Package 'childeswordfreq'

May 8, 2026

Type Package

Title Word and Phrase Frequency Tools for CHILDES

Version 0.2.0

Description Tools for extracting word and phrase frequencies from the Child Language Data Exchange System (CHILDES)
database via the 'childesr' API. Supports type-level word counts,
token-mode searches with simple wildcard patterns and part-of-speech
filters, optional stemming, and Zipf-scaled frequencies. Provides
normalization per number of tokens or utterances, speaker-role
breakdowns, dataset summaries, and export to Excel workbooks for
reproducible child language research.
The CHILDES database is maintained at <https://talkbank.org/childes/>.

License MIT + file LICENSE

URL https://github.com/n-albudoor/childeswordfreq

BugReports https://github.com/n-albudoor/childeswordfreq/issues

Depends R (>= 4.4.0)

Imports cachem, childesr, dplyr, memoise, rappdirs, readr, rlang,
stats, tibble, tidyr, utils, writexl

Suggests testthat (>= 3.0.0), knitr, rmarkdown

VignetteBuilder knitr

Encoding UTF-8

RoxygenNote 7.3.3

Config/testthat/edition 3

NeedsCompilation no

Author Nahar Albudoor [aut, cre]

Maintainer Nahar Albudoor <n.albudoor@gmail.com>

Repository CRAN

Date/Publication 2025-11-15 22:40:09 UTC

Contents

`cwf_cache_disable` *Disable caching*

Description

Disable caching

Usage

```
cwf_cache_disable()
```

`cwf_cache_enable` *Enable on-disk caching of CHILDES queries*

Description

Enable on-disk caching of CHILDES queries

Usage

```
cwf_cache_enable(cache_dir = NULL)
```

Arguments

`cache_dir`	Directory for cached results; defaults to user cache dir.

`cwf_cache_enabled` *Return TRUE if caching is enabled*

Description

Return TRUE if caching is enabled

Usage

```
cwf_cache_enabled()
```

phrase_counts *Count phrase matches in CHILDES utterances (experimental)*

Description

Matches surface phrases in utterance text and outputs counts, plus dataset summary and run metadata. Supports simple wildcards in phrases: * (any chars), ? (one char). Normalization is per number of utterances.

Usage

```
phrase_counts(
  phrases,
  collection = NULL,
  language = NULL,
  corpus = NULL,
  age = NULL,
  sex = NULL,
  role = NULL,
  role_exclude = NULL,
  wildcard = FALSE,
  ignore_case = TRUE,
  normalize = FALSE,
  per_utts = 10000L,
  db_version = "current",
  cache = FALSE,
  cache_dir = NULL,
  output_file = NULL
)
```

Arguments

phrases	Character vector of phrases or patterns.
collection, language, corpus, age, sex, role, role_exclude	CHILDES filters.
wildcard	Logical; enable * and ? in phrases.
ignore_case	Logical; case-insensitive matching.
normalize	Logical; if TRUE, add per-N utterance rates.
per_utts	Integer; denominator for utterance rates (default 10000).
db_version	CHILDES DB version (recorded).
cache	Logical; cache CHILDES queries on disk.
cache_dir	Optional cache directory.
output_file	Optional .xlsx path; if NULL, returns a tibble.

Details

Tier targeting is not applied in phrase mode. Phrases are matched in the main utterance text. For tier-constrained contexts around words, use `contexts_for(..., mode = "word", tier = "mor")`.

Value

If output_file is NULL, returns a tibble of phrase counts; otherwise writes an Excel file and returns the file path (invisibly).

`word_counts` *Get word counts by speaker role*

Description

Reads a CSV with a `word` column or an in-memory character vector and writes an Excel file with Word_Frequencies, Dataset_Summary, File_Speaker_Summary, and Run_Metadata. If no word list is provided, all types in the selected slice are counted (FREQ-style "all words" mode).

Usage

```
word_counts(
  word_list_file = NULL,
  output_file,
  words = NULL,
  collection = NULL,
  language = NULL,
  corpus = NULL,
  age = NULL,
  sex = NULL,
  role = NULL,
  role_exclude = NULL,
  wildcard = FALSE,
  collapse = c("none", "stem"),
  part_of_speech = NULL,
  tier = c("main", "mor"),
  normalize = FALSE,
  per = 1000L,
  zipf = FALSE,
  include_patterns = NULL,
  exclude_patterns = NULL,
  sort_by = c("word", "frequency"),
  min_count = 0L,
  freq_ignore_special = TRUE,
  db_version = "current",
  cache = FALSE,
  cache_dir = NULL,
  ...
)
```

Arguments

word_list_file	Optional path to a CSV file with a column named word. If NULL and words is also NULL, all types in the slice are counted.
output_file	Path to the output .xlsx file.
words	Optional character vector of target words/patterns. Ignored if word_list_file is provided. If both are NULL, all types are counted.
collection	Optional CHILDES filter.
language	Optional CHILDES filter.
corpus	Optional CHILDES filter.
age	Optional numeric: single value or c(min, max) in months.
sex	Optional: "male" and/or "female".
role	Optional character vector of roles to include.
role_exclude	Optional character vector of roles to exclude.
wildcard	Logical; treat "%" as any number of characters and "_" as one character (token mode).
collapse	Either "none" or "stem". Using "stem" triggers token mode.
part_of_speech	Optional POS filter, e.g., c("n","v") (token mode).
tier	Which tier to count from: "main" or "mor".
normalize	Logical; if TRUE, add per-N rate columns.
per	Integer denominator for rates (for example 1000 for per-1k).
zipf	Logical; if TRUE, also add Zipf columns (log10 per-billion).
include_patterns	Optional character vector of CHILDES-style patterns, using "%" and "_" to restrict output to matching words (FREQ-style +s).
exclude_patterns	Optional character vector of CHILDES-style patterns to drop from the output.
sort_by	Final sort order: "word" (alphabetical) or "frequency" (descending Total).
min_count	Integer; drop rows with Total < min_count (after counting).
freq_ignore_special	Logical; if TRUE, drop "xxx", "www", and any word starting with 0, &, +, -, or # (FREQ default ignore rules).
db_version	CHILDES database version label to record in metadata.
cache	Logical; if TRUE, cache CHILDES queries on disk.
cache_dir	Optional cache directory when cache = TRUE.
...	Reserved for future extensions; currently unused.

Details

Uses exact type counts by default; switches to token mode when wildcards, stems, or POS filters are requested. Optional MOR-only tier.

Value

Invisibly returns `output_file` after writing the workbook.

Examples

```
## Not run:
# Minimal example (not run during R CMD check)
tmp_csv <- tempfile(fileext = ".csv")
write.csv(data.frame(word = c("the","go")), tmp_csv, row.names = FALSE)

out_file <- tempfile(fileext = ".xlsx")
word_counts(
  word_list_file = tmp_csv,
  output_file    = out_file,
  language       = "eng",
  corpus         = "Brown",
  age            = c(24, 26)
)

# All-words mode (no word list; counts every type in the slice)
out_all <- tempfile(fileext = ".xlsx")
word_counts(
  word_list_file = NULL,
  words          = NULL,
  output_file    = out_all,
  language       = "eng",
  corpus         = "Brown",
  age            = c(24, 26)
)

## End(Not run)
```

Index