Beyond frequency counts: Novel conceptual recurrence analysis metrics to index semantic coordination in team communications

Michael T. Tolston¹ · Michael A. Riley² · Vincent Mancuso³ · Victor Finomore⁴ · Gregory J. Funke⁴

Published online: 3 October 2018

© Psychonomic Society, Inc. (This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply) 2018

Abstract

Semantic alignment is a key process underlying interpersonal and team communication. However, semantic similarity is difficult to quantify, and statistical approaches designed to measure it often rely on methods that make the identification of the relative importance of key words difficult. This study outlines how conceptual recurrence analysis (CRA) can address these issues and can be used to detect conceptual structure in interpersonal communication. We developed several novel CRA metrics to analyze communication data reported previously by Mancuso, Finomore, Rahill, Blair, and Funke (Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58, 405–409, 2014), gathered from teams who worked cooperatively on a logic puzzle under different cognitive biasing contexts. CRA, like other measures of semantic coordination, relies on parameters whose values affect estimates of semantic alignment. We evaluated how the dimensionality of semantic spaces affects metrics quantifying the conceptual similarity of communicative exchanges, and whether metrics calculated from top-down, a priori semantic spaces or bottom-up semantic spaces empirically derived from each data set were more sensitive to biasing context. We found that the novel CRA measures were sensitive to manipulations of cognitive bias, and that higher-dimensional, bottom-up semantic spaces generally yielded more sensitivity to the experimental manipulations, though when the communication was evaluated with respect to specific key concepts, lower-dimensional, top-down spaces performed nearly as well. We conclude that CRA is sensitive to experimental manipulations in ways consistent with prior findings and that it presents a customizable framework for testing predictions about interpersonal communication patterns and other linguistic exchanges.

Keywords Conceptual recurrence analysis · Text analysis · Semantic similarity · Semantic coordination · Team communication · Interpersonal communication

Language use during interpersonal conversation offers unique insight into the processes that support coordinated behaviors (Clark, 1996; Tylén, Weed, Wallentin, Roepstorff, & Frith, 2010). Importantly, the basis for conversation has been thought of as a continuous, oftentimes implicit, alignment of multiple scales of behaviors and mental processes, with nonlinear interactions between levels (Spevack, Falandays, Batzloff, & Spivey, 2018). For example, the interactive alignment model (Pickering

& Garrod, 2004) posits that establishment of the mutual understanding required for comprehension of linguistic exchanges occurs across multiple levels, ranging from phonetic, phonological, lexical, syntactic, semantic, and finally a situation model that represents specific contexts, with complex interactions occurring between these scales. Many of these types of interpersonal alignment are readily measurable. For instance, phonological (e.g., Shockley, Sabadini, & Fowler, 2004), syntactical (e.g., Healey, Purver, & Howes, 2014), and lexical and morphological alignment (e.g., Dale, Duran, & Coco, 2018; Dale & Spivey, 2005, 2006) have all been successfully quantified.

However, when evaluating alignment of the meaning of language—semantics—the ambiguity of terms often makes it difficult to accurately quantify semantic content without relying on human coders (Iliev, Dehghani, & Sagi, 2014). In part, this is because the meanings of words lie not in the words themselves, but in their relation to their external concepts or entities (i.e., the referential theory of meaning; Bunnin & Yu, 2004). Since measures of semantic alignment are difficult to

✉ Michael T. Tolston
michael.tolston.ctr@us.af.mil

¹ Ball Aerospace and Technologies Corporation, Dayton, OH, USA

² Center for Cognition, Action, & Perception, University of Cincinnati, Cincinnati, OH, USA

³ MIT Lincoln Laboratory, Lexington, MA, USA

⁴ Air Force Research Laboratory, Wright Patterson Air Force Base, Wright-Patterson AFB, OH, USA

obtain but provide insight into interpersonal alignment and its interactions across various scales (e.g., via network approaches; Paxton, Dale, & Richardson, 2016), the continued development of automated techniques that can quantify conceptual alignment to complement other measures of interpersonal alignment is advantageous (see also Cooke, Gorman, Myers, & Duran, 2013; Cooke, Salas, Kiekel, & Bell, 2004). While it is true that many existing techniques already serve this purpose, they are typically resource-intensive (e.g., Carletta et al., 1997), account for only a limited amount of the rich conceptual content of language-based communications (e.g., Orsucci et al., 2006), or are based on methods that can obfuscate the specific nature of the semantic relationship between similar words (see Broniatowski & Magee, 2012). In this article we evaluate a relatively new tool in the domain of semantic analysis—conceptual recurrence analysis (CRA)—that has properties that may address these concerns. Specifically, we determined the ability of CRA to detect the influence of the presentation of cognitively biasing information on interpersonal communications in a team problem-solving context. In doing so, we developed a set of metrics and methods to quantify specific aspects of semantic coordination and tested how variations in CRA parameters affected analysis outcomes.

Quantifying semantic content

Commonly used metrics that quantify communication, such as frequency counts of specific terms and speaking durations, can provide insight to the amount of information being shared during a team task (Wildman, Salas, & Scott, 2013) and are relatively simple to calculate. Additionally, it is possible to quantify functional content through compound measures of team interactions such as the anticipation ratio, which captures how often team members explicitly request information versus having that information prospectively supplied by teammates (Entin & Entin, 2001). However, both term-based and function-based metrics are limited to predefined dictionaries that may not be sensitive to idiosyncratic patterns of communication, nor do they account for the detailed semantic content of the communication. Such semantic content is traditionally evaluated through human raters (e.g., Volpe, Cannon-Bowers, Salas, & Spector, 1996). Because of the expertise humans have dealing with ambiguities in semantics, it has been recognized that hand-coding of dialogue for semantic content is the “gold standard” method of analyzing human communication (Iliev et al., 2014). However, the process is labor-intensive and time-consuming, and moreover, it has long been recognized that human bias and fatigue can undermine such efforts (cf. Indulska, Hovorka, & Recker, 2012; Weber, 1990). Issues of bias may be addressed with appropriate training and high standards of inter-rater reliability (Weber, 1990), but those solutions potentially exacerbate the costs of coding.

To avoid difficulties associated with subjective rater evaluations, researchers have turned to automated approaches to quantify semantic content. Some of these methods, such as latent semantic analysis (LSA), perform as well as human raters in certain contexts (e.g., Foltz, Laham, & Landauer, 1999). However, these approaches also have limitations. For instance, although LSA is a powerful tool for evaluating team communications (e.g., Foltz & Martin, 2009; Gorman, Foltz, Kiekel, Martin, & Cooke, 2003; Gorman, Martin, Dunbar, Stevens, & Galloway, 2016), it can require sizeable semantic spaces built from a large corpus of relevant documents prior to analyzing the data of primary interest. These semantic spaces can be nontrivial to build (Gefen, Endicott, Fresneda, Miller, & Larsen, 2017; Quesada, 2007). For instance, it is generally advisable to experiment with multiple dimensions, ways of normalizing the text, and different types of weights that index word-context co-occurrence (e.g., simple matching, entropy, inverse-document frequency). Furthermore, such constructed spaces may have only domain-specific applicability (Iliev et al., 2014), potentially reducing their usability.

In addition to LSA, there have also been several recent advances in conceptual analysis options, perhaps the most notable being Google’s word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Though powerful and fairly easy to implement with specialized packages (e.g., the Gensim library; Rehurek & Sojka, 2010), these new methods still suffer in part from a crucial drawback shared with LSA, in that the embeddings used to assess semantic similarity are high-dimensional mathematical spaces whose intrinsic meaning can be challenging to apprehend (Smalheiser & Bonifield, 2018). Though there has been research into techniques that attempt to address this issue (e.g., Luo, Liu, Luan, & Sun, 2015; Park, Bak, & Oh, 2017), generally these approaches make both the interpretation of the dimensions of the semantic space and understanding of the influence of specific keywords difficult. Further, though some of the simplicity of using word2vec comes from using pretrained embeddings, these spaces may not be optimal for particular applications, and training new embeddings can present several challenges (Smalheiser & Bonifield, 2018). CRA offers an enticing alternative that addresses these concerns.

Conceptual recurrence analysis

Conceptual recurrence analysis (CRA) is similar in several regards to categorical variations of recurrence quantification analysis that have been usefully applied to linguistic data to quantify linguistic and communicative coordination (Dale & Spivey, 2005, 2006; see also Coco & Dale, 2014; Dale et al., 2018; Orsucci et al., 2006). However, CRA differs from these approaches in that it evaluates the semantic similarity of utterances, or the alignment of the *meaning* of the content of the

utterances, rather than syntactical, morphological, or lexical alignment (however, see Dale et al., 2018, for an extension of their categorical recurrence approach in combination with word2vec to analyze semantic coordination). CRA is similar in this regard to other semantic similarity measures like LSA (Foltz, 1996; Gorman et al., 2003; Gorman et al., 2016).

CRA implemented using the software package Discursis has shown promise in qualitative (Angus, Smith, & Wiles, 2012a; Angus, Watson, Smith, Gallois, & Wiles, 2012) and quantitative (Angus, Smith, & Wiles, 2012b; Watson, Angus, Gore, & Farmer, 2015) analyses of relatively small data sets, utilizing the data set itself as the corpus for evaluating semantic content. For example, Watson et al. (2015) showed that there are several relationships between semantic coordination and effective patient–caregiver interactions. In effective interactions, the caregiver tends to align the conceptual content of his or her utterances to those of the patient, and these interactions are typically characterized by a strong leader–follower relation. It has also been shown that Discursis can be used to quantify content engagement between individuals with dementia and their caretakers, and that metrics associated with short-term topic repetition are related to subsets of communicative interactions manually coded as “active listening” (Atay et al., 2015). These relations show the potential of CRA metrics in the systematic evaluation of communicative content.

Importantly, Discursis uses a semantic space model based on the distribution of words in the analyzed text, meaning that issues related to idiosyncratic language usage are largely avoided. Additionally, the resulting semantic space has an intuitive structure; groups of words are classified according to a conceptual thesaurus that measures statistical similarity to important key words identified in the text (Smith, 2000), rather than abstract mathematical spaces with no such lexical anchors (as in LSA) that can be difficult for users to decipher (Broniatowski & Magee, 2012). In other words, Discursis provides a semantic space whose dimensions are key words chosen due to their relative importance in the discourse at hand.

The foundation of CRA: Computing the semantic similarity matrix

Discursis implements CRA by first determining the conceptual content of each utterance (e.g., the individual content of each speaker’s turn in the conversation). It then computes the relative similarity between all pairs of utterances, which yields a *similarity matrix* that describes the degree of semantic overlap among all utterances (see Fig. 1). The specific nature of computing semantic similarity has been reported elsewhere (Angus, Smith, & Wiles, 2012a, 2012b), but the general process is described briefly below.

The process of CRA (outlined in Fig. 1) starts with text preprocessing to remove stop words (articles, prepositions, and vocal gestures like “hmm”). The preprocessed text is then submitted to a concept-building routine, which starts by identifying a lexicon—the set of all the unique concepts (i.e., words) in the corpus. Key concepts to serve as the bases of dimensionality reduction and projection are then identified (e.g., as the most frequently occurring concepts), and the degree of similarity between each concept in the lexicon and each key concept is evaluated (e.g., using a naïve Bayes technique; Salton, 1989). The naïve Bayes measure quantifies any two concepts’ semantic similarity via the extent of their occurrence in the same context. For instance, interchangeable terms used in similar contexts have high semantic similarity (e.g., “rotation” and “shift” used in the context of discussing a personnel change). This bottom-up approach does not require any predetermined set of conceptual categories (although it does permit this; we evaluated using predetermined concepts in one of the analyses reported below). These similarity scores form a matrix, Key-Concepts $\times$ Concepts, with rows corresponding to key concepts and each concept corresponding to a column. Following this, each utterance is coded for the presence of each concept, forming a Boolean matrix that has the dimensions of Concepts $\times$ Utterances. Next, a Key-Concepts $\times$ Utterances matrix is calculated by multiplying the Key-Concepts $\times$ Concepts similarity matrix and a Concepts $\times$ Utterances Boolean matrix—which has entries equal to 1 if the concept occurs in a given utterance, and 0 otherwise. The resulting matrix codes the amount of similarity each utterance has with each key word.

The outlined dimensionality reduction yields multidimensional feature vectors that encode the degree of presence of each key concept in each utterance. Calculating the normalized vector dot product (cosine) between all pairs of feature vectors yields a final matrix of the amount of conceptual similarity between each utterance—in the degree to which the relative proportions of key terms occur—of the dimensions Utterances $\times$ Utterances. The pairwise similarity of utterances is presented in Discursis in the form of an Utterances $\times$ Utterances *conceptual recurrence plot* (see Figs. 3 and 4 in the Method section; Angus, Watson, et al., 2012), which makes it possible to visualize conceptual coordination within or across team members. These matrices give a score ranging continuously between 0 (*no similarity*) and 1 (*perfect alignment*) for each possible pair of utterances in a transcript coded for speaker.

The present study

In the present study we investigated whether automated analyses based on CRA can circumvent the limitations of

Fig. 1 A flowchart illustrating the process that generates the semantic similarity matrix used in conceptual recurrence analysis

manual coding and detect conceptual coordination in interpersonal communications consisting of a relatively small number of interactions, without the need to build an a priori semantic space from a preexisting corpus, and whether CRA-based metrics would be more sensitive to experimental manipulation than are simple frequency counts. Specifically, we evaluated the applicability of CRA, conducted using Discursis (Angus, Smith, & Wiles, 2012a, 2012b; Angus, Watson, et al., 2012), to analyze team communication data. In doing so, we developed a novel set of CRA metrics, including measures that capture the proportion of similarity accounted for by a given concept, broken down by the similarity between utterances that originate from the same speaker and those shared among speakers.

Since CRA in Discursis offers parameters that users can adjust, a goal of our analysis approach was to determine how the parameter settings in Discursis influence the identified semantic spaces, and how those modifications may subsequently influence the detection of group-level differences in the proposed metrics due to experimental manipulations. Specifically, since it has been shown that analyzing team communications using task-specific terms might provide a better basis for predicting communication effectiveness than does general vocabulary alignment (Fusaroli et al., 2012), we tested how Discursis can focus on the amount of semantic similarity that can be accounted for by projecting onto task-specific keywords. We did this by calculating the similarity of utterances similar to specific task-critical keywords and by contrasting

metrics from intrinsic semantic spaces to those obtained from a priori spaces created from a task-relevant corpus. We also tested the effects of using larger versus smaller semantic spaces on the sensitivity of the proposed CRA measures.

Method

Communication data set

To test the utility of summary measures obtained from CRA, as well as the influence of semantic-space parameters on these measures, we analyzed previously collected team chat data (Mancuso, Finomore, Rahill, Blair, & Funke, 2014). The data were obtained using the Experimental Laboratory for Investigating Collaboration Information-Sharing and Trust (ELICIT), a simulated intelligence, surveillance, and reconnaissance task in which teams work to solve a logic puzzle framed as an investigation of a pending terrorist attack (Finomore et al., 2013; Ruddy & Nissen, 2008). In ELICIT, each team member receives a unique set of 15 complimentary text-based statements (factoids) about an expected terrorist attack. These factoids hold information that varied in importance: expert factoids (~5%) hold explicit information about the specifics of the pending attack; key factoids (~19%) help identify which information is meaningful; support factoids (~26%) supplement expert and key factoids; and noise factoids (~50%) contain no information needed to solve the puzzle. To succeed, team members must share their information and come to a joint decision on the details of the upcoming attack (i.e., *Who will be attacking*, *What will be attacked*, *When will the attack take place*, and *Where will the attack occur*). All the information needed to solve the task is dispensed to the team; by sharing information and following the logic of the factoids teams can arrive at the correct conclusions. The factoids were presented to participants in this task at a fairly quick pace: Every 5 s, a set of four factoids was introduced (one factoid per person), meaning that within 75 s each team received all the factoids.

The purpose of the study conducted by Mancuso et al. (2014) was to analyze the effect of confirmation bias on team decision making. To this end, distractor factoids (nine in total) were introduced in ELICIT that implicated the wrong terrorist group in the pending attack—the factoids implicated the “green”-Incorrect (I) group, though the correct answer was the “blue”-Correct (C) group. The distractor factoids did not make the puzzle unsolvable. Rather, they introduced a counter narrative that potentially named the “green”-I group as the answer to the *Who* question, but with insufficient evidence to justify that answer. In the experiment, teams of four individuals communicated via Internet relay chat (IRC) to solve the logic puzzle, whereas misleading information was conditionally presented either early in participants’ information queue, late in the queue, or mixed (both early and late).¹ It is

important to note that if misleading information about “green”-I was presented early, then most information related to “blue”-C was introduced late in the queue, and vice versa. This biasing information was expected to complicate decision making and diminish team performance by inducing a confirmation bias, which is ultimately a failure to update beliefs in the presence of new information (Schippers, Edmondson, & West, 2014). Specifically, introducing the “green”-I distractor information early in the queue was expected to reduce conversation about later “blue”-C information and lead to an incorrect conclusion about the *Who* component of the puzzle.

Additionally, factoids about the “green”-I group were intentionally designed to be noisy (i.e., thematically incoherent), to allow teams to infer disjoint conclusions. For instance, consider the following factoids, and note that the last two were designed to create an implication that the “green”-I group may be behind the impending terrorist attack:

The target is a bank or military general.

Both cabinet members and a military general will be attending a political convention in New Zealand.

Rumor has it that the Green group has stolen uniforms from military bases in New Zealand and Australia.

Security forces for New Zealand’s political convention are providing extensive protection with military personnel only.

Since the target is possibly a military general and security is provided by military personnel only, a plausible inference is that the “green”-I group stole the uniforms to get access to the target. Other disjoint inferences are also possible—for example, regarding the “green”-I group communicating with locals about being oppressed by the military—that could plausibly establish motivation, but are either unrelated or only tangentially related to the other clues provided to the participant teams. Another difference between conditions was that information regarding the “blue”-C group given to specific individuals in the distractor-first condition was not always paired with its supported information, though it was in the distractor-last condition. Finally, the distractor-first condition contained an injected factoid that linked “green”-I to stolen uniforms but that did not do so in the distractor-last condition. Thus, the implication is that the biasing factoids would not only reduce the discussion of “blue”-C relevant information but would also create a noisy semantic profile.

Analyses of the number of exchanges referencing biasing information showed that teams were susceptible to it when it was introduced early in the queue (Mancuso et al., 2014). Specifically, participants in the early and mixed conditions

¹ The original study by Mancuso et al. (2014) included a control condition which was excluded from the present research because it used factoids that differed in content from the other conditions.

discussed the misleading factoids equally often as they discussed those leading to the correct answer, whereas participants who received biasing information late in their queue were more likely to focus on relevant factoids. Additionally, teams who received biasing information early in their queue were less likely to answer the *Who* part of the puzzle correctly than were teams in the other conditions. Given that the biasing factoids were found to influence team performance and communication in the Mancuso et al. study, we expected that the CRA analyses in the present study would also be sensitive to those experimental manipulations. Importantly, Mancuso and colleagues' findings regarding the effects of the order of biasing information on communication were obtained from analyses of hand-coded data that quantified how utterances were related to categories of factoids. We predicted that CRA would circumvent the need for such hand-coding and would be able to quantify the influence of the biasing factoids on team discourse. In particular, according to the hypothesis that teams will be susceptible to a confirmation bias, we expected a discounting of information that relates to “blue”-C, meaning that information semantically related to “blue”-C would be not be as prevalent in the distractor-first group.

Participants

The sample in the original study consisted of 64 paid participants (29 females), with ages ranging from 18–31 years ($M = 24.5$, $SD = 5.5$). Participants completed the study in four-person teams, yielding a total of 16 teams. All participants had normal hearing and normal or corrected-to-normal vision.

Analysis procedure

CRA parameter settings

The process of constructing a semantic space in Discursis (see Fig. 2) depends upon several user-defined settings that include: the upper limit of the number of key terms extracted; the identity of the corpus from which the key concepts are extracted (if the discourse being analyzed is not to be used to generate a set of intrinsic concepts); whether stemming—a preprocessing procedure to remove prefixes and suffixes—is used; and the list of stop words, among other options. For this project, we examined the relative influences of two of these settings: (1) the *source of key concepts* and (2) the *number of key concepts*. To do so, we created different sets of conceptual recurrence plots for all groups using two values for both parameters (see the example in Figs. 3 and 4). Each of these plots is summarized with the set of metrics introduced in the next section.

For the *source of key concepts*, we compared key concepts extracted from the communications themselves (intrinsic) against key concepts extracted from the factoid sets (factoid). Using the factoid sets to create an a priori conceptual space

might give insight into how task specific concepts constrained team conversations. We also believe that testing a priori bases might be a first step toward future implementation of CRA as a tool for real-time analysis of team communication. For *number of key concepts*, we compared two methods: automatic identification of the number of key concepts (automatic) and a set upper limit of 100 (N100, the default setting in Discursis). We note here that the term “automatic concept extraction” is potentially misleading, because for both this method and the method that sets a ceiling limit on the number of identified concepts, the concepts are extracted automatically rather than manually. The potential importance of this setting lies in the granularity of the constructed semantic space; larger semantic spaces have more specific terms (Smith & Humphreys, 2006) that can distinguish otherwise similar utterances. In the data we analyzed, the average number of key concepts identified in the N100 analysis was almost three times greater than in the automatic analysis.

We ran CRA with all concepts or only certain concepts activated to test the influence of order of biasing information on utterances related to task-specific keywords. Specifically, we created exclusive-or filters that passed utterances conceptually similar to the names of either of two ELICIT terrorist groups, “green”-I or “blue”-C, but not both simultaneously. In both cases, we ran the analyses with intrinsic and factoid concept bases and using the N100 and automatic concept extraction methods.

For all analyses, the “merge word variants” option was selected in Discursis, which identified plural and singular words (e.g., “general” and “generals”) or words with different tense suffixes (e.g., “attack” and “attacked”) as the same. Such stemming procedures are commonly used in preparing texts for semantic analysis (Turney & Pantel, 2010), and preliminary evaluations of ELICIT data indicated that this option essentially filters the data and provides a clearer picture of conceptual coordination (although in some cases there may be good reasons to leave this option unchecked; Smith, 2000). As with many data preprocessing procedures, stemming alters the form of the data in ways intended to clarify the signal, but that can also introduce noise (i.e., by mapping words to the same root that should be separated—overstemming—or mapping words to different roots that should be the same—understemming). For instance, the above example regarding the semantic similarity of “general” and “generals” may raise concerns about varying definitions of “general” (e.g., “a commander of an army” vs. “widespread”). In the present context, both words frequently refer to the same type of entity (a commander of an army—e.g., “Generals in New Zealand have private guards” and “The target is a bank or military general”), but it bears noting that such morphological limitations are at the crux of automatic processing of semantic content of language and that no stemming algorithm is perfect (Paice, 1996).

Fig. 2 A flowchart illustrating workflow in Discursis

Metrics

Our analyses focused on global aspects of team conceptual cohesion, where by “global” we mean average measures of whole conversations. Such measures have been used to quantify team (Gorman et al., 2003, 2016) and dyadic (Babcock, Ta, & Ickes, 2014) communications, but here we show that CRA as implemented in Discursis adds concept-specific and speaker-specific flexibility to the available analyses. We report a set of novel summary statistics based on partitioning similarity matrices into self and shared-with-others similarity, and by similarity with particular concepts. The metrics described below are not those automatically output by Discursis (which does provide its own set of automatically generated measures), but they are, we believe,

among the most basic measures that can be computed from the Discursis output of the similarity matrices. One measure in particular—mean similarity—appeared to be the most sensitive to the biasing information experimental manipulation, but other metrics may also prove useful in different circumstances.

Total similarity Total similarity (TS) is the grand sum of all cells (where each cell contains the conceptual similarity between a pair of utterances) above the diagonal of the similarity matrix, S . The lower triangular part is disregarded due to symmetry and the similarity of an utterance with itself is ignored. TS serves as measure and as a normalization factor for other measures and ratios. It is defined as

$$TS = \sum_{i=1}^{N-1} \sum_{j=i+1}^N S(i, j). \quad (1)$$

Fig. 3 Conceptual recurrence plot for ELICIT task performance. Each of the four team members is shown in a different color. Elements along the diagonal represent distinct conversation turns, with the element size proportional to the utterance length. Off-diagonal elements indicates conceptual matches, and empty (white) elements are utterances that do not match. Long vertical bands of shaded elements indicate when a concept influenced later utterances. Long horizontal bands indicate when earlier concepts were revisited or summarized. The degree of conceptual similarity is indicated by the degree of shading. Conversation themes are indicated by verbal labels above the diagonal. We note that the formulas presented in our text calculate values of similarity using cells above the diagonal; since the matrix is symmetric, the upper and lower parts are equivalent

Fig. 4 A closer inspection of the utterances represented in Fig. 3. The utterance number precedes the actual utterance, and the keywords of each utterance follow in brackets. Highly saturated cells indicate that the utterances are closely aligned in content (in terms of the relative proportion of the presence of key concepts), and light cells indicate

As with all reported measures, the similarity sum may be divided into self and shared (Angus, Smith, & Wiles, 2012a) by multiplying each entry in the similarity matrix by either a self-matrix,

$$TS_{self} = \sum_{i=1}^{N-1} \sum_{j=i+1}^N self(i, j)S(i, j), \quad (2)$$

or a shared matrix,

$$TS_{shared} = \sum_{i=1}^{N-1} \sum_{j=i+1}^N shared(i, j)S(i, j), \quad (3)$$

where the values in the self-matrix are 1 when the rows and columns are coded as originating from the same speaker, and 0 otherwise; the shared matrix has the opposite coding.

Recurrence Recurrence (REC) is a count of the number of cells above the diagonal of the similarity matrix S that exceed a threshold of similarity, denoted by ε . REC is a normalization factor for several other metrics, or is itself normalized by the number of utterances to yield a separate metric—percent recurrence. For our purposes, we chose ε to be equal to the smallest nonzero similarity score in the matrix so that all nonzero entries were counted,

$$\varepsilon = \inf\{x \in S : 0 < x\}. \quad (4)$$

utterances that are aligned along only a subset of the key concepts in either utterance. The three utterances following the top question focus on the “green”-I group and uniforms (a key piece of misleading information implicating “green”-I), whereas the bottom four utterances show a transition to discussing date and place

Once the threshold has been chosen, REC is computed as the sum of entries in S above the diagonal with a similarity equal to or greater than the threshold,

$$REC = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \Theta(S(i, j) - \epsilon), \quad (5)$$

where Θ denotes the Heaviside function,

$$\Theta(x) = \begin{cases} 1 & |x \geq 0; \\ 0 & |x < 0 \end{cases}. \quad (6)$$

Overall similarity Overall similarity (OS) is a measure of the average amount of similarity in cells above the diagonal of S , including zero cells. It is obtained by dividing total similarity by the total number of observations,

$$OS = \frac{TS}{\frac{N^2 - N}{2}}, \quad (7)$$

where N is the number of utterances. OS indexes the overall average density of similarity between all counted utterances, and can be broken down in terms of utterances that are made by the same person (OS_{self}),

$$OS_{self} = \frac{TS_{self}}{\frac{N^2 - N}{2}}, \quad (8)$$

and utterances that are conceptually similar but that originated from different speakers (OS_{shared}),

$$OS_{shared} = \frac{TS_{shared}}{\frac{N^2 - N}{2}}. \quad (9)$$

Mean similarity Mean similarity (MS) is a measure of the average similarity of utterances similar to at least one key concept. It is the total similarity divided by the number of pairs of utterances that have some minimum amount of similarity between them,

$$MS = \frac{TS}{REC}. \quad (10)$$

Conceptually, MS indexes the average degree of alignment between utterances that are projected onto at least one shared dimension in the semantic space. In other words, this value indicates how similar, on average, similar utterances tend to be, and can again be analyzed in terms of utterances that are made by the same person (MS_{self}) and utterances made by different people (MS_{shared}).

Percent recurrence Percent recurrence (PREC) is the percentage of similar utterances to all utterances,

$$PREC = \frac{REC}{\frac{N^2 - N}{2}} \times 100. \quad (11)$$

This measure correlates highly, though not entirely, with OS. It can also be divided into utterances that are similar and repeated by the same person, *PREC-Self*, and utterances that are similar but made by different people, *PREC-Shared*.

Filtering by concepts

A conceptual recurrence matrix can be filtered by a subset of concepts (Angus, Smith, & Wiles, 2012b). Furthermore, this process may be extended by evaluating the exclusive appearance of a particular concept with respect to another. For instance, as part of the present analyses, we evaluated the occurrence of utterances similar to a particular concept (i.e., “blue”-C a correct answer to the *Who* component in the ELICIT task), but in which similarity to another specific concept was not present (i.e., “green”-I, an incorrect answer that was emphasized in the biasing factoids). We thus obtained an estimate of how tightly the groups focused on the correct or incorrect answer.

To filter a similarity matrix by a given concept means to eliminate all utterances from the analysis save those similar to the specified concept or concepts. The first step is to create a Boolean column vector, (of size Utterances $\times$ 1) containing a 1 if the specified concept is present in a given utterance, and a 0 otherwise. This vector may be calculated from the Discursis “concepts” output, an exportable file that gives the projection of every utterance onto the set of key concepts. To inclusively select more concepts, it is sufficient to take the Boolean of the vector sum of desired vectors. To create an exclusive-or filter, we applied the Heaviside function to the difference of the two concept vectors,

$$U_{concept} = \Theta(U_i - U_j) \quad (12)$$

where U_i is the vector mapping of utterances to concepts to be kept, and U_j is a vector mapping of utterances to concepts to be excluded. The filter is then created by taking the product of the concept vector with its transpose,

$$M_{concept} = U_{concept} \times U_{concept}^T, \quad (13)$$

resulting in an Utterances $\times$ Utterances matrix. The filter is then applied to the original similarity matrix by taking the Hadamard (element-wise) product of the original similarity matrix and the filter,

$$S_{concept} = S \circ M_{concept}. \quad (14)$$

This similarity matrix then forms the basis for obtaining concept-specific derivations of the previously specified metrics ($REC_{concept}$, $OS_{concept}$, and $MS_{concept}$), each of which can

Table 1 Proportion of similarity accounted for by concept-specific measures

Proportion	Supporting Equation	Description
Proportion-overall-similarity-concept	$\frac{OS_{concept}}{OS}$	The proportion of total similarity that is accounted for by cells containing the specified concept or concepts
Proportion-mean-similarity-concept	$\frac{MS_{concept}}{MS}$	The proportion of similarity that is accounted for by cells containing the specified concept or concepts
Proportion-self-similarity-concept	$\frac{MS_{concept-self}}{MS_{self}}$	The proportion of <i>self-similarity</i> that is accounted for by cells containing the specified concept or concepts
Proportion-shared-similarity-concept	$\frac{MS_{concept-shared}}{MS_{shared}}$	The proportion of <i>shared similarity</i> that is accounted for by cells containing the specified concept or concepts
Proportion-REC-concept	$\frac{REC_{concept}}{REC}$	The proportion of similar cells that contain the specified concept or concepts
Proportion-REC-self-concept	$\frac{REC_{concept-self}}{REC_{self}}$	The proportion of self-similar cells that contain the specified concept or concepts
Proportion-REC-shared-concept	$\frac{REC_{concept-shared}}{REC_{shared}}$	The proportion of shared similar cells that that contain the specified concept or concepts

be used to extract ratios that quantify the proportion of similarity accounted for by various partitions of the filtered matrix (see Table 1). We note that care should be taken in creating such exclusive-or filters, since utterances that are similar to both concepts are filtered out.

Preprocessing

Although CRA in Discursis can automate many of the steps involved in conceptual analyses of textual data, it cannot easily deal with misspelled words, since each spelling variation would be identified as a unique concept. Furthermore, abbreviations occasionally used in place of semantically identical terms may have a similar effect of inflating the number of concepts. We treated both events as noisy processes that increase the dimensionality of a semantic space without adding meaningful conceptual discrimination. As such, prior to submission to CRA in Discursis, the data were processed by correcting misspellings and by mapping semantically identical abbreviations to common terms. For example, New Zealand—mentioned in some factoids—was referred to as “Zealand,” “New Zealand,” and “NZ” in the participant discussions; these were all changed to the term “Zealand.” We note that the latter step of mapping abbreviations to common terms is not necessary in Discursis but was an a priori decision made in order to minimize noise in a relatively small data set.

Analyses

To test the sensitivity of simple frequency counts to the order of biasing information, the relative frequencies of “blue”-C and “green”-I were calculated from the proportion of utterances produced by each team that contained either term. These results were submitted to a Concept (“blue”-C, “green”-I) × Order of Biasing Information (distractor-first, distractor-last,

and distractor-mixed) mixed ANOVA, with concept as a within-groups factor and order of biasing information as a between-groups factor.

As we described previously, it is possible to filter the conceptual recurrence matrix to emphasize or remove the influence of specific concepts in the analysis. In present analyses, the concepts “blue”-C and “green”-I were of import. As such, we conducted our analyses first by including all key concepts identified by Discursis (all key concepts active), and then by filtering out all key concepts except “blue”-C and “green”-I (selected key concept active). In addition to deriving key concepts for each individual set of communications, it is possible to supply a separate corpus to Discursis to identify key concepts as a basis for analyzing semantic similarity. This may be of interest for

Fig. 5 Significant Source × Number of Concepts Extracted interaction on the number of key concepts identified. Error bars correspond to ± 1 standard deviation. A significant simple effect for source of key concepts was observed: More key concepts were identified in the intrinsic N100 settings than in the factoid N100 settings

evaluating the extent to which utterances align along predetermined dimensions. For this study, we tested the difference in sensitivity to experimental conditions provided by projecting conversations onto bases that were extracted from the sets of factoids (factoid), or onto bases that were determined individually for each group from their actual communications (intrinsic). Finally, it is possible to set an upper limit on the number of key concepts that Discursis may extract from the corpus. We first verified that this setting made a difference in the number of extracted key concepts for both the intrinsic and factoid bases, and then we evaluated whether allowing the number of concepts to be automatically determined by Discursis (automatic) resulted in output that showed more or less sensitivity to the manipulations than did setting the maximum number of concepts to 100 (N100; the default setting).

To verify that the number of key concepts identified by Discursis was higher in the N100 than in the automatic condition, the number of key concepts identified for each group was submitted to a Number of Concepts Extracted (automatic, N100) $\times$ Source of Key Concepts (intrinsic, factoid) $\times$ Order of Biasing Information (distractor-first, distractor-last, and distractor-mixed) mixed ANOVA, with number of extracted concepts and source of key concepts as within-groups factors and order of biasing information as a between-groups factor.

The influences of three parameters on CRA similarity measures are reported: (1) the number of active key concepts (all key concepts, selected key concepts), (2) the source of the key concepts (intrinsic, factoid), and (3) the number of concepts extracted (N100, automatic). For each parameter setting, dependent measures of conceptual similarity were submitted to one-way between-groups ANOVAs, with order of biasing information as the factor. Measures included *MS*, the average degree of alignment of similar utterances; *OS*, the average amount of semantic similarity; and percent recurrence (*PREC*), the percentage of semantically similar utterances. When the source of similarity was investigated, similarity type (self, shared) was added as a within-groups ANOVA factor. When the degree of similarity as a function of specific terms was investigated (e.g., *PREC*-Concept—the proportion of similar utterances exclusively similar to either “green”-I or “blue”-C), term (“blue”-C, “green”-I) was added as a within-groups ANOVA factor. Due to the similarity of the latter analyses to those obtained from the “blue”-C and “green”-I specific proportions of *MS*, *TS*, *OS*, and *PREC*, results from the measures listed in Table 1 are not reported.

Prior to conducting planned inferential statistics, outlier analyses were performed using a repeated Grubb’s test within each condition separately for the analyses based on intrinsically defined concepts and for the analyses with concepts derived from the factoid lists. If a team was an outlier on more than one metric, that team was dropped from all analyses. Overall, three teams were removed (two from the distractor-first condition and one from the distractor-mixed condition).

This resulted in a final sample of nine teams, with three teams in each of the bias conditions.

Results

Since our goals in these analyses were to evaluate the utility of CRA and to demonstrate the effects of several parameters on CRA metrics, omnibus tests were conducted with $\alpha = .05$, and Tukey’s HSD test was used for post-hoc comparisons ($\alpha_{fw} = .05$). Only statistically significant results are reported. No significant results were found for *OS*.

Frequency counts of “blue”-C and “green”-I

The results showed a main effect of concept: The proportion of utterances containing “blue”-C was higher ($M = .179$, $SD = .049$) than of those containing “green”-I ($M = .124$, $SD = .056$), $F(1, 6) = 10.486$, $p = .018$, $\eta_p^2 = .636$.

Number of key concepts identified

Mean results for the effects of source and number of concepts extracted on number of key concepts identified can be seen in Fig. 5. We found a significant main effect of source of key concepts on the number of key concepts identified, $F(1, 6) = 120.484$, $p < .001$, $\eta_p^2 = .953$: The intrinsic extraction method resulted in a larger number of key concepts than did the factoid method. There was also a significant main effect of the number of concepts extracted, $F(1, 6) = 282.245$, $p < .001$, $\eta_p^2 = .979$: The number of key concepts was higher using the N100 than using the auto parameter setting. A significant interaction between source and number of concepts extracted was also visible, $F(1, 6) = 117.990$, $p < .001$, $\eta_p^2 = .952$. Simple-effects analyses showed that the interaction arose because source affected the number of key concepts identified only in the N100 condition, $F(1, 6) = 144.980$, $p < .001$, $\eta_p^2 = .960$, with the intrinsic bases having a larger number of identified key concepts than the factoid bases.

Semantic similarity: All key concepts active

Intrinsic concept analyses

N100 concepts extraction analyses Mean similarity (*MS*) was sensitive to the order of biasing information, $F(2, 6) = 21.74$, $p = .002$, $\eta_p^2 = .88$. *MS* in the distractor-first condition ($M = .373$, $SD = .003$) was significantly lower than in both the distractor-mixed ($M = .409$, $SD = .018$, $p = .001$), and the distractor-last ($M = .433$, $SD = .006$, $p = .019$) conditions. This means that utterances that were conceptually similar were overall less so

when the biasing information was introduced early in the factoid queue.

MS was parsed into self and shared similarity, and similarity type (self, shared) was added as a within-group analysis of variance (ANOVA) factor. Mean results for the effects of similarity type and order of biasing information on MS can be seen in Fig. 6. There was a significant main effect of similarity type: Self-similarity was greater than shared similarity, $F(2, 6) = 388.29$, $p < .001$, $\eta_p^2 = .98$. A significant Similarity Type $\times$ Order of Biasing Information interaction was present, $F(2, 6) = 9.67$, $p = .001$, $\eta_p^2 = .76$, and an order of biasing information effect was present for both self-similarity, $F(2, 6) = 28.48$, $p = .001$, $\eta_p^2 = .905$, and shared similarity, $F(2, 6) = 20.486$, $p = .002$, $\eta_p^2 = .872$. The interaction arose because the order of biasing information had a stronger effect on self- than on shared similarity. For self-similarity, post-hoc tests showed that all conditions differed significantly from each other, with $ps = .001$, $.048$, and $.011$ for differences between the distractor-first and distractor-last, distractor-first and distractor-mixed, and distractor-last and distractor-mixed conditions. For shared similarity, distractor-first was significantly lower than both distractor-last, $p = .002$, and distractor-mixed, $p = .015$. In general, participants were less aligned in the semantic content of their utterances in the distractor-first than in the distractor-last condition.

Automatic concept extraction analyses The only significant effect using automatic concept extraction was a main effect of similarity type on MS: Self-similarity ($M = .515$, $SD = .018$) was greater than shared similarity ($M = .483$, $SD = .028$), $F(2,$

$6) = 77.25$, $p < .001$, $\eta_p^2 = .928$. We observed that automatic concept extraction yielded higher similarity metrics, overall, than the N100 analysis. This was expected, since lowering the number of key concepts reduces the ways for utterances to be differentiated.

Factoid-derived concept analyses

N100 concept extraction analyses The only significant factor in the factoid-derived concept N100 analyses was of similarity type on MS: Self-similarity ($M = .475$, $SD = 0.03$) was greater than shared similarity ($M = .441$, $SD = .034$), $F(2, 6) = 90.19$, $p < .001$, $\eta_p^2 = .938$.

Automatic concept extraction analyses Consistent with the N100 method, for automatic concept extraction MS was affected by similarity type: Self-similarity ($M = .541$, $SD = 0.018$) was significantly greater than shared similarity ($M = .511$, $SD = .022$), $F(2, 6) = 34.347$, $p = .001$, $\eta_p^2 = .851$.

All key concepts active: Summary

To summarize thus far, when all key concepts were included in the calculations of mean similarity, differences were detected between self- and shared similarity of utterances, regardless of the source of key concepts (intrinsic vs. factoid) and the number of concepts extracted (N100 vs. automatic). Only the intrinsic N100 analyses revealed differences that were linked to the experimental manipulations. In other words, in the analyses using the larger bases derived empirically from the participants' conversations, bias condition influenced both self and shared semantic content. This difference is reflected in

Fig. 6 Significant Similarity Type $\times$ Condition interaction for mean similarity, with an upper limit of concepts extracted from the data set (intrinsic) set to 100 (N100). Error bars correspond to ± 1 standard deviation. Significant simple effects of both condition and similarity type were

observed. Condition affected self-similarity more than shared-similarity, and the similarity difference was greater in the distractor-first than in the distractor-last and -mixed conditions

Fig. 7 Conceptual recurrence plots for a distractor-first team with selective activation of the concepts blue (left) and green (right). The same data with all concepts activated are shown in Fig. 3. Blue does not recur

frequently, whereas green recurs prominently throughout, clearly showing the effect of the order-of-biasing-information manipulation. This team incorrectly answered “green”-I to the “Who will be attacking?” question

the interaction between source and the number of key concepts extracted on the number of key concepts identified: Using the N100 setting, significantly more key concepts were identified from the intrinsic than from the factoid sources. For all analyses, self-similarity was significantly greater than shared similarity.

Semantic similarity: Selected key concepts activated

Two specific concepts were chosen to investigate in detail: “green”-I—the terrorist group incorrectly implicated by the biasing factoids—and “blue”-C—the name of the terrorist group planning the attack. Conceptual filters were designed that passed only utterances similar to one keyword but not the other—utterances could have some similarity to the concept “blue”-C but not to “green”-I, or vice versa. Two different similarity matrices were obtained from each team using the exclusive-or filters. The dependent variables from these matrices were then submitted to a Similarity Type (self, shared) $\times$ Term (“green”-I, “blue”-C) $\times$ Order of Biasing Information (distractor-first, distractor-last, distractor-mixed) mixed ANOVA, with similarity type and term as within-groups factors and order of biasing information as a between-groups factor. Two example conceptual recurrence plots with either “blue”-C only or “green”-I only activated are shown in Fig. 7 (Fig. 3 shows the same trial with all concepts active).

Intrinsic concept analyses

N100 concept extraction analyses There was a main effect of type on MS, $F(1, 6) = 8.698, p = .026, \eta_p^2 = .592$. Average MS was higher for self-similar utterances ($M = .505, SD = .032$) than for utterances that were shared ($M = .468, SD = .062$).

We also found a significant effect of term on percent recurrence (PREC), $F(1, 6) = 16.195, p = .007, \eta_p^2 = .730$. PREC-Green ($M = .043, SD = .038$) was significantly lower than PREC-Blue ($M = .126, SD = .065$). There was a significant Term $\times$ Order of Biasing Information interaction, $F(2, 6) = 6.126, p = .036, \eta_p^2 = .671$. Simple-effects analysis showed that the order of biasing information affected PREC-Blue, $F(1, 6) = 7.386, p = .024, \eta_p^2 = .711$. Post-hoc analyses showed that utterances similar to “blue”-C accounted for a smaller portion of the total number of similar utterances in the distractor-first condition ($M = .059, SD = .033$) than in the distractor-last condition ($M = .185, SD = .061$), $p = .02$. Neither distractor-first nor distractor-last was different from the distractor-mixed condition ($M = .134, SD = .009$), $ps = .137$ and $.334$, respectively.

Automatic concept extraction analyses Here we found a main effect of similarity type on MS, $F(1, 6) = 7.133, p = .037, \eta_p^2 = .543$. Average MS was higher for self-similar utterances ($M = 0.61, SD = 0.035$) than for utterances that were shared ($M = .574, SD = .047$).

There was also a significant effect of term on the PREC, $F(1, 6) = 15.856, p = .007, \eta_p^2 = .725$. PREC-Green ($M = .045, SD = .040$) was significantly lower than PREC-Blue ($M = .134, SD = .067$). A significant term $\times$ order of biasing information interaction, $F(2, 6) = 5.647, p = .042, \eta_p^2 = .653$. Simple-effects analysis showed that the order of biasing information had no effect on PREC-Green ($p = .297, \eta_p^2 = .333$). However, order did affect PREC-Blue, $F(1, 6) = 7.271, p = .025, \eta_p^2 = .708$. Post-hoc analyses showed that utterances similar to “blue”-C in the distractor-first condition ($M = .064, SD = .039$) accounted for a smaller portion of the total

number of similar utterances than in the distractor-last condition ($M = .193$, $SD = .060$), $p = .022$.

Factoid-derived concept analyses

N100 concept extraction analyses There was a significant Term $\times$ Order of Biasing Information interaction for MS, $F(2, 6) = 8.673$, $p = .017$, $\eta_p^2 = .743$. Simple-effects analysis showed that, when “green”-I was the activated term, MS was significantly affected by condition, $F(2, 6) = 6.898$, $p = .028$, $\eta_p^2 = .697$. Post-hoc analyses showed that similar utterances made by teams in the distractor-last condition ($M = .632$, $SD = .044$) were more aligned than similar utterances made by teams in the distractor-mixed condition ($M = .473$, $SD = .07$), $p = .023$. When “blue”-C was the activated term, MS was also significantly affected by condition, $F(2, 6) = 6.083$, $p = .036$, $\eta_p^2 = .67$. Post-hoc analyses showed that the similarity between aligned utterances in teams in the distractor-first condition ($M = .509$, $SD = .021$) was significantly lower than that of teams in the distractor-mixed condition ($M = .578$, $SD = .033$), $p = .03$.

There was a main effect of term on PREC-Concept, $F(1, 6) = 15.994$, $p = .007$, $\eta_p^2 = .727$: PREC-Blue ($M = .138$, $SD = .07$) was significantly higher than PREC-Green ($M = .045$, $SD = .039$). There was also a significant Term $\times$ Order of Biasing Information interaction $F(2, 6) = 5.501$, $p = .044$, $\eta_p^2 = .647$. Simple-effects analysis showed that the order of biasing information significantly affected PREC-Blue, $F(2, 6) = 7.449$, $p = .024$, $\eta_p^2 = .713$. Post-hoc analyses showed that the distractor-first condition ($M = .064$, $SD = .038$) was significantly lower than the distractor-last condition ($M = .199$, $SD = .062$), $p = .009$.

Automatic concept extraction analyses We found a significant Term $\times$ Order of Biasing Information interaction on MS, $F(2, 6) = 8.54$, $p = .018$, $\eta_p^2 = .74$. Simple-effects analysis of the effect of the order of biasing information broken down by term showed that when “blue”-C was the activated concept, a significant effect of order of biasing information emerged, $F(2, 6) = 11.944$, $p = .008$, $\eta_p^2 = .799$. Post-hoc analyses showed that there was a significant difference between the distractor-first ($M = .596$, $SD = .026$) and distractor-mixed ($M = .655$, $SD < .001$) conditions, $p = .009$, and between the distractor-mixed and distractor-last conditions ($M = .601$, $SD = .008$), $p = .023$.

There was a main effect of term on PREC-Concept, $F(1, 6) = 14.479$, $p = .009$, $\eta_p^2 = .707$: PREC-Blue was significantly higher ($M = .147$, $SD = .071$) than PREC-Green ($M = .050$, $SD = .044$).

Selected key concepts: Summary

In the selected concepts active analyses, automatic concept extraction offered sensitivity to the experimental manipulation largely on par with the N100 method. Furthermore, in addition

to MS, PREC-Concept—the proportion of similar utterances exclusively similar to either “green”-I or “blue”-C—was also affected by the order of biasing information.

Teams in the distractor-first condition produced similar utterances that were less aligned than similar utterances in the distractor-last condition. Looking at PREC-Blue, teams in the distractor-first condition produced significantly fewer utterances similar to “blue”-C than did teams in the distractor-last condition. In all conditions, the concept “green”-I was discussed less than “blue.” Exchanges related to the concept “blue”-C were typically disrupted by bias condition more than were those related to “green.” However, in the N100 concept extraction of the factoid-derived concepts, utterances about “green”-I were likely to be more similar in the late than in the mixed condition.

The effect of similarity type was not observed in the factoid-derived concept analyses, meaning the analyses were unable to differentiate the average similarity of utterances made by the same person and similar utterances produced by different individuals. This suggests that utterances focusing on these concepts were more homogeneous across individuals than the average utterance.

Discussion

The overarching goal of this article was to test the utility of CRA, performed within the Discursis framework and supplemented with novel metrics that we developed, as a method for analyzing semantic alignment in interpersonal communication. Specifically, CRA has been shown to yield insight into conversations in non-experimental settings but has yet to be applied to analysis of experimental manipulations in team settings.² We calculated CRA metrics from team communications collected from a previous experiment investigating cognitive biases in team problem solving. We sought to show how CRA might be extended to test specific hypotheses about the distribution of semantic similarity across utterances as a function of the source of the utterances and relation to specific key concepts. We also sought to test two parameter settings available in Discursis, the source of key concepts and the number of key concepts. In pursuit of these goals, we introduced a novel set of recurrence- and similarity-based metrics for analyzing the semantic alignment of discourse or communication records and compared inferential analyses of these to those from simple summary measures from frequency counts of key words. In line with the findings reported by Mancuso et al. (2014), we expected there to be an influence of bias on team communications related to the key concepts “blue”-C and “green”-I. Specifically, we expected there to be a decrease

² A conference report regarding a follow-up study to the one presented in this article, by Tolston et al. (2016), temporally preceded this work.

in semantic similarity accounted for by utterances related to “blue”-C in the distractor-first condition relative to the distractor-last condition. We briefly summarize our findings below and then discuss some broader implications of CRA.

All key concepts active

Analyses of the relative frequency of the words “blue”-C and “green”-I showed that these were unaffected by the experimental manipulation, but that “blue”-C was mentioned more on average than “green”-I. With respect to CRA, when all concepts were active, the N100 analyses using intrinsic bases differentiated bias conditions. In contrast, the factoid-derived bases were insensitive to the order of biasing information, and this was true for both N100 and automatic concept extraction. This difference, combined with the fact that significantly more key concepts were identified from the intrinsic than from the factoid sources in the N100 parameter setting, indicates a high degree of task-relevant semantic richness of group communications relative to the factoid set.

When using intrinsic bases and the N100 setting, partitioning MS according to self and shared showed that presenting even a small number of biasing factoids early in the information queue (i.e., the distractor-mixed condition) was enough to disrupt communication patterns at the individual level. However, in communications between teammates, only the distractor-first condition significantly affected MS. In other words, either amount of early biasing factoids shifted the average conceptual content of utterances made by individuals, but only the larger amount of early biasing in the distractor-first condition altered the focus of responses teammates made to one another. Self-similarity of MS was greater than shared similarity of MS for both source of key concepts (intrinsic vs. factoid) and both number of concepts extracted (N100 vs. automatic).

Selected key concepts

When we modified our analyses to look only at utterances similar to either “blue”-C or “green”-I, many of the differences in sensitivity between the sizes of the bases of key concepts were eliminated. In general, similar utterances from teams in the distractor-first condition were less aligned than similar utterances in the distractor-last condition. We found that PREC-Concept—the proportion of similar utterances exclusively similar to either “green”-I or “blue”-C—was impacted by bias condition. Teams in the distractor-first condition produced fewer utterances similar to “blue”-C than teams in the distractor-last condition. Regardless of the parameters chosen, results showed that the concept “green”-I was discussed less than “blue”-C. Utterances with content similar to the concept “blue”-C were affected by bias condition more than those similar to “green”-I. This partly confirmed our hypothesis and is consistent with the notion that teams were susceptible

to a confirmation bias and did not adequately discuss “blue”-C relevant information. The CRA metrics showed a sensitivity to the experimental manipulation that exceeded what we observed with simple frequency counts. Specifically, by taking semantic information into account, CRA metrics showed that discussions related to “blue”-C and “green”-I were affected by the distribution order of factoids, even though frequency counts of specific mentions of the words “green” and “blue” were the same between experimental conditions.

When all concepts were active, the average amount of similarity for utterances made by the same person was usually higher than the average amount of similarity of utterances made by separate people, meaning that individuals in general displayed some degree of idiosyncratic language use. Interestingly, the effect of similarity type was not seen in the factoid-derived concepts when only “blue”-C or “green”-I were active, meaning that semantic idiosyncrasies were diminished within these subsets of utterances on the coarsest basis. This may mean that utterances that mentioned different terrorist groups may have focused on a smaller number of key concepts than did other utterances.

In addition, when all concepts were active, sensitivity of CRA metrics to the experimental manipulation depended on the number of extracted key concepts: The N100 concept extraction proved more sensitive to the bias manipulation than automatic concept extraction. However, when looking at the subset of concepts “green”-I and “blue”-C, the number of concepts used to construct semantic spaces made less of a difference. This suggests utterances about who was planning the fictional attack were more tightly focused on a smaller number of key concepts than utterances in general.

Utterances in the distractor-first condition exhibited a lower degree of MS, on average, than utterances in the distractor-last condition. MS measures the degree to which utterances contain similar mixtures of key concepts; when MS is low, this may indicate that utterances tend to be less focused on particular subsets of concepts, but lower MS could also occur when a larger number of concepts are active, as the chance for utterances to be indirectly related increases. It could be argued that such circumstances were induced in the distractor-first condition if more topics were competing for discussion early on in the experiment. However, the fact that factoids that were related to “blue”-C were introduced later in the queue in the distractor-first condition means that the rate of topic introduction was approximately balanced between the distractor-first and distractor-last conditions. In light of this, we believe these results show that teams in the distractor-first condition were less likely to converge on a common set of concepts, perhaps because of the thematically disjointed nature of the factoids implicating the “green”-I team and the less concentrated distribution of “blue”-C information. We find it interesting to note that noisy information profiles are conceptually similar to encoding errors that are thought to be at the heart of several

cognitive biases (Hilbert, 2012). Our findings also revealed a lower prevalence of utterances similar to the key concept “blue” in the distractor-first condition, providing evidence for a confirmation bias. We believe that using techniques like CRA to explore sources and mechanisms of cognitive bias has promising implications.

One of the most reliable findings was that similar utterances that were made by the same individual were more aligned than similar utterances that were made by different individuals. This effect was seen in all analyses, except when using key concepts derived from the factoid sets and analyzing similarity in a very restricted subset of concepts (e.g., “blue”-C or “green”-I). It remains an open question whether this effect will hold in general or if it is contextually driven by individuals sharing prespecified pieces of information. Regardless, the ability to easily partition semantic similarity according to sources (e.g., self or shared) promises to provide insight into situations in which it is desirable to evaluate the relative convergence or divergence of semantic patterns (e.g., Mills, 2014). For instance, the difference between self and other similarity might provide insight into the relative degree of conceptual alignment between individuals that can complement measures of morphological or lexical alignment (e.g., Fusaroli & Tylén, 2016). Furthermore, contrasting differences in conceptual alignment between self and other over the course of a conversational interaction has implications for evaluating the interactive alignment model (Garrod & Pickering, 2004).

In addition to partitioning semantic similarity by type, we showed that partitioning similar utterances by term and determining the proportion of all utterances that such subsets account for is an effective method to evaluate the relative importance of particular key concepts. Early introduction of biasing information reduced the extent to which teams were conceptually coordinated regarding the blue group (the correct answer), and this parallels the significant decline in performance for the distractor-first condition reported previously (Mancuso et al., 2014). By adding together filters for more than one key concept, the technique can be expanded to test hypotheses about the relative importance of groups of key concepts. The statistical nature of the conceptual similarity algorithm of this method is likely to offer more nuanced outcomes than what would be obtained by frequency counts alone. In fact, the present analysis provides insight into the reexamined data set that goes beyond what the authors were able to infer in their initial publication (Mancuso et al., 2014). In their article, they discussed how the biasing information may have served as a hidden knowledge profile (e.g., Stasser & Titus, 1985) causing a confirmation bias in which teams focused on earlier noncritical distractor information and discounted new information that did not conform to their established viewpoint.

Generally speaking, hidden knowledge is suggested to create noise, and this is consistent with our findings of decreased conceptual coherence in the distractor-first condition. In addition, as our new findings show (see Fig. 6), the distractor factoids created a systematic bias that permeated through the group and their communicative interactions and resulted in a decrease in the discussion of the relevant “blue” information.

We see many potential future applications and extensions of these methods and metrics. For instance, the metrics outlined in this article may be of use in evaluating pathological speech, where dense and oblique relationships (Cabana, Valle-Lisboa, Elvevåg, & Mizraji, 2011; Mota et al., 2012) between concepts in utterances indicate a semantic space that might differentiate under refinement in ways substantially different from spaces constructed from the speech of typical individuals. This suggests that the techniques presented in this article could be modified to produce a conceptual analogue of a scaling exponent (Kello et al., 2010) by evaluating the change in the degree of alignment between utterances as a function of the number of concepts in the basis. Such a measure might index what has been identified as the nested complexity of natural language (Hodges & Fowler, 2015). Future directions also include development of online analysis techniques and algorithms that may allow a continuous evaluation of communication in near-real-time.

Another potential application is use of the CRA metrics presented in this article to create graphical models to evaluate potential causal pathways between levels of interpersonal coordination (Runge, Petoukhov, Donges, Hlinka, Jajcay, Vejmelka, ... Kurths, 2015; Tolston et al., 2017), such as postural alignment (e.g., Shockley, Santana, & Fowler, 2003), speech alignment of speaking rate (e.g., Shockley et al., 2004), and lexical alignment (e.g., Dale & Spivey, 2006). We believe the techniques presented in this article provide a means for integrating semantic alignment into models of interpersonal coordination to build upon previous related work using network approaches (see Paxton et al., 2016).

It is important to mention several limitations of the present study. First, we acknowledge that the relatively small sample is a possible limitation. However, we believe this concern is allayed, to some degree, by results that are consistent with both prior research and our hypotheses and by the observed effect sizes. Despite this potential limitation, we believe our results successfully demonstrate both the experimental utility of the proposed novel CRA metrics and how they may be affected by parameter choices. We also note that CRA is a relatively new tool, and some of the metrics derived from it require further research before they can be clearly interpreted (though some work along these lines has been recently conducted; e.g., Atay et al., 2015). Furthermore, a variety of parameter values must be set during data analysis with CRA, including the number of key terms and the source

of key terms. Although our results were robust against small parameter changes, it is possible that different parameter settings could lead to different outcomes. Creating objective criteria or even “rules of thumb” for parameter choices similar to those used in RQA (Marwan, Romano, Thiel, & Kurths, 2007) is a potentially fruitful area for future research. Additionally, the sheer number of possible metrics is a double-edged sword—there are many ways to characterize discourse, but as more measures are deployed the volume of data becomes unwieldy, and the risk of making Type I errors increases. The careful choice of the minimal set of meaningful measures prior to analysis is, as always, a requirement of judicious research.

Conclusions

This project was the first to utilize CRA to quantify conceptual coordination in distributed team communication during team decision making. We applied a new quantitative framework to re-examine and extend prior research on the influence of biasing information on team communications using an automated technique. Our results show that CRA was sensitive to the experimental manipulations. We demonstrated how novel metrics derived from the similarity matrix supplied by the Discursis software package might be calculated and have shown the utility of several of these that we believe to be among the most useful. We conclude by listing some suggestions for conducting CRA in Discursis on the basis of the results of our evaluation of the effects of varying two CRA parameters, the number of key concepts, and the source of key concepts.

Based on our results, we advocate caution against using bases including only a small number of key concepts when conducting CRA, though we did also find evidence that using smaller bases might be acceptable if particular *a priori* concepts are being evaluated. Measures obtained in the N100 analysis generally proved more sensitive to the experimental manipulation, suggesting that the smaller conceptual bases did not capture as much of the information contained in the exchanges as the larger bases. We note, however, that increasing dimensionality of semantic spaces must be limited as well, as they can be expanded up to the point of essentially fitting noise (Quesada, 2007).

With respect to the source of the semantic spaces, if no *a priori* concepts are critical for the analysis of data, it might be better to analyze the data using intrinsically defined key concepts. However, if key concepts are of interest, *a priori* conceptual bases may offer similar sensitivity. This is not a surprising result and is consistent with the conditions in which an *a priori* basis might be used (e.g., to determine the relative influence of particular key concepts). This has implications for studies that might use external training corpora, though we note the data set we used to identify key concepts was relatively small (i.e., the 64 factoids distributed to team

members in the ELICIT task). We also note that the *a priori* set we used consisted of prewritten statements from a logic puzzle, and that, under general circumstances, creating *a priori* bases from a semantically richer database might give better results. That *a priori* bases provide sensitive measures is a promising finding for the continued development of methods for real-time analysis of data, given that having *a priori* bases eliminates some time-consuming data-processing steps. In particular, it might prove useful for online analysis of communications for the presence or absence of structure in communications related to task-critical key concepts.

Overall, we believe that the results of our evaluation of CRA suggest it may be a useful additional lens through which to investigate communication, and which provide information that complements other methods, such as frequency counts. Of added benefit, CRA needs minimal data preprocessing, and does not rely on extensive amounts of costly human coding, further increasing its attractiveness.

Author note This research was funded by grant no. 15RHCOR233 from the Air Force Office of Scientific Research (Benjamin Knott, Program Officer).

References

- Angus, D., Smith, A., & Wiles, J. (2012b). Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics*, 18, 988–997. <https://doi.org/10.1109/tvcg.2011.100>
- Angus, D., Smith, A. E., & Wiles, J. (2012a). Human communication as coupled time series: Quantifying multi-participant recurrence. *IEEE Transactions on Audio, Speech, and Language Processing*, 20, 1795–1807. <https://doi.org/10.1109/tasl.2012.2189566>
- 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, e38014. <https://doi.org/10.1371/journal.pone.0038014>
- Atay, C., Conway, E. R., Angus, D., Wiles, J., Baker, R., & Chenery, H. J. (2015). An automated approach to examining conversational dynamics between people with dementia and their carers. *PLoS ONE*, 10, e0144327. <https://doi.org/10.1371/journal.pone.0144327>
- Babcock, M. J., Ta, V. P., & Ickes, W. (2014). Latent semantic similarity and language style matching in initial dyadic interactions. *Journal of Language and Social Psychology*, 33, 78–88. <https://doi.org/10.1037/e578192014-548>
- Broniatowski, D. A., & Magee, C. L. (2012). Studying group behaviors: A tutorial on text and network analysis methods. *IEEE Signal Processing Magazine*, 29, 22–32. <https://doi.org/10.1109/msp.2011.942680>
- Bunnin, N., & Yu, J. (2004). *The Blackwell dictionary of Western philosophy*. Malden: Blackwell.
- Cabana, Á., Valle-Lisboa, J. C., Elvevåg, B., & Mizraji, E. (2011). Detecting order-disorder transitions in discourse: Implications for schizophrenia. *Schizophrenia Research*, 131, 157–164. <https://doi.org/10.1016/j.schres.2011.04.026>

- Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23, 13–31.
- Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511620539>
- Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: An R package. *Frontiers in Quantitative Psychology and Measurement*, 5, 510.
- Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, 37, pp. 255–285. <https://doi.org/10.1111/cogs.12009>
- Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 83–106). Washington, DC: American Psychological Association. <https://doi.org/10.1037/10690-005>
- Dale, R., Duran, N. D., & Coco, M. (2018). Dynamic natural language processing with recurrence quantification analysis. arXiv Preprint arXiv:1803.07136.
- Dale, R., & Spivey, M. J. (2005). Categorical recurrence analysis of child language. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 530–535). Mahwah: Lawrence Erlbaum.
- Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56, 391–430. <https://doi.org/10.1111/j.1467-9922.2006.00372.x>
- Entin, E. E., & Entin, E. B. (2001). Measures for evaluation of team processes and performance in experiments and exercises. Paper presented at the 6th International Command and Control Research and Technology Symposium, Annapolis. Retrieved from http://dx.dodccrp.org/events/6th_ICCRTS/Tracks/Papers/Track6/090_tr6.pdf
- Finomore, V., Sitz, A., Blair, E., Rahill, K., Champion, M., Funke, G., ... Knott, B. (2013). Effects of cyber disruption in a distributed team decision making task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, pp. 394–398). <https://doi.org/10.1177/1541931213571085>
- Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, 28, 197–202. <https://doi.org/10.3758/bf03204765>
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. Paper presented at the ED-MEDIA 99 World Conference on Educational Multimedia, Hypermedia & Telecommunications, Seattle. Retrieved from www-psycho.nmsu.edu/~pfoltz/reprints/Edmedia99.html
- Foltz, P. W., & Martin, M. J. (2009). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 411–432). New York: Taylor & Francis.
- 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, 937–939. <https://doi.org/10.1177/0956797612436816>
- Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, 40, 145–171. <https://doi.org/10.1111/cogs.12251>
- Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8, 8–11. <https://doi.org/10.1016/j.tics.2003.10.016>
- Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in R with annotated code: Studying online reviews and the stock exchange community. *Communications of the Association for Information Systems*, 41, 450–496. <https://doi.org/10.17705/1cais.04121>
- Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. J., & Cooke, N. J. (2003). Evaluation of Latent Semantic Analysis-based measures of team communications content. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47, 424–428. <https://doi.org/10.1037/e577042012-037>
- Gorman, J. C., Martin, M. J., Dunbar, T. A., Stevens, R. H., Galloway, T., Amazeen, P. G., & Likens, A. D. (2016). Cross-level effects between neurophysiology and communication during team training. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(1), 181–199. <https://doi.org/10.1177/0018720815602575>
- Healey, P. G., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PLoS ONE*, 9, e98598. <https://doi.org/10.1371/journal.pone.0098598>
- Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, 138, 211. <https://doi.org/10.1037/a0025940>
- Hodges, B. H., & Fowler, C. A. (2015). Fields, waves, and particles: Finding common ground in understanding language as a public activity. *Ecological Psychology*, 27, 175–189. <https://doi.org/10.1080/10407413.2015.1068647>
- Iliev, R., Dehghani, M., & Sagi, E. (2014). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, 7, 1–26. <https://doi.org/10.1017/langcog.2014.30>
- Indulska, M., Hovorka, D. S., & Recker, J. (2012). Quantitative approaches to content analysis: Identifying conceptual drift across publication outlets. *European Journal of Information Systems*, 21, 49–69. <https://doi.org/10.1057/ejis.2011.37>
- Kello, C. T., Brown, G. D. A., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14, 223–232. <https://doi.org/10.1016/j.tics.2010.02.005>
- Luo, H., Liu, Z., Luan, H., & Sun, M. (2015). Online learning of interpretable word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1687–1692). 10.18653/v1/d15-1196
- Mancuso, V. F., Finomore, V. S., Rahill, K. M., Blair, E. A., & Funke, G. J. (2014). Effects of cognitive biases on distributed team decision making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58, 405–409. <https://doi.org/10.1177/1541931214581083>
- Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5), 237–329. <https://doi.org/10.1016/j.physrep.2006.11.001>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 3111–3119). New York: Curran Associates.
- Mills, G. J. (2014). Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, 32, 158–173. <https://doi.org/10.1016/j.newideapsych.2013.03.006>
- Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., ... Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE*, 7, e34928. <https://doi.org/10.1371/journal.pone.0034928>
- Orsucci, F., Giuliani, A., Webber, C., Zbilut, J., Fonagy, P., & Mazza, M. (2006). Combinatorics and synchronization in natural semiotics. *Physica A*, 361, 665–676. <https://doi.org/10.1016/j.physa.2005.06.044>
- Paice, C. D. (1996). Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information*

- Science, 47, 632–649. [https://doi.org/10.1002/\(sici\)1097-4571\(199608\)47:8<632::aid-asi8>3.0.co;2-u](https://doi.org/10.1002/(sici)1097-4571(199608)47:8<632::aid-asi8>3.0.co;2-u)
- Park, S., Bak, J., & Oh, A. (2017). Rotated word vector representations and their interpretability. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (pp. 401–411). 10.18653/v1/d17-1041
- Paxton, A., Dale, R., & Richardson, D. C. (2016). Social coordination of verbal and nonverbal behaviours. In P. Passos, K. Davids, & J. Y. Chow (Eds.), *Interpersonal coordination and performance in social systems*. New York: Routledge.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169–190. <https://doi.org/10.1017/s0140525x04000056>
- Quesada, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 71–85). Mahwah: Erlbaum. <https://doi.org/10.4324/9780203936399.ch4>
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 46–50).
- Ruddy, M., & Nissen, M. (2008). New software platform capabilities and experimentation campaign for ELICIT. Paper presented at the 13th International Command and Control Research and Technology Symposium, Seattle. Retrieved from www.dtic.mil/dtic/tr/fulltext/u2/a487140.pdf
- Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., ... Kurths, J. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6, 8502. <https://doi.org/10.1038/ncomms9502>
- Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Boston: Addison-Wesley Longman.
- Schippers, M. C., Edmondson, A. C., & West, M. A. (2014). Team reflexivity as an antidote to team information-processing failures. *Small Group Research*, 45, 731–769. <https://doi.org/10.1177/1046496414553473>
- Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, 66, 422–429. <https://doi.org/10.3758/bf03194890>
- Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 326–332. <https://doi.org/10.1037/0096-1523.29.2.326>
- Smalheiser, N. R., & Bonifield, G. (2018). *Unsupervised low-dimensional vector representations for words, phrases and text that are transparent, scalable, and produce similarity metrics that are complementary to neural embeddings*. arXiv preprint. arXiv: 1801.01884
- Smith, A. E. (2000). Machine mapping of document collections: The leximancer system. Paper presented at the 5th Australasian Document Computing Symposium, Sunshine Coast. Retrieved from www.leximancer.com/wiki/images/5/54/Adcs2000.pdf
- Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods*, 38, 262–279. <https://doi.org/10.3758/bf03192778>
- Spevack, S. C., Falandays, J. B., Batzloff, B., & Spivey, M. J. (2018). Interactivity of language. *Language and Linguistics Compass*, e12282. <https://doi.org/10.1111/lnc3.12282>
- Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6), 1467–1478. <https://doi.org/10.1037/0022-3514.48.6.1467>
- Tolston, M., Strang, A. J., Funke, G. J., Miller, B., Brown, R., & Menke, L. (2017). Identifying physio-behavioral drivers of performance using multivariate causal networks. Paper presented at the 19th International Symposium on Aviation Psychology, Dayton.
- Tolston, M. T., Finomore, V., Funke, G. J., Mancuso, V., Brown, R., Menke, L., & Riley, M. A. (2016). Effects of biasing information on the conceptual structure of team communications. In K. S. Hale & K. M. Stanney (Eds.), *Advances in neuroergonomics and cognitive engineering* (pp. 433–445). Cham: Springer.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188. <https://doi.org/10.1613/jair.2934>
- Tylén, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C. D. (2010). Language as a tool for interacting minds. *Mind & Language*, 25, 3–29. <https://doi.org/10.1111/j.1468-0017.2009.01379.x>
- Volpe, C. E., Cannon-Bowers, J. A., Salas, E., & Spector, P. E. (1996). The impact of cross-training on team functioning: An empirical investigation. *Human Factors*, 38, 87–100. <https://doi.org/10.1518/001872096778940741>
- Watson, B. M., Angus, D., Gore, L., & Farmer, J. (2015). Communication in open disclosure conversations about adverse events in hospitals. *Language & Communication*, 41, 57–70. <https://doi.org/10.1016/j.langcom.2014.10.013>
- Weber, R. P. (1990). *Basic content analysis*. Newbury Park: Sage.
- Wildman, J. L., Salas, E., & Scott, C. P. (2013). Measuring cognition in teams: A cross-domain review. *Human Factors*, 56, 911–941. <https://doi.org/10.1177/0018720813515907>