Package ‘wordbankr’

November 14, 2020

Type Package

Title Accessing the Wordbank Database

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

Version 0.3.1

Depends R (>= 4.0)

License GPL-3

URL https://github.com/langcog/wordbankr

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

Imports assertthat (>= 0.2.1), DBI (>= 1.1.0), dbplyr (>= 1.4.4), dplyr (>= 1.0.2), purrr (>= 0.3.4), quantregGrowth (>= 0.4), rlang (>= 0.4.8), RMySQL (>= 0.10.20), robustbase (>= 0.93), stringr (>= 1.4.0), tidyr (>= 1.1.2)

Suggests ggplot2, knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 7.1.1

Encoding UTF-8

NeedsCompilation no

Author Mika Braginsky [aut, cre],
Daniel Yurovsky [ctb],
Michael Frank [ctb],
Danielle Kellier [ctb]

Maintainer Mika Braginsky <mika.br@gmail.com>

Repository CRAN

Date/Publication 2020-11-13 23:10:02 UTC

R topics documented:

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 `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

```
## Not run:
eng_ws_data <- get_instrument_data(language = "English (American)",
                                    form = "WS",
                                    items = c("item_1", "item_42"),
                                    administrations = TRUE)
eng_ws_aoa <- fit_aoa(eng_ws_data)

## End(Not run)
```

`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 = NULL, 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

```
## Not run:
eng_ws <- get_administration_data("English (American)", "WS")
fit_vocab_quantiles(eng_ws, production)
fit_vocab_quantiles(eng_ws, production, sex)
fit_vocab_quantiles(eng_ws, production, quantiles = "quartiles")

## End(Not run)
```

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,
  original_ids = FALSE,
  mode = "remote"
)
```

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.
`original_ids`	A logical indicating whether to include the original ids provided by data contributors. Wordbank provides no guarantees about the structure or uniqueness of these ids. Use at your own risk!
`mode`	A string indicating connection mode: one of `"local"`, or `"remote"` (defaults to `"remote"`).

Value

A data frame where each row is a CDI administration and each column is a variable about the administration (`data_id`, `age`, `comprehension`, `production`), its instrument (`language`, `form`), its child (`birth_order`, `ethnicity`, `sex`, `mom_ed`, `zygosity`), and its dataset source (`source_name`, `source_dataset`, `norming`, `longitudinal`). Also includes an `original_id` column if the `original_ids` flag is `TRUE`.

Examples

```
## Not run:
english_ws_admins <- get_administration_data("English (American)", "WS")
all_admins <- get_administration_data()

## End(Not run)
```

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, mode = "remote")
```

Arguments

uni_lemmas	A character vector of uni_lemmas.
mode	A string indicating connection mode: one of "local", or "remote" (defaults to "remote").

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

```
## Not run:
crossling_data <- get_crossling_data(uni_lemmas = c("hat", "nose"))

## End(Not run)
```

`get_crossling_items` *Get the uni_lemmas available in Wordbank*

Description

Get the uni_lemmas available in Wordbank

Usage

```
get_crossling_items(mode = "remote")
```

Arguments

`mode`	A string indicating connection mode: one of `"local"`, or `"remote"` (defaults to `"remote"`).

Value

A data frame with the column `uni_lemma`.

Examples

```
## Not run:
uni_lemmas <- get_crossling_items()

## End(Not run)
```

`get_instruments` *Get the Wordbank instruments*

Description

Get the Wordbank instruments

Usage

```
get_instruments(mode = "remote")
```

Arguments

`mode`	A string indicating connection mode: one of `"local"`, or `"remote"` (defaults to `"remote"`).

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

```
## Not run:
instruments <- get_instruments()

## End(Not run)
```

`get_instrument_data` *Get the Wordbank administration-by-item data*

Description

Get the Wordbank administration-by-item data

Usage

```
get_instrument_data(
  language,
  form,
  items = NULL,
  administrations = FALSE,
  iteminfo = FALSE,
  mode = "remote"
)
```

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`.

`administrations` Either a logical indicating whether to include administration data or a data frame of administration data (from `get_administration_data`).

`iteminfo` Either a logical indicating whether to include item data or a data frame of item data (from `get_item_data`).

`mode` A string indicating connection mode: one of `"local"`, or `"remote"` (defaults to `"remote"`).

Value

A data frame where each row is the result (`value`) of a given item (`num_item_id`) for a given administration (`data_id`), with additional columns of variables about the administration and item, if specified.

Examples

```
## Not run:
eng_ws_data <- get_instrument_data(language = "English (American)",
                                   form = "WS",
                                   items = c("item_1", "item_42"))

## End(Not run)
```

`get_item_data` *Get the Wordbank by-item data*

Description

Get the Wordbank by-item data

Usage

```
get_item_data(language = NULL, form = NULL, mode = "remote")
```

Arguments

`language`	An optional string specifying which language's items to retrieve.
`form`	An optional string specifying which form's items to retrieve.
`mode`	A string indicating connection mode: one of `"local"`, or `"remote"` (defaults to `"remote"`).

Value

A data frame where each row is a CDI item and each column is a variable about it (`item_id`, `definition`, `language`, `form`, `type`, `category`, `lexical_category`, `lexical_class`, `uni_lemma`, `complexity_category`, `num_item_id`).

Examples

```
## Not run:
english_ws_items <- get_item_data("English (American)", "WS")
all_items <- get_item_data()

## End(Not run)
```

get_sources *Get the Wordbank data sources*

Description

Get the Wordbank data sources

Usage

```
get_sources(language = NULL, form = NULL, admin_data = FALSE, mode = "remote")
```

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 summary-level statistics on the administrations within a dataset.
mode	A string indicating connection mode: one of "local", or "remote" (defaults to "remote").

Value

A data frame where each row is a particular dataset and its characteristics: dataset id and name (source_id, name, dataset), language (instrument_language), form (instrument_form), contributor and affiliated institution (contributor), provided citation (citation), whether dataset includes longitudinal participants (longitudinal), and licensing information (license). Also includes summary statistics on a dataset if the (administrations) flag is TRUE: number of children (n_children) and age range (age_min, age_max).

Examples

```
## Not run:
english_ws_sources <- get_sources(language = "English (American)",
                                  form = "WS",
                                  admin_data = TRUE)

## End(Not run)
```

summarise_items *Get item-by-age summary statistics*

Description

Get item-by-age summary statistics

Usage

```
summarise_items(lang_items, mode = "remote")
```

Arguments

lang_items	A dataframe as returned by get_item_data().
mode	A string indicating connection mode: one of "local", or "remote" (defaults to "remote").

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, definition, uni_lemma, lexical_category, lexical_class).

Examples

```
## Not run:
italian_dog <- get_item_data(language = "Italian", form = "WG") %>%
  dplyr::filter(uni_lemma == "dog")
italian_dog_summary <- summarise_items(italian_dog)

## End(Not run)
```

Index