Measuring Synchrony in Task-based Dialogues

Justine Reverdy, Carl Vogel

ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland

reverdyj@tcd.ie, vogel@cs.tcd.ie

Abstract

In many contexts from casual everyday conversations to formal discussions, people tend to repeat their interlocutors, and themselves. This phenomenon not only yields random repetitions one might expect from a natural Zipfian distribution of linguistic forms, but also projects underlying discourse mechanisms and rhythms that researchers have suggested establishes conversational involvement and may support communicative progress towards mutual understanding. In this paper, advances in an automated method for assessing interlocutor synchrony in task-based Human-to-Human interactions are reported. The method focuses on dialogue structure, rather than temporal distance, measuring repetition between speakers and their interlocutors last n-turns (n = 1, however far back in the conversation that might have been) rather than utterances during a prior window fixed by duration. The significance of distinct linguistic levels of repetition are assessed by observing contrasts between actual and randomized dialogues, in order to provide a quantifying measure of communicative success. Definite patterns of repetitions where identified, notably in contrasting the role of participants (as information giver or follower). The extent to which those interacted sometime surprisingly with gender, eye-contact and familiarity is the principal contribution of this work.
Index Terms: alignment, human-human interaction, computational linguistics, repetitions, dialogue structure

1. Introduction

Here we focus on one aspect of dialogue structure, the repetition of linguistic choices as cues of an alignment process. Numerous studies have repeatedly given evidence that the phenomenon of alignment is a strong component of communicative success — notably suggested by [1] as an unconscious process in their the Interactive Alignment Model. This tendency toward alignment can be observed at many levels, such as syntactic and lexical [2, 3, 4], or in terms of speech rate or phonetic realisations [5]. According to these studies, interlocutors tend to align their representation of the world using different linguistic and non-linguistic strategies [6] to construct mutual understanding throughout the conversation [7]. Among these strategies, repetition mechanisms play a crucial role, and holds multiple functions [8] in a process that leads interlocutors to establish a common ground [9]. Those mechanisms can also take place to avoid miscommunication, and as indicator of involvement or engagement in an interaction. Gumperz [10], by exploring discourse strategies, argued that understanding depends on conversational involvement. Reidsma et al. [11] automatically calculate synchrony between speakers motion by using a time-lagged cross-correlation technique from [12]. In their method, they randomly shuffled interaction measures and compared them to the actual level of the synchrony measure to assess whether the actual levels were higher than what would be considered by chance. A similar Monte Carlo approach is adopted here. Our hypothesis is that, by measuring repetition in dialogues, a degree of involvement can be assessed, which we assume relates to mutual understanding. Dialogues display internal structures where different communication strategies are used with the usual aim to achieve mutual understanding. However, a pessimistic view of this achievement, is that it can not be formally proven [13] is assumed. The existence of a null hypothesis as described by Vogel [14, pp. 384] in communication is therefore adopted. As the certainty of mutual understanding could never be established, the hypothesis that communication was not achieved unless a certain amount of specific cues were detected within the dialogue is the basis of the method [15, 14] that is expended here. In order to measure the degree of syncrony or mutual understanding, we chose, in this study, to link these concepts to the notion of task success. To explore this notion in relation with repetition structures, we used the HCRC Map Task corpus, described in § 2.1, that in addition to the deviation score, which is a measure of task success of the map task, contains non-linguistic features, such as gender, eye-contact and familiarity, that play a crucial role in the distribution of repetitions. Concerning Gender distribution, Branigan et al. [16] found male subjects to be overall more disfluent than female subjects, with repetitions included in their definition of disfluency. Also Colman et al. [17] showed that repair mechanisms are approximately double in task-oriented dialogues than in everyday conversations. Colman et al. additionally pointed out that "two different task roles are associated with divergent patterns of repair" [17, p. 1567]. Those patterns may be considered through the scope of who is holding information in an interaction. In everyday interaction it could be considered that participants hold information by turns, while in a task-based map task interaction this distinction is maintained thorough the course of the dialogue, hence our hypothesis is that if short-term repetition play a role in communication in relation with task-success, then distinctive patterns can appear in interacting linguistic features of repetitions and non-linguistics features of role, gender, eye-contact and familiarity.

2. Method

2.1. Data Set

The Human Communication Research Centre (HCRC) Map Task corpus consist of 128 dialogues released in 1992 [19]. This corpus uses the map task technique to elicit spontaneous communicative behaviours in the frame of Human-to-Human task based interactions. Two subjects per dialogue, almost all participants were native Scottish speakers of English, with either the role of Information Giver (IG) or Information Follower (IF) were given each A3 maps containing landmarks. The IG had a route drawn on the map with a START and a FINISH, and was tasked with guiding the IF through a map containing only landmarks. To add to the difficulty of the task, landmarks from the two maps and their placement differed a little. The subjects could not see their interlocutor's map at any point. The

settings of the recordings were divided into two, with half the subjects being able to see their interlocutor's face (i.e., with eye-contact), while the other half had screens placed between them (i.e., without eye-contact). The IF used on average 393.31 tokens per dialogue and the IG 858.10. The participants were 64 in total (32 females, 32 males), and would participate in the task four times, twice as IG and twice as IF, and in each role once with a familiar partner and once with an unfamiliar one. As the IG has to guide the IF along a predefine route, any deviation from that route were assumed to be the result of less successful communication between the two participants, as the subjects' were precisely told not to stray from the route. *Deviation from path* scores (deviation score) were then computed by the authors of the corpus as a measure of task success.[1]

They 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 score, which this study uses, ranges from 4 (best) to 227 (worst). The higher the score, the more the route deviate from the original route, which is taken as an indication of less successful communication.

2.2. Analysis by conversations

2.2.1. Base Method

The base method [15, 14] consist of counting the repetition of tokens of a contribution and the immediately preceding contribution, assimilated as a dialogue turn of each speaker. A REGISTER is created for each participant containing his most recent contribution. A count is made of each repetitions of a token into the REGISTER, for other-repetitions (repetition of a token uttered by another participant) and self-repetitions. This count is made up to a length of $n = 5$, n-grams. The turns are then assigned a time-stamp and then randomly re-ordered (but without reordering within a turn) ten times and the other-repetitions and the self-repetitions are counted again for each re-ordered dialogue, with the intent to observe if a significant contrast between the actual dialogues and the shuffled ones emerge. The focus is on the ratio of the total number of n-grams that could have been shared (NON-OTHERSHARED, NON-SELFSHARED) and the ones that were repeated (OTHERSHARED, SELFSHARED), both in actual and randomised dialogues. Detailed descriptions are given in [15, 14]. Results observed of previous use led notably to highlight the importance of social role in conversations, that was suggested to prime over the individual personality in task-based interactions.

2.2.2. Extended Method

We are working toward providing a confident measure that mutual understanding was reached or not, thus quantifying a degree of mutual understanding. We have extended the method for two reasons. We are particularly interested by the scope in which different linguistic levels of repetitions provide information reliably as an indicator of synchrony, and secondly, to which extent success in communication is associated with repetitions, and how are those link articulated. Previous works by Reitter et al. [3] have taken an interest in measuring repetitions in relation with task accomplishment, focusing particularly on phrase-structure, i.e. syntactic analysis and proportion of repetitions within a short time window. As they stated, the repetitions due to the priming effect and due to a specific topic in a conversation are difficult to distinguish in an automated method. Using their method, they stated that short-term priming effects, while being present in the corpus, did not correlate with task-success. Nonetheless using different linguistic features rather than focusing on token level allowed them to look at more variation in repetition structure, and is the motivation of this extension, while adopting a different approach. The interest here is to assess the contrasting proportion of repetition in *Actual* versus *Randomised* dialogues, rather than in a specific temporal distance. The extension consist of a preprocessing 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. To keep uniformity with previous works, we labelled the HRCR Map Task with the default English training set of the TreeTagger [20]. The aim is to observe the additional information given by these different levels of repetitions in interaction with the other variables described below. 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=1 [length 1]; N2+: n-grams>1 [length 2 to 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). Comparing the proportions of repetitions of Tokens vs. Lemmas was of particular interest to us, as we might expect a different distribution of repetitions for this conventional representation of lexemes. As only repetitions of the exact same token were taken into account in previous uses, observing a contrastive effect in repetitions in variations of the same lexeme might provide a higher qualitative observation of the repetitions. Tannen [8] distinguish instances of repetitions along a fixity scale. Repetitions of lexemes can be viewed as midway between a scale that goes from exact same word to paraphrasing an idea. This allow us to capture variations and inflections in repetitions. While it might not be considered as a method designed to look at syntactic repetitions, the POS labelling allow us to observe two different form of repetitions; lexical categories for N1: n-gram=1, and structural repetitions for N2+: n-gram>1 in combination with Level 4(POS).

2.3. Hypothesis

To explore the influence of the variables (DialogType, 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 [21]. Therefore, we tested the following hypothesis:

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

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 potential role in the communication could be assumed. For an even more fine grained observation, the Tukey's tests were also made for:

N1: n-gram=1:

$$H_0 : Random.Speaker.Level.N1 - Actual.Speaker.Level.N1 \geq 0$$
$$H_1 : Random.Speaker.Level.N1 - Actual.Speaker.Level.N1 < 0$$

As well as N2+: n-gram>1:

[1] http://groups.inf.ed.ac.uk/maptask/maptask-description.html (Last consulted: 20/03/2017)

H_0 : *Random.Speaker.Level.N2+ − Actual.Speaker.Level.N2+* ≥ 0
H_1 : *Random.Speaker.Level.N2+ − Actual.Speaker.Level.N2+* < 0

2.4. 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), first including all *n*-grams , then for *n*-gram=1, and finally for *n*-gram>1. We opted for a threshold of ($p \leq 0.05$), dividing the results of the tests into a factor TRUE or FALSE (TRUE: $p \leq 0.05$, the null hypothesis is rejected; FALSE: $p > 0.05$, the null hypothesis was not rejected). This factor distinguishing the dialogues were repetitions happened above chance is the basis of our meta-analysis, and the variable against which the non-linguistic features of the map task corpus are tested.

3. Results

3.1. Overview and roles

Following a 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: *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* ≥ 0

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

N1: *n*-gram=1 (OtherShared)
H_0 : *Random.Speaker.Level.N1 − Actual.Speaker.Level.N1* ≥ 0

Level	1	2	3	4	5	Mean
IF	78	78	74	46	75	70.2
IG	49	47	51	18	54	43

N2+: *n*-gram>1 (OtherShared)
H_0 : *Random.Speaker.Level.N2+ − Actual.Speaker.Level.N2+* ≥ 0

Level	1	2	3	4	5	Mean
IF	108	104	105	81	107	101
IG	90	91	88	58	89	83.2

A closer look at the number of times the null hypothesis was rejected depending on the Role and Level is given in Table 1 and 2. Within those two tables, we observe the higher rate of rejection for OTHERSHARED than SELFSHARED, for both Information Giver (IG) and Information Follower (IF). Nonetheless, a significant asymmetry between the different roles arise, the IF repeating himself and the IG significantly more in Actual dialogue than the Randomized ones, except for N2+ SELFSHARED repetitions. We notice the low rates of rejection for SELFSHARED for the IG, that only repeated himself significantly in five dialogues for the Level 4 (POS) in particular. Yet, the IG overall rate of rejection for N2+: *n*-gram>1 is slightly higher than the IF, which could signal that when the IG do repeat himself, he tend to repeat longer utterances.

Table 2: *Rejections of H0 for Selfshared, 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 (Selfshared)
H_0 : *Random.Speaker.Level − Actual.Speaker.Level* ≥ 0

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

N1: *n*-gram=1 (Selfshared)
H_0 : *Random.Speaker.Level.N1 − Actual.Speaker.Level.N1* ≥ 0

Level	1	2	3	4	5	Mean
IF	8	10	11	4	11	8.8
IG	4	4	4	0	5	3.4

N2+: *n*-gram>1 (Selfshared)
H_0 : *Random.Speaker.Level.N2+ − Actual.Speaker.Level.N2+* ≥ 0

Level	1	2	3	4	5	Mean
IF	38	38	39	26	37	35.6
IG	44	49	43	16	46	39.6

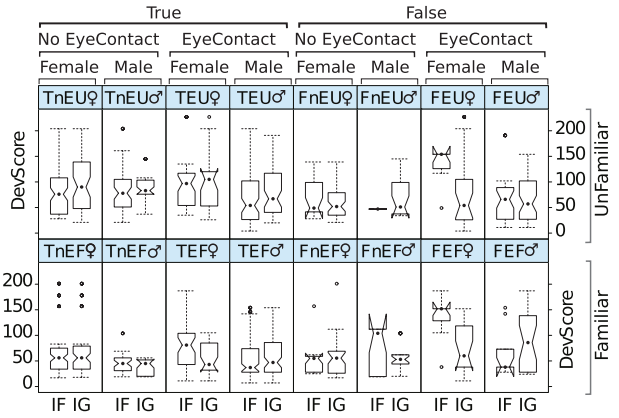

Figure 1: *Distribution of Deviation Score by Role (IG: Information Giver—IF: Information Follower), significant* OTHERSHARED *p-values (T: TRUE—F: FALSE), Eye-contact (nE: no Eye-contact—E: Eye-contact), Familiarity (U: Unfamiliar—Fa: Familiar), and Gender (♀: Female—♂: Male)*

3.2. Non-linguistics features

Mann-Whitney-Wilcoxons tests for population distribution and Hedge tests for effect-size showed overall non-significant differences and negligible effect-sizes of the distribution of Deviation scores neither between Male and Female (W= 9049.5, $p = 0.14$, g estimate= 0.17), nor between Eye-Contact and no-Eye-Contact (W= 8278, $p = 0.88$), and only showed significant difference (W= 6572, $p = 0.006$) between Familiar ($\bar{x} = 64.37$) and Unfamiliar ($\bar{x} = 79.28$) participants. Those tests showed both non-significantly different distributions for Information follower (Gender: $p = 0.10$; Eye-Contact: $p = 0.92$; Familiarity: $p = 0.053$) and Information giver, even if a small effect size appeared between gender for the IF ($g = 0.30$). However, significant differences appeared at the introduction of the factors TRUE or FALSE resulting from our previous Tukey's tests described in § 2.4, and not only with Familiarity. The Figures 1 and 2, show the distribution of the dialogues along the deviation score depending on Role, in interaction with significant OTHERSHARED and SELFSHARED p-values, Eye-contact, Familiarity and Gender for All *n*-grams. A large effect-size ($g = -0.92$)

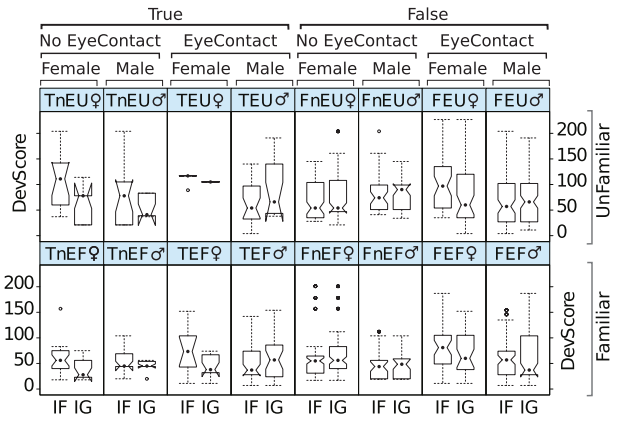

Figure 2: *Distribution of Deviation Score by Role (IG: Information Giver—IF: Information Follower), in interaction with significant* SELFSHARED *p-values (T: TRUE—F: FALSE), Eye-contact (nE: no Eye-contact—E: Eye-contact), Familiarity (U: Unfamiliar—Fa: Familiar), and Gender (♀: Female—♂: Male)*

was shown between female and male IF OTHERSHARED non-significant p-values. We observe in Figure 1 and 2 that the combination no Eye-Contact and Familiar subjects is related to lower deviation scores ($\bar{x} = 61.2$), without strong effects from Gender, Role or Shared type, except for the combination OTHERSHARED FNEF♂, where the male IF not repeating significantly the IG have an average deviation score ($\bar{x} = 72$) higher than any other Familiar—no Eye-Contact combination. However, in Unfamiliar—no Eye-Contact condition, a large effect-size was found ($g = 0.30$), for OTHERSHARED repetitions of the IG, between male repeating the IF significantly (TNEU♂) and female not repeating the IF significantly (FNEU♀), with an average deviation score of ($\bar{x} = 91.14$) and ($\bar{x} = 58.62$) respectively.

In addition to observe the distribution for All *n*-grams, we took a closer look at the notion we introduced in § 2.2.2, "structural repetitions". Indeed, a large effect-size ($g = 0.89$, W= 44.5, $p = 0.18$) was detected for self-repetitions of the female IG at Level 4 (POS) for *n*-gram>1. A medium effect ($g = -0.71$, W= 57, $p = 0.10$) was seen between significant self-repetition of the IG at Level 1(Token) for all n-grams. That effect can be seen in Figure 1 where TRUE SELFSHARED combination relate to lower scores for the IG, except for male participants in Eye-Contact settings. This tendency seem to indicate that if the participants cannot see each-other, self-repetitions from the IG are playing a role toward a lower deviation score. Significant differences in distribution associated with a large effect-size (W= 49, $p = 0.01$, $g = -1.03$) was also found in Eye-Contact condition between male IF repeating significantly the IG ($\bar{x} = 62.17$) and female IF non-significantly repeating the IG ($\bar{x} = 105.44$).

4. Discussion

From the simple observation that the Information Giver (IG) had a much higher volume of speech than the Information Follower (IF) and tended to produce longer utterances (see§ 2.1), it is interesting to see that while talking less, the Information Follower (IF) repeated himself and the IG more often significantly in almost all the tested conditions. The results show consistency with previous finding in the sense that in task-based interaction significant Othershared and Selfshared repetitions have an impact on task-success. It seem that overall, the IG self-repeating more often significantly for N2+ could be interpreted as keeping the same structure in providing information, which tends to relate to lower deviation scores, or in other terms, to higher communicative success. Despite Familiarity appearing at first as the most distinguishing factor, significant differences in Gender, in particular in interaction with Eye-Contact, corresponded to clearly different communicative behaviours. Even thought the results have to be taken consciously given the sometime small size of the samples, the results suggested that for the IF, non-significant OTHERSHARED repetitions mattered less for male than female, for a small portion of which it meant less successful communication. "Structural" self-repetitions of the female IG was related to lower deviation score than male. No Eye-Contact and Familiarity were related to lower deviation scores, which seem the best combination for task-success, even if men IF not repeating the IG were the one performing the least in those conditions. However, in Eye-contact situation, female Information Follower (IF) not repeating themselves performed in average less well than men. The patterns founds with the method are encouraging as it could be potentially used in a predictive machine learning task in the perspective of a use in dialogue systems. In particular in the case of SELFSHARED repetitions, the Level 1(Token Only) did not always display significant differences to allow the rejection of the Null Hypothesis, but other Levels did so, hence indicating the additional information those Level divisions are bringing. The labelling step that leads to counting repetitions for other linguistic Levels is therefore giving additional information that can be used in the interpretation of the variations of communicative behaviours. It has to be noted that no significant difference appeared between the Levels Lemma and Token, as well as the association Lemma+POS (L3) and Token+POS (L5) which often showed little variations. Thus could be explained in two ways. First, the nature of the task did not allowed an important variety of inflexions to appear, the participants used a simple vocabulary. It is also possible to imagine that the influence of inflexion would be more significant in a different language than English.

5. Conclusion

This article has described the extension of a method of interaction analysis based on repetitions that is distinct in analytical details from other analytical methods in the literature. The possibilities given by this extended method to find specific patterns of significant OTHERSHARED and SELFSHARED repetitions in relation with task-success in various settings is a promising step toward the establishment of an automatic quantifying measure of communicative success without the need for annotated data. As many factors were taken into account, other potential effects might have been over-looked. In particular, the effects the linguistic Levels in interaction with the non-linguistic features of Gender, Eye-contact and Familiarity are yet to be examined. As one of the possible future application of this measure of synchrony to quantify mutual understanding reside in its use in dialogue systems, and the possible determination of change in communicative strategies within them, further exploration of the method is required.

6. Acknowledgements

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.

7. 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] H. P. Branigan, M. J. Pickering, J. Pearson, J. F. McLean, and C. Nass, "Syntactic alignment between computers and people: The role of belief about mental states," in *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, R. Alterman and D. Kirsh, Eds. Boston, Massachusetts, USA: Cognitive Science Society, 2003, pp. 186–191.

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

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

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

[10] J. J. Gumperz, *Discourse strategies*. Cambridge University Press, 1982, vol. 1.

[11] 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.

[12] 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.

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

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

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

[16] H. Branigan, R. Lickley, and McKelvieDavid, "Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech," in *Proceedings of ICPhS XIV (14th International Congress of Phonetic Sciences)*, San Francisco, California, USA, 1999, pp. 387–390.

[17] M. Colman, P. G. Healey *et al.*, "The distribution of repair in dialogue." in *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 2011, pp. 1563–68.

[18] M. Swerts, H. Koiso, A. Shimojima, and Y. Katagiri, "On Different Functions of Repetitive Utterances," in *Proceedings of ICSLP 1998: the Fifth International Conference on Spoken Language Processing*. Sydney, Australia: ISCA, 1998.

[19] 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.

[20] H. Schmid, "Treetagger— a language independent part-of-speech tagger," *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, vol. 43, p. 28, 1995.

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