Human understanding of information represented in natural versus artificial language (Poster)

Erin Zaroukian
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD
erin.g.zaroukian.ctr@mail.mil

Jonathan Z. Bakdash
U.S. Army Research Laboratory South at the University of Texas Dallas
Richardson, TX
jonathan.z.bakdash.civ@mail.mil

***Abstract*—In this paper we compare human understanding of information represented in a natural language (NL) to a type of artificial language, called a Controlled Natural Language (CNL). Potential applications for CNLs include decision support and conversational agents, but currently there is limited empirical research on the understandability of CNLs for untrained humans. We investigate a particular type of CNL, called Controlled English (CE), which was designed to be a simplified, artificial subset of natural language that is both human readable and unambiguous for fast and accurate machine processing. We quantify and compare human understanding of NL and CE using accuracy and speed for language statements. The statements described entities (people and objects) and relations (actions) among entities with the ground-truth represented using visual diagrams. Participants responded whether the statements matched the diagram (yes/no). In Experiment I, we found accuracy for NL and CE was comparable, although the speed for understanding CE was slower. To further examine the role of speed, we induced time pressure in Experiment II. We found both the accuracy and speed for CE was lower than NL. These results indicate that if people have sufficient time, understanding for CE can be equivalent to NL. However, with limited time the accuracy and speed for understanding NL is better than CE. Our findings indicate that both accuracy and speed of CNLs should be evaluated. Furthermore, under time pressure there can be meaningful differences in accuracy and speed between different ways of representing information. Understanding for methods of representing machine information has potential implications for situation understanding and management with human-machine interaction and collaboration.**

Keywords— information representation, controlled natural language, natural language, understanding, human-machine interaction, human-computer collaboration

I. Introduction

Computational systems can be used to represent information for human-machine interaction (e.g., decision support systems) and collaboration (e.g., conversational agents) with the potential to aid human information processing and enhance decision-making [1]–[5]. However, these systems have mixed effectiveness, and even decrements, for human performance [6], [7]. The inconsistent findings have been attributed to a variety of factors (e.g., automation bias and complacency, lack of transparency and opaque rationale in computational information and reasoning) [7]–[10]. The effective representation of information for humans has implications for situation understanding and management.

Here, we investigate human understanding for information represented using two types of language: Natural language (NL) and an artificial language called a Controlled Natural Language (CNL) [11], [12]. A CNL is "… a subset of natural language that can be accurately and efficiently processed by a computer, but is expressive enough to allow natural usage by non-specialists" [12: p. 1]. Applications for CNLs include knowledge-based systems such as decision-support systems and conversational interfaces [13]–[15]. Typical users are not expected to write CE on their own. However, we have shown that with minimal training users can effectively with a conversational agent that parses their NL input into CE for inclusion in a knowledge base or for knowledge base querying [13]

We focus on comparing understanding of information represented in NL versus a type of CNL called International Technology Alliance Controlled English (CE) [16]. CE was designed to be human readable and unambiguous for fast and accurate machine processing of information, bridging the gap between NL and programming languages [16]. There are numerous types of CNLs with tradeoffs in human understandability and flexibility [17].

Despite the large number of different CNLs, empirical research on the human understanding of CNLs is limited. Prior work tends to only assess accuracy, not speed, and it typically overlooks comparisons with NL [17]–[19]. We measure both speed and accuracy for NL and CE statements. Statements were either correct or incorrect using a ground-truth visual representation of relationships among entities. This visual representation is called an ontograph [17].

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-17-2-0003. 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 Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Using ontographs paired with NL and CE statements, we compared understanding for each type of language in two experiments. In Experiment I, we assessed baseline performance for NL versus CE. There was no specific hypothesis for the comparison because while CE was designed to be human readable, it is also more verbose. Thus, there were multiple possibilities (e.g., lower accuracy and a slower speed for CE, higher accuracy and lower speed for CE because its extra information could draw attention to specific elements in the ontograph [such as persons], or accuracy or speed alone for CE would be negatively impacted). Overall, we found accuracy for NL and CE was comparable but speed for CE was slower. In Experiment II, we followed up on the effect of speed by experimentally manipulating it using time pressure. Time pressure results in shallower human information processing [20][20], which can negatively impact accuracy. We hypothesized the reduced depth of processing caused by time pressure would result in lower accuracy for CE than NL. This is because under limited time, the fluency to process CE would be degraded by it being more verbose and less familiar than NL. There were two distinct hypotheses for speed: Either speed would be similar (comparable use of limited time to understand information in NL and CE) or slower for CE. Under time pressure in Experiment II, we found CE had lower accuracy and a slower speed than NL.

II. Experiment I

A. Methods and Procedure

In Experiment I, participants provided *yes/no* judgments for ontograph-statement pairs in Controlled English (CE, a CNL, [16]) and a Natural Language (NL, here English), as well as separate subjective usability ratings for the languages [21]. Accuracy and response time (speed) for the two languages, along with their usability ratings, are compared.

1) Participants

One hundred three participants (48 female, 55 male) were recruited through Amazon Mechanical Turk to participate in this experiment online. Participants were aged 21-63 (Median = 32). Participants were compensated $0.75. We did not ask about the native language of participants. Participants were required to have a HIT approval rate of at least 95% and to have at least 50 HITs approved in order to participate.

2) Materials and Equipment

The experiments were prepared using the Ibex tool for running behavioral psycholinguistic experiments (https://code.google.com/archive/p/webspr/) and run online through Amazon Mechanical Turk. Data and full results are available online: osf.io/bkx8d.

3) Procedure

In this experiment, participants were presented with an ontograph paired with a written statement, and their task was to respond "yes" if the statement matched the ontograph and "no" otherwise. The complexity of the statements was varied within subjects, where complexity was increased by introducing a Boolean operator ("and", "not"). Participants were instructed to respond as quickly and as accurately as possible. An example trial is shown in Fig. 1. Prior to starting each block, participants completed 6 practice trials using simple statements, see (1)-(2). Feedback on accuracy was provided for the practice trials. These practice trials were included primarily to familiarize participants with ontographs and how to evaluate them, but this practice also provides participants with some exposure to CE. We nonetheless consider participants untrained with respect to CE.

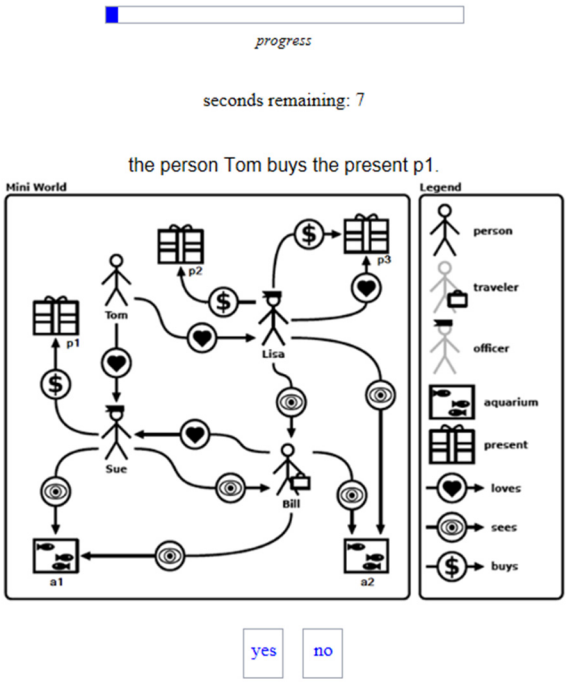

Fig. 1. Example of trial showing from top to bottom: Progress bar, time pressure (only used in Experiment II), CE statement, ontograph (mini world and legend), and response boxes (yes/no). Participants could respond with the keyboard or mouse.

The experiment used a within-participants design, 2 statement language (English/NL and Controlled English/CE) by 2 complexity (simple and complex). That is, every participant completed trials in all four conditions. Statement language was blocked and counterbalanced. There were two types of simple statements (identity and relation) and two types of complex statements (conjunction and negation). Participants completed a total of 48 trials: 24 trials for NL and CE each with 12 trials for simple (5 identity and 7 relation) and complex (6 conjunction and 6 negation) in each block. Six different ontographs were selected from [21][1] for use in this experiment. Four ontographs were used in each block, with two of the six ontographs appearing in both blocks. Each ontograph was presented a total of 8 times.

Examples statements corresponding to the ontographs in Fig. 1 are shown in (1)–(4). Note the paired NL and CE

[1] Ontographs are available at: http://attempto.ifi.uzh.ch/site/docs/ontograph/

statements are equivalent and the correct response appears in italics following each statement. In the actual experiment, statements were true/false on half the trials in each block and the trial order was randomized in block (statement language and simple/complex).

1. Simple – Identity
    a. NL: Sue is an officer. *yes*
    b. CE: there is an officer named Sue. *yes*
2. Simple – Relation
    a. NL: Tom buys a present. *no*
    b. CE: the person Tom buys the present p1. *no*
3. Complex – Conjunction
    a. NL: Sue loves Tom and Tom loves Lisa. *no*
    b. CE: the person Tom loves the person Tom and the person Tom loves the person Lisa. *no*
4. Complex – Negation
    a. NL: Tom doesn't see an aquarium. *Yes*
    b. CE: it is false that the person Tom sees the aquarium a1. *yes*

B. *Results*

Upon examining response times, 866/4944 trials (18%) were excluded from analysis using the following exclusion criteria: (1) All trials faster than 2 seconds (861 trials) were excluded as this was not enough time to realistically perform the task. (2) All trials slower than 2 minutes (5 trials) were excluded as being exceptionally slow. Note the exclusion criteria were solely based on response time; accuracy was not considered.

We used R [23] and lme4 [24] to perform linear mixed effects analyses of the relationship of language on the accuracy and speed of responses for the remaining trials.

1) Accuracy

Mean accuracy is shown in Fig. 2. A generalized linear mixed-effects model was performed to analyze the relationship between language and accuracy. Language (NL vs. CE) and Complexity (Simple: Identity and Relation; Complex: Conjunction and Negation), with their interactions, were entered as fixed effects and participants as random intercepts.

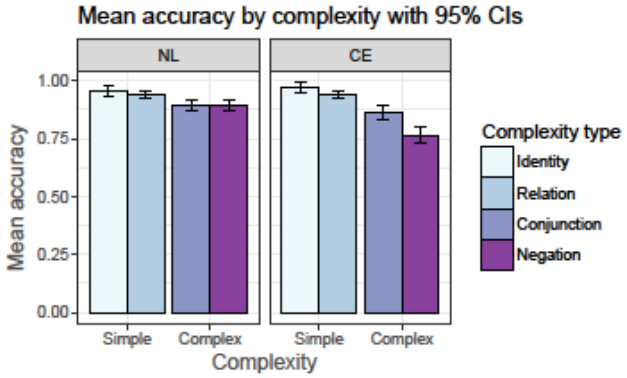

Fig 2. Mean accuracy for NL vs. CE with the different types of simple and complex statements. Error bars depict 95% confidence intervals (CIs).

Model results are summarized in Fig. 3 using the odds ratio for parameter estimates. Here, the odds ratio (OR) is the odds or probability for the proportions of accuracy/inaccuracy for the first factor *relative* to the second factor (or between two factor levels). Specifically, ORs are interpreted as:

a) OR = 1 is equal probability, the same proportion of accuracy/inaccuracy between two factors (i.e., no effect)

b) OR > 1 is the first factor has higher proportion (accuracy/inaccuracy), that is higher accuracy and thus lower inaccuracy, relative to proprotion (accuracy/inaccuracy) for the second factor. For example, an OR = 3 can be calculated from accuracy/inaccuracy for factor 1 (accuracy = 0.90/inaccuracy = 0.10 = 9) divided by factor 2 (accuracy of 0.75/inaccruacy = 0.25 = 3) which is odds of 9 (factor 1 proportion)/odds of 3 (factor 2 proportion) = odds ratio of 3. Note that the OR is relative, not absolute, because other proportions of accuracy/inaccuracy can produce an identical OR.

c) OR < 1 is the opposite b), the first factor has a lower proportion (accuracy/inaccuracy), that is lower accuracy and thus higher inaccuracy, relative to proportion (accuracy/inaccuracy) for the second factor.

No meaningful difference in overall accuracy was found for language (NL vs. CE); the OR is near 1 and the confidence intervals extend below 1. However, reliable interactions were found for Language x Complexity as well as Language x Complex, indicating a difference of differences in patterns for NL and CE (see Fig. 2). Other meaningful differences, across languages, included higher accuracy for Simple over Complex, Conjunction over Negation, and Identity over Relation. These results demonstrate the Simple vs. Complex manipulation was effective and that particular types of simple and complex statements were easier to understand than other ones.

The absolute fit of the model was assessed using two pseudo-R^2 values [25]. For only fixed effects, the marginal pseudo-R^2 was 0.10. For fixed and random effects, the conditional pseudo-

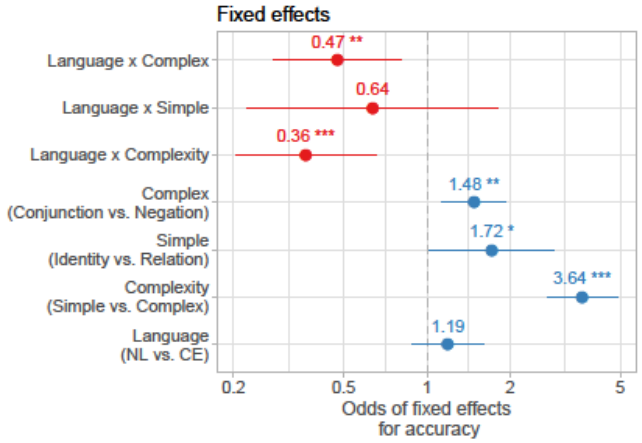

Fig 3. Summary of model results for accuracy. The x-axis depicts the odds ratio and the y-axis the name of each parameter estimate. The dot indicates the odds ratio of the fixed effect and the width of the line is the 95% confidence interval. Significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

R^2 was 0.29. ANOVA results for the fixed effects are shown in Table I.

TABLE I. ANOVA RESULTS FOR ACCURACY FIXED EFFECTS

Fixed effects	χ^2	p-value
Language	1.34	0.25
Complexity type	99.81	< 0.001
Language x Complexity type	22.13	< 0.001

2) Speed

Mean response time is shown in Fig. 4. A linear mixed-effects model was run with Language (NL vs. CE) and Complexity (Simple: Identity and Relation; Complex: Conjunction and Negation) as predictors with random intercepts for participants and response time as the dependent variable.

Model results for speed are summarized in Fig. 5 using fixed effects parameter estimates. Parameter estimates are regression coefficients using seconds and interpreted as follows:

a) Estimate equal to 0 seconds indicates no difference between factors (or factor levels)

b) Estimate > 0 seconds indicates a slower speed (higher response time) for the first factor compared to the second factor

c) Estimate < 0 seconds indicates a faster speed (lower response time) for the first factor compared to the second factor; the opposite of b)

We found a meaningful difference in language (NL vs. CE), where response times were faster (i.e., lower by 2.09 seconds plus a small value for the random intercept [0.07 seconds]) for NL. Across languages, other meaningful differences included faster speeds for Simple over Complex statements, Identity over Relation statements, and Negation over Conjunction statements. In addition, we found an interaction for Language x Complexity (see Fig. 5).

For only fixed effects, the marginal pseudo-R^2 was 0.12. For fixed and random effects, the conditional pseudo-R^2 was 0.33. ANOVA results for the fixed effects are shown in Table II.

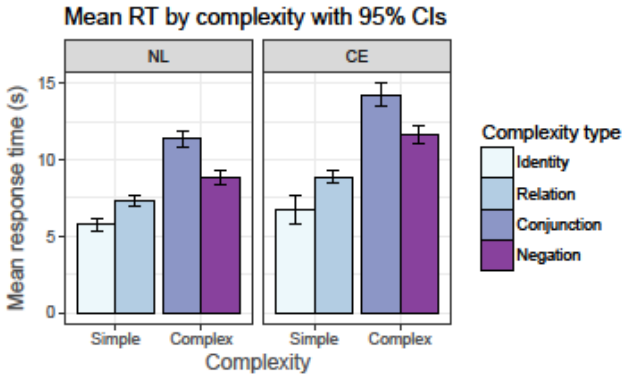

Fig 4. Mean response time in seconds (lower values indicate faster speeds) for NL vs. CE with the different types of simple and complex statements. Error bars depict 95% confidence intervals.

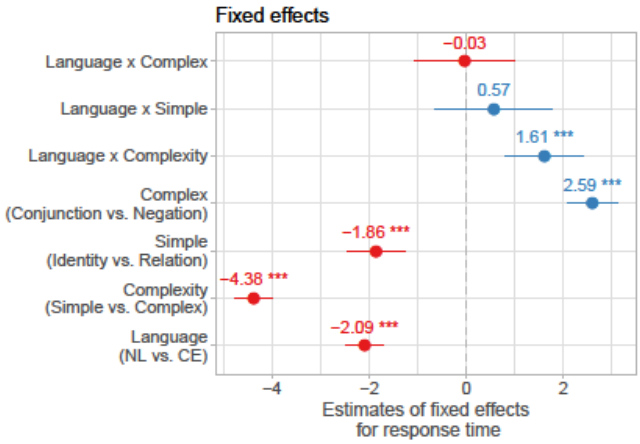

Fig 5. Summary of model results for speed. The x-axis depicts the value of the estimated parameters in seconds and the y-axis the name of each parameter. The dot indicates the mean fixed effect and the width of the line is the 95% confidence interval. Significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

TABLE II. ANOVA RESULTS FOR RESPONSE TIME FIXED EFFECTS

Fixed effects	F	p-value
Language	105.71	<0.001
Complexity type	189.13	< 0.001
Language x Complexity type	5.39	0.001

3) Usability

The distribution of participant ratings for the usability of the system was higher for NL (*Mean* SUS = 69.83, *Median* SUS = 72.5 *SD* = 8.46) than for CE (*Mean* SUS = 64.51, *Median* SUS = 62.5, *SD* = 8.46), $p = 0.001$ (Wilcoxon signed-rank test).

C. Discussion

Overall, participants in Experiment I performed with high accuracy. Accuracy for complex statements was understandably lower than for simple statements, but this difference is more pronounced in CE than in NL due to the interaction for language and complexity. Within the complex trials, accuracy for negation was lower than for conjunction overall, though again this appears to be driven by the CE trials, suggesting that there may be something dangerously unintuitive in the way CE expresses negation. However, overall high accuracy suggests that CE was understandable to participants with only minimal training.

CE trials were overall slower than NL trials, which is not surprising given that participants were less familiar with CE and that CE is often more verbose (see (1)-(4)). Also unsurprising are the slower response times for complex trials than for simple trials. This effect is more pronounced in CE, which mirrors the lower accuracy for complex trials driven by CE. We also found that, within simple trials, relation trials were slower than identity trials, which may be due to relations requiring more visual searching (identifying two entities and a relationship *between*

them). Within complex trials, conjunction trials were slower than negation trials. While negation, especially in CE, can be tricky (e.g., determining scope of negation, needing to exhaust the search space before determining that negation is correct), conjunction is often more verbose and often requires the participant to check the truth of two propositions instead of one, which may explain its relative slowness.

While CE trials lead to slower response times (approximately 2 seconds) than NL trials, differences in accuracy between the languages were quite small because of variability (for fixed effects alone, the marginal Pseudo-R^2 values for accuracy and response time were in the medium effect size range: 10% and 12%, respectively). It is possible that without extra time, accuracy on CE trials would drop below that of NL trials. In Experiment II we test this.

III. Experiment II

Given that the primary difference in Experiment I was in response time, we sought to experimentally decrease time in Experiment II by introducing a 10-second response deadline. Time pressure has been shown to limit the depth of human information processing, resulting in decreased accuracy, and this may differentially affect NL and CE.

A. Methods and Procedure

Methods and procedures were identical to Experiment I with the exception of an added 10-second deadline. Remaining seconds were displayed at the top of the screen for each trial.

1) Participants

An additional 101 participants (41 female, 59 male, 1 no response) were recruited through Amazon Mechanical Turk to participate in this experiment. Participants were aged 19-66 (*Median* = 31). Participants were compensated $0.75.

2) Materials and Equipment

Materials and equipment were identical to those in Experiment I with the exception of the added 10-second deadline.

B. Results

Upon examining response times, 550/4848 trials (11%) were excluded from analysis for being faster than 2 seconds.

1) Accuracy

Mean accuracy is shown in Fig. 6. Note the accuracy for complex statements (conjunction and negation) is nearing chance performance (50%) for CE, indicating poor understanding.

As in Experiment I, a generalized linear mixed-effects model was performed to analyze the relationship between language and accuracy. Language (NL vs. CE) and Complexity (Simple: Identity and Relation; Complex: Conjunction and Negation), with their interactions, were entered as fixed effects and participants as random intercepts.

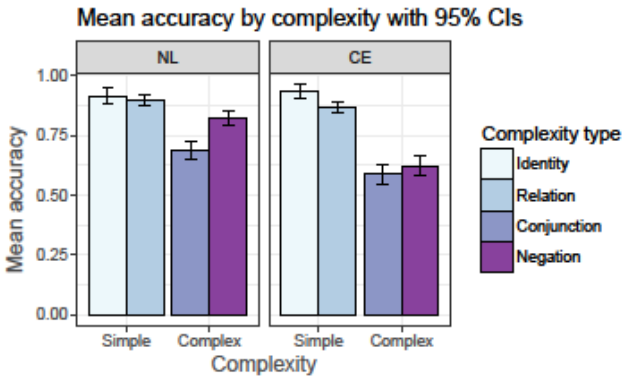

Fig 6. Mean accuracy for NL vs. CE with the different types of simple and complex statements. Error bars depict 95% confidence intervals.

Model results are summarized in Fig. 7 using the odds ratio for parameter estimates. For only fixed effects, the marginal pseudo-R^2 was 0.16. For fixed and random effects, the conditional pseudo-R^2 was 0.31. These effect sizes were similar to accuracy and speed in Experiment I. ANOVA results for the fixed effects are shown in Table III.

TABLE III. ANOVA Results for Accuracy Fixed Effects

Fixed effects	χ^2	p-value
Language	253.18	< 0.001
Complexity type	15.04	< 0.001
Language x Complexity type	310.51	< 0.001

With the addition of a 10-second deadline, we found overall accuracy for NL higher than CE, as well as reliable interactions for Language x Complexity and Language x Complex. Across languages, we again found higher accuracy for Simple over Complex, Identity over Relation, and Conjunction over Negation. These results demonstrate that additional time

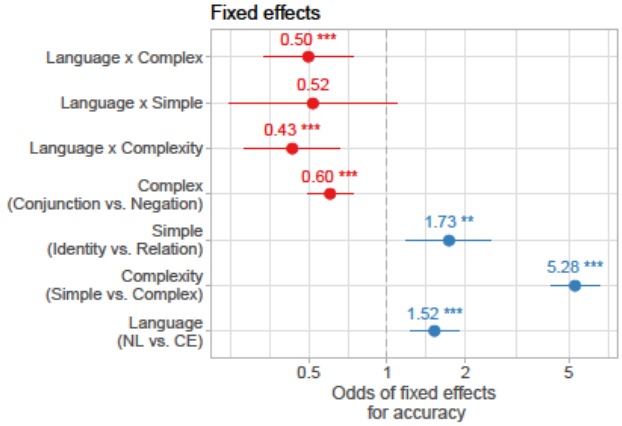

Fig 7. Summary of model results for accuracy. The x-axis depicts the odds ratio and the y-axis the name of each parameter estimate. The dot indicates the odds ratio of the fixed effect and the width of the line is the 95% confidence interval. Significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

pressure was enough to uncover differences in accuracy between NL and CE while replicating the remaining effects found in Experiment I.

2) Speed

Mean response time is shown in Fig. 8. A linear mixed-effects model was again run with Language (NL vs. CE) and Complexity (Simple: Identity and Relation; Complex: Conjunction and Negation) as predictors with random intercepts for participants and response time as the dependent variable.

Model results for speed are summarized in Fig. 9 using fixed effects parameter estimates. We again found a meaningful difference in language (NL vs. CE) with faster response times for NL (by approximately 0.66 seconds). Across languages, we again found faster response times for Simple over Complex statements, Identity over Relation statements, and Negation over Conjunction statements, with significant interactions for Language x Complexity, Language x Simple, and Language x Complex.

For only fixed effects, the marginal pseudo-R^2 was 0.22. For fixed and random effects, the conditional pseudo-R^2 was 0.50. The magnitudes of the effects for speed were larger than accuracy, suggesting that time pressure has more of an impact on response time than accuracy. ANOVA results for the fixed effects are shown in Table VI.

TABLE IV. ANOVA RESULTS FOR RESPONSE TIME FIXED EFFECTS

Fixed effects	F	p-value
Language	132.93	< 0.001
Complexity type	489.13	< 0.001
Language x Complexity type	18.47	< 0.001

3) Usability

As in Experiment I, the distribution of participant ratings for the usability of the system was higher for NL (*Mean* SUS = 65.79, *Median* SUS = 67.50, *SD* = 14.71) than CE (*Mean* SUS = 51.81, *Median* SUS = 50.00, *SD* = 14.71), $p < 0.001$ (Wilcoxon signed-rank test).

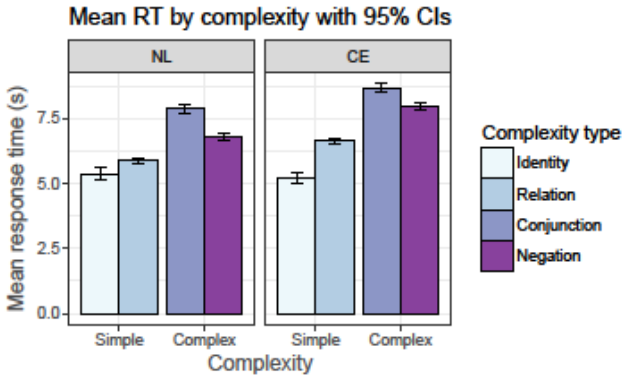

Fig 8. Mean response time in seconds (lower values indicate faster speeds) for NL vs. CE with the different types of simple and complex statements. Error bars depict 95% confidence intervals.

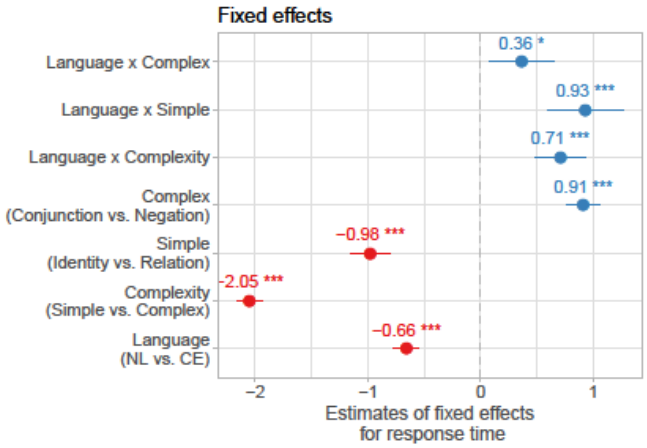

Fig 9. Summary of model results for response time. The x-axis depicts the value of the estimated parameters in seconds and the y-axis the name of each parameter. The dot indicates the mean fixed effect and the width of the line is the 95% confidence interval. Significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

C. Discussion

With the addition of the 10-second deadline, a meaningful difference in accuracy between NL and CE emerges that was not seen in Experiment I. Overall, CE accuracy suffers more from the deadline than NL accuracy, and this is primarily driven by lower CE scores for complex statements, while CE simple statements appear to be on par with NL.

In addition, accuracy for complex statements for NL and, particularly, CE were impeded by time pressure. Decisions are frequently made under time pressure in the military and in other safety critical environments (e.g., healthcare, aviation). Consequently, for time-sensitive decisions, caution should be used with information represented via text statements for conjunction and negation.

While in Experiment I we saw lower accuracy for negation than conjunction in CE, we now see lower accuracy for conjunction than negation in NL. This is attributable to conjunction suffering more from time pressure than negation across languages, possibly due to the additional time needed to verify two propositions.

IV. CONCLUSION

A common representation of information may help facilitate human-machine interaction and collaboration. We assessed human understanding (accuracy and speed) for text statements written in natural language vs an artificial language, CE. Without time pressure (Experiment I), the NL and CE had similar accuracy, although NL was faster and assigned a higher usability score by participants. With time pressure (Experiment II), NL had higher accuracy and speed than CE and was assigned a higher usability score by participants. However, time pressure caused lower accuracy in both NL and especially CE.

These experiments suggest that CE, especially in its more complex forms, may be less reliable than NL for information transfer in heavily time-constrained situations. Lowered

comprehension accuracy in such situations must be weighed against the advantages of a CNL (e.g., machine readability). It may be possible, however, to redesign aspects of a CNL, such as CE's conjunction and negation, that have been identified as problematic through an investigation like the one presented in this paper (see [26] for comprehension comparisons of synonymous CNL statements). Care should be used, however, in differentiating CNL statements that simply require longer to read from statements that are difficult to comprehend without training. A limitation here is we did not examine highly complex statements (such as multiple [nested] Boolean operators).

For future work, this research illustrates that information understanding should be assessed using accuracy and speed. Furthermore, it suggests the importance of time pressure for understanding, revealing the challenges in interpreting findings in speed, accuracy, and their relationship: Low accuracy may be due to a statement simply taking too long to read, or the statement may be difficult to understand even with unlimited time.

ACKNOWLEDGMENT

We thank Paul Smart for sharing his knowledge of CNLs and pointing us to the relevant literature.

REFERENCES

[1] M. M. Cummings, "Man versus Machine or Man + Machine?," *Intell. Syst. IEEE*, vol. 29, no. 5, pp. 62–69, 2014.

[2] A. Preece, T. Norman, G. de Mel, D. Pizzocaro, M. Sensoy, and T. Pham, "Agilely Assigning Sensing Assets to Mission Tasks in a Coalition Context," *IEEE Intell. Syst*., vol. 28, no. 1, pp. 57–63, 2013.

[3] A. Preece, T. Norman, G. de Mel, D. Pizzocaro, M. Sensoy, and T. Pham, "Agilely Assigning Sensing Assets to Mission Tasks in a Coalition Context," *IEEE Intell. Syst.*, vol. 28, no. 1, pp. 57–63, 2013.

[4] D. Pizzocaro, C. Parizas, A. Preece, D. Braines, D. Mott, and J. Z. Bakdash, "CE-SAM: A conversational interface for ISR mission support," in *SPIE Defense, Security, and Sensing*, Baltimore, MD, 2013.

[5] K. Kawamoto, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *BMJ*, vol. 330, no. 7494, pp. 765–0, Apr. 2005.

[6] A. N. Garg AX, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review," *JAMA*, vol. 293, no. 10, pp. 1223–1238, Mar. 2005.

[7] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Hum. Factors J. Hum. Factors Ergon. Soc*., vol. 39, no. 2, pp. 230–253, Jun. 1997.

[8] L. G. Terveen, "Overview of human-computer collaboration," *Knowl.-Based Syst*., vol. 8, no. 2, pp. 67–81, 1995.

[9] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

[10] E. Zaroukian, J. Z. Bakdash, A. Preece, and W. Webberley, "Automation Bias with a Conversational Interface." *IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2017.

[11] T. Kuhn, "A survey and classification of controlled natural languages," *Comp Ling*, vol. 40, no. 1, pp. 121–170, 2014.

[12] N. E. Fuchs and R. Schwitter, "Specifying logic programs in controlled natural language," *ArXiv Prepr. Cmp-Lg9507009,* 1995.

[13] A. Preece, W. Webberley, D. Braines, E. G. Zaroukian, and J. Z. Bakdash, "SHERLOCK: Experimental evaluation of a conversational agent for mobile information tasks," *IEEE Trans. Hum.-Mach. Syst*., vol. 47, no. 6, pp. 1017–1028, 2017.

[14] P. Xue et al., "Information extraction using controlled english to support knowledge-sharing and decision-making," *Boeing Co Seattle WA Research and Technology*, 2012.

[15] A. Wyner et al., "On controlled natural languages: Properties and prospects," in *Controlled Natural Language,* Springer, 2010, pp. 281–289.

[16] D. Mott, "Summary of controlled English," *ITACS* https://www.usukitacs.com, 2010.

[17] T. Kuhn, "An evaluation framework for controlled natural languages," in *Controlled Natural Language,* Springer, 2010, pp. 1–20.

[18] T. Kuhn, "The understandability of OWL statements in controlled English," *Semantic We*b, vol. 4, no. 1, pp. 101–115, 2013.

[19] R. Schwitter, K. Kaljurand, A. Cregan, C. Dolbear, and G. Hart, "A Comparison of three Controlled Natural Languages for OWL 1.1.," in *OWLED (Spring*), 2008.

[20] A. Edland and O. Svenson, "Judgment and decision making under time pressure," in *Time pressure and stress in human judgment and decision making*, Springer, 1993, pp. 27–40.

[21] J. Brooke, SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor & Francis, 1996.

[22] T. Kuhn, Controlled English for knowledge representation. Doctoral thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich, 2010.

[23] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2013.

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv Prepr. ArXiv14065823*, 2014.

[25] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R^2 from generalized linear mixed-effects models," *Methods in Ecology and Evolution*, 4(2), 2013, pp. 133-142.

[26] E. Zaroukian, "Human Understanding of Controlled Natural Language in Simulated Tactical Environments," in *IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2016, 126-130.