

# Package ‘ConversationAlign’

November 13, 2025

**Type** Package

**Title** Process Text and Compute Linguistic Alignment in Conversation Transcripts

**Version** 0.4.0

**Maintainer** Jamie Reilly <jamie\_reilly@temple.edu>

**Description** Imports conversation transcripts into R, concatenates them into a single dataframe appending event identifiers, cleans and formats the text, then yokes user-specified psycholinguistic database values to each word. 'ConversationAlign' then computes alignment indices between two interlocutors across each transcript for >40 possible semantic, lexical, and affective dimensions. In addition to alignment, 'ConversationAlign' also produces a table of analytics (e.g., token count, type-token-ratio) in a summary table describing your particular text corpus.

**License** LGPL (>= 3)

**Encoding** UTF-8

**Depends** R (>= 3.5)

**Imports** DescTools, dplyr (>= 0.4.3), httr, magrittr, purrr, rlang, stringi, stringr, textstem, tibble, tidyr, tidyselect, stats, utils, YRmisc, zoo

**Suggests** devtools, knitr, rmarkdown, testthat (>= 3.0.0)

**URL** <https://github.com/Reilly-ConceptsCognitionLab/ConversationAlign>

**RoxygenNote** 7.3.3

**LazyData** true

**VignetteBuilder** knitr

**Collate** 'ConversationAlign-package.R' 'compute\_auc.R'  
'compute\_lagcorr.R' 'corpus\_analytics.R' 'data.R'  
'generate\_shams.R' 'globals.R' 'prep\_dyads.R' 'read\_1file.R'  
'read\_dyads.R' 'replacements\_25.R' 'summarize\_dyads.R'  
'utils.R' 'zzz.R'

**Config/testthat/edition** 3

**NeedsCompilation** no

**Author** Jamie Reilly [aut, cre] (ORCID:  
    <<https://orcid.org/0000-0002-0891-438X>>),  
    Virginia Ulichney [aut],  
    Ben Sacks [aut],  
    Sarah Weinstein [ctb],  
    Chelsea Helion [ctb],  
    Gus Cooney [ctb]

**Repository** CRAN  
**Date/Publication** 2025-11-12 23:30:31 UTC

Contents

corpus_analytics . . . . .	2
generate_shams . . . . .	3
load_github_data . . . . .	3
MaronGross_2013 . . . . .	4
NurseryRhymes . . . . .	4
NurseryRhymes_Prepped . . . . .	5
prep_dyads . . . . .	5
read_1file . . . . .	6
read_dyads . . . . .	7
summarize_dyads . . . . .	7
<b>Index</b>	<b>9</b>

---

corpus_analytics	<i>corpus_analytics</i>
------------------	-------------------------

---

Description

Produces a table of corpus analytics including numbers of complete observations at each step, word counts, lexical diversity (e.g., TTR), stopword ratios, etc. Granularity of the summary statistics are guided by the user (e.g., by conversation, by conversation and speaker, collapsed all)

Usage

```
corpus_analytics(dat_prep)
```

Arguments

dat\_prep            takes dataframe produced from the df\_prep() function

Value

dataframe with summary statistics (mean, SD, range) for numerous corpus analytics (e.g., token count, type-token-ratio, word-count-per-turn) for the target conversation corpus. Summary data structured in table format for easy export to a journal method section.

---

generate_shams	<i>generate_shams</i>
----------------	-----------------------

---

**Description**

Generates a permutation of each individual dyad. Shuffled dyads may act as controls to their originals.

**Usage**

```
generate_shams(df_prep, seed = NULL)
```

**Arguments**

df_prep	Output dataframe of prep_dyads().
seed	(Optional) a seed for reproducibility in random sampling

**Value**

A dataframe similar to prepped dyads, with each participant's time series randomly shuffled.

---

load_github_data	<i>Load all .rda files from a GitHub data folder into the package environment</i>
------------------	---

---

**Description**

Load all .rda files from a GitHub data folder into the package environment

**Usage**

```
load_github_data(
  repo = "Reilly-ConceptsCognitionLab/ConversationAlign_Data",
  branch = "main",
  data_folder = "data",
  envir = parent.frame()
)
```

**Arguments**

repo	GitHub repository (e.g., "username/repo")
branch	Branch name (default: "main")
data_folder	Remote folder containing .rda files (default: "data/")
envir	Environment to load into (default: package namespace)

**Value**

nothing, loads data (as rda files) from github repository needed for other package functions

---

MaronGross_2013	<i>Sample Dyadic Interview Transcript: Marc Maron and Terry Gross Radio Interview 2013</i>
-----------------	--

---

**Description**

Text and talker information delineated, raw transcript, multiple lines per talker

**Usage**

MaronGross\_2013

**Format**

## "MaronGross\_2013" A data.frame with 546 obs, 2 vars:

**text** text from interview

**speaker** speaker identity ...

---

NurseryRhymes	<i>Text and talker information delineated, 3 separate nursery rhymes, good for computing analytics and word counts</i>
---------------	--

---

**Description**

Text and talker information delineated, 3 separate nursery rhymes, good for computing analytics and word counts

**Usage**

NurseryRhymes

**Format**

## "NurseryRhymes" A data.frame with 100 observations, 2 vars:

**Event\_ID** factor 3 different simulated conversations

**Participant\_ID** fictional speaker names, 2 each conversation

**Text\_Raw** simulated language production, actually looped phrases from nursery rhymes ...

---

NurseryRhymes\_Prepped *Text and talker information delineated, 3 separate nursery rhymes, good for computing analytics and word counts*

---

### Description

Text and talker information delineated, 3 separate nursery rhymes, good for computing analytics and word counts

### Usage

NurseryRhymes\_Prepped

### Format

## "NurseryRhymes\_Prepped" A data.frame with 1507 x 7 observations, 5 vars:

**Event\_ID** factor 3 different simulated conversations

**Participant\_ID** fictional speaker names, 2 each conversation

**Exchange\_Count** sequential numbering of exchanges by conversation, 1 exchange = 2 turns

**Turn\_Count** sequential numbering of turns by conversation

**Text\_Clean** content words

**emo\_anger** raw value of anger salience yoked to each word ...

---

prep_dyads	<i>prep_dyads</i>
------------	-------------------

---

### Description

Cleans, vectorizes and appends lexical norms to all content words in a language corpus. User guides options for stopword removal and lemmatization. User selects up to three psycholinguistic dimensions to yoke norms on each content word in the original conversation transcript.

### Usage

```
prep_dyads(
  dat_read,
  lemmatize = TRUE,
  omit_stops = TRUE,
  which_stoplist = "Temple_stops25",
  remove_backchannel = FALSE,
  verbose = TRUE
)
```

**Arguments**

<code>dat_read</code>	dataframe produced from <code>read_dyads()</code> function
<code>lemmatize</code>	logical, should words be lemmatized (switched to base morphological form), default is TRUE
<code>omit_stops</code>	option to remove stopwords, default TRUE
<code>which_stoplist</code>	user-specified stopwords removal method with options including "none", "SMART", "MIT_stops", "CA_OriginalStops", or "Temple_Stopwords25". "Temple_Stopwords25" is the default list
<code>remove_backchannel</code>	logical, should turns that are full of stopwords (e.g., "Uhm yeah") be preserved as NAs or removed. Removal will 'squish' the turn before and after together into one. If NAs are preserved they are later interpolated.
<code>verbose</code>	display detailed output such as error messages and progress (default is TRUE)

**Value**

dataframe with text cleaned and vectorized to a one word per-row format. Lexical norms and metadata are appended to each content word. Cleaned text appears under a new column called 'Text\_Clean'. Any selected dimensions (e.g., word length) and metadata are also appended to each word along with speaker identity, turn, and Event\_ID (conversation identifier).

---

<code>read_1file</code>	<i>read_1file</i>
-------------------------	-------------------

---

**Description**

Reads pre-formatted dyadic (2 interlocutor) conversation transcript already imported into your R environment.

**Usage**

```
read_1file(my_dat)
```

**Arguments**

<code>my_dat</code>	one conversation transcript already in the R environment
---------------------	--

**Value**

a dataframe formatted with 'Event\_ID', "Participant\_ID", "Text\_Raw" fields – ready for `clean_dyads()`

---

read\_dyads

*read\_dyads*


---

### Description

Reads pre-formatted dyadic (2 interlocutor) conversation transcripts from your machine. Transcripts must be either csv or txt format. IF you are supplying a txt file, your transcript must be formatted as an otter.ai txt file export. Your options for using csv files are more flexible. ConversationAlign minimally requires a csv file with two columns, denoting interlocutor and text. Each separate conversation transcript should be saved as a separate file. ConversationAlign will use the file names as a document ID. Within the read dyads function, set the my\_path argument as the directory path to the local folder containing your transcripts on your machine (e.g., "my\_transcripts"). Please see our github page for examples of properly formatted transcripts: <https://github.com/Reilly-ConceptsCognitionLab/ConversationAlign>

### Usage

```
read_dyads(my_path = "my_transcripts")
```

### Arguments

my\_path                      folder of conversation transcripts in csv or txt format

### Value

a dataframe where each individual conversation transcript in a user's directory has been concatenated. read\_dyads appends a unique document identifier to each conversation transcript appending its unique filename as a factor level to 'Event\_ID'.

---

summarize\_dyads

*summarize\_dyads*


---

### Description

Calculates and appends 3 measures for quantifying alignment. Appends the averaged value for each selected dimension by turn and speaker. Calculates and Spearman's rank correlation between interlocutor time series and appends by transcript. Calculates the area under the curve of the absolute difference time series between interlocutor time series. The length of the difference time series can be standardized the shortest number of exchanges present in the group using an internally defined resampling function, called with resample = TRUE. Spearman's rank correlation and area under the curve become less reliable for dyads under 30 exchanges.

**Usage**

```
summarize_dyads(  
  df_prep,  
  custom_lags = NULL,  
  sumdat_only = TRUE,  
  corr_type = "Pearson"  
)
```

**Arguments**

df_prep	produced in the align_dyads function
custom_lags	integer vector, should any lags be added in addition to -2, 0, 2
sumdat_only	default=TRUE, group and summarize data, two rows per conversation, one row for each participant, false will fill down summary statistics across all exchanges
corr_type	option for computing lagged correlations turn-by-turn covariance (default='Pearson')

**Value**

either: - a grouped dataframe with summary data aggregated by conversation (Event\_ID) and participant if sumdat\_only=T. - the original dataframe 'filled down' with summary data (e.g., AUC, turn-by-turn correlations) for each conversation if sumdat\_only=F.



# Index

## \* **datasets**

MaronGross\_2013, [4](#)

NurseryRhymes, [4](#)

NurseryRhymes\_Prepped, [5](#)

corpus\_analytics, [2](#)

generate\_shams, [3](#)

load\_github\_data, [3](#)

MaronGross\_2013, [4](#)

NurseryRhymes, [4](#)

NurseryRhymes\_Prepped, [5](#)

prep\_dyads, [5](#)

read\_1file, [6](#)

read\_dyads, [7](#)

summarize\_dyads, [7](#)