COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13398

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13398

Multi-Level Linguistic Alignment in a Dynamic Collaborative Problem-Solving Task

Nicholas D. Duran,[a] Amie Paige,[b] Sidney K. D'Mello[c]

[a]*School of Social and Behavioral Sciences, Arizona State University*
[b]*Department of Psychology, Stony Brook University*
[c]*Institute of Cognitive Science and Department of Computer Science, University of Colorado Boulder*

Received 12 April 2022; received in revised form 13 October 2023; accepted 8 December 2023

Abstract

Cocreating meaning in collaboration is challenging. Success is often determined by people's abilities to coordinate their language to converge upon shared mental representations. Here we explore one set of low-level linguistic behaviors, linguistic alignment, that both emerges from, and facilitates, outcomes of high-level convergence. Linguistic alignment captures the ways people reuse, that is, "align to," the lexical, syntactic, and semantic forms of others' utterances. Our focus is on the temporal change of multi-level linguistic alignment, as well as how alignment is related to communicative outcomes within a unique collaborative problem-solving paradigm. The primary task, situated within a virtual educational video game, requires creative thinking between three people where the paths for possible solutions are highly variable. We find that over time interactions are marked by decreasing lexical and syntactic alignment, with a trade-off of increasing semantic alignment. However, greater semantic alignment does not translate into better team performance. Overall, these findings provide greater clarity on the role of linguistic coordination within complex and dynamic collaborative problem-solving tasks.

Keywords: Linguistic alignment; Collaborative problem-solving; Interactive alignment; Dialog; Virtual learning environments

1. Introduction

Conversation provides people the opportunities to create shared understandings of concepts, assumptions, and expectations for whatever communication goals are at hand

Correspondence should be sent to Nicholas D. Duran, School of Social and Behavioral Sciences, Arizona State University, 4701 W Thunderbird Rd, Faculty Administration Bldg FAB, Glendale, AZ 85306-4908, USA. E-mail: nicholas.duran@asu.edu

(Brennan & Clark, 1996; Clark, 1996; Clark & Brennan, 1991). But even when the goals seem simple and the process easy (e.g., just shooting the breeze), verbal communication is still challenging (Brennan, Galati, & Kuhlen, 2010). There are demands on individual memory, perspective-taking, emotional regulation, and other cognitive skills (Galati, Dale, & Duran, 2019; Horton & Gerrig, 2005; Roche & Arnold, 2018), as well as challenges in dealing with another who might not be as skilled or motivated to meet communication goals (Brennan & Clark, 1996; Slocome et al., 2013). Such demands are only compounded in collaborative problem-solving (CPS) situations, a type of real-world interaction that is becoming increasingly common across workplaces and classrooms in the 21st century (Levy & Murnane, 2012; Wüstenberg, Greiff, & Funke, 2012). In CPS interactions, multiple people must collaborate to identify a solution path(s) to get to a goal state from a current state (i.e., to solve a problem). This entails developing transactive knowledge structures that consist of situational and task-appropriate strategies for solving nonroutine problems (Chen et al., 2020). Often, this goal must be accomplished in contexts that involve a high degree of uncertainty, varied domain knowledge, multiple solution paths, both optimal and suboptimal, and a constant updating of task priorities and perspectives (Graesser et al., 2018). Communications in CPS are thus complex, leading to the critical question of how to assess the quality of communication in CPS interactions when outcomes are multidimensional (e.g., solving the problem, being pleased with the level of teamwork) and varied.

In this study, we explore linguistic alignment behaviors that are critical for conversation and establishing shared meaning, but in the novel context of triadic remote CPS. These behaviors have been largely studied in association with various process models of dialogue, where the primary outcome of communication is the convergence of mental representations between conversational partners (Clark, 1992; Pickering & Garrod, 2004). This process is both supported by—and is a result of—the alignment of linguistic expressions and structures, particularly that of lexical and syntactic forms (Branigan, Pickering, & Celeland, 2000; Brennan & Clark, 1996; Branigan & Pickering, 2017; Cleland & Pickering, 2003). In what follows, we provide a brief overview of the role of lexical and syntactic alignment in task-based interactions and touch upon semantic alignment as a relatively new domain of analysis. We then turn to our main aim of exploring conceptual extensions of linguistic alignment within a dynamic, complex, and naturalistic CPS task involving spontaneous open-ended speech communication.

1.1. Linguistic alignment across multiple channels

1.1.1. Lexical and syntactic alignment

Lexical alignment refers to a phenomenon whereby a speaker tends to reuse the same words and phrases from the recent discourse. According to prominent process models of dialogue, one of the major purposes for lexical alignment is to establish localized referential precision (Clark & Wilkes-Gibbs, 1986; Dideriksen, Fusaroli, Tylén, Dingemanse, & Christiansen, 2019; Pickering & Garrod, 2004). Conversational partners tend to rely heavily on lexical alignment to draw attention to salient elements within an environment (Dideriksen et al., 2019), and to form what are known as "conceptual pacts," a common way of referring to

objects strongly associated with a particular dialogue partner (Brennan & Clark, 1996). Evidence for lexical alignment tends to be most pronounced in tasks where there is a high degree of interactivity with a limited number of possible objects. The widely employed "Map task" and tangram-based paradigms are two such examples (Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987). In these tasks, participants typically have asymmetric knowledge of a static environment (e.g., one participant alone having a privileged view; landmarks in the case of the Map task, ambiguous figures in tangram studies), and success is predicated on establishing shared terms. Although the mechanisms by which participants arrive at this convergence are not absolute, whether through low-level priming that does not require explicit monitoring of others' mental states (Pickering & Garrod, 2004, "interactive alignment"), through a more active seeking and confirming of mutual understanding (Clark, 1992, "common ground"), or on more context-sensitive and functionally driven mechanisms outside of a priming/common ground continuum (Rasenberg, Özyürek, & Dingemanse, 2020; Duran, Dale, & Galati, 2016; Dale, Fusaroli, Duran, & Richardson, 2013), the same assumption of outcome in most cases remains: referential language should become simpler, less variable, and more coordinated over time (Brennan & Clark, 1996; Foltz et al., 2015; Mills, 2014).

Syntactic alignment is another discourse phenomenon that underscores the interdependencies between speakers. This alignment occurs when people reuse recently heard syntactic structures of their conversational partners in their own speech (Branigan et al., 2000). Syntactic alignment is typically taken as a behavior driven by cross-person priming (Bock, 1986; Mahowald, James, Futrell, & Gibson, 2016). It mostly occurs outside of conscious awareness, and reflects an implicit, collaborative attunement to a conversational partner that is unique from lexical alignment, insofar that the latter involves more explicit tracking of what another may or may not know. Given the same idea/meaning can be conveyed with divergent syntactic structures, a lack of syntactic alignment does not necessarily interfere with negotiating a shared knowledge, but it can signal more shallow processing of what another is saying (Branigan, Pickering, McLean, & Cleland, 2007; Heyselaar & Segaert, 2019). Indeed, in corpus studies of naturalistic dialogue, problem-solving tasks that require greater attention to another speaker show stronger syntactic priming effects (Reitter, Keller, & Moore, 2006). This increased attunement to another, as expressed in greater syntactic alignment, helps explain the findings across other studies where speakers' comprehension is facilitated by syntactic alignment (Noppeney & Price, 2004; Schoot, Menenti, Hagoort, & Segaert, 2014; Thothathiri & Snedeker, 2008).

1.1.2. Semantic alignment

Semantic alignment attempts to capture how people converge on similar meanings *without* the use of strict lexical (symbolic) repetition. Although these explorations have traditionally been pursued through qualitative analysis (Schegloff, 2007), recent insights have been made with the use of computational models for generating word embeddings (Angus, Smith, & Wiles, 2012a; Duran, Paxton, & Fusaroli, 2019; Sagi & Diermeier, 2017). These models represent word meaning based on the co-occurrence statistics of words across unique contexts within a vary large corpus of language (Foltz, Kintsch, & Landauer, 1998; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). For example, a word like "street" will be similar to "road,"

as compared to "eggplant," because street and road appear in more contexts that contain the same types of words. Word meaning thus becomes a distribution pattern of occurrences across many possible contexts, that is, a high-dimensional vector, based on the neighbors a word keeps (and, depending on the algorithm used, those it selectively does not keep; see Angus, Smith, & Wiles, 2012b). Because the unit of analysis in examining alignment is typically at the level of a conversational turn, the words within each turn can be mathematically combined to create a higher-order semantic vector. It is these semantic vectors that are then compared to gauge similarity across conversational turns.

Angus, Watson, Smith, Gallois, and Wiles (2012) are among the first to use a distributional semantic models to evaluate turn-by-turn semantic alignment in open-ended and spontaneous conversation. They evaluated physician consultations with patients through the lens of communication accommodation theory (CAT; Gallois, Ogay, & Giles, 2005), where communication effectiveness was determined, in part, through a greater frequency of semantic repetition by interlocutors throughout an interaction. In related work, caregivers who engaged in more short-term topic repetition with dementia patients were rated as having more advanced active listening skills (Baker et al., 2015). Expanding to multi-party interactions involving a negotiation task, Sagi and Diermeier (2017) also applied a distributional semantic modeling approach to track semantic alignment as a function of dialogue turns over time (albeit within a text-based chat environment). They found that those who formed an alliance in the negotiation task showed greater semantic alignment as they converged on an outcome, whereas others who were also involved in the linguistic exchange, but excluded from the alliance, did not align. In other words, local convergence (alignment) does not automatically occur when participating in a linguistic exchange, instead, it is associated with interactions that involve greater shared understanding and knowledge.

1.2. Trade-offs in alignment during CPS

Although there is much evidence for linguistic alignment in dialogue, open questions abound on its functional role, whether alignment varies for particular types of linguistic forms, and how the expression of alignment changes over time given changing problem-solving needs. Many of the expectations for how alignment should be expressed are obtained in tasks that constrain variation, such as with strict turn-taking and the use of simple referents and goals that do not require a diverse vocabulary (Gries, 2005; Howes, Healey, & Purver, 2010). Many of these tasks also overemphasize the need for convergence of mental representations. They typically begin with a high degree of asymmetric knowledge between conversational partners, where common ground needs to be accumulated over time in order to succeed (Rączaszek-Leonardi, Dębska, & Sochanowicz, 2014). However, as others have observed, in complex problem-solving scenarios, dialogue often breaks from convergence and takes on a divergent process involving contrasting views, elaborations, and updating of prior meanings across turns (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014; Mills, 2014; Rączaszek-Leonardi et al., 2014). Such variation helps move along a conversation to explore various possible outcomes, and given these processes are likely to influence lexical and syntactic alignment, it supports observations that interlocutors find persistent

amounts of high referential precision to be overly redundant and unnatural, and in some cases, counterproductive to meeting problem-solving solutions (Fusaroli et al., 2014). Indeed, Amon, Vrzakova, and D'Mello (2019) recently reported that irregularity (i.e., lack of routine patterns) in teammates' nonverbal signals better reflected expert-coded CPS skills, whereas Eloy et al. (2019) found that it better predicted CPS outcomes.

In the current work, we focus on a CPS interaction that we believe represents a more balanced tension between the dialogue goals of convergence and divergence. Our theoretical motivation is to better understand how various forms of linguistic alignment (lexical, syntactic, and semantic), as a contextually driven and temporally evolving phenomenon, are expressed within these more complex interactions. Specifically, the CPS task used here, explained in greater detail below, requires three people to work together to generate creative and often ill-defined solutions in visually shared dynamic environments. Each environment contains a diverse set of referents and possible actions that can be taken on the referents. Often multiple possibilities need to be entertained and revised to find solutions and avoid impasses. We argue that the process involved reflects an alternative "synergistic" approach to dialogue where the primary goal is not to necessarily achieve a mutually shared mental model and to overcome individuals' privileged knowledge (Rączaszek-Leonardi et al., 2014). Rather, the aim of the interaction is to maximize functional outcomes, which is best served by multiple ways of coordinating meaning (Brown-Schmidt, 2012). Depending on the temporal scale in which communicative demands are examined, as well as the behavioral channel in focus, evidence for greater convergence *and* greater divergence of overlapping knowledge should be found.

We argue that such evidence will be revealed in linguistic alignment when examined in terms of its presence and magnitude (i.e., does it occur or not, and if so, to what degree) across conversational utterances. Importantly, simultaneous markers of convergence and divergence require an assessment of linguistic alignment across its submodalities, separately tracking the repetition of lexical and syntactic forms, as well as the similarity of semantic information, across speech turns. Moreover, to reveal changes over longer timescales, it is critical that analysis occurs within extended, task-based dialogues. In the current study, the dialogue occurs over 45 min, in 15 min rounds, with sequences of speech naturally sequenced within rounds into several possible problem-solving trials of varying length (determined by whether participants are successful in solving individual problems or quit and move on to the next problem). We quantify change in alignment as a linear trend over time within rounds, as well as within the shorter trials (i.e., individual problems embedded in rounds). This characterization targets a type of system change where the behavior of interest, in this case various submodalities of lexical alignment, evolves toward distinct stable states over different periods of observation (i.e., *System I Dynamics* as discussed in Gorman & Wiltshire, 2024).

To offer predictions on how these various changes will be expressed, we take inspiration from predominant mechanistic accounts without attempting to promote one over the other, recognizing that certain accounts, like priming, more naturally explain syntactic alignment, whereas a grounding perspective is particularly salient for lexical and semantic alignment. We also recognize that the degree to which this is true is largely dictated by how alignment is measured and analyzed (Rasenberg et al., 2020), as well as the contextual and

task demands imposed on speakers. Given both elements are novel in the current study, predictions are somewhat speculative in nature but draw from relevant research. With this caveat in mind, one reasonable expectation for lexical alignment is that in creative and unconstrained problem-solving tasks, alignment will remain constant throughout the interaction or even decrease over time. As previously discussed, a major role of lexical alignment is to establish a common way of referring to objects or actions. But once established, there is also evidence that repeatedly referring to the same elements within a problem space in the same way might not be particularly persuasive (Duran et al., 2019; Healey, Purver, & Howes, 2014; Mills, 2014). Other ways of coordinating meaning may take precedent, such as convergence within a semantic space of shared topics and themes without the need for precise lexical repetition, that is, semantic alignment. As such, we expect that as lexical alignment decreases, there will be compensation by an increase of semantic alignment. For syntactic alignment, one expectation is that it will simply covary with lexical alignment. From an interactive alignment account, cascading priming effects are assumed across the two linguistic levels (Pickering & Branigan, 1998; Pickering & Garrod, 2004, the so-called lexical boost effect). However, there are other reasons to think alignment will decrease for task-specific reasons. Because priming is a cognitively resource-minimizing process, where preactivated syntactic structures are easier to produce, its effects should be most pronounced in cognitively demanding tasks (Pickering & Garrod, 2004). Insofar that these demands decrease with practice, so too will the presence of syntactic alignment (Foltz et al., 2015). Likewise, in considering an account of syntactic alignment as influenced by one's attunement to others, as demands of social engagement decrease with greater familiarity, again, an associated decrease of alignment is expected. Although the current work is not able to tease apart these various mechanisms, they all point to a similar prediction of how alignment might unfold over time.

As linguistic alignment becomes more or less convergent, it should also be associated with the likelihood of problem-solving success. A common expectation, one that is supported by priming and common ground models, is that greater alignment across individual linguistic channels predicts positive effects on communicative outcomes, such as in ratings of others' comprehensibility or whether interlocutors come to a desired agreement (Angus et al., 2012; Sagi and Diermeier, 2017). However, whether this expectation should also hold in extended CPS contexts is not as straightforward given conversational goals are not necessarily the same as CPS goals. For CPS, decreasing lexical alignment, or more breaks in lexical repetition, with a simultaneous increase of semantic alignment, could be functionally productive for better problem-solving solutions. And insofar that arriving at better solutions is marked by less cognitive demand, syntactic alignment could also be expected to be less pronounced.

Here, we add to a growing body of observations by taking a unique approach in evaluating multiple levels of linguistic alignment at once in a single paradigm, and examining their importance to a relatively more challenging and open-ended problem-solving space (e.g., CPS involving triads). Moreover, success in the current task is not binary, but incorporates gradations of quality that can be taken into account. We can do this, as explained next in greater detail, because our task is composed of multiple and clearly delineated problem-solving scenarios where a solution is achieved or not, and the quality of the solution is automatically assessed as more or less elegant.

2. Methods and setup

2.1. Data collection

Data analyzed here were collected as part of a multiyear study on CPS. Only aspects germane to the present analysis on linguistic alignment, which has not been previously published with these data, are presented here.

Participants were 282 undergraduates (originally 288 but six participants were removed due to recording errors) from two large public universities in the United States (61% from University 1). Of the 282, 56% were female, with an average age of 21.73 years, and 75% reporting English as their first language. Participants self-reported the following race/ethnicities: 50% Caucasian, 26% Hispanic/Latino, 18% Asian, 3% Black or African American, 1% American Indian or Alaska Native, and 2% "Other." Participants were assigned to 94 triads based on scheduling constraints. Forty-three participants from 21 teams (22%) indicated they knew at least one person from their team prior to participation. Participants were compensated with a $50 Amazon gift card (96%) or course credit (4%) at the end of the study.

Data were collected in a CPS task involving teams of three participants playing an educational video game called Physics Playground. This game was developed to support and measure the learning of conceptual physics (Shute, Ventura, & Kim, 2013). Game play occurs in a dynamic virtual environment across multiple levels. In each level, the main goal is to move a small ball to a balloon in the midst of fixed obstacles via the creation and manipulation of simple objects (ramps, pendulums, springboards, etc.) drawn into the environment with a computer mouse (as shown in Fig. 1). These objects, as well as the environment as a whole, obey basic rules of physics relating to Newton's laws of force and motion, mass, gravity, potential and kinetic energy, and conservation of momentum. Participants can immediately see the consequences of their actions and revise them accordingly by adding and deleting objects as they see fit. Teams can also spend as much time on each level as they like and they can quit and attempt a new level at any time or reattempt a previous level.

Prior to the day of the scheduled in-lab collaborative task, participants were asked to independently complete a brief tutorial on how to play Physics Playground and to complete a few levels for familiarization with the game. Both components were embedded in a web-based application that allowed participants to work in an informal setting of their choosing (i.e., outside of the laboratory). Although there was no recording of any speech or video during this initial training, tutorial and practice-level attempts were verified for completion.

For the second phase of the study involving a laboratory-based session, participants were assigned to work with two other people based entirely on scheduling constraints. They were placed at separate computers and conversational interaction was made possible through video- and voice-enabled virtual collaboration (using the Zoom software). One participant was deemed the Controller who interacted with the game and shared their screen with the other participants who played the role of Contributors, providing hints, suggestions, and/or encouragement. Participants could freely talk to each other throughout the collaborative interaction via headset-mounted microphones. Audio was recorded as separate channels for each participant.

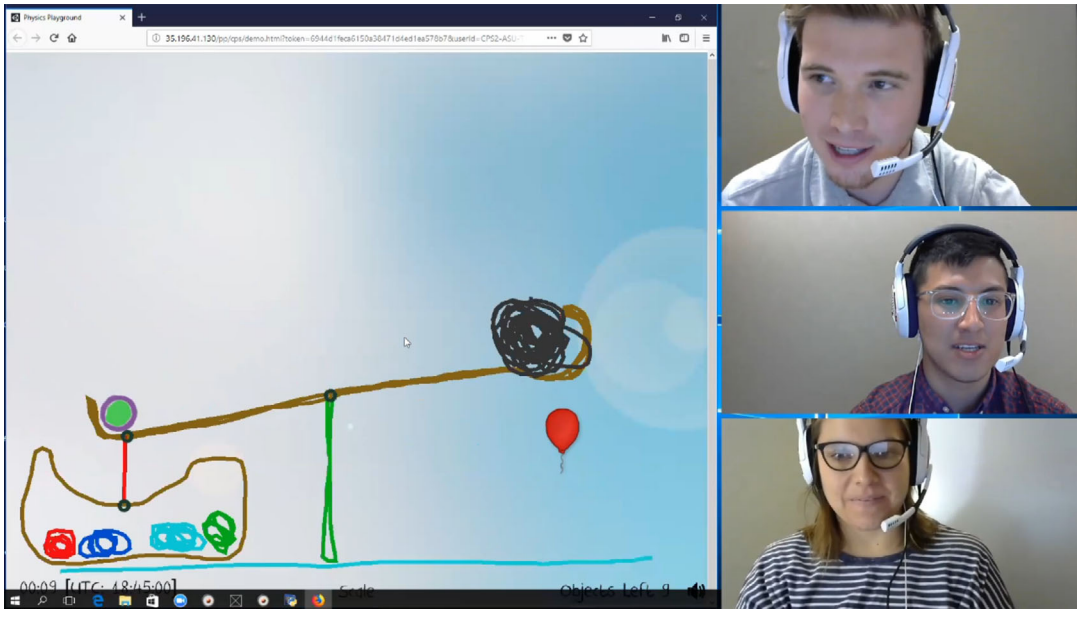

Fig. 1. A game level of Physics Playground where participants employed lever-related physics concepts to lead a ball to a specified target to solve the level. The team was virtually collaborative over Zoom with video, audio, and screen sharing enabled.

Collaboration proceeded in three rounds, with each round lasting 15 min and each participant randomly assigned to act in the Controller role for each round. There was also a short break between rounds so participants could individually provide subjective self-report ratings of the previous interaction (not analyzed in the current study). The first round consisted of five possible levels that served as both practice with Physics Playground and as an opportunity to become acquainted with each other. The following two rounds consisted of either six or seven possible unique levels, where each level set contained challenges, objects, and operations meant to demonstrate a particular physics concept ("properties of torque" or "energy transfer"). The order in which the concepts were presented across rounds 2 and 3 was counterbalanced across teams (and added as covariate in all statistical models), whereas there was a mix of concepts across levels for the initial warm-up round. In addition, we also manipulated and counterbalanced the CPS goal for rounds 2 and 3, instructing participants to either solve as many levels as possible in one round, and in another round to attempt to solve levels using the fewest objects as possible in their solutions (this was an experimental manipulation for another purpose).

Based on expert-ratings of difficulty (comprising the game mechanics and underlying physics concepts), easier levels were assigned to the initial warm-up round, whereas moderate/difficult levels were assigned to the next two rounds. Participants were free to choose any of the levels within each round in any order, though they tended to follow the prescribed order. Each attempt could end without resolution, be abandoned for a new level and returned to later, or a level could be solved (i.e., the ball strikes the balloon). Based on algorithms

established in previous validation studies (Shute et al., 2013), solutions are automatically evaluated, resulting in immediate feedback in the form of a virtual gold or silver coin. An optimal (elegant) solution, corresponding to a gold coin, is achieved through drawing a limited number of objects to solve the problem, emphasizing efficiency/creativity, whereas any solution earns a silver coin.

To prepare the speech data for linguistic analysis, conversations were segmented into participant-level turns and transcribed (by humans) using a third-party service called Rev (www.rev.com). The Rev-recruited transcriptionists are vetted for basic English competency and trained to identify unique speakers and to write down words as they hear them (some transcribers also included speech fillers, noted laughter, and noted long pauses, although this was done inconsistently across transcribers and these annotations were not used in the current analysis). This data is stored and transcribed within a secure portal where it cannot be downloaded onto personal devices. Transcriptionists are also under a nondisclosure agreement to protect participants' confidentiality.

Upon completion of each transcription, an internal team of three research assistants compared the transcription to the corresponding video to ensure that (a) each conversational turn consisted of an utterance from a single participant, (b) a new turn was generated when a new participant began speaking, and (c) the transcription was accurate (e.g., high fidelity to what was said, correct spelling). Research assistants corrected minor errors, but for consistent problems, video/audio files were returned to Rev and reassigned to a new transcriber until all issues were resolved.

2.2. *Quantifying linguistic alignment*

Linguistic alignment was calculated using ALIGN, an open-source natural language processing tool developed in our previous work (see Duran et al., 2019 for details). The tool computes the amount of lexical and syntactic alignment across conversational turns, and employs a novel method for computing semantic alignment over time. To prepare the data for analysis, the transcripts for each game-play level, for each team, needed to be converted into an N x 2 matrix, where each N row is a speech turn of the current speaker in the order it occurred in the dialogue interaction. In this way, each row is a record of alternating speakers, with associated data of who is speaking (Speaker A, Speaker B, or Speaker C) and a word-level transcription of the spoken utterance. Several standardized preprocessing options were applied to the dialogue, including: (a) removal of all numbers, punctuation, and other non-ASCII characters; (b) removal of common speech fillers (e.g., "um" and "uh"); and (c) removal of short utterances of two or fewer words (typically back-channels and simple affirmations/negations). Given alignment scores are generated across pair-wise contiguous turns, these preprocessing settings help maximize linguistic content and variation across comparisons.

- *Sample of a typical exchange and formatting after processing:*
  C: build something right there by the apple
  A: that it is an apple
  C: yeah it is supposed to be an apple i guess i am just going to call it apple there we go yes

A: yes okay perfect got it
B: and now you can build on the other side there we go yes
A: there you go back to the other one yes
B: wait go back to the other one okay

The ALIGN tool then makes use of Python's Natural Language Toolkit (NLTK, Version 3.2.5; Bird, Klein, & Loper, 2009) to generate a tokenized and lemmatized version of each conversational turn, where tokens are simply words in their original form and lemmas correspond to a grouping together of the inflected forms of a word to create a single unit (e.g., "are" and "is" become "be," "cats" becomes "cat"). Each turn-based token and lemma is then classified according to its part of speech (PoS; e.g., noun, preposition) using the Penn Treebank tagset and two well-established PoS taggers: NLTK's default "averaged perceptron" method and the Stanford Natural Language Processing Group's log-linear implementation (Toutanova, Klein, Manning, & Singer, 2003). The taggers will produce slightly varied results given differences in underlying training corpora and prediction algorithms.

- *Example of the first three turns from the example collaborative exchange as lemmas with Stanford POS tags.*
  C: [('build', 'VB'), ('something', 'NN'), ('right', 'RB'), ('there', 'RB'), ('by', 'IN'), ('the', 'DT'), ('apple', 'NN')]
  A: [('that', 'IN'), ('be', 'VB'), ('an', 'DT'), ('apple', 'NN')]
  C: [('yeah', 'VB'), ('it', 'PRP'), ('be', 'VB'), ('suppose', 'VB'), ('to', 'TO'), ('be', 'VB'), ('an', 'DT'), ('apple', 'NN'), ('i', 'FW'), ('guess', 'NN'), ('i', 'FW'), ('be', 'VB'), ('just', 'RB'), ('go', 'VB'), ('to', 'TO'), ('call', 'VB'), ('it', 'PRP'), ('apple', 'NN'), ('there', 'RB'), ('we', 'PRP'), ('go', 'VBP'), ('yes', 'RB')

To compute linguistic alignment scores, starting with lexical alignment, for each contiguous pair of turns within a larger dialogue, the ALIGN tool generates a vectorized frequency count of the occurrences of each token or lemma for each conversational turn. The vectorized frequency counts for the first two turns of the example collaborative exchange are represented in the following format (based on lemmatization; note the repetition of "apple," where Speaker A aligns to Speaker C):

C: [build: 1, something: 1, right: 1, there: 1, by: 1, the: 1, **apple**: 1]
A: [oh: 1, that: 1, be: 1, an: 1, **apple**: 1]

The cosine similarity between the two turns is then generated, resulting in a score ranging from 1 to 0, with higher scores indicating greater repetition of lexical items. This process is repeated such that each pair-wise turn has an alignment cosine score associated with tokens and lemmas. These two token/lemma values are then averaged to create a single lexical alignment score at each turn.

To compute syntactic alignment, a similar procedure as lexical alignment was used except for the PoS sequences being first converted into *n*-grams, specifically bi- and tri-grams, to capture variation in structural complexity. For each *n*-gram representation of each turn, the tool computes a vectorized frequency count. For example, the following is the vectorized

Table 1
Example of cosine scores computed across pair-wise turns in a dialogue

		Semantic	Lexical (uni-gram)	Syntactic (bi-gram)
C:	And now make it really heavy.	–	–	–
B:	I wonder if we just add a weight.	0.698	0	0.482
A:	It does not have to be too big.	0.608	0	0.378
C:	Is that the one where we have to make it heavier on the other side?	0.695	0.343	0.216

Note. Each value, which can range from 0 to 1, represents alignment of the speaker to the previous speaker (second row values: Speaker B aligned to Speaker C). This example also shows how semantic alignment can be elevated (higher cosine values) while having lexical alignment at zero.

frequency counts for a bi-gram representation for a pairwise exchange in the sample dialogue (based on tokens; note the repetition of the [RB PRP] bigram, where Speaker A aligns to Speaker B):

B: [[and now, CC RB: 1], **[now you, RB PRP**: 1], [you can PRP MD: 1], [can build, MD VB: 1]...]
A: **[[there you RB PRP**: 1], [you go PRP VBP: 1], [go back, VBP RB: 1], [back to, RB TO: 1]...]

Before a cosine comparison between pair-wise turns is made, we also applied a stricter test of syntactic alignment by removing *n*-gram sequences that were also lexically identical across turns, thereby minimizing conflation in determining whether syntactic alignment is merely lexical overlap (i.e., a so-called "lexical boost"; Healey et al., 2014; Pickering & Branigan, 1998; Reitter et al., 2011). Because there are two PoS taggers applied to either token or lemma representation and further divided into either bi- or tri-grams, a total of eight cosine scores (2 taggers x 2 lexical representations x 2 n-gram types) are generated for each pair-wise conversational turn sequence. These eight values were averaged to create a single syntactic alignment score at each turn.

Semantic alignment makes use of Google's *word2vec* model to attain high-dimensional vectors for each word in a conversational turn. This model involves a massive semantic space, with a vocabulary of 3 million words trained on the Google News dataset (about 100 billion words). The vector representations were then combined by simple additive composition, resulting in a new "turn-level" projection in the semantic space. We also used ALIGN's default settings to exclude those words that occurred only once in the entire Physics Playground transcripts and words that were extremely frequent (exceeding three standard deviations of the mean frequency count of all words). Staying consistent with lexical and syntactic alignment, cosine scores were generated for each contiguous turn pair to create a running sequence of values over the course of a conversational transcript. Higher values correspond to turn pairs that are more semantically related given colocation in similar regions of the imported semantic space (see Table 1).

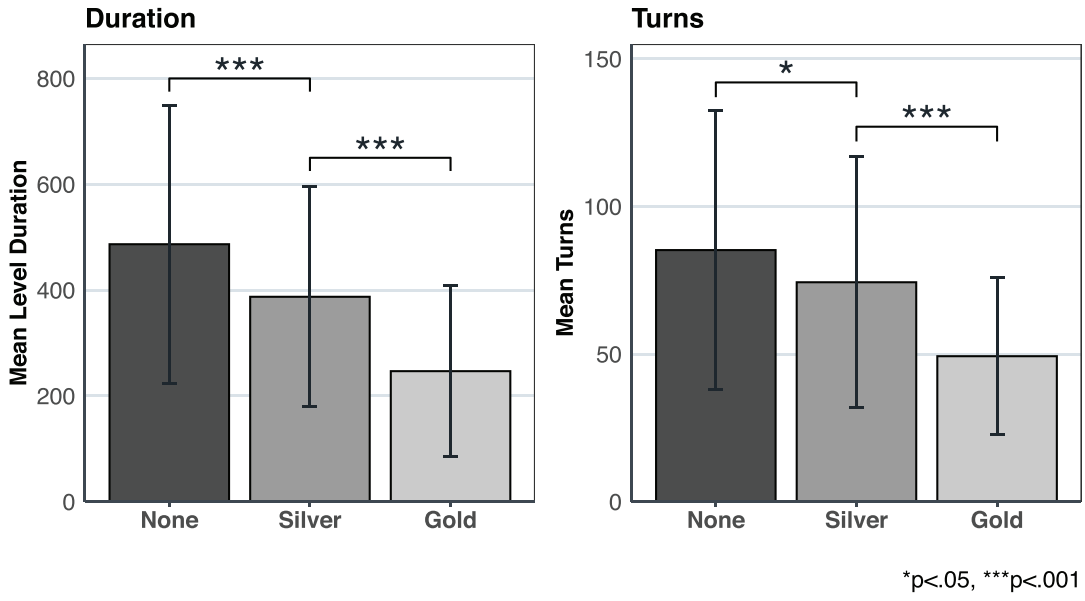

Fig. 2. Mean duration and mean number of turns for levels that ended in a gold, silver, or no coin (none) outcome. Error bars indicate standard deviations.

It is important to note that this approach of using word2vec to capture word and utterance-level meaning allows a more flexible way of capturing referential overlap, akin to that which supports theories of "conceptual pacts" (Brennan & Clark, 1996), but in a more expansive and flexible way than going about it by lexical repetition alone. In our CPS task, as is the case in many such tasks, the focus is largely on a set of referents and possible actions among the referents. To complete the task, it is within this space that the vast amount of language is directed. Interlocutors can use lexically isomorphic expressions to engage in shared reference, but they are also not obligated to do so.

2.3. Level and outcome characteristics

We analyzed the language from 94 teams (282 participants) who played three rounds of Physics Playground. A total of 1089 levels were played by teams across all three 15-min rounds. Not all levels could be used to compute linguistic alignment because of an insufficient number of conversational turns in each. To ensure a reliable signal for measuring amount of alignment and its temporal change, while also preserving as much data as possible, we set an inclusion threshold for levels of 20 turns or higher. This allowed 633 (58.13%) of the levels to be retained for evaluation (number of turns: $M = 54.25$, $SD = 34.92$; words per turn: $M = 11.48$, $SD = 3.03$).

Of the 633 reduced levels, 107 resulted in a gold coin, 242 in a silver coin, and 283 went unsolved. As shown in Fig. 2, teams generally took more time and required more turns in levels that went unsolved as compared to levels where they earned a gold or where they

earned a silver coin. In comparing between gold and silver, silver coins took more time and required more turns. These differences were confirmed with a linear mixed model with team as a random effect (see Appendix A for full results).

2.4. Recap

Before turning to our more detailed findings, we pause here to reiterate our operationalization for how we are measuring, and ultimately analyzing, linguistic alignment. We do so by way of a framework developed by Rasenberg et al. (2020). This framework consists of five theory-agnostic dimensions that nearly all behavioral alignment research can be described by, and is useful for establishing the boundary conditions by which this work should be compared, as well as making more transparent how our methodological commitments might privilege particular mechanistic interpretations.

Using a similar setup as Rasenberg et al. (2020), Fig. 3 provides a descriptive snapshot of this study across the five relevant dimensions: modality, form, meaning, time, and sequence (top panel). The prose for each dimension is linked to a visualization (bottom panel) that represents an interaction between three speakers, depicting how alignment across speakers' utterance turns can be related to the Rasenberg et al. (2020) dimensions. For the three linguistic channels (lexical, syntactic, and semantic) that comprise modality, the underlying forms in each utterance are automatically extracted as either count occurrences or as word embeddings that can be represented in a common vectorization scheme, thus allowing the same cosine similarity metric to be equally applied for quantifying alignment as an ensemble of submodalities. In terms of meaning, the use of an automated approach emphasizes the repetition of lexical units, inclusive of many types of non-differentiated meaning (e.g., opposed to Fusaroli et al. (2012) targeting of a specific speech act), and syntactic units that are primarily independent of meaning.

This characterization is further complemented with an additional automated approach for capturing nonrepeated lexical forms, incorporating a form of meaning in terms of high-level, semantic convergence. For time and sequence, utterance turns are targeted for analysis based on the temporal order in which they are spoken (e.g., Speaker A follows Speaker B; Speaker B follows Speaker C) rather than on any specific coordination adjacencies (e.g., repair statements, question-answer pairs). Given the current work focuses on naturalistic and extended conversational interactions, the durations between paired utterances are quite short, with occasional longer lags.

The dimensions of time and sequence also carry privileged consideration in this work. Going beyond the Rasenberg et al. (2020) framework, which mostly focuses on the relationship between paired, localized behaviors, the current study prioritizes the evolving nature of alignment as a linear change within distinct game-play levels and also collapsed across levels within longer rounds. Alignment is, therefore, examined at two timescales: one where change is bounded by encountering specific novel problems, and another that represents an accruing familiarity with each other and general adaptation to study-imposed role changes (as a different participant was assigned to be the controller in each round).

Modality	Three linguistic channels (lexical, syntactic, semantic); alignment values (magnitude and presence) computed across contiguous, paired speaker turns within channel (as indicated by arrows and x, y, z values)
Form	Lexical units consist of token/lemma uni-grams, syntactic units of POS bi- and tri-grams, and semantic as composite word embeddings
Meaning	Meaning between speaker turns emphasizes the repetition of lexical and POS units independent of meaning; semantic vector emphasizes conceptual similarity in meaning
Time	Short temporal distance between speaker turns; varies based on natural rhythms of conversational interaction
Sequence	Sequential relation between speaker turns loosely bounded by rounds and game-levels within rounds (as indicated by large dashed rectangle)

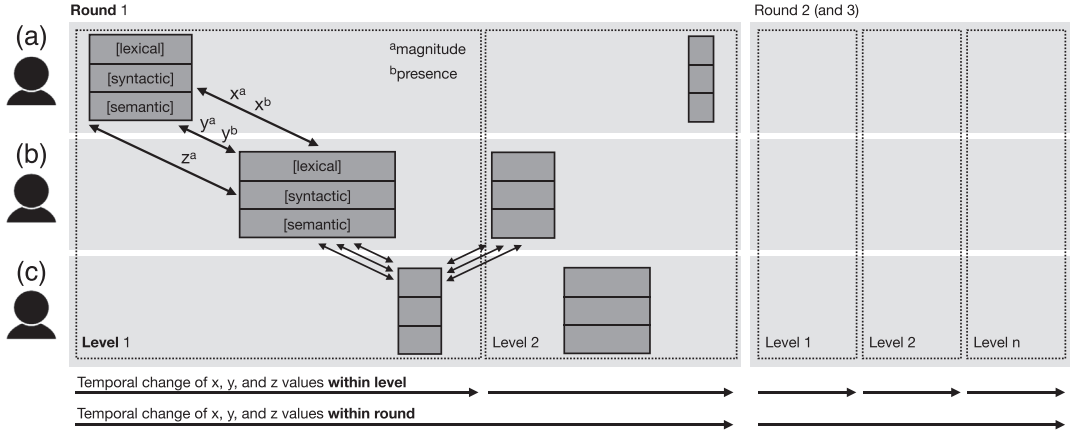

Fig. 3. Top panel: The table provides the distinctive properties of the current study as organized based on the multidimensional Rasenberg et al. (2020) framework for understanding and investigating alignment. Bottom panel: Visualization of an interaction between three participants that captures framework features as well as unique study-specific considerations. The embedded darker rectangles represent a spoken utterance of varying length for a particular participant.

3. Analysis 1: Alignment over time

3.1. Statistical models

In exploratory analysis, lexical and syntactic alignment scores showed many instances of nonoccurrence across contiguous turns, resulting in a distribution with clumping at zero and a skew in the positive values (see Fig. 4). To account for these data, we make use of a hurdle model that separates the analysis into two parts. In the first part, we use a binomial generalized linear mixed model (GLMM) to predict whether alignment occurred to any degree versus the absence of alignment. Then, in the second part, we use a gamma-distributed GLMM to evaluate just the nonzero data that captures the magnitude of alignment. Given the nonzero data involves positive and continuous nonintegers, the gamma distribution is most appropriate in this instance, opposed to more common models that require a truncated count distribution.

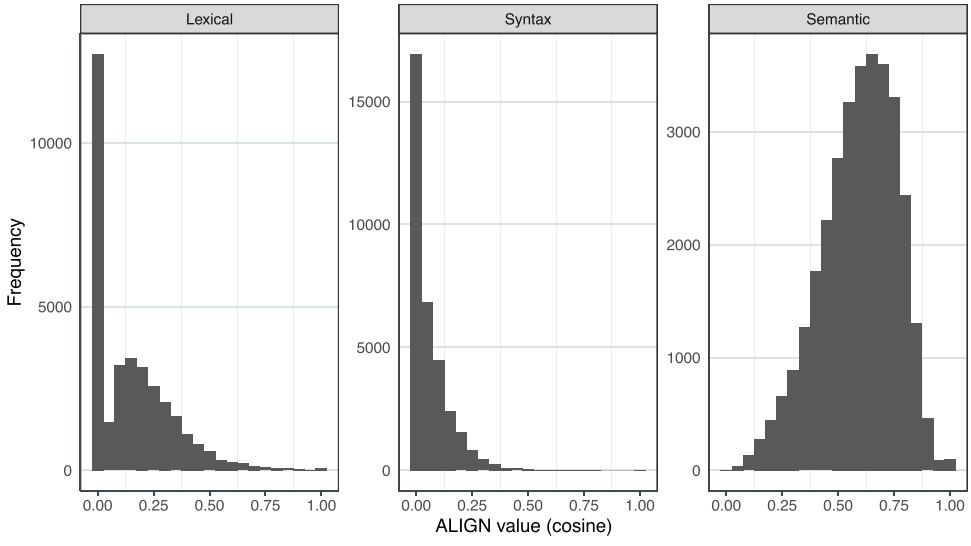

Fig. 4. Distribution of alignment scores for each linguistic type.

In the binomial component of the hurdle model, the logit link was used in predicting the presence of alignment (nonzero vs. zero occurrences). For each independent variable, the resulting coefficients are the expected change in log odds of presence for a one-unit increase in the independent variable. In the case of turn order, a one-unit increase is from the start to the end of the level, and for categorical variables, change is relative to the reference category. For ease of interpretability, the coefficients are typically exponentiated to generate an odds ratio, and subtracting 1 from the odds ratio (odds ratio –1) and multiplying by 100 gives the percentage change in the presence of linguistic alignment for a one unit increase in the corresponding independent variable.

For the gamma component, which examines just the nonzero values of alignment, we use a log link that allows us to predict the model-adjusted mean of alignment. For each independent variable, being on the log scale, we can also exponentiate the resulting coefficients to produce a rate ratio, that is, a multiplicative factor, expressing the expected change of alignment magnitude for a one-unit increase in the independent variable. By subtracting 1 from the rate ratio and multiplying by 100, we can also interpret the change on magnitude of linguistic alignment as a percentage.

Unlike lexical and syntactic alignment, distributions for semantic alignment rarely contained zero values and are normally distributed. Accordingly, we used linear mixed-effects regression to evaluate the time-course of semantic alignment magnitude within levels and across rounds. Estimates are interpreted as indicating how a 1-unit change in each independent variable causes a linear change in semantic alignment (i.e., effectively, the magnitude of alignment).

All analyses were performed with R statistical software using the lme4 package (version 1.1-23; Bates, Maechler, Bolker, & Walker, 2015). All materials, the collected data,

and analyses scripts can be accessed from a Github public repository: https://github.com/nickduran/cps-ling-align

3.2. Analytical approach

Each of the three linguistic alignment variables were entered as response outcomes in their respective models. For all, we included fixed effects for within-level progression of time, represented as turn order within each level, and an ordered factor (linear progression) for round. Given each Physics Playground level varies in duration and number of turns, turn order was transformed as a proportion of total turns (e.g., accruing values from 0.00 and 1.00).

We also included additional fixed-effects covariates to account for subject and team-level characteristics: duration of each level (measured in seconds), start time of each level within each round (measured in seconds elapsed from start of the round), number of words spoken in each turn (measured as a count), the aligner's role (either a contributor or controller of the interface), and the concept being targeted in each level (either "properties of torque" or "energy transfer"). The continuous scores for level duration, level start time, and turn length variables were z-score standardized before being entered into the model. As a reminder, for lexical and syntactic alignment, these covariates are reported as odds ratios, whereas for semantic alignment, they are reported as regression coefficients.

For all models, to examine the appropriateness of their random effect structures, we tested whether a more complex model with by-subject random slopes for role provided a better fit than by-subject random intercepts alone. We used Akaike information criteria (AIC) to compare the models via the AICtab function from the bbmle package (version 1.0.24). The function reports the differences in AIC from the best model, with lower AIC scores being desirable. We also report a likelihood-ratio (LR) test between models using the lrtest function from the lmtest package (version 0.9-39). Next, having selected a model with a relatively higher quality random effects structure, we compared the estimates of this model with a standard version that contains the same predictors but omits the random effects. Again, we used AIC to compare the two models to determine whether the inclusion of subject-level variance produces a relatively better fit to the data.

Lastly, to evaluate overall model significance, we compared the full version of the final model with a null version that omits the predictors (but retains the same covariates). Model comparisons are made and reported as an LR test.

3.3. Results

For all models, the more complex random effect structure involving by-subject random intercepts and slopes was ultimately selected over random intercepts only (Appendix B: Table B.1 for results). The mixed effects model also provided a better fit than a model that does not account for the dependency among data points via random effects parameters (Appendix B: Table B.2). The comparison of the full models versus their respective null version showed that each was statistically significant at the $p < .001$ level (Appendix B: Table B.3).

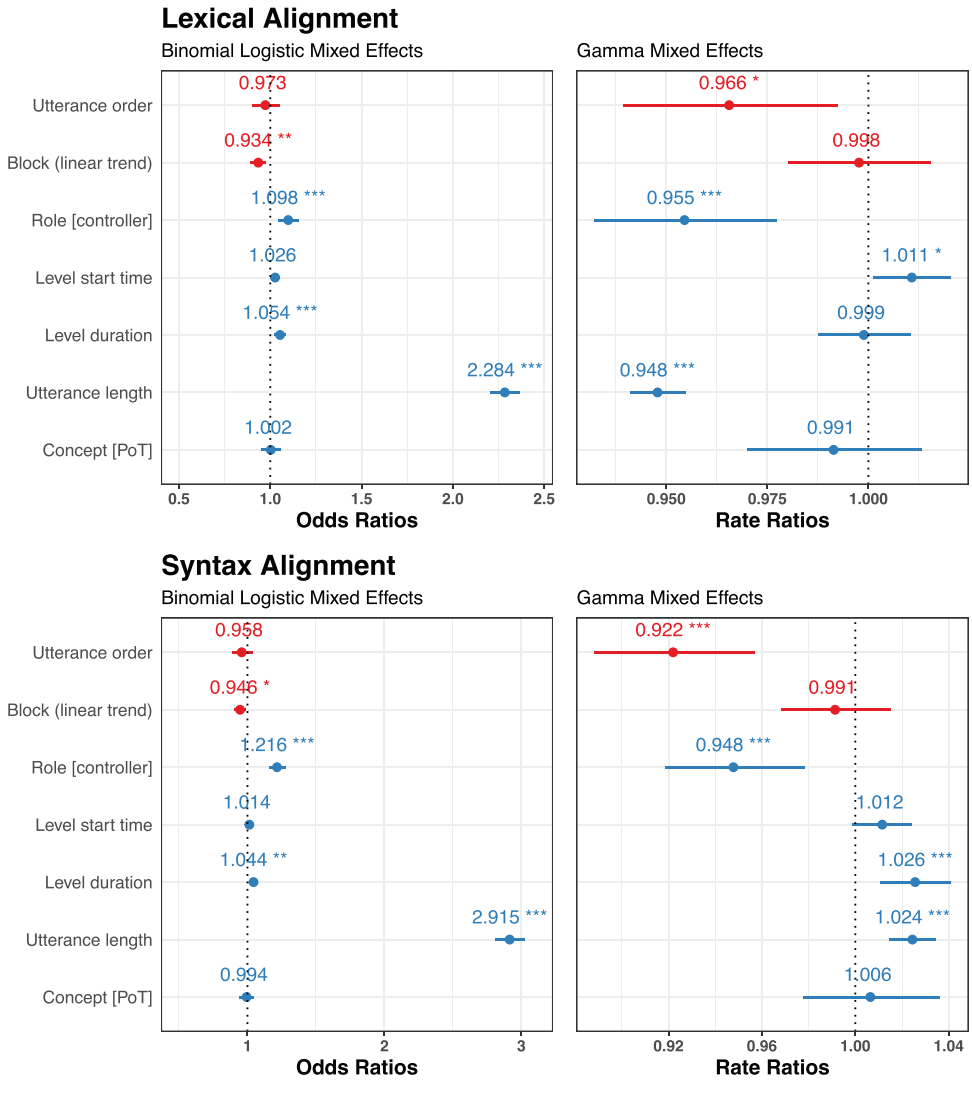

Fig. 5. Forest plots with exponentiated results for lexical alignment and syntactic alignment. The binomial logistic mixed effects models correspond to the presence of alignment, whereas gamma mixed effects models correspond to the magnitude of alignment.

3.3.1. Presence: Lexical and syntactic alignment

The exponentiated model coefficients (to facilitate interpretation) are reported in Fig. 5. For the binomial logistic mixed effects models, there was no statistically significant difference for presence of alignment from the start to end of the Physics Playground level (lexical: log effect estimate = –0.027, Wald CI –0.11, 0.05; syntactic: log effect estimate = –0.043, Wald CI –0.12, 0.03). However, the linear trend across rounds showed a lower presence of alignment as rounds progressed, with a decrease of approximately 6.57% for lexical alignment (log effect

estimate: –0.068, Wald CI –0.11, –0.02, $p = .002$) and 5.45% for syntactic alignment (log effect estimate: –0.056, Wald CI –0.10, –0.01, $p = .013$).

Estimates of between-subject standard deviations were 0.187 and 0.182 for lexical and syntactic alignment, respectively, indicating low to moderate variation among subjects.

The statistically significant covariates also showed that the presence of alignment was more likely for participants in the controller role (vs. contributor) (lexical: log effect estimate = 0.094, Wald CI 0.04, 0.14, $p < .001$; syntactic: log effect estimate = 0.195, Wald CI 0.15, 0.24, $p < .001$, and as the duration of a level increased (lexical: log effect estimate = 0.052, Wald CI 0.02, 0.08, $p < .001$; syntactic: log effect estimate = 0.043, Wald CI 0.01, 0.07, $p = .004$). And as might be expected, as a turn increased in length, the likelihood of alignment increased (lexical: log effect estimate = 0.826, Wald CI 0.79, 0.86, $p < .001$; syntactic: log effect estimate = 1.070, Wald CI 1.03, 1.11, $p < .001$). Level start time and concept (PoT vs. EcT) did not predict either lexical or syntactic alignment presence.

3.3.2. Magnitude: Lexical and syntactic alignment

Based on the gamma-distributed mixed effects models, for the magnitude of alignment from the start to the end of the Physics Playground levels, there was a 3.44% reduction in the mean of lexical alignment (log effect estimate = –0.035, Wald CI –0.06, –0.01, $p = .012$) and a 7.78% reduction in the mean of syntactic alignment (log effect estimate = –0.081, Wald CI –0.12, –0.04, $p < .001$). However, unlike the presence of alignment, there was no statistically significant difference for the magnitude of alignment as rounds progressed (lexical: log effect estimate = –0.002, Wald CI –0.02, 0.02; syntactic: log effect estimate = –0.009, Wald CI –0.03, 0.01).

As with presence, the standard deviations of between-subjects variance values (.075 and.088 for lexical and syntactic alignment, respectively) indicate minimal variation among subjects for alignment magnitude.

Several covariates for lexical and syntactic alignment were also statistically significant. For both, when participants were assigned to the controller versus contributor role, the likelihood of greater magnitude of alignment decreased (lexical: log effect estimate = –0.047, Wald CI –0.07, –0.02, $p < .001$; syntactic: log effect estimate = –0.054, Wald CI –0.09, –0.02, $p < .001$) and for longer turns, the likelihood of greater magnitude of alignment decreased for lexical (log effect estimate = –0.054, Wald CI –0.06, –0.05, $p < .001$) but increased for syntactic (log effect estimate = 0.024, Wald CI 0.01, 0.03, $p < .001$). Further, as a level appeared later in a round (i.e, start time), there was a higher likelihood of greater lexical (but not syntactic) alignment (log effect estimate = 0.011, Wald CI 0.00, 0.02, $p = .027$), whereas syntactic (but not lexical) alignment was likely to be greater as the duration of a level increased (log effect estimate = 0.025, Wald CI 0.01, 0.04, $p < .001$). Once again, concept did not predict lexical or syntactic alignment magnitude.

3.3.3. Magnitude: Semantic alignment

Table 2 shows the results for all predictor variables. Of the primary predictors, there was a statistically significant increase of semantic alignment magnitude as turns progressed within Physics Playground levels. However, the linear trend of semantic alignment across rounds

Table 2
Predicting semantic alignment scores for adjacent turns across time

	Semantic		
Predictors	*Estimates*	*CI*	*p*
Utterance order	0.010	0.004–0.016	**.002**
Block (linear trend)	0.000	–0.003 to 0.004	.830
Role [controller]	0.016	0.011–0.020	**<.001**
Level start time	0.000	–0.002 to 0.002	.803
Level duration	0.001	–0.001 to 0.004	.216
Utterance length	0.051	0.050–0.053	**<.001**
Concept [PoT]	0.001	–0.003 to 0.005	.629

Note. The table presents the estimates, confidence intervals, and *p*-values, as predicted by turn order and round. Values in bold indicate significance at the $p < .05$ level. Participant and game-level related covariates are also reported.

was not statistically significant. For covariates in the model, the effect of role was statistically significant, such that the controller showed a tendency to align to a greater magnitude than the contributors, and the effect of turn length was also statistically significant; longer turns tended to have higher semantic alignment values.

For the overall model, the estimate of among-subject standard deviation was low, with a value of 0.019.

3.3.4. Baseline data for comparison

In examining patterns of linguistic alignment, it is assumed that the repetition of lexical items and syntactic phrases between contiguous turns are due to sensitivities of what a particular partner is saying as it is being said in real time. Nevertheless, it is still possible that similar frequency patterns are observed that are not partner-specific but rather driven by contextual constraints, where a limited set of referential items and the structure of the environment restrict how language can be used. To control for this possibility, we created mock conversations for each game level that preserved the original speakers and the order in which they spoke, but the utterance attributed at each turn to the speaker was a randomly selected utterance from the same speaker within the same dialogue. Because no utterance was repeated, the overall frequency of lexical items and syntactic phrases were identical to the original. The crucial disruption was in the assumed structural information driven by the potential dependencies of what was said by one partner as an immediate (contiguous) response to another. The ALIGN analysis was rerun on these mock conversations to generate baseline turn-by-turn linguistic alignment scores. The same statistical analyses as performed on the original data were replicated with the baseline data to determine whether similar statistical results could be reproduced.

When the same analyses were run using the baseline data of mock conversations, there were no statistically significant fixed effects for the main predictors. Interestingly, the overall amount of alignment between the real and mock conversations did not vary greatly, indicating

what matters most between real and mock conversations is how alignment changes over time. Detailed results from statistical models are reported in Appendix C.

4. Analysis 2: Linguistic alignment and CPS outcomes

For each team, we also identified the start and stop times of each attempted level, and recorded whether the team had achieved a gold, silver, or no coin for that level as our outcome measure. We then regressed CPS outcome on the different types of linguistic alignment scores.

4.1. Statistical modeling

We built an ordinal regression mixed-effects models to predict the likelihood of solving a level (i.e., silver/gold coin vs. no coin) based on alignment amount (mean-centered), with an additional interaction term with within-level progression of alignment over turns/time (mean-centered). Models were adjusted for game-play level characteristics hypothesized to most influence the potential for earning a coin, including which round the level occurred, the level duration (z-scored), start time of each level within each round (z-scored), whether the level was revisited (vs. first attempt), and the concept (EcT or PoT) being presented in each level. A random effect term for the clustered variance of team was included (given coin attainment is fundamentally the product of team interaction). For the primary model, all linguistic alignment variables were entered at once to predict performance. Doing so provides insight as to whether each predictor is uniquely informative when we take into account the presence of the other linguistic variables. We also report the results of separate models for each linguistic variable. Analyses for performance prediction were conducted with the use of the "ordinal" R package (version 2019.12-10).

4.2. Results

As can be seen in Table 3, only semantic alignment contributes unique information to CPS performance. For every one unit increase in semantic alignment, the odds of being more likely to earn a coin decreases ($OR = 0.841$, $p = .035$). For the random effect, the estimate of among-team standard deviation was 1.033, with an inter-class correlation of 0.234. There were no interactions with turn order, suggesting that neither an increase or decrease of alignment across conversational turns was predictive of performance outcome.

For models run separately for each linguistic predictor, a similar pattern emerged where only semantic alignment was associated with decreased odds in earning a coin ($OR = 0.845$, Wald CI 0.735, 0.972, $p = .019$). However, for syntactic alignment, when its effects are examined without taking into account the frequency of the other variables, there was a marginally statistically significant effect also showing an association between greater alignment and decreased odds of earning a coin ($OR = 0.784$, Wald CI 0.599, 1.026, $p = .076$). Lexical alignment remained statistically insignificant ($OR = 0.936$, Wald CI 0.819, 1.069, $p = .327$)

Table 3
Predicting task outcomes based on linguistic modality alignment and interactions with turn order

	Trophy		
Predictors	*Odds ratios*	*CI*	*p*
Lexical	1.050	0.894–1.234	.551
Syntax	0.854	0.639–1.142	.287
Semantic	0.841	0.716–0.988	**.035**
Lexical * Time	0.836	0.484–1.443	.519
Syntactic * Time	0.795	0.296–2.134	.649
Semantic * Time	0.978	0.568–1.685	.937
Block (linear trend)	0.633	0.604–0.663	**<.001**
Level Duration	0.286	0.276–0.297	**<.001**
Level Start Time	0.381	0.369–0.393	**<.001**
Revisit [yes]	0.667	0.571–0.779	**<.001**
Concept [PoT]	2.733	2.578–2.898	**<.001**

Note. To interpret the odds ratios, the odds of obtaining a coin is either increased (above 1) or decreased (below 1) as amount of alignment increases. Level- and turn-related covariates are also included in each model.

All covariates entered into the models were statistically significant. Again, as shown in Table 3, the odds of earning a coin was less likely for levels that occurred later within rounds (Level Start Time), that occurred later across rounds (Block [linear trend]), that were longer in duration (Level Duration), and in levels that the teams revisited after already experiencing a prior impasse (Revisit [yes]). Levels dealing with the concept of "property of torque (PoT)" over "transfer of energy" were also more likely to receive an earned coin (Concept [PoT]).

5. Discussion

There is a need to better understand the various forms of linguistic alignment as contextually driven and temporally evolving phenomena in complex tasks. The goal here was to do so within a CPS scenario. Not only is this important theoretically, but also practically, as the future of work and education has undergone a pronounced shift toward nonroutine analytical tasks that require team-based solutions (Levy & Murnane, 2012; Wüstenberg et al., 2012). These tasks often involve solutions that can be arrived at from many possible directions and require creative exploration (Graesser et al., 2018). It is also where people's abilities to use language to communicate and coordinate meaning is often maximally challenged (Dillenbourgh & Traum, 2006; Fiore & Salas, 2004).

In the current work, we explored how interactive linguistic alignment changes in CPS interactions: across more localized levels (turns) to more large-scale changes across rounds. We also evaluated whether the amount of alignment of different linguistic types predicted the likelihood of success in each task. Our exploration took place in a virtual environment where some of the control common to lab-based, dyadic problem-solving tasks was sacrificed for nonroutine, analytical, and dynamic interactions involving teams of three. Despite the

complexity, we found unique statistically significant patterns of alignment. Of most interest are those patterns that provide a glimpse into how teams linguistically ground meaning and its implications for associated cognitive processes.

5.1. Trends in alignment over time within levels and across rounds

Notably, and consistent with our predictions, there was a decrease in lexical alignment over time within a level suggesting that precision, that is, referring to the same things in the same way, diminished as participants gained greater familiarity with the problem domain and each other's understanding. There was also an apparent trade-off of lexical repetition with an increase in semantic alignment, pointing to greater convergence of shared topics and themes over time.

The above pattern was most evident with alignment measures of magnitude (opposed to presence) and as expressed within levels (opposed to across rounds). This is likely because each level resets with a new and unfamiliar problem to solve, and thus the need for precision should also reset in a more pronounced way. That this occurs with magnitude alone suggests that precision is not established by the simple frequency (presence) of alignment events (coded as a binary outcome of present or not), but instead by the depth of the alignment; where greater magnitude (continuous, positive values) is associated with an increased number of lexical items or longer phrases being repeated.

Although the alignment measure for presence did not vary within levels, it did decrease across rounds. It is important to note that when evaluating alignment from round to round, a more general behavior is being captured that has less to do with the particular demands of any level, but more on how participants are coordinating meaning globally. Presumably, after each 15-min round of interaction, participants are becoming more familiar with each other and potentially more adaptive. It appears that participants needed to "check in" with each other less with the repetition of phrases or lexical items, thus a decreasing frequency (presence) of alignment across rounds. But when participants did touch base, there is apparently no appreciable change in the depth (magnitude) of that alignment.

Another identified pattern was a decrease of syntactic alignment over time for magnitude within levels, and a decrease over time for presence across rounds. Given our computation of syntactic alignment excluded lexical overlap (thus minimizing a lexical boost effect), it suggests mechanisms at play that are less directly tied to meaning. Based on a cognitive processing or partner attunement account, decreasing syntactic alignment might be associated with less cognitive demands with practice, or a decreasing attunement to others, or both. Although the current study is unable to directly assess the exact mechanism, it does raise interesting possibilities for future work. For example, coordination processes that require greater cognitive demands or attunement might occur earlier within our problem-solving levels, thus higher syntactic alignment, that then give way to decreased demand or attunement as coordination shifts to new processes. Indeed, by recognizing that CPS interactions generally involve qualitative phases of team-level (distributed) processes (Bales & Strodtbeck, 1951; Fiore et al., 2010), a closer examination of linguistic alignment might help reveal otherwise hidden trajectories.

The patterns of alignment found here also find support across diverse theoretical perspectives in how shared meaning is forged during communication. For example, in the area of experimental semiotics, a notable finding is that in highly interactive tasks where partners are asked to create and interpret novel communication systems, the shared depictions that emerge become simpler and less complex over time, similar to what happens with the creation of conceptual pacts during experimental tasks using natural language (Galantucci & Garrod, 2011; Nolle & Galantucci, 2023). Despite this reduction of precision, communication remains efficient so long as the depictions maintain a high degree of semantic complexity (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007). This process is not so different in kind from what we see in our study. There is a strong possibility that partners were trading one form of precision for another. Across each of the game-play levels, although repetition of lexical referents tended to diminish, the referential cohesion supported by this repetition was still present but now captured in the overlap of semantically similar utterances. This interpretation also underscores the significance of using distributional semantic models, such as word2vec, for alignment research. Although the meanings they represent are incomplete approximations of the conceptual richness that emerges during conversational interaction, they do serve as useful tools that allow for a more comprehensive and flexible understanding of referential overlap that is not captured by lexical repetition alone.

5.2. Patterns of alignment associated with collaborative success

For each CPS interaction in our dataset, we also measured whether a solution was achieved and the quality of the solution. For the relationship between alignment and task performance, the predominant pattern found across CPS interactions (i.e., with respect to decreased lexical and syntactic alignment over time) did not translate into better performance. Indeed, not only was semantic alignment alone predictive, it also ran counter to intuitions based on simple communicative tasks that greater convergence should lead to better shared understanding and thus better problem-solving. But much depends on unique task goals and the communicative context. In our study, it is plausible that less semantic alignment predicts better performance. In tasks which require multiple possibilities to be entertained, too great of perseveration on any single solution, as captured by high semantic alignment, leads to detrimental dead-ends. What is instead needed is greater complementarity and exploration of ideas, where team members contribute different content and perspectives (also see Dideriksen et al., 2019).

This interpretation is one that is bolstered by recent brain-based evidence examining how communication partners establish "conceptual alignment" (Stolk, Verhagen, & Toni, 2016). In this framework, partners converge on a shared understanding by exploring and finding diverse connections within a large web of conceptual, and often unrelated, possibilities. Much of this process is marked by mutual inference of what others might know given the evolving interactional context (at many timescales, including the individual and shared histories of those involved). Inferences are achieved by using words to probe and bias the fleeting conceptual structures, but the utterances themselves do not contain a priori the meaning that emerges at the conceptual level. This is evidenced in neural regions of mutual understanding that are invoked at temporal scales independent from the occurrence of the linguistic

signal itself. Relating this to the current findings, success in our interaction task required participants to problem solve within a highly generative communicative scenario where the set of open-ended possibilities is greater than that of more rigorously controlled experimental approaches. If "conceptual alignment," as the work by Stolk et al. (2016) suggests, is associated with a superior ability in navigating a shared conceptual space through the exploration-exploitation of many possibilities, in which the words used are fleeting reflections of this process rather than constitutive of it, than alignment at the signal level (lexical and syntactic) should not carry much information. Moreover, the process of "conceptual alignment" is one where semantic alignment—a measure that is fundamentally capturing a type of referential consistency—would see greater variation (and thus a decrease in overall magnitude).

5.3. Limitations and additional considerations

To structure the current work and motivate our analyses, we provided a set of expectations from existing theory on how linguistic alignment and CPS success might be related: specifically, an expectation that there will be a decrease in lexical and syntactic alignment with an increase in semantic. What we found instead was that decreased semantic alignment overall to be associated with success. Although we have already discussed an alternative theoretical viewpoint to address these findings, there is always the possibility that our task simply fails to capture CPS in a meaningful way, and this itself could explain the unexpected relationships. There is some support to this view as only one participant had control over the computer mouse during game play, possibly curtailing collaboration. Acknowledging this possible limitation, we would also point out that the task still has all the elements identified in related work as necessary for successful collaboration (Szewkis et al., 2011). Through screen sharing, actions performed by participants with control over the mouse were observed by their partners and thus they were accountable for these actions, and there was continual awareness among all participants about the current state of their partners and the team. Participants also submitted joint answers and received joint outcomes/rewards in return (e.g., trophies awarded, levels completed), and they indicated positive interdependence in follow-up assessments. Overall, the task required team members to work together and communicate effectively in pursuit of shared objectives.

Although there is a good reason to assert collaboration was essential for our task, what is less clear is the extent in which collaboration was driven solely by the linguistic signal. The task demands were such that participants controlling the computer mouse could signal understanding of a solution by actions taken in the visually shared workspace. If collaboration is offloaded in this way, there is less of a need to signal and coordinate meaning across verbal turns. Indeed, the problem-solving levels with the most elegant solutions were also those where language was least used, as seen with the "gold coin" solutions that were solved with the fewest turns (even after retaining only those levels with 20 or more turns; see Fig. 2). Nevertheless, this sensitivity to the broader communicative domain points to a reality of complex communicative interactions where cognition is distributed among people and into the task environment. Optimal CPS is inherently a multimodal phenomenon where individual

behavioral channels become more or less pronounced (like language) given changing demands and available resources.

The reality of this situation also points to the need for more comprehensive analytical approaches that go beyond single communicative channels and automated techniques. Although a great deal of information can be gained with using a tool like ALIGN that focuses only on transcribed utterances, integrating with qualitative investigations would allow for more nuanced understandings of how the nonverbal cues, pauses, hesitations, gestures, and drawings within the visually shared workspace play a pivotal role in shaping shared understanding (Cornejo, Cuadros, Morales, & J, 2017; Hadley, Naylor, & Hamilton, 2022; Obhi & Cross, 2016). These insights range from the emotional, social, and cognitive; be it in how participants nod their heads in agreement, raise their eyebrows in surprise, or cross their arms defensively; to how participants might emphasize a point, invite a response, or signal uncertainty through a well-timed pause; or in how participants signal cognitive processing, doubt, or a desire with hesitations via various "speech fillers" (which are currently ignored given the difficulty of consistently transcribing accurately). There are also rich gestural and visually drawn depictions afforded by our task to consider. Physics concepts naturally lend themselves to gestures and visual aids to indicate direction, size, or movement, which are all integral for conveying complex concepts and in clarifying ambiguities (Johnson-Glenberg & Megowan-Romanowicz, 2017). Future work will need to more effectively capture this information with a combination of quantitative and qualitative methods, a challenge that generally applies to all research programs that examine the shared understanding-building processes in conversation.

As another consideration raised by the current study, there is an important distinction between "conceptual alignment" and "semantic alignment" that we have indirectly drawn thus far but is worth explicating as it calls into question the ability of distributional semantic models to ever adequately capture mutual understanding in human communication. Although we believe that distributional semantic models like word2vec are useful tools for shedding new light on communicative processes (Günther, Rinaldi, & Marelli, 2019; Kumar, 2021, also see), their capabilities should not be confused with how meaning is actually established and deployed in conversation. What cannot be captured is a conceptual alignment that is fundamentally informed by the specific interactions in which they occur (Stolk et al., 2016). That is, meaning is very much constituted by the demands of the interaction itself, and what words mean within and across conversational turns are tuned to maintaining the coherence of the ongoing interactive system (Rączaszek-Leonardi et al., 2014). What is said, or what one anticipates saying, is useful insofar, that is, organizes experience for those involved in mutually relevant ways. From this perspective, word-level meaning is emergent and participatory. It is jointly structured by environmental and contextual constraints. It is determined by its ability to coordinate individuals into a functional system. Word2vec and similar models, no matter their computational complexity and impressive feats of approximating semantic knowledge, fail to capture the full extent of conceptual complexity during real-time interactions because they are based on algorithms that encode/decode information from stored representations established prior to that interaction.

What then do distributional semantic models provide for us in the context of our study? Although they do not capture the conceptual richness of the words as they functioned when originally uttered, nor for that matter nonlinguistic, sensorimotor experiences that undoubtedly shape meaning (Binder et al., 2016; Frisby, Halai, Cox, Lambon Ralph, & Rogers, 2022), they do allow insight into the semantic relatedness between utterances that are not restricted to strict lexical overlap. But it is important to keep in mind that this semantic relatedness is ultimately that which exists in a high-dimensional vector space derived from pre-existing text (in the case of our use of word2vec, billions of words from Google News).

As mentioned earlier, it has also been shown that CPS interactions involve distinct qualitative phases where various competencies are more or less pronounced. We know with certainty from previous analyses that participants in our study are engaging in a number of critical competencies, such as actively constructing a shared knowledge, negotiating among possible solutions, and testing/revising agreed-upon solutions (Sun et al., 2020, 2022). What is yet unexplored is how these are possibly clustered. This leads to possibilities for future work in analyzing episodes of convergence and divergence as distinct phases in CPS coordination processes. New techniques need to be devised that go beyond our current scale of linguistic alignment analysis —as something that linearly changes over time—to something that captures distributed patterns that map onto critical phases (Wiltshire, Butner, & Fiore, 2018, for an example using nonlinguistic data). Even so, we believe our general approach lends support to process models of dialogue that prioritize the role of parallel (graded) and distributed mechanisms in coordinating meaning (Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Rączaszek-Leonardi et al., 2014). In what is a "synergistic" process model, complex communicative interactions are best characterized by an interplay of convergent and divergent mechanisms. The current work suggests that it is possible for these trade-offs to be simultaneously expressed across different channels of linguistic alignment (at least at one level of analysis).

We also recognize that the methods for studying social interaction exist along a continuum, from maximally naturalistic observation on one end to completely controlled experimentation on the other (Kendrick, 2017). For the current work, our objective was aimed somewhere near the middle, where social interaction was placed in a controlled situation that allowed detailed measurements and a wide range of known variables, and at the same time, with reduced constraints on the generative possibilities of naturalistic conversational communication. A limitation of this approach is that the study is not necessarily targeted to a focused set of predictions where the manipulations allow clear-cut causal claims. Although this study was originally designed to allow the analyses that we have reported, the results are only correlational in nature. Future work will need to systematically manipulate the need for precise information sharing to evaluate its impact on alignment. It will also need to better account for a range of task goals and team members' individual differences. And more precise measures of linguistic alignment are necessary that go beyond immediate turn-taking. However, it is also necessary that knowledge be advanced in more naturalistic and complex problem-solving scenarios. Although this introduces new challenges, it will provide a richer picture of how collaborative dialogue unfolds, and how it potentially can be improved.

Acknowledgments

This work was supported by the US National Science Foundation (award number 1660859; DRL 2019805; DUE 1745442/1660877). We would also like to thank Angela Stewart, Amanda Michaels, and Caroline Reinhardt for their help collecting the data used in this study.

References

Amon, M. J., Vrzakova, H., & D'Mello, S. K. (2019). Beyond dyadic coordination: Multimodal behavioral irregularity in triads predicts facets of collaborative problem solving. *Cognitive Science*, *43*, e12787.

Angus, D., Smith, A., & Wiles, J. (2012a). Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics*, *18*, 988–997.

Angus, D., Smith, A., & Wiles, J. (2012b). Human communication as coupled time series quantifying multi-participant recurrence. *IEEE Transactions on Audio, Speech and Language Processing*, *20*, 1795–1807.

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising conversation structure across time: Insights into effective doctor–patient consultations. *PLoS ONE*, *7*, 22693629.

Baker, R., Angus, D., Smith-Conway, E. R., Baker, K. S., Gallois, C., Smith, A., Wiles, J., & Chenery, H. J. (2015). Visualising conversations between care home staff and residents with dementia. *Ageing & Society*, *35*, 270–297.

Bales, R. F., & Strodtbeck, F. L. (1951). Phases in group problem-solving. *Journal of Abnormal and Social Psychology*, *46*, 485.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*, 130–174.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. Sebastopol, CA: O'Reilly Media.

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Branigan, H., & Pickering, M. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*, 1–18.

Branigan, H., Pickering, M., & Celeland, A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, B13–25.

Branigan, H., Pickering, M., McLean, J., & Cleland, A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, *104*, 163–197.

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, *22*(6), 1482–1493.

Brennan, S., Galati, A., & Kuhlen, A. (2010). Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, *53*, 301–344.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*, 62–89.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–1134.

Chen, S., Shute, V., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672.

Clark, H. (1992). *Arenas of language use*. University of Chicago Press.

Clark, H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H., & Brennan, S. (1991). *Grounding in communication*. Washington, DC: APA Books.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Cleland, A., & Pickering, M. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*, 214–230.

Cornejo, C., Cuadros, Z., Morales, R., & Javiera, P. (2017). Interpersonal coordination: Methods, achievements, and challenges. *Frontiers in Psychology*, *8*, 42–54.

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. In B. Ross (Ed.), *Psychology of learning and motivation* (pp. 43–95). Elsevier.

Dideriksen, C., Fusaroli, R., Tylén, K., Dingemanse, M., & Christiansen, M. (2019). Contextualising conversational strategies: Backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations. In *CogSci'19*, Cognitive Science Society.

Dillenbourgh, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in mutlimodal collaborative problem solving. *Journal of the Learning Sciences*, *15*(1), 121–151.

Duran, N., Dale, R., & Galati, A. (2016). Toward integrative dynamic models for adaptive perspective taking. *Topics in Cognitive Science*, *8*, 761–779.

Duran, N., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techniques–A Python library. *Psychological Methods*, *24*(4), 419–438.

Eloy, L., EB Stewart, A., Jean Amon, M., Reinhardt, C., Michaels, A., Sun, C., Shute, V., Duran, N. D., & D'Mello, S. (2019). Modeling team-level multimodal dynamics during multiparty collaboration. In *2019 International Conference on Multimodal Interaction (ICMI'19)* (pp. 244–258).

Fiore, S., Rosen, M., Smith-Jentsch, K., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, *52*(2), 203–222.

Fiore, S., & Salas, E. (2004). *Why we need team cognition*. Washington, DC: American Psychological Association.

Foltz, A., Gaspers, J., Meyer, C., Thiele, K., Cimiano, P., & Stenneken, P. (2015). Temporal effects of alignment in text-based, task-oriented discourse. *Discourse Processes*, *52*, 609–641.

Foltz, P., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*, 285–307.

Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A., & Rogers, T. T. (2022). Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences*, *27*, 258–281.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, *23*(8), 931–939.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience*, *5*, 1–15.

Galati, A., Dale, R., & Duran, N. (2019). Social and configural effects on the cognitive dynamics of perspective-taking. *Journal of Memory and Language*, *104*, 1–24.

Gallois, C., Ogay, T., & Giles, H. (2005). Communication accommodation theory: A look back and a look ahead. In W. Gudykunst (Ed.), *Theorizing about intercultural communication* (pp. 121–148). Thousand Oaks, CA: Sage

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 187–218.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*, 961–987.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*, 1006–1033.

Gorman, J. C., & Wiltshire, T. J. (2024). A typology for the application of team coordination dynamics across increasing levels of dynamic complexity. *Human Factors*, *66*, 5–16.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, *19*, 59–92.

Gries, S. (2005). Syntactic priming: A corpus-based approach. *Psycholinguistic Research*, *34*, 365–399.

Hadley, L., Naylor, G., & Hamilton, A. (2022). A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, *1*, 42–54.

Healey, P., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PLoS ONE*, *9*(6), e98598.
Heyselaar, E., & Segaert, K. (2019). Memory encoding of syntactic information involves domain-general attentional resources: Evidence from dual-task studies. *Quarterly Journal of Experimental Psychology*, *72*, 1285–1296.
Horton, W., & Gerrig, R. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*, 127–142.
Howes, C., Healey, P., & Purver, M. (2010). Tracking lexical and syntactic alignment in conversation. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 2004–2009).
Johnson-Glenberg, M., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality: How gesture and motion capture affect physics education: Principles and implications. *Cognitive Research: Principles and Implications*, *2*, 1–28.
Kendrick, K. (2017). Using conversation analysis in the lab. *Research on Language and Social Interaction*, *50*, 1–11.
Kumar, A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*, 40–80.
Levy, F., & Murnane, R. (2012). *The new division of labor: How computers are creating the next job market*. New York: Princeton University Press.
Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
Mills, G. (2014). Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, *32*, 158–174.
Nolle, J., & Galantucci, B. (2023). Experimental semiotics: Past, present, and future. In A. García& A. Ibáñez (Eds.), *The Routledge Handbook of Semiosis and the Brain* (pp. 1–16). New York: Routledge
Noppeney, U., & Price, C. J. (2004). An fMRI study of syntactic adaptation. *Journal of Cognitive Neuroscience*, *16*, 702–713.
Obhi, S. S., & Cross, E. S. (2016). *Shared representations: Sensorimotor foundations of social life*. Cambridge University Press.
Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–226.
Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633–651.
Rączaszek-Leonardi, J., Dębska, A., & Sochanowicz, A. (2014). Pooling the ground: Understanding and coordination in collective sense making. *Frontiers in Psychology*, *5*, 1233.
Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*, 1–29.
Reitter, D., Keller, F., & Moore, J. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*, 587–637.
Reitter, D., Keller, F., & Moore, J. D. (2006). Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL* (pp. 121–124).
Roche, J., & Arnold, H. (2018). The effects of emotion suppression during language planning and production. *Journal of Speech, Language, and Hearing Research*, *61*, 2076–2083.
Sagi, E., & Diermeier, D. (2017). Language use and coalition formation in multiparty negotiations. *Cognitive Science*, *41*, 259–271.
Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge University Press.
Schoot, L., Menenti, L., Hagoort, P., & Segaert, K. (2014). A little more conversation–The influence of communicative context on syntactic priming in brain and behavior. *Frontiers in Psychology*, *5*, 1–16.

Shute, V., Ventura, M., & Kim, Y. (2013). Assessment and learning of qualitative physics in Newton's playground. *Journal of Educational Research*, *106*, 423–430.

Slocome, K., Alvarez, I., Brenigan, H., Jilema, T., Burnett, H., Fischer, A., Li, Y., Garrod, S., & Levita, L. (2013). Linguistic alignment in adults with and without Asperger's syndrome. *Journal of Autism and Developmental Disorders*, *43*, 1423–1436.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, *20*, 180–191.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672.

Sun, C., Shute, V. J., Stewart, A. E., Beck-White, Q., Reinhardt, C. R., Zhou, G., Duran, N., & D'Mello, S. K. (2022). The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior*, *128*, 107120.

Szewkis, E., Nussbaum, M., Rosen, T., Abalos, J., Denardin, F., Caballero, D., Tagle, A., & Alcoholado, C. (2011). Collaboration within large groups in the classroom. *Computer Supported Learning*, *6*, 561–575.

Thothathiri, M., & Snedeker, J. (2008). Give and take: Syntactic priming during spoken language comprehension. *Cognition*, *108*, 51–68.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 173–180).

Wiltshire, T. J., Butner, J. E., & Fiore, S. M. (2018). Problem-solving phase transitions during team collaboration. *Cognitive Science*, *42*, 129–167.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving-more than reasoning? *Intelligence*, *40*(1), 1–14.

Appendix A: Comparison of turns and duration for levels by outcome

Linear mixed-effects models were used to assess the overall duration of levels and number of turns with level outcome (gold, silver, and none) as the fixed effect, with a random effect for team. The overall effect for level outcome was statistically significant based on a likelihood ratio test between versions of the model with and without the fixed effect, duration: $\chi^2(2) = 55.951$, $p<.001$; and number of turns: $\chi^2(2) = 30.726$, $p<.001$. Planned contrasts between outcome types are shown in Tables A.1 and A.2.

Table A.1
Model results involving **duration** of levels based on gold, silver, and none outcomes

Contrast	Estimate	SE	*t*	*p*
None—Silver	56.4	16.9	3.346	.0009
None—Gold	161.3	21.7	7.416	<.0001
Silver—Gold	104.8	22.6	4.642	<.0001

Table A.2
Model results involving **number of turns** in levels based on gold, silver, and none outcomes

Contrast	Estimate	SE	*t*	*p*
None—Silver	5.84	2.94	1.988	.0472
None—Gold	20.97	3.79	5.532	<.0001
Silver—Gold	15.13	3.92	3.858	.0001

Appendix B: Statistical tests for model selection considerations

Table B.1
Tests to compare whether a more complex model with by-subject random slopes for role (role|subject) provided a better fit than a model with by-subject random intercepts (subject) alone

	Lexical				Syntax				Semantic	
	Binomial logistic		Gamma		Binomial logistic		Gamma		linear	
	role|subject	subject	role|subject	subject	role|subject	subject	role|subject	subject	role|subject	subject
dLogLik	3.3	0	23.5	0	0.3	0	29.3	0	7.6	0
dAIC	0	2.6	0	43	3.4	0	0	54.5	0	11.1
df	12	10	13	11	12	10	13	11	13	11
LR-test	6.602, $p = .037$	–	46.995, $p < .001$	–	0.607, $p = .738$	–	58.513, $p < .001$	–	15.132, $p = 005$	–

Table B.2
Tests to compare whether models with or without random effects structure (GLMM vs. GLM) are a better fit to the data

	Binomial logistic		Gamma		Binomial logistic		Gamma		Linear	
	GLMM	GLM	GLMM	GLM	GLMM	GLM	GLMM	GLM	LME	LM
dLogLik	18.6	0	95.4	0	28.6	0	48	0	6.9	0
dAIC	0	31.9	0	184.7	0	51.2	0	90	0	7.8
df	12	9	13	10	12	9	13	10	13	10

Note. Nonzero values indicate the model with the higher log-likelihood/AIC.
Abbreviations: dLogLik, difference of log-likelihoods; dAIC, difference of AIC.

Table B.3
Tests of overall statistical significance of full model, comparing the null model (without key predictors) with the full model (with predictors)

	Lexical				Syntax				Semantic	
	Binomial logistic		Gamma		Binomial logistic		Gamma		linear	
	Null	Full	Null	Full	Null	Full	Null	Full	Null	Full
dLogLik	0	1498.1	0	110.1	0	2412.7	0	37	0	1510.7
dAIC	2980.2	0	204.2	0	4809.4	0	58	0	3009.3	0
df	4	12	5	13	4	12	5	13	5	11
LR-test	–	2996.2, $p<.001$	–	220.16, $p<.001$	–	4825.4, $p<.001$	–	74.026, $p<.001$	–	3021.3, $p<.001$

Note. Nonzero values indicate the model with the higher log-likelihood/AIC.
Abbreviations: dLogLik, difference of log-likelihoods; dAIC, difference of AIC.

Appendix C: Mock conversations (baseline)

C.1Overall amount of alignment between real and mock conversations

Table C.1 shows the average amount of lexical, syntactic, and semantic alignment between real and mock conversations collapsed across temporal ordering of turns and rounds.

C.2Presence: Lexical and syntactic alignment

Tables C.2 and C.3 show the results for the models based on binomial component of the hurdle model (using the logit link) that predict the presence of alignment. The two main predictors "Utterance order" and "Block" are statistically nonsignificant for both lexical and syntactic alignment.

C.3Magnitude: Lexical and syntactic alignment

Tables C.4 and C.5 show the results for the models based on gamma-distributed component of the hurdle model that predict mean (i.e., magnitude) of alignment. The two main predictors "Utterance order" and "Block" were statistically nonsignificant for both lexical and syntactic alignment.

C.4Magnitude: Semantic alignment

Table C.6 shows the results for the models based on a linear mixed-effects regression to evaluate the mean (i.e., magnitude) of alignment. The two main predictors "Utterance order" and "Block" were statistically nonsignificant for semantic alignment.

Table C.1
Mean amount of alignment for each linguistic type (standard deviations in parenthesis)

	Real	Mock
Lexical	0.157 (0.174)	0.115 (0.137)
Syntax	0.06 (0.086)	0.057 (0.062)
Semantic	0.592 (0.17)	0.571 (0.167)

Note. Nonzero values indicate the model with the higher log-likelihood/AIC.
Abbreviations: dLogLik, difference of log-likelihoods; dAIC, difference of AIC.

Table C.2
Predicting the **presence** of **lexical** alignment based on mock conversations

Parameter	log-odds	SE	95% CI	z	p
Utterance order	−0.04	0.039	[−0.12, 0.04]	−1.025	.305
Block [linear trend]	−0.025	0.023	[−0.07, 0.02]	−1.123	.261
Role [controller]	0.058	0.026	[0.01, 0.11]	2.257	.024
Level start time	0.014	0.013	[−0.01, 0.04]	1.006	.314
Level duration	0.032	0.015	[0.00, 0.06]	2.113	.035
Utterance length	0.956	0.018	[0.92, 0.99]	52.961	<.001
Concept [PoT]	0.011	0.028	[−0.04, 0.07]	0.382	.702

Table C.3
Predicting the **presence** of **syntactic** alignment based on mock conversations

Parameter	log-odds	SE	95% CI	*z*	*p*
Utterance order	−0.051	0.042	[−0.13, 0.03]	−1.223	.221
Block [linear trend]	−0.021	0.026	[−0.07, 0.03]	−0.816	.414
Role [controller]	−0.079	0.033	[−0.14, −0.02]	−2.421	.015
Level start time	0.046	0.014	[0.02, 0.07]	3.224	.001
Level duration	0.13	0.017	[0.10, 0.16]	7.695	<.001
Utterance length	0.103	0.014	[0.08, 0.13]	7.601	<.001
Concept [PoT]	−0.012	0.032	[−0.08, 0.05]	−0.39	.697

Table C.4
Predicting the **magnitude** of **lexical** alignment based on mock conversations

Parameter	Coefficient	SE	95% CI	*t*	*p*
Utterance order	0.022	0.014	[−0.01, 0.05]	1.591	.112
Block [linear trend]	0.008	0.009	[−0.01, 0.02]	0.852	.394
Role [controller]	−0.042	0.011	[−0.06, −0.02]	−3.696	<.001
Level start time	0.007	0.005	[0.00, 0.02]	1.338	.181
Level duration	0.005	0.006	[−0.01, 0.02]	0.815	.415
Utterance length	0.004	0.004	[0.00, 0.01]	1.137	.255
Concept [PoT]	0.007	0.011	[−0.01, 0.03]	0.665	.506

Table C.5
Predicting the **magnitude** of **syntactic** alignment based on mock conversations

Parameter	Coefficient	SE	95% CI	*t*	*p*
Utterance order	−0.012	0.015	[−0.04, 0.02]	−0.792	.428
Block [linear trend]	−0.011	0.01	[−0.03, 0.01]	−1.154	.248
Role [controller]	−0.083	0.013	[−0.11, −0.06]	−6.341	<.001
Level start time	0.01	0.005	[0.00, 0.02]	1.84	.066
Level duration	0.017	0.006	[0.00, 0.03]	2.675	.007
Utterance length	0.017	0.005	[0.01, 0.03]	3.769	<.001
Concept [PoT]	−0.003	0.012	[−0.03, 0.02]	−0.236	.813

Table C.6
Semantic

Parameter	Coefficient	SE	95% CI	*t*	*p*
Utterance order	5.63E−04	0.003	[−0.005, 0.006]	0.19	.849
Block [linear trend]	7.52E−04	0.002	[−0.003, 0.004]	0.419	.675
Role [controller]	0.012	0.002	[0.008, 0.017]	5.583	<.001
Level start time	−4.62E−04	0.001	[−0.002, 0.002]	−0.453	.65
Level duration	7.28E−04	0.001	[−0.001, 0.003]	0.66	.509
Utterance length	0.057	8.81E−04	[0.055, 0.058]	64.4	<.001
Concept [PoT]	0.003	0.002	[−0.001, 0.007]	1.288	.198