ConversationAlign: An R package for Computing Linguistic Alignment and Corpus Analytics in Dyadic Conversation Transcripts

Jamie Reilly[1, 2], **Benjamin Sacks**[2], **Virginia Ulichney**[2], **Gus Cooney**[3], **and Chelsea Helion**[1,2]

1 Department of Communication Sciences and Disorders, Temple University, United States **2** Department of Psychology and Neuroscience, Temple University, United States **3** Wharton School, University of Pennsylvania, United States

Abstract

`ConversationAlign` is an R package that executes a series of operations upon one or more conversation transcripts (i.e., two-person dialogues). Transcripts nominally contain at least two variables (speaker identity and text).`ConversationAlign` will retain all other meta-data such as timestamps, demographics, and grouping variables. `ConversationAlign` imports raw transcripts into R, appends unique document identifiers, and concatenates all conversations into a single dataframe. `ConversationAlign` generates corpus analytics characterizing the conversation transcript(s) of interest. Users guide a number of text cleaning operations such as stopword removal and lemmatization. The package ultimately vectorizes the original text into a one-word-per-row format. `ConversationAlign` yokes published norms to each content word spanning more than 40 lexical, affective, and semantic dimensions (e.g. word length, morphological complexity, arousal, valence). `ConversationAlign` outputs summary data for each conversation including main effects and indices of local and global alignment for each specified dimension of interest.

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Statment of Need

Conversation is among the most complex social and linguistic behaviors that humans routinely undertake. In a dyadic interaction, two conversation partners modify the form and content of their own production to align with each other (Pickering & Garrod, 2021). This process, known as linguistic alignment, occurs across many dimensions (e.g., affective coloring, prosody, gesture, formality, semantic and syntactic complexity). `ConversationAlign` offers an automated approach to computing empirical indices alignment between conversation partners across one or more conversations with coverage of over 40 distinct psycholinguistic dimensions (e.g., word length, valence, concreteness, morphological complexity). Although other open-source software applications exist for text corpus analyses, such as Quanteda (Benoit et al., 2018) and Korpus (Michalke, Brown, Mirisola, Brulet, & Hauser, 2018), we know of R-specific packages that offer a comprehensive text processing pipeline capable of cleaning, formatting, transforming, and summarizing raw conversation transcripts (but for Python see ALIGN (Duran, Paxton, & Fusaroli, 2019)).

Recent advances in Natural Language Processing (NLP) and the dissemination of corpora such as CANDOR (Reece et al., 2023) are creating fertile for the study of dialogue. `ConversationAlign` is a open-source R package that leverages many of these advances as a tool for examining conversation dynamics at an unprecedented scale. `ConversationAlign`

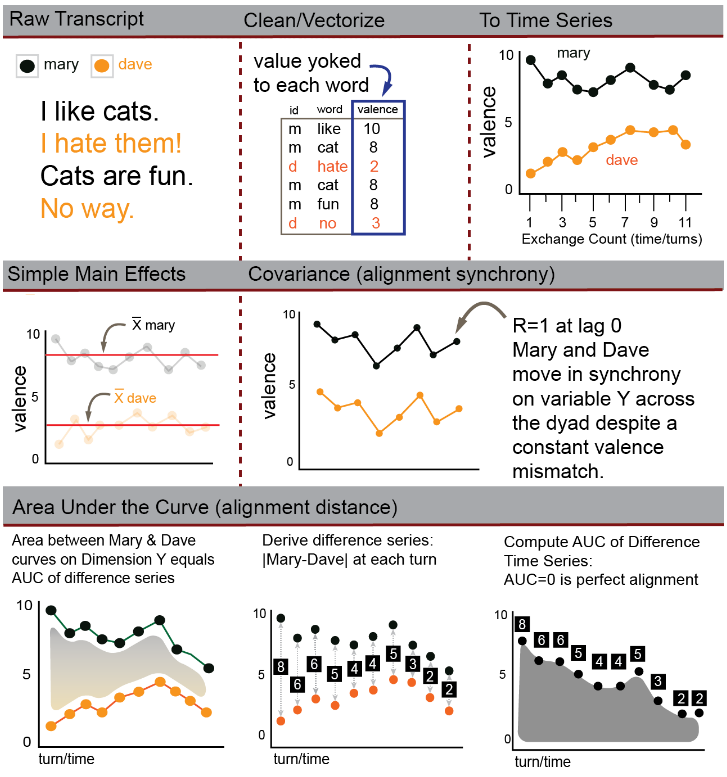
Raw Transcript
Clean/Vectorize
To Time Series
mary
dave
I like cats.
I hate them!
Cats are fun.
No way.
value yoked to each word
id word valence
m like 10
m cat 8
d hate 2
m cat 8
m fun 8
d no 3
valence
10
5
0
mary
dave
1 3 5 7 9 11
Exchange Count (time/turns)
Simple Main Effects
Covariance (alignment synchrony)
X̄ mary
X̄ dave
R=1 at lag 0 Mary and Dave move in synchrony on variable Y across the dyad despite a constant valence mismatch.
Area Under the Curve (alignment distance)
Area between Mary & Dave curves on Dimension Y equals AUC of difference series
Derive difference series: |Mary-Dave| at each turn
Compute AUC of Difference Time Series: AUC=0 is perfect alignment
8 6 6 5 4 4 5 3 2 2
turn/time

Figure 1: Overview of ConversationAlign Pipeline

is **NOT** a large language model (e.g., GPT, DeepSeek). Its algorithms do not use generative artificial intelligence and instead index an internal lexical lookup database with coverage of over 100,000 English words across more than 40 unique dimensions spanning affective (e.g., happiness, valence), semantic (e.g., concreteness, semantic density), lexical (e.g., age-of-scqusition, morphological complexity), and phonological (e.g., word length, syllable length) information. `ConversationAlign` joins content words from conversation transcripts of any length to their corresponding values in this lookup database, effectively transforming words into quantitative time series objects aggregated by speaker, turn, and conversation. Figure 1 illustrates the primary steps undertaken by `ConversationAlign` in executing these transformations.

Key Components of the ConversationAlign Pipeline

`ConversationAlign` processes dyadic (2-person) conversation transcripts (`*.csv`, `*.txt`) via a series of four customizable functions:

1. **read_dyads()**: imports one or more conversation transcripts into R, concatenating all transcripts into a single dataframe marked with its unique filename as a document identifier.
2. **prep_dyads()**: executes numerous text cleaning and formatting operations (e.g., to lowercase, expand contractions, remove special characters, squish whitespace). Options include stopword removal, stopword list specification, and lemmatization. `prep_dyads()' splits the raw text into a one word per row format then prompts the user to select up to three dimensions for computing main effects and alignment.`prep_dyads()’ returns a dataframe with values for the

variables of interest (e.g., word length, word frequency, valence) to each running content word.

3. **summarize_dyads()**: produces a summary dataframe with main effects and alignment indices for the user-specified variables of interest summarized by conversation (Event_ID) and participant (Participant_ID). Alignment indices include: a) lagged spearman R correlation values reflecting turn-by-turn covariance between interlocutors across each dimension of interest (e.g., Mary uses unpleasant words, Dave immediately responds with unpleasant words); b) dAUC: global distance between partners by conversation across each variable of interest (e.g., 'pleasantness' distance between Dave and Mary across all turns). `summarize_dyads()` produces raw AUC and AUC normalized to a fixed conversation length (i.e., 50 exchanges, 100 turns) to promote standardization/comparison across different conversation durations.
4. **corpus_analytics()**: produces text analytics and descriptive statistics for your conversation corpus, including total number of tokens, average number of turns per conversation, average number of words-per-turn by conversation, average word length (letter count) by conversation, type token ratio by converation (for comprehensive list see package documentation). Summary dataframe readily exportable to a table for journal submission.

`ConversationAlign's` core algorithm(s) transform raw text data into numeric time series yoked to specific dimensions of interest selected by the user. This approach moves beyond simple comparisons of means to potentially capture causal relationships (e.g., John mirrors increases in Mary's word length, but Mary does not recirpocate this pattern). The alignment indices generated by `ConversationAlign` measure both local (e.g., turn-by-turn) and global (AUC) distance between conversation partners across each dimension of interest.

Uses of `ConversationAlign`

We anticipate that `ConversationAlign` will have many theoretical and applied applications for measuring and modeling conversation dynamics. Example applications include: - Assesing alignment dynamics between conversation partners across individual difference factors (e.g., age, culture, education level, socio-economic status). - Asessing pre/post changes in naturalistic language use as a function of a specific intervention (e.g., metacognitive training for traumatic brain injury). - Measuring alignment dynamics between friends (and rivals) to elucidate semantic, affective, and lexical dynamics that mark 'good' conversations. - Examining alignment (and misalignment) between people with neurological disorders and their significant others (spouses, friends, children) to improve the quality of communication and reduce the prevalence of communication breakdown. - Syncrhonizing language with physiological data (e.g., biosignals) to examine real-time coupling between interacting people and brains.

Concluding Remarks

We look forward to supporting the user community in their efforts to better understand, measure, and model communication using `ConversationAlign`.

Acknowledgements

We are grateful to Sarah Weinstein, Tania Giovannetti, and Anna Duncan for conceptual and methodological advice.

This work was supported by US Public Health Service grant R01 DC013063 (JR) from the National Institute on Deafness and Communicative Disorders (NIH/NIDCD).

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. doi:10.21105/joss.00774

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques—a python library. *Psychological Methods*, *24*(4), 419–438. doi:10.1037/met0000206

Michalke, M., Brown, E., Mirisola, A., Brulet, A., & Hauser, L. (2018, October). koRpus: An r package for text analysis. Retrieved from https://CRAN.R-project.org/package=koRpus

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=3RgXEAAAQBAJ&oi=fnd&pg=PR7&dq=pickering+garrod+understanding+dialogue&ots=0qe68OV8Xs&sig=ulM_ibE3lLJewbmVg-UvQvfEHhM

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., et al. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, *9*(13), eadf3197. doi:10.1126/sciadv.adf3197