Temporal effects of alignment in text-based, task-oriented discourse

Foltz, A.; Gaspers, J.; Meyer, C.; Thiele, K.; Cimiano, P.; Stenneken, P.

Discourse Processes

DOI:
10.1080/0163853X.2014.977696

Published: 22/12/2014

Peer reviewed version

Temporal effects of alignment in text-based, task-oriented discourse

Anouschka Foltz

Bangor University

Judith Gaspers

Bielefeld University

Carolin Meyer

Fresenius University of Applied Sciences

Kristina Thiele

University of Cologne

Philipp Cimiano

Bielefeld University

Prisca Stenneken

University of Cologne

Author Note

Corresponding author: Anouschka Foltz, School of Linguistics and English Language, Bangor University, Bangor, Gwynedd, LL57 2DG, UK. Phone: +44 (0)1248 38 2267. Fax: NA. Email: a.foltz@bangor.ac.uk.

This work was supported by the CRC 673 "Alignment in Communication" and the Cognitive Interaction Technology - Center of Excellence (CITEC) at Bielefeld University. Parts of this research have been presented at the 11th International Symposium of Psycholinguistics in Tenerife, Canary Islands, Spain. We would like to thank Maximilian Panzner and Janik Jaskolski for designing the chat game and running the experiment. We would also like to thank Said Abir Anbari for his help in annotating and preparing the data for this study.

Temporal effects of alignment in text-based, task-oriented discourse

Abstract

Communicative alignment refers to adaptation to one's communication partner. Temporal aspects of such alignment have been little explored. This paper examines temporal aspects of lexical and syntactic alignment (i.e. tendencies to use the interlocutor's lexical items and syntactic structures) in task-oriented discourse. In particular, we investigate whether lexical and syntactic alignment increases throughout the discourse, and whether alignment contributes to speedy task completion. We present data from a text-based chat game, where participants instructed each other on where to place objects in a grid. Our methodological approach allows calculating a robust baseline and revealed reliable lexical and syntactic alignment. However, only lexical alignment, but not syntactic alignment, was sensitive to temporal aspects in that only lexical alignment increased throughout the discourse and positively affected task completion time. We discuss how these results relate to the communicative task and mention implications for models of alignment.

Introduction

Alignment in conversation refers to a speaker's tendency to adopt various aspects of their interlocutors' linguistic and non-linguistic behavior. Thus, speakers may adopt their interlocutors' syntactic structures (Bock, 1986), lexical items (Garrod & Anderson, 1987), phonological properties (Nilsenová & van Amelsvoort, 2010), formality level (Westbrook, 2007) etc. Alignment is pervasive and has been argued to decrease processing load during conversation and contribute to successful communication (cf. Pickering & Garrod, 2004). Much research in the linguistic domain has focused on syntactic alignment (see Pickering & Branigan, 1999; Pickering & Ferreira, 2008, for reviews). Such linguistic alignment can be due to linguistic conventions (e.g. printed sheets bound together into a volume are typically called a *book*), the task at hand (e.g. saying *next* when giving a set of instructions) etc., or constitute what we call *communicative* alignment (e.g. calling a long piece of seating furniture for two or more persons a *seat* because that's the term the interlocutor used). Alignment studies are typically concerned with the latter type of alignment, which constitutes communicative adaptation to the structures and words used by the interlocutor.

In this paper, we address temporal aspects of lexical and syntactic alignment in task-oriented discourse, that is, in a situation where two people communicate in order to complete a certain task. We focus on a simple task involving simple geometric shapes with conventional labels, e.g. a green circle. This allows us to study alignment in a context where no syntactic alternatives are deliberately introduced and where speakers can use conventionally-given object labels to successfully complete the task. In this context, it is possible that linguistic conventions and task constraints determine many of the interlocutors' syntactic and lexical choices. For example, when discussing geometric shapes, interlocutors may both call a round shape a *circle* because that's how both conventionally refer to a round shape, not because *circle* was the term among several lexical alternatives that their interlocutor had previously used for the round shape. For a situation with relatively little room for such communicative alignment, which may occur alongside alignment due to linguistic conventions and task constraints, we ask three research questions: Does alignment occur at all? Does alignment increase over the course of the discourse? And, does alignment contribute to faster successful task completion?

Does alignment increase with the evolving discourse?

This paper focuses on lexical and syntactic alignment. There is ample evidence that speakers align to their conversation partner on both the lexical and the syntactic level (e.g. Brennan & Clark, 1996; Branigan, Pickering, & Cleland, 2000). Most accounts of language coordination assume or imply that alignment should increase throughout the discourse. In terms of conceptual pacts (Brennan & Clark, 1996), for example, speakers should negotiate a conceptualization of the objects they are referring to and lexical alignment should increase as this conceptualization is being negotiated. In an interactive-alignment approach (Pickering & Garrod, 2004), alignment relies on a priming mechanism that allows percolation between different linguistic levels. As speakers start aligning to their conversation partner, this percolation process should increase both lexical and syntactic alignment as the conversation evolves. An implicit learning approach (Chang et al., 2006) models alignment in terms of weight changes in a connectionist network. Throughout a conversation the relevant weights should change so as to facilitate alignment with the conversation partner, also potentially leading to an increase in alignment throughout the discourse. There is also evidence that alignment increases (or at least systematically varies) over the course of discourse. For example, Garrod and Anderson (1987) found that alignment of spatial descriptions increased over the course of the experiment. Friedberg, Litman, & Paletz (2012) found a more nuanced effect: Lexical alignment of groups with high task success increased over time, whereas lexical alignment of groups with low task success decreased over time. At the syntactic level, Kaschak, Loney, & Borreggine's (2006) and Jaeger and Snider's (2013) results show that the more frequently a certain prime structure was encountered or produced, the more likely a speaker was to produce this structure. Thus, they found evidence for cumulative priming, suggesting that alignment should increase over time (but see Carbary & Tanenhaus, 2011, who found no increase of syntactic alignment over the course of their experiment).

An increase in alignment relies on languages' flexibility. The more conventions in language determine how to refer to an object or how to describe an action, the less room there is for additional communicative alignment (i.e. in addition to alignment due to conventions or task constraints) and for such communicative alignment to increase over the course of a conversation. As an example, let us consider lexical alignment and the question whether lexical alignment increases with the evolving discourse. (Note that the reasoning in the upcoming example also applies to syntactic alignment and the question whether syntactic alignment increases throughout a conversation.) When no conventional object labels are available and labels need to be explicitly negotiated, speakers and listeners can quite quickly agree upon labels for the objects. Consider the following example from Brennan (2010, p. 204), where two participants match duplicate tangram figures. The example shows how participants refer to one of many tangram shapes in the experiment the first three times that they encounter this particular tangram shape:

(1) Example from Brennan (2010, p. 204).

Encounter 1:

A: ah boy this one ah boy alright it looks kinda like, on the right top there's a square that looks diagonal

B: uh huh

A: and you have sort of another like rectangle shape, the-like a triangle, angled, and on the bottom it's ah I don't know what that is, glass shaped

B: alright I think I got it

A: it's almost like a person kind of in a weird way

B: yeah like like a monk praying or something

A: right yeah good great

B: alright I got it

Encounter 2:

B: 9 is that monk praying

A: yup

Encounter 3:

A: number 4 is the monk

B: ok

Here, the participants relatively quickly settle on the term *monk* to describe the tangram figure in question. Such negotiation and eventual agreement on a referring expression would lead to an increase in alignment as the discourse evolves. But what happens when object labels are conventionally given? In this case, a speaker can propose a conceptualization that the listener is likely to accept and act upon correctly in a given task, for example, introduce a colored round shape as *the green circle*. The interlocutor may then immediately accept the conceptualization and no further negotiation is necessary. The question we are concerned with in this paper is whether lexical alignment increases over the course of task-oriented discourse if the objects involved in the task have conventional names and thus may not need negotiation beyond the proposal and acceptance of a conventional name. If throughout the discourse both conversation partners stick to the referring expression introduced when an object is first mentioned, we would observe no increase in alignment over time since both partners are fully aligned from the beginning. However, even in the case of objects with conventional names, languages tend to be flexible enough to allow for other conceptualizations, e.g. *the light green ball*. It is thus also possible that even such seemingly obvious and conventionally-given referring expressions sometimes receive further negotiation. In such a situation alignment should increase throughout the discourse as referring expressions are being negotiated.

Does alignment contribute to faster task performance?

Most accounts of language coordination also assume or imply that alignment should contribute to task success. Since we are interested in temporal aspects of alignment, we measure task success in terms of how quickly a task can be performed. Within a conceptual pacts framework, referring expressions are negotiated and often shortened during negotiation. That is, aligned expressions are often shorter and thus take less time to produce and comprehend than non-aligned expressions. In addition, negotiation takes time, such that interlocutors should perform a task faster once they have agreed upon referring expressions. It would thus be reasonable to assume that alignment contributes to faster task performance. Within an interactive-alignment framework, Pickering and Garrod (2004) assume that alignment is the basis for successful communication. In task-oriented discourse, successful communication should translate into successful task completion. In other words, people who are more aligned should communicate more successfully and should complete the task more accurately and in less time. An implicit learning approach, on the other hand, currently makes no predictions about how alignment relates to task performance.

There is evidence from previous studies that alignment contributes to task success. For example, Reitter and Moore (2007) used the HCRC Map Task corpus (Anderson, Bader, Bard, Boyle, Doherty, Garrod, Isard, Kowtko, McAllister, Miller, Sotillo, Thompson, & Weinert, 1991) and showed that syntactic and lexical repetition reliably predicted task success, measured as how well participants' paths fit the intended path. Nenkova, Gravano and Hirschberg (2008) found that alignment of high-frequency words in the Columbia Games Corpus (Benus, Gravano, & Hirschberg, 2007) positively correlated with task success, measured as the achieved game score. Metzing and Brennan (2003) showed that participants took more time to find an object when their interlocutor suddenly switched referring expressions, for example, by first referring to an object as *the shiny cylinder* and later as *the silver pipe*. Here, the interlocutor intentionally misaligned and this slowed down participants' responses. In contrast to these results, Carbary and Tanenhaus (2011) found no correlation between syntactic alignment and task completion time.

Let us now consider how alignment may contribute to faster task performance in a simple task with simple geometric shapes, where referring expressions may only need minimal negotiation. It seems that in such a situation there is little room for shortening of referring expressions or simplification of syntactic structures. We propose that in such a situation communicative alignment may contribute to faster task performance not by decreasing negotiation time or through shortened expressions, but by decreasing the time it takes to comprehend the interlocutor's utterances. For example, there should be no need for explicit negotiation, regardless of whether an interlocutor refers to an object as *the green circle* or as *the green ball* and regardless of whether the interlocutor says *Place the green ball...* or *The green ball goes....* Both expressions are also equally succinct, and both syntactic alternatives have similar complexity. However, if someone expects their communication partner to use the word *circle* to refer to the object in question, then hearing the object referred to as *ball* may slow down his or her response, regardless of how succinct and appropriate the expression may be. Similarly, if one expects a certain syntactic routine, hearing an unexpected syntactic structure may slow down one's response regardless of how appropriate that syntactic structure is.

Studying syntactic and lexical alignment in highly-structured, text-based, task-oriented discourse

To study how alignment develops over the course of the discourse and how it may affect task success when participants discuss simple geometric shapes, we used a highly structured task that is very similar to a task developed by Gergle and colleagues (Gergle, Kraut, & Fussell, 2006; Kraut, Gergle, & Fussell, 2002). In particular, we present data from conversations derived from a chat matching game, where two naïve participants instructed each other on where to place simple geometric shapes in a 6x6 grid. Such matching tasks are used rather frequently in communication research (e.g. Carletta, Hill, Nicol, Taylor, de Ruiter, & Bard, 2006).

For the purposes of our study, a highly-structured matching task had advantages over paradigms that are more frequently used to study communicative alignment: Many previous studies on syntactic and lexical alignment have used carefully controlled psycholinguistic picture-description experiments (e.g. Branigan et al., 2000; Branigan, Pickering, Pearson, McLean, & Brown, 2011). Such experiments typically involve a confederate of the experimenter, who follows a script, and carefully selected pictures that allow for a syntactic alternation or alternative referential expressions. The interaction thus presents quite a departure from natural dialog. Alignment is then measured at certain experimentally-induced points in the discourse by counting how often speakers adopt the previously encountered syntactic structures or lexical items. Such an approach would not work for our purposes since we are interested in communicative alignment in situations where there are no obvious experimentally introduced alternatives. In addition, we would like to track alignment over the course of a conversation, which is difficult when oppportunities to align are experimentally determined.

Corpus studies have also been employed to study syntactic and lexical alignment (e.g. Gries, 2005; Howes, Healey & Purver, 2010). Such studies track recurrence of syntactic structures and lexical items in available corpora. Some corpus studies tracked only the syntactic alternations also used in picture-description experiments (e.g. Dubey, Keller, & Sturt, 2005; Gries, 2005; Howes et al., 2010; Szmrecsanyi, 2005), others tracked all syntactic structures and lexical items (Reitter, Moore, & Keller, 2006). These corpus studies investigate alignment in rather unconstrained, natural conversation. Our task resembles such corpus studies in that participants could produce language rather freely. However, unlike many of the previous corpus studies, we used a highly constrained task. In particular, we tried to minimize choices (and thus opportunities to communicatively align) by using a simple task and simple geometric shapes. For our purposes, this task has the following advantages: It creates enough experimental control to estimate alignment due to linguistic conventions and task constraints as well as communicative alignment due to coordination with one's interlocutor, but it is also unrestricted enough to allow participants to produce language rather freely.

The current study

In the following sections we describe the creation of the corpus as well as the analysis method in more detail. We then present results from both lexical and syntactic alignment. In particular, we investigate whether we find reliable lexical and syntactic alignment at all in a corpus created from a highly structured task using basic colored geometric shapes. Then, we present results on temporal aspects of alignment. In particular, we ask whether alignment increases over the course of the discourse and whether alignment contributes to speedy task completion. The results are then discussed with respect to the experimental task, current models of alignment, and flexibility in language.

Methods

Participants

A total of 14 pairs of adult native-German speakers participated in the study (8 male, 20 female with a mean age of 26 (*sd* = 10)). They were paid and/or received course credit for their participation. Data from two further participant pairs were excluded because they did not finish the task within an hour.

Materials

We designed and implemented a two-player, online, browser-game. In this game, a director described the positions of colored, geometric shapes in a 6x6 grid to a matcher. The matcher’s task was to use the director’s descriptions to replicate the arrangement of shapes on the director’s screen. Figure 1 shows sample grids seen by the director (left) and by the matcher (right). The director saw two 6x6 grids. The grid on the right shows the target arrangement of five colored, geometric shapes. The grid on the left shows the arrangement of shapes on the matcher’s grid. A box below the two grids reminded the director when it was his/her turn to play (*Sie sind am Zug: geben Sie dem anderen Spieler Anweisungen* (*It is your turn: give instructions to the other player*.)) and provided white space to type a chat message to the matcher. Furthermore, a button (*Runde Beenden* (*End Round*)) was displayed which enabled the director to end a round, i.e. to indicate that all shapes were positioned correctly. Thus, the director could see the target arrangement of shapes as well as the matcher’s grid and any changes made to it. The matcher saw an initially empty 6x6 grid and the five target shapes positioned outside of the grid. A box below the two grids reminded the matcher when it was his/her turn to play (*Sie sind am Zug: bewegen Sie die Figuren* (*It is your turn: move the shapes.*)) and displayed the chat messages typed in by the director. Furthermore, a button (*Zug Beenden* (*End Move*)) was displayed which enabled the matcher to indicate that he/she had positioned a shape. Thus, the matcher could see his/her own screen, but not the director's screen.

(insert Figure 1 about here)

The game consisted of eight rounds. Each round consisted of moves. Directors' moves consisted of typing and sending chat messages. Matchers' moves involved following the instructions in the chat message to position an object in the grid.

For each round, five color-shape combinations were randomly chosen and randomly placed in the director's right-hand 6x6 grid, yielding the round's arrangement of shapes. The same five color-shape combinations were placed to the left of the matcher's 6x6 grid. The geometric shapes were randomly chosen out of 50 possible color-shape combinations. There were five possible shapes (triangles, circles, ellipses, rectangles, and squares) and ten possible colors (black, blue, cyan, green, magenta, orange, pink, red, yellow, and gray). Arrangements were created like this for all eight rounds and then saved to create a single version of the game. Thus, randomization was done only once, so that all pairs of participants played identical games. The colored geometric shapes used in each round are given in Table 1.

(insert Table 1 about here)

Procedure

Two participants were each seated in front of a computer screen, such that they could not see each other's screens. One participant was assigned the role of director, the other the role of matcher. After each round, participants switched roles. Participants were instructed not to talk to each other during the game and to avoid absolute positions (e.g. purple square to (1,4)) and linear dimension units (e.g. 2 centimeters) in the chat descriptions. We put these restrictions on the description schemes to adequately capture alignment with the computational methods that we used. For example, if participants were to use absolute dimensions, only references to the same cell (e.g. (1,4) and (1,4)) would be considered aligned, but not references to two different cells (e.g. (1,4) and (2,3)), even though in both cases participants would be using the same description scheme. The first eight rounds that each pair of participants played make up the data for this study.

During each round, the director had to describe the position of the same object until the matcher had placed it in the correct position in the grid. Since the director could see the matcher's current arrangement of shapes, s/he could correct erroneous positions. Objects were moved solely based on the director's written descriptions, as the matcher could not see the director's screen, type messages to ask questions, or talk to the director. Thus, the matcher provided concurrent feedback to the director only by placing objects in the grid. That is, the matcher's feedback was exclusively nonverbal. Typing chat messages and positioning objects could not occur at the same time. Thus, at any given time only one of the players could execute a game action. The director ended the round once all objects were correctly positioned by the matcher.

Data Preprocessing

For each game, all chat messages and corresponding object moves were recorded, time-stamped, and saved in a game-log file in XML-format. Time was reset to zero for each round. The chat messages contained in the game-log files were annotated for further processing. In a first step, the texts were manually corrected and formatted. In particular, spelling and punctuation errors were corrected, capitalization was made uniform (e.g. an upper-case letter at the start of each sentence), and sentence fragments were divided by commas. This manual correction was necessary so that later steps could be performed automatically. Since we were not interested in chat spelling conventions, abbreviations were also written out during this step.

In a second step, we derived measures pertinent to our analyses of lexical and syntactic alignment from the text: The hand-corrected text was run through the Stanford parser (Klein & Manning, 2002), in particular, a German parser (Rafferty & Manning, 2008) based on the NEGRA corpus. The parser generated a parse tree including part of speech information for each sentence or sentence fragment. The output of this automatic process was manually corrected. We then computed the complexity of generated parse trees, i.e. tree breadth and tree depth. Tree depth was computed as the maximum depth of the generated parse tree, i.e. the length of the longest path from the root to a leaf node. Tree breadth was computed as the average number of child nodes in the parse tree, i.e. the number of child nodes divided by the number of all nodes. For instance, the parse tree of the noun phrase *the red circle* is shown in Figure 2. Its depth and breadth would be 2 and 0.8, respectively.

(insert Figure 2 about here)

Furthermore, lemmas were manually annotated for all words which were tagged as nouns, adjectives or verbs. Lemmatization was important to detect all relevant aligned lexical items and to create a lexical alignment measure that had no syntactic component. Without lemmatization many aligned lexical items may not be detected or reflect syntactic choices. For example, without lemmatization the words *rote* (*red*) and *roten* (*red*) in the sentences *[Der rote Kreis]$_{NOM}$ kommt...* (*The red circle goes...*) and *Platziere [den roten Kreis]$_{ACC}$...* (*Place the red circle...*) would not be considered to be aligned because of case marking differences (nominative vs. accusative) associated with the different syntactic structures. To avoid that our lexical alignment measure actually captures syntactic alignment, lemmatization was necessary, even though it may obscure potentially interesting distinctions between word forms. Finally, we hand-annotated sentence types, using the types given in Table 2.

(insert Table 2 about here)

Measures

We extracted two main measures from the preprocessed data: alignment and performance. These measures were compared across game rounds since only one player could produce messages in any given round. Thus, we measured alignment of all chat messages sent in one round to all messages sent in the following round. We calculated the phenomena of lexical and syntactic alignment using the following concrete measures: lemma alignment (lexical) and sentence type alignment, parse tree breadth alignment, and parse tree depth alignment (all syntactic). The sentence type measure captures the grammatical constructions participants used, whereas the parse tree breadth and parse tree depth measures capture the complexity of the sentences participants used. Performance was measured as the average time it took the matcher to perform a move, i.e. to place an object in response to the director's chat message. Since the measures lemma, sentence type, parse tree breadth and parse tree depth have different characteristics (e.g. sentence type is nominal while parse tree breadth is numeric), we employed two different notions of alignment, both based on distances. In the following, the alignment and performance measures are described in more detail.

Nominal values: Lemmas and sentence types. Since lemmas and sentence types are nominal measures, we estimated alignment for these measures using the *cosine similarity* (Manning, Raghavan, & Schütze, 2008, p 111), a measure often used in text mining/information retrieval to compute similarity across texts. Similarity is computed based on word counts (ignoring word order), in particular, on word frequency vectors. Table 3 uses some examples to illustrate the intuition to use the cosine similarity to measure alignment.

(insert Table 3 about here)

The table shows example sentence type counts, which reflect how frequently given sentence types occurred in a given round. Now imagine that in round 1 of the game player *A* produced sentences with the distribution of sentence types shown in the table. Alignment of sentence types between rounds 1 and 2 is then calculated based on how similar this distribution is to player *B*'s distribution of sentence types in round 2. To illustrate, let's estimate alignment between *A*'s round 1 and *B*'s rounds 2 for four possible scenarios, which are labeled 1 through 4 in the table. The distribution of sentence types is identical for player *A* and player *B*'s scenario 1. With such maximum similarity, the two players should intuitively be measured as maximally aligned, and computing the cosine similarity between both frequency distributions indeed yields its maximum value of 1. In contrast to scenario 1, the sentence type frequencies for players *A* and *B* in scenario 2 are not identical since *B* typed twice as many sentences as *A*. However, the proportion of the sentence types used is still exactly the same. Intuitively, the two players should therefore be measured as maximally aligned as well, and the cosine similarity again yields its maximum value of 1 (due to count normalization). Now consider the sentence types used in scenario 3. While *B* used exactly the same sentence types as *A*, their relative frequencies differ, and thus one would intuitively measure player *B* as less aligned to player *A* in scenario 3 than in scenarios 1 and 2. Since the cosine similarity takes the frequency distribution of the used sentence types into account, it again captures this intuition. Finally, comparing *A* and *B* in scenario 4, both players used completely different sentence types, and thus intuitively show no alignment. Again, the minimum possible value of 0, as estimated by the cosine similarity, reflects this intuition.

Taken together, using the cosine measure yields several major advantages: It measures lexical and syntactic alignment based on all relevant words and syntactic structures in the corpus (see also Reitter et al., 2006; Howes et al., 2010), and not merely for select target words or syntactic structures (e.g. Gries, 2005; Howes et al., 2010). In addition to reflecting if certain choices are adopted, it also incorporates the frequency distribution of adoption. It also normalizes counts and thus abstracts away from the number of occurrences. In particular, it abstracts away from the length of chat messages and number of words and syntactic constructions used. In the following, we will introduce the cosine similarity more formally, and describe how it is applied to measure alignment.

In general, the cosine similarity is defined as the cosine angle between two given vectors $\vec{x}$ and $\vec{y}$ by

$$cos(\vec{x},\vec{y})=\frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|}=\frac{\sum_{i=1}^{n}\vec{x}_i*\vec{y}_i}{\sqrt{\sum_{i=1}^{n}(\vec{x}_i)^2}*\sqrt{\sum_{i=1}^{n}(\vec{y}_i)^2}}, \quad (1)$$

where $\cdot$ denotes the dot product, and $\|\vec{x}\|$ and $\|\vec{y}\|$ denote the magnitudes of $\vec{x}$ and $\vec{y}$, respectively. Resulting values range from -1 to 1, where higher values denote higher similarity between two vectors. A value of 1 expresses exactly the same vector and -1 the maximal difference between two vectors.

Similarly to what is done in text mining/information retrieval, we estimated lexical alignment as similarity between texts, in our case chat message texts of two players. Since we were interested in alignment of open-class words, we extracted word frequency vectors consisting only of lemmas which were annotated either as noun, adjective, or verb. In particular, for each round r_x a frequency vector $\vec{v}_{L(r_x)}$ was extracted by counting lemma frequencies for all words tagged as noun, verb, or adjective. Given the frequency vectors for two rounds r_x and r_y, we measured lexical alignment using the cosine similarity as:

$$ALIGN^{lexical}(r_x,r_y)=cos(\vec{v}_{L(r_x)},\vec{v}_{L(r_y)}). \quad (2)$$

As word counts cannot be negative, resulting similarity values range between 0 and 1, where higher values denote higher similarity between the two feature vectors. Accordingly, we regarded higher values as indicating more lexical alignment between the instructor in round r_x and the instructor in the following round r_y.

We measured syntactic alignment based on sentence types analogously to lexical alignment. In particular, given the annotated data of a round r_x, a frequency vector capturing sentence type frequencies $\vec{v}_{S(r_x)}$ was created by counting the number of occurrences of each individual sentence type. Alignment was again measured using the cosine similarity. In particular, given the annotated sentences for two rounds r_x and r_y, we measured sentence type alignment as

$$ALIGN_{type}^{sentence}(r_x,r_y)=cos(\vec{v}_{S(r_x)},\vec{v}_{S(r_y)}). \quad (3)$$

Again, values range between 0 and 1, and we measured higher values as indicating more alignment between the director in round r_x and the director in round r_y.

Numeric Values: Parse tree breadth and parse tree depth. Unlike lemmas and sentence types, parse tree breadth and parse tree depth are numeric, allowing us to directly compute distances between any two corresponding values. We therefore estimated alignment based on absolute distances. In particular, given the chat data for any round r_x, we first computed the arithmetic mean of parse tree breadth and parse tree depth values as *avg_breadth*(r_x), and *avg_depth*(r_x), respectively. We then estimated alignment for two successive rounds r_x and r_y based on these mean values. In particular, alignment was calculated as the absolute distance between the mean values, using the equations in (4) and (5), where lower values denote lower distances between two values. Thus, lower values correspond to more alignment with 0 being maximum alignment (denoting identical values).

$$ALIGN^{breadth}(r_x,r_y)=|avg\_breadth(r_x)-avg\_breadth(r_y)|, \tag{4}$$

$$ALIGN^{depth}(r_x,r_y)=|avg\_depth(r_x)-avg\_depth(r_y)|, \tag{5}$$

Note that the lemma and sentence type values described in the previous section range from 0 (no alignment) to 1 (complete alignment). In contrast, parse tree breadth and parse tree depth alignment values range from some positive number (minimal alignment) to 0 (complete alignment). This positive number represents the largest difference between the average parse tree breadth or parse tree depth of one round and of the following round. In our data set, this positive number was 5.1 for breadth alignment and 7 for depth alignment. Thus, individual values for breadth and depth alignment ranged from 5.1 and 7 (minimal alignment), respectively, to 0 (complete alignment). For easy comparison with lexical and sentence type alignment values, we transformed the individual breadth and depth alignment values, using the equations in (6) and (7), so that they also range from 0 (no alignment) to 1 (complete alignment).

$$ALIGN\_TRANS^{breadth}(r_x,r_y)=\frac{5.1-ALIGN^{breadth}(r_x,r_y)}{5.1} \tag{6}$$

$$ALIGN\_TRANS^{depth}(r_x,r_y)=\frac{7-ALIGN^{depth}(r_x,r_y)}{7} \tag{7}$$

Performance. As a performance measure, we used move completion time. For each round r_x we computed the average move completion time *avg_time*(r_x) as the average time that passed between sending a chat message and executing the last move in response to that chat message. The time used for typing a message was excluded to avoid that individuals' typing speed and message length affected our performance measure.

Baseline Values

In order to interpret the alignment values described above, we created alignment baseline values. These baseline values allowed us to disentangle communicative alignment from what participants would have said anyway, for example, because of lexical and other linguistic conventions. For example, when writing an instruction about where to move a green circle, participants are likely to use the words *green* and *circle* and they may do this not because they are aligning with their interlocutor, but because this is how they would conventionally refer to the object, independently of their conversation partner. We created baseline values by calculating alignment values for participants who did not actually play together. Thus, we estimated how similar participants' descriptions were when not playing together. Such a baseline provides an estimate of alignment due to linguistic conventions and the task. Creating a baseline like this was possible because all pairs of players had to place the same geometric figures in each round of the game. Thus, the opportunities to use certain structures and lexical items are the same in the baseline and the actual interaction If two people who played together had significantly higher alignment values than two people who did not play together, we could assume that these higher alignment values reflect communicative alignment, not merely linguistic conventions or constraints imposed by the game task.

This baseline does have a limitation that should be mentioned: Interleaving rounds from two speakers who did not interact with each other may artificially enlarge the alignment effect. This is the case because both speakers did interact with another partner and may have aligned to that partner. However, this problem is less pronounced than in most experimentally controlled settings, where alignment to a certain primed word or structure has most commonly been compared to a baseline that measures the use of the target word or structure following an alternative prime word or structure. Thus, all baseline trials in such settings involve the alternative lexical item or syntactic structure. In contrast, in our study participants are randomly matched to create a baseline and are thus not necessarily matched with someone who used all the possible alternatives. In fact, this situation is rather unlikely. Nevertheless our baseline may be somewhat biased towards alternative lexical items and syntactic structures.

To create the baseline, we first paired the 14 games randomly. Then, for each game pair, two baseline games were created by interleaving rounds 1, 3 ,5 and 7 of game one (i.e. the rounds played by player *A* of game one) with rounds 2, 4, 6 and 8 of game two (i.e. the rounds played by player *B* of game two), thus simulating games between players who did not actually play together. Likewise, rounds 2, 4, 6 and 8 of game one (i.e. the rounds played by player *B* of game one) were interleaved with rounds 1, 3, 5 and 7 of game two (i.e. the rounds played by player *A* of game two). Taken together, our baseline consisted of 14 (pseudo)games. We then computed baseline aligment values as described above.

Results

This section first provides a short description of the data generated by the chat task. We then present results for the following research questions: Did players align at all? Does alignment increase with the evolving discourse? Does higher alignment lead to faster task performance?

Data Description

Before we delve into the results, a brief description of the kind of data our task generated is in order. The following examples illustrate the raw, uncorrected data: They show directors' instructions from rounds one and two for one pair of participants. Loose translations are provided in square brackets. The examples illustrate that the task yielded rather natural instructions.

(2a) Round one for participant pair 1: Player A is the director.

1. Ganz links in der Mitte (unten) ist ein pinkes Quadrat [to the very left in the middle (bottom) is a pink square]
2. Es ist in der ganz äußeren, linken Reihe, nicht unten. [it is in the very outside, left row, not in the bottom]
3. Genau. In der selben Zeile ist links in der äußersten Reihe ein rotes Dreieck [exactly. In the same row is a red triangle on the left in the outermost row]
4. Die Zeile ist richtig, du musst das Dreieck jetzt nur noch nach ganz rechts schieben [The row is correct, now you only need to slide the triangle to the very right]
5. Genau. Unten links vom pinken Quadrat ist ein türkiser Kreis (in dem Feld, welches das Feld des Quadrats mit der Ecke berührt)) [Exactly. To the bottom left of the pink square is a turquoise circle (in the cell which touches the cell with the square in the corner))]
6. ja genau (sorry, habe rechts gemeint). Aber du hast es ja richtig gemacht :) In der gleichen Spalte ist fast ganz oben ein grauer Kreis [yes exactly (sorry, I meant on the right). But you did do it correctly :) In the same column almost at the top is a gray circle]
7. Ja, die letzte Figur ist ein hautfarbendes Rechteck, welches mit dem pinken Quadrat und dem türkisen Kreis eine Diagonale bildet (rechts unten von den beiden)) [Yes, the last figure is a skin-colored rectangle, which forms a diagonal with the pink square and the turquoise circle]

(2b) Round two for participant pair 1: Player B is the director.

1. das gelbe dreieck muss nach links oben. aber nicht nach ganz oben. [the yellow triangle needs to go in the top left. but not the very top.]
2. ja. :) das andere dreick muss ganz rechts. mittig. die untere mitte. [yes. :) the other triangle needs to go to the very right. centered. the bottom middle.]
3. der graue kreis muss rechts neben das pinke dreieck. [the gray circle needs to go to the right of the pink triangle.]
4. mein fehler. tausch bitte den grauen kreis mit dem pinken dreieck. sorry. [my mistake. please exchange the gray circle with the pink triangle. sorry.]
5. sorry [sorry]
6. der grüne kreis kommt über den grauen [the green circle goes above the gray one]
7. das rechteck ist in der spalte der kreise ganz unten [the rectangle is at the very bottom of the column of the circles]

Alignment was estimated from the kind of raw data presented in example (2) above. For example, lexical alignment between two rounds was estimated by first counting all noun, verb, and adjective lemmas based on the parse trees. For rounds (2a) and (2b) above, this results in the two frequency vectors $L(r_{1a})$= ("Mitte" = 1, "ist" = 7, "pink" = 3, "Quadrat" = 4, …) and $L(r_{1b})$= ("Mitte" = 1, "ist" = 0, "pink" = 2, "Quadrat" = 0, …), respectively. Lexical alignment would then be estimated by computing the cosine distance between those frequency vectors, using equation (2) introduced in the methods section.

The preprocessed data revealed that participants produced a total of 234 different lemma types tagged as either noun, adjective or verb. There were 79 lemma types tagged as nouns, 90 tagged as adjectives, and 65 tagged as verbs. Table 4 shows the ten most frequent lemma types tagged as noun, adjective or verb. Notice that all of the nouns most frequently used by participants referred either to the shape of a geometric figure or to a location. Most of the top ten adjectives referred to an object's color, and many of the most frequently used verbs denote a location or movement to a location. Thus, the choice of lexical items seems to be rather constrained by the experimental task of moving geometric figures in different colors to various locations.

(insert Table 4 about here)

To see whether linguistic conventions and the experimental task allowed for opportunities to communicatively align at the lexical level at all or whether lexical choices were largely conventionally determined, we extracted all the lexical items that participants used to refer to the colors and shapes of the objects. Table 5 shows the expressions used to refer to the different colors and how often these expressions denoted the relevant colors. Table 6 shows the expressions used to refer to the different shapes and how often the expressions denoted the relevant shapes. Note that objects were sometimes referred to with a pronoun or a nominalized color, such as *das Gelbe* (*the yellow one*). These nominalizations are not listed in the tables because they refer to the object as a whole, not just to the color or shape. It was therefore difficult to determine which table they should appear in. Table 5 shows that half the colors had one fully conventionally-determined name, such that they presented no opportunities for communicative alignment. For the remaining colors, participants used at least two different referring expressions. However, in some cases (e.g. blue and pink), one referring expression was clearly preferred. Table 6 reveals a different picture: All shapes were referred to with at least two different referring expressions. Thus, none of the expressions referring to the objects' shapes were fully conventionally determined. However, most shapes had one clearly preferred referring expression.

(insert Tables 5 and 6 about here)

Participants further produced a total of 781 sentences and sentence fragments, which could be grouped into eight different sentence type categories. Table 7 shows the eight sentence types and how often they occurred in the data. The table shows that the vast majority of sentences were either indirect indicative sentences (e.g. *The circle goes to the right of the triangle*) or fragments without a verb (e.g. *Circle to the right of triangle*). Together, these make up about 83% of all sentences. Here, we observe that the linguistic choices are constrained by the experimental task. In particular, the large number of fragments without verbs is typical for the abbreviated language used in chat conversations. Finally, participants produced fragments and sentences which ranged in parse tree breadth from 1.4 to 6.5 and in parse tree depth from 1 to 8. Altogether, the data description suggests that the experimental task serves to considerably constrain both the lexical items and the sentence types used in the chat conversations.

(insert Table 7 about here)

Did players align at all?

We first analyzed whether players aligned to their conversation partner at all in the kind of highly-structured task oriented discourse with basic geometric shapes examined here. We will consider both lexical alignment (lemma alignment) and syntactic alignment (sentence type, parse tree breadth, and parse tree depth alignment). The mean lexical alignment values for participants who actually played together (participants) and participants who did not play together (baseline) are shown in Figure 3. Individual lexical alignment values can range from 0 (no alignment) to 1 (complete alignment). The baseline values indicate how much alignment likely occurred based on the chosen objects and the nature of the communicative task, i.e. on the words and structures conventionally used to refer to such objects and actions in such a task. The participants' values capture alignment driven by the task and conventions as well as communicative alignment between conversation partners. Thus, if participants' alignment values are higher than baseline values, we take this increase in alignment to be due to communicative adaptation.

(insert Figure 3 about here)

Figure 3 shows that the lemma alignment value is higher for participants than for the baseline. Mixed-effects models with lemma alignment values as response variable, group (participants vs. baseline) as fixed effect (centered, with treatment coding), and subjects and round number (2 though 8) as random effects were calculated. Such mixed-effects models allow modeling more than one random effect within the same analysis. The statistical results are presented in Table 8 and show that participants' lexical alignment values are statistically significantly higher than the baseline lexical alignment values. We take this increase in alignment to stem from the interlocutors' communicative adaptation. We thus find reliable lemma alignment in a setting where participants produced language rather freely, but where linguistic conventions and the task substantially constrained lexical choices. Figure 3 illustrates this constraint on the choices of lexical items: alignment due to lingustic and task constraints is numerically larger (the 0.3 of the baseline) than the additional communicative alignment (the 0.21 that the participants value is larger than the baseline value). Nevertheless, participants showed communicative alignment in that they aligned beyond the lexical choices imposed by linguistic conventions and the experimental task.

(insert Table 8 about here)

Mean syntactic alignment values for participants and the baseline are also displayed in Figure 3. Alignment values for all three syntactic measures are numerically higher for participants than for the baseline. To determine whether these increases are statistically significant, we fit mixed-effects models analogous to the ones described above, but with the different syntactic measures (sentence type, breadth, depth) as response variables. Statistical models for breadth and depth were calculated using both raw and transformed values, always yielding the same results. Therefore, only results for the transformed values are reported and graphically displayed. Table 8 shows that participants aligned reliably more than the baseline for sentence type and breadth, but not for depth, which shows only a marginal effect. Notice also that sentence type and breadth, the two measures which did show reliable alignment compared to the baseline, show rather high overall alignment rates: Sentence type and breadth have mean participant alignment values of 0.89 and 0.93, respectively, with a possible range from 0 to 1. Altogether, we observed statistically significant alignment on both the lexical and the syntactic level.

(insert Figure 3 about here)

The above analyses showed statistically significantly higher alignment values for participants who played together compared to the baseline for lemmas, sentence types, and parse tree breadth. What we are interested in is this *difference* between baseline values and participant values. In all the following analyses, we therefore measure alignment as the difference between baseline alignment values and participant alignment values. This difference reflects the additional alignment that is at least partially due to communicative alignment. In particular, we calculated alignment for each round by subtracting each baseline value from its corresponding participant value. Thus, in all the following analyses, higher alignment values correspond to more alignment. Note also that subtracting each baseline value from its corresponding participant value can yield negative numbers when an individual baseline value is larger than its corresponding participant value. Thus, in all the following analyses, individual alignment values may be negative.

We conducted further analyses to explore whether any of our syntactic alignment measures could predict lemma alignment, i.e. whether being highly aligned at the lexical level also coincided with being highly aligned at the syntactic level. However, since participants did not reliably align at the level of parse tree breadth, we will omit this measure from all following analyses. Mixed-effects models with lemma alignment values as response variable, sentence type alignment and parse tree breadth alignment as fixed effects (centered, with sum coding), and subjects and round number (2 though 8) as random effects were calculated. Redundant fixed-effects were removed until the model was minimally optimized. The final model included only sentence type as fixed effect and showed a reliable effect of sentence type alignment on lemma alignment (*estimate* = 0.05544, $t = 3.209$, $p < .001$). Thus, lower lemma alignment coincided with lower sentence type alignment and higher lemma alignment coincided with higher sentence type alignment. The same does not hold for lemma alignment and parse tree breadth. Thus, we find some evidence that the degree of alignment at one linguistic level affects the degree of alignment at other linguistic levels. A further analysis suggests that this effect could possibly be due to some participants being generally high aligners (i.e. showing high lemma and sentence type alignment) and others being generally low aligners (i.e. showing low lemma and sentence type alignment): A comparison of the final model, which models only the variance of mean lemma alignment between subjects, with a model that allows for differences between subjects in sentence type alignment effects (cf. Kliegl, Wei, Dambacher, Yan, & Zhou, 2011) suggests that the more complex model provides a marginally better fit of the data (χ (df = 1) = 3.0538, p = 0.08). In other words, individual differences are marginally associated with the sentence type alignment effect.

Did alignment increase with the evolving discourse?

Next, we investigated whether alignment increased with the evolving discourse, i.e. over the course of the experiment. We considered both lexical and syntactic alignment. Figure 4 shows communicative alignment values for rounds two through eight for lexical, sentence type, and parse tree breadth alignment. The x-axis shows the individual rounds and the y-axis shows communicative alignment (calculated as the difference between participant alignment and baseline alignment). The figure shows an increase in lexical alignment over the second half of the experiment. In particular, rounds seven and eight show a substantial increase in alignment compared to round two. No such increase can be seen for sentence type and parse tree breadth alignment, which remain relatively stable across rounds. Thus, it seems that only lexical, but not syntactic alignment increases over the course of the discourse.

(insert Figure 4 about here)

To confirm these observations, we fit linear models with the different alignment measures (lemma, sentence type, and breadth) as response variable and round number as predictor variable. The results of these models are shown in Table 9 and confirm that lemma alignment, but not sentence type or parse tree breadth alignment, increases over the course of the discourse. Thus, participants become more aligned at the lexical level over the course of the chat discourse.

However, we find no evidence that participants' syntactic alignment increases over the course of the discourse.

(insert Table 9 about here)

Did alignment contribute to faster task performance?

Finally, we investigated whether alignment contributed to faster task performance, i.e. whether alignment correlated negatively with the time it took to complete a move in the game. Move completion times above 25 seconds were considered outliers and were excluded from the current analyses. Three move completion times (2.7% of the data) were excluded based on this criterion. (Recall that the time it took to type a message was not included in the move completion time measure.) To see if move completion time decreased as alignment increased, we fit mixed-effects models with move completion time as response variable, lexical alignment, sentence type alignment, and breadth alignment as fixed effects, and subjects as random effects. Redundant fixed factors were removed from the initial model until the model was minimally optimized. The final model contained only lexical alignment as fixed effect (estimate = -6.3608; $t = -3.194$; $p < .01$). Thus, participants who were lexically more aligned finished the task of positioning five geometric figures in a grid faster than participants who were lexically less aligned. However, we find no evidence that syntactic alignment contributes to move completion time.

We performed additional analyses to test whether lexical alignment served as a mediator (Baron & Kenny, 1986), that is, whether round number affected lexical alignment, which in turn affected completion time. Such a result would suggest that as rounds went by, participants showed more lexical alignment and that this in turn made them faster. For lexical alignment to serve as a mediator, the following must hold: (1) round number must reliably affect lexical alignment, (2) round number must reliably affect completion time, (3) lexical alignment must reliably affect completion time, and (4) the reliable effect must be smaller in (3) than in (2) (cf. Baron & Kenny, 1986). We therefore fit the following mixed-effects models with subjects as random effects: (1) a model with lexical alignment as response variable and round number as fixed effect (estimate = 0.07798; $t = 4.448$; $p < .001$), (2) a model with completion time as response variable and round number as fixed effect (estimate = -1.5815; $t = -5.294$; $p < .001$), and (3) a model with completion time as response variable and lexical alignment as fixed effect (estimate = -1.1437; $t = -3.194$; $p < .01$). The results from these models reveal that lexical alignment does indeed serve as a mediator such that as rounds went by, participants showed more lexical alignment, which in turn made them faster.

Discussion

The current study used highly structured task-oriented discourse to investigate temporal aspects of lexical and syntactic alignment. Overall, the results suggest that participants did align in such highly structured task-oriented discourse, both at the lexical and syntactic levels. In addition, lexical alignment, but not syntactic alignment, increased over the course of the discourse and modulated task completion time.

Alignment in highly structured task-oriented discourse

The task and analysis we chose allowed us to measure alignment in rather natural, yet highly structured, chat communication. Overall, participants in this study showed reliable lexical and syntactic alignment in task-oriented discourse. Importantly, our baseline allowed us to disentangle alignment due to linguistic conventions and task constraints from alignment that is (most likely) communicative or interactive. The data description shows that there are considerable effects of the task: The most frequently-used lexical items referred to object shapes, colors, and positions. In addition, the high occurrence of fragments without a verb reflect the abbreviations typical for chat communication. The baseline data further reveal considerable effects of the task and linguistic conventions: For both lexical and sentence type alignment, a higher proportion of alignment seems to be due to the task and linguistic conventions (baseline) than participants' communicative adaptation (participant alignment minus baseline). For both parse tree breadth and depth, alignment is already extremely high in the baseline, such that the task and linguistic conventions overwhelmingly contribute to parse tree breadth and depth. That is, people just seem to have a tendency to write chat sentences with a certain complexity, regardless of how complex their game partner's sentences are. Overall, the majority of the observed alignment between players that actually played together is due to the task and linguistic conventions, not due to interactive, communicative adaptation. This result highlights the need for a good and reliable baseline in studies of communicative alignment.

We did find communicative lexical and syntactic alignment in addition to alignment due to the task and linguistic conventions: Participants who played together had reliably higher alignment values than participants who did not play together for lexical, sentence type, and parse tree breadth alignment. Only parse tree depth revealed no alignment effect. There are several possible reasons why we found reliable communicative alignment despite rather high baseline alignment values. Participants needed to achieve similar situation models to successfully complete the task. Thus, the need to align may have been higher in our task than in free natural speech. This assumption is in line with studies that showed higher alignment in task-oriented discourse compared to free natural speech (Reitter et al., 2006; Reitter & Moore, 2014) and may explain why Healey, Purver, & Howes (2014) found divergence rather than alignment in free natural speech, whereas we found rather strong alignment in our task. In addition, the structure of the task encouraged cumulative priming, which has also been found to increase alignment (Kaschak et al., 2006). In particular, the director in any given round gave on average six instructions to the matcher and thus had ample opportunity to repeatedly produce prime syntactic structures and lexical items.

Notice further that all three reliable communicative alignment effects look rather different. Lexical alignment and, in particular, sentence type alignment showed much larger communicative alignment effects (participant alignment minus baseline) than parse tree breadth. In addition, sentence type alignment is numerically higher than lexical alignment, both in the baseline (0.57 vs. 0.3) and for participants (0.89 vs. 0.51). The reason for this may be the nature of lexemes and syntactic constructions. In particular, there are many more lexemes than there are syntactic constructions. There are thus many more lexical choices than sentence type choices, resulting in a lower probability of two people making the same choice, both in the baseline and when aligning communicatively.

Our additional analyses showed that sentence type alignment reliably predicted lemma alignment and that this effect is marginally related to individual differences. Thus, this effect may be due to some participants being generally high aligners and others being generally low aligners. Alternatively, this effect may be due to the percolation of alignment between linguistic levels, as proposed by Pickering and Garrod (2004).

Increase of alignment over the evolving discourse

Our results showed that lemma alignment increased over the course of the discourse. This increase is in line with previous studies (e.g. Garrod & Anderson, 1987). Furthermore, all three accounts of language coordination that we mentioned in the introduction are compatible with an increase of alignment over the course of the discourse. Recall that lemma alignment increased mostly during the last two rounds of the game. This is potentially in contrast to studies using figures without conventional names, such as tangrams. Krauss and Weinheimer (1966) showed that the number of words in the name given to a novel shape most drastically declined during the first four of fifteen trials if the listener could provide feedback. Similarly, Clark and Wilkes-Gibbs (1986) showed that the number of turns and the number of words per tangram figure most drastically declined between trials one and two of six trials. This early reduction in the number of words used to refer to tangram shapes may reflect the interlocutors' early agreement on referring expressions for figures without conventional names. Our results would then suggest that when the objects that interlocutors refer to are simple and have conventional names, alignment may increase more gradually than when object names need to be explicitly negotiated. There are several possible explanations for this observation: It seems that alignment as a result of negotiating a referring expression for a figure that has no conventional name is rather stragetic and represents a process that is well described in terms of a conceptual pacts approach. In contrast, the alignment that we observe in our task with simple geometric shapes may be less stragetic and more implicit and may represent a process that is better described in terms of interactive alignment or implicit learning. If a process is implicit rather than strategic, it may take longer for alignment to increase over the course of the discourse. In addition, the risk of misunderstanding and moving an incorrect object is lower in a task using objects with conventional names compared to without. When there is no already existing terminology for the figures to be described, interlocutors may need to quickly agree upon a term for each figure in order to avoid communicative breakdown. In our task, there are existing terms for the figures and misalignment is unlikely to lead to communicative breakdown. As a result, there may be less of a need to align referring expressions. The interlocutor is likely to move the correct geometric shape regardless of whether it is referred to as *the pink square* or *the rose-colored quadrangle*. It is possible that alignment increases more slowly over the course of the discourse when the risk of misunderstanding is low compared to when it is high. Note also that matchers gave no verbal, but only nonverbal feedback in our task. Krauss and Weinheimer (1966) showed that the number of words in the name given to a novel shape declined more drastically at the beginning of the interaction when listeners could give feedback compared to when not. It is thus possible that lemma alignment would have its most prominent increase earlier in the discourse if we had allowed verbal feedback too.

In contrast to our expectations, we found that syntactic alignment did not increase over the course of the discourse. To explain the lack of an effect in our data, recall that sentence type alignment values were much higher than lexical alignment values, both in the baseline and for participants. It is possible that syntactic alignment did not increase throughout the discourse due to this rather high overall sentence type alignment. In particular, alignment could have been so high overall that there was little room for it to increase over time. In this situation, participants were likely maximally aligned from the beginning of the discourse. Such maximal alignment from the beginning may have been a result of the structure of the game. Sentence type choices were task-specific in that they reflect the way participants gave each other instructions. Since the task never changed throughout the discourse (participants had to repeatedly instruct their interlocutor on where to place an object), it would not be surprising if participants agreed on one way of giving instructions after only a few moves. An informal inspection of the data suggests that this was indeed the case. In addition, if speaker *A* chose one sentence type in round one to give instructions, then speaker *B* was exposed to several examples of this sentence type before giving his or her first instruction. This cumulative priming from round one could also have led to maximal sentence type alignment values from very early on in the game. In addition, syntactic alignment in our task may be higher than in highly interactive tasks, where speakers alternate on a sentence-by-sentence basis, due to this cumulative priming since participants are not primed by just one occurrence of the prime structure, but by an average of six occurrences. Compare this to lexical alignment, which was not task-specific, but rather configuration-specific, in that lexical choices were influenced by the type, color, and locations of the objects to be placed in each round. Thus, lexemes could repeat with less frequency than sentence types and were thus less susceptible to cumulative priming, leaving more room to observe temporal effects.

Contribution of alignment to task performance

We showed that lexical alignment affected move completion time. In particular, lexical alignment functioned as a mediator, such that as rounds went by, participants were more lexically aligned, which in turn made them faster. This result is compatible with the interactive-alignment model and a conceptual pacts approach. Notice that lexical choices, in particular, shape, color, and location terms, were important for task success. Such successful synchronization of lexical choice routines likely reduced cognitive load and allowed faster object placement. Notice, however, that our measure of success, move completion time, is a rather course-grained measure and thus had the odds stacked against finding an effect of alignment on task success.

In contrast, syntactic alignment did not affect move completion time, a result which is neither in line with the interactive-alignment model nor with a conceptual pacts approach. However, this result may again be explained in terms of the task. While lexical choices were critical for task performance, syntactic choices were largely redundant for task performance, since the task was predetermined and never changed. The frequent use of fragments without a verb (such as *circle to the right of triangle*) best illustrates this: Such fragments, which only provided two object names with relational information, were possible because the task was given and did not have to be negotiated. Thus, syntactic alignment was not critical for task performance. In addition, sentence types differed in length. Compare *circle to the right of triangle* and *you should place the circle to* *the right of the triangle*. Even though we measured move completion time excluding the time it took to type a message, it is still possible that the mere length of the fragments or sentences affected the time it took to place object. In particular, reading time could have affected task performance.

Overall, the results show temporal effects of lexical alignment, but not syntactic alignment. The lack of syntactic effects, however, may be related to the task. Future studies could, for example, test whether syntactic alignment increases throughout the discourse and leads to faster task completion when using more than one task (e.g. placing objects, moving objects, deleting objects, changing objects in some way etc.). In such a scenario, there may be more syntactic choices and less cumulative priming, such that alignment could increase over the course of the discourse. In addition, syntax would be less redundant and may thus contribute reliably to task performance.

A more global measure of alignment

In this study, we used a more global measure of alignment than most carefully-controlled alignment studies. In typical psycholinguistic studies of alignment, participants alternate utterances on a sentence-by-sentence basis. Our data is different from such highly interactive scenarios. In particular, participants in our task gave on average six instructions before it was their interlocutor’s turn to give instructions. Thus, our approach produces a more global measure of alignment. This is also reflected in the measures we used to calculate alignment: The cosine similarity is often used to compute similarities across longer texts. We believe that our study provides a useful complement to studies with highly interactive tasks: In natural conversations, interlocutors do have highly interactive exchanges, where speakers alternate on a sentence-by-sentence or even phrase-by-phrase basis. However, in natural conversations, interlocutors also produce longer stretches of speech, for example, when talking about a recent situation, telling a story, providing information, giving instructions etc. Thus, in natural conversations, speaker and listener may switch roles very frequently at times and less frequently at other times. Our study thus nicely complements studies that found alignment in highly interactive settings and shows that similar alignment phenomena are observed in settings where speaker and hearer switch roles less frequently.

Conclusions

This study explored temporal aspects of lexical and syntactic alignment in task-oriented discourse. The results add to the sparse data currently available regarding alignment over the course of the discourse and the relationship of alignment and task success. Both these temporal aspects of alignment are critical for informing models of alignment. Our study also highlights how object choices and task constraints may affect temporal measures of lexical and syntactic alignment.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51(6)*:1173-1182.

Benus, S., Gravano, A., & Hirschberg, J. (2007). The prosody of backchannels in American English. *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany*, pages 1065-1068.

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

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition*, 75:B13–B25.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121:41–57.

Brennan, S. E. (2012). Conversation and dialogue. In H. Pashler (Ed.), *Encyclopedia of the Mind* (vol. 1, pp. 202-205). Los Angeles, CA: SAGE Publications.

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

Carbary, K. & Tanenhaus, M. (2011). Conceptual pacts, syntactic priming, and referential form. In K. van Deemter, A. Gatt, R. van Gompel, E. Krahmer (Eds.) *Proceedings of the CogSci Workshop on the Production of Referring Expressions: Bridging the Gap Between*

Computational, Empirical and Theoretical Approaches to Reference (PRE-CogSci 2011). Boston: Cognitive Science Society.

Carletta, J., Hill, R. L., Nicol, C., Taylor, T., de Ruiter, J. P., & Bard, E. G. (2006). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods, 42(1)*, 254-265.

Chang, F., Dell, G. S., and Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113(2):234–272.

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

Dubey, A., Keller, F., & Sturt, P. (2005). Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing. Vancouver, Canada*, pages 827–834.

Friedberg, H., Litman, D., & Paletz, S. B. F. (2012). Lexical entrainment and success in student engineering groups. *Proceedings of the Spoken Language Technology Workshop (SLT), 2012 IEEE, pages 404-409*.

Garrod, S. & Anderson, A. (1987). Saying What You Mean in Dialoge: A Study in Conceptual and Semantic Coordination. *Cognition*, 27:181–218.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2006). The impact of delayed visual feedback on collaborative performance. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. Olson (Eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*, ACM, New York, NY, USA, 1303-1312.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4):365–399.

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58:214–238.

Healey, P. G. T., Purver, M., & Howes, C. (2014). Divergence in Dialogue. PLoS ONE 9(6): e98598.

Howes, C., Healey, P. G. T., & Purver, M. (2010). Tracking lexical and syntactic alignment in conversation. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*, pages 2004–2009.

Jaeger, T. F. & Snider, N. E. (2013). Alignment as a consequence of expectation adaption: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127:57–83.

Kaschak, M., Loney, R., & Borreggine, K. (2006). Recent experience affects the strength of structural priming. *Cognition*, 99(3):B73-B82.

Klein, D. & Manning, C. D. (2002). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, & K. Obermayer (Eds.) *Advances in neural information processing systems 15* (pp. 3–10). Cambridge, MA: MIT Press.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1:1–12.

Krauss, R. M. & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–346.

Kraut, R. E., Gergle, D., & Fussell, S. R. (2002). The use of visual information in shared visual spaces: Informing the development of virtual co-presence. In *Proceedings of the 2002 ACM* *conference on Computer supported cooperative work* (CSCW '02). ACM, New York, NY, USA, 31-40.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Metzing, C. & Brennan, S. E. (2003). When conceptual pacts are broken:Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49:237–246.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 169–171.

Nilsenová, M. & van Amelsvoort, M. (2010). Syntax drives phonological choice - even independently of word choice. *Proceedings of the 32th Annual Meeting of the Cognitive Science Society*, 32:212–217.

Pickering, M. J. & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, 3(4):136–141.

Pickering, M. J. & Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological Bulletin*, 134(3):427–459.

Pickering, M. J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–225.

Rafferty, A. & Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. *Proceedings of the Workshop on Parsing German*, pages 40-46.

Reitter, D. & Moore, J. D. (2007). Predicting success in dialogue. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815.

Reitter, D. & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1–6.

Szmrecsanyi, B. (2005). Creatures of habit: A corpus-linguistic analysis of persistence in spoken english. *Corpus Linguistics and Linguistic Theory*, 1(1):113–149.

Westbrook, L. (2007). Chat reference communication patterns and implications: applying politeness theory. *Journal of Documentation*, 63(5):638–658.

Table 1: Colored geometric shapes chosen for each round

Round	**Geometric shapes**	**Round**	**Geometric shapes**
1	gray circle cyan circle red triangle magenta square pink rectangle	5	yellow circle cyan square blue square pink rectangle magenta ellipsis
2	gray circle green circle yellow triangle pink triangle pink rectangle	6	yellow triangle blue triangle green ellipsis red circle black rectangle
3	pink circle red ellipsis orange triangle gray triangle black rectangle	7	blue ellipsis cyan ellipsis magenta triangle blue rectangle green square
4	orange circle orange rectangle magenta circle black triangle black rectangle	8	black circle black ellipsis pink triangle pink rectangle green ellipsis

Table 2: Employed sentence types

Sentence type	**Example**
fragment without verb	Kreis rechts neben Dreieck (Circle to the right of triangle.)
fragment with verb	Kreis rechts neben Dreieck platzieren ((To) place circle to the right of triangle.)
sentence imperative	Platziere den Kreis rechts neben dem Dreieck. (Place the circle to the rigth of the triangle.)
sentence declarative	Das nächste Objekt ist ein Kreis. (The next object is a circle.)
sentence indicative	Du sollst den Kreis rechts neben dem Dreieck platzieren. (You should place the circle to the right of the triangle.)
sentence conjunctive	Würdest Du den Kreis rechts neben dem Dreieck platzieren? (Would you place the circle to the right of the triangle? ’)
sentence indirect indicative	Der Kreis kommt rechts neben das Dreieck. (The circle goes to the right of the triangle.)
sentence indirect indicative passive	Der Kreis muss rechts neben dem Dreieck platziert werden. (The circle must be placed to the right of the triangle.)

Table 3: Example sentence counts which illustrate the cosine similarity measure.

	Player/Scenario				
Sentence type	*A*	*B*/1	*B*/2	*B*/3	*B*/4
frequency fragment without verb	1	1	2	4	0
frequency sentence indicative	1	1	2	1	0
frequency sentence imperative	4	4	8	1	0
frequency sentence conjunctive	0	0	0	0	4
cosine similarity to A	–	1	1	0.5	0

Table 4: Top ten lemma types.

Nouns		**Adjectives**		**Verbs**	
item	**count**	**item**	**count**	**item**	**count**
Kreis		gelb		sein	
(circle)	231	(yellow)	117	(to be)	109
Dreieck		schwarz		befinden	
(triangle)	192	(black)	105	(to be located)	88
Rechteck		blau		liegen	
(rectangle)	118	(blue)	94	(to be located)	88
Spalte		rosa		setzen	
(column)	88	(rose-colored)	87	(to place)	40
Quadrat		unten		ist	
(square)	77	(at the bottom)	69	(is)	34
Ellipse		rot		müssen	
(ellipsis)	70	(red)	65	(must)	27
Mitte		pink		kommen	
(middle)	66	(pink)	64	(here: to go)	24
Reihe		grau		meinen	
(row)	49	(gray)	61	(to mean)	17
Viereck		grün		bewegen	
(quadrangle)	46	(green)	54	(to move)	15
Ecke		ganz		werden	
(corner)	44	(complete)	52	(will)	13

Table 5: Colors and expressions used to refer to them. Expressions that were likely produced in error are given in italics. Numbers of occurrences are given in parentheses.

Color	**Expressions**
blue	blau (blue, 63), dunkelblau (dark blue, 13)
cyan	blau (blue, 31), türkis (turquoise, 29), hellblau (light blue, 17), anderes blau (other blue, 1), türkisfarben (turquoise-colored, 1)
yellow	gelb (yellow, 51)
gray	grau (gray, 61), *schwarz (black, 1)*
green	grün (green, 64)
magenta	pink (pink, 61), rosa (rose-colored, 10), rot (red, 4), violett (violet, 1), lila (purple, 1)
orange	gelb (yellow, 66), gold (gold, 1)
pink	rosa (rose-colored, 80), hautfarben (skin-colored, 8), hellrot (light red, 8), pink (pink, 5), rosafarben (rose-colored, 5), rot (red, 4), beige (beige, 3)
red	rot (red, 57)
black	schwarz (black, 106)

Table 6: Shapes and expressions used to refer to them. Expressions that were likely produced in error are given in italics. Numbers of occurrences are given in parentheses.

Shape	**Expressions**
circle	Kreis (circle, 218), Punkt (dot, 19), Ball (ball, 4), Symbol (symbol, 1)
ellipsis	Ellipse (ellipsis, 69), Oval (oval, 30), Symbol (symbol, 8), Ding (thing, 7), Etwas (something, 3), ovale Figur (oval figure, 3), Figur (figure, 2), Teil (piece, 2), ovales Ding (oval thing, 1), ovaler Kreis (oval circle, 1)
rectangle	rechteck (rectangle, 115), viereck (quadrangle, 11), Balken (bar, 7), Symbol (symbol, 2), Kästchen (little box, 1), quadrat (square, 1)
square	quadrat (square, 65), viereck (quadrangle, 25), rechteck (rectangle, 2), *dreieck (triangle, 1)*
triangle	dreieck (triangle, 191), Figur (figure, 1), *circle (Kreis, 1), Rechteck (rectangle, 1), viereck (quadrangle, 1)*

Table 7: Sentence type categories and numbers of occurrence.

Sentence or Fragment Type	**N**
sentence: indirect indicative	340
fragment: without verb	306
sentence: imperative	50
sentence: declarative	34
fragment: with verb	32
sentence: indirect indicative passive	16
sentence: conjunctive	2
sentence: indicative	1

Table 8: Results of the mixed-effects models for lemma, sentence type, breadth, and depth alignment. All models included group (participants vs. baseline) as fixed effect, and subjects and round number as random effects.

response variable	**estimate**	**t-value**	**p-value**
lemma alignment	0.20857	9.359	< .0001
sentence type alignment	0.31951	4.494	< .0001
breadth alignment	0.021213	2.55	< .05
depth alignment	0.020205	1.92	= .0566

Table 9: Results of the linear models with lemma, sentence type, and breadth alignment as response variable. All models included round number as predictor variable.

response variable	**estimate**	**t-value**	**p-value**	**Adjusted R^2**
lemma alignment	0.034386	4.095	< .0001	0.1399
sentence type alignment	-0.005215	-0.249	= .8042	-0.009767
parse tree breadth alignment	-0.007174	-0.318	= .751	-0.009353

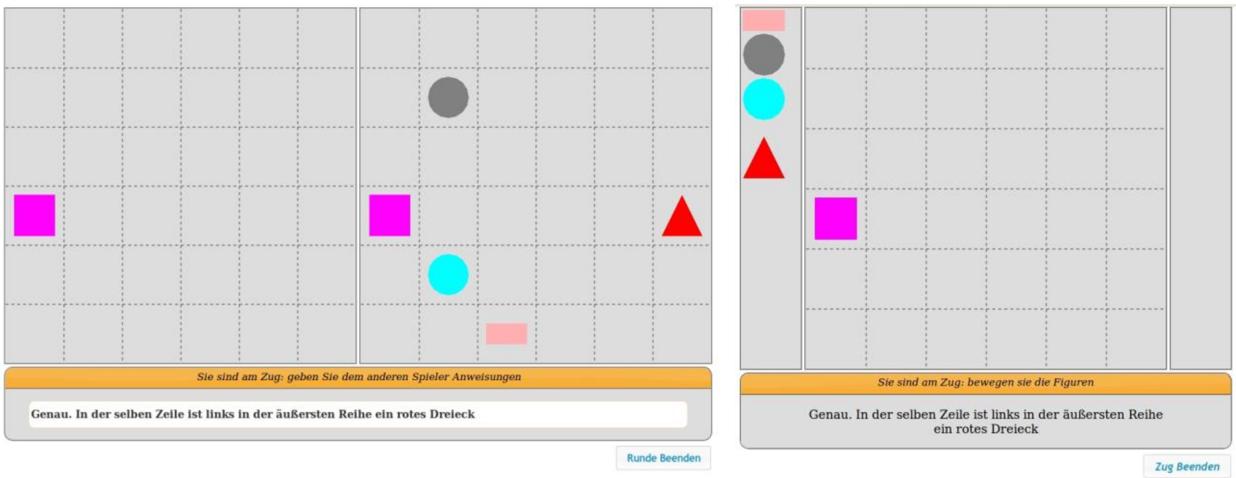

Figure 1: Left: Sample screen with two 6x6 grids, as seen by the director. The grid on the right shows the target arrangement of five colored, geometric shapes. The grid on the left shows the arrangement of shapes on the matcher's grid. Right: Sample screen, as seen by the matcher. Initially the 6x6 grid is empty and the target shapes are positioned outside of the grid.

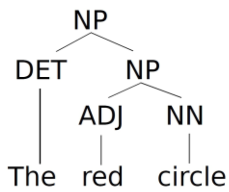

Figure 2: Example parse tree for the noun phrase "the red circle".

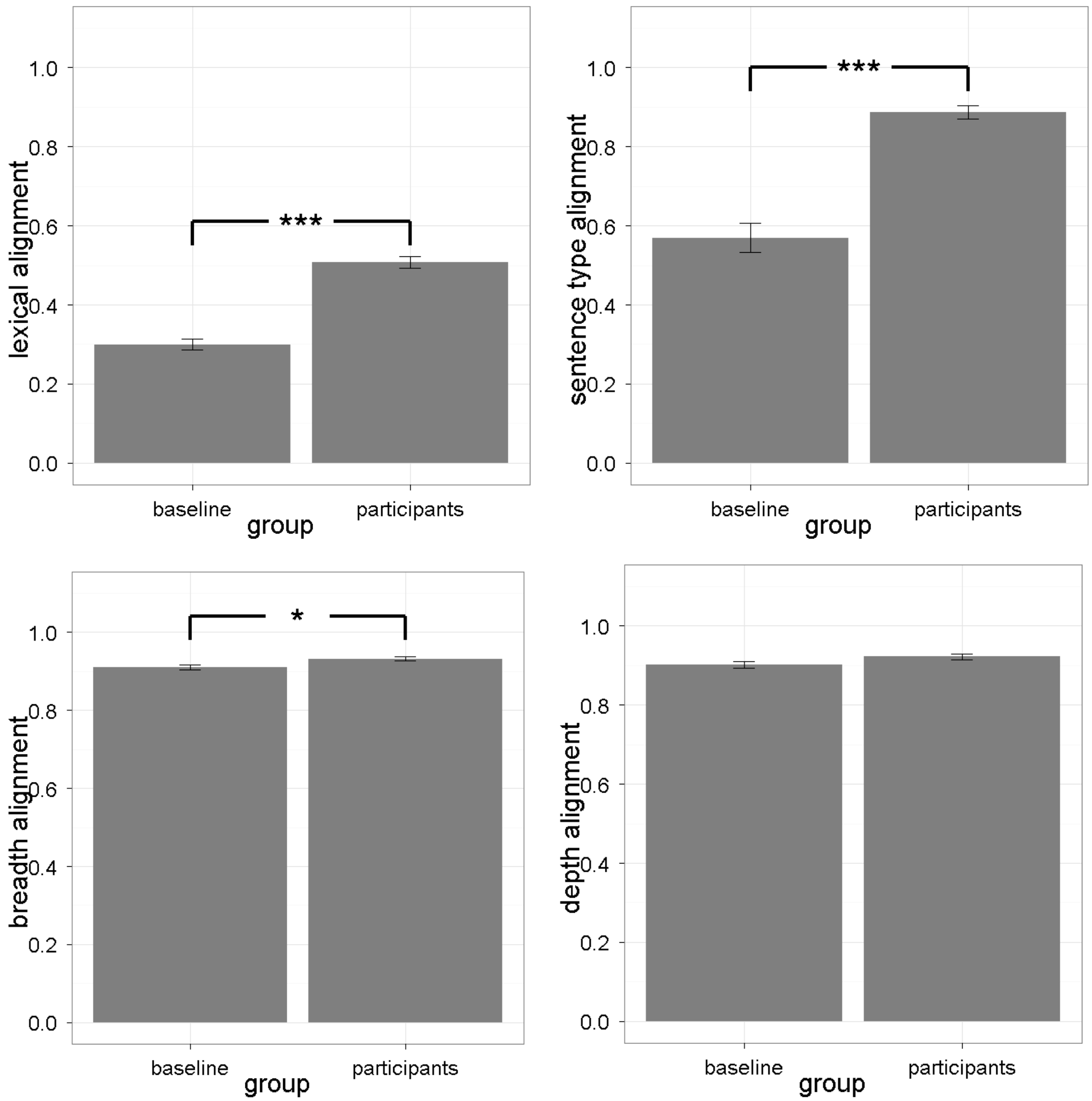

Figure 3: Means and standard errors for lemma, sentence type, breadth, and depth alignment in the baseline and for participants. The y-axes range from 0 (no alignment) to 1 (complete alignment).

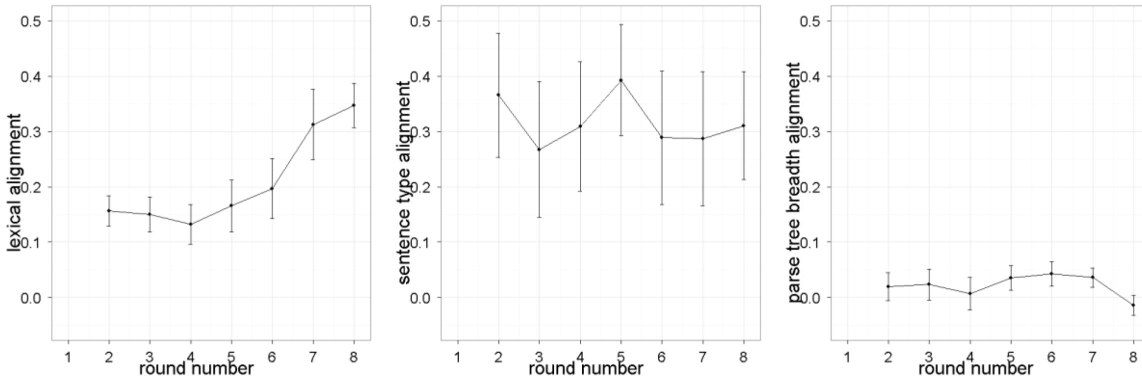

Figure 4: Mean lexical, sentence type and parse tree breadth alignment values with standard errors for rounds two through eight of the experiment. The y-axis range is kept constant across graphs for better comparability.