Package 'multicastR'

October 13, 2022

Type Package

Title A Companion to the Multi-CAST Collection

Version 2.0.0

URL https://multicast.aspra.uni-bamberg.de/

Description Provides a basic interface for accessing annotation data from the Multi-CAST collection, a database of spoken natural language texts edited by Geoffrey Haig and Stefan Schnell. The collection draws from a diverse set of languages and has been annotated across multiple levels. Annotation data is downloaded on request from the servers of the University of Bamberg. See the Multi-CAST website <https://multicast.aspra.uni-bamberg.de/> for more information and a list of related publications.

License CC BY 4.0

Encoding UTF-8

LazyData true

Depends R (>= 3.0.0),

Imports

RoxygenNote 7.1.1

Suggests

NeedsCompilation no

Author Nils Norman Schiborr [aut, cre]

Maintainer Nils Norman Schiborr <nils-norman.schiborr@uni-bamberg.de>

Repository CRAN

Date/Publication 2021-02-22 19:20:02 UTC

R topics documented:

`mc_clauses` *Count clauses in a multicastR table*

Description

`mc_clauses` counts the number of clause units (bounded by the <##> or <#> GRAID annotation symbols) in a multicastR table.

Usage

```
mc_clauses(x, bytext = FALSE, printToConsole = FALSE)
```

Arguments

`x`	A `data.frame` in multicastR format. This table minimally requires the `corpus` and `graid` columns with the names of the corpora and the GRAID annotation values, respectively, as well as the `text` column if `bytext` is set to `TRUE`.
`bytext`	Logical. If `FALSE`, calculate the number of clause units for each corpus. If `TRUE`, count for each text separately. `FALSE` by default.
`printToConsole`	Logical. If `TRUE`, prints the table to the console (using `message`). `FALSE` by default.

Value

A `data.frame` with the `corpus`, `text` (if `bytext` is `TRUE`), the number of valid clause units in each corpus (`nClause`), the total number of clause units (`nAll`), the number of clause units not analyzed (`nNC`), and the percentage the later make up of the total (`pNC`).

See Also

`multicast`, `mc_index`, `mc_metadata`, `mc_referents`, `mc_clauses`

Examples

```
## Not run:
  # count clause units in the most recent version
  # of the Multi-CAST data, by corpus
  n <- mc_clauses(multicast())

  # count by text instead
  m <- mc_clauses(multicast(), bytext = TRUE)
```

```
  # number of clauses units in the whole collection
  sum(n$nClauses)

## End(Not run)
```

`mc_index` *Access the Multi-CAST version index*

Description

`mc_index` downloads a tabular index of the versions of the Multi-CAST corpus data from the servers of the University of Bamberg. The value in the leftmost `version` column may be passed to the `multicast` method for access to earlier versions of the annotations.

Usage

```
mc_index()
```

Value

A `data.frame` with five columns:

`[, 1]` `version` Version key. Used for the `vkey` argument of other functions in this package.

`[, 2]` `date` Publication date in YYYY-MM-DD format.

`[, 3]` `corpora` Number of corpora (languages).

`[, 4]` `texts` Number of texts.

`[, 5]` `size` Total file size in kilobytes.

See Also

`multicast`, `mc_metadata`, `mc_referents`, `mc_clauses`

Examples

```
## Not run:
  # retrieve version index
  mc_index()

## End(Not run)
```

`mc_metadata` *Access the Multi-CAST metadata*

Description

`mc_metadata` downloads a table with metadata on the texts and speakers in the Multi-CAST collection from the servers of the University of Bamberg.

Usage

```
mc_metadata(vkey = NULL)
```

Arguments

`vkey`	A four-digit number specifying the requested version of the metadata. Must be one of the version keys listed in the first column of `mc_index`, or empty. If empty, the most recent version of the metadata is retrieved automatically.

Value

A `data.frame` containing metadata on the Multi-CAST collection. The table has the following eight columns:

`[, 1]` `corpus` The name of the corpus.

`[, 2]` `text` The title of the text.

`[, 3]` `type` The text type, either `TN` 'traditional narrative', `AN` 'autobiographical narrative', or `SN` 'stimulus-based narrative'.

`[, 4]` `recorded` The year (YYYY) the text was recorded.

`[, 5]` `speaker` The identifier for the speaker.

`[, 6]` `gender` The speaker's gender.

`[, 7]` `age` The speaker's age at the time of recording. Approximate values are prefixed with a `c`.

`[, 8]` `born` The speaker's birth year (YYY). Approximate values are prefixed with a `c`.

See Also

`multicast`, `mc_index`, `mc_referents`, `mc_clauses`

Examples

```
## Not run:
  # retrieve the most recent version of the Multi-CAST metadata
  mc_metadata()

  # retrieve the lists of referents published in January 2021
  mc_metadata(2101)
```

```
  # join the metadata to a table with annotation values
  mc <- multicast()
  merge(mc, mc_metadata(),
        by = c("corpus", "text"))

## End(Not run)
```

`mc_referents` *Access the Multi-CAST list of referents*

Description

`mc_referents` downloads a tabular list of all discourse referents occuring in those texts in the Multi-CAST collection that have been annotated with the RefIND scheme (Schiborr et al. 2018). The data are downloaded from the servers of University of Bamberg.

Usage

```
mc_referents(vkey = NULL)
```

Arguments

`vkey`	A four-digit number specifying the requested version of the list of referents. Must be one of the version keys listed in the first column of `mc_index`, or empty. If empty, the most recent version of the list of referents is retrieved automatically. Note that the first annotations with RefIND were added with version 1905 (May 2019) of Multi-CAST, and hence no lists of referents exist for earlier versions (i.e. 1505 and 1606).

Value

A `data.frame` containing a list of referents for all texts with RefIND annotations in the Multi-CAST collection. The table has the following eight columns:

`[, 1]` `corpus` The name of the corpus.

`[, 2]` `text` The title of the text.

`[, 3]` `refind` The four-digit referent index, unique to each referent in a text.

`[, 4]` `label` The label used for the referent.

`[, 5]` `description` A short description of the referent.

`[, 6]` `class` The semantic class of the referent. Legend: `hum` = human, `anm` = animate, `inm` = inanimate, `bdp` = body part, `mss` = mass, `loc` = location, `tme` = time, `abs` = abstract.

`[, 7]` `relations` Relations of the referent to other referents. Legend: `<` = set member of (partial co-reference), `>` = includes (split antecedence), `M` = part-whole.

`[, 8]` `notes` Annotators' notes on the referent and its properties.

See Also

multicast, mc_index, mc_metadata, mc_clauses

Examples

```
## Not run:
  # retrieve the most recent version of the Multi-CAST list of referents
  mc_referents()

  # retrieve the lists of referents published in January 2021
  mc_referents(2021)

  # join the list of referents to a table with annotation values
  mc <- multicast()
  merge(mc, mc_referents(),
        by = c("corpus", "text", "refind"),
        all.x = TRUE)

## End(Not run)
```

---

multicast *Access Multi-CAST annotation data*

Description

multicast downloads corpus data from the Multi-CAST collection (Haig & Schnell 2015) from the servers of the University of Bamberg. As the Multi-CAST collection is continuously evolving through the addition of further data sets and the revision of older annotations, the multicast function takes an optional argument vkey to select earlier versions of the annotation data, ensuring scientific accountability and the reproducibility of results.

Usage

```
multicast(vkey = NULL)
```

Arguments

vkey	A four-digit number specifying the requested version of the metadata. Must be one of the version keys listed in the first column of mc_index, or empty. If empty, the most recent version of the metadata is retrieved automatically.

Value

A data.frame with eleven columns:

[, 1] corpus The name of the corpus.

[, 2] text The name of the text.

`[, 3]` `uid` The utterance identifier. Uniquely identifies an utterance within a text.

`[, 4]` `gword` Grammatical words. The tokenized utterances in the object language.

`[, 5]` `gloss` Morphological glosses following the Leipzig Glossing Rules.

`[, 6]` `graid` Annotations with the GRAID scheme (Haig & Schnell 2014).

`[, 7]` `gform` The form symbol of a GRAID gloss.

`[, 8]` `ganim` The person-animacy symbol of a GRAID gloss.

`[, 9]` `gfunc` The function symbol of a GRAID gloss.

`[, 10]` `refind` Referent tracking using the RefIND scheme (Schiborr et al. 2018).

`[, 11]` `isnref` Annotations of the information status of newly introduced referents.

Licensing

Citing Multi-CAST

Data from the Multi-CAST collection should be cited as:

- Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilinguial Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/) (Accessed *date*.)

If for some reason you need to cite this package specifically, please refer to `citation(multicastR)`.

References

- Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators.* Version 7.0. (https://multicast.aspra.uni-bamberg.de/#annotations)
- Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. *RefIND – Referent Indexing in Natural-language Discourse: Annotation guidelines.* Version 1.1. (https://multicast.aspra.uni-bamberg.de/#annotations)

See Also

mc_index, mc_metadata, mc_referents, mc_clauses

Examples

```
## Not run:
  # retrieve and print the most recent version of the
  # Multi-CAST annotations
  multicast()

  # retrieve the version of the annotation data published
  # in January 2021
```

```
  multicast(2021)

## End(Not run)
```

`multicastR` *multicastR: A companion to the Multi-CAST collection.*

Description

The `multicastR` package provides a basic interface for accessing the annotated corpus data in the Multi-CAST collection (edited by Geoffrey Haig and Stefan Schnell), a database of spoken natural language texts that draws from a diverse set of languages. The corpus data are downloaded on command from the servers of the University of Bamberg via the `multicast` method. Details on the Multi-CAST project and a list of publications can be found online at https://multicast.aspra.uni-bamberg.de/.

Licensing

The Multi-CAST annotation data accessed by this package are published under a *Create Commons Attribution 4.0 International* (CC-BY 4.0) licence (https://creativecommons.org/licenses/by-sa/4.0/). Please refer to the Multi-CAST website for information on how to give proper credit to its contributors.

Citing Multi-CAST

Data from the Multi-CAST collection should be cited as:

- Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilinguial Corpus of Annotated Spoken Texts*. (https://multicast.aspra.uni-bamberg.de/) (Accessed *date*.)

If for some reason you need to cite this package specifically, please refer to `citation(multicastR)`.

See Also

`multicast`, `mc_index`, `mc_metadata`, `mc_referents`, `mc_clauses`

Index