

Package: wordbankr (via r-universe)

September 30, 2026

Type Package

Title Accessing the 'Wordbank' Database

Description Connecting to 'Wordbank' <<https://wordbank.stanford.edu>>, an open repository for developmental vocabulary data from the MacArthur-Bates Communicative Development Inventories (Frank et al. 2017 <[doi:10.1017/S0305000916000209](https://doi.org/10.1017/S0305000916000209)>). Data are read from a versioned dataset hosted on 'Redivis', so analyses can be pinned to a specific release.

Version 2.0.0

Depends R (>= 4.0)

License GPL-3

URL <https://langcog.github.io/wordbankr/>,
<https://github.com/langcog/wordbankr/>

BugReports <https://github.com/langcog/wordbankr/issues/>

Imports assertthat (>= 0.2.1), dplyr (>= 1.1.3), glue (>= 1.6.2),
jsonlite (>= 1.8.7), lifecycle (>= 1.0.3), purrr (>= 1.0.2),
quantregGrowth (>= 1.7-0), rlang (>= 1.1.1), robustbase (>= 0.99-0), stringr (>= 1.5.0), tidyr (>= 1.3.0)

Suggests redivis (>= 0.12.12), ggplot2, knitr, rmarkdown, testthat (>= 3.0.0)

Additional_repositories <https://langcog.r-universe.dev>

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

Config/roxygen2/version 8.1.0

RoxygenNote 7.3.3

Repository <https://mikabr.r-universe.dev>

Date/Publication 2026-09-30 17:02:40 UTC

RemoteUrl <https://github.com/langcog/wordbankr>

RemoteRef HEAD

RemoteSha 86df76ff835fae30374eb432a45d94b96a62786e

Contents

fit_aoa	2
fit_vocab_quantiles	3
get_administration_data	4
get_aoa	5
get_crossling_data	6
get_crossling_items	7
get_datasets	7
get_embeddings	8
get_instrument_data	9
get_instruments	10
get_item_data	10
summarise_items	11
wb_dataset	12

Index	13
--------------	-----------

fit_aoa	<i>Fit age of acquisition estimates for Wordbank data</i>
---------	---

Description

For each item in the input data, estimate its age of acquisition as the earliest age (in months) at which the proportion of children who understand/produce the item is greater than some threshold. The proportions used can be empirical or first smoothed by a model.

Usage

```
fit_aoa(  
  instrument_data,  
  measure = "produces",  
  method = "glm",  
  proportion = 0.5,  
  age_min = min(instrument_data$age, na.rm = TRUE),  
  age_max = max(instrument_data$age, na.rm = TRUE)  
)
```

Arguments

instrument_data	A data frame returned by get_instrument_data, which must have an "age" column and a "num_item_id" column.
measure	One of "produces" or "understands" (defaults to "produces").
method	A string indicating which smoothing method to use: empirical to use empirical proportions, glm to fit a logistic linear model, glmrob a robust logistic linear model (defaults to glm).

proportion	A number between 0 and 1 indicating threshold proportion of children.
age_min	The minimum age to allow for an age of acquisition. Defaults to the minimum age in instrument_data
age_max	The maximum age to allow for an age of acquisition. Defaults to the maximum age in instrument_data

Value

A data frame where every row is an item, the item-level columns from the input data are preserved, and the aoa column contains the age of acquisition estimates.

Examples

```
eng_ws_data <- get_instrument_data(language = "English (American)",
                                   form = "WS",
                                   items = c("item_1", "item_42"),
                                   administration_info = TRUE)
if (!is.null(eng_ws_data)) eng_ws_aoa <- fit_aoa(eng_ws_data)
```

fit_vocab_quantiles	<i>Fit quantiles to vocabulary sizes using quantile regression</i>
---------------------	--

Description

Fit quantiles to vocabulary sizes using quantile regression

Usage

```
fit_vocab_quantiles(vocab_data, measure, group, quantiles = "standard")
```

Arguments

vocab_data	A data frame returned by get_administration_data.
measure	A column of vocab_data with vocabulary values (production or comprehension).
group	(Optional) A column of vocab_data to group by.
quantiles	Either one of "standard" (default), "deciles", "quintiles", "quartiles", "median", or a numeric vector of quantile values.

Value

A data frame with the columns "language", "form", "age", group (if specified), "quantile", and measure, where measure is the fit vocabulary value for that quantile at that age.

Examples

```
eng_wg <- get_administration_data(language = "English (American)",
                                form = "WG",
                                include_demographic_info = TRUE)

if (!is.null(eng_wg)) {
  vocab_quantiles <- fit_vocab_quantiles(eng_wg, production)
  vocab_quantiles_sex <- fit_vocab_quantiles(eng_wg, production, sex)
  vocab_quartiles <- fit_vocab_quantiles(eng_wg, production, quantiles = "quartiles")
}
```

get_administration_data

Get the Wordbank by-administration data

Description

Get the Wordbank by-administration data

Usage

```
get_administration_data(
  language = NULL,
  form = NULL,
  filter_age = TRUE,
  include_demographic_info = FALSE,
  include_birth_info = FALSE,
  include_health_conditions = FALSE,
  include_language_exposure = FALSE,
  include_study_internal_id = FALSE,
  version = "current"
)
```

Arguments

language	An optional string specifying which language's administrations to retrieve.
form	An optional string specifying which form's administrations to retrieve.
filter_age	A logical indicating whether to filter the administrations to ones in the instrument's age range.
include_demographic_info	A logical indicating whether to include the child's demographic information (birth_order, caregiver_education, ethnicity, race, sex).
include_birth_info	A logical indicating whether to include the child's birth information (birth_weight, born_early_or_late, gestational_age, zygosity).

include_health_conditions	A logical indicating whether to include the child's health condition information (a nested dataframe under health_conditions with the column health_condition_name).
include_language_exposure	A logical indicating whether to include the child's language exposure information at time of administration (a nested dataframe under language_exposures with the columns language, exposure_percentage, age_of_first_exposure).
include_study_internal_id	A logical indicating whether to include the child's ID in the original study data.
version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A data frame where each row is a CDI administration and each column is a variable about the administration or the corresponding child, including which dataset_version it came from.

Examples

```
english_ws_admins <- get_administration_data("English (American)", "WS")
```

get_aoa	<i>Get cached age-of-acquisition estimates</i>
---------	--

Description

Age-of-acquisition estimates for every word item on every instrument, precomputed with [fit_aoa](#) (glm method, 50 data release. aoa is NA for items that do not reach the threshold within the instrument's age range.

Usage

```
get_aoa(language = NULL, form = NULL, measure = NULL, version = "current")
```

Arguments

language	An optional string specifying which language's estimates to retrieve.
form	An optional string specifying which form's estimates to retrieve.
measure	An optional string ("produces" or "understands") to filter by measure.
version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A data frame with one row per instrument item and measure: language, form, item_id, item_definition, category, uni_lemma, measure, aoa, dataset_version.

Examples

```
danish_aoa <- get_aoa(language = "Danish", form = "WS")
```

get_crossling_data	<i>Get item-by-age summary statistics for items across languages</i>
--------------------	--

Description

Get item-by-age summary statistics for items across languages

Usage

```
get_crossling_data(uni_lemmas, version = "current")
```

Arguments

uni_lemmas	A character vector of uni_lemmas.
version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A dataframe with a row for each combination of language, item, and age, and columns for summary statistics for the group: number of children (n_children), means (comprehension, production), standard deviations (comprehension_sd, production_sd); and item-level variables (item_id, definition, uni_lemma, lexical_category, lexical_class, dataset_version).

Examples

```
## Not run:
# downloads item-level data from every instrument, which takes minutes
crossling_data <- get_crossling_data(uni_lemmas = "dog")

## End(Not run)
```

get_crossling_items	<i>Get the uni_lemmas available in Wordbank</i>
---------------------	---

Description

Get the uni_lemmas available in Wordbank

Usage

```
get_crossling_items(version = "current")
```

Arguments

version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.
---------	--

Value

A data frame with the columns uni_lemma and dataset_version.

Examples

```
uni_lemmas <- get_crossling_items()
```

get_datasets	<i>Get the Wordbank data sources</i>
--------------	--------------------------------------

Description

Get the Wordbank data sources

Usage

```
get_datasets(  
  language = NULL,  
  form = NULL,  
  admin_data = FALSE,  
  version = "current"  
)
```

Arguments

language	An optional string specifying which language’s datasets to retrieve.
form	An optional string specifying which form’s datasets to retrieve.
admin_data	A logical indicating whether to include the number of administrations in the dataset.
version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A data frame where each row is a particular dataset and its characteristics, including which dataset_version it came from.

Examples

```
english_ws_datasets <- get_datasets("English (American)", "WS")
```

get_embeddings	<i>Get multilingual item embeddings</i>
----------------	---

Description

Semantic embeddings for every unique word item definition, computed with Google’s multilingual gemini-embedding-001 model (768 dimensions). All languages share one embedding space, so cosine similarities are meaningful both within and across languages.

Usage

```
get_embeddings(language = NULL, version = "current")
```

Arguments

language	An optional string specifying which language’s embeddings to retrieve.
version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A data frame with one row per unique item definition: language, item_definition, embedding (a list-column of numeric vectors), and dataset_version.

Examples

```
danish_embeddings <- get_embeddings(language = "Danish")
```

get_instruments	<i>Get the Wordbank instruments</i>
-----------------	-------------------------------------

Description

Get the Wordbank instruments

Usage

```
get_instruments(version = "current")
```

Arguments

version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.
---------	--

Value

A data frame where each row is a CDI instrument and each column is a variable about the instrument (instrument_id, language, form, form_type, age_min, age_max, has_grammar, unilemma_coverage, dataset_version).

Examples

```
instruments <- get_instruments()
```

get_item_data	<i>Get the Wordbank by-item data</i>
---------------	--------------------------------------

Description

Get the Wordbank by-item data

Usage

```
get_item_data(language = NULL, form = NULL, version = "current")
```

Arguments

language	An optional string specifying which language's items to retrieve.
form	An optional string specifying which form's items to retrieve.
version	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A data frame where each row is a CDI item and each column is a variable about it: `item_id`, `item_kind`, `item_definition`, `english_gloss`, `language`, `form`, `form_type`, `category`, `lexical_category`, `lexical_class`, `complexity_category`, `uni_lemma`, `dataset_version`.

Examples

```
english_ws_items <- get_item_data("English (American)", "WS")
```

<code>summarise_items</code>	<i>Get item-by-age summary statistics</i>
------------------------------	---

Description

Get item-by-age summary statistics

Usage

```
summarise_items(item_data, version = "current")
```

Arguments

<code>item_data</code>	A dataframe as returned by <code>get_item_data()</code> .
<code>version</code>	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.

Value

A dataframe with a row for each combination of item and age, and columns for summary statistics for the group: number of children (`n_children`), means (`comprehension`, `production`), standard deviations (`comprehension_sd`, `production_sd`); also retains item-level variables from `lang_items` (`item_id`, `item_definition`, `uni_lemma`, `lexical_category`) and `dataset_version`.

Examples

```
italian_items <- get_item_data(language = "Italian", form = "WG")
if (!is.null(italian_items)) {
  italian_dog <- dplyr::filter(italian_items, uni_lemma == "dog")
  italian_dog_summary <- summarise_items(italian_dog)
}
```

`wb_dataset`*The Wordbank dataset on Redivis*

Description

Returns a reference to the Wordbank Redivis dataset (<https://stanford.redivis.com/datasets/627v-9ewzpdvz0>).

Usage

```
wb_dataset(version = "current")
```

Arguments

<code>version</code>	A string specifying which version of the Wordbank dataset to use, e.g. "v1.2" to pin a released version for reproducibility. Defaults to "current", the most recent release.
----------------------	--

Value

A redivis dataset reference.

Index

`fit_aoa`, [2](#), [5](#)
`fit_vocab_quantiles`, [3](#)

`get_administration_data`, [4](#)
`get_aoa`, [5](#)
`get_crossling_data`, [6](#)
`get_crossling_items`, [7](#)
`get_datasets`, [7](#)
`get_embeddings`, [8](#)
`get_instrument_data`, [9](#)
`get_instruments`, [10](#)
`get_item_data`, [10](#)

`summarise_items`, [11](#)

`wb_dataset`, [12](#)