The Journal of Open Source Software

ConversationAlign: 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

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

`ConversationAlign` executes a series of operations upon one or more conversation transcripts (i.e., two-person dialogues). `ConversationAlign` imports raw language transcripts into R, appends unique document identifiers, and concatenates all conversations into a single dataframe. `ConversationAlign` generates corpus analytics and executes text cleaning operations such as stopword removal and lemmatization. The package vectorizes text and yokes published norms to each content word spanning more than 40 lexical, affective, and semantic dimensions. `ConversationAlign` outputs summary statistics each conversation including main effects and indices of local and global alignment for each specified dimension of interest.

Statment of Need

Many excellent text analysis applications exist (e.g., `Quanteda` (Benoit et al. 2018) and `Korpus` (Michalke et al. 2018)). However, few applications are tailored to the unique demands of conversation analysis (but for Python see ALIGN (Duran, Paxton, and Fusaroli 2019)). `ConversationAlign` offers a comprehensive text processing pipeline and novel algoirthms for computing linguistic alignment in 2-person dialogues. This software offers standardization and automation advantages that are in great need in a field that has historically relied heavily upon manual coding systems and subjective human judgment.

State of the Field

The fundamental challenges involved in conversation analysis involve not only measuring individual behavior but also characterizing the dynamics of alignment between two or more partners. Conversation analysis has historically been undertaken by a variety of fields (e.g., linguistics, psychology). Recent advances in 2-person neuroscience (e.g., hyperscanning) and natural language processing (NLP) (e.g., time series analysis) are driving development of new methods for modeling human interaction and understanding how people synchronize language and brain activity. `ConversationAlign` is poised to make important contributions to measuring and modeling alignment and integrating physiological with linguistic processes as simultaneous time series data.

Software Design

`ConversationAlign` was designed as a user-friendly R package with no proprietary components or input from artificial intelligence (e.g., large language models). Our goal was to make the software accessible to users who do not have extensive backgrounds in computational linguistics or Natural Language Processing.

Research Impact Statement

`ConversationAlign` has supported two peer-reviewed publications to date in the cognitive neuroscience and psychological methods journals `Cortex` (Reilly et al. 2025) and `Behavior Research Methods` (article in press). The software is relatively new. Reliable use metrics are not yet available.

Background

Conversation is among the most complex behaviors that humans routinely undertake. In a dyadic interaction, conversation partners modify the form and content of their own production to align with each other (Pickering and Garrod 2021). This process, known as linguistic alignment, occurs across many dimensions. `ConversationAlign` offers an automated approach to computing linguistic alignment across >40 distinct psycholinguistic dimensions (e.g., word length, valence, concreteness), leveraging recent advances in natural language processing to examine dynamics of human interaction at an unprecedented scale.

AI Usage

`ConversationAlign` is **NOT** a large language model (LLM). It instead indexes an static lexical lookup database populated with published norms for >100,000 English words across more than 40 unique dimensions. We used DeepSeek to troubleshoot elements of code and to generate regular expressions (regex) for complex pattern matching. We did not use AI to write this paper or generate segments of code.

References

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An r Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Duran, Nicholas D., Alexandra Paxton, and Riccardo Fusaroli. 2019. "ALIGN: Analyzing Linguistic Interactions with Generalizable techNiques—a Python Library." *Psychological Methods* 24 (4): 419–38. https://doi.org/10.1037/met0000206.

Michalke, Meik, Earl Brown, Alberto Mirisola, Alexandre Brulet, and Laura Hauser. 2018. "koRpus: An r Package for Text Analysis." https://CRAN.R-project.org/package=koRpus.

Pickering, Martin J., and Simon Garrod. 2021. *Understanding Dialogue: Language Use and Social Interaction.* Cambridge University Press. 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.

Reilly, Jamie, Virginia Ulichney, Benjamin Sacks, Anna Duncan, Sarah M. Weinstein, Tania Giovannetti, Chelsea Helion, and Gus Cooney. 2025. "Abstract Word Dropout and Cross-Speaker Misalignment of Word Concreteness Are Features of Conversation in Aging." *Cortex* 190 (September): 286–303. https://doi.org/10.1016/j.cortex.2025.07.003.