automatic because there is no simple method of determining what utterances are of the same type. While in principle it is possible to use basic features (e.g. the presence of a question mark indicates a *question* type), not all utterances have such features. As such, for the purposes of this evaluation, the annotation will be manual but with a view to developing an automated approach in the future.

Following the annotation of utterances, patterns will be identified to unearth the protocol encoded within. For instance, if every *question* utterance is followed by a (series of) *explain* utterances, that represents a rule in the protocol. This protocol will be then be expressed as a dialogue game, in the same way as was reported in D5.2.

4.4 Implementation of dialogue game and content

Once a protocol has been identified, it will be implemented in Dialogue Game Description Language. This implementation will take the same form as was reported in D5.3, where the dialogue games from D5.2 were implemented.

The content extracted from the WOOL representation of the dialogue will be fed into the Utterance Generator (see D5.4) for populating dialogue moves during execution.

5 Evaluation Method 1: Recreating concrete dialogues

5.1 Description

The aim of this evaluation method is to use the dialogue games and content produced in Section 4 to first reproduce the input scripted dialogue(s), then generate new dialogues. This process will contain a mixture of both fully-autonomous dialogues (between coaches), and dialogues that involve the user.

5.2 Process

5.2.1 Dialogue game execution

The first stage in this method is to execute the dialogue game produced in Section 4 using the Dialogue Game Execution Platform (DGEP). This execution will take place twice in two different contexts.

The first context is to fully manually simulate a dialogue using the Dialogue and Argumentation Framework demonstrator reported in D5.4. This approach places the choice of all dialogue moves (i.e. those for both the coach(es) and the user) entirely in the hands of the researchers. This will allow us to determine if it is at least *theoretically* possible to recreate the original input dialogue based on the available moves and content at any given moment, removing any unpredictability associated with (semi-) autonomous move selection when the DAF presents two or more alternatives. This evaluation will require knowledge of the input dialogue, with the researcher selecting moves based on it; if an expected move is not available at a given time, this will represent a failure in the evaluation.

The second context is to run repeated executions of the dialogue game within the full Council of Coaches technical prototype, thus allowing the coaches to determine the choice of move based on the alternatives provided by the DAF. This will allow us to examine different ways in which the dialogue unfolds and evaluate them using the metrics described in below Section 5.2.2.

5.2.2 Evaluation

Following the second execution context, the outputs from the resultant dialogues will be evaluated across a range of metrics, including but not necessarily limited to:

- **Soundness:** are the dialogues structurally sound with respect to the original scripted dialogue (i.e. do the DGEP-driven dialogues all follow the same structure)?
- **Correctness:** do the dialogues provide the same outcome(s) (e.g. agreed goal) predicted by the scripted dialogue?
- **Completeness:** do all dialogues terminate successfully? (Where “success” is defined in structural terms, i.e. reaching an identified termination move)

It is not the aim of this method to measure how “natural” or “realistic” the dialogues are; that is left to the second method, described below in Section 6.

6 Evaluation Method 2: Expert-based evaluation

6.1 Description

This method will ask a panel of experts to evaluate DGEP-driven dialogues along several dimensions including, but not limited to, flow, naturalness, and usefulness.

6.2 Process

The user-based evaluation will have two broad stages: dialogue execution, and the user evaluation itself.

6.2.1 Dialogue execution

This stage will involve executing autonomous dialogues between two or more agents and recording the resultant transcripts. These dialogues will come from Method 1 described in Section 5 above.

6.2.2 Expert evaluation

The expert evaluation stage will take the output from the dialogue execution and present it to an expert in dialogue and argumentation, using an interface modelled on the one used in the functional demonstrator. Experts will be instructed that they are watching a coaching session of a third-party user (rather than experiencing a live coaching session with themselves as user). The rationale for replaying the dialogues, instead of having participants directly observe them as they are executed is to avoid any unexpected communication issues that might occur between the interface and the dialogue and argumentation framework itself. Using transcripts allows the evaluation setup to stay self-contained.

While observing the dialogues, the expert participants will be asked to provide feedback based on established principles for evaluating conversational agents (e.g., (Radziwill & Benton, 2017)).

Evaluation will take place using a matrix such as the one shown below in Figure 1, in which the columns represent different aspects of the discussion, and the rows represent different metrics along which dialogues (both humans and hybrid, such as with conversational agents) are typically evaluated. Each cell will be completed with a number from a Likert scale: in addition, experts will be encouraged to comment on metrics of interest, using free text. Dialogical aspects might include chronological aspects such as pre-dialogue, opening, coaching, agreement, closing and post-dialogue, or organisational aspects such as turn-taking or interruptions, etc. Evaluation metrics will include aspects such as Flow, Naturalness, Effectiveness, Usefulness, Sensibleness and Specificity.² This will be performed by a panel of 3-5 people who are experts in dialogue.

² For the latter two, see <https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html>

		Aspects of a Dialogue				
		Opening	Coaching	Closing	Turn-taking	Interruptions
Evaluation Metrics	Flow					
	Natural					
	Effective					
	Useful					
	Sensible					
	Specific					

Figure 1: Expert evaluation to record different evaluation metrics during different stages of a conversation.

We expect that the evaluation will proceed along the following six stages:

1. Assemble a panel of 3-5 experts in dialogue (all meet together at the same time). These have not been involved in the Council of Coaches project.
2. Show them a preliminary matrix of evaluation (for instance, as shown above in Figure 1) and discuss together. Changes can be made at this stage to the matrix.
3. Once the matrix is finalised, give each expert a set of real-world dialogues to evaluate (selected from our video corpus). This will help us to calibrate each expert's score on the generated dialogues. It will not be used as a discrimination test (we will be open that these are human dialogues).
4. Once calibration has occurred, we will ask each expert to independently complete the evaluation matrix for a selection of our generated dialogues.
5. The focus group will re-convene and discuss their assessments and in particular, their in-depth qualitative data regarding the various metrics. This will allow us to report on the expert's views on aspects of our system without directly comparing it with the baseline.
6. Analysis of both (human and generated dialogues) evaluation matrices and discussion of performance.

7 Bibliography

Beinema, T., op den Akker, H., Kosterink, S., ter Stal, S., & van den Boer, J. (2019). *D3.4: Final coaching actions and content*. The Council of Coaches Consortium.

Radziwill, N., & Benton, M. (2017). *Evaluating Quality of Chatbots and Intelligent Conversational Agents*. Retrieved from arXiv:1704.04579v1

Roessingh Research and Development. (2020, Februari). *The WOOL Dialogue Framework*. Retrieved from WOOL: <http://www.woolplatform.eu/>

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 License³.

Additional text in this document uses the Roboto font, designed by Christian Robertson and licensed under the Apache License, Version 2.0⁴.

The Council of Coaches logo and Blobmen graphics were *drawn freely* in Inkscape, licensed under the GNU General Public License⁵.

³ Open Font License: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=OFL_web

⁴ Apache License, Version 2.0: <http://www.apache.org/licenses/LICENSE-2.0>

⁵ Inkscape License Information: <https://inkscape.org/about/license/>