

# <sup>1</sup> ConversationAlign: Computing Linguistic Alignment and Corpus Analytics in Dyadic Conversation Transcripts

<sup>4</sup> **Jamie Reilly**  <sup>1,2¶</sup>, **Benjamin Sacks**  <sup>2\*</sup>, **Virginia Ulichney**  <sup>2</sup>, **Gus Cooney**  <sup>3</sup>, and **Chelsea Helion**  <sup>1,2</sup>

<sup>6</sup> **1** Department of Communication Sciences and Disorders, Temple University, United States **2**

<sup>7</sup> Department of Psychology and Neuroscience, Temple University, United States **3** Wharton School,

<sup>8</sup> University of Pennsylvania, United States ¶ Corresponding author \* These authors contributed equally.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 09 July 2025

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## <sup>9</sup> Summary

<sup>10</sup> ConversationAlign executes a series of operations upon one or more conversation transcripts (i.e., two-person dialogues). ConversationAlign imports raw language transcripts into R, <sup>11</sup> appends unique document identifiers, and concatenates all conversations into a single dataframe. <sup>12</sup> ConversationAlign generates corpus analytics and executes text cleaning operations such as stopword removal and lemmatization. The package vectorizes text and yokes published <sup>13</sup> norms to each content word spanning more than 40 lexical, affective, and semantic dimensions. <sup>14</sup> ConversationAlign outputs summary statistics each conversation including main effects and <sup>15</sup> indices of local and global alignment for each specified dimension of interest. <sup>16</sup>

## <sup>17</sup> Statement of Need

<sup>19</sup> Many excellent text analysis applications exist (e.g., Quanteda ([Benoit et al., 2018](#)) and Korpus <sup>20</sup> ([Michalke, Brown, Mirisola, Brulet, & Hauser, 2018](#))). However, few applications are tailored <sup>21</sup> to the unique demands of conversation analysis (but for Python see ALIGN ([Duran, Paxton, & Fusaroli, 2019](#))). ConversationAlign offers a comprehensive text processing pipeline and <sup>22</sup> novel algorithms for computing linguistic alignment in 2-person dialogues. This software offers <sup>23</sup> standardization and automation advantages that are in great need in a field that has historically <sup>24</sup> relied heavily upon manual coding systems and subjective human judgment. <sup>25</sup>

## <sup>26</sup> State of the Field

<sup>27</sup> The fundamental challenges involved in conversation analysis involve not only measuring <sup>28</sup> individual behavior but also characterizing the dynamics of alignment between two or more <sup>29</sup> partners. Conversation analysis has historically been undertaken by a variety of fields (e.g., <sup>30</sup> linguistics, psychology). Recent advances in 2-person neuroscience (e.g., hyperscanning) and <sup>31</sup> natural language processing (NLP) (e.g., time series analysis) are driving development of new <sup>32</sup> methods for modeling human interaction and understanding how people synchronize language <sup>33</sup> and brain activity. ConversationAlign is poised to make important contributions to measuring <sup>34</sup> and modeling alignment and integrating physiological with linguistic processes as simultaneous <sup>35</sup> time series data.

## <sup>36</sup> Software Design

<sup>37</sup> ConversationAlign was designed as a user-friendly R package with no proprietary components  
<sup>38</sup> or input from artificial intelligence (e.g., large language models). Our goal was to make the  
<sup>39</sup> software accessible to users who do not have extensive backgrounds in computational linguistics  
<sup>40</sup> or Natural Language Processing.

## <sup>41</sup> Research Impact Statement

<sup>42</sup> ConversationAlign has supported two peer-reviewed publications to date in the cognitive  
<sup>43</sup> neuroscience and psychological methods journals *Cortex* ([Reilly et al., 2025](#)) and *Behavior*  
<sup>44</sup> *Research Methods* (article in press). The software is relatively new. Reliable use metrics are  
<sup>45</sup> not yet available.

## <sup>46</sup> Background

<sup>47</sup> Conversation is among the most complex behaviors that humans routinely undertake. In a  
<sup>48</sup> dyadic interaction, conversation partners modify the form and content of their own production  
<sup>49</sup> to align with each other ([Pickering & Garrod, 2021](#)). This process, known as linguistic  
<sup>50</sup> alignment, occurs across many dimensions. ConversationAlign offers an automated approach  
<sup>51</sup> to computing linguistic alignment across >40 distinct psycholinguistic dimensions (e.g., word  
<sup>52</sup> length, valence, concreteness), leveraging recent advances in natural language processing to  
<sup>53</sup> examine dynamics of human interaction at an unprecedented scale.

## <sup>54</sup> AI Usage

<sup>55</sup> ConversationAlign is **NOT** a large language model (LLM). It instead indexes a static lexical  
<sup>56</sup> lookup database populated with published norms for >100,000 English words across more than  
<sup>57</sup> 40 unique dimensions. We used DeepSeek to troubleshoot elements of code and to generate  
<sup>58</sup> regular expressions (regex) for complex pattern matching. We did not use AI to write this  
<sup>59</sup> paper or generate segments of code.

## <sup>60</sup> References

- <sup>61</sup> Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018).  
<sup>62</sup> Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open*  
<sup>63</sup> *Source Software*, 3(30), 774. doi:[10.21105/joss.00774](https://doi.org/10.21105/joss.00774)
- <sup>64</sup> Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions  
<sup>65</sup> with generalizable techNiques—a python library. *Psychological Methods*, 24(4), 419–438.  
<sup>66</sup> doi:[10.1037/met0000206](https://doi.org/10.1037/met0000206)
- <sup>67</sup> Michalke, M., Brown, E., Mirisola, A., Brulet, A., & Hauser, L. (2018, October). koRpus: An  
<sup>68</sup> r package for text analysis. Retrieved from <https://CRAN.R-project.org/package=koRpus>
- <sup>69</sup> Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social*  
<sup>70</sup> *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=uIM\\_ibE3ILJewbmVg-UvQvfEHhM](https://books.google.com/books?hl=en&lr=&id=3RgXEAAAQBAJ&oi=fnd&pg=PR7&dq=pickering+garrod+understanding+dialogue&ots=0qe68OV8Xs&sig=uIM_ibE3ILJewbmVg-UvQvfEHhM)
- <sup>73</sup> Reilly, J., Ulichney, V., Sacks, B., Duncan, A., Weinstein, S. M., Giovannetti, T., Helion, C., et  
<sup>74</sup> al. (2025). Abstract word dropout and cross-speaker misalignment of word concreteness are  
<sup>75</sup> features of conversation in aging. *Cortex*, 190, 286–303. doi:[10.1016/j.cortex.2025.07.003](https://doi.org/10.1016/j.cortex.2025.07.003)