Paper to appear in the 2016 conference proceedings of the 21st International Command and Control Research and Technology Symposium.
Link: http://www.dodccrp-test.org/2016-info-central/

Human Computer Collaboration at the Edge: Enhancing Collective Situation Understanding with Controlled Natural Language

Alun Preece*, William Webberley*, Dave Braines†, Erin G. Zaroukian‡, and Jonathan Z. Bakdash‡
*School of Computer Science and Informatics, Cardiff University, Cardiff, UK; contact email: PreeceAD@cardiff.ac.uk
†Emerging Technology Services, IBM United Kingdom Ltd, Hursley Park, Winchester, UK
‡US Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, USA

***Abstract*—Effective coalition operations require support for dynamic information gathering, processing, and sharing at the network edge for Collective Situation Understanding (CSU). To enhance CSU and leverage the combined strengths of humans and machines, we propose a conversational interface using Controlled Natural Language (CNL), which is both human readable and machine processable, for shared information representation. We hypothesize that this approach facilitates rapid CSU when assembled dynamically with machine assistance, via social sensing, from local observations, with information rapidly disseminated among people at the network edge. We report a behavioural experiment wherein small groups of users attempted to build CSU via social sensing, interacting with the machine via Natural Language (NL) and CNL. To simulate a tactical environment, participants answered 36 questions (operationalized as CSU) by visiting various locations and describing their discoveries to a mobile conversational agent. To test our hypothesis, we compared the performance of groups of users between the:**

1) **Online Condition: CSU, the status of all questions, dynamically updated by the machine as users collect information.**
2) **Offline Condition: No dynamic machine-supported CSU, simulating unreliable connectivity at the edge. Each participant was restricted to their own information until the end of the experiment.**

Results indicated the Online Condition had greater agreement in CSU, but individual participants answered significantly fewer questions than the Offline Condition. In other words, the Offline Condition group provided more answers, but there was more consistency among the answers provided by the Online Condition group.

***Index Terms*—collective situation understanding; controlled natural language; conversational interface; human-machine interaction, groups**

I. Introduction

Over the past four years we have been researching technologies to support human-machine collaboration in the context of coalition intelligence, surveillance, and reconnaissance (ISR) tasks [1]. We have focused on approaches using Controlled Natural Language (CNL) [2] to provide representations of information and knowledge that are human readable and writable, as well as machine-processable. Such approaches allow the machine to perform computational reasoning over a knowledge base while expressing rationale that is human-understandable. The overall goal of this research is to use human-machine collaboration, also called Human Computer Collaboration (HCC), to enhance human cognition.

Our recent focus has been on behavioural research to test the effectiveness of the technology for people using simulated ISR tasks, specifically in the context of Collective Situation Understanding (CSU). To this end, we have designed a platform for running a series of experiments in which human participants work alone or collectively on gathering synthetic and natural information either in situ or online. Participants address CSU tasks by interacting with a CNL agent through dialogues in which they can use both Natural Langauge (NL) and CNL. The tasks involve the collection of locally observed information, such as simulating activities humans would perform on patrol or while operating a remote sensing system.

Tasks in these experiments are simplified to allow participation without any specific ISR training and to ease some aspects of the NLP performed by the agent. This was done because NLP was not the focus of the current work. NLP is "...computational techniques for the automatic analysis and representation of human language" [3, p.48]. In prior research, we have demonstrated that with minimal training most people could effectively use variants of the conversational interface in simulated tactical intelligence tasks with imagery [4] and in a real-world environment [5]. We hypothesize that the approach with the conversational agent facilitates rapid CSU when assembled dynamically with machine assistance, via social sensing, from local observations, with information rapidly disseminated among people at the network edge.

In this paper, we report a behavioural experiment wherein small groups of users attempted to build CSU via social sensing, interacting with the machine via NL and CNL. To simulate a tactical environment, participants answered 36 questions (operationalized as CSU) by visiting posters hung in various locations in a building complex and describing their discoveries to a mobile conversational agent. To test our hypothesis, we compared the performance of groups of users between the:

1) Online Condition: CSU, the status of all questions, was dynamically updated by the machine as users collected information. This information was presented in a dashboard, which served as a real-time common operating picture (COP). A COP is a unified display of relevant information shared shared by one or more users.
2) Offline Condition: No dynamic machine-supported CSU,

simulating unreliable connectivity at the edge. Each participant was restricted to their own information until the end of the experiment.

The experiment uses novel technology for CSU: a decentralised platform for knowledge capture and sharing based on NL and CNL, called CENode (Controlled English Node)[1]. Being decentralised makes CENode robust in settings where network connectivity is limited or unreliable: users are able to work offline with local knowledge bases, sharing information when connectivity is available. While the primary goal of the work was to compare the CSU performance of human teams in the Online and Offline Conditions, a secondary goal was to test the effectiveness of the CENode software in enabling CSU at the network edge, which we refer to as *Edge CSU*.

This paper is structured as follows: Section II establishes the context of our research in terms of related work; Section III introduces the CENode platform and the capabilities of the CNL agent built using CENode for the experiment; Section IV details the experiment design; Section V provides analysis of the results; finally, Section VI concludes the paper and points to future work.

II. Related Work

Networked Enabled Operations allow rapid information sharing and communication, making it possible to have decentralized or edge groups, teams, and organisations rather than hierarchical ones [6], [7]. Edge Command and Control (C2) is characterised by enriched peer-to-peer interactions such as horizontal exchanges and interactions with peer contributing partners in a coalition, where the resulting increases in information sharing improve the quality and accessibility of available intelligence [8].

Edge C2 has well-known limits and even detriments to group performance. For example, as group size increases the potential for loss of motivation increases [9]. Also, the number of communication links in a fully connected network increases exponentially as a function of group size: $n(n-1)/2$ where n = number of individuals [10], making complete connectivity difficult and expensive to maintain as networks grow. Finally, more information, even if task relevant, can impair human decision-making [11]. A key motivation for human-computer collaboration in this context is to ameliorate these negative effects, where machine affordances in data manipulation can reduce cognitive burdens on humans [12]. Consequently, we sought to minimize human efforts for lower levels of information fusion (i.e., information pre-processing and refinement) so users could focus on high-level inferences to improve effectiveness [13]. In our work, the group members are assisted in CSU tasks by agents performing information fusion and simple visualisations that indicate where information is currently lacking, while also managing communication in order to avoid the 'cost' of links as group size grows.

Understanding phenomena in CSU requires multiple levels of analysis [14]. For example, groups comprise a lower level of analysis: individuals. Knowing how individuals behave does not necessarily fully inform how the group behaves and vice-versa [15]. Consequently, we examined group performance for information quality using CSU and information quantity using total messages. Individual performance for information quantity was assessed using the quantity of messages each person in each group added to the knowledge base.

Our conversational approach using NL and CNL is intended support HCC where natural communication, shared representation and manipulation of knowledge and problem-solving entities, and balanced representation and reasoning between human and machine are key principles [16].

III. Approach to Supporting Edge CSU

Our approach to supporting CSU at the network edge is founded on the use of a CNL as a means to define information models as well as structured instance data. Model and instance elements collectively form a knowledge base. The CNL used in this work is International Technology Alliance (ITA) Controlled English (CE) [17], which offers approximately the same expressivity in terms of information modelling as the Web Ontology Language (OWL) [18]. Model elements and instance data are defined via CE sentences. For example, the first sentence below defines the model concept `character` as a child concept of the parent concept `locatable thing`. This definition allows instances of `character` to inherit a relationship `is in` that associates instances of `locatable thing` with instances of the `location` concept. The second example sentence below is a piece of instance data asserting that a specific instance of `character` (named 'Dr Finch') is associated with a specific instance of `location` (named 'Gold Room').

```
conceptualise a ~ character ~ C that is a
locatable thing.

the character 'Dr Finch' is in the location
'Gold Room'.
```

CENode is a lightweight CE processing environment implemented in JavaScript so as to be easily deployable in a variety of contexts, including web browsers, mobile apps, and servers[2]. CENode is lightweight in the sense that it does not aim to be a fully fledged CE engine — for example, offering only limited inference and NL processing — and requires relatively little network bandwidth to download and operate. Once loaded, a CENode instance can function independently without any network connection, maintaining a local knowledge base (KB) and communicating with other CENode instances only when connectivity is available, via the CE Card conversational protocol [19] and blackboard mechanism. This makes it well-suited to deployments at the network edge, and in settings where a centralised client-server model is not the most appropriate configuration. (In centralised settings, the CE Store [20] offers a far richer set of CE knowledge representation and reasoning capabilities.)

[1] http://cenode.io

[2] For example, via Node.js: https://nodejs.org

CENode instances can either be run independently or as part of a multi-node system. In a multi-node system, at least one of the nodes needs to be run as a service (e.g., via Node.js). All CENode instances in a multi-node system are, by default, equal in terms of functionality and behaviour. This is the case even if each node is deployed in a different way (e.g., some nodes may be running as a service, some as a web application, and some as a programmatic JavaScript application). Providing information to (and retrieving information from) a node is always done via CE. Using CE as the only means of communication enables support for distributed systems including humans, CENode agents, and a CE Store.

CENode is intended to offer a number of key benefits in an edge CSU setting:

- Users have access to, and can interact with, a CENode agent directly on their device. Any CE provided to the agent can be parsed locally and any local knowledge stored can later be relayed ('told') to other agents once a network connection is (re)established.
- Because the local node is a CE processing environment, features such as CE 'autocorrect' and 'spellchecking' can be provided at no bandwidth cost and in the absence of a network connection. The local agent can quickly check the validity of any CE as it is being typed in order to guide the user towards inputting correct CE and also giving insight into the concepts and instances stored in the local CE model.
- Local NL processing of input means that only validated CE is transmitted between nodes, at a saving of bandwidth and time.
- Instead of relying on a single CE Store server with a centralised knowledge base, CENode supports a network of peers with different local knowledge base variants. This is particularly important in a coalition context where different partners may hold different knowledge.

The CNL grammar understood by CENode has been extended from standard ITA CE, supporting various 'shorthands' for easier input and querying of information to and from the KB. Input made in this way (as with the NL processing) can be guided by the node's own KB, and predictions for intended sentences can be provided. Whilst not standard CE, CENode's understanding of the grammar means that the following types of sentences can safely be sent within CE cards to a CENode agent. For example, the CE instance sentence above can more concisely be written:

```
Dr Finch is in the location 'Gold Room'.
```

Another useful 'shorthand' is the ability to ask questions to provide users and agents with the ability to make who/what/where queries of the node's KB. As well as supporting questions such as 'What is a character?', 'What is the Gold Room?', and 'Where is Dr Finch?' the interface can be used to query about relationships and properties. For example, the query `What is 'is in'?`
results in the response:

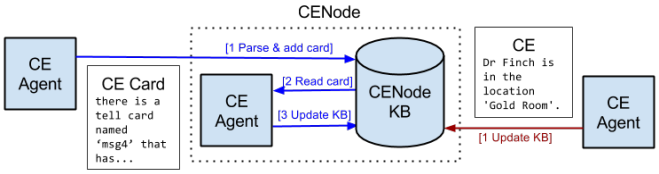


Fig. 1. Manipulating a node's KB - *left: through cards; right: directly.*

```
'is in' describes the relationship between a
locatable thing and a location.
```

A. Manipulating a CENode Knowledge Base

Each CENode comprises a KB and a local *CE agent* that maintains the KB, shown in Figure 1. A CENode will try to process and update its KB when any CE is received. As with the CE Store, CENode instances also support the blackboard architecture, which enables users and agents to submit CE sentences wrapped in CE Cards that are addressed to the local agent. If a card addressed to the agent is received, then the agent can find the card and read it. If the card contains valid CE, then the agent can then use this to modify its KB. If the card is not addressed to the local agent, then it will remain in the KB unread. The addressee node may eventually find this card as a result of policies (see Section III-B).

For example (illustrated by Figure 1), assuming a node's local agent is named `agent1`, the following two sentences received by the node would have equal effect:

```
there is a tell card named 'msg4' that is to
the agent 'agent1' and has "Dr Finch is in
the location 'Gold Room'." as content.

Dr Finch is in the location 'Gold Room'.
```

The CE agents identified in Figure 1 represent any entity that is able to emit CE and communicate with the node. These might be human agents inputting information through a text messaging interface, or machine agents which are communicating with the node as a result of policies (see Section III-B). CENode provides RESTful and programmatic APIs for supplying CE. The APIs are exposed to JavaScript applications (e.g., within web apps or Node.js applications) and the RESTful endpoints are exposed when CENode is run as a web service (e.g., again via Node.js).

B. Agents and Policies

Each CENode instance includes a local agent (see Figure 1), which is normally responsible for updating the local KB when cards are received. Agents in multi-agent setups are also able to send cards with respect to *policies*. Policies are instructions, written in CE, that, when applied to a node, may cause the local agent to try and communicate with another agent.

For example, consider a *tell policy*, which instructs the agent to forward any *tell card*s received on to another target agent and is useful for propagating information through a network of node instances:

```
there is a tell policy named 'p1' that targets
```

Fig. 2. Sample poster design (left), shown in situ (right).

```
the agent 'agent2'.
```

Other policy types include a *listen policy* (for retrieving cards from other agents) and a *feedback policy* (for governance over responses provided to received cards).

If policies are active on an agent, but there is no network route to other nodes, then the local node will still function as normal in the meantime, but will attempt to re-establish connections with other nodes once a network becomes available. Combining policies in different ways allows for the deployment of various network topologies of nodes that might be useful in different coalition settings.

IV. Experiment Method

The goal of the experiment was to compare the ability of small groups performing CSU tasks at the network edge under different connectivity conditions; specifically, to compare performance between groups (i) with 'good' connectivity allowing real-time sharing of the COP, and (ii) with no connectivity outside the environs of the base location allowing sharing of the collective picture only when the group returns to base at the end of their experiment run. A physical rather than virtual environment was chosen for the experiment in order to gather performance data on the operation of a network of CENode instances running in situ, as well as to provide a richer and more immersive setting for the participants, with human-human as well as human-machine collaboration opportunities. Participants were tasked to explore a set of given locations in a building complex and use mobile devices running CENode agents to capture information that the agents would use to assemble a COP in the form of a shared CE KB.

The situation was entirely synthetic, with elements depicted in a set of 16 stylised posters distributed in the vicinities of six given locations. The participants were tasked to provide essential elements of information (EEOIs) on six persons of interest (POIs): their location, the colour of their shirt, what sport they play, and what fruit they eat. Each poster depicted 2 or 3 EEOIs. An example is shown in Figure 2. Here, the POI Rev Hawk is shown located in the Emerald Room, wearing a red shirt, and with a pear (3 EEOIs). Participants were given a set of 'mugshots' of the POIs so they could recognise them on the posters, and they were tasked with answering 36 questions, such as:

```
What character eats pears?
What character is in the Emerald Room?
What character is wearing a red shirt?
What fruit does Rev Hawk eat?
Where is the pear?
```

Note that in some cases, the answer to a question can be inferred from the answer to another question. Such inferences are performed automatically by the CE agents as part of their task in assembling the COP. Given that our research focus is not on NLP, the elements of the synthetic situation were designed to be easily recognised and relatively easily described by users to the CE agent, with distinct shirt colours (black, white, red, green, blue, yellow), items of fruit (apple, banana, lemon, pear, orange, pineapple), and sports (baseball, cricket, golf, rugby, soccer, tennis). The experiment runs were conducted in the UK so some cultural background was assumed (recognisability of a rugby ball and cricket bat, for example) though there was potential for ambiguity in that 'football' is often used in the UK in preference to 'soccer', and the British game 'rounders' uses the same type of bat as baseball. Further potential for ambiguity existed in the fact that some of the characters looked superficially similar.

Participants used their own mobile devices to run instances of the CENode-based conversational agent, typically smartphones; an option was provided for them to use a tablet or PC at the base location if they experienced technical problems with running the agent, though a negligible number of participants took up this option. A full description of the capabilities of the agent is given in [19]. The core capabilities support (i) information capture where the user provides NL text input which the agent 'confirms' in CE and (ii) question answering of simple who/what/where queries. Figure 3 shows a screenshot from the agent. The conversation between the user (blue messages) and the agent (grey messages) is rendered as a conventional mobile app text message thread. The user's input ('Rev Hawk plays baseball') is in NL, which the agent maps to CE via simple NLP. The user can either confirm that the CE is an acceptable interpretation of his or her input (shown as the 'Yes' message here) or reject the agent's interpretation and try again. The user is permitted to enter the same piece of information only once. The user can also ask the agent questions, as shown in the bottommost message. The agent will use the current contents of its KB to try to answer the question (in this case, it has no information).

The top-left button on the agent user interface allows the user to view a 'dashboard' showing the current status of the COP in terms of the 36 questions. An example is shown in Figure 4. The meaning of the colours was explained to participants as follows:

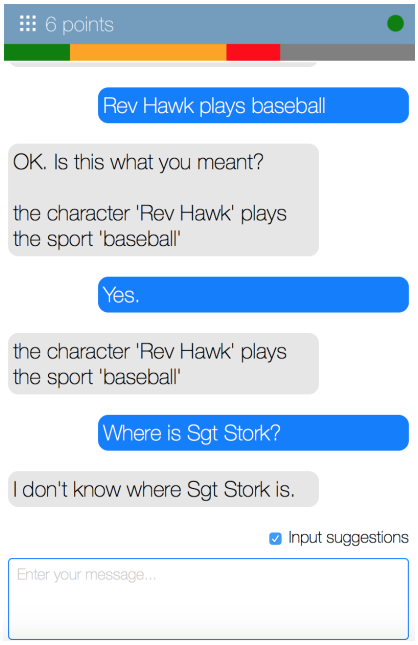


Fig. 3. Conversational agent user interface.

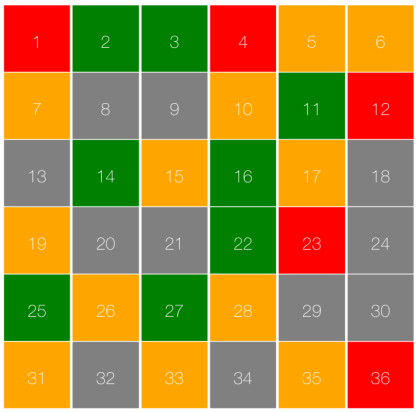


Fig. 4. Conversational agent dashboard display.

Grey	No information received to answer this question
Amber	Some information received, but insufficient to give a conclusive answer to this question
Green	Sufficient information received to give a conclusive answer to this question
Red	Conflicting information received in answer to this question

The indicator in the top-right corner of the agent user interface indicates whether the user's agent is currently online (green) or offline (grey). This is under the control of the experimenters, as explained below.

A. Experimental Design and Hypotheses

The experiment used a single factor, two-level between-participants design. In the Online Condition, the group members were given access to a conversational agent with full network connectivity, able to exchange collected information with all other participants' agents in real time using *tell* and *listen* policies as described in Section III-B. In the Offline Condition, the group members were given access to a version of the conversational agent with information exchange disabled until the end of the experiment run. The meaning of the indicator in the top-right of the agent user interface was explained to participants, making clear to participants in the Offline Condition that their indicator would remain grey until the end of the run (for Online participants, the indicator would reflect their actual network connectivity at any time during the run). The meaning of the dashboard display was also explained, highlighting the following:

- Online Condition: the dashboard would update in real time to show the current collective state of the COP in terms of information submitted by all group members.
- Offline Condition: the dashboard would reflect only the information submitted by the individual user (because no information would be exchanged) until the end of the run. Therefore, every square on the grid would be grey or amber depending on whether the individual had submitted any answers to that question. At the end, connectivity would be enabled and the dashboard would update to show the collective state for the group as for the Online Condition.

Participants were drawn from a sample of convenience: they were first and second year UK undergraduate students studying computer science. They had no prior knowledge of CE. The experiment was run over two days, with two groups on the first day and two on the second:

Group	Condition	Participants
A	Online	30 1st year undergraduates
B	Offline	15 1st year undergraduates
C	Online	13 2nd year undergraduates
D	Offline	8 2nd year undergraduates

The posters were distributed around the building complex as shown in Figure 5. The layout of the buildings was generally familiar to the participants. They were given an instruction sheet with the 36 questions, 'mugshots' of the six POIs, locations of the six 'room' posters (Amber Room, Emerald Room, Gold Room, Ruby Room, Sapphire Room, Silver Room) and told that the location of other posters needed to be discovered in the vicinities of the 'rooms'.

Prior to the experiment, participants were given a 10 minute briefing on the CSU task and the use of the conversational agent, but were given no opportunity to practice using the agent before participating in the experiment. A summary of the instructions for use of the agent was also on their sheet. Each group was briefed separately and each pair of groups (A/B, C/D) was told that they were in competition. Participants were instructed: "Your group is in competition with the other group for the highest group score. Your group gets one point for each answer you get to green on the dashboard."

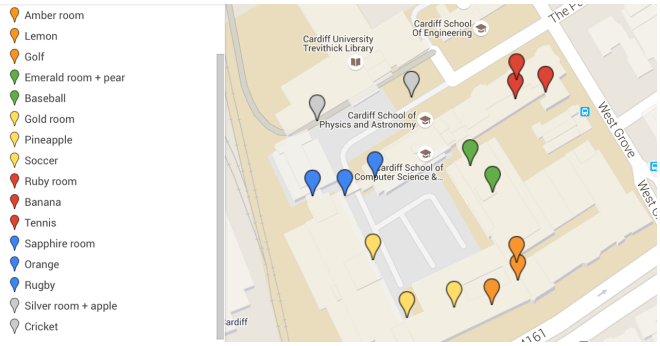


Fig. 5. Map showing approximate locations of the characters and objects (Groups A and B).

Following the briefing, groups were given 40 minutes to perform their task and instructed to return to the starting location at the end, where they were given a short debrief on their performance and the final state of their dashboard. In the case of the Offline Condition groups, this debriefing was the first time they were able to see the collective state of the dashboard, as connectivity was enabled for their agents at that point. After each pair of runs (A/B, C/D), all participants were told the final scores for both groups, and the winning group was revealed.

The primary hypothesis was that the Online Condition participants would build a more 'settled' CSU picture as measured by a greater proportion of 'green questions' than the Offline Condition participants, because the real-time connected status of these participants would allow them to collectively identify and focus their efforts on parts of the COP that required more information (grey or amber) or a higher degree of consistency (red).

V. Experiment Results

The experimenters recorded the following qualitative observations on the four runs:

- Group A was very noticeably more energetic than the other three, being faster to mobilise in leaving the starting location.
- Members of all four groups worked in sub-groups to some extent, with few apparently working entirely alone. Sub-groups followed different routes around the buildings.
- Some members of Group C seemed to be foraging for information and reporting back to friends who stayed in the starting location.
- Although there was no intention on the part of the designers to have any hidden meaning in the scenes depicted in the posters, nevertheless some participants believed that some of the POI names (e.g. 'Capt Falcon', 'Prof Crane') were references to characters in popular culture and therefore clues to some hidden situation.
- It was clear from observation of participants' behaviour in the Online Condition that they were trying to turn red squares to green by working collectively (each individual could only answer each question once so they were reliant on colleagues to provide additional corroborating information in order to resolve inconsistencies).
- Some participants in Group D (Offline Condition) attempted to supply further information after the dashboard was revealed!
- There was an example of groupthink with Group D: over half the members returned to the starting location shortly after leaving, to seek clarification on how to answer the questions. It seemed that a few dominant group members had taken a view on how to interact with the agent and, when their approach hadn't worked, they all returned to the starting location to see the briefing note again rather than try alternatives.

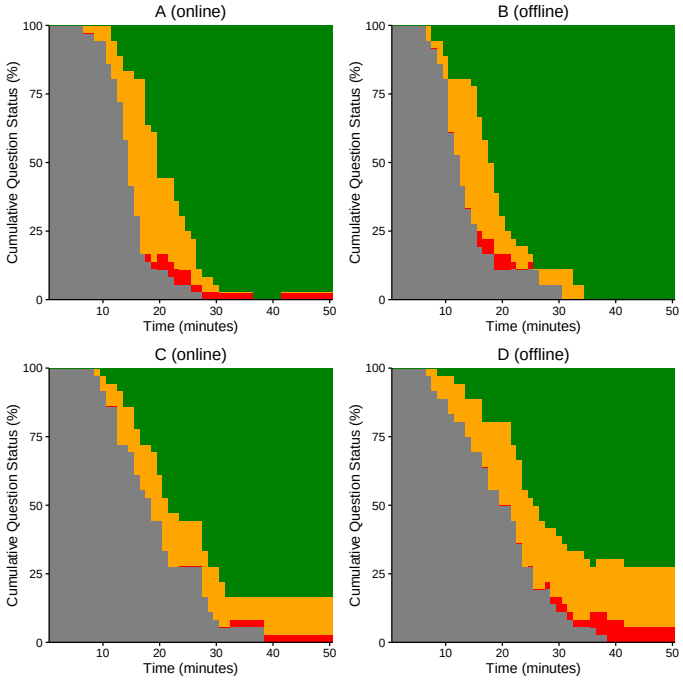


Fig. 6. Dashboard development over time for cumulative question status; colours correspond to definitions provided in Section IV

During the runs, each participant's conversational agent logged all cards generated, including the NL input from the user, confirmatory messages from the agent to the user, confirmed CE added to the KB, and any queries input by the user. Figure 6 shows a reconstruction of the progression of the dashboard for each of the groups, based on the logged cards. Reconstruction of the global state of the dashboard was necessary because this state did not exist at run-time. In the Online Condition, each participant only sees their local view of the dashboard which, due to distributed system effects such as intermittent network connectivity or delays in information sharing, may not be identical to other participants. In the Offline Condition, the state of the dashboard is computed only when participants come online at the end.

The visualisations show a period of 50 minutes, starting approximately 5 minutes before the start of each 40-minute run, and ending approximately 5 minutes after the end of

the run. The x-axis shows time and the y-axis shows the cumulative question status. Each run of the experiment starts with all questions grey (no information). Then question status begins to shifts to amber (some information), occasionally red (conflicting information), with green (sufficient information) growing over time.

A. CSU Results: Group Information Quality

In terms of group scores, Group B was the only group to achieve a fully settled set of 36 questions. Group A settled all but question 36 which was: 'What sport does Capt Falcon play?' and answers conflicted because some players used the name 'soccer' while others used 'football'. It was evident from observation that members of Group A were aware of this issue and were trying to coordinate their efforts to resolve this conflict, but were not able to do so. In fact, additional information received between minutes 40 and 41 in Figure 6(a) meant that the state of the KB with respect to question 36 became conflicted again. It is worth noting that participants were not told the internal rules that the agents used to determine which state to display on the dashboard. Where multiple conflicting answers were submitted for a question (e.g. 'soccer' vs. 'football') the green/red state was determined by counting the frequency of all submitted answers. If the count for the most frequently submitted answer was at least 3 higher than the count for the next most frequently submitted answer, the dashboard square for that question would be shown as green; otherwise it would be red. For example, 6 users answering 'soccer' and 3 users answering 'football' would result in green, while 6 users answering 'soccer' and 4 users answering 'football' would result in red.

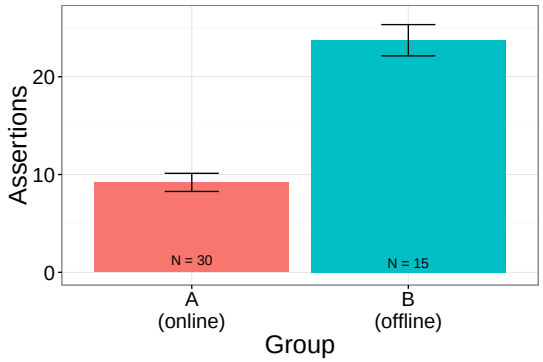
The smaller groups, C and D achieved a less settled state overall, with more questions still in the amber and red states at the end, particularly for Group D. The Offline Groups, B and D, had more questions in the red state during play — participants of course were unaware of this since their individual responses were not aggregated until connectivity was enabled at the end.

These visualisations suggest that:

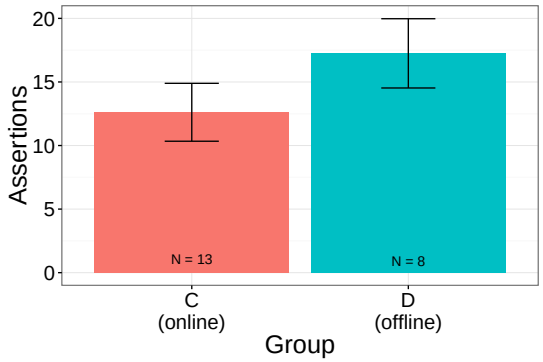
- The Online Condition groups maintained a more settled picture during the runs though did not markedly outperform the Offline Condition groups.
- The larger groups (A/B) achieved a more settled picture more rapidly than the smaller groups (C/D).

B. CSU Results: Group Information Quantity

Figure 7 reveals differences between the groups in terms of the mean number of assertions — statements added to the KB — made by the participants. Scores were analysed using separate binomial regressions. There was a significant difference between Groups A and B ($p < 0.001$) but not between C and D ($p = 0.30$). The likely explanation for this difference is participants in the Online Condition, being aware of the current state of the COP in terms of the real-time dashboard updates, reduced their efforts in the latter part of the run once the dashboard was predominately green.



(a) Groups A and B



(b) Groups C and D

Fig. 7. Mean participant assertions. Error bars represent one standard error of the mean. Note the y-axes differ.

Participants in the Offline Condition, being unaware only of the state of their individual dashboard, continued to make assertions until their dashboard was predominantly amber. One member of Group A commented afterwards that members of their group had realised that, once a question was settled and the dashboard showed it was green, it was counterproductive to continue making assertions in relation to that question as to do so risked introducing conflict in the collective KB and turning the dashboard red for that question.

To quantify the magnitude and confidence intervals for the effect sizes between the Online and Offline Conditions, a meta-analytic approach was used [21], see Figure 8. Note the very large effect size for Groups A/B (*Pseudo-*R^2 = 0.62) and, in contrast, the almost medium effect size, albeit with a wide confidence interval, for C/D (*Pseudo-*R^2 = 0.08). The overall, pooled effect size was large (*Pseudo-*R^2 = 0.53). That is, 53% of the variance in assertions can be explained by the online versus offline manipulation.

The effect size was calculated using a linear model for the correlation between the actual and fitted values [22]. This is denoted here with *Pseudo-*R^2. *Pseudo-*R^2s provide an effect size estimate for non-linear and other complex models which do not otherwise have an measure of absolute fit, see [23]. For Groups A/B and C/D, the confidence intervals for the *Pseudo-*R^2s were determined using a quantile bootstrap, a robust

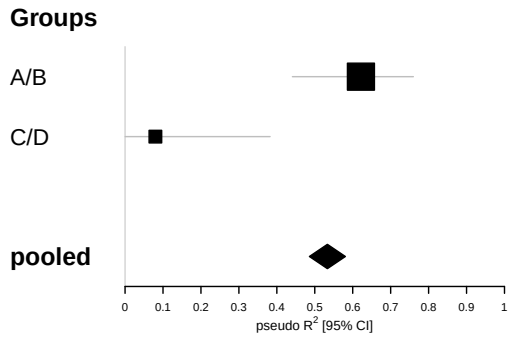


Fig. 8. Forest plot of effect sizes for online vs. offline assertions with bootstrapped 95% confidence intervals. The pooled effect combines A/B and C/D.

random resampling method for parameter estimation [24]. The pooled effect size was calculated by combining these two effects, weighted by the number of participants [25].

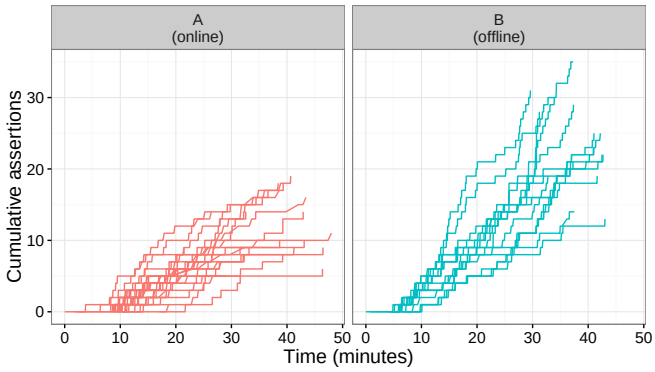
Examining the cumulative progression of individual numbers of assertions during the runs in Figure 9 we see a markedly higher reach in Group B compared to Group A, suggesting again that members of Group A reduced their effort in the latter stages of the game as their dashboard became predominately green.

The lesser difference in individual performance between Groups C and D is likely due to the smaller group size and the fact that it took them longer to achieve settled (collective or individual) states than the larger groups. Looking at histograms of the individual performance in terms of assertions between Groups A/B and C/D, Figure 10, we see that members of the offline groups appeared to 'work harder' with a frequency shift to higher numbers of individual assertions. This is corroborated by the number of messages each group submitted: the number of messages submitted per person in the Offline Condition is nearly double that of the Online Condition.

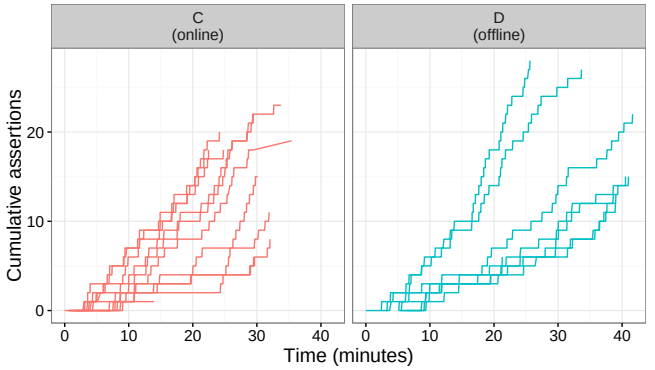
	Online $N = 43$	Offline $N = 23$
Total messages	1031	1040
Average messages per participant	24.0	45.2

Figure 11 shows the volume of cards generated in the system based on the logged data during each of the four runs as a measure of group activity. The volumes show reducing activity in the Online Condition groups in the latter stages. Comparing Groups A and B, we see activity in Group A reduces from around the halfway point (20 min into the run) whereas the reduction occurs only in the final 5 minute period for Group C compared to Group D. The results for Group D in particular is sensitive to the performance of specific individuals as this group was the smallest.

Results are fully reproducible. The data, analyses, and graphs are available from: https://osf.io/5fhsb/



(a) Groups A and B



(b) Groups C and D

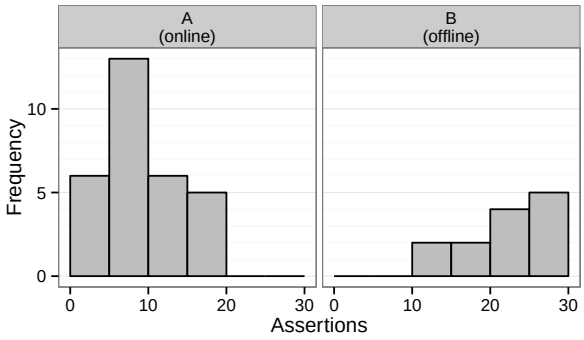
Fig. 9. Individual participants' cumulative assertions. Note the y-axes differ.

VI. Conclusion and Future Work

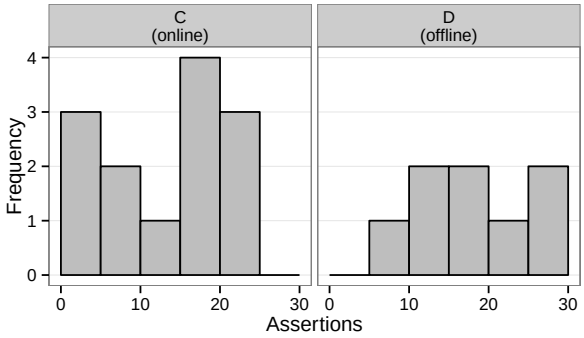
In this paper, we introduced CENode as a novel technology for HCC at the network edge, and presented results from a behavioural experiment comparing the performance of groups using CENode to achieve CSU in online and offline settings. In our experiment, the COP increased the quality, information agreement, of CSU but the quantity of information was greater without the COP. Other research has conceptually noted limitations of COPs [26]; our research empirically illustrates the tradeoffs with a COP even if all users have a shared goal.

Because of the time and personnel needed to conduct this experiment in the real-world, and analysis at the group level, the overall sample size was limited. However, this design has solid external validity because it was a simulation of tactical intelligence with relevant aspects of actual tasks such as time pressure, uncertainty, dynamic interactions. The heterogeneity in the effect sizes for CSU with the COP may be attributable to differences in group size and dynamics. To further assess the effects of a COP on information quality and quantity, we plan to run a conceptual replication of the experiment online using a large sample from Amazon's Mechanical Turk — see below.

The results suggest CENode is a promising technology for supporting rich human-machine interactions in situations

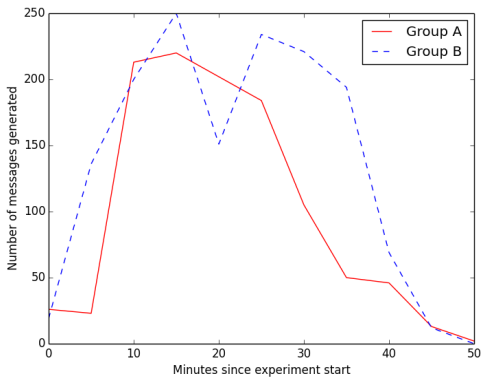


(a) Groups A and B

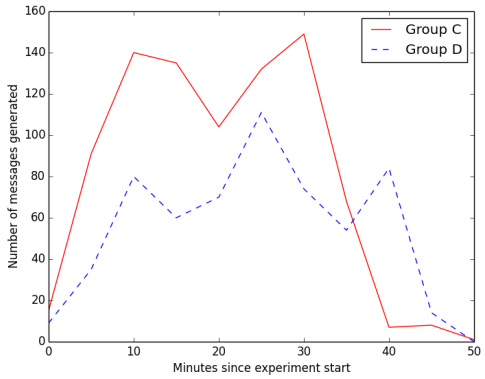


(b) Groups C and D

Fig. 10. Histogram of participants' total assertions.



(a) Groups A and B



(b) Group C and D

Fig. 11. Message volumes at 5-minute intervals.

where users predominantly collect local knowledge (the offline setting) as well as situations where they are able to assemble a global COP in real-time.

Given the initial success, we plan to develop CENode further and use it as a basis for additional behavioural experiments with scenarios that require human-machine conversational interactions for solving tactical and crowd-sourced intelligence, surveillance, and reconnaissance tasks, including the following:

- Instead of a real-world CSU exercise where participants directly experience elements of the situation, participants will gather situational information via simulated 'sensor feeds'. This design would be well-suited to delivery via online platforms such as Amazon's Mechanical Turk[3] or Volunteer Science[4] allowing access to greater number of participants, although it would favour a single-participant rather than group exercise, due to difficulties in coordinating groups of participants via such platforms.
- Enrich the CSU task with the addition of features such as 'hidden' or 'anomalous' objects, that are not explicitly referenced in the participants' tasking. This design would assess the extent to which participants are steered by the specifics of the tasking versus the situation. In other words, are they approaching the task as being 'open' or 'closed', and does this differ between Online or Offline Conditions?
- During the initial experiments, participants were requested to supply location data along with their submitted information, based on the Global Positioning System (GPS) coordinates of their device at the time each input was made. The experiments were conducted indoors so the GPS data collected was noisy and inconclusive; nevertheless, we plan to analyse these data in detail and consider ways to collect more accurate location data in future experiments in real-world settings. The data may provide insights as to how the various groups tackled the CSU task (e.g., dispersing vs. staying together), the extent to which they backtracked (e.g., to revisit a location to gather missing information or collect confirmatory information), and compare behaviours between Online and Offline Conditions.
- Assess human, with minimal training, and machine performance with CENode capabilities for agile knowledge representation and benefits of hybrid human-machine reasoning [27]. Agile knowledge representation would allow users to dynamically create new entities and concepts. For

[3]https://www.mturk.com
[4]https://volunteerscience.com

example, the name of a person ('Mike'), their age ('45'), height ('6 feet'), hair colour ('black'), location ('Hursley Village') and social connections ('brother of John'). This can be further extended to hybrid human-machine reasoning, where humans can understand and leverage the power of machine reasoning for large amounts of information. For example, what people are from Hursley Village or who is in Mike's family?

ACKNOWLEDGEMENTS

This research was sponsored by the US Army Research Laboratory and the UK Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. Dr. Zaroukian was supported by an appointment to the US Army Research Laboratory Postdoctoral Fellowship Program administered by the Oak Ridge Associated Universities. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the US Government, the UK Ministry of Defence or the UK Government. The US and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] A. Preece, D. Pizzocaro, D. Braines, D. Mott, G. de Mel, and T. Pham, "Integrating hard and soft information sources for D2D using controlled natural language," in *Proc 15th International Conference on Information Fusion*, 2012.

[2] T. Kuhn, "A survey and classification of controlled natural languages," *Computational Linguistics*, vol. 40, pp. 121–170, 2014.

[3] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *Computational Intelligence Magazine, IEEE*, vol. 9, no. 2, pp. 48–57, 2014.

[4] A. Preece, C. Gwilliams, C. Parizas, D. Pizzocaro, J. Z. Bakdash, and D. Braines, "Conversational sensing," in *Proc Next-Generation Analyst II (SPIE Vol 9122)*. SPIE, 2014.

[5] A. Preece, W. Webberley, D. Braines, E. G. Zaroukian, and J. Z. Bakdash, "SHERLOCK: Experimental evaluation of a conversational agent for mobile information tasks," Under Review.

[6] M. S. Vassiliou, D. S. Alberts, and J. R. Agre, *C2 Re-envisioned: The Future of the Enterprise*. CRC Press, 2014.

[7] C. J. Alberts, A. J. Dorofee, J. H. Allen, and N. L. Compton, *Power to the Edge: Command and Control in the Information Age*. Command and Control Research Program, 2003.

[8] NATO-SAS-065, *NATO NEC C2 Maturity Model (N2C2M2)*. CCRP, 2010.

[9] A. G. Ingham, G. Levinger, J. Graves, and V. Peckham, "The Ringelmann effect: Studies of group size and group performance," *Journal of Experimental Social Psychology*, vol. 10, no. 4, pp. 371–384, 1974.

[10] F. P. Brooks Jr, *The Mythical Man-Month: Essays on Software Engineering*, 2nd ed. Pearson, 1995.

[11] L. R. Marusich, J. Z. Bakdash, E. Onal, S. Y. Michael, J. Schaffer, J. O'Donovan, T. Höllerer, N. Buchler, and C. Gonzalez, "Effects of information availability on command-and-control decision making performance, trust, and situation awareness," *Human Factors*, vol. 58, no. 2, pp. 301–321, 2016.

[12] R. Crouser and R. Chang, "An affordance-based framework for human computation and human-computer collaboration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18(12), pp. 2859–2868, 2012.

[13] E. Blasch, P. Valin, and E. Bosse, "Measures of effectiveness for high-level fusion," in *Information Fusion (FUSION), 2010 13th Conference on*. IEEE, 2010, pp. 1–8.

[14] H. A. Simon, *The Sciences of the Artificial*. MIT Press, 1996.

[15] J. R. Hackman, "Learning more by crossing levels: Evidence from airplanes, hospitals, and orchestras," *Journal of Organizational Behavior*, vol. 24, no. 8, pp. 905–922, 2003.

[16] L. Terveen, "Overview of human-computer collaboration," *Knowledge-Based Systems*, vol. 8(2), pp. 67–81, 1995.

[17] D. Mott, "Summary of ITA Controlled English," 2010. [Online]. Available: https://www.usukita.org/papers/5658/details.html

[18] W3C, "OWL 2 Web Ontology Language document overview (second edition)," World Wide Web Consortium, Dec. 2012. [Online]. Available: http://www.w3.org/TR/owl2-overview/

[19] A. Preece, D. Braines, D. Pizzocaro, and C. Parizas, "Human-machine conversations to support multi-agency missions," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 18(1), pp. 75–84, 2014.

[20] D. Braines, D. Mott, S. Laws, G. de Mel, and T. Pham, "Controlled english to facilitate human/machine analytical processing," in *Proc Next-Generation Analyst (SPIE Vol 8758)*. SPIE, 2013.

[21] G. Cumming, "The new statistics why and how," *Psychological Science*, p. 0956797613504966, 2013.

[22] B. Zheng and A. Agresti, "Summarizing the predictive power of a generalized linear model," *Statistics in Medicine*, vol. 19, no. 13, pp. 1771–1781, 2000.

[23] M. R. Veall and K. F. Zimmermann, "Pseudo-R^2 measures for some common limited dependent variable models," *Journal of Economic Surveys*, vol. 10, no. 3, pp. 241–259, 1996.

[24] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*. Cambridge University Press, 1997, vol. 1.

[25] C. Harris, L. Hedges, and J. Valentine, *Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, New York, 2009.

[26] D. Mendonça, T. Jefferson, and J. Harrald, "Collaborative adhocracies and mix-and-match technologies in emergency management," *Communications of the ACM*, vol. 50, no. 3, pp. 44–49, 2007.

[27] D. Braines, J. Ibbotson, D. Shaw, and A. Preece, "Building a living database for human-machine intelligence analysis," in *Information Fusion (Fusion), 2015 18th International Conference on*. IEEE, 2015, pp. 1977–1984.