

# Package: wordbankr (via r-universe)

August 31, 2024

**Type** Package

**Title** Accessing the Wordbank Database

**Description** Connecting to Wordbank, an open repository for developmental vocabulary data. For more information on the underlying data, see <<http://wordbank.stanford.edu>>.

**Version** 1.0.3.9000

**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), DBI (>= 1.1.3), dbplyr (>= 2.3.4),  
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), RMySQL (>= 0.10.26), robustbase (>= 0.99-0), stringr (>= 1.5.0), tidyr (>= 1.3.0)

**Suggests** ggplot2, knitr, rmarkdown

**VignetteBuilder** knitr

**RoxygenNote** 7.3.1

**Encoding** UTF-8

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

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

**RemoteRef** HEAD

**RemoteSha** 35077eb3652f58367f85b83c74dc58b36a6c4ad0

## Contents

connect_to_wordbank . . . . .	2
fit_aoa . . . . .	3

fit_vocab_quantiles . . . . .	4
get_administration_data . . . . .	4
get_crossling_data . . . . .	6
get_crossling_items . . . . .	6
get_datasets . . . . .	7
get_instruments . . . . .	8
get_instrument_data . . . . .	8
get_item_data . . . . .	9
get_wordbank_args . . . . .	10
summarise_items . . . . .	10
<b>Index</b>	<b>12</b>

---

connect_to_wordbank	<i>Connect to the Wordbank database</i>
---------------------	---

---

## Description

Connect to the Wordbank database

## Usage

```
connect_to_wordbank(db_args = NULL)
```

## Arguments

db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).
---------	---

## Value

A src object which is connection to the Wordbank database.

## Examples

```
src <- connect_to_wordbank()
```

fit\_aoa

*Fit age of acquisition estimates for Wordbank data***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 <code>get_instrument_data</code> , 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: <code>empirical</code> to use empirical proportions, <code>glm</code> to fit a logistic linear model, <code>glmrob</code> a robust logistic linear model (defaults to <code>glm</code> ).
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 <code>instrument_data</code>
age_max	The maximum age to allow for an age of acquisition. Defaults to the maximum age in <code>instrument_data</code>

**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     *Fit quantiles to vocabulary sizes using quantile regression*

---

### 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 <code>get_administration_data</code> .
measure	A column of <code>vocab_data</code> with vocabulary values (production or comprehension).
group	(Optional) A column of <code>vocab_data</code> 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,
  db_args = NULL
)

```

**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 valid age range for their instrument.
include_demographic_info	A logical indicating whether to include the child's demographic information (birth_order, ethnicity, race, sex, caregiver_education).
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_proportion, age_of_first_exposure).
include_study_internal_id	A logical indicating whether to include the child's ID in the original study data.
db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).

**Value**

A data frame where each row is a CDI administration and each column is a variable about the administration (data\_id, date\_of\_test, age, comprehension, production, is\_norming), the dataset it's from (dataset\_name, dataset\_origin\_name, language, form, form\_type), and information about the child as described in the parameter specification.

**Examples**

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

---

get\_crossling\_data      *Get item-by-age summary statistics for items across languages*

---

**Description**

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

**Usage**

```
get_crossling_data(uni_lemmas, db_args = NULL)
```

**Arguments**

uni_lemmas	A character vector of uni_lemmas.
db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).

**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).

**Examples**

```
crossling_data <- get_crossling_data(uni_lemmas = "dog")
```

---

get\_crossling\_items      *Get the uni\_lemmas available in Wordbank*

---

**Description**

Get the uni\_lemmas available in Wordbank

**Usage**

```
get_crossling_items(db_args = NULL)
```



---

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

---

**Description**

Get the Wordbank instruments

**Usage**

```
get_instruments(db_args = NULL)
```

**Arguments**

db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).
---------	---

**Value**

A data frame where each row is a CDI instrument and each column is a variable about the instrument (instrument\_id, language, form, age\_min, age\_max, has\_grammar).

**Examples**

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

---

get_instrument_data	<i>Get the Wordbank administration-by-item data</i>
---------------------	---

---

**Description**

Get the Wordbank administration-by-item data

**Usage**

```
get_instrument_data(  
  language,  
  form,  
  items = NULL,  
  administration_info = FALSE,  
  item_info = FALSE,  
  db_args = NULL,  
  ...  
)
```



**Arguments**

language	A string of the instrument's language (insensitive to case and whitespace).
form	A string of the instrument's form (insensitive to case and whitespace).
items	A character vector of column names of instrument_table of items to extract. If not supplied, defaults to all the columns of instrument_table.
administration_info	Either a logical indicating whether to include administration data or a data frame of administration data (as returned by get_administration_data).
item_info	Either a logical indicating whether to include item data or a data frame of item data (as returned by get_item_data).
db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).
...	<[‘dynamic-dots’][rlang::dyn-dots]> Arguments passed to get_administration_data().

**Value**

A data frame where each row contains the values (value, produces, understands) of a given item (item\_id) for a given administration (data\_id), with additional columns of variables about the administration and item, as specified.

**Examples**

```
eng_ws_data <- get_instrument_data(language = "English (American)",
                                  form = "WS",
                                  items = c("item_1", "item_42"),
                                  item_info = TRUE)
```

---

get\_item\_data                      *Get the Wordbank by-item data*

---

**Description**

Get the Wordbank by-item data

**Usage**

```
get_item_data(language = NULL, form = NULL, db_args = NULL)
```

**Arguments**

language	An optional string specifying which language's items to retrieve.
form	An optional string specifying which form's items to retrieve.
db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).

**Value**

A data frame where each row is a CDI item and each column is a variable about it: `item_id`, `item_kind` (e.g. `word`, `gestures`, `word_endings`), `item_definition`, `english_gloss`, `language`, `form`, `form_type`, `category` (meaning-based group as shown on the CDI form), `lexical_category`, `lexical_class`, `complexity_category`, `uni_lemma`).

**Examples**

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

---

<code>get_wordbank_args</code>	<i>Get database connection arguments</i>
--------------------------------	--

---

**Description**

Get database connection arguments

**Usage**

```
get_wordbank_args()
```

**Value**

List of database connection arguments: `host`, `db_name`, `username`, `password`

**Examples**

```
get_wordbank_args()
```

---

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

---

**Description**

Get item-by-age summary statistics

**Usage**

```
summarise_items(item_data, db_args = NULL)
```

**Arguments**

item_data	A dataframe as returned by <code>get_item_data()</code> .
db_args	List with arguments to connect to wordbank mysql database (host, dbname, user, and password).

**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`).

**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)
}
```

# Index

`connect_to_wordbank`, 2

`fit_aoa`, 3

`fit_vocab_quantiles`, 4

`get_administration_data`, 4

`get_crossling_data`, 6

`get_crossling_items`, 6

`get_datasets`, 7

`get_instrument_data`, 8

`get_instruments`, 8

`get_item_data`, 9

`get_wordbank_args`, 10

`summarise_items`, 10