Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1721–1731
Marseille, 20-25 June 2022

The Brooklyn Multi-Interaction Corpus for Analyzing Variation in Entrainment Behavior

Andreas Weise[1], **Matthew McNeill**[1], **Rivka Levitan**[1,2]
[1]Department of Computer Science, The Graduate Center, CUNY, New York, NY, USA
[2]Department of Computer and Information Science, Brooklyn College, CUNY, Brooklyn, NY, USA
{aweise, mmcneill}@gradcenter.cuny.edu, rlevitan@brooklyn.cuny.edu

Abstract

We present the Brooklyn Multi-Interaction Corpus (*B-MIC*), a collection of dyadic conversations designed to identify speaker traits and conversation contexts that cause variations in entrainment behavior. B-MIC pairs each participant with multiple partners for an object placement game and open-ended discussions, as well as with a Wizard of Oz for a baseline of their speech. In addition to fully transcribed recordings, it includes demographic information and four completed psychological questionnaires for each subject and turn annotations for perceived emotion and acoustic outliers. This enables the study of speakers' entrainment behavior in different contexts and the sources of variation in this behavior. In this paper, we introduce B-MIC and describe our collection, annotation, and preprocessing methodologies. We report a preliminary study demonstrating varied entrainment behavior across different conversation types and discuss the rich potential for future work on the corpus.

Keywords: Entrainment, Variation, Dialogue Corpus

1. Introduction

Entrainment is the phenomenon of conversational partners adapting to one another to become more similar. This affects a wide variety of linguistic dimensions, including lexical choice (Brennan and Clark, 1996), syntax (Reitter et al., 2006), phonetics (Pardo, 2006), and prosody (Levitan and Hirschberg, 2011; Gravano et al., 2014). It has also been found in many conversational contexts, from collaborative, personal interactions, in pairs (Brennan and Clark, 1996; Levitan and Hirschberg, 2011; Lubold and Pon-Barry, 2014) and in groups (Rahimi et al., 2017), to conversations over the phone (Cohen Priva and Sanker, 2020) or on Twitter (Danescu-Niculescu-Mizil et al., 2011), to even human-computer interaction (Coulston et al., 2002; Suzuki and Katagiri, 2007; Thomason et al., 2013). Notably, it correlates with, among others, success in tasks (Reitter and Moore, 2007) and learning (Thomason et al., 2013), rapport (Lubold and Pon-Barry, 2014) and social behavior (Levitan et al., 2012), and the perceived trustworthiness and likability of artificial interlocutors (Levitan et al., 2016; Metcalf et al., 2019).

The degree and valence with which entrainment occurs vary substantially by the speakers and context (Lubold and Pon-Barry, 2014; Pardo et al., 2018; Weise et al., 2019). Despite theories purporting to broadly explain the behavior (Giles et al., 1991; Chartrand and Bargh, 1999; Pickering and Garrod, 2004), the exact mechanisms governing its emergence in individual conversations are still poorly understood. Attempts to attribute differences to gender, for instance, have yielded disparate results or no effect (Pardo et al., 2018; Weise et al., 2019). This lack of theoretical understanding also hinders broader practical application, as appropriate settings for entraining conversational avatars are difficult to predict. This can even result in a shift of user preference from entraining avatars to *dis*entraining ones (Levitan et al., 2016), without the reason – avatar gender, modified feature, etc. – being clear.

In this paper, we present the Brooklyn Multi-Interaction Corpus (*B-MIC*) designed to identify speaker traits and conversation contexts that cause variations in entrainment behavior. B-MIC pairs each participant with multiple partners for dyadic conversations of two different types, as well as with a wizarded dialogue system for a baseline of their speech. In addition to fully transcribed recordings, it contains demographic information and four completed psychological questionnaires for each subject and turn annotations for perceived emotion and acoustic outliers. This enables the study of speakers' entrainment behavior in different contexts and the sources of variation in this behavior.

Much research on entrainment has involved dialogue corpora originally collected for other purposes, such as the Switchboard Corpus (Godfrey and Holliman, 1993; Cohen Priva and Sanker, 2020), the Columbia Games Corpus (Beňuš et al., 2007; Levitan and Hirschberg, 2011), and the Fisher Corpus (Cieri et al., 2004; Weise et al., 2019). Other researchers have collected corpora specifically for the analysis of entrainment. Pardo et al. recorded subjects in interactive conversation as well as a non-interactive speech shadowing task (Pardo et al., 2018). Another example is the SibLing Corpus (Kachkovskaia et al., 2020) of game conversations between a core group of speakers and interlocutors with whom they have varying degrees of familiarity. B-MIC differs in focus from both of these new corpora and offers a more controlled setting than the Switchboard Corpus – including a consistent recording environment and a consistent number of interactions per subject. It also provides more information about the subjects and conversations, potentially allowing for the attribution of behavior to speaker traits and conversation context. This paper is organized as follows. First we introduce B-MIC and discuss its experiment design, data collection, pre-processing, and annotation. Then we conduct a preliminary analysis comparing entrainment behaviors across the two different interaction types, as well as the Objects Game portion of the Columbia Games Corpus (*CGC*), whose data collection task we adopted for the task-oriented portion of this corpus. Lastly, we discuss the results as well as research questions we plan to address with the corpus in the future.

2. The Brooklyn Multi-Interaction Corpus

The Brooklyn Multi-Interaction Corpus (*B-MIC*) is designed to facilitate entrainment research through its structure, the kinds of dialogue it includes, and its participant surveys and annotations.

- Each participant interacts with three other speakers, as well as a wizarded dialogue interface, so we can observe which aspects of their behavior are consistent and which vary across interlocutors.

- The corpus includes two different registers of dialogue - conversational and task-oriented - so we can analyze how entrainment behavior differs based on dialogue context, with all other factors held constant.

- Each participant completed a set of psychological questionnaires measuring traits that have been associated with entrainment in prior work.

- Recordings have been segmented, orthographically transcribed, and annotated at the turn level for phenomena associated with entrainment.

2.1. Corpus and experiment design

The corpus is designed to include 48 speakers, divided into 12 groups of four, two males and two females each. Speakers participate in two types of dyadic conversation with each other as well as a Wizard of Oz and complete five questionnaires.

To elicit spontaneous conversation, we ask participants to discuss starting a business and going back in time to change something they have done, respectively. We chose these hypotheticals from the Fisher Corpus (Cieri et al., 2004) because we expected them to be engaging but not polarizing. Each speaker participates in one *conversational* session with a female interlocutor and one with a male interlocutor, each lasting ten minutes.

For the *task-oriented* conversations, speakers take turns describing the placement of target objects among arrangements of several others, a setup modeled after the Objects Game portion of the Columbia Games Corpus (Beňuš et al., 2007). Participants complete four *sessions* of this, each consisting of 14 placement *tasks* lasting one minute apiece.[1] After completing three sessions with three different human partners, participants are asked to play one more session with an "automated partner", a "computer" recording and responding to their speech in text form. In reality, research assistants hear and respond to the subjects' utterances with standardized written messages. For a screenshot of the user interface and further details, see Figure 1. Subjects are informed about this Wizard of Oz setup during debriefing and given opportunity to withdraw their consent. Note that responses are in text to avoid acoustic-prosodic entrainment to the system output (Coulston et al., 2002; Suzuki and Katagiri, 2007), so a baseline of the subjects' speech can be recorded.

At the end of each session with a human partner, subjects are asked to rate the perceived likability of the interlocutor and the smoothness of the interaction on a five-point Likert scale.

Finally, participants provide basic demographic information – age, sex, and gender and racial identification, each with an option not to report – and complete four psychological questionnaires. Specifically, these are:

1. A short form (Reynolds, 1982) of the *Marlowe-Crowne Social-Desirability Scale* (Crowne and Marlowe, 1960) of subjects' need for social approval,

2. the *Interpersonal Reactivity Index* (perspective-taking subscale) (Davis, 1983), assessing subjects' tendency to consider and adopt their interlocutor's point of view,

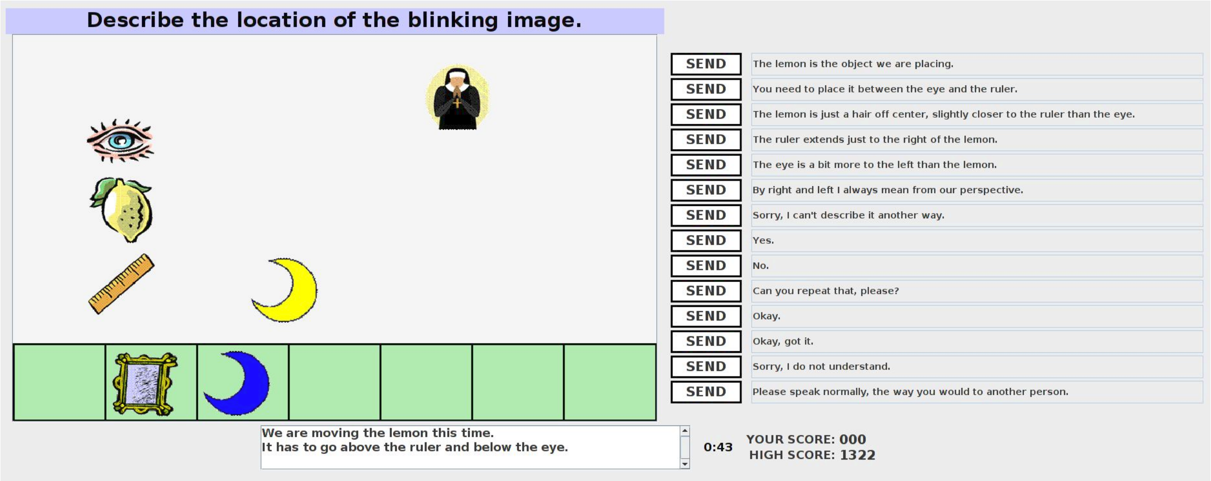
3. the *Ten Item Personality Inventory* (Gosling et al., 2003) of the "Big Five" personality dimensions,

4. and the *Reading the Mind in the Eyes Test* (Baron-Cohen et al., 2001), measuring Theory of Mind (ToM), the ability to model the emotional state of the interlocutor.

We note that this sensitive data is not linked with subjects' names or other personal identifiers. Subjects also provide informed consent at the beginning of the experiment and are informed of the option to withdraw consent at any time without any penalty. The research protocol has been submitted to and approved by the University Integrated Institutional Review Board of the City University of New York under file #2018-1568.

The choice of the questionnaires is informed by prior work on entrainment. Natale (1975) and Chartrand and Bargh (1999) found social desirability and perspective-taking, respectively, to be significant moderators of entrainment behavior, which motivates us to investigate them as well. While both results are frequently cited, to our knowledge neither has ever been replicated. Various studies, meanwhile, have found the "Big Five" to significantly influence entrainment (Gill et al., 2004; Yu et al., 2011; Lewandowski and Jilka, 2019). The

[1] In *CGC*, task duration varied substantially.



Describe the location of the blinking image.

The lemon is the object we are placing.

You need to place it between the eye and the ruler.

The lemon is just a hair off center, slightly closer to the ruler than the eye.

The ruler extends just to the right of the lemon.

The eye is a bit more to the left than the lemon.

By right and left I always mean from our perspective.

Sorry, I can't describe it another way.

Yes.

No.

Can you repeat that, please?

Okay.

Okay, got it.

Sorry, I do not understand.

Please speak normally, the way you would to another person.

We are moving the lemon this time.
It has to go above the ruler and below the eye.

0:43

YOUR SCORE: 000
HIGH SCORE: 1322

Figure 1: The Objects Game interface for the Wizard of Oz during a task in which the Wizard takes the role of the describer. The lemon would be blinking for the assistant seeing this interface. Meanwhile, the subject, having to place the lemon in this case, would see it as solid in the inventory at the bottom, next to the blue moon and the mirror. The message log at the bottom is only visible during the Wizard of Oz session and the message bar on the right is only ever visible to the Wizard, never to subjects. The remaining time for the current task (out of one minute), the overall score for the current session, and the high score are always visible at the bottom right.

Reading the Mind in the Eyes Test, lastly, is often used in the study of autism, which is linked to impaired ToM. Autism has been considered as a possible source of variation in entrainment in the past (e.g., (Allen et al., 2011; Slocombe et al., 2013; Branigan et al., 2016); for a review, see (Kruyt and Beňuš, 2021)). We include the test to analyze the impact of ToM on entrainment within a neurotypical range and to facilitate future comparison with autistic speakers.

The corpus design includes counterbalancing. Speakers in half of the groups engage in free conversation first; for the other half, speakers play the game first. The Wizard of Oz session and the completion of the questionnaires always happen directly after the game sessions with human partners.

2.2. Data collection

So far, four of the planned 12 groups of subjects have been recruited and recorded. Further data collection has been prevented by public health measures due to the COVID-19 pandemic. The corpus currently includes 56 sessions with roughly 4.5k annotated turns (each annotated three times) and over 27k speech segments, roughly 7 hours of speech (excluding silent pauses, see details below).[2]

During recording, subjects were seated inside separate audiometric booths designed to reduce ambient noise. They were not able to see each other and only able to communicate through Microsoft LifeChat LX-3000 headsets over a local network connection. Subjects were recruited mainly from the Brooklyn College student body (age $\mu = 22.7$). All 16 are native speakers of American English, 8 male and 8 female. Six identify as Black or African American. Table 1 lists statistics for the personality traits of our 16 speakers to enable comparisons with other populations. Results for Theory of Mind, for instance, are in line with general, neurotypical populations (Baron-Cohen et al., 2001).

2.3. Pre-processing and annotation

2.3.1. Segmentation and transcription

Recorded conversations have been split into interpausal units, *IPUs*, – maximal speech segments from each speaker without interruption by the interlocutor or a pause – and transcribed. This was done automatically but with manual correction.

The automatic segmentation into sounding intervals separated by at least 50ms of silence was done based on energy levels measured with openSMILE (Eyben et al., 2013, v2.3.0) and an empirically chosen threshold around 34dB. This was possible thanks to low levels of ambient noise around 15dB. Manual corrections were performed by the first author to finalize IPUs. Specifically, non-speech sounds, including laughter, were removed; incorrect boundaries of low-intensity speech sounds were corrected; and intervals were merged when the silence separating them was within a word or less than 100ms long and before a stop consonant such as [p] or [k]. The latter indicates a hold phase (Crystal and House, 1988) and is thus merely articulatory in nature. Thresholds of 50 to 100ms for delimiting pauses capture even brief interruptions, for instance for repairs, which are common in spontaneous speech. As a result, they have been used in prior acoustic-prosodic entrainment research (Levitan and Hirschberg, 2011;

[2] The corpus is available on request. Please contact the third author at rlevitan@brooklyn.cuny.edu. Samples can be found at `http://www.sci.brooklyn.cuny.edu/~levitan/speechlab/bmic/`.

Trait	Survey range	Statistics min	 max	 avg	 std
Social desirability	0-13	3	11	7.13	2.33
Perspective-taking	0-28	15	28	21.5	4.13
Big 5: openness	2-14	10	14	13	1.46
Big 5: conscientiousness	2-14	8	14	11.5	1.86
Big 5: extraversion	2-14	3	13	8.94	3.79
Big 5: agreeableness	2-14	7	14	9.94	2.67
Big 5: neuroticism	2-14	2	14	8.38	4.69
Theory of Mind	0-36	20	31	26.8	3.34

Table 1: Minima, maxima, averages, and standard deviations for the personality traits of the 16 speakers recorded so far. "Survey range" specifies the possible values in the underlying questionnaire. Note that the Big 5 are based on Likert scales with a minimum value of 1 for each item. For details on the questionnaires, see Section 2.1

Lubold and Pon-Barry, 2014; Reichel et al., 2018) and we employ them here.

For automatic transcription, all IPUs were sent to the speech to text component of IBM's Watson cloud services.[3] Transcriptions were then manually corrected by two students from CUNY's Graduate Center, including the second author, both native speakers of American English. Finally, all corrected transcripts were examined a second time by the first author to ensure consistency and accuracy.

2.3.2. Emotion and outlier annotation

In addition to the orthographic transcription, each speaker turn is annotated for perceptual acoustic-prosodic outliers and emotional expression. The recordings collected so far were annotated by three linguistics students at the CUNY Graduate Center, two female, one male, all native speakers of American English. Each of them annotated all conversations except the Wizard of Oz sessions. For each conversation, annotators listen to all turns for both speakers in sequence and annotate all those with at least 1 second of non-silent speech. Perceived emotions are annotated with regard to the dimensions of valence and arousal (Russell, 1980). Annotators denoted emotions on a $[-100, 100]$ scale by placing points in a 2-D grid, with an option for multiple emotions per turn. For outlier annotation, they note if part or all of a turn seems to be unusually breathy or creaky, and unusually high or low with respect to intensity, pitch, or speech rate.

For acoustic outliers, at least two out of three annotators agree on outliers for roughly 87 percent of turns for intensity, pitch, and speech rate[4] and roughly 98 percent of the turns for creakiness and breathiness, counting only turns where at least one annotator perceived an outlier. The female pair agrees roughly twice as often as the male annotator agrees with either one of them. Acoustic outliers have previously been found to result in greater entrainment (Levitan, 2014). We intend to replicate this result but also expand it beyond objective outliers to perceived ones, which might not be identical and appear even more directly relevant to entrainment.

For perceived emotion, the Pearson r between pairs of annotators are 0.15 (female pair), 0.3 (mixed pair 1), and 0.45 (mixed pair 2) for valence and 0.35 (female pair), 0.48 (mixed pair 1), and 0.27 (mixed pair 2) for arousal. The Krippendorff α for overall emotion annotation agreement is 0.24 for both valence and arousal. We note that this is a low level of inter-annotator agreement, both with regard to accepted practices around Krippendorff's α generally and when compared to corpora of affective speech specifically. B-MIC's experimental design is not intended to elicit explicitly emotional speech. Rather, participants converse naturally with strangers in relatively low-stakes interactions. While successes and failures in the games or recollections and aspirations in the free conversations lead to occasional peaks in emotion, overall our annotators were not listening to clearly affective speech. Instead, emotion in the recordings is mostly subtle, sometimes barely perceptible. While this results in lower levels of agreement, it more closely resembles every-day conversational speech which is the target of practical applications of entrainment research, e.g., for the evaluation of call-center agents.

A simpler, categorical annotation scheme might have resulted in higher inter-annotator agreement, but we do not believe that categorical labels capture sufficient detail to represent the subtle emotions we seek to collect. For example, sentiment labels like those used to annotate the Switchboard Corpus (Chen et al., 2020) only capture the general positivity or negativity of an utterance, while categorical labels like the Ekman categories (Ekman and Friesen, 1971) describe concrete emotional states such as "happy" or "sad" that an annotator might struggle to hear in our recordings. Our novel aim is to assess how limited, occasional emotion interacts with entrainment. We intend further efforts to establish ground-truth valence and arousal values and, alternatively, to explicitly account for the disagreements among annotators (e.g., (Sethu et al., 2019)).

[3] https://www.ibm.com/cloud/watson-speech-to-text

[4] Exact matches: high or low *and* part or all of the turn.

	Characteristics				
	#sessions per subject	types of sessions	speech baseline	personality data	emotion annotation
Switchboard Corpus (Godfrey and Holliman, 1993)	1-32	free conversation	No	No	No
Columbia Games Corpus (Beňuš et al., 2007)	1-2	task-oriented	No	No	No
SibLing Corpus (Kachkovskaia et al., 2020)	5	task-oriented	No	No	No
Montclair Map Task Corpus (Pardo et al., 2018)	2	task-oriented & speech shadowing	Yes	No	No
B-MIC	5	task-oriented & free conversation	Yes	Yes	Yes

Table 2: Comparison of the Brooklyn Multi-Interaction Corpus (B-MIC) with other dialogue corpora that have been used for entrainment research and contain multiple sessions for at least some subjects. Note that B-MIC is the first corpus we are aware of in which subjects interact with the same partners in different registers, namely free conversation and task-oriented interaction. The speech shadowing component of the Montclair Map Task Corpus is non-interactive and involves a distinct group of model speakers. Additionally, while B-MIC is the first corpus of its type we are aware of that contains emotion annotations, the Switchboard Corpus was annotated for sentiment with positive, negative, and neutral labels by Chen et al. (2020).

2.4. Comparison with other corpora

Table 2 lists dialogue corpora that have been used for entrainment research in the past and compares them with B-MIC regarding core characteristics. Subjects in all these corpora either participated in multiple interactions or, in the case of the Montclair Map Task Corpus (Pardo et al., 2018), in an interaction and a non-interactive speech shadowing session.

To our knowledge, B-MIC is the first corpus that has subjects interact with the same partners in different registers, both free conversation and task-oriented interaction, rather than, for instance, just different tasks. It also provides personality data on the subjects that should help explain observed differences in entrainment behavior across subjects, partners, and registers.

3. Entrainment measures and features

We apply five established acoustic-prosodic entrainment measures (Levitan and Hirschberg, 2011) that have been used extensively in the literature (Levitan et al., 2012; Thomason et al., 2013; Lubold and Pon-Barry, 2014; Rahimi et al., 2017; Weise et al., 2019). **Global similarity** compares speakers' mean feature values for entire sessions. It is said to be present for a feature if speakers are significantly more similar to their respective partners than to comparable non-partners with whom they never interacted. Meanwhile, **global convergence** is found for a feature if the mean feature values of paired speakers are significantly more similar in the second half of a session than in the first. Besides these global measures whose significance is assessed for the corpus as a whole, three local measures are applied to each session individually. **Local similarity** compares the similarity of adjacent IPUs at turn exchanges – i.e., a turn-final IPU A and the immediately following turn-initial IPU B – to the similarity between IPU B and a random selection of other, non-adjacent, turn-final IPUs from the same session and by the same speaker as IPU A. **Local convergence** measures whether the similarity between adjacent IPUs at turn exchanges correlates with the index of those exchanges within the session, i.e., whether feature values at turn exchanges become more or less similar over time. **Synchrony**, lastly, measures the correlation between interlocutors' feature values at turn exchanges. For further details, see (Levitan and Hirschberg, 2011).[5]

Each measure is applied to eight features which are commonly used in acoustic-prosodic entrainment research (Levitan and Hirschberg, 2011; Lubold and Pon-Barry, 2014; Truong and Heylen, 2012; Weidman et al., 2016; Rahimi et al., 2017). Specifically, for each IPU, we extract mean and max pitch, mean and max intensity, and jitter, shimmer, and noise-to-harmonics ratio (NHR) using Praat (Boersma and Weenink, 2001) and measure speaking rate as syllables per second. All features are z-score normalized by speaker.

4. Results

In this section, we report the results of applying the five measures to each of the eight features in the *game* and *conversation* sessions of B-MIC as well as the Objects Games portion of the Columbia Games Corpus (*CGC*), whose data collection paradigm we adopted for the B-MIC *game* sessions. We present a highlighted sum-

[5] Our exact implementation and other code related to the project is available at https://github.com/andreas-weise/bmic.

	B-MIC games			B-MIC conversations			Columbia Games Corpus		
	+	−	+/−	+	−	+/−	+	−	+/−
global similarity	0%	0%	n/a	25%	25%	n/a	0%	0%	n/a
global convergence	0%	0%	n/a	0%	0%	n/a	0%	0%	n/a
local convergence	4.2%	8.3%	4.2%	6.2%	37.5%	0%	25%	8.3%	16.7%
local similarity	20.8%	0%	0%	0%	12.5%	0%	16.7%	0%	0%
synchrony	29.2%	12.5%	25%	12.5%	37.5%	0%	33.3%	0%	0%

Table 3: Overview of results: Percentage of features, out of 8, exhibiting global similarity or global convergence with positive or negative valence for each (sub-)corpus; and percentage of sessions, out 24/16/12, resp., with local convergence, local similarity, or synchrony for at least one feature, with all positive, all negative, or mixed valence.

mary of the results here.[6] Further details on the results per session for our local measures can be found in Appendix A in Tables 4 to 12.

To account for multiple testing, we control the false discovery rate (Benjamini and Hochberg, 1995). We treat each group of eight tests for a session (for local measures) or group of sessions of the same type (for global measures) as a family and apply thresholds of a family-wise error of $\alpha = 0.05$ for significance and $\alpha = 0.1$ for approaching significance.

4.1. Brooklyn Multi-Interaction Corpus

In the *game* sessions, we find no significant evidence of global entrainment, neither similarity nor convergence. The only result with $p < 0.1$, for global similarity of mean pitch ($t(47) = 2.23, p = 0.03$), does not even approach significance when accounting for multiple testing. For the local measures, there are at least some significant results. Local convergence is rare, with only four of the 24 sessions approaching or reaching significance for at least one feature. Moreover, three of these actually showed *di*vergence on mean and max intensity. Local similarity is also rare, occurring in only five of the 24 sessions, with no clear feature pattern, but always positive valence. Synchrony, lastly, is common, with 16 of the 24 sessions reaching or approaching significance for at least one feature, most often on max intensity, mean pitch, and NHR. While most of these results have positive valence, for max intensity they are almost exclusively negative. This means speakers tend to respond to louder utterances with a quieter response, a complementary behavior which can also be beneficial (Pérez et al., 2016).

For the *conversation* sessions, global similarity approaches significance for mean pitch ($t(31) = 2.48, p = 0.019$), shimmer ($t(31) = 2.56, p = 0.015$), mean intensity ($t(31) = -2.14, p = 0.026$) and max intensity ($t(31) = -2.34, p = 0.041$). Note that the valence is negative for mean and max intensity, indicating that speakers are *less* similar to their partners than others with whom they never spoke. While no feature even approaches significance for *global* convergence, *local* convergence reaches or approaches significance on at least one feature for seven out of 16 sessions, most frequently on mean and max intensity. However, almost all of these results have negative valence, indicating *di*vergence, i.e., speakers becoming less similar at turn exchanges over time. Similarly, the rare cases of local similarity – in two sessions – and almost all of the cases of synchrony – across eight sessions – are negative, suggesting complementary behavior.

In summary, we find no evidence of global convergence in either type of sessions in B-MIC, while global similarity at least approaches significance in *conversation* sessions, with negative valence in two cases. Local entrainment is found in both session types, with notable differences. Local similarity and synchrony are more common in *game* than in *conversation* sessions, with much more positive valence. Local divergence, on the other hand, is more common in *conversation* sessions.

4.2. Comparison with Columbia Games Corpus

Neither global similarity nor global convergence even approaches significance in *CGC*. Of the local forms of entrainment, similarity is least common, found in only two of the twelve sessions. Synchrony is present in four sessions, for up to five features simultaneously. Local convergence is found in three sessions. All of the results for local similarity and synchrony have positive valence, while some for convergence are negative.

Table 3 summarizes our findings. The results for *CGC* are comparable to those for the B-MIC *games*, whose experimental design is based on that of the *CGC* sessions. *CGC* and the *games* share a lack of global entrainment while B-MIC *conversation* sessions contain evidence of global similarity. Local similarity is found at a similar rate and with the same positive valence in *CGC* and *games*, but only with negative valence in *conversations*. Synchrony is less common in *CGC*, but with the same tendency for positive valence as in the *games*, unlike for the *conversation* sessions. For local convergence, lastly, no two of our (sub-)corpora show similar trends.

This comparison highlights the important role of register and social context in moderating entrainment be-

[6] Full results are available at `http://www.sci.brooklyn.cuny.edu/~levitan/speechlab/bmic/`.

havior. The *game* sessions exhibited similar entrainment behavior across corpora, while the B-MIC *conversation* sessions tend to differ from the B-MIC *games* even though they involved the same set of speakers. This suggests that conversational and speaker *states* may be more important moderators of entrainment than speaker *traits*.

5. Discussion and future work

We introduce the Brooklyn Multi-Interaction Corpus, a new dialogue corpus designed for the analysis of variations in entrainment behavior with different interlocutors and across dialogue registers – conversational and task-oriented. Psychological questionnaires and turn-level annotations included in the data provide the basis for the attribution of variations to speaker traits as well as states within interactions.

A preliminary analysis presented in this paper demonstrates the variation across conversation contexts. Speakers in task-oriented conversation tend to show entrainment behavior more like that of *other* speakers in a similar setting than like *themselves* in free conversation. That is, the same speaker pairs can exhibit profoundly different entrainment behavior in otherwise identical conditions, on the same day, based solely on the interaction type. One pair, for instance, became less similar at turn exchanges for five of our eight acoustic-prosodic features over the course of their free conversation. Yet in their task-oriented interaction, the same speakers did not show such local divergence, or convergence, for *any* feature. In fact, there is only a single instance of any of our local entrainment measures reaching the level of significance for the same feature in both types of interaction for the same speaker pair. This is a clear indication of the impact of conversation context on entrainment behavior.

Prior work has shown that cognitive load inhibits entrainment (Abel and Babel, 2017). Our task-oriented setting is designed to be more mentally challenging than the free conversations. However, we find *more* closely matching and synchronous features at turn exchanges in our task-oriented sessions. In our future work, we intend to investigate alternative explanations, such as the emotional state of the speakers in the different conversation types. Speakers often expressed frustration or concern regarding their score in game sessions, but reminisced about the past, discussed personal ambitions, and offered advice and support to one another in conversation sessions.

The new corpus will enable two broad paths for future work. Firstly, to use the multiple interactions in which each speaker participates, including the wizarded baseline, to develop an analysis of inter- and intra-speaker variations in entrainment behavior, and attempt to explain these variations in terms of the demographic and psychological data associated with each speaker.

Secondly, we will use the turn-level annotations for perceived outliers and emotional state, as well as session information regarding task success and partner liking, to analyze variations in entrainment behavior associated with speaker and conversation *state*.

The ultimate goal of these analyses is to achieve an understanding of how all these moderators of entrainment behavior interact in a single integrated model that can explain the behavior observed in human-human communication, and generate behavior in an embodied conversational agent that is appropriate for a particular context and persona. The richness of the multiple kinds of interactions in this corpus, the participant demographic and psychological data, and the turn-level annotations, should help make such a model possible.

6. Acknowledgements

This work was supported by the Alfred P. Sloan Foundation [Award No. G-2018-11286] and the National Science Foundation [Grant No. 1845710]. The authors thank Agustín Gravano of the Universidad de Buenos Aires, Argentina, for providing the code of the Columbia Games Corpus interface. We are also very grateful to the Diana Rogovin Davidow Speech Language Hearing Center at Brooklyn College and its Director Michael Bergen for allowing us to use their facilities and for their helpful assistance. Finally, we thank Tiffany Davis and Brad Pringle for their assistance in collecting the corpus.

7. Bibliographical References

Abel, J. and Babel, M. (2017). Cognitive Load Reduces Perceived Linguistic Convergence Between Dyads. *Language and Speech*, 60(3):479–502.

Allen, M. L., Haywood, S., Rajendran, G., and Branigan, H. (2011). Evidence for syntactic alignment in children with autism. *Developmental science*, 14(3):540–548.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The “reading the mind in the eyes” test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2):241–251.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Beňuš, Š., Gravano, A., and Hirschberg, J. (2007). The Prosody of Backchannels in American English. In *ICPhS XVI*, pages 1065–1068.

Boersma, P. and Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, 5(9-10):341–347.

Branigan, H. P., Tosi, A., and Gillespie-Smith, K. (2016). Spontaneous lexical alignment in children with an autistic spectrum disorder and their typically developing peers. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 42(11):1821.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482–1493.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: The Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

Chen, E., Lu, Z., Xu, H., Cao, L., Zhang, Y., and Fan, J. (2020). A large scale speech sentiment corpus. In *LREC 2020*, pages 6549–6555.

Cieri, C., Miller, D., and Walker, K. (2004). The Fisher corpus: a Resource for the Next Generations of Speech-to-Text. *LREC*, 4:69–71.

Cohen Priva, U. and Sanker, C. (2020). Natural leaders: Some interlocutors elicit greater convergence across conversations and across characteristics. *Cognitive Science*, 44(10).

Coulston, R., Oviatt, S., and Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. In *ICSLP2002*, volume 4, pages 2689–2692.

Crowne, D. P. and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4):349–354.

Crystal, T. H. and House, A. S. (1988). The duration of american-english stop consonants: An overview. *Journal of Phonetics*, 16(3):285–294.

Danescu-Niculescu-Mizil, C., Gamon, M., and Dumais, S. (2011). Mark My Words! Linguistic Style Accommodation in Social Media. In *WWW 2011*, pages 745–754.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113–126.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *ACM Multimedia*, pages 835–838.

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1–68. Cambridge University Press.

Gill, A. J., Harrison, A. J., and Oberlander, J. (2004). Interpersonality: Individual differences and interpersonal priming. *CogSci 2004*, 26(26):464–469.

Godfrey, J. J. and Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. Web download. Linguistic Data Consortium, Philadelphia.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Gravano, A., Beňuš, Š., Levitan, R., and Hirschberg, J. (2014). Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In *Spoken Language Technology (SLT), 2014 IEEE Workshop on*, pages 578–583.

Kachkovskaia, T., Chukaeva, T., Evdokimova, V., Kholiavin, P., Kriakina, N., Kocharov, D., Mamushina, A., Menshikova, A., and Zimina, S. (2020). SibLing Corpus of Russian Dialogue Speech Designed for Research on Speech Entrainment. In *LREC 2020*, pages 6556–6561.

Kruyt, J. and Beňuš, Š. (2021). Prosodic entrainment in individuals with autism spectrum disorder. *Topics in Linguistics*, 22(2):47–61.

Levitan, R. and Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *INTERSPEECH 2011*, pages 3081–3084.

Levitan, R., Willson, L., Gravano, A., Beňuš, Š., Hirschberg, J., and Nenkova, A. (2012). Acoustic-Prosodic Entrainment and Social Behavior. In *NAACL HLT 2012*, pages 11–19.

Levitan, R., Beňuš, Š., Gálvez, R. H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., and Hirschberg, J. (2016). Implementing acoustic-prosodic entrainment in a conversational avatar. In *INTERSPEECH 2016*, pages 1166–1170.

Levitan, R. (2014). *Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue*. Ph.D. thesis, Columbia University.

Lewandowski, N. and Jilka, M. (2019). Phonetic Convergence, Language Talent, Personality and Attention. *Frontiers in Communication*, 4(May):1–19.

Lubold, N. and Pon-Barry, H. (2014). Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12.

Metcalf, K., Theobald, B. J., Weinberg, G., Lee, R., Jonsson, I. M., Webb, R., and Apostoloff, N. (2019). Mirroring to build trust in digital assistants. In *INTERSPEECH 2019*, pages 4000–4004.

Natale, M. (1975). Convergence of Mean Vocal Intensity in Dyadic Communication as a Function of Social Desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., and Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, 69:1–11.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Pérez, J. M., Gálvez, R. H., and Gravano, A. (2016). Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement. In *INTERSPEECH 2016*, pages 1270–1274.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences*, 27(2):169–190.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., and Yu, M. (2017). Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels. In *INTERSPEECH*, pages 1696–1700.

Reichel, U. D., Beňuš, Š., and Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication*, 100:46–57.

Reitter, D. and Moore, J. D. (2007). Predicting Success in Dialogue. In *ACL 2007*, pages 808–815.

Reitter, D., Moore, J. D., and Keller, F. (2006). Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In *CogSci 2006*, pages 685–690.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the marlowe-crowne social desirability scale. *Journal of Clinical Psychology*, 38(1):119–125.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

Sethu, V., Provost, E. M., Epps, J., Busso, C., Cummins, N., and Narayanan, S. (2019). The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*.

Slocombe, K. E., Alvarez, I., Branigan, H. P., Jellema, T., Burnett, H. G., Fischer, A., Li, Y. H., Garrod, S., and Levita, L. (2013). Linguistic alignment in adults with and without asperger's syndrome. *Journal of autism and developmental disorders*, 43(6):1423–1436.

Suzuki, N. and Katagiri, Y. (2007). Prosodic Alignment in Human-Computer Interaction. *Connection Science*, 19(2):131–141.

Thomason, J., Nguyen, H. V., and Litman, D. (2013). Prosodic entrainment and tutoring dialogue success. In *International Conference on Artificial Intelligence in Education*, pages 750–753.

Truong, K. P. and Heylen, D. (2012). Measuring prosodic alignment in cooperative task-based conversations. In *Interspeech 2012*, pages 843–846.

Weidman, S., Breen, M., and Haydon, K. C. (2016). Prosodic speech entrainment in romantic relationships. In *Speech Prosody 2016*, pages 508–512.

Weise, A., Levitan, S. I., Hirschberg, J., and Levitan, R. (2019). Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, 115(April):78–87.

Yu, A. C., Abrego-Collier, C., Baglini, R., Grano, T., and Martina, M. (2011). Speaker Attitude and Sexual Orientation Affect Phonetic Imitation. *Proceedings of the 34th Annual Penn Linguistics Colloquium*, 17(1):234–242.

A. Session results

Tables 4 to 12 list the results per local measure and (sub-)corpus for all sessions with at least one feature at least approaching significance. Valence per session is represented through plus or minus symbols and their number reflects the level of significance α after adjusting for repeated testing per session and measure (each "family" consists of eight tests): "+++": $\alpha < 0.001$, "++": $\alpha < 0.01$, "+": $\alpha < 0.05$, "(+)": $\alpha < 0.1$ (analogous for "–").

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
2	(–)	(–)						
29	– –	–						
32				++				
34	– –	– –			+	+		

Table 4: B-MIC *game* session results for local convergence.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
12	– –	– –			– –	– – –	– –	
14	(–)	(–)		(–)				
17	++	+						
39			– –					
40			–					
41	– –	– –						
45	– –	– –						

Table 5: B-MIC *conversation* session results for local convergence.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
3	–	–						
6							+	
7	(+)							
9			(+)	(–)				
10	+					(+)		
11			(+)					(–)

Table 6: *Columbia Games Corpus* session results for local convergence.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
1			++	++				
20	(+)							
22						+++	+	
47	+++	+						
52	+		++					

Table 7: B-MIC *game* session results for local similarity.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
44	(–)	–						
45	–	– –						

Table 8: B-MIC *conversation* session results for local similarity.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
3								(+)
9	+				(+)		+	

Table 9: *Columbia Games Corpus* session results for local similarity.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
1			+++	+++				
3		– –						
4							(+)	
5					++	++	++	
6							(+)	(–)
20	(+)		++					
22		–			+	+++	++	
23		(–)			(+)			
30								(–)
31		–			–		+	
32			++					
34	– – –	– – –					+	– – –
47	+++	++	(–)			+	(+)	
49			+			+		
51		(–)						
52	(+)		+					

Table 10: B-MIC *game* session results for synchrony.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
13			(+)					
15							–	
16							(–)	
17	–	–						
39	(–)	–						(–)
44	– –	– – –						
45	(–)	– – –						
46				+				

Table 11: B-MIC *conversation* session results for synchrony.

ses	mean intensity	max intensity	mean pitch	max pitch	jitter	shimmer	nhr	rate
2	+++	+			+++	+	++	
3				+			+	+
7					+			
9	+				+		(+)	

Table 12: *Columbia Games Corpus* session results for synchrony.