Linguistic Repetitions, Task-based Experience and A Proxy Measure of Mutual Understanding

Justine Reverdy
Computational Linguistics Group
ADAPT Centre
School of Computer Science and Statistics
Trinity College Dublin, The University of Dublin
Ireland
Email: reverdyj@tcd.ie

Carl Vogel
Computational Linguistics Group
Trinity Centre for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin, The University of Dublin
Ireland
Email: vogel@tcd.ie

***Abstract*—The way dialogue partners collaborate to achieve a joint task is dependent on the way they construct a common ground of knowledge. Diverse conversational mechanisms are involved in developing a common ground, and repetition phenomena appear to be strongly connected to these processes. This article describes the use of an automatic method to detect, within dialogue transcripts, linguistic cues of engagement and synchrony, by observing repetitions at different linguistic levels. We focus on the relationship between repetition patterns and task-based success in interaction with task-based experience and partner familiarity. We conduct our analysis on the data of the HCRC Map Task corpus. Results suggests that, among other patterns, significant amounts of repetitions play a role for unfamiliar participants, with greater success, in particular, at first attempts.**

I. Introduction

A dense literature on human communication in dialogue agrees on the existence of a phenomenon of alignment — repetition of linguistic choices — as a sometimes unconscious mechanism underlying discourse structure. The Interactive Alignment Model (IAM) [1] has been taken as the basis of studies exploring alignment at different linguistic levels, for instance, lexical and syntactic [2], [3], [4] or phonetic realisations [5]. Interlocutors construct a mutual representation of the world [6] and establish a common ground [7] through verbal and non-verbal exchanges that could potentially lead them to mutual understanding. Whether mutual understanding can actually be achieved by interlocutors can never be established for a certainty; however, interlocutors can achieve a state in which they lack direct evidence of misunderstanding [8]. Further, in situations in which interlocutors desire the communication to feel successful, they may use repetition effects as evidence of mutual understanding through normal confirmation bias. Therefore, repetition may be a signal of the degree to which interlocutors think they have understood each other.[1] If dialogue participants think that they have understood each other, they may well have; however, if they think that they have not understood each other, then it is an analytic truth that they have not. Regarding the default hypothesis of mutual understanding among people speaking a shared language, one may make the optimistic assumption that they have understood each other or the pessimistic assumption that they have not. In the present study, we adopt the null hypothesis as described by Vogel [10, pp. 384]: unless a significant amount of communicative cues, like repetition, are evident, mutual understanding cannot be reliably asserted. The safest null hypothesis about a communication act is that it has not succeeded in achieving mutual understanding.

Repetitions are frequent within communicative behaviours and possess multiple functions [11]. In addition to sometimes being consciously used to diminish chances of miscommunication [12], they can signal engagement or involvement in an interaction. Involvement can be seen as entailing a desire for mutual understanding to be achieved. Conversely, in a exploration of discourse strategies, it has been asserted that understanding is dependent upon conversational involvement [13]. In 2012, [14] Beňus et al. explored entrainment of acoustic features. They examined filled pauses between lawyers in Supreme Court hearings and Justices, in relation with the favorability of the Justice vote. They found that when above chance adjacent filled pauses (pauses contained in previous contribution) occurred between a lawyer and a Justice, it related to more favorable decision of the Justice for the case being discussed. Whereas when observing over-all dialogue filled pauses means, no relation with favorability was found. Hence, their findings support the hypothesis of short-term accommodation having an effect on communication.

By measuring repetitions in dialogues at different linguistic levels, a degree of involvement can be estimated, and we try to assess whether the measures of repetition can be linked to estimates of mutual understanding. To assess the relation between synchrony and mutual understanding, we chose in this study to tie mutual understanding to the notion of task success in a collaborative task. We used HCRC Map Task corpus [15], described in § II-A. The Map Task makes random success unlikely, and non-random success achievable mainly through collaborative dialogue. This corpus was chosen because it contains transcripts of collaborative task oriented dialogue and meta-data regarding the conditions assigned to each participant, and equally importantly for us, a measure of success in the underlying task achieved in each dialogue. The conditions included alternation between dialogue roles (across dialogues) as well as features such as familiarity between participants.

[1] Consider the case of clarification questions (which include reprise fragments): the fact that a question is asked strongly suggests a lack of shared meaning (yet, note that these rather often go unanswered [9]), but equally strongly (particularly in the case of reprise fragments) suggests that the interlocutors agree on what has been uttered.

Over the past decade, Reitter et al. [3], [16], [17], with an emphasis on syntactic analysis (phrase-structure), have undertaken studies relating task success measures given in the HCRC Map Task corpus and proportion of repetitions in specific time window. With their method, they did not find direct evidence that the short-term priming effects they defined correlated with task-success. Yet, they established a link between repetition proportions (long-term priming effects) and task success using different linguistic features, which allowed them to capture more variation in repetition structures.

Instead of exploring specific temporal durations for possible repetitions, an alternative is to construe dialogue structurally. In relation to a speaker's current turn, one may note the speaker's own immediately prior turn and the last turn of each other dialogue participant. Given the flow of multi-party dialogue, one speaker's contribution may be their last contribution for a longer period than the temporal duration of interest. Of course, one may also consider the speaker's and other interlocutors' prior *n*-turns, and boundary conditions can be identified for which the methods are the same, but in general, they are different. The structural construal of repetitions has been explored before [18], [10], and scrutiny of the HCRC Map-Task data has addressed levels of linguistic representation at which repetition may occur [19]. We adopt this structural approach to analysing dialogues, using memory registers for each interlocutors' last contribution, and comparing the content of the speaker's current utterance to the contents of each register (rather than within a specific temporal distance).

Our approach uses a technique also adopted by Ramseyer et al. [20], [21], taking measurements of repetition in actual dialogue and comparing those with measurements taken from random shufflings of the turns of the actual dialogues. We explore Self-Repetition and Other-Repetition. The former has an interpretation in maintenance of a rolling dialogue plan and the latter, in grounding and engagement [10]. Our analyses focus on cases where repetition in the *Actual* dialogue exceeds that of *Randomized* counterparts. It would also be interesting to understand cases in which *Actual* repetition is significantly *less* than that of *Randomized* counterparts, but in our initial explorations, based on notions of engagement, grounding and miscommunication risk minimization, as discussed above, our focus is here in the dialogues where *Actual* repetition significantly exceeds *Randomized*, and thus hypothetically signal alignment and mutual understanding.

Where communication is in task-related settings, it is tempting to take task-based success as an indication of communication success. However, it must be granted that joint tasks may often be successfully completed (salt passed, windows closed, doors opened, etc.) without successful information sharing in support of the task having been achieved through dialogue. Individuals frequently improve at tasks with practice. Therefore, it is interesting to test whether signals of communication success correlate with task-based success while taking into account task-based experience. We focus on the relationship between repetition patterns and task success, taking into account experience with the task itself (participants each have four attempts at the task over the course of the experiment) and interlocutor familiarity (in two of their four attempts, participants interact with a friend and in the other two, with someone previously unknown to them). In her practical method to operationalize behavioural dialogue principles, Davies [22, p. 48] asserted both that "Speakers will improve at tasks" under the Principle of Gricean cooperation [23], and "Task success will improve (as speakers negotiate trade-off more successfully)" under the principle of Parsimony. We therefore expect an improvement in the task over time. Our hypothesis is that if short-term repetition plays a role in communication in relation to task-success, then distinctive patterns will appear interacting with linguistic features of repetitions and non-linguistics features of Experience and Familiarity. In particular, where interlocutors are not familiar with each other, then we expect the presence of significant amounts repetition in dialogue to correlate with task-based success. Similarly, where interlocutors are not familiar with their task, we expect significant amounts of repetition to correlate with task-based success.

II. Method

A. Data Set

Released in 1992 [15], the Human Communication Research Centre (HCRC) Map Task corpus contain 128 dialogues of Human-to-Human task based interactions. Two participants per dialogue are requested to communicate to achieve a map task, one having the role of the Information Giver (IG) in charge of guiding an Information Follower (IF) in reproducing a path on a map containing various landmarks. The map of the IG would have a specific route drawn, while the IF's contained only landmarks, landmarks that slightly differed to add to the difficulty of the task (and elicit linguistic phenomena, like contrastive focus). The IF and IG could not see each other's map. However, half of the participants were assigned to a condition in which eye contact was possible, while the other half could not see each other. Each subject ($n = 64$) participated in four dialogues in total, twice as IG and twice as IF, and in each role once with a familiar partner and once with an unfamiliar one. The IF used on average 393.31 tokens per dialogue and the IG, 858.10.

The corpus designers computed *Deviation from path* scores[2] (deviation score). The higher the score, the more the path drawn by the IF had deviated from the original route given on the IG's map, which is assumed to be the sign of a less successful communication. The scores are described as the centimetre square difference between the map of the IG and the IF, having the map divided into a 1 centimetre square grid. The HCRC Map Task corpus deviation scores, which this study uses, ranges from 4 (best) to 227 (worst). The corpus was recorded entirely in English; all participants were students of the University of Glasgow, 61 of them from Scotland.

B. Analysis by conversations

1) Base Method: The base method first described in [18] consists of counting the repetition of tokens of a contribution and the immediately preceding contribution of each participant in the dialogue. A count up to a length of $n = 5$, *n*-grams, is made of each repetition, for other-repetitions (repetition of a token uttered by another participant) and self-repetitions (repetition of a token uttered by the same participant). Once the count is made in the actual dialogue, each contribution is indexed and

[2]http://groups.inf.ed.ac.uk/maptask/maptask-description.html (Last consulted: 07/08/2017)

randomly shuffled within each dialogue, and a count is made again in the randomly re-ordered dialogue. Those re-orderings and countings are made ten times, to observe if a significant contrast emerges between the actual dialogues and the shuffled ones in repetition counts. The focus is on the proportion of the total number of *n*-grams that could have been shared but were not (NON-OTHERSHARED, NON-SELFSHARED) and the ones that were repeated (OTHERSHARED, SELFSHARED), both in actual and randomised dialogues. For more detailed descriptions, see [10] and [18].

2) Extended Method: The method was extended in [19] with the aim of exploring more levels of linguistic description than the tokens as transcribed from the dialogue (lemmas and part-of-speech labels were considered alongside tokens, and sequences thereof) and doing so in the context of the HCRC Map Task data in order to relate repetition effects to task success (and other variables controlled in the maptask experiment [15]). It was extended for two reasons: first to observe the scope in which different linguistic levels of repetitions provide information reliably as indicators of synchrony within the frame of the base method, and second, to which extent success in communication is associated with repetitions. The extension described below is retained in this study.

The extension consists of a pre-processing labelling designed to measure five linguistics type of repetitions (referred to as 'Levels'): Token (which was the only unit previously analysed), Lemma, Part-Of-Speech (POS), and a combination of Token with POS and Lemma with POS. We labelled the HRCR Map Task transcripts with the default version of the TreeTagger as trained for English [24]. For each dialogue, proportions of repetitions were extracted, per Dialogue type (*Actual* versus *Randomised*), per speakers (IF: Information Follower and IG: Information Giver), per *n*-grams (All *n*-grams [up to length 5]; N1: [*n*-grams, $n = 1$]; N2+: [*n*-grams, $1 < n \leq 5$]), per type of sharing (OTHERSHARED and SELFSHARED), and per Level: TOKEN (Level 1), LEMMA (Level 2), LEMMA+POS (Level 3), POS (Level 4), TOKEN+POS (Level 5). As we observed in [19], although the method is not designed directly to look at syntactic repetitions, the POS labelling allows us to observe two different form of repetitions; lexical categories for N1 and structural repetitions for N2+ in combination with Level 4 (POS).

C. Hypothesis

To explore the influence of the variables (Dialog-Type, Speaker, Level) depending on the type of repetitions (OTHERSHARED and SELFSHARED) we computed single-step Tukey HSD (honest significant difference) multiple comparison tests using a general linear model with a binomial error family [25]. We tested the following hypothesis for each level:

$$H_0 : Random.Speaker.Level - Actual.Speaker.Level \geq 0$$
$$H_1 : Random.Speaker.Level - Actual.Speaker.Level < 0$$

This H_0 hypothesis states that if repetitions are due to chance in a dialogue, the difference between the proportion of repetition should be equal (or exceed) in the randomised dialogues than in the actual dialogues. While if they are happening significantly more in actual dialogues (H_1), a functional role for repetition in the communication could be assumed.

D. Meta Analysis across the Conversations

The Tukey's tests were performed on each dialogue, resulting in 1280 comparisons of the three variables against the two repetition type (OTHERSHARED, SELFSHARED), including all *n*-grams, $1 \leq n \leq 5$. We opted for a threshold of $p \leq 0.05$, dividing the results of the tests into a factor (SIGNIFICANTREPETITION) for each of OTHERSHARED and SELFSHARED with levels TRUE and FALSE (TRUE: $p \leq 0.05$, the null hypothesis was rejected; FALSE: $p > 0.05$, the null hypothesis was not rejected). This factor distinguishing the dialogues where repetitions happened significantly more than chance would lead one to expect is the basis of our meta-analysis, and the variable against which the non-linguistic features of Experience and Familiarity of the Map Task corpus are tested.

E. Summary

For each dialogue, the method described above established a proxy measure of mutual understanding (whether repetition levels exceeded chance, leading to H_0 rejection, $p \leq 0.05$: this yields the meta-analysis factor: SIGNIFICANTREPETITION with levels TRUE or FALSE, TRUE corresponding to H_0 rejection in the underlying dialogues). Within each dialog, this measure is given per speaker, linguistic level and *n*-gram lengths. We expect that a significant amount of repetition (TRUE) will correlate with greater task success. This study focuses on how this measure interacts with Experience and Familiarity within task success, success represented by the Deviation Score. As we established in previous study that familiar pairs had lower path deviation scores than unfamiliar pairs, and one may expect that experience will positively increase success over task attempt, we asked these questions:

- Do measurements at different linguistic levels impact the detection of significant amounts of repetition in the first attempt at a dialogue task?
- Do familiar partners display alignment during first task attempts?
- Do proxy measures of repetition differ for familiar and unfamiliar partners in their first attempts and in correlation with task success measures?

To answer those questions in relation to the counts and proportions we have described, we use non-parametric tests, namely, Mann-Whitney-Wilcoxons for distribution differences, Chi-square, and we also rely on Pearson's standardized residuals from log-linear models.

III. RESULTS

A. Overview

Following the threshold of ($p \leq 0.05$), the Null Hypothesis was rejected 902 times for OTHERSHARED and 281 for SELFSHARED, for all *n*-grams, which shows that across all variables, there was a much higher proportion of significant OTHERSHARED repetitions in that task-based corpus. Table 1 highlight the higher rate of rejection for OTHERSHARED than SELFSHARED, for both Information Giver (IG) and Information Follower (IF).

TABLE I: Rejections of H0 for OtherShared, in relation to roles (IF:Information follower;IG:Information Giver) and means of rejections by roles. In each case the Null Hypothesis can potentially be rejected 128 times

All n-grams (OtherShared)

$H_0 : Random.Speaker.Level - Actual.Speaker.Level \geq 0$

Level	1	2	3	4	5	Mean
IF	112	109	109	82	107	103.8
IG	88	87	80	47	81	76.6

All n-grams (Selfshared)

$H_0 : Random.Speaker.Level - Actual.Speaker.Level \geq 0$

Level	1	2	3	4	5	Mean
IF	36	35	37	19	38	33
IG	27	26	30	5	28	23.2

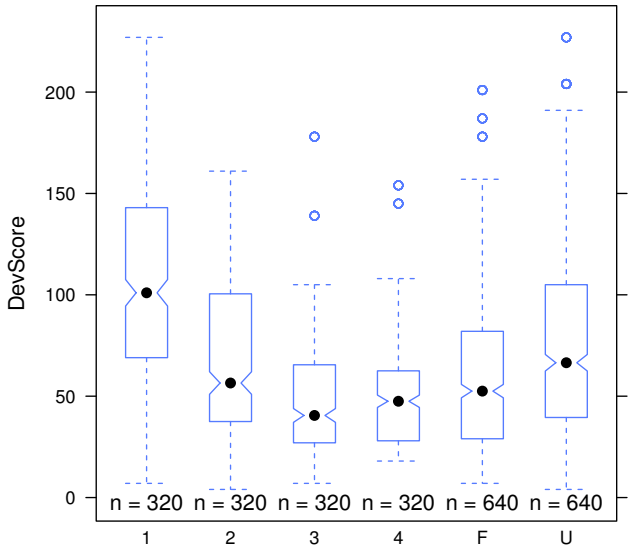

Fig. 1: Distribution of Deviation Score by Experience (Attempt 1,2,3,4), along with Familiarity (U: Unfamiliar|F: Familiar)

Figure 1 shows the importance of Experience on the Deviation score, the first attempt having the highest average Deviation Score ($\bar{x} = 109.4$) by far in comparison to the next three attempts (Second: ($\bar{x} = 69$), Third: ($\bar{x} = 54.2$), Fourth: ($\bar{x} = 54.5$)). This phenomenon is also visible in Figure 2, with the first attempt displayed in the darkest shade of grey. We also note that for OTHERSHARED, SIGNIFICANTREPETITION was TRUE 27 times for familiar participants and 28 times for unfamiliar, and only FALSE 5 and 4 times respectively. Figure 1 also shows the difference in Deviation Score between familiar and unfamiliar pairs of participants. A Mann-Whitney-Wilcoxon test for population distribution found significant difference in Deviation Score between familiar and unfamiliar partners (W= 6572, $p = 0.00625$).

B. Linguistic Levels

We observe that for Level 1 (Token only), no significant differences between SIGNIFICANTREPETITION=TRUE and SIGNIFICANTREPETITION=FALSE within the Deviation Score during the first attempt was found, neither OTHERSHARED (W= 78.5, $p = 0.45$) nor SELFSHARED (W= 73.5, $p = 0.14$). We also tested if significant differences between

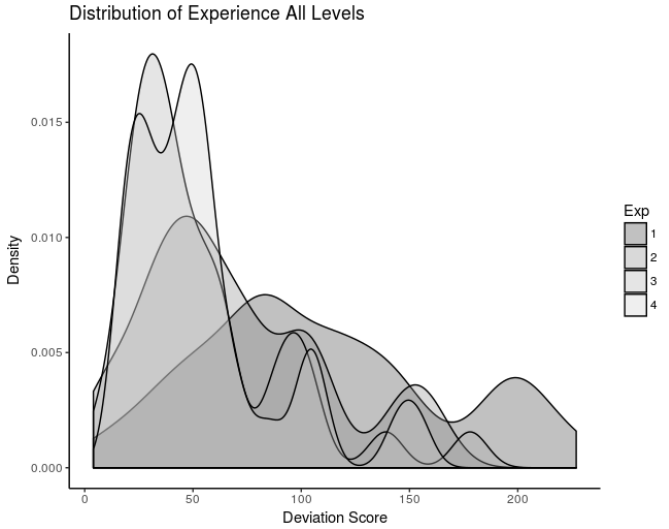

Fig. 2: Density plot of Deviation Score per Experience (By grey shading, First Attempt: Dark grey to Fourth Attempt: Light grey). For each distribution n = 32

SIGNIFICANTREPETITION=TRUE and SIGNIFICANTREPETITION=FALSE within the Deviation Score appeared for each linguistic level in isolation and none was detected. However, when testing with all linguistics Levels (TOKEN (Level 1), LEMMA (Level 2), LEMMA+POS (Level 3), POS (Level 4), TOKEN+POS (Level 5) in combination, a significant difference was found (W= 7015.5, $p = 0.03$), correlating TRUE to a lower Deviation score ($\bar{x} = 105$) than FALSE ($\bar{x} = 122.59$) and thus higher success. Then we observed if the first attempt being an Information Giver (IG) or an Information Follower (IF) had an impact on success following significant amount of repetition detected at all linguistic levels. It was the case for OTHERSHARED ($p = 0.02$) and SELFSHARED ($p = 5.035e-05$) IG, and for SELFSHARED ($p = 0.031$), but not for OTHERSHARED ($p = 0.45$) IF.

C. Familiarity and Experience

The association plots in Figure 3 (OTHERSHARED) and 4 (SELFSHARED) display the relationship between the sum of Deviation Scores, Familiarity and SIGNIFICANTREPETITION detected at all linguistic Levels. The Pearson's standardized residuals point out that for Unfamiliar pairs where SIGNIFICANTREPETITION=TRUE, the observed value is under the expected value and for Familiar pairs, the observed value is above the expected value. Those figures display the sum of Deviation Scores: results under the baseline for independence imply greater than expected task success (in correspondence with lower than expected Deviation Score sums).

For both cases Chi-square tests indicate the association present between the variables (OTHERSHARED: $p = 8.7316e-15$; SELFSHARED: $p =< 2.22e-16$). This might indicate that the repetitions levels detected by the method have a higher impact on Unfamiliar pairs task success than on Familiar pairs.

Figure 5 (First Attempt), and Figure 6 (Attempts 2 to 4) display the distribution of deviation score between Familiar and Unfamiliar pairs. We observe that if Unfamiliar pairs

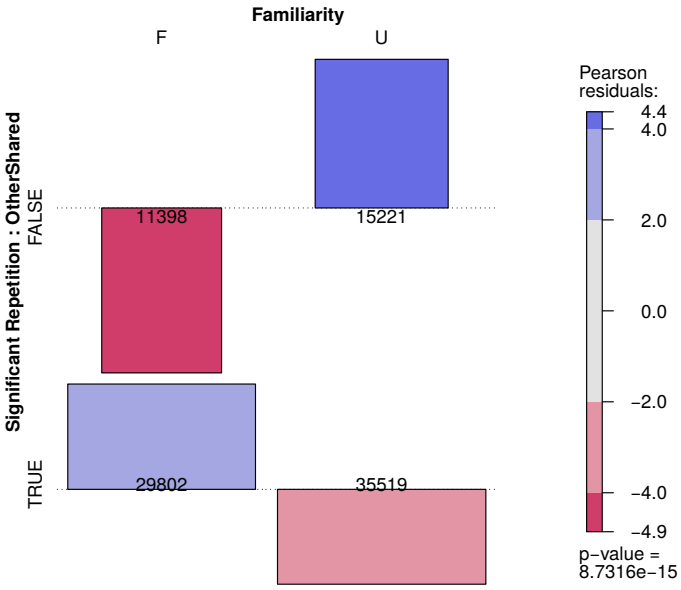

Fig. 3: Association Plot of OTHERSHARED SIGNIFICANTREPETITION (TRUE|FALSE) and Familiarity (U: Unfamiliar|F: Familiar)

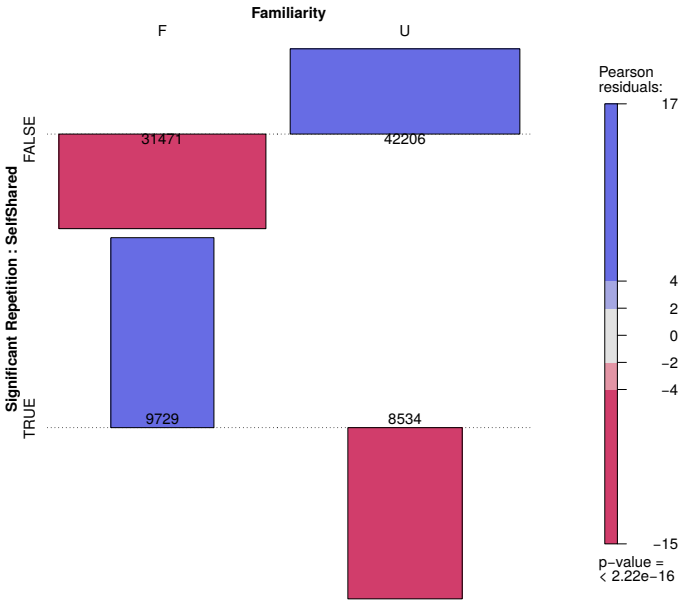

Fig. 4: Association Plot of SELFSHARED SIGNIFICANTREPETITION (TRUE|FALSE) and Familiarity (U: Unfamiliar|F: Familiar)

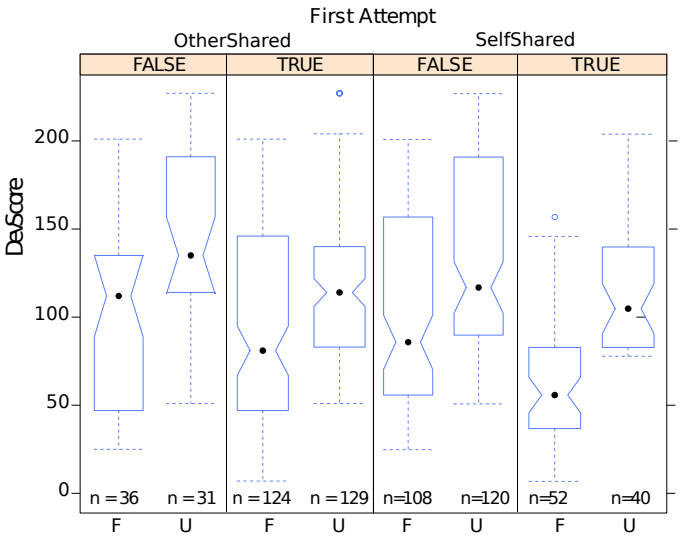

Fig. 5: Distribution of Deviation Score for First Attempt in interaction with Familiarity (U: Unfamiliar|F: Familiar) for the two values of SIGNIFIANTREPETITION (TRUE|FALSE) for OTHERSHARED and SELFSHARED

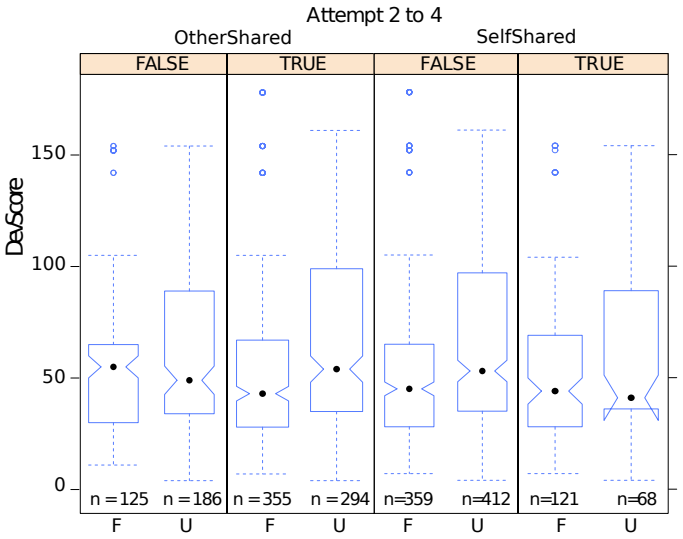

Fig. 6: Distribution of Deviation Score for Attempt 2 to 4 in interaction with Familiarity (U: Unfamiliar|F: Familiar) for the two values of SIGNIFIANTREPETITION (TRUE|FALSE) for OTHERSHARED and SELFSHARED

seem to have on average always a higher deviation score in all conditions, the distinction between SIGNIFICANTREPETITION=TRUE and FALSE is clearly observable at First Attempt. For OTHERSHARED SIGNIFICANTREPETITION, no significant difference was found between familiar pairs at First Attempt ($p = 0.106$), however, a difference was found for Unfamiliar pairs ($p = 0.039$), with TRUE having a lower mean ($\bar{x} = 123.41$) than FALSE ($\bar{x} = 141.29$). A significant difference was found at second attempt for Familiar pairs ($p = 0.004$), with TRUE having a lower mean ($\bar{x} = 68.19$) than FALSE ($\bar{x} = 93.74$), but not for Unfamiliar pairs ($p = 0.106$). A combination of Deviation Score distribution from attempt 2 to 4 was found significant for both Familiar and Unfamiliar pairs, with TRUE having a lower mean ($\bar{x} = 53.42$) than FALSE ($\bar{x} = 60.03$) for Familiar pairs, and with FALSE having a lower mean ($\bar{x} = 58.43$) than TRUE ($\bar{x} = 66.97$).

For SELFSHARED, no significant difference was found between TRUE and FALSE, except for First Attempt of Familiar pairs ($p = 1.393e-05$), with TRUE having a lower mean ($\bar{x} = 63.03$) than FALSE ($\bar{x} = 105.85$).

As we stated in § III-B, among all levels and participants, significant differences in Deviation Score distribution

are observed between SIGNIFICANTREPETITION=TRUE and FALSE for both OTHERSHARED and SELFSHARED, in First Attempt in isolation. Those distinctions are highlighted when compared to attempts 2 to 4. Once the first attempt happened, the differences between SIGNIFICANTREPETITION=TRUE and FALSE for familiar and unfamiliar pairs becomes more difficult to interpret, as the relation is not always when SIGNIFICANTREPETITION=TRUE correlated to higher success.

IV. DISCUSSION

Returning to the questions posed in §II-E, the meta-analysis shows that the effects described in §III hold when all linguistic levels are taken into account but not each level in isolation. It remains to assess the extent to which this follows from individual repetition events being counted at *each* level of linguistic representation versus repetitions counting as such at some levels without simultaneous manifestation at others. The First Attempt, when participants discover the task, represents the closest observation of an untrained pair of participants in real task-solving conditions. Therefore, associated outcomes in this study are particularly interesting. The fact that both Familiar and Unfamiliar partners display a high level of TRUE significant repetition for OTHERSHARED during the First Attempt is a sign of alignment. In the First Attempt, Unfamiliar partners who repeat each other to a significant degree (summing across levels of linguistic representations), and thus align to their partner, have greater levels of task success than Unfamiliar partners without a significant degree of repetition. Although both Familiar and Unfamiliar pairs align to each other, alignment does not correlate with task-success at the first attempt for Familiar pairs, in contrast to Unfamiliar pairs. However, familiar pairs with significant self-repetition in the First Attempt, compared to familiar pairs without significant self-repetition, achieved greater task-success.

V. CONCLUSION

Both Familiar and Unfamiliar partners display alignment, but it is related to task-success with statistical significance for Unfamiliar pairs, in particular at first attempts at the task. The patterns highlighted here have promise as a step towards quantifying engagement and mutual understanding using the automatic method described. Further exploration is needed to establish repetition's relevance in other languages and possible application to computer-mediated interactions and dialogue systems. Those possible uses make the study of the concept of alignment particularly relevant to the CogInfoCom line of research as it links cognitive science and linguistics [26] with likely use of the method in speech technologies. This is another step toward technologies for inter-cognitive communication.

ACKNOWLEDGMENT

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

REFERENCES

[1] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 02, pp. 169–190, 2004.

[2] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic coordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[3] D. Reitter and J. D. Moore, "Predicting Success in Dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 808–815.

[4] S. Garrod and A. Anderson, "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination," *Cognition*, vol. 27, no. 2, pp. 181–218, 1987.

[5] H. Giles, J. Coupland, and N. Coupland, *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, 1991.

[6] W. Turnbull, *Language in action: Psychological models of conversation*. Psychology Press, 2003.

[7] H. H. Clark and S. E. Brennan, *Grounding in communication*. Washington, DC, US: American Psychological Association, 1991, pp. 127–149.

[8] T. J. Taylor, *Mutual Misunderstanding: Scepticism and the theorizing of language and interpretation*. Duke University Press, 1992.

[9] M. Purver, P. G. T. Healey, J. King, J. Ginzburg, and G. J. Mills, "Answering clarification questions," in *SIGDIAL Workshop*, 2003, pp. 23–33.

[10] C. Vogel, "Attribution of Mutual Understanding," *Journal of Law and Policy*, vol. 21.2, pp. 377–420, 2013.

[11] D. Tannen, *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge University Press, 2007, vol. 26.

[12] S. Cushing, *Fatal Words: Communication Clashes and Aircraft Crashes*. University of Chicago Press, 1994.

[13] J. J. Gumperz, *Discourse strategies*. Cambridge University Press, 1982.

[14] Š. Beňuš, R. Levitan, and J. Hirschberg, "Entrainment in spontaneous speech: the case of filled pauses in supreme court hearings," in *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*. IEEE, 2012, pp. 793–797.

[15] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[16] D. Reitter, J. D. Moore, and F. Keller, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proceedings of the Cognitive Science Society*, 2006, pp. 685–690.

[17] D. Reitter and J. D. Moore, "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, pp. 29–46, 2014.

[18] C. Vogel and L. Behan, "Measuring Synchrony in Dialog Transcripts," *Cognitive Behavioural Systems. Lecture Notes in Computer Science*, vol. 7403, pp. 73–88, 2012.

[19] J. Reverdy and C. Vogel, "Measuring synchrony in task-based dialogues," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20–24, Stockholm, Sweden, Proceedings*, 2017.

[20] F. Ramseyer and W. Tschacher, "Synchrony: A core concept for a constructivist approach to psychotherapy," *Constructivism in the human sciences*, vol. 11, no. 1, pp. 150–171, 2006.

[21] D. Reidsma, A. Nijholt, W. Tschacher, and F. Ramseyer, "Measuring multimodal synchrony for human-computer interaction," in *Cyberworlds (CW), 2010 International Conference on*, Oct 2010, pp. 67–71.

[22] B. L. Davies, "Testing dialogue principles in task-oriented dialogues: An exploration of cooperation, collaboration, effort and risk," *Leeds Working Papers in Linguistics and Phonetics*, vol. 11, pp. 30–64, 2006.

[23] H. P. Grice, P. Cole, J. Morgan *et al.*, "Logic and conversation," *1975*, pp. 41–58, 1975.

[24] H. Schmid, "Probablistic part-of-speech tagging using decision trees," in *Proceedings of The First International Conference on New Methods in Natural Language Processing (NemLap-94)*. Manchester, U.K., 1994, pp. 44–49.

[25] F. Bretz, T. Hothorn, and P. Westfall, *Multiple comparisons using R*. CRC Press, 2016.

[26] P. Baranyi and A. Csapo, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.